Biologia molecular e evolução. Segunda edição.

Book · October 2012

CITATIONS
7

READS
14,576

2 authors:

Sergio R. Matioli
University of São Paulo
70 PUBLICATIONS **1,212** CITATIONS
SEE PROFILE

Marie-Anne Van Sluys
University of São Paulo
390 PUBLICATIONS **9,414** CITATIONS
SEE PROFILE

BIOLOGIA MOLECULAR E EVOLUÇÃO

2ª edição

Editores:
Sergio Russo Matioli
Flora M. C. Fernandes

SBG
Holos, Editora

Apoio Cultural:

Biologia Molecular e Evolução
(2a Edição)

Sergio Russo Matioli
Flora Maria de Campos Fernandes
(editores)

Sergio Russo Matioli. Departamento de Genética e Biologia Evolutiva, Instituto de Biociências, Universidade de São Paulo. Caixa Postal 11.461, 05422-970 São Paulo SP, BRASIL. Endereço eletrônico: srmatiol@ib.usp.br Fone: +55.11.3091.7552.

Flora Maria de Campos Fernandes. Departamento de Biologia Geral, Instituto de Biologia, Universidade Federal da Bahia, Salvador BA, BRASIL. Endereço eletrônico: florapangea@gmail.com

Site do livro: https://srmatioli.ib.usp.br/biolmolevol/

Dados Internacionais de Catalogação da Publicação (CIP)

575
M433b Matioli, Sergio Russo (ed.).
Biologia Molecular e Evolução / editores Sergio Russo Matioli e Flora Maria de Campos Fernandes. Ribeirão Preto: Holos, Editora / Sociedade Brasileira de Genética. 2012.
256 p.; il.; 28 cm --.

1. Genética. 2. Biologia Molecular.
3. Evolução. I. Título.

ISBN 85-86699-75-7
9788586 699757 C.D.U.

Revisão
Flora Maria de Campos Fernandes
José Geraldo Aquino Assis
Vera Cristina Silva

Arte Final
Sergio Russo Matioli

2012

Holos, Editora Ltda-ME
Av. Coronel Fernando Ferreira Leite 102
14.026-020 Ribeirão Preto SP
tele: 0.++16.3234.8083 / fax: 016.3234.8084
holos@holoseditora.com.br

www.holoseditora.com.br

Sumário

Apresentação 7
Prof. Antonio Brito da Cunha

Prefácio 9
Flora Maria de Campos Fernandes e Sergio Russo Matioli

Prefácio da primeira edição 11
Sergio Russo Matioli

Capítulo 1. Origem da vida: um tempo curto para uma experiência bem-sucedida 13
Carlos Frederico Martins Menck, Eduardo Gorab e Mariana Cabral de Oliveira

Capítulo 2. O mundo de RNA e a origem da complexidade da vida 21
Mariana Cabral de Oliveira e Carlos Frederico Martins Menck

Capítulo 3. Genoma não codificante — uma breve introdução 33
Alysson Renato Muotri e Cassiano Carromeu

Capítulo 4. O papel da interferência de RNA na célula eucariótica 39
Stephano Spanó Mello, Luciana Nogueira de Sousa Andrade e Carlos Frederico Martins Menck

Capítulo 5. Estabilidade do material genético: mutagênese e reparo 45
Luis Eduardo Soares Netto e Carlos Frederico Martins Menck

Capítulo 6. Sexo, por quê? 55
Sergio Russo Matioli e Anita Wajntal

Capítulo 7. Taxas de evolução e relógios moleculares 61
Daniela Calcagnotto

Capítulo 8. Evolução dos genes nucleares de RNA ribossômico 75
Eduardo Gorab

Capítulo 9. O genoma instável, sequências genéticas móveis 79
Marie-Anne Van Sluys, Nathalia de Setta, Katia C. Scortecci e Ana Paula Pimentel Costa

Capítulo 10. Evolução dos genomas 91
Laila Alves Nahum

Capítulo 11. Biologia evolutiva do desenvolvimento 105
Luis Paulo de Moura Andrioli

Capítulo 12. Reconstrução filogenética. Introdução e o método da máxima parcimônia 113
Cristina Yumi Miyaki, Cláudia A. de Moraes Russo e Sergio Luiz Pereira

Capítulo 13. Reconstrução filogenética: Métodos geométricos 123
Cláudia A. de Moraes Russo, Cristina Yumi Miyaki, Sergio Luiz Pereira

Capítulo 14. Reconstrução filogenética: Métodos probabilísticos 133
Sergio Luiz Pereira, Cristina Yumi Miyaki, Cláudia A. de Moraes Russo

Capítulo 15. Reconstrução filogenética: Inferência bayesiana 147
Sergio Luiz Pereira

Capítulo 16. Como escolher genes para problemas filogenéticos específicos 157
Claudia A. M. Russo, Carolina Moreira Voloch e Carlos G. Schrago

Capítulo 17. Polimorfismos de isozimas 165
Vera Nisaka Solferini e Denise Selivon Scheepmaker

Capítulo 18. RFLP: O emprego de enzimas de restrição para a detecção de polimorfismos no DNA 171
Maria Cristina Arias e Maria Elena Infante-Malachias

Capítulo 19. Métodos baseados em PCR para análise de polimorfismos de ácidos nucléicos 181
Sergio Russo Matioli e Maria Rita dos Santos e Passos-Bueno

Capítulo 20. Introdução às árvores genealógicas e à teoria da coalescência 191
Flora Maria de Campos Fernandes

Capítulo 21. Análise filogeográfica 197
Haydée A. Cunha e Antonio M. Solé-Cava

Capítulo 22. A genética e a conservação da Natureza 217
Antonio M. Solé-Cava e Haydée A. Cunha

Glossário 239
Sergio Russo Matioli

Índice Remissivo 247

Apresentação

O progresso no conhecimento sobre a estrutura do material genético, de sua função no desenvolvimento e na fisiologia dos organismos e de seu papel nos processos evolutivos tem ocorrido num ritmo impressionante e acelerado. Seria espantoso, para alguém formado na década de 1940, por exemplo, comparar o que se sabia então nessa área com o que se sabe hoje em dia. Felizmente, a ciência brasileira, graças a algumas Universidades e a órgãos nacionais e estrangeiros financiadores da atividade científica, está conseguindo acompanhar e participar desse processo. Este livro é uma demonstração dessas considerações. Todos os autores deste livro ensinam e pesquisam no Brasil e tratam em seus capítulos com muita clareza e atualidade de muitos aspectos da evolução e da adaptação dos organismos no nível molecular.

Em uma apresentação do livro como um todo, não seria justo salientar este ou aquele aspecto do texto, pois o importante é seu conjunto, próprio de um livro em padrão internacional. Mais justo a fazer é recomendar a sua leitura e o seu estudo a todos os interessados em evolução e, especialmente, no nível molecular.

O organizador e todos os autores estão de parabéns pela ótima realização desta obra que será responsável, entre nós, por grande parte do progresso dos estudos das áreas tratadas. A eles os meus cumprimentos e aos leitores, minhas felicitações.

Antonio Brito da Cunha
Professor Emérito do Instituto de Biociências da Universidade de São Paulo

Página deixada em branco

Prefácio à segunda edição

Bem vindos!

Chegarmos até aqui não foi, digamos, uma tarefa fácil. O sonho e a vontade desta segunda edição, ampliada e atualizada, de Biologia Molecular e Evolução, nos fez trabalhar muito, mas com muito prazer. Mais de dez anos depois da publicação da primeira edição, todos tivemos, editores e autores, acréscimos de tarefas em nossas carreiras acadêmicas, o que tornou nosso tempo cada vez mais escasso. Por outro lado, a experiência de todos aumentou e isso somente pôde acrescentar um brilho especial a essa segunda edição. Nosso objetivo continua o mesmo da primeira edição: levar a estudantes e profissionais informações importantes e em nossa língua pátria a respeito da biologia molecular e das suas relações com a evolução biológica. A primeira edição esgotou rapidamente e foi com grande satisfação que vimos exemplares sendo utilizados e disputados por todo o Brasil. Sim, o objetivo havia sido atingido, mas não podíamos parar por ali. Assim, recrutamos novamente nossos colaboradores, desta vez acrescentando outros mais, para dar continuidade a nosso sonho. Somos muito gratos a todos vocês, pela presteza e profissionalismo. Queremos agradecer também ao Dr José Geraldo Aquino Assis pelo seu minucioso trabalho de revisão científica. Agradecemos também aos amigos Dr Dalton de Souza Amorim e Dra Vera Cristina da Silva, pelo seu empenho fantástico nas sugestões e correções muitas que fizeram nesta obra. Queremos também deixar registrado que esse livro somente pôde ter um preço bem acessível para os leitores graças ao apoio financeiro das empresas listadas na contracapa.

Mais uma vez, para vocês, queridos estudantes, colegas e amigos, Biologia Molecular e Evolução repaginado. Obrigada pelo estímulo à nossa realização!

> *"O Mundo é tão complexo, e os talentos necessários para apreendê-lo tão variados, que até os melhores intelectos muitas vezes precisam de um parceiro que lhes complemente com o talento de que carecem."* (Stephen Jay Gould, em "Seta do tempo, ciclo do tempo")

Continuemos!
Obrigado, amigos!

Flora Fernandes e Sergio Matioli
Editores

Página deixada em branco

Prefácio à primeira edição

A Biologia está vivendo um momento fascinante neste início de milênio. A consecução de projetos cooperativos em nível mundial para o conhecimento completo de genomas inteiros de organismos com diferentes complexidades, aliada ao desenvolvimento de computadores cada vez mais poderosos, está tornando possível a compreensão em detalhes cada vez maiores de como a informação genética está relacionada aos demais aspectos dos organismos.

No entanto, toda a diversidade biológica tem uma origem histórica através da evolução dos seres vivos. Tentar compreender os genomas de organismos que vivem atualmente sem levar a evolução em consideração seria, no mínimo, uma atitude ingênua.

A proposta deste livro é apresentar a pesquisadores e estudantes de Biologia de graduação ou de pós-graduação o estado-da-arte da pesquisa que é desenvolvida na área de Genômica Evolutiva. Esse texto pode ser utilizado também por outros profissionais de áreas biológicas, biomédicas ou até mesmo por leigos esclarecidos que tenham a curiosidade de compreender os genomas atuais como resultado dos processos evolutivos.

Ao contrário de outros livros que tratam de Biologia molecular apenas descritivamente, mesmo que através de experimentos elegantes, tentamos mostrar que a maneira como as células atuais "funcionam" resulta de eventos que vêm ocorrendo há bilhões de anos. Os dois primeiros capítulos tratam justamente da origem da vida e de seu desenvolvimento inicial para aquilo próximo do que conhecemos atualmente. Evidentemente, o tratamento experimental desses passos iniciais é muito difícil e, mesmo, sujeito a críticas sob os mais diversos pontos de vista. No caso, podemos comparar a investigação com o trabalho do detetive que tenta desvendar o ocorrido através de evidências minuciosamente colhidas, junto com o depoimento de testemunhas. O mais interessante da investigação nessa área é que as principais testemunhas somos nós, os seres vivos da atualidade que carregamos marcas de um passado longínquo. No terceiro capítulo, mostra como os processos evolutivos podem ser utilizados até com finalidades industriais, na confecção de fármacos que poderão ser, em um futuro próximo, talvez tão comuns como a aspirina, com a vantagem de que as ribozimas podem continuar evoluindo, assim como os organismos patogênicos o fazem continuamente. No quarto capítulo, será mostrada como a estabilidade do material genético –ou melhor, sua instabilidade– propicia a matéria prima para a evolução. Esse tema raramente é abordado em livros que tratam de evolução molecular, talvez por se tratar de área de intenso desenvolvimento, onde talvez a maior parte do conhecimento ainda está por ser obtida. No quinto capítulo, o surpreendente comportamento das macromoléculas como um relógio evolutivo é discutido em seus detalhes. O capítulo seis faz uma análise de como a estrutura dos RNAs influencia seus padrões de evolução, alertando contra análises simplificadas, especialmente na questão dos alinhamentos de macromoléculas. A questão da evolução dos elementos genéticos móveis, sua origem e seus efeitos nos genomas está discutida no capítulo sete. O capítulo oito trata da área em grande desenvolvimento atual, a Genômica comparada, onde os padrões gerais conhecidos são sistematizados, com um delicioso sabor de que o melhor ainda está por vir. Os capítulos nove a doze, organizado por uma equipe, tratam do problema de reconstrução de filogenias com dados moleculares. Optamos por dividir a abordagem do problema em blocos temáticos, que se constituem nos capítulos propriamente ditos, tentando manter uma certa uniformidade na exposição de assunto que tem sido um tanto quanto desnecessariamente polêmico. Nos capítulos de treze a dezessete, os polimorfismos moleculares são tratados, tanto através das metodologias empregadas (polimorfismos de isozimas, RFLP e métodos baseados em PCR), como dos métodos de análise (com o emprego da teoria da coalescência) e de uma aplicação na conservação de recursos naturais (capítulo 17), tema tão importante na época atual.

Todos os esforços foram feitos no sentido de propiciar aos interessados uma obra de custo bastante acessível. Para atingir esse fim, contamos com a colaboração dos autores, que abriram mão dos direitos autorais, com o apoio financeiro da Applied Biosystems do Brasil e com o trabalho da própria Holos, Editora, que se esforçou para encontrar a melhor relação custo/benefício. Ao Dr. Antonio Brito da Cunha fica aqui um agradecimento especial pela sugestão de procurar apoio junto à iniciativa privada após tentativas infrutíferas junto a financiadoras e também por sua disposição em apresentar a obra.

Finalmente ressaltamos que os capítulos foram escritos por autores que trabalham com temas relacionados aos assuntos tratados. Foi solicitado explicitamente que os autores apresentassem os resultados de seus trabalhos como exemplos dentro do contexto temático de cada capítulo, para que não continuemos, nos cursos ministrados no Brasil, simplesmente a apresentar os exemplos citados nas obras estrangeiras. Esse procedimento, no entanto, não prejudicou a generalidade da obra. Ao invés disso, temos certeza de que ela ficou enriquecida. Afinal, como disse Leon Tolstoi, "Se queres ser universal, canta a tua aldeia".

Sergio Russo Matioli
Editor

Página deixada em branco

Capítulo 1

Origem da vida: um tempo curto para uma experiência bem-sucedida

Carlos Frederico Martins Menck (cfmmenck@usp.br)
Departamento de Microbiologia
Instituto de Ciências Biomédicas
Universidade de São Paulo

Eduardo Gorab (egorab@usp.br)
Departamento de Genética e Biologia Evolutiva
Instituto de Biociências
Universidade de São Paulo

Mariana Cabral de Oliveira (mcdolive@ib.usp.br)
Departamento de Botânica
Instituto de Biociências
Universidade de São Paulo

"*A origem da vida parece (...) ser quase um milagre, sendo que muitas condições tiveram que ser satisfeitas para mantê-la.*" (Francis Crick, 1981)

1.1. O Ambiente Pré-Biótico

Há uma concordância entre cosmologistas de que o sistema solar tenha se originado há cerca de 4,6 bilhões de anos, a partir de uma nuvem de gases e de poeira interestelar. Um processo de colapso gravitacional dessa nuvem formou o Sol. Por motivos desconhecidos, parte dessa nuvem não se juntou ao Sol, mas os agregados formados geraram os embriões dos planetas atuais. Esses agregados eram muito quentes, mas a Terra primitiva esfriou em algumas centenas de milhões de anos, com a formação de uma crosta externa, exceto pelas erupções vulcânicas do núcleo interno, que fustigavam sua superfície. Os gases que provinham do interior da Terra rapidamente formaram a atmosfera primitiva.

Apesar de alguma discordância quanto à composição dessa atmosfera, a presença de água na forma de vapor é certa. Com o esfriamento do planeta, houve a formação de água na forma líquida, resultando na formação dos primeiros mares e oceanos. No entanto, durante os primeiros 500 milhões de anos, nosso planeta foi violentamente castigado por imensos meteoros, cujos impactos devem ter provocado sua evaporação (Kerr, 1999). Para se ter uma idéia da violência desses impactos, acredita-se que, em um deles, a Terra se partiu originando o nosso único satélite, a Lua. Nessas condições extremamente hostis, surgiram formas de vida que constituem ancestrais diretos dos seres vivos atuais. Várias questões são levantadas sobre quais seriam as condições que propiciariam esse início de vida, mas sem dúvida haveria necessidade da existência de água liquefeita para origem e manutenção de formas de vida tais como as conhecemos. Assim, é de se esperar que os contínuos impactos de meteoros fossem esterilizantes, ou seja, se eventuais formas de vida estivessem sendo então geradas, elas seriam extintas, abrindo a possibilidade de que a vida na Terra tenha se originado mais de uma vez. De qualquer forma, hoje em dia temos evidências concretas de que a vida primitiva já existia em nosso planeta há mais de 3,5-3,8 bilhões de anos, justamente quando os impactos eram mais violentos e quase simultaneamente à formação dos mares e oceanos.

1.2. Evidências das Primeiras Formas de Vida na Terra

Fósseis são restos do processo de mineralização que transforma matéria viva em rocha. Normalmente, apenas porções mais duráveis, como conchas, ossos, penas e folhas, são conservadas em fósseis. Em algumas raras circunstâncias, no entanto, detalhes pequenos podem ser preservados. No final da década de 1950, paleobiologistas descobriram estruturas similares a células com alguns mícrons de diâmetro em formações rochosas antigas (datadas de 3,5 bilhões de anos) na África do Sul e na Austrália. Essas formações rochosas são compostas por estromatólitos fossilizados, que são estruturas desenvolvidas por colônias de cianobactérias e que ainda hoje podem ser observadas em regiões da costa da Austrália, Bahamas, México e outros sítios. As diferentes progênies bacterianas produzem camadas sucessivas umas sobre as outras e a análise de rochas antigas com essas estruturas especiais revelou microfósseis com formas e tamanhos similares a bactérias atuais. Vários tipos de microfósseis (filamentosos ou células individualizadas) foram encontrados juntos, como esperado de um ambiente biológico em que diferentes espécies convivem lado a lado, mas que, considerando a época em questão, é no mínimo surpreendente. Em alguns casos, os fósseis apresentavam estruturas que pareciam processos de divisão celular, o que não se esperaria encontrar em casos de bolhas triviais, casualmente formadas no processo de mineralização, o que confirma que algumas propriedades daqueles organismos eram similares às bactérias atuais.

O material preservado em microfósseis revela apenas a morfologia geral da parede celular, com poucas informações da célula no nível molecular, mas a similaridade entre estromatólitos antigos e os atuais poderiam indicar que já há 3,5 bilhões de anos as células seriam capazes de realizar fotossíntese, como as cianobactérias atuais. Esses dados são impressionantes, visto que processos como divisão celular, formação de parede celular definida e mesmo fotossíntese são complexos. Entretanto, é importante destacar que não existe evidência química de que esses organismos fósseis eram realmente fotossintetizantes (Pierson,

1994). De fato, as evidências mais conclusivas da presença de organismos com capacidade fotossintetizante são de 2,5 a 2,8 bilhões de anos atrás (Canfield, 1999).

O grau de complexidade observado nos microfósseis indica que a vida pode ter se originado mesmo em períodos anteriores a 3,5 bilhões de anos. De fato, mais recentemente, dados referentes à composição isotópica de carbono em rochas ainda mais antigas (3,8 bilhões de anos) indicam que as inclusões de carbono são isotopicamente leves, resultado de atividade biológica (Mojzsis *et al.*, 1996). Essas rochas são compostas de grãos de apatita, fosfato de cálcio básico, encontradas a oeste da Groenlândia e os dados obtidos, apesar de indiretos, não podem ser explicados por atividade abiótica, sendo considerados as evidências mais precoces, de 3,8 bilhões de anos, de existência de vida na Terra.

Assim, o intervalo para que o nosso planeta tivesse condições mínimas para manutenção de vida (água na forma líquida, de 4,0 a 3,8 bilhões de anos) até a origem da vida propriamente dita foi aparentemente curto em relação à história geológica da Terra (não mais do que 300 milhões de anos, em um total de 4 bilhões de anos, ou seja, menos que 8%). Soma-se a isso o fato de que as evidências das primeiras formas de vida já apresentarem um grau de complexidade similar a algumas bactérias atuais, tendo sido originadas em um ambiente extremamente hostil, em que extinções em massa poderiam ser comuns. Essas questões sobre a origem da vida são conhecidas como o paradoxo do tempo e estão longe de serem resolvidas, pois a estruturação de uma primeira célula parece ser um dos pontos mais difíceis de explicar de toda a evolução. Além disso, o curto período de tempo pode indicar simplesmente que a origem da vida não é um processo complexo em si e que poderia ter acontecido várias vezes em nosso planeta e mesmo no universo (Damineli e Damineli, 2007). Certamente, o processo que levou ao surgimento da primeira forma de vida na Terra é desconhecido, mas tem aguçado a curiosidade humana há séculos. Algumas idéias serão discutidas a seguir.

1.3. Síntese Pré- Biótica e a Sopa Primitiva

A composição da atmosfera primitiva é ainda tema de debate (Lazcano e Miller, 1996). Sabe-se que o oxigênio deveria estar ausente ou presente em quantidades extremamente pequenas nessa atmosfera, uma vez que a maior parte do oxigênio atmosférico atual é proveniente da água, produzido por organismos fotossintetizantes. Níveis próximos aos atuais foram atingidos apenas há cerca de 2 bilhões de anos. Compostos redutores poderiam estar presentes, tais como hidrogênio molecular (H_2), metano (CH_4), amônia (NH_3) e monóxido de carbono (CO). No entanto, hoje em dia acredita-se que a atmosfera primitiva tinha uma composição menos redutora, incluindo a presença de dióxido de carbono (CO_2) e nitrogênio molecular (N_2). A energia proveniente de descargas elétricas (relâmpagos) e da luz solar (principalmente de alta energia, como a luz ultravioleta) pode ter permitido a realização de reações químicas que formariam vários compostos intermediários importantes para a formação de moléculas orgânicas mais complexas, que coexistiriam no oceano primitivo. Essa idéia, conhecida como sopa primitiva, foi proposta independentemente por Oparin e Haldane na década de 1920, tendo como base uma atmosfera altamente redutora.

Essa hipótese foi testada experimentalmente apenas em 1953 por Stanley Miller, então estudante de Harold Urey (Miller, 1953). Para isso, Miller e Urey construíram um sistema fechado, com refluxo (Figura 1.1), que mimetizaria as condições da atmosfera primitiva, incluindo a presença de água, amônia, hidrogênio e metano, sendo que descargas elétricas simulavam relâmpagos. A análise da água condensada ("oceanos") após esses experimentos revelou que cerca de 10% do carbono adicionado na forma de metano produz várias moléculas orgânicas, incluindo aminoácidos como glicina, alanina, aspartato, valina e leucina (nas formas D e L, ou seja, seus isômeros ópticos). A formação de outros compostos, como formaldeído (H_2CO), nitratos e cianeto (HCN), também foi verificada, sendo que esses podem ter sido intermediários na formação de outros aminoácidos e de componentes dos ácidos nucléicos.

De fato, após os experimentos iniciais de Miller e Urey, vários outros foram feitos simulando diferentes condições presentes na Terra primitiva. De interesse, a síntese de polipeptídeos pode ser realizada a altas temperaturas (120º C), condições que poderiam ser encontradas, por exemplo, próximo a vulcões. A síntese abiótica de alguns componentes dos ácidos nucléicos, como purinas e, em menor quantidade, de pirimidinas, também é possível a partir da condensação de HCN (Figura 1.2). A adenina, de fato, é a base nitrogenada encontrada em maiores concentrações nos experimentos de simulação de síntese abiótica. Pequenas quantidades de ATP (trifosfato de adenosina) também podem ser produzidas em condições abióticas, sobretudo na presença de um mineral bastante comum conhecido como apatita (fosfato de cálcio). A síntese de vários açúcares a partir de formaldeído (H_2CO) também foi descrita em condições que simularam a Terra primitiva. Por exemplo, a polimerização de formaldeído resulta em ribose e não em desoxirribose. Como será descrito adiante, essa é apenas uma das muitas observações que sugerem que o RNA precedeu o DNA na evolução da vida. Apesar da possível formação de nucleosídeos polifosfatados em determinadas condições (principalmente na presença de apatita), a polimerização de oligômeros de RNA não é evidente, uma vez que misturas de isômeros (D e L) e ligações envolvendo as hidroxilas 2', 3' e 5' da ribose podem gerar moléculas ramificadas, diferentes das que conhecemos atualmente (contendo apenas enantiômeros D da ribose e ligações 5'-3'). É possível que condições especiais tenham existido que permitissem a geração de polinucleotídeos similares

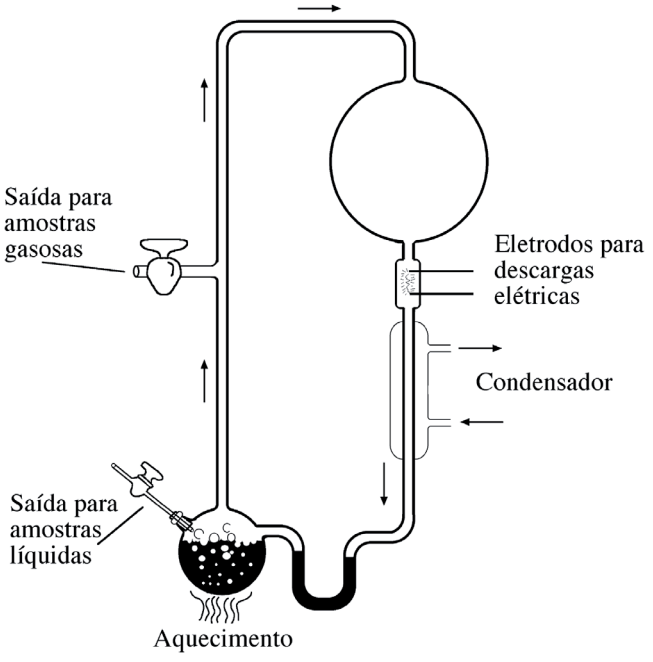


Figura 1.1. O equipamento empregado por Muller e Urey, para simular as condições da Terra Primitiva (modificado de Miller, 1953).

aos que formam a base dos organismos vivos conhecidos, que teriam sido selecionados positivamente.

Outra evidência de que moléculas orgânicas podem ser geradas em condições abióticas provém do espaço. Análises de meteoritos têm demonstrado em posições internas do meteoro a presença de aminoácidos e outras moléculas consideradas os monômeros da vida, descartando-se, portanto, a possibilidade de contaminação com material terrestre.

A formação de lipídeos também seria fundamental para o processo de compartimentalização de compostos eventualmente necessários para a criação do ser vivo, através da composição de membranas de dupla camada lipídica. Embora a síntese prébiótica de seus componentes (ácidos graxos, glicerol e fosfato) seja plausível na sopa primitiva, não está claro como foram formados lipídeos de cadeias longas, lineares, indispensáveis para a formação de membranas.

Assim, os compostos orgânicos correspondendo aos menores blocos fundamentais para a vida (aminoácidos, nucleotídeos e açúcares) provavelmente puderam ser formados através de síntese abiótica. A formação de compostos maiores, resultado da condensação (eliminação da água) destes compostos deve ter gerado, de forma ainda desconhecida, moléculas poliméricas através de ligações fosfodiéster entre nucleotídeos dos ácidos nucléicos e amida entre aminoácidos de proteínas. Pela inexistência de oxigênio na atmosfera primitiva os compostos gerados poderiam ser estáveis, o que levaria ao seu acúmulo nos oceanos com concentrações suficientemente altas para reações mais complexas. Exatamente como previsto por Haldane e Oparin.

1.4. Quiralidade e Origem da Vida

Moléculas biologicamente importantes podem se apresentar em duas formas no que se refere à sua atividade óptica. São os chamados isômeros dextrógiros (D, desviam o plano de vibração da luz polarizada para a direita) e os levógiros (L, desvio para a esquerda). São, portanto, enantiômeros que apresentam propriedades quirais. O conceito quiralidade não se aplica a moléculas desprovidas de atividade óptica e a misturas equimolares de enantiômeros, denominadas misturas racêmicas.

A origem da homoquiralidade, isto é, a geração de um tipo predominante de molécula com atividade óptica e o uso

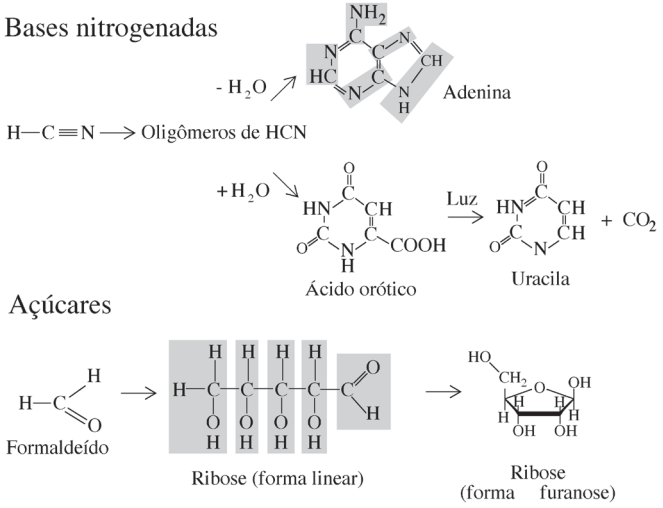


Figura 1.2. Esquema para formação de purinas e pirimidinas e riboses, precursores básicos dos ácidos nucléicos.

preferencial de uma destas formas enantioméricas pelos seres vivos é um dos problemas importantes no estudo da origem da vida. Exemplificando, aminoácidos levógiros são as unidades estruturais das proteínas em sistemas biológicos. Já os açúcares, presentes na estrutura dos ácidos nucléicos e em várias vias metabólicas, são dextrógiros. Esta característica dos organismos torna-se ainda mais intrigante quando a síntese em laboratório de moléculas que apresentam isomeria óptica normalmente leva à produção de misturas racêmicas ao invés de ocorrência preferencial de um dos enantiômeros.

Quanto à sua origem, a homoquiralidade poderia ter sido abiótica, isto é, ela pode ter surgido antes da emergência dos primeiros sistemas biológicos. Neste caso, excessos enantioméricos teriam aparecido no decorrer da evolução química do universo precedendo a origem da vida. A hipótese biótica do surgimento da homoquiralidade leva em conta que os primeiros seres vivos utilizaram misturas racêmicas de moléculas e que a especialização dos organismos na direção do uso de enantiômeros específicos foi uma aquisição evolutivamente tardia.

No estudo do problema da quiralidade, são evidentemente importantes as pesquisas de química orgânica utilizando compostos de origem terrestre. No entanto, se a mesma abordagem é feita com compostos de origem extraterrestre, dados significativamente informativos podem ser obtidos particularmente pelo seu horizonte temporal mais amplo. Para isto, meteoritos da classe dos condritos carbonáceos são a fonte dos compostos extraterrestes. Conhecendo sua idade, estimada em aproximadamente 4,5 bilhões de anos, e estudando sua composição, é possível obter registros da evolução química do sistema solar desde fases precoces de sua formação. Com esta abordagem, os dados procedentes do meteorito de Murchinson chamaram a atenção. Tendo sido observada em sua composição uma ligeira predominância de certos aminoácidos levógiros, os resultados sugeriram que a homoquiralidade seria um fenômeno abiótico com possíveis implicações para a origem da vida (Cronin e Pizzarello, 1997). Estes dados têm sido discutidos até os dias de hoje, sendo ainda matéria controversa.

É importante ressaltar que, nos estudos sobre origem da vida na Terra, tem ganhado importância a participação de compostos extraterrestres trazidos por meteoros, cometas e asteróides. Nesta direção, outra abordagem de estudo da quiralidade, derivada dos métodos acima descritos, consiste em simular no laboratório a química interestelar para verificar a possibilidade de formação de compostos importantes para a origem da vida, com ênfase nos enantiômeros, vindos de fora do sistema solar. São também denominados estudos de síntese nos "análogos de gelo interestelar". Outros processos terrestres ou extraterrestres também podem enriquecer enantiômeros de um dos tipos, entre eles a polimerização (Gleiser e Walker, 2008), sublimação (Cintas, 2008), cristalização (Weissbuch *et al.*, 2009) e adsorção (Hazen *et al.*, 2001) diferenciais. Tais processos não são mutuamente exclusivos, podendo portanto ter ocorrido em sinergia.

1.5. Coacervados

A evolução da matéria orgânica aos primeiros organismos requer, entre outros passos, uma separação clara entre os componentes orgânicos e o meio circundante. Esta barreira limitante assegura não somente que os constituintes das primeiras estruturas biológicas não se diluam no ambiente, mas também garante a possibilidade de trocas entre o meio externo e o interno. A separação da matéria orgânica numa solução em corpúsculos, que poderiam ser vistos como possíveis precurso-

res dos primeiros organismos, foi estudada pioneiramente por Oparin (Oparin, 1957). Durante muitos anos, Oparin ocupou-se em observar a tendência de soluções aquosas de polímeros segregar espontaneamente na forma de coacervados, que são corpúsculos de natureza coloidal ricos em polímeros e que podem ser mantidos em equilíbrio com o meio circundante rico em água. Oparin testou várias combinações de polímeros biológicos capazes de formar coacervados, que não devem necessariamente ser vistos como os ancestrais dos primeiros organismos. Isto porque alguns dos polímeros empregados nos experimentos de Oparin são complexos, como por exemplo o RNA. As dimensões dos coacervados variam de 1 a 500 μm e muitos deles apresentam, ao microscópio, uma delimitação que lembra a de uma membrana celular atual. Entretanto, os coacervados são instáveis de maneira geral e podem se desfazer em minutos. Oparin e seus colaboradores estudaram condições para aumentar sua estabilidade. Uma propriedade importante dos coacervados, ou de qualquer outro sistema difásico, é a de que substâncias, de solubilidade diferente nas duas fases, ficarão concentradas preferencialmente em uma das duas fases. Oparin observou que, quando fosforilase era adicionada a uma solução contendo histona e goma arábica, a enzima se concentrava nos coacervados formados por estes dois componentes. Se glicose-1-fosfato era adicionada à solução, ela se difundia para o interior dos coacervados e a glicose era enzimaticamente polimerizada em amido. A energia para a polimerização vem do grupo fosfato que após a reação se difunde para fora do coacervado (Figura 1.3). Como a goma arábica é também um polímero, a produção de amido incrementava a massa do coacervado e este crescia em tamanho. Quando os coacervados se tornam muito volumosos, eles tendem à fragmentação espontânea produzindo uma "descendência". Os coacervados "filhos" que recebem a enzima podem continuar crescendo embora a uma taxa menor já que a quantidade inicial de enzima estará dispersa entre um número crescente de coacervados. Esta limitação seria superada se os coacervados tivessem meios de produzir sua própria enzima. Oparin introduziu nos coacervados outras etapas conhecidas do metabolismo celular como por exemplo a quebra do amido pela amilase produzindo maltose, ou ainda a produção de RNA poli-A a partir de ADP e RNA polimerase. É interessante notar que a estabilidade dos coacervados aumentava se "dotados de algum tipo de metabolismo" como exemplificado acima.

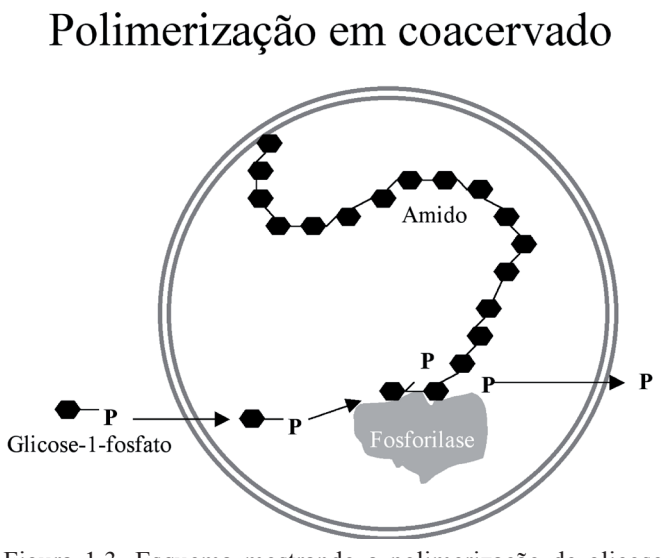


Figura 1.3. Esquema mostrando a polimerização de glicose-1 fosfato a amido no interior de um coacervado (modificado de Dickerson, 1978).

1.6. Hipóteses Alternativas

Embora diversos compostos necessários à vida possam ser formados abioticamente, a existência em quantidades significativas de muitas outras moléculas consideradas fundamentais para a vida não está estabelecida. Além disso, a suposição de que a atmosfera primitiva fosse realmente redutora tem sido contestada mais recentemente. Mas a natureza da atmosfera primitiva seria de fato irrelevante se acreditarmos na hipótese de que a vida teria surgido nas fontes térmicas oceânicas de profundidade (Damineli e Damineli, 2007). Fosseis datados em 3,2 bilhões de anos foram encontrados em uma formação na Austrália considerada como um sistema hidrotérmico do fundo oceânico. Esses fosséis foram interpretados como procariotos filamentosos, quimiotróficos e termofílicos (Rasmussen, 2000), além disso, a presença de resquícios de petróleo nessa formação indica que os oceanos primitivos estariam repletos de vida há mais de 3 bilhões de anos (Rasmussen, 2005).

Uma alternativa para a "sopa" primitiva é a hipótese de que a vida teria surgido em um substrato sólido. A existência de ciclos de polimerização e catálise do ponto de vista cinético é mais facilmente explicada sobre um substrato sólido onde as interações entre as moléculas poderiam ocorrer de uma forma mais organizada e específica (Maynard-Smith e Szathmáry, 1995). Segundo Wächtershäuster (1988), o substrato ideal seria uma superfície mineral carregada positivamente, sendo que os compostos estariam ancorados nesta superfície por porções carregadas negativamente, o que elimina a argila da lista de possíveis superfícies. Wächtershäuster favorece a pirita (sulfeto de ferro, o chamado "ouro dos tolos") como superfície possível. Suponha que um determinado ciclo químico que chamamos de ciclo 1 usasse o composto A como substrato e produzisse o composto B. Se o ciclo 2 usasse o composto B como substrato e produzisse o A, o estabelecimento de uma relação entre estes dois ciclos seria vantajosa para ambos e seria muito mais fácil ocorrer em um substrato do que em meio líquido. A especificidade da reação é crucial para a manutenção de um ciclo. Nos seres vivos esta especificidade é mantida pelas capacidades catalíticas das enzimas e pela compartimentalização provida pelas membranas. Nos processos abióticos os problemas da falta de especificidade e de compartimentalização poderiam em parte ser minimizados por um metabolismo de superfície.

Outras alternativas para o local de origem da vida foi proposta por Woese (1979) que levantou a hipótese de que a evolução dos ciclos químicos que deram origem aos primeiros seres vivos tenha ocorrido em gotículas de água presentes nas nuvens. Esta alternativa foi proposta por duas razões: 1. a temperatura da superfície terrestre pode ter sido elevada demais para permitir a formação de oceanos; 2. gotículas de água proveriam uma vasta área para a ocorrência destas reações químicas. Estas gotículas frequentemente são formadas ao redor de partículas sólidas, o que poderia combinar a hipótese das gotículas com a de uma superfície sólida.

Outra sugestão interessante é a proposta por Juan Oró (1961) na qual ele considera que a presença de matéria orgânica em meteoritos não só comprova sua origem abiótica como também sugere que o bombardeamento da Terra com meteoros e cometas trouxe vários dos componentes necessários para a origem da vida. De fato, a presença de matéria orgânica no universo já é bem aceita na comunidade científica e uma parte importante dos compostos orgânicos encontrados na Terra primitiva pode ter tido origem interestelar (Bernstein *et al.,* 1999). Como corolário dessa idéia, a recente e polêmica observação de evidências de células microbianas em meteorito (ALH84001) vindo de Marte

é consistente com a proposta de Oró e também traz de volta a discussão de uma possível origem de vida na Terra a partir de uma colonização vinda do espaço, idéia antiga conhecida como Teoria da Panspermia cósmica.

A exobiologia, ou o estudo da possibilidade de vida fora da Terra, não é uma preocupação recente da humanidade. Ela tem suas raízes na antiguidade clássica, passando pela Idade Média e chegando ao século XIX com a idéia de que a vida havia sido semeada na Terra vinda do espaço. Este processo foi defendido por nomes ilustres da Ciência até o século XX. Naquele século, a exobiologia ganhou impulso considerável graças à corrida espacial. A NASA, na década de 1960, entendendo que estes estudos seriam importantes com o desenvolvimento da exploração do espaço, decidiu criar uma divisão de exobiologia que estudaria problemas relacionados às origens da vida na Terra e fora dela. Desde então até os dias de hoje, projetos de pesquisa envolvendo agências espaciais e laboratórios de vários países têm sido feitos, tendo como um dos intuitos a busca de evidências que sustentem a possibilidade de vida fora da Terra. Provavelmente o trabalho de maior impacto nos meios de comunicação nesta área foi aquele realizado em 1996 no meteorito ALH84001, no qual os pesquisadores sugerem a possibilidade de atividade biológica no planeta Marte (McKay *et al*., 1996). É importante destacar, no entanto, que as estruturas que foram encontradas nesse meteorito têm um tamanho cerca de 10 vezes menor do que as menores células conhecidas. Outra linha de investigação exobiológica, além dos estudos continuados sobre a composição de meteoritos e sua possível relação com atividade biogênica, é a pesquisa sobre a composição química de planetas, do sistema solar ou fora dele. Estes dados têm, como uma das finalidades, por exemplo, simular em laboratório se tal composição permite evolução química, como aquela observada nos experimentos clássicos de Miller em 1953, e quais os resultados. Há poucos anos, a existência de água em planetas também passou a ser uma referência para a NASA nesta busca por indícios de vida fora da Terra (Ball, 2004).

1.7. A Questão da Vida

Darwin sabia da importância da questão da origem da vida para a elaboração de sua teoria. Em uma carta a um amigo ele sugere que "em algum lago, pequeno e quente, com todo tipo de amônia e sais fosfóricos, luz, calor, eletricidade, etc. presentes, (…) um composto protéico teria sido formado quimicamente, pronto para sofrer mudanças mais complexas (...)".

Há um consenso de que a protocélula seria envolvida por uma membrana lipídica de duas camadas, criando um sistema compartimentalizado que isolaria algumas reações químicas do meio. Esse isolamento não poderia ser total, uma vez que haveria a necessidade de permeabilidade da membrana para algumas moléculas, permitindo alguma troca com o meio. No compartimento interior moléculas orgânicas estariam dissolvidas em água, incluindo moléculas energéticas (ex. ATP). Além disto, seriam necessárias moléculas catalisadoras (ex. RNA e/ou proteínas) e moléculas para armazenamento de informação (provavelmente RNA, ver Capítulo 2). A origem desta protocélula foi precedida pela chamada evolução química, que produziu os blocos básicos para a vida, como discutido anteriormente.

As seguintes condições são necessárias para a formação da vida: 1. Energia: necessária para a formação de moléculas orgânicas mais complexas. A energia na Terra primitiva estava presente sob diferentes formas (descargas elétricas, raios cósmicos, raios UV, impactos de meteoros, radioatividade, vulcanismo); 2. Proteção: após as moléculas complexas terem se formado, elas tinham que ser protegidas ou então seriam destruídas pelo contínuo fluxo de energia. Esta proteção poderia ser em águas mais profundas, ou em fendas nas rochas, gelo, sedimentos; 3. Concentração: a diluição dos compostos levaria a reações muito lentas, portanto devem ter existido mecanismos que favoreceriam a maior concentração de compostos (evaporação ou congelamento, por exemplo); 4. Catálise: muitas reações químicas são favorecidas por catalisadores, uma substância que auxilia nas reações mas não toma parte nela (ex. enzimas). Inicialmente, alguns catalisadores podem ter favorecido a origem de moléculas mais complexas, posteriormente, determinadas reações podem ter sido favorecidas por catalisadores presos no interior de membranas.

A definição de vida proposta por Muller (1966) é que entidades vivas teriam as propriedades de multiplicação, variação e hereditariedade. Estas propriedades seriam importantes, pressupondo um armazenamento de informação herdável, que poderia evoluir, sofrendo os efeitos da seleção natural. Uma outra característica fenotípica essencial a vida é o metabolismo – uma série de reações químicas relacionadas. O metabolismo supriria os monômeros necessários para a confecção das moléculas replicativas e, por sua vez, as moléculas replicativas alterariam as reações químicas do metabolismo. Assim, a seleção natural, agindo nas moléculas replicativas, influenciaria o metabolismo.

1.8. Vida Artificial

Se por um lado os coacervados não são entidades biológicas e não podem ser vistos necessariamente como ancestrais dos organismos primitivos, eles certamente foram os precursores de uma linha de investigação atual conhecida como "vida artificial". Iniciada no final da década de 1980, estes experimentos procuravam estabelecer os componentes de um sistema com um número mínimo de componentes para ser considerado vivo (Luisi *et al.*, 2006). Isto naturalmente remete ao significado do conceito vida. Neste caso, a pesquisa sobre vida artificial considera vivo um sistema macromolecular capaz de se auto-sustentar com um metabolismo mínimo, capaz de se reproduzir e, mais importante, dotado de um potencial evolutivo. Este potencial evolutivo é a noção darwinista de que populações compostas por estes sistemas devam sofrer mudanças estruturais sujeitas à seleção. A chamada "proto-célula artificial" é o conceito, comumente citado na pesquisa na área, que busca atender às exigências acima mencionadas. Neste caso, a construção mais simples, de natureza ainda puramente teórica, é a da "protocélula artificial de RNA". As propriedades genéticas e enzimáticas das ribozimas possuem papel central neste modelo (ver capítulos 2 e 3). As "células de RNA" consistem de uma vesícula contendo duas ribozimas. Uma delas teria atividade de uma replicase, isto é, ela é capaz de se auto-replicar e a outra possuiria atividade catalisadora da síntese dos componentes do envoltório celular. Esta construção parte também da premissa de que a permeabilidade do envoltório seria total e que os precursores das estruturas celulares estariam generosamente presentes no ambiente externo. Além disto, parte-se do pressuposto de que estas entidades se dividem e que as duas ribozimas estariam presentes nas "células-filhas" (Figura 1.4). Trata-se de um modelo hipotético, uma vez que que estas ribozimas ainda não foram identificadas. Apesar desta limitação, duas subáreas de estudo emergiram com o uso de vesículas de lipídios (lipossomos) como modelos proto-celulares. A primeira se concentra no estudo de possíveis analogias entre as vesículas e

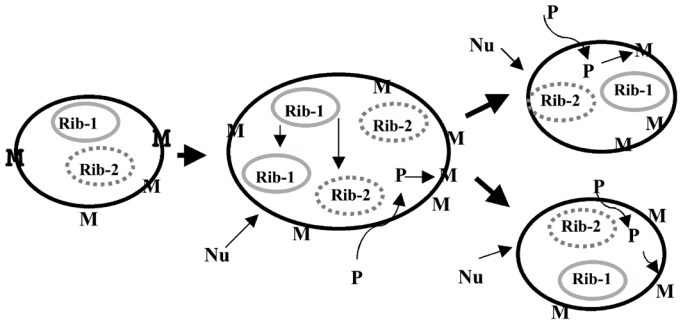


Figura 1.4. Estrutura esquemática (modificada de Luisi et al., 2006) de uma protocélula artificial de RNA. A ribozima 1 (rib-1) é uma RNA replicase capaz de se autoreplicar e de produzir cópias da ribozima 2 (rib-2) utilizando precursores de RNA (Nu). A ribozima 2 é capaz de converter moléculas precursoras de membrana (P) em componentes de membrana (M).

membranas celulares em termos de propriedades físico-químicas tais como estabilidade, permeabilidade e auto-reprodução. Foi demonstrado que as vesículas podem se multiplicar na presença de precursores, retendo em certas condições as dimensões da matriz. Por outro lado, problemas podem surgir quanto à permeabilidade, visto que foi observada certa resistência de membranas fosfolipídicas a certas moléculas. A segunda subárea investiga o uso das vesículas como hospedeiras de reações bioquímicas complexas para verificar se elas comportam a química biológica. Os resultados têm apontado para várias possibilidades como a síntese enzimática de RNA poli-A no interior das vesículas, síntese de polipeptídeos em lipossomos que incorporaram complexos ribossômicos, além da amplificação de DNA por PCR no interior dos lipossomos. Um resultado que merece destaque foi o da replicação do RNA no interior das vesículas de ácido oléico, com concomitante auto-replicação das vesículas a partir de precursores de RNA, RNA molde e uma replicase do RNA. Apesar de ser o resultado mais próximo ao conceito de "protocélula artificial de RNA" citado anteriormente, a replicase não pode ser produzida continuamente e, consequentemente, o sistema está limitado à quantidade de replicase disponível.

Outra frente de pesquisa relacionada à criação de vida artificial ou sintética é o projeto "Genoma Mínimo" iniciado em meados dos anos 1990, pelo Instituto J. Craig Venter (JCVI), cujo objetivo era entender quais são os componentes genéticos mínimos necessários para sustentar a vida através da criação de uma célula bacteriana baseada inteiramente em um genoma obtido em laboratório. Entre as etapas para atingir esse objetivo, o grupo deu dois passos importantes, em 2007 mostrou ser possível o transplante de um genoma inteiro, possibilitando transformar uma bactéria em outra (Lartigue *et al*., 2007). O transplante de genoma foi o 1º passo essencial no campo da genômica sintética, uma vez que é o mecanismo pelo qual um cromossomo artificial inteiramente sintetizado em laboratório pode ser introduzido numa célula originando um novo ser vivo. Outro passo foi a reconstrução, a partir da síntese de DNA inteiramente em laboratório, de um cromossomo de *Mycoplasma genitalium* (Gibson *et al*., 2008), que é a bactéria que pode ser crescida em uma cultura pura e possui menor genoma conhecido (582.970 pares de bases e 485 genes). Embora a síntese de DNA em laboratório seja possível desde a década de 1980, a síntese de grandes moléculas de DNA não é trivial e a construção desse cromossomo foi conseguida a partir de fragmentos de DNA quimicamente sintetizados em laboratório e a partir do desenvolvimento de novas metodologias para a montagem desses fragmentos na ordem correta. Esses trabalhos abrem uma série de possibilidades e aplicações, mas geram também uma série de preocupações e questões éticas.

1.9. As Formas Simples de Vida

Todas as hipóteses apresentadas acima referem-se aos diferentes ambientes químicos nos quais surgiram os primeiros polímeros capazes de replicação. Experimentos laboratoriais trouxeram grande contribuição na explicação de como determinados compostos podem ter surgido, entretanto ainda existem diversos problemas a serem resolvidos. Sem dúvida, ao lado de dados experimentais, é possível que a busca de evidências naturais, seja pelo registro fóssil, seja por resquícios da vida primitiva em organismos atuais, tragam pistas mais consistentes de qual tipo de organismo seria o progenota (nome dado ao primeiro organismo).

Como teria surgido esta relação entre moléculas replicativas e metabolismo? O mais simples seria assumir que inicialmente o metabolismo tenha consistido de reações químicas abióticas que supriam monômeros a partir dos quais as moléculas replicativas eram sintetizadas, mas estes replicadores não influenciavam o metabolismo. A seleção natural selecionaria apenas os replicadores que se multiplicavam melhor. Este é o modelo de origem da vida "replicadores precoces" ("replicator-first", em inglês). Apenas mais tarde os replicadores teriam desenvolvido a habilidade de alterarem seu ambiente químico. Na busca de modelos apropriados nas células atuais, verifica-se que a molécula de RNA apresenta uma série de características que a tornam candidata a ter atuado nas primeiras formas de vida. Como discutido no próximo capítulo, moléculas de RNA poderiam guardar informações e catalisar vários processos químicos, incluindo auto-replicação. Entretanto, nas células atuais o RNA tem uma função subalterna ao DNA em termos de guarda de informação genética, com exceção de alguns vírus, cujo genoma é basicamente RNA. Vírus, aliás, pela sua simplicidade e diversidade de genoma (incluindo RNA) são comumente lembrados como eventuais precursores de vida na Terra. Apesar de existirem muitos tipos de vírus e estes serem ubíquos, ou seja, infectarem todo tipo de célula viva conhecida, estes são parasitas celulares obrigatórios e, portanto, dependem de um metabolismo celular onde poderiam se replicar. Sendo assim, é improvável que os vírus tenham sido as primeiras formas de vida na Terra, mesmo que eles tenham surgido logo após as primeiras células vivas.

A origem da vida teria requerimentos geológicos muito restritivos, ou seja, uma série de condições favoráveis seria necessária para o surgimento da vida, como água na forma líquida; seriam necessários ciclos químicos básicos; a vida primitiva poderia ter mecanismos que não estão mais presentes nos seres vivos atuais, mas o princípio básico de operação deveria ser semelhante; os primeiros sistemas biológicos deveriam ter sido pouco eficientes em comparação com os sistemas atuais.

Referências Bibliográficas

Bernstein, M.P., Sanford, A.S. e Allamandola, L.J. (1999). Life's far-flung Raw materials. **Scientific American, 799:**26-31.

Ball, P. (2004) Astrobiology: water, water, everywhere? **Nature 427:** 19-20.

Canfield, D.E. (1999). A breath of fresh air. **Nature 400**: 503-505.

Cintas, P. (2008). Sublime arguments: Rethinking the generation of homochirality under prebiotic conditions. Angew. Chem. Int. Ed. 47:2918 – 2920.

Cronin, J.R. e Pizzarello, S. (1997) Enantiomeric excesses in meteoritic amino acids. **Science 275**: 951-955

Gibson, D.G., Benders, G.A., Andrews-Pfannkoch, C., Denisova, E.A., Baden-Tillson, H., Zaveri, J., Stockwell, T.B., Brownley, A., Thomas, D.W., Algire, M.A., Merryman, C., Young, L., Noskov, V.N., Glass, J.I., Venter, J.C., Hutchison, III, C.A. e Smith, H.O. (2008). Complete Chemical Synthesis, Assembly, and Cloning of a *Mycoplasma genitalium* Genome. **Science 319**: 1215-1220.

Gleiser, M. e Walker, S.I.(2008). An extended model for the evolution of prebiotic homochirality: A bottom-up approach to the Origin of Life. Orig. Life Evol. Biosph. 38:293-315.

Hazen, R.M., Filley, T.R. e Goodfriend, G.A.(2001). Selective adsorption of L- and D-amino acids on calcite: Implications for biochemical homochirality. Proc. Natl. Acad. Sci. USA, 98(10):5487–5490.

Damineli, A. e Damineli, D.S.C. (2007). Origens da vida. **Estudos Avançados 21:** 263-284.

Dickerson R.E. (1978) Chemical evolution and the origin of life. **In Evolution** (A Scientific American Book): 30-46.

Kerr, R.A. (1999). Early life thrieved despite earthly travails. **Science 284**:2111-2113.

Lartigue, C., Glass, J.I., Alperovich, N., Pieper, R., Parmar, P.P., Hutchison, III, C.A., Smith, H.O., e Venter, J.C. (2007). Genome Transplantation in Bacteria: Changing One Species to Another. **Science 317:** 632-638.

Lazcano, A. e Miller, S.L. (1996). The origin and early evolution of life: prebiotic chemistry, the pre-RNA world, and time. **Cell 85:** 793-798.

Luisi P.L., Ferri, F., Stano, P. (2006) Approaches to semi-synthetic minimal cells: a review. **Naturwissennschaften 93:** 1-13.

McKay, D.S., Gibson, E.K., Thomas-Keprta, K.L., Vali, H., Romanek, C.S., Clemett, S.J., Chillier, D.F., Maechling, C.R., Zare, R.N. (1996). Search for Past Life on Mars: Possible relic biogenic activity in martian meteorite ALH84001. **Science 273**: 924-930.

Maynard Smith, J. e Szathmáry, E. (1995). **The major transitions in evolution**. WH Freeman Spektrum, Oxford, 346 pp.

Miller, S. L. (1953). A production of amino acids under possible primitive Earth conditions. **Science 117**:528-529.

Mojzsis, S.J., Arrhenius, G., McKeegan, K.D., Harrison, T.M., Nutman, A.P. e Frind, C.R.L. (1996). Evidence for life in Earth before 3,800 million years ago. **Nature 384**:55-59.

Muller, H.J. (1966). The gene material as the initiator and organizing basis of life. **American Naturalist 100**: 493-517.

Oparin, A.I. (1957). **The Origin of Life on the Earth**, 3rd Edition. Academic Press Inc. Pub. New York.

Oró, J. (1961). Comets and the formation of biochemical compouds on the primitive Earth. **Nature 190**:389.

Pierson, B.K. (1994). The emergence, diversification, and role of photosynthetic eubacteria. **In** Bengtson, S. (ed) **Early Life on Earth, Nobel Symposium, no 84**. Columbia Univ. Press, New York, 70-80.

Rasmussen, B. (2000). Filamentous microfossils in a 3,235-million-year-old volcanogenic massive sulphide deposit. **Nature 405**: 676-679.

Rasmussen, B. (2005). Evidence for pervasive petroleum generation and migration in 3.2 and 2.63 Ga shales. **Geology 33:** 497–500

Wächtershäuster, G. (1988). Before enzymes and templates: theory of surface metabolism. **Microbiological Reviews 52**: 452-84.

Weissbuch, I, Illos, R.I., Bolbach, G. e Lahav, M. (2009). Racemic beta-sheets as templates of relevance to the origin of homochirality of peptides: Lessons from crystal chemistry. Acc. Chem. Res., 42(8):1129-1140

Woese, C.R. (1979). A proposal concerning the origin of life on the planet Earth. **J. Mol. Evol**. **13**:95-101.

Página deixada em branco

Capítulo 2

O mundo de RNA e a origem da complexidade da vida

Mariana Cabral de Oliveira (mcdolive@ib.usp.br)
Departamento de Botânica
Instituto de Biociências
Universidade de São Paulo

Carlos Frederico Martins Menck (cfmmenck@usp.br)
Departamento de Microbiologia
Instituto de Ciências Biomédicas
Universidade de São Paulo

"*Nada em Biologia faz sentido exceto à luz da Evolução.*" (Theodosius Dobzhansky)

2.1. Fundamentos do "Mundo de RNA"

Até os anos 1950, prosperou a idéia de que as proteínas foram as primeiras moléculas associadas à vida na Terra. Essa idéia baseava-se no fato de que as proteínas são estruturas flexíveis, muito diversas e possuem propriedades catalíticas. Além disso, havia sido demonstrada a possibilidade de síntese abiótica de aminoácidos e até mesmo de polipeptídeos (veja Capítulo 1). No entanto, no final da década de 1960, Orgel (1968), Crick (1968) e Woese (1967) propuseram independentemente e a partir de evidências complementares que o RNA seria a primeira molécula da vida. Esses autores perceberam o papel central dessa molécula em relação às proteínas e ao DNA, além de chamarem a atenção para as várias funções que já se sabia, à época, eram exercidas pelo RNA nas células: mensageiro (mRNA), transportador (tRNA) e ribossômico rRNA (Jeffares *et al.*, 1998; Landweber *et al.*, 1998; Szathmáry, 1999). Descobertas mais recentes somente vieram a fortalecer essa hipótese: a existência de moléculas de RNA com capacidade catalítica, uma propriedade que era considerada exclusiva das proteínas; a presença de algumas moléculas idênticas ou muito semelhantes aos monômeros de RNA em todos os seres vivos, que atuam como cofatores; o fato de o DNA não ser quimicamente tão flexível; e o fato de os desoxirribonucleotídeos serem derivados dos ribonucleotídeos—além de um número crescente de funções celulares que estão sendo associadas às moléculas de RNA (Soares e Valcárcel, 2006, veja também Capítulo 4).

O termo "mundo de RNA" (*RNA world*, em inglês) foi cunhado por Gilbert (1986) para descrever um cenário onde a principal molécula ativa na origem da vida era o RNA. Isso foi proposto com base nas descobertas de moléculas de RNA com propriedades catalíticas, somando-se sua capacidade de guardar a informação genética e de evolução, conceitos diretamente ligados à própria definição de vida.

Com a descoberta de RNA catalítico, resolveu-se um paradoxo do tipo "o ovo ou a galinha" sobre a origem da vida: ácidos nucléicos são essenciais à vida, mas parecem necessitar das proteínas para funcionar. Entretanto, se o RNA funcionasse como fonte de informação e também como enzima, as proteínas poderiam ter surgido posteriormente. A hipótese do mundo de RNA afirma que a reprodução e o metabolismo das primeiras formas de vida dependiam das atividades catalíticas e replicativas do RNA, e que tanto o DNA quanto as proteínas teriam assumido suas funções atuais posteriormente (Gilbert, 1986).

O RNA é único na sua capacidade de armazenar informação genética (em diversos vírus, como o HIV e *Influenzae*) e de executar uma série de atividades catalíticas (introns autocatalíticos, ribonuclease P, entre outros), uma propriedade que, até alguns anos atrás, se acreditava limitada às proteínas. Nas células atuais, o RNA está envolvido em uma série de processos, como a síntese protéica, a replicação de DNA e o processamento de RNA. As múltiplas funções exercidas pelo RNA dão apoio indireto para a hipótese do mundo de RNA, que considera os RNAs catalíticos atuais como remanescentes de uma época em que a vida teria o RNA como principal mediador de processos informacionais e catalíticos, ou seja, seriam verdadeiros fósseis moleculares (Joyce, 1989).

Neste capítulo, pretendemos descrever e discutir o panorama de como um ancestral universal (progenota) com metabolismo baseado em RNA deu origem à diversidade de seres vivos atuais, que têm como material genético o DNA e o restante do metabolismo realizado por RNA e proteínas.

2.2. O RNA Provavelmente Precedeu o DNA

Apesar das dificuldades para se conseguir formar, em condições pré-bióticas, polímeros similares aos ácidos nucléicos atuais, há consenso em que a existência da molécula de RNA precede a de DNA. Várias são as linhas de evidências que apontam nesse sentido. Algumas dessas evidências são estruturais: maior flexibilidade resultante de sua composição, em geral como simples fita; e do açúcar componente dos nucleotídeos de RNA, a ribose (o DNA tem desoxirribose), que é mais reativa devido à hidroxila na posição 2'. Certamente, a enorme quantidade de funções desempenhadas pela molécula de RNA na célula é impressionante. Além de funções de mensageiro, transportador e ribossômico, já se sabe que a molécula de RNA desempenha vários outros papéis na célula e ainda há muito a ser descoberto. Entre as funções conhecidas, alguns exemplos merecem destaque: a molécula de RNA é parte essencial da telomerase, que adiciona o DNA das extremidades dos cromossomos; o RNA 7S é parte do sistema secretor celular; pequenas moléculas de RNA nucleares participam diretamente no processamento do mRNA (*splicing*); moléculas duplas fita de RNA participam diretamente na regulação da expressão de genes, em um processo conhecido como RNA de interferência (Fire, 2007, ver também Capítulo 4). Em algumas situações do metabolismo celular, também encontramos evidências de que o RNA é anterior ao DNA na evolução biológica. É o

caso do RNA iniciador ("primer"), necessário para que ocorra a própria síntese do DNA. Também na síntese dos precursores de DNA, verificamos que os desoxirribonucleotídeos são derivados de ribonucleotídeos a partir de reação catalisada por redutases de ribonucleotídeo. Da mesma forma, a síntese de timidina é realizada a partir da metilação de uridina. Finalmente, a atividade de transcriptase reversa encontrada em retrovírus sintetiza DNA a partir de RNA, em um processo que se acredita tenha ocorrido no início da evolução biológica, levando à origem das primeiras células contendo DNA.

2.3. O RNA Catalítico

Em 1977, foi descoberto que as sequências codificadoras de vários genes eram interrompidas por sequências não codificadoras. Essa descoberta foi baseada na comparação entre a sequência do DNA e de seu RNA correspondente. Essas sequências intercalantes foram denominadas introns, enquanto que as sequências codificantes foram denominadas exons. Após a transcrição, os introns têm que ser removidos do pré-RNA (*RNA splicing*, em inglês) para originar o RNA "maduro", que vai servir de molde para a tradução de uma proteína. Essa remoção tem que ser extremamente acurada para assegurar que os códons sejam lidos corretamente. Cada códon é composto por três nucleotídeos e corresponde a um aminoácido, de modo que, se for inserido ou retirado um nucleotídeo naquela sequência, isso acarretará um erro de leitura dos trios de nucleotídeos. As primeiras evidências da existência da descontinuidade do gene eucarionte pela existência de introns foi observada em adenovírus, no qual se verificou que os mRNAs maduros hibridizam em diferentes regiões do genoma de DNA do vírus. Essas descobertas deram o Prêmio Nobel a Richard J. Roberts e Philip A. Sharp em 1993.

Introns são comuns em eucariotos (no núcleo e nas organelas), mas também já foram encontrados em arquéias, eubactérias e em bacteriófagos. Nos eucariotos multicelulares, a maioria dos genes é interrompida e geralmente os introns são muito mais longos que os exons. Entretanto, não há uma regra e em levedura, por exemplo, a grande maioria dos genes não é interrompida. Além disso, a distribuição, o número e o tamanho dos introns variam enormemente (Roy e Gilbert, 2006).

Na década de 1980, foi descrito um grande número de introns, que foram separados em diferentes categorias, de acordo com suas características estruturais e os mecanismos de remoção dos introns do pré-RNA. No início da década de 1980, Cech e seus colaboradores mostraram que alguns introns são capazes de catalisar sua própria remoção do pré-RNA sem a ajuda de proteínas. Esses introns foram denominados de autocatalíticos (*self-splicing*, em inglês; Cech, 1988, 1990). Cech criou o termo ribozima para RNAs com propriedades catalíticas. O primeiro intron autocatalítico foi descrito no ciliado *Tetrahymena thermophila* e pertence ao chamado grupo I. Introns do grupo I são caracterizados por uma estrutura secundária altamente conservada e por seu mecanismo de excisão, em que o intron catalisa diretamente as duas reações de transesterificação consecutivas requeridas para sua excisão do transcrito primário (Cech e Bass, 1986). No primeiro estágio da reação de autocatálise, o grupo 3'-OH de uma guanosina (G) livre ataca a ligação fosfodiéster 3' do último nucleotídeo (em geral uma uridina) do exon 5'. Essa uridina está emparelhada com um nucleotídeo do intron, em geral uma guanina. Esse ataque resulta na quebra da ligação 5' entre o exon 5' e o intron. No segundo estágio, o grupo 3'-OH do exon 5' ataca a ligação fosfodiéster após o resíduo G terminal do intron, rompendo a ligação 3' intron/exon e se ligando ao exon 3' (Figura 2.1). As duas reações de transesterificação são catalisadas pelo próprio intron, que funciona como uma enzima (Michel e Westhof, 1990). Doudna *et al.* (1989) mostraram que a capacidade catalítica de introns do grupo I reside em suas estruturas secundária e terciária, e não na sua estrutura primária (sequência de nucleotídeos). Os introns do grupo II também podem ser autocatalíticos, mas esses introns apresentam sequências consenso e o mecanismo de excisão semelhantes aos introns que são removidos pela maquinaria riboprotéica (*spliceosome*, em inglês). O grupo de Sidney Altman, que compartilhou o prêmio Nobel de química em 1989 com Thomas Cech pela descoberta do RNA catalítico, demonstrou que a fração de RNA da ribonuclease P de *E. coli* era importante na catálise (Stark *et al.*, 1978)

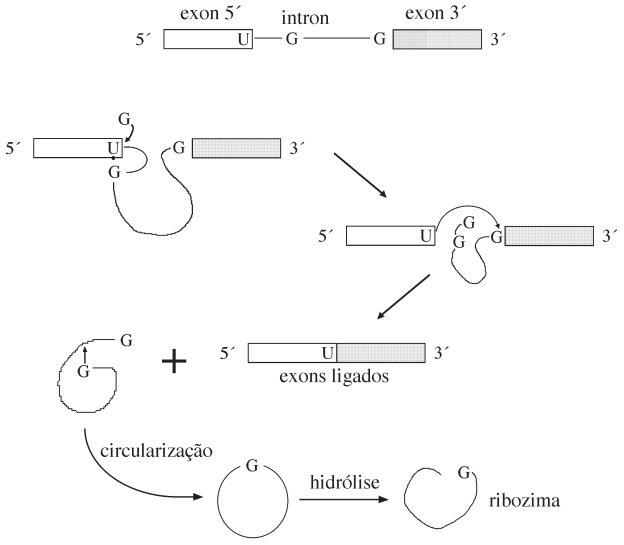


Figura 2.1. Esquema da reação de autoprocessamento de um intron do grupo I. O grupo OH-3' de uma guanina (G) livre ataca a ligação fosfodiéster 3' do último nucleotídeo (em geral uma uridina) do exon 5'. Essa uridina está pareada com uma guanina do intron. O ataque resulta na quebra da ligação 5' entre o exon 5' e o intron. O grupo OH 3' do exon 5' ataca a ligação fosfodiéster após o resíduo G terminal do intron, rompendo a ligação 3' intron/exon e se ligando ao exon 3'. As duas reações de transesterificação são catalisadas pelo próprio intron, que funciona como uma enzima e pode sofrer processos de circularização e hidrólise.

Apesar de algum ceticismo inicial, hoje a capacidade catalítica do RNA está plenamente comprovada. Existe uma crescente quantidade de dados experimentais demonstrando que moléculas de RNA são catalisadores surpreendentes e que sua ação não está confinada a substratos de ácidos nucléicos (Lazcano, 1994; Jeffares *et al.*, 1998; Landweber *et al.*, 1998). Vários trabalhos têm mostrado a participação de moléculas de RNA em diferentes atividades celulares. Potter *et al.* (1995) verificaram a existência de uma endonuclease na arquéia *Sulfolobus* que contém uma molécula de RNA que catalisa a excisão e a maturação de rRNA. Essa molécula de RNA é muito semelhante ao RNA U3 envolvido na maturação do mRNA em eucariotos e, segundo os autores, estaria presente antes da divergência entre as arquéias e os eucariotos, um verdadeiro fóssil molecular! Young *et al.* (1991) mostraram que a polimerase do RNA III do bicho-da-seda requer um fator de transcrição composto por RNA. Fung *et al.* (1995) apresentaram indícios de que pequenas moléculas de RNA citoplasmáticas (RNA G8) estão envolvidas em tolerância térmica do ciliado *Tetrahymena thermophila*.

Além disso, existem diversas coenzimas e grupos prostéticos compostos por ribonucleotídeos (como NAD e FAD) presentes em todos os seres vivos, os quais, na ausência da proteína correspondente, catalisam reações químicas similares àquelas que tomam parte como cofatores (Lazcano, 1994; Szathmáry, 1999). As várias atividades catalíticas das ribozimas serão detalhadamente discutidas no Capítulo 3. As capacidades de autoprocessamento, clivagem, elongação e ligação colocam o RNA num papel central na evolução pré-celular.

Obviamente, nem todas as moléculas de RNA são remanescentes do mundo de RNA. Para Jeffares *et al.* (1998), as moléculas consideradas fósseis teriam que apresentar uma ou mais das seguintes características: (1) ser catalítica—como as proteínas são melhores catalisadoras do que o RNA, é improvável que o RNA catalisador seja uma aquisição recente do metabolismo; (2) ser ubíqua, indicando que já estava presente no último ancestral comum de todos os seres vivos; (3) ter função central no metabolismo—qualquer RNA que ocupe uma posição central no metabolismo celular dificilmente seria substituído.

A existência do mundo de RNA requer ribozimas capazes de replicar RNA. Esse tipo de atividade catalítica foi demonstrado por Doudna e Szostak (1989). Além disso, seriam necessárias as capacidades de tomar matéria-prima do meio (supondo que as moléculas auto-replicativas de RNA já estivessem envoltas por uma membrana) e de coletar energia de outras moléculas com ligações de alta energia (Lazcano, 1994). O RNA não estaria solitário, mas acompanhado de diversas moléculas que poderiam funcionar como cofatores e substratos, incluindo íons metálicos, aminoácidos, polipeptídeos, açúcares e lipídios. Essa coexistência levaria à aquisição de outros grupos funcionais. A autoclivagem de alguns introns atuais, por exemplo, é dependente de Mg^{++}, o que sugere a existência de metalo-ribozimas (Gilbert, 1987). Outro exemplo de associação é a existência de um terpenóide hidrofóbico ligado a um ribonucleotídeo existente na membrana da bactéria púrpura *Rhodopseudomonas acidophila* (Neunlist e Rohmer, 1985). A existência desse composto sugere que uma associação direta entre RNA e lipídios pode ter existido, sendo que esse tipo de associação pode ter facilitado a encapsulação das moléculas de RNA (Lazcano, 1994).

A sequência de eventos que levaram à síntese protéica direcionada por RNA provavelmente começou com uma simples interação química entre aminoácidos e ribozimas, mas eventualmente levou a uma transformação completa da célula baseada em RNA (Lazcano, 1994). A primeira protocélula é, por definição, um sistema envolto por membranas, composto de macromoléculas capazes de auto-replicação e de catálise, com mecanismos de tomada de matéria-prima do meio e de obtenção de energia (Deamer *et al.*, 1994) (Figura 2.2). Os genomas de RNA das primeiras células teriam as seguintes propriedades, segundo Ohta (1994): a replicase do RNA seria ineficiente devido a uma alta taxa de erro em termos de substituição de nucleotídeos; o material genético e funcional seria o mesmo, mas as duas formas de RNA deveriam ter-se diferenciado logo no início, sendo o material genômico formado por RNA dupla fita (uma vez que RNA simples fita tem alta taxa de hidrólise, além do fato de que a transição de RNA para DNA seria facilitada se o RNA fosse dupla fita); várias funções genéticas já deveriam existir a partir da diversificação da primeira replicase do RNA; uma estrutura semelhante a um tRNA teria servido como marcação para a transcrição (nesse caso, copiar o RNA genômico para RNA funcional); o genoma de RNA deveria ter aumentado gradualmente para permitir uma maior diversidade funcional.

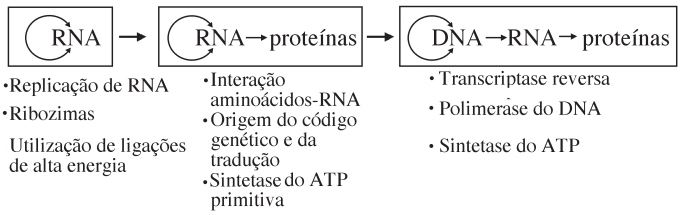


Figura 2.2. Esquema de um possível cenário de transição entre um organismo com metabolismo baseado em RNA até um organismo atual com metabolismo baseado em DNA, RNA e proteínas (modificado de Lazcano,1994).

2.4. O Aparecimento do Código Genético e a Transição para o "Mundo de RNP"

A origem do código genético e do sistema de tradução foi uma das principais transições na evolução e diversificação da vida. Essa transição modificou radicalmente os sistemas vivos, permitindo a divisão de trabalho entre os ácidos nucléicos (informação) e as proteínas (catálise).

A possibilidade de que genomas de DNA tenham aparecido antes do surgimento da síntese protéica não pode ser completamente descartada. De acordo com a hipótese do mundo de RNA, a síntese protéica mediada por ribossomos surgiu a partir da interação entre aminoácidos e RNA. Existem evidências de que aminoácidos e oligopeptídeos estavam presentes na "sopa" prébiótica na Terra primitiva (ver Capítulo 1). A síntese protéica é um processo complexo que requer muitos componentes, como rRNA, tRNA, mRNA e diversas proteínas, como fatores de elongação e iniciação, sintetases de aminoacil-tRNA, proteínas ribossômicas, entre outras.

Ainda não se sabe com certeza como as ligações peptídicas são formadas no ribossomo, porém existem algumas evidências de que o rRNA é o responsável pela catálise (Noller, 1991; Nitta *et al.*, 1998). Recentemente, foi selecionada uma ribozima capaz de catalisar a formação de uma ligação amida (Szathmáry, 1999). Um dos passos mais críticos na origem da síntese protéica é a formação de uma estrutura altamente complexa como o ribossomo (Poole *et al.*, 1998).

Geralmente é assumido que, no início do mundo de RNA, a precisão da replicação era limitada e que, por isso, as moléculas de RNA não deveriam ultrapassar algumas centenas de bases. À medida que a precisão da replicação foi aumentando, moléculas maiores puderam ser formadas. É possível que os vários sítios ativos dos ribossomos tenham se formado como ribozimas individuais, posteriormente reunidos por recombinação, formando os rRNA (Jeffares *et al.*, 1998; Poole *et al.*, 1998).

Segundo Poole *et al.* (1998), a síntese protéica baseada numa molécula-molde de RNA teria se originado a partir de uma ribozima com atividade de polimerase do RNA e que adicionasse trinucleotídeos (Figura 2.3). Considere uma molécula semelhante a um tRNA com trinucleotídeos na posição do anticódon; se estes forem complementares aos nucleotídeos presentes na fita de RNA-molde, poderiam emparelhar com ela e o trinucleotídeo poderia ser incorporado na nova fita. Uma vantagem de se adicionarem trinucleotídeos, ao invés de nucleotídeos isolados, seria que um número maior de ligações de hidrogênio manteria os nucleotídeos mais tempo no lugar. Como a catálise realizada por ribozimas é mais lenta do que a realizada por proteínas, esse maior tempo de emparelhamento seria bastante vantajoso. O problema da precisão de replicação, discutido anteriormente, poderia ser minimizado em parte com a adição de mais sítios de reconhecimento, como, por exemplo, a adição de um aminoácido ao tRNA. Ou seja, o

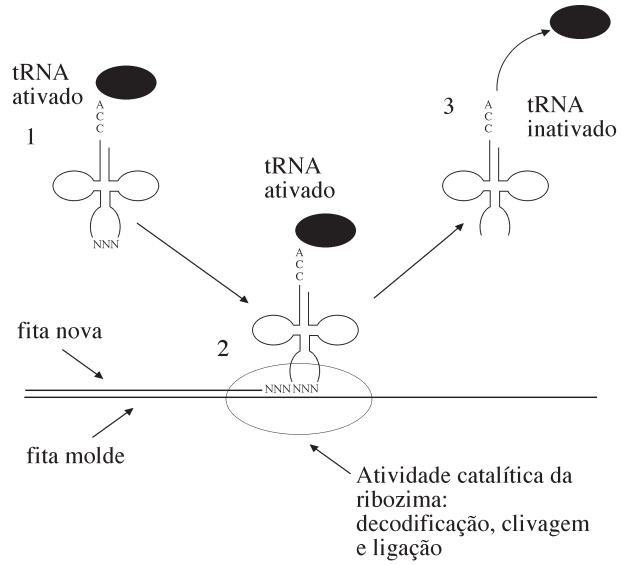


Figura 2.3. Modelo para a origem da síntese protéica baseada numa molécula molde de RNA. 1. Um aminoácido com carga positiva ajudaria uma replicase de RNA a reconhecer o tRNA, aproximando os dois; 2. O trinucleotídeo no anticódon é adicionado à nova fita de DNA pela replicase de RNA através de clivagem e ligação, semelhantes às realizadas pelos introns atuais; 3. O aminoácido é então clivado do tRNA e este é liberado, semelhante à atividade do 23S rRNA atual (modificado de Poole *et al.* 1998).

código genético já poderia ter sido estabelecido no mundo de RNA (Nagel e Doolitle, 1995; Wetzel, 1995). A afinidade do complexo de replicação por um tRNA ligado a um aminoácido poderia ser revertida com a clivagem do aminoácido, o que liberaria o tRNA. Esse complexo de replicação seria o proto-ribossomo.

A vantagem desse modelo é que várias das funções catalíticas presumidas podem ser testadas com a evolução *in vitro* de ribozimas. Uma ribozima desenvolvida *in vitro*, por exemplo, foi capaz de ligar um aminoácido a um tRNA (Illangasekare *et al.*, 1995).

A origem da informação (mRNA) provavelmente é o passo mais difícil de se explicar. Segundo Poole *et al.* (1998), os mRNAs podem ter surgido como produtos secundários do processamento de RNA. Nesse modelo, as ribozimas seriam removidas do pré-RNA e as regiões flanqueadoras seriam reunidas, dando origem a novas sequências. As ribozimas atuais estão, de modo geral, localizadas nos introns, motivo pelo qual os autores chamam essa hipótese de introns precoces (*introns-first*, em inglês). Ou seja, os introns que corresponderiam às moléculas de RNA catalíticas teriam surgido antes que seus exons flanqueadores. A função dos mRNAs teria surgido a partir de fragmentos de sequências que teriam sido juntados secundariamente.

Pequenas moléculas de RNA nucleolar (snoRNAs, *small nucleolar RNA*, em inglês) são processadas a partir de introns encontrados nos genes que codificam para proteínas ribossômicas e para chaperonas. Com base nesse fato, Poole *et al.* (1998) propõem que essas proteínas –que são universais– estariam provavelmente entre as primeiras proteínas a ter surgido na transição entre o mundo de RNA e o metabolismo das células atuais, baseado em proteínas.

As primeiras proteínas deveriam interagir com o RNA com baixa especificidade e deveriam atuar como chaperonas, ou seja, deveriam auxiliar ou facilitar o correto dobramento da molécula catalítica de RNA (ribozima). Muitas das chaperonas atuais estão envolvidas na resposta ao choque térmico e são denominadas de HSP (*Heat Shock Proteins*, em inglês). Poole *et al.* (1998) incluem na categoria de chaperonas as moléculas que se ligam a RNA e que não são em si catalíticas, como as proteínas ribossômicas e aquelas ligadas à remoção de introns, entre outras.

Polipeptídeos carregados positivamente ligar-se-iam às moléculas de RNA carregadas negativamente, aumentando sua estabilidade. O aumento da estabilidade nas estruturas terciárias das ribozimas, que sem as proteínas seriam bastante dependentes das concentrações de íons no meio (por exemplo, Mg^{++}), permitiria um aumento na precisão da replicação e, consequentemente, um aumento do tamanho das moléculas de RNA sendo replicadas. Esse aumento na precisão de replicação da informação é fundamental para o surgimento da síntese protéica.

As proteínas são catalisadores mais eficientes e rápidos que o RNA, pois possuem um número muito maior de grupos funcionais (20 aminoácidos) e a capacidade de manter uma estrutura terciária precisa (Jeffares *et al.*, 1998). Atualmente, são raros os catalisadores formados unicamente por RNA (alguns introns autocatalíticos e ribozimas virais). Na maioria dos casos, os RNAs catalisadores estão associados a proteínas que ajudam a manter uma estrutura terciária correta. A estrutura terciária de moléculas de RNA varia com a concentração de íons no meio, daí a necessidade de interação com proteínas. Os ribossomos atuais parecem ser exatamente isso, ribozimas estabilizadas por proteínas.

Uma vez que as proteínas se apresentam muito mais eficientes como catalisadores do que as ribozimas, sua síntese e utilização seriam vantajosas para os organismos (Jeffares *et al.*, 1998). A partir dessa interação entre os polipeptídeos e as moléculas de RNA, teria surgido o chamado mundo de RNP (RNP = RNA + proteínas; Figura 2.2).

2.5. Transição para o "Mundo de DNA"

Atualmente, há uma grande aceitação da hipótese do mundo de RNA. Apesar disso, todas as células vivas conhecidas têm como material informativo o DNA. Dos seres vivos conhecidos, apenas os vírus podem apresentar o RNA como portador de informação genética, podendo existir com fitas simples ou duplas dessa molécula. Proteínas podem ser sintetizadas na ausência de DNA, mas não na de RNA. É, portanto, razoável assumir que os genomas de DNA surgiram posteriormente à síntese protéica e que seriam monofiléticos, desenvolvendo-se antes da divergência das três linhagens celulares (eubactérias, arquéias e eucariotos). O DNA dupla fita é uma molécula extremamente resistente; os genomas de DNA teriam sido selecionados, ao invés dos genomas de RNA, pela simples razão de serem mais estáveis (Lazcano, 1994). Além disso, a informação está duplicada (em cada uma das fitas de DNA), o que facilitaria o reparo com precisão, em caso de dano em uma das fitas. Por serem mais estáveis, os genomas de DNA puderam aumentar de tamanho através de duplicação gênica. A duplicação de RNA é um processo intrinsecamente pouco fiel, o que limita o tamanho das cópias, uma vez que o número de mutações pontuais é proporcional ao tamanho do molde (Lazcano, 1994).

O surgimento de genomas de DNA e de polimerases do DNA de alta fidelidade possibilitaram o desenvolvimento de genomas maiores, com capacidade de codificação aumentada. O aparecimento do DNA possibilitou a duplicação de genes em grande escala e o embaralhamento dos exons, gerando proteínas com novas capacidades catalíticas. Isso, por sua vez, possibilitou uma grande diversificação das formas de vida e seu aumento de

complexidade. Assim, acredita-se que células contendo RNA como material genético devem ter existido com uma capacidade metabólica limitada e lenta, o que as restringiram na competição com as células emergentes contendo DNA, resultando em sua extinção gradual.

O processo de transferência de informação genética do RNA para o DNA ocorreu graças à atividade de enzimas conhecidas como transcriptase reversa. Essas enzimas foram inicialmente encontradas em retrovírus, que constituem tipos de vírus que possuem, em seu ciclo de vida, moléculas de RNA empregadas como molde para intermediários de DNA. A existência de transcriptase reversa, no entanto, não está restrita a retrovírus. Esse tipo de atividade enzimática tem sido descrito também em células eucarióticas e procarióticas, indicando sua ancestralidade. É possível, portanto, que, em um mundo de células contendo RNA como material genético, as proteínas já existissem como determinantes importantes do metabolismo celular (mundo de RNP) e que atividades de enzimas como a transcriptase reversa converteriam o genoma, ou parte dele, em DNA. Um fato interessante é que a telomerase, que constitui uma enzima importante na síntese das extremidades repetitivas dos cromossomos de eucariotos (os telômeros), realiza sua função empregando uma molécula de RNA como molde da região repetitiva. Essa ribonucleoproteína sintetiza DNA a partir de RNA, sendo, portanto, uma transcriptase reversa. Assim, acredita-se que essa enzima seja um dos fósseis moleculares remanescentes do mundo de RNP, além de indicar que a função de transcriptase reversa pode também ser realizada com atividades catalíticas de RNAs, ou seja, ribozimas. Essa conversão de RNA em DNA permitiu a origem de células com metabolismo próximo ao que conhecemos hoje e deve ter tido um papel determinante na origem das células atuais, constituindo o que chamamos hoje de *progenota* (ver Capítulo 1).

2.6. O Aumento da Complexidade

Estudos de filogenia molecular utilizando o gene que codifica para o RNA da subunidade pequena do ribossomo (SSU rDNA, também chamado 16S nos procariotos e 18S nos eucariotos) feitos por Woese (1987) e Woese *et al.* (1990) transformaram a dicotomia eucarioto/procarioto em um sistema de três domínios: Bacteria, Archaea e Eucaria (neste capítulo, serão usadas as designações eubactéria, arquéia e eucariotos, respectivamente—Figura 2.4).

Atualmente, muitos caracteres moleculares e fenotípicos indicam que os eucariotos e as arquéias formam um grupo monofilético, irmão das eubactérias. Entre as evidências que indicam uma ancestralidade comum exclusiva entre arquéias e eucariotos, podemos citar a presença de proteínas semelhantes às histonas associadas ao DNA, a presença de moléculas semelhantes a esteróides em um grupo de arquéias (também denominados de eócitos), a semelhança de várias proteínas e a semelhança de vias metabólicas. Entretanto, as relações entre os três domínios ainda é bastante controversa (Katz, 1998; Doolittle, 1999a; Nelson *et al.*, 1999; Pace, 2004).

Frequentemente é assumido que os procariotos são anteriores aos eucariotos devido a sua aparente simplicidade, sua presença anterior no registro fóssil e também com base em estudos filogenéticos. Nesse cenário, as características complexas dos eucariotos, como a compartimentalização nuclear, organelas membranosas e processamento de mRNA para remoção de introns, seriam aquisições tardias. Os procariotos obviamente antecedem os eucariotos modernos que possuem mitocôndria (Forterre e Philippe, 1999). Entretanto, Poole *et al.* (1998) sugerem que o genoma do ancestral comum mais antigo seria linear, capaz de recombinação, fragmentado e repleto de fósseis moleculares, ou seja, mais parecido com um eucarioto do que com procariotos.

As moléculas atuais de RNA que apresentam capacidade catalítica seriam relíquias do mundo de RNA (Jeffares *et al.*, 1998), ou seja, de um período anterior à divisão do último ancestral comum que deu origem às linhagens dos organismos atuais. Poole *et al.* (1998) utilizam essas moléculas fósseis para enraizar a origem da árvore da vida. Segundo esses autores, o genoma do tipo eucariótico seria anterior ao tipo procariótico, visto que o genoma eucariótico contém um maior número de fósseis moleculares (introns autocatalíticos, *spliceosomes*, snoRNAs, telomerase, entre outros—ver Jeffares *et al.*, 1998). Além disso, a transcrição e a tradução seriam muito mais rápidas e eficientes nos procariotos. A origem de um genoma procariótico a partir de um eucariótico seria relativamente simples e direta se uma forte seleção para ambientes termofílicos fosse considerada. O ambiente termofílico favoreceria um rápido processamento do RNA e sua subsequente tradução, considerando que as taxas de hidrólise do RNA aumentam com a temperatura. Outra possibilidade seria uma forte seleção no sentido de altas taxas reprodutivas, pequeno tamanho e ciclos de vida curtos, que frequentemente são encontrados em ambientes instáveis. Os efeitos combinados de uma adaptação à termofilia e uma pressão seletiva para ciclo de vida rápido teria levado à perda dos fósseis moleculares e a uma simplificação no processamento e tradução dos RNAs nos procariotos (Darnell e Doolitle, 1986; Poole *et al.*, 1998).

Alguns autores (Poole *et al.*, 1998; Forterre e Philippe, 1999) argumentam que a confiabilidade de métodos filogenéticos (que indicariam uma origem procariótica da vida) na recuperação de divergências tão antigas está sujeita a controvérsias e que a semelhança dos fósseis mais antigos (estromatólitos de cerca de

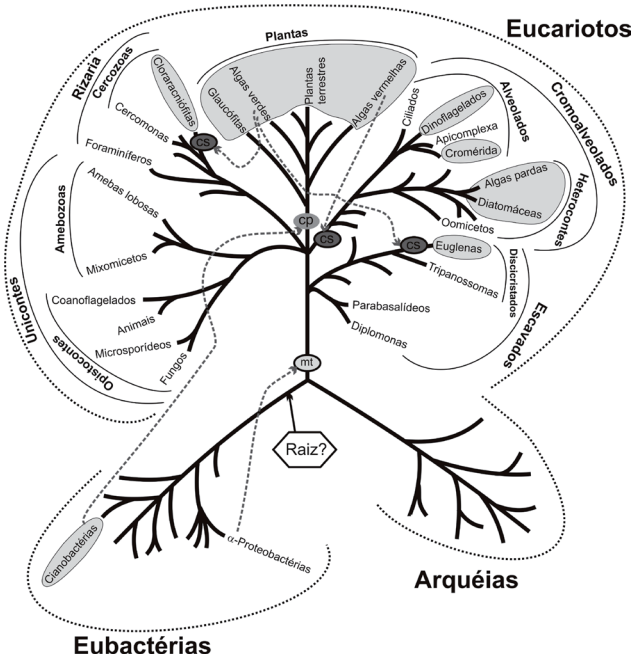


Figura 2.4. Árvore filogenética esquemática baseada em dados moleculares e fenotípicos mostrando os três domínios de seres vivos (Eubactéria, Arquéia e Eucariotos) com ênfase nas principais linhagens eucarióticas. As linhagens filogenéticas fotossintetizantes estão circundadas em fundo cinza. As linhas pontilhadas indicam os eventos de endossimbiose que deram origem à mitocôndria (mt), ao cloroplasto primário (cp) e aos cloroplastos secundários (cs). A possível posição da raiz, indicando um ancestral comum para Arquéia e Eucariotos, está indicada na árvore (modificado de Baldauf *et al.*, 2004; Bellorin e Oliveira, 2006).

3,8 bilhões de anos, ver Capítulo 1) às atuais cianobactérias não seria conclusiva.

Os genomas dos eucariotos podem apresentar uma enorme complexidade, com regiões espaçadoras, introns, regiões repetitivas, elementos de transposição e famílias multigênicas (ver Capítulos 9 e 10). Essa grande complexidade é em parte possível através da duplicação de genes e processos de recombinação (Ohta, 1994; ver também Capítulo 10). A duplicação gênica possibilita a ocorrência de variabilidade e a diversificação das cópias. Se essas cópias ainda mantêm funções relacionadas, originam as chamadas famílias gênicas. Quando a diversificação é muito grande, novos genes são gerados, que codificam proteínas com novas funções. Novos genes podem ser gerados também pelo mecanismo chamado embaralhamento de exons (*exon-shuffling*, em inglês), onde os exons funcionariam como módulos que, mudando de posição no genoma, formariam genes de propriedades diferentes. A hipótese de embaralhamento de exons foi proposta por Gilbert (1978), mas os mecanismos pelos quais isso acontece permaneceram obscuros até recentemente, quando Moran *et al.* (1999) mostraram que o embaralhamento de exons pode ocorrer através da mobilização de retrotransposons.

Lazcano e Miller (1996) sugerem que a maioria dos genes teria surgido a partir de duplicação gênica. Baseados nas semelhanças entre vias metabólicas e nas funções relacionadas de proteínas, os autores estimam que entre 20 e 100 genes iniciais devem ter coexistido no progenota, de onde vem o interesse em se estudarem os genomas mínimos para a manutenção da vida, mencionados no Capítulo 1.

Além do aumento da complexidade na estrutura genética, ocorreu também um aumento da complexidade celular, com o surgimento de diferentes organelas, que delimitam distintos compartimentos internos. Pelo menos duas dessas organelas, mitocôndrias e cloroplastos, são derivadas de associações endossimbióticas entre os eucariotos e outros organismos (procariotos e eucariotos). Esses eventos endossimbióticos introduziram genomas inteiros no interior da célula hospedeira, possibilitando a transferência lateral de genes. A transferência lateral de genes (também chamada de transferência horizontal) é o processo pelo qual o material genético é transferido de um outro organismo para outro por uma outra via que não seja a reprodução, em contraposição à transferência vertical, que ocorre quando um organismo recebe seu material genético de seu ascendente. A genética clássica considerava a via de transferência vertical de genes como única relevante para a origem e evolução das espécies, mas recentemente essa visão está mudando em função dos dados obtidos a partir do sequenciamento de genomas.

Os Projetos Genoma têm acumulado uma quantidade gigantesca de dados. As análises e comparações de genomas como um todo estão ainda no seu início, mas já têm causado agitação em diversas áreas. A quantidade de eventos de transferência lateral sugerida a partir dessas análises é surpreendente e provavelmente teve um papel fundamental na evolução dos organismos.

Nelson *et al.* (1999) fizeram uma comparação de 33 genes dos quais foram encontradas cópias homólogas em todas as espécies já sequenciadas à época. A maioria dos genes estudados sugere que as arquéias constituem um grupo monofilético, separado das eubactérias, padrão também encontrado para o SSU rDNA. A maioria dos genes de levedura (eucarioto) agrupa-se com os genes de arquéias, resultado também encontrado para o SSU rDNA. Entretanto, as árvores geradas para os diferentes genes apresentam uma falta de congruência significativa entre si. Os autores supracitados atribuem essa falta de congruência principalmente a mecanismos como duplicação, perda e transferência lateral de genes. Verificaram ainda que, na eubactéria *Thermotoga maritima*, 52% de seus genes são mais semelhantes a genes de outras eubactérias, mas 24% são mais semelhantes aos genes de arquéias. Eles atribuem essa alta similaridade entre arquéias a uma extensiva transferência lateral de genes, argumentando que esses não estão distribuídos uniformemente nas diferentes categorias, nem nas diferentes regiões do genoma. Além disso, a ordem de distribuição de alguns genes e também algumas regiões repetitivas só foram encontradas em arquéias. Apesar de *T. maritima* ter um genoma essencialmente eubacteriano, quase um quarto de seu genoma parece ser resultado de um ou mais eventos de transferência lateral de genes provenientes de arquéias.

A maioria dos genes envolvidos na estrutura do genoma, na transcrição e na tradução claramente separa as arquéias das eubactérias. Os produtos desses genes apresentam mais interações macromoleculares, o que dificulta enormemente sua fixação em novos hospedeiros após eventos de transferência lateral (Shi e Falkowski, 2008). Nesse caso, as filogenias refletem aqueles marcadores que são menos propensos à transferência lateral e explicam a presença de muitos genes em arquéias que são próximos às eubactérias como resultado de transferência lateral (Gogarten *et al.*, 1999). A principal questão atual é estabelecer o impacto desses eventos na evolução microbiana e, em particular, na nossa habilidade de reconstruir a história evolutiva dos organismos. Alguns autores propõem que, ao invés da herança vertical, a transferência lateral seja o principal determinante taxonômico em procariotos (Olendzenski *et al.* 2002). Doolittle (1999a) afirma que a transferência lateral de genes teve e tem um papel crucial na formação dos seres vivos e que as relações filogenéticas formam uma rede intrincada (Figura 2.5). Discutiremos em detalhes, a seguir, a questão da origem de células eucarióticas.

2.6.1. A origem do núcleo

A hipótese mais antiga—e talvez ainda a mais aceita—para a origem do núcleo é a hipótese autógena, em que o núcleo teria se originado a partir de uma organização gradual de membranas ao redor do material genético. A membrana nuclear é, em muitos casos, contínua ao retículo endoplasmático e teria se originado diretamente a partir deste (Dyer e Obar, 1994). Uma evidência que favorece a hipótese autógena é que, em muitos eucariotos, a membrana nuclear é completamente desintegrada durante a divisão celular e é formada novamente nas células-filhas. Além

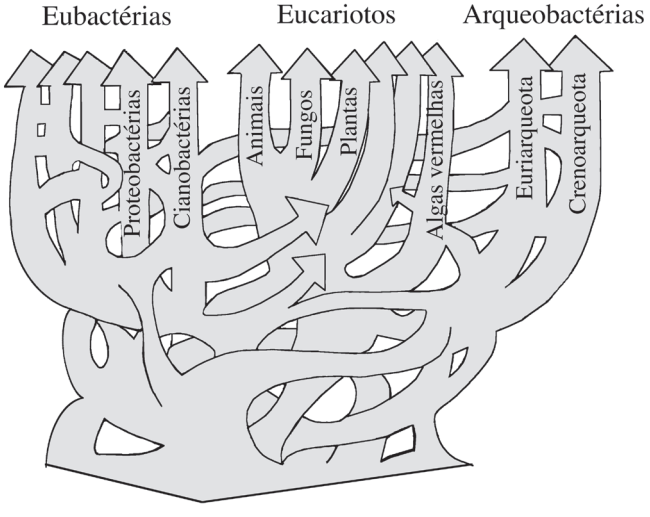


Figura 2.5. Árvore filogenética reticulada representando possíveis eventos de transferência lateral entre os três domínios (adaptado de Doolittle, 1999a).

disso, o envelope nuclear não é composto por duas membranas, mas por uma série de vesículas achatadas (Martin, 1999).

Esse sistema de membranas internas que teria circundado o material genético teria se originado por invaginações da membrana plasmática pelo mesmo sistema que permitia a fagocitose. Ou seja, a célula que deu origem aos eucariotos não deveria apresentar parede celular e já deveria ter um sistema de microtúbulos (citoesqueleto) antes da formação do núcleo. Nos últimos anos, o acúmulo de sequências genômicas levou à caracterização de diversas proteínas, tanto em arquéias quanto em eubactérias, que apresentam homologia com as proteínas do citoesqueleto eucariótico, incluindo actinas e tubulinas bacterianas. A detecção dessas proteínas através de marcadores citológicos revelou um sistema complexo de citoesqueleto bacteriano e mostrou que essas proteínas poliméricas estão envolvidas em diferentes processos nas células procarióticas, incluindo divisão celular, manutenção da forma celular e segregação de moléculas de DNA. Entretanto, não se conhece qualquer procarioto que tenha a capacidade de realizar fagocitose (Pogliano, 2008). Nos eucariotos, o citoesqueleto, formado por actina, tubulina e os filamentos intermediários, é responsável pela manutenção da forma celular, pela movimentação da célula e de seus componentes internos (por exemplo, a movimentação dos cromossomos durante a divisão celular).

Sogin (1994) propôs um modelo de origem nuclear quimérico para explicar as divergências encontradas nas árvores filogenéticas, geradas usando-se sequências de diferentes genes para estabelecer a relação dos três domínios, eubactéria, arquéia e eucariotos. Nesse modelo, o progenota já apresentava um sistema primitivo de tradução, mas os eventos celulares ainda eram dominados pelo RNA. Uma linhagem celular, a partir do progenota, teria adquirido uma complexidade do citoesqueleto suficiente para permitir a transição para uma célula nucleada.

Ainda de acordo com o modelo de Sogin, uma outra linhagem celular teria desenvolvido um sistema sofisticado de tradução e possivelmente teria substituído o RNA catalisador pelas proteínas e o RNA repositório da informação pelo DNA. Essa segunda linhagem teria se diferenciado nas arquéias e eubactérias. O citoesqueleto da primeira linhagem teria permitido o englobamento em outros organismos. Essa linhagem proto-eucariótica teria então englobado uma arquéia, a qual teria dado origem a um núcleo quimérico, que incluía o genoma de DNA das arquéias (contribuindo principalmente com os genes para a tradução e para as proteínas) e o genoma de RNA do proto-eucarioto (contribuindo com os rRNAs e a informação para o citoesqueleto; Figura 2.6). Margulis *et al.* (2000) propuseram uma idéia semelhante, em que o ancestral comum a todos os eucariotos teria se originado da fusão dos genomas de dois ou mais procariotos distintos a partir de uma associação simbiótica.

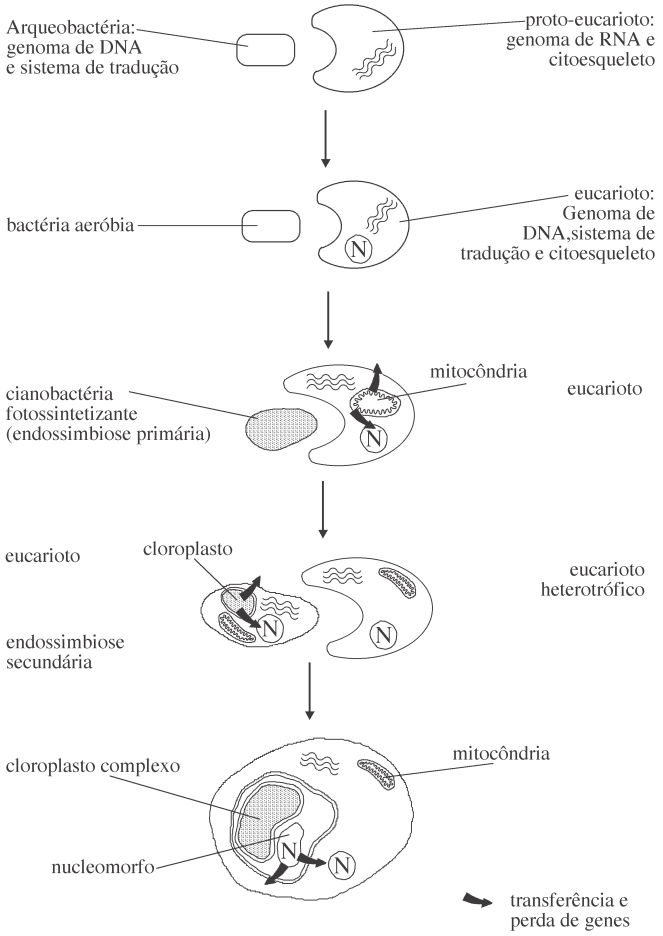


Figura 2.6. Esquema de um possível cenário para a origem e evolução dos eucariotos, baseado em várias hipóteses referidas no texto.

2.6.2 A origem das organelas

A fotossíntese baseada em clorofila **a**, com desprendimento de oxigênio na sua forma molecular (O_2), é possivelmente o evento mais significativo na história da vida na Terra, depois da origem da vida (Canfield, 2005). O oxigênio é um oxidante potente, cujo acúmulo alterou a atmosfera terrestre e o rumo da história evolutiva da vida no planeta.

A fotossíntese baseada em clorofila **a** é restrita a um grupo de eubactérias, as cianobactérias. Os eucariotos fotossintetizantes adquiriram essa capacidade através de endossimbiose através das cianobactérias. Dentro das eubactérias, cinco linhagens diversas são fotossintetizantes (as cianobactérias e outras quatro linhagens que possuem bactério-clorofilas, que não produzem oxigênio), levando à especulação de que o ancestral desse domínio seria fotossintetizante (Woese, 1987; Pierson, 1994). Entretanto, a ausência de fotossíntese baseada em clorofila nas arquéias argumenta contra a presença de fotossíntese no progenota. Além disso, a molécula de clorofila **a** e a sua associação em fotossistemas são extremamente complexas para estarem presentes no progenota. Isso não significa, no entanto, que o progenota não usava luz como fonte de energia, o que poderia ser feito através de moléculas captadoras de energia radiante bem mais simples (Deamer *et al.*, 1994).

Com o advento da fotossíntese, o O_2 começou a ser liberado para a atmosfera terrestre e foi se acumulando gradativamente. Por volta de 2,5 a 2,2 bilhões de anos atrás, o oxigênio já deveria estar presente em pequena quantidade (em torno de 0,5%), possibilitando a respiração aeróbica (Knoll, 1992; Bekker *et al.*, 2004). Essa transição para uma atmosfera rica em oxigênio foi denominada de "o grande evento de oxidação" (Canfield, 2005). Nessa época, devem ter-se originado os ancestrais da linhagem que, mais tarde, originaria as mitocôndrias. Nesse período, teriam surgido também as primeiras defesas celulares contra os efeitos tóxicos do oxigênio, então um gás letal para a maioria das formas de vida existentes no planeta. Apenas há cerca de 2 bilhões de anos é que o oxigênio deve ter-se acumulado na atmosfera em quantidades suficientes para formar uma camada de ozônio (O_3), a qual diminuiu a incidência de raios ultravioleta sobre a superfície terrestre (Dyer e Obar, 1994).

As células eucarióticas, como conhecemos hoje, isto é, células nucleadas e com organelas, teriam surgido de eventos de endossimbiose (simbiogênese sendo o surgimento de uma nova linhagem de organismos como consequência de uma associação simbiótica estável) entre uma célula hospedeira e células procarió-

ticas que deram origem às mitocôndrias e aos cloroplastos (Figura 2.6). O termo *simbiogênese* foi introduzido pelo biólogo russo Mereschkovsky em 1909 (Margulis e Cohen, 1994). A principal implicação da simbiogênese é que os eucariotos seriam, de fato, quimeras produzidas pela combinação de distintos genomas.

As evidências que apóiam a hipótese de simbiogênese são: (1) as proteínas presentes nas organelas são mais semelhantes a seus análogos procarióticos que aos eucarióticos; (2) existem procariotos de vida livre com forte semelhança estrutural, bioquímica e genética com as organelas; (3) as organelas possuem genoma próprio, com organização semelhante ao genoma procariótico; (4) os RNAs (ribossômico, transportador e mensageiro) das organelas também são mais semelhantes aos de procariotos; (5) as organelas são semi-independentes, com capacidade de replicação; (6) as organelas e suas funções estão, alternativamente, presentes ou ausentes das células eucarióticas, não sendo encontrados intermediários (Gray, 1992; Dyer e Obar, 1994; Martin, 1999).

Uma vez ocorrida a endossimbiose, genes do endossimbionte podem ser transferidos lateralmente para o núcleo da célula hospedeira. Os produtos desses genes seriam, então, direcionados às organelas. Mitocôndrias e cloroplastos são, portanto, semi-independentes, já que necessitam dos produtos de genes que agora são codificados no núcleo. Em alguns complexos enzimáticos, uma parte das subunidades é codificada pelo genoma nuclear e uma parte pelo genoma da organela (como a sintetase do ATP na mitocôndria e Rubisco—carboxilase/oxigenase da ribulose-bisfosfato—no cloroplasto). Transferências de genes entre mitocôndrias e cloroplastos na mesma célula também podem ter ocorrido. A transferência específica de genes entre os compartimentos celulares varia nos diferentes organismos. O mecanismo pelo qual se dá esta transferência lateral no interior das células não está estabelecido, mas, em alguns casos, elementos de transposição poderiam estar envolvidos. Esses eventos de transferência intensificam a dependência entre organelas e núcleo, e provavelmente são essenciais para a manutenção da associação endossimbiótica (Dyer e Obar, 1994). Apesar de eventos de transferência e perda de genes ocorrerem provavelmente ao acaso, aparentemente existe uma direcionalidade, visto que o núcleo apresenta uma tendência de adquirir genes, enquanto que as organelas parecem perder genes redundantes. Uma explicação para essa direcionalidade seria que o núcleo é um ambiente geneticamente mais estável que as organelas.

A maioria das linhagens eucarióticas tem mitocôndrias que devem ter sido adquiridas através de um evento de endossimbiose entre uma célula eucariótica e uma bactéria púrpura (α-proteobactéria), provavelmente há cerca de 2,5 bilhões de anos (Dyer e Obar, 1994). O parente atual mais próximo das mitocôndrias é a α-proteobactéria *Rickettsia*, um parasita intracelular causador do tifo, cujo genoma foi sequenciado há pouco mais de uma década (Andersson *et al.*, 1998).

As mitocôndrias apresentam duas membranas, a externa, normalmente lisa, e a interna, com dobramentos que podem assumir diversas conformações (tubular, vesicular ou lamelar). Apesar da variação de forma, número e tamanho das mitocôndrias, as sequências moleculares têm mostrado uma origem única, indicando não mais de um evento de aquisição de mitocôndrias por endossimbiose, seguido de algumas raras perdas secundárias (Dyer e Obar, 1994).

Existem linhagens eucarióticas que não possuem mitocôndrias e muitas dessas linhagens parecem ter divergido antes da maioria das linhagens atuais de eucariotos (Figura 2.4). Entretanto, estudos mais recentes questionam a posição basal nas árvores filogenéticas de alguns arqueozoários (Keeling, 2007). Esses protistas sem mitocôndria foram reunidos em um grupo denominado Arqueozoa (reunindo ameboflagelados, diplomonas, retortomonas, microsporídios e tricomonas), que se acreditava ter divergido antes da aquisição de mitocôndrias. Entretanto, nos últimos anos, genes tipicamente mitocondriais têm sido encontrados no núcleo de todas essas linhagens, indicando que ocorreu uma perda secundária da mitocôndria (Doolittle, 1999b) e, portanto, esses não seriam grupos basais na evolução dos eucariotos. Além disso, proteínas codificadas por esses genes foram localizadas em organelas envoltas por duas membranas, que produzem hidrogênio e ATP—os hidrogenossomos (ou mitossomos) — em todos os principais grupos de arqueozoa. Os hidrogenossomos estão relacionados ao metabolismo fermentativo em vários protistas anaeróbicos. Apesar de a maioria dos hidrogenossomos não possuir um genoma, foi encontrado DNA nessas organelas em alguns ciliados anaeróbicos. Com base nesses genes, foi verificado que os hidrogenossomos parecem ser um tipo de mitocôndria anaeróbica (Embley e Martin, 1998; Martin, 1999).

Atualmente não existe um forte candidato a um eucarioto que nunca tenha possuído uma mitocôndria. Segundo Clark (1999), talvez a origem das mitocôndrias, inversamente, tenha sido a força motriz para a origem da membrana nuclear e, consequentemente, dos eucariotos propriamente ditos. Muitos dos organismos sem mitocôndria são parasitas internos (por exemplo, *Giardia*). Isso sugere que tenha havido nesses grupos uma perda secundária da mitocôndria, já que os ambientes nos quais esses parasitas vivem são pobres em oxigênio (Taylor, 1994). Martin (1999) argumenta que, se não forem encontrados eucariotos atuais que nunca tiveram mitocôndrias, a ordem dos eventos que levaram à origem da célula eucariótica deve ser revista.

A hipótese tradicional sustenta que a célula hospedeira, quando adquiriu sua mitocôndria, já possuía um núcleo. Martin e Müller (1998) sugerem que o endossimbionte que deu origem às mitocôndrias e hidrogenossomos era uma α-proteobactéria anaeróbica facultativa, com considerável flexibilidade metabólica. A célula hospedeira seria semelhante aos atuais metanogêneos (uma arquéia), os únicos procariotos conhecidos que possuem histonas verdadeiras (as proteínas associadas ao DNA, encontradas nos eucariotos).

Baseado no que foi visto até o momento, uma ordem de origem das estruturas eucarióticas mais lógica seria: primeiro, o citoesqueleto, possibilitando a fagocitose; depois, a origem da mitocôndria; a seguir, o sistema de endomembranas; e, por fim, o compartimento nuclear. Todos os organismos que possuem cloroplastos também possuem mitocôndrias, o que sugere que as mitocôndrias precederam os plastos. Entretanto, isso também pode indicar que as mitocôndrias são obrigatórias para a manutenção de plastos (Dyer e Obar, 1994).

A época de origem dos eucariotos tem sido estimada com base no tamanho das células do registro fóssil (Runnegar, 1994). A descoberta de um fóssil denominado *Grypania spiralis*, com idade estimada em 2,1 bilhões de anos, interpretado como uma alga eucariótica fotossintetizante, indica que a origem do cloroplasto por endossimbiose teria ocorrido antes dessa data (Han e Runnegar, 1992). O fato de as algas serem relativamente derivadas entre os eucariotos, entretanto, apontaria para uma origem dos eucariotos ainda um tanto anterior.

A associação simbiótica entre organismos autótrofos e heterótrofos é extremamente comum na natureza. Diversos animais e protozoários apresentam associações com microalgas (como corais, esponjas, ascídias e foraminíferos), os fungos associam-se às algas, formando os líquens. Em ambientes iluminados e ricos em nutrientes, a fotossíntese tende a ser superprodutiva, a ponto de haver excesso de produção para o organismo hospedeiro (Dyer e Obar, 1994; Margulis e Cohen, 1994).

Existem vários tipos de cloroplastos, que diferem em sua forma, ultraestrutura e pigmentação. Entretanto, sequências de rRNA e de vários outros genes também indicam uma origem única para todos os cloroplastos, os quais teriam surgido a partir de um evento endossimbiótico entre uma célula eucariótica (célula hospedeira) e uma cianobactéria semelhante a *Synechococcus* (Dyer e Obar, 1994; Delwiche e Palmer, 1997). Esse resultado foi uma surpresa para alguns pesquisadores, que consideravam a diversidade pigmentar e estrutural encontrada nos organismos fotossintetizantes atuais como indicação de uma origem polifilética dos cloroplastos. Atualmente, a origem endossimbiótica dos cloroplastos já está plenamente estabelecida, embora ainda exista controvérsia quanto ao número de eventos de endossimbiose que levou à formação dos diferentes tipos de cloroplastos. As árvores filogenéticas indicam um único evento primário de endossimbiose. Entretanto, esses dados também poderiam ser interpretados como eventos múltiplos, que ocorreram em um grupo de cianobactérias filogeneticamente próximas. Além do evento de endossimbiose primário, múltiplos eventos de endossimbiose secundários teriam ocorrido nas linhagens eucarióticas fotossintetizantes. Um evento de endossimbiose secundário é aquele em que uma célula eucariótica engloba uma outra célula eucariótica que já continha um cloroplasto (produto do evento primário). Eventos de endossimbiose secundários geraram cloroplastos complexos, com mais de duas membranas (três ou quatro). Em alguns casos, ainda é possível verificar a presença de um núcleo vestigial (chamado de *nucleomorfo*) entre a segunda e a terceira membranas do cloroplasto (Figura 2.6). Na endossimbiose secundária, o núcleo da célula englobada sofre o mesmo processo de redução que aquele postulado para as mitocôndrias, com perda e transferência lateral de genes para o núcleo da célula hospedeira.

O número de membranas ao redor dos cloroplastos é interpretado como indicativo de uma origem primária ou secundária. Os cloroplastos com duas membranas são considerados como produto de endossimbiose primária, sendo que as duas membranas nesses plastos primários têm sido explicadas de diferentes maneiras. A explicação mais tradicional diz que a membrana mais interna é de origem procariótica (membrana plasmática) e a mais externa de origem eucariótica (membrana do fagossomo). Uma explicação alternativa é que ambas as membranas do cloroplasto representam a dupla membrana típica de bactérias Gram-negativas e assume que o fagossomo foi perdido (Bellorin e Oliveira, 2006). As bactérias Gram-negativas, entre elas as cianobactérias, apresentam duas camadas de membrana lipoprotéica e, entre elas, existe uma parede celular de peptidoglicano. A evidência para essa segunda hipótese vem de um pequeno grupo de algas unicelulares flageladas, as glaucófitas, que possuem um cloroplasto com grande semelhança ultraestrutural a uma cianobactéria, incluindo a presença de uma parede de peptidoglicano entre as membranas interna e externa da organela (Löffelhardt e Bohnert, 2002). O cloroplasto secundário, além das duas membranas, possui ainda uma terceira (membrana do eucarioto que foi englobado) e uma quarta (do fagossomo) membranas eucarióticas (Figura 2.6). Em alguns casos, parece ter ocorrido a perda de uma das membranas, como nas euglenófitas e nos dinoflagelados (Tabela 2.1).

Entre os eucariotos fotossintetizantes atuais, reconhecemos uma linhagem filogenética que tem sido chamada de Plantas (Figura 2.4), originada a partir de um evento primário de endossimbiose e composta por três grupos: (1) as algas verdes (Chlorophyta) e plantas terrestres; (2) as algas vermelhas (Rhodophyta); (3) glaucófitas (Glaucocystophyta). Esses grupos, por sua vez, teriam dado origem aos plastos de outras linhagens de eucariotos através de eventos de endossimbiose secundária. Essa hipótese baseia-se em evidências de ultraestrutura, bioquímicas e na comparação de sequências macromoleculares. Por exemplo, as algas verdes são apontadas como o grupo que deu origem aos cloroplastos das euglenófitas e de um pequeno grupo de algas amebóides (Chlorarachniophyta) cujo plasto ainda apresenta um nucleomorfo. Por outro lado, o cloroplasto das criptófitas, um pequeno grupo de algas unicelulares flageladas cujo plasto possui um nucleomorfo, deve ter tido origem a partir de uma alga vermelha (Douglas *et al.* 1991; Douglas e Penny, 1999). As algas vermelhas também deram origem aos cloroplastos das algas com clorofila **c** (heterocontes, haptófitas e dinoflagelados; Oliveira e Bhattacharya, 2000) e, possivelmente, ao plasto não fotossintetizante dos apicomplexa, grupo que inclui *Plasmodium*

Tabela 2.1. Características dos cloroplastos dos eucariontes. cl-a, clorofila a; cl-b, clorofila b; cl-c, clorofila c; fb, ficobilinas. As linhagens com cloroplastos de duas membranas, típicos de um evento primário de endossimbiose são: as algas verdes (Chlorophyta) e plantas terrestres; as algas vermelhas (Rhodophyta); e um pequeno grupo de algas unicelulares flageladas (Glaucocystophyta). As Chlorophyta deram origem, através de endossimbiose secundária, ao cloroplasto complexo (com mais de duas membranas) de um pequeno grupo de algas amebóides (Chlorarachniophyta) cujo plasto, que ainda apresenta um nucleomorfo, e ao plasto das euglenófitas. As Rhodophyta deram origem através de endossimbiose secundária ao cloroplasto das Cryptophyta, algas unicelulares flageladas cujo plasto também possui um nucleomorfo, aos plastos dos demais dos grupos de algas com clorofila c (incluindo heterocontes, haptófitas e dinoflagelados), ao plasto do alveolado Chromerida e ao plasto não fotossintetizante dos Apicomplexa, grupo que inclui Plasmodium (causador da malária) e Toxoplasma (causador da toxoplasmose) (modificado de Delwiche & Palmer, 1997).

Eucariotos com cloroplastos	**nº de membranas do cloroplasto**	**Nucleomorfo**	**Pigmentos fotossintéticos**	Endossimbiose
Chlorophyta e Plantas terrestres	2	Ausente	**cl-a, cl-b**	primária
Rhodophyta	2	Ausente	**cl-a, fb**	primária
Glaucocystophyta	2	Ausente	**cl-a, fb**	primária
Chlorarachniophyta	4	Presente	**cl-a, cl-b**	secundária
Euglenophyta	3	Ausente	**cl-a, cl-b**	secundária
Cryptophyta	4	Presente	**cl-a, cl-c, fb**	secundária
Heterocontophyta	4	Ausente	**cl-a, cl-c**	secundária
Haptophyta	4	Ausente	**cl-a, cl-c**	secundária
Dinophyta	3	Ausente	**cl-a, cl-c**	secundária
Chromerida	4	Ausente	**cl-a**	secundária
Apicomplexa	4	Ausente	**Não fotoss.**	secundária

(causador da malária) e *Toxoplasma* (causador da toxoplasmose). Recentemente, Moore *et al.* (2008) descreveram um novo grupo de alveolados (Chromerida) que possui um plasto fotossintetizante envolto por quatro membranas e é proximamente relacionado aos apicomplexa parasitas, indicando que o ancestral desse grupo era realmente fotossintetizante.

Um segundo evento independente de endossimbiose primária tem sido recentemente postulado para *Paulinella chromatophora,* uma ameba tecada (com carapaça) do grupo dos cercozoários. Esse organismo era conhecido por ter duas células de cianobactérias como endossimbiontes, mas foi demonstrado que ambos não podem ser cultivados independentemente e que os endossimbiontes se dividem sincronicamente com a célula hospedeira. Além disso, parece haver a importação de ao menos algumas proteínas produzidas pela célula hospedeira para o endossimbionte. Esse e outros exemplos de endossimbioses estáveis parecem indicar que novas organelas estão em processo de integração ao hospedeiro e podem ajudar no entendimento de como novas organelas podem ser adquiridas (Bodyl *et al.*, 2007).

A origem endossimbiótica dos plastos e mitocôndrias (incluindo hidrogenossomos) já está firmemente estabelecida. Entretanto, endossimbiose também tem sido proposta para explicar a origem de praticamente todas as demais organelas nas células eucarióticas. Para essas outras organelas—como o sistema relacionado à motilidade (sistema microtubular e flagelo), o retículo endoplasmático, peroxissomos, glicossomos, entre outros—, não existe evidência molecular ou bioquímica conclusiva de origem endossimbiótica. Ao contrário, os dados existentes favorecem a hipótese de origem autógena, em que essas estruturas teriam se originado e se organizado gradativamente na célula eucariótica (Martin, 1999). Uma possível presença de DNA e RNA nos centros organizadores de microtúbulos não foi confirmada. Cavalier-Smith (1975) propôs que o mesmo sistema de fagocitose, em células com membranas flexíveis e sem parede, poderia ter dado origem a vários dos sistemas internos de membranas, como o retículo endoplasmático, a membrana nuclear e o complexo de Golgi.

Os organismos eucarióticos diversificaram-se em várias linhagens filogenéticas principais (Figura 2.4) (chamadas, em inglês, de *crown lineages*): Excavata, um grupo aparentemente basal dentro dos eucariotos, que inclui algumas linhagens amitocondriadas; Opistocontes que incluem todos animais (Metazoa, incluindo invertebrados e vertebrados) e os fungos verdadeiros; Amebozoa (mixomicetos e alguns grupos de amebas); Rhizaria, que inclui os Cercozoa (cloraracniófitas e grupos afins) e foraminíferos; as plantas, incluindo a linhagem das algas verdes e plantas terrestres (com clorofila **a** e **b**), as algas vermelhas e as glaucófitas; os Alveolata (dinoflagelados, ciliados e os apicomplexa); e os Heterocontes ou Estramenopilas (oomicetos, labirintulomicetos e as algas heterocontes—pardas, diatomáceas e outras algas com clorofilas **a** e **c**). Os Alveolados e Heterocontes parecem ter tido um ancestral comum, e são frequentemente referidos como Chromoalveolados (Parfrey *et al.* 2006). Algumas dessas linhagens ainda não estão firmemente estabelecidas e outros grupos ainda tem o seu posicionamento filogenético bastante incerto (por exemplo, as haptófitas). Além disso, novas linhagens de eucariotos, principalmente entre os picoeucariotos (eucariotos com células bastante diminutas, com alguns μm de diâmetro), têm sido descobertas recentemente (Baldauf, 2003).

A rápida diversificação das principais linhagens dos eucariotos deve ter ocorrido em torno de 1 a 1,5 bilhão de anos atrás e pode ter sido ocasionada por diversos fatores, tais como alterações ambientais. O aumento de oxigênio na atmosfera, por exemplo, pode ter atingido patamares que possibilitaram a ocupação de novos nichos, assim como o descongelamento da Terra, segundo a hipótese "Terra como bola de neve" (Hoffman *et al.*, 1998). Essa diversificação também pode ter sido causada por mecanismos internos, como, por exemplo, o surgimento de genes homeóticos, que permitiram padrões de diferenciação celular mais complexos em organismos multicelulares, culminando na chamada "explosão do Cambriano", por volta de 540 milhões de anos atrás, havendo um abundante registro fóssilífero de organismos multicelulares desse período (Sogin, 1994). A transição para uma organização multicelular e a diferenciação de tecidos a partir de uma única célula é um processo altamente complexo, que ocorreu independentemente em diferentes linhagens filogenéticas, bem como outras importantes transições que ocorreram após o surgimento das células eucarióticas, que não serão abordadas neste capítulo.

2.7. Conclusões

Os RNAs catalíticos são considerados fósseis moleculares que remontam à origem da vida, quando o RNA era a molécula principal e atuava no armazenamento de informações e na atividade catalítica. Entretanto, existe ainda uma grande distância entre o que sabemos das condições na Terra nos primórdios da vida e as propriedades atuais do RNA. Apesar dessas limitações, o grande sucesso da hipótese do mundo de RNA está no fato de que, atualmente, ela é a mais abrangente para explicar a origem da vida e, de certa maneira, vários de seus pressupostos são passíveis de experimentação. Um número cada vez maior de experimentos tem dado suporte à hipótese do mundo de RNA, demonstrando as inúmeras capacidades das ribozimas (Lazcano, 1994; Jeffares *et al.*, 1998; Poole *et al.*, 1998; Szathmáry, 1999).

No futuro, um maior conhecimento sobre as interações entre ácidos nucléicos e proteínas e experimentos de simulação de sistemas de RNA *in vitro* poderão esclarecer pontos ainda obscuros. Análises filogenéticas de genes e genomas poderão trazer maiores informações sobre o progenota e quais os prováveis genes e proteínas presentes no início da vida. Ainda falta muito para que possamos ter um cenário mais claro da origem e evolução dos primeiros seres vivos, mas nas duas últimas décadas houve avanços muito significativos.

Embora a existência do mundo de RNA talvez nunca venha a ser comprovada, sua plausibilidade pode ser testada no laboratório investigando as possibilidades de as moléculas de RNA de catalisarem as reações e armazenarem as informações necessárias à vida.

Referências Bibliográficas

Andersson, S.G.E., Zomoeodipour, A., Andersson, I.O. *et al.* (1998). The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. **Nature 396**:133-140.

Baldauf, S.L. (2003). The Deep Roots of Eukaryotes. **Science 300**: 1703-1706.

Baldauf, S.L., Bhattacharya, D., Cockrill, J., Hugenholtz, P., Pawlowski, J. e Simpsin, A.G.B. (2004). The tree of life. **In** Cracraft J., Donohue M.J. (eds) **Assembly the Tree of Life.** Oxford University Press, Oxford, 43-75.

Bekker, A., Holland, H.D., Wang, P.-L., Rumble III, D., Stein, H.J., Hannah, J.L., Coetzee, L.L. e Beukes, N.J. (2004). Dating the rise of atmospheric oxygen. **Nature 427**: 117-120.

Bellorin, A. e Oliveira, M.C. (2006). Plastid origin: A driving force for the evolution of algae. **In** Sharma, A.K. e Sharma, A. (eds) **Plant Genome Biodiversity and Evolution**, Vol 2, Part B, Science Publishers, Enfield, 39-87.

Bodyl, A., Mackiewicz, P. E Stiller, J.W. (2007). The intracellular cyanobacteria of Paulinella chromatophora: endosymbionts or organelles? **Trends in Microbiology 15**: 295-296.

Canfield, D.E. (2005). The early history of atmospheric oxygen: homage to Robert M. Garrels. **Annu. Rev. Earth Plant Sci. 33**: 1-36.

Cavalier-Smith, T. (1975). The origin of nuclei and the eukaryotic cells. **Nature 256**: 463-468.

Cech, T.R. (1988). Conserved sequences and structure of group I introns: building an active site for RNA catalists – A review. **Gene 73**: 259-271.

Cech, T.R. (1990). Self-splicing of group I introns. **Annu. Rev. Biochem. 59**: 543-568.

Cech, T.R. e Bass, B.L. (1986). Biological catalysis of RNA. **Annu. Rev. Biochem. 55**: 599-629.

Clark, C.G. (1999). The effect of secondary loss on our views of eukaryotic evolution. **Biol. Bull. 196**: 385-388.

Crick, F.H.C. (1968). The origin of the genetic code. **J. Mol. Biol. 38**:367–379.

Darnell, J.E. e Doolittle, W.F. (1986). Speculations on the early course of evolution. **Proc. Natl. Acad. Sci. USA 83**: 1271-1275.

Deamer, D.W., Mahon, E.H. e Bosco, G. (1994). Self-assembly and function of primitive membrane structures. **In** Bengtson, S. (ed) Early Life on Earth, **Nobel Symposium, no 84**. Columbia Univ. Press, New York, 107-123.

Delwiche, C.F. e Palmer, J. D. (1997). The origin of plastids and their spread via secondary symbiosis. **Pl. Syst. Evol. [suppl.] 11**: 53-86.

Doolittle, W.F. (1999a). Phylogenetic classification and the universal tree. **Science 284**: 2124-2128.

Doolittle, W.F. (1999b). Rethinking the origins of eukaryotes. **Biol. Bull. 196**: 378-380.

Doudna, J.A. e Szostak, J.W. (1989). RNA-catalyzed synthesis of complementary-strand RNA. **Nature 339**: 519-522.

Doudna, J.A., Cormack, B.P. e Szostak, J.W. (1989). RNA structure, not sequence, determines the 5' splice-site specificity of a group-I intron. **Proc. Nat. Acad. Sci. USA 86**: 7402-7406.

Douglas, S.E., Murphy, C.A., Spencer, D.F. e Gray, M.W. (1991). Cryptomonad algae are evolutionary chimaeras of two phylogenetically distinct unicellular eukaryotes. **Nature 350**:148-151.

Douglas, S.E. e Penny, S.L. (1999). The plastid genome of the cryptophyte alga, *Guillardia theta*: complete sequence and conserved synteny groups confirms its common ancestry with red algae. **J. Mol. Evol. 48**: 236-244.

Dyer, B.D. e Obar, R.A. (1994). **Tracing the history of eukaryotic cells. The enigmatic smile**. Columbia University Press, new York, 259 pp.

Embley, T.M. e Martin, W. (1998). A hydrogen-producing mitochondrion. **Nature 396**: 517-519.

Fire, A.Z.. (2007) Gene silencing by double-stranded RNA. **Cell Death Differ. 14**: 1998-2012

Forterre, P. e Philippe, H. (1999). The last universal common ancestor (LUCA), simple or complex? **Biol. Bull. 196**: 373-377.

Fung, P.A., Gaertig, J., Gorovsky, M.A. e Hallberg, R.L. (1995). Requirement of a small cytoplasmic RNA for the establishment of thermotolerance. **Science 268**: 1036-1039.

Gilbert, W. (1978). Why genes in pieces? **Nature 271**: 501.

Gilbert, W. (1986). The RNA world. **Nature 319**: 618.

Gilbert, W. (1987). The exon theory of genes. **Cold Spring Harbor Symposia on Quantitative Biology 52**: 901-905.

Gogarten, J.P., Murphey, R.D. e Olendzenski, L. (1999). Horizontal gene transfer: pitfalls and promisses. **Biol. Bull. 196**: 359-362.

Gray, M.W. (1992). The endosymbiont hypothesis revisited. **Int. Rev. Cytol. 141**: 233-357.

Han, T.-M. e Runnegar, B. (1992). Megascopic eukaryote algae from the 2.1 billion-year-old Negaunee Iron-Formation, Michigan. **Science 257**: 232-235.

Hoffman, P.., Kaufman, A.J., Halverson, G.P. e Schrag, D.P. (1998). A neoproterozoic snowball Earth. **Science 281**(5381):1342-1346.

Illangasekare, M., Sanchez, G., Nickles, T. e Yarus, M. (1995). Aminoacyl-RNA synthesis catalyzed by an RNA. **Science 267**:643-647.

Jeffares, D.C., Poole, A.M. e Penny, D. (1998). Relics from the RNA world. **J. Mol. Evol. 46**: 18-36.

Joyce, G.F. (1989). RNA evolution and the origins of life. **Nature 338**: 217-224.

Katz, L.A. (1998). Changing perspectives on the origin of eukaryotes. **TREE 13**: 493-497.

Keeling, P.J. (2007). Deep questions in the tree of life. **Science 317:** 1875-1876.

Knoll, A. (1992). The early evolution of eukaryotes: a geological perspective. **Science 256**: 622-627.

Landweber, L.F., Simon, P.J. e Wagner, T.A. (1998). Ribozyme engineering and early evolution. **Bioscience 48**: 2-103.

Lazcano, A. (1994). The RNA world, its predecessors, and its descendants. **In** Bengtson, S. (ed) **Early Life on Earth, Nobel Symposium, no 84**. Columbia Univ. Press, New York, 70-80.

Lazcano, A. e Miller, S.L. (1996). The origin and early evolution of life: prebiotic chemistry, the pre-RNA world, and time. **Cell 85**: 793-798.

Löffelhardt W. e Bohnert, H.J. (2002) The Cyanelle (muroplast) of *Cyanophora paradoxa*: a paradigm for endosynbiotic organelle evolution. **In** Seckbach J. (ed.) **Symbiosis: Mechanisms and Model Systems**. Cellular Origin and Life in Extreme Habitats Series, vol. 4. Kluwer Academic Publishers, Dordrecht, 113-130.

Margulis, L. e Cohen, J.E. (1994). Combinatorial generation of taxonomic diversity: implication of symbiogenesis for the proterozoic fossil record. **In** Bengtson, S. (ed) **Early Life on Earth, Nobel Symposium, no 84**. Columbia Univ. Press, New York, 327-333.

Margulis, L., Dolan, M.F. e Guerrero, R. (2000). The chimeric eukaryote: Origin of the nucleus from the karyomastigont in amitochondriate protists. **PNAS 97:** 6954-6959.

Martin, W. (1999). A briefly argued case that mitochondria and plastids are descendants of endosymbionts, but that the nuclear compartment is not. **Proc. R. Soc. Lond. 266**: 1387-1395.

Martin, W. e Müller, M. (1998). The hydrogen hypothesis for the first eukaryote. **Nature 392**: 37-41.

Maynard Smith, J. e Szathmáry, E. (1999). **The origins of life. From the Birth of Life to the Origin of Language**. Oxford University Press, Oxford, 180 pp.

Michel, F. e Westhof, E. (1990). Modelling of the 3-dimensional architecture of group-I catalytic introns based on comparative sequence-analysis. **J. Mol. Biol. 216**: 585-610.

Moore, R.B., Oborník, M., Janouskovec, J., Chrudimský, T., Vancová, M., Green D.H., Wright, S.W., Davies N.W., Bolch, C.J.S., Heimann, K., Slapeta, J., Hoegh-Guldberg, O., Logsdon, J.M. e Carter, D.A. (2008). A photosynthetic alveolate closely related to apicomplexan parasites. **Nature 451**: 959-963.

Moran, J.V., DeBerardinis, R.J. e Kazazian, H.H. (1999). Exon shuffling by L1 retrotransposition. **Science 283**: 1530-1534.

Nagel, G.M. e Doolittle, R.F. (1995). Phylogenetic analysis of the aminoacyl-transfer-RNA synthetases. **J. Mol. Evol. 40**: 487-498.

Nelson, K.E., Clayton, R.A., Gill, S.R., Gwinn, M.L. *et al.* (1999). Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermotoga maritima*. **Nature 399**: 323-329.

Neunlist, S. e Rohmer, M. (1985). A novel hopanoid, 30-(5'-adenosyl) hopane, from the purple non-sulphur bacterium *Rhodopseudomonas acidophila*, with possible DNA interations. **Bioch. J. 228**: 769-771.

Nitta, I., Kamada, Y., Noda, H., Ueda, T. e Watanabe, K. (1998). Reconstitution of peptide bond formation with *Escherichia coli* 23S RNA domains. **Science 281**: 666-669.

Noller, H.F. (1991). Drugs and the RNA world. **Nature 353**: 302-303.

Ohta, T. (1994). Early evolution of genes and genomes. **In** Bengtson, S. (ed) **Early Life on Earth, Nobel Symposium, no 84**. Columbia Univ. Press, New York, 70-80.

Olendzenski, L., Zhaxybayeva, O. e Gogarten, J.P. (2002). What's in a tree? Does horizontal gene transfer determine microbial taxonomy? **In** Seckbach J. (ed.) **Symbiosis: Mechanisms and Model Systems**. Cellular Origin and Life in Extreme Habitats Series, vol. 4. Kluwer Academic Publishers, Dordrecht, 67-78.

Oliveira, M.C. e Bhattacharya, D. (2000). Phylogeny of the Bangiophycidae (Rhodophyta) and the secondary endosymbiotic origin of algal plastids. **Am J. of Botany 87**: 482-492.

Orgel, L.E. (1968). Evolution of the genetic apparatus. **J. Mol. Biol. 38**:381–393.

Pace, N.R. (2004). The early branches in the tree of life. **In** Cracraft J., Donohue M.J. (eds) **Assembly the Tree of Life.** Oxford University Press, Oxford, 76-85.

Parfrey, L.W., Barbero, E., Lasser, E., Dunthorn, M., Bhattacharya, D., Patterson, D.J. e Katz, L.A. (2006). Eavaluating support for the current classification of eukaryotic diversity. **PLoS genetics 2**: 2062-2073.

Pierson, B.K. (1994). The emergence, diversification, and role of photosynthetic eubacteria. **In** Bengtson, S. (ed) **Early Life on Earth, Nobel Symposium, no 84**. Columbia Univ. Press, New York, 70-80.

Pogliano, J. (2008). The bacterial cytoskeleton. **Current Opnion in Cell Biology 20:**19-27.

Poole, A.M., Jeffares, D.C. e Penny, D. (1998). The path from the RNA world. **J. Mol. Evol. 46**: 1-17.

Potter, S., Durovic, P. e Dennis, P.P. (1995). Ribosomal-RNA precursor processing by a eukaryotic U3 small nucleolar RNA-like molecule in an archaeon. **Science 268**: 1056-1060.

Roy, S.W. e Gilbert, W. (2006). The evolution of spliceosomal introns: patterns, puzzles and progress**. Nat. Rev. Gent. 7**: 211-221.

Runnegar, B. (1994). Proterozoic eukaryotes: evidence from biology and geology. **In** Bengtson, S. (ed.) **Early Life on Earth, Nobel Symposium, no 84**. Columbia Univ. Press, New York, 287-297.

Shi, T. e Falkowski, P.G. (2008). Genome evolution in Cyanobacteria: The stable core and the variable shell. **PNAS 105**: 2510-2515.

Soares, L.M.M. e Valcárcel, J. (2006). The expanding transcriptome: the genome as the 'Book of Sands'. **The EMBO Journal 25**:923-931.

Sogin, M.L. (1994). The origin of eukaryotes and evolution into major kingdoms. **In** Bengtson, S. (ed.) **Early Life on Earth, Nobel Symposium, no 84**. Columbia Univ. Press, New York, 181-192.

Stark, B.C., Kole, R., Bowman, E.J. e Altman, S. (1978). Ribonuclease P: an enzyme with an essential RNA component. Proc. Natl. Acad. Sci. USA, 75(8):3717-3721.

Szathmáry, E. (1999). The origin of the genetic code. **Trends in Genetics 15**: 223-229.

Taylor, F.J.R. (1994). The role of phenotypic comparisons in the determination of protist phylogeny. **In** Bengtson, S. (ed) **Early Life on Earth, Nobel Symposium, no 84**. Columbia Univ. Press, New York, 312-326.

Wetzel, R. (1995). Evolution of the aminoacyl-transfer-RNA synthetases and the origin of the genetic-code. **J. Mol. Evol. 40**: 545-550.

Woese, C. (1967). **The Genetic Code, the molecular basis for genetic expression**. Harper and Row, New York.

Woese, C.R. (1987). Bacterial Evolution. **Microbiol. Rev. 51**: 221-271.

Woese, C.R., Kandler, O. e Wheelis, M.L. (1990). Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria and Eucarya. **Proc. Natl. Acad. Sci. USA 87**: 4576-4579.

Young, L.S., Dunstan, H.M., Witte, P.R., Smith, T.P., Ottonello, S. e Sprague, K.U. (1991). A class III transcription factor composed of RNA. **Science 252**: 542-252.

Capítulo 3

Genoma não codificante — uma breve introdução

Alysson R. Muotri (muotri@ucsd.edu)
Departamento de Pediatria/Medicina Molecular e Celular
Faculdade de Medicina
Universidade da Califórnia – San Diego

Dr. Cassiano Carromeu (ccarromeu@ucsd.edu)
Departamento de Pediatria/Medicina Molecular e Celular
Faculdade de Medicina
Universidade da Califórnia – San Diego

3.1. Introdução

Até pouco tempo atrás, acreditava-se que os papéis das moléculas DNA, RNA e proteína estavam bem estabelecidos. Segundo a definição universalmente aceita, o DNA era o detentor da informação genética, enquanto que as proteínas eram responsáveis pelas atividades catalíticas e estruturais, assim como a regulação de todo o sistema. Ao RNA, cabia o mero papel de mensageiro da informação genética, sendo considerado um remanescente primitivo da evolução molecular que sofremos. Logo, grande esforço cientifico era direcionado a entender como o DNA guardava a informação e era regulado, e à maneira como as proteínas exercem suas atividades catalíticas e regulação necessária para constituir um organismo vivo. Essa visão tem raízes no início dos anos 50, logo depois que James Watson e Francis Crick decifraram a estrutura da dupla hélice do DNA (Watson & Crick, 1953). Em 1958, Francis Crick enunciou o popular "Dogma Central da Biologia Molecular" (Fig. 3.1), segundo o qual o DNA é capaz de se replicar (transmitindo a informação genética para células-filha), assim como traduzir sua informação genética em forma de RNA (Crick, 1958). As proteínas são o produto final dessa informação traduzida. Ao RNA, coube o papel singelo de fazer essa ponte entre o detentor de toda a informação genética (DNA) e o efetor dessa informação (proteína). Esse é um papel um tanto quanto modesto, pelo que se sabe hoje, para uma das moléculas que pode ter dado origem a vida como conhecemos hoje, por meio do "mundo de RNA" (Gilbert, 1986; Müller, 2006) (Capítulo 2).

Com os avanços das técnicas de sequenciamento, diversos projetos surgiram com a proposta de sequenciar o genoma humano. Desses, dois destacam-se: o consórcio de sequenciamento do genoma humano (Lander *et al.*, 2001) e o sequenciamento financiado pela Celera (Venter *et al.*, 2001). O primeiro envolvia o sequenciamento de fragmentos do genoma de vários doadores, enquanto o segundo focou no consenso do sequenciamento de cinco indivíduos. Ambos projetos eram baseados na idéia de que todos os seres humanos compartilham uma alta identidade genética. Foi postulado que a identidade genética entre dois indivíduos quaisquer, sem grau de parentesco próximo, é de 99,9%, enquanto que, com o nosso "primo" próximo chimpanzé, é de 98,8%. Outro dado que emergiu desses projetos foi que apenas 2% do genoma humano codifica proteínas (Lander *et al.*, 2001; Venter *et al.*, 2001; Szymanski *et al.*, 2005). O restante do genoma é constituído de sequências reguladoras (como promotores para transcrição), sequências repetitivas, regiões intronicas e genes de RNA (RNAt, RNAr). Com exceção das regiões reguladoras e dos genes de RNA, o restante do genoma era intitulado de "entulho" ou "DNA-lixo", considerado "resquício" da evolução. Os projetos de sequenciamento do genoma reforçaram essa visão.

Neste capítulo, vamos revisar o que se conhece do genoma, particularmente suas regiões não-codificantes. Mostraremos que esse "entulho" é, na verdade, uma mina de ouro para a vida como a conhecemos. Propõem-se situá-lo no cenário cientifico atual, mostrando como o RNA pode ser uma molécula-chave quando se fala em diversidade genética. Revisaremos o conceito de gene, assim como as diferenças genéticas entre os indivíduos.

3.2. Exorcisando um Dogma

A palavra dogma vem do grego e significa, filosoficamente, "ponto fundamental e indiscutível de uma crença religiosa". Por sua definição, podemos excluir essa palavra de qualquer explicação cientifica. Ciência faz-se com discussão e o critério de demarcação de ciência de Karl Popper denota isso claramente (Popper, 1959). Uma teoria cientifica, na visão popperiana, tem duas características principais: 1. ser passível de ser falseada ou refutada; 2. ser passível de ser testada (por meio de experimentação). A ciência trabalha testando hipóteses e muitas vezes essas

Figura 3.1. Visão esquemática do Dogma Central da Biologia Molecular. O DNA é capaz de se autocopiar num fenômeno denominado replicação, assim como copiar seu conteúdo na forma de RNA mensageiro (RNAm). Este fenômeno denomina-se transcrição. O RNAm é traduzido dando origem às proteínas.

podem levar a derrubar uma teoria ou restringir sua capacidade explicativa. Nesse sentido, pode-se dizer que não existem teorias comprovadas, mas apenas teorias que ainda não foram derrubadas. Quanto mais uma teoria sobrevive a hipóteses que possam derrubá-la, maior sua robustez ou veracidade. Alguns anos após propor o Dogma Central da Biologia Molecular, Francis Crick o modificou para adequá-lo as novas descobertas, de que o fluxo da informação poderia ser feito no sentido RNA → DNA (Crick, 1970). O Dogma popularizou-se por sua simplicidade e beleza extrínseca. Fácil de ensinar, fácil de aprender, essa simplicidade explicaria toda a vida como a conhecemos.

O problema do Dogma Central da Biologia Molecular são suas consequências. Ele tem uma visão centrada na proteína e no DNA, o que se refletiu nas definições que surgiram a partir dele, como a definição de gene. Uma definição geral, encontrada na maioria dos livros de Biologia Molecular, é de que "gene é a sequência inteira de ácido nucléico necessária para a síntese de um polipeptídeo funcional" (Lodish *et al.*, 1999). Ou seja, gene = polipeptídio = função.

Outro ponto importante para a popularização do Dogma foi o advento das técnicas de biologia molecular. Quanto mais simples o organismo, mais fácil sua manipulação e grande parte de nosso conhecimento nesse campo se deve a estudos feitos na bactéria *Eschirichia coli*. Muitos desses estudos podiam ser aplicados para organismos mais complexos, como cavalos e seres humanos. Logo se popularizou o conceito de que "o que é verdade para uma bactéria também é verdade para um elefante", generalizando os achados em *E.coli* para outros organismos vivos.

3.2.1. Ouro no Entulho: do Lixo ao Luxo

Com o surgimento dos projetos de sequenciamento e, consequentemente, a disponibilização de genomas inteiros de organismos, constatou-se que o tamanho das regiões codificadoras entre diferentes organismos se mantinha similar (Fig. 3.2). Por exemplo, o verme *Caenorhabidtis elegans*, que possui aproximadamente 1000 células, possui um tamanho muito similar de regiões codificadoras da espécie humana, que possui aproximadamente 100 trilhões de células (Mattick, 2007). Na verdade, não se vê um crescimento significativo em tamanho dessas regiões, se comparado com a explosão de diversidade que a natureza apresenta. Isso foi uma surpresa, uma vez que, dado que as proteínas são responsáveis pela maior parte da regulação gênica, era esperado um crescimento exponencial de regiões que as codificam. No entanto, se olharmos as regiões não codificadoras do genoma, podemos ver um crescimento exponencial nessas regiões, acompanhando uma maior complexidade do organismo (Fig. 3.2) (Szymanski *et al.*, 2005; Mattick, 2007). Evidentemente definir complexidade de um organismo não é algo trivial. Nesse caso, consideramos a complexidade organizacional aparente do organismo como um todo. Por exemplo, um homem ou um cachorro são mais complexos que uma bactéria ou fungo, dada, por exemplo, a quantidade de células, sua organização e as habilidades do organismo. É necessário notar que isso não é o mesmo que considerar grupos mais ou menos evoluídos.

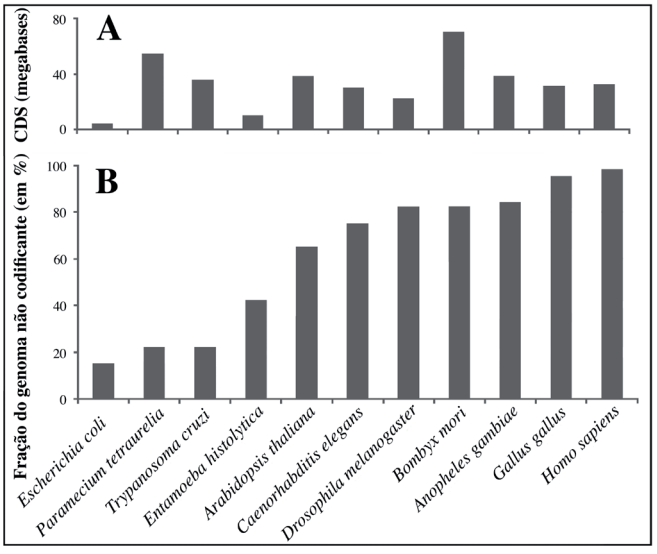


Figura 3.2. Analise das regiões codificadoras versus regiões não codificantes do genoma de diferentes organismos. A, regiões codificantes (em Megabases); B, regiões não codificantes (em porcentagem do genoma). Modificado do trabalho de Mattick (2007).

Em 2007, o projeto ENCODE publicou seus primeiros resultados (ENCODE Project Consortium, 2007). Tratava-se de um grande consórcio, envolvendo várias instituições ao redor do mundo, que tinha como objetivo uma análise rigorosa de determinadas porções do genoma humano. Ele analisou o correspondente a cerca de 1% do genoma humano e, dentre outros resultados, constatou-se que uma grande quantidade de DNA dessas regiões é transcrito em RNA (em torno de 90%). Ainda há muita controvérsia, no entanto, sobre quanto realmente é transcrito. De qualquer forma, o número de RNAs transcrito é, com certeza, muito maior que o número de regiões codificadoras preditas (que correspondem a aproximadamente 2% do genoma humano). A esses RNAs transcritos que não codificam proteínas foi dada a denominação de RNAs não-codificantes (ncRNA). Alguns desses RNAs possuíam função já estabelecida (como o RNAr e o RNAt), mas muitos permanecem sem função descrita. Mesmo não codificando para proteínas ou RNAs da maquinaria de transcrição, algumas evidências apontam que esses RNAs não são dispensáveis na célula:

- Alguns ncRNAs são expressos diferentemente, alguns sob o controle de fatores de transcrição;
- Pelo menos alguns ncRNAs tem uma localização subcelular específica;
- Alguns se mostram funcionais.

Note que o uso da palavra "alguns" deve somente ao fato de as pesquisas estarem engatinhando no campo. É provável que a maioria desses RNAs tenha funções especificas na célula. Dentre as funções que começam a ser atribuídas a essa fração de RNAs, temos:

- controle da dinâmica do cromossomo;
- edição de RNAs;
- inibição da tradução;
- degradação de RNAm.

Dependendo da função a ser exercida, os RNAs apresentam algumas vantagens sobre as proteínas. Um ponto-chave é a especificidade do RNA por moléculas de RNA e DNA em relação às proteínas. Por pareamento de Watson-Crick de bases, moléculas pequenas de RNA conseguem uma especificidade por regiões complementares (sejam elas RNA ou DNA) muito maior que proteínas complexas. Outro ponto forte a favor dos RNA versus proteínas é o fato de a informação ser codificada economicamente. Ou seja, ao invés de passar por um processo de transcrição e tradução, passa-se por um processo de transcrição somente, sendo mais rápido e energeticamente mais eficiente. Um exemplo pode ser encontrado no sistema nervoso. A abundância de ncRNA no cérebro tem sido vista como um mecanismo de resposta rápida à

interação com o ambiente. Por último, essa camada do fluxo de informação pode ser alterada, gerando alterações fenotípicas sem causar uma "quebra do sistema". Ou seja, diferentemente de uma proteína, em que uma mutação pode levar à completa inativação, mutações nos ncRNA são mais flexíveis, uma vez que podem levar a uma menor afinidade pela molécula-alvo, mas dificilmente à não interação.

3.2.2. O Cozinheiro Celular

Uma analogia que ajuda a explicar a importância dos ncRNA é a comparação da célula com uma cozinha. Para executar uma receita, precisa-se de um livro de receitas e dos ingredientes. Muitos ingredientes são comuns a diversas receitas. Com água, leite, açúcar, fermento, óleo e ovos, é possível fazer diversos alimentos (pão, bolo, massa para pizza e macarrão). As mesmas regras aplicam-se para a célula. Neste caso, o livro de receitas é o DNA, contendo todo o protocolo para se fazer um organismo. As proteínas entram como os ingredientes necessários para preparar esse organismo. E o ncRNA? Bem, esse é o cozinheiro, um elemento fundamental para a receita funcionar. É esse cozinheiro que vai ler o livro de receitas e seguir suas ordens, adicionando todos os ingredientes na ordem correta (não se faz um bolo misturando-se os ingredientes aleatoriamente). Tem-se de prepará-los, pesar, misturar e assar. Tudo numa sequência determinada. O cozinheiro faz toda essa "regulação" necessária para que a receita de bolo dê origem a um bolo. Outra coisa interessante dessa analogia: pegue uma receita qualquer de bolo e conte quantas vezes as palavras que significam ingredientes (farinha, leite, ovos, etc) aparecem. Agora olhe para a receita como um todo e verá que se têm muito mais palavras (misture 100 g, acenda o forno, prepare a forma) que não são ingredientes. É tudo informação que deve ser seguida na ordem e como aparece para finalizar seu bolo. Da mesma forma ocorre na célula. Essa informação está presente lá na forma de DNA e é o cozinheiro celular (ncRNA) que organiza e coordena tudo para que a "receita" funcione.

Genes eucariotos são formados por segmentos de éxons e íntrons, sendo os éxons os detentores das regiões codificadoras e os introns removidos do RNAm maduro por um processo conhecido como *splicing* (para uma revisão, ver Licatalosi, 2010). Por muito tempo, acreditou-se que esses introns, após remoção, eram degradados. Atualmente, está se atribuindo função a muitos deles e um exemplo do que eles poderiam estar fazendo está mostrado na Figura 3.3. Seguindo a mesma analogia empregada anteriormente, um gene codificando para um ingrediente (como farinha) é expresso e seus introns removidos. No entanto, ao invés de serem degradados, esses introns são processados e geram ncRNA com informação importante para o processo que se inicia (por exemplo, regulando vias em que esse produto gênico irá se inserir). Ou seja, para cada produto gênico, a célula geraria uma quantidade de informação adicional. Essa informação pode ter funções diversas, atuando em diversas vias celulares, inibindo-as ou ativando-as, dependendo da natureza da informação.

Assim como numa receita, não podemos usar um ingrediente estragado (por exemplo, uma proteína mutada de tal forma a perder sua função normal). No entanto, podemos variar um pouco os elementos regulatórios (como colocar 105 g de farinha, ao invés de 100 g ou assar um pouco mais), sem grandes prejuízos para a receita. Dado isso, nosso genoma pode acomodar mais mutações em regiões não-codificantes que em regiões que codificam proteínas. O resultado final pode explicar um pouco a variabilidade que encontramos entre indivíduos de uma mesma espécie, dando a gama de fenótipos observados existentes. É como se diferentes cozinheiros fizessem a mesma receita. Todos usam os mesmos ingredientes básicos, mas cada qual dá um toque especial, de tal forma que, no final, o gosto e a aparência são ligeiramente diferentes.

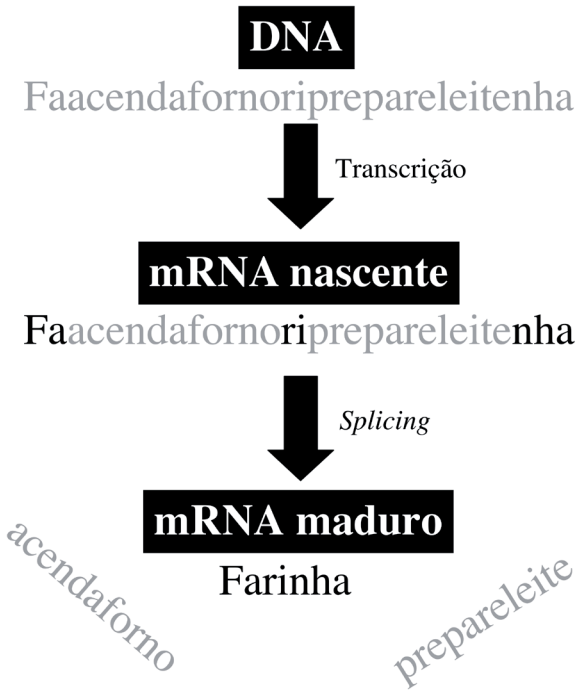


Figura 3.3. Esquematização do processo de transcrição de uma região codificante num ser eucarioto. A região transcrita do DNA aparece em cinza. O RNAm nascente contem tanto exons (em preto) quanto introns (em cinza). Após edição desse RNA pelo processo de *splicing*, tem-se o RNAm maduro (em preto) que é exportado para o citoplasma e traduzido em proteína. Os introns, removidos do RNAm maduro, estão mostrados em cinza.

3.3. Uma Bactéria Não é um Elefante

A principal função dos ncRNA seria a de regular e direcionar os complexos caminhos do desenvolvimento celular e do organismo. Isso requer uma enorme quantidade de informação num organismo tão precisamente esculpido como o ser humano. Dentre as implicações dos ncRNA para a biologia molecular, temos que "nem tudo que é verdade para uma bactéria é verdade para um elefante". Muitas similaridades podem ser vistas entre diferentes organismos e é indiscutível a importância das bactérias para o melhor entendimento de eucariotos complexos. No entanto, fica claro que seres mais complexos apresentam um genoma com peculiaridades que devem ser levadas em conta. O fato de transcrevermos grande parte de nosso genoma certamente representa algo importante, que começa a ser levado em conta hoje em dia.

Considerando o que vem sendo descoberto, um novo conceito de gene tem sido proposto (Fig. 3.4). Ainda não é consenso, mas um dos novos conceitos define um gene como a união de sequências genômicas (derivados de DNA ou RNA) codificando um conjunto de produtos funcionais (que podem ser de origem senso, antisenso ou senso+antisenso) (Gerstein *et al.*, 2007). Ou seja, muda-se o conceito "gene = proteína" para "gene = função", com ênfase no fato de um gene poder codificar diferentes funções.

Similar ao conceito de gene, também o conceito de como se dá o fluxo da informação gênica dentro da célula vem mudando. Novos conceitos vêm sendo propostos, um deles mostrado na Figura 3.5. Note que, no topo da hierarquia, se tem a cromatina.

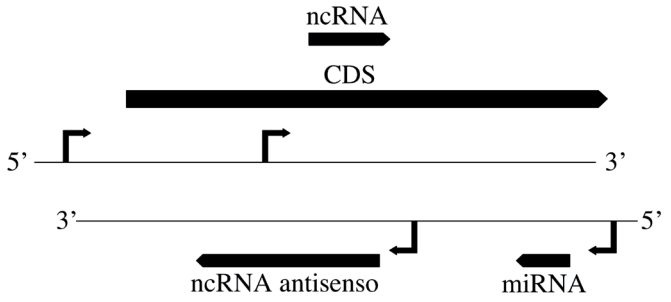


Figura 3.4. Esquematização de um gene eucarioto. A região codificante para proteína (CDS) deixa de ser vista como a única formadora de um gene e este passa a ser visto como um conjunto de produtos funcionais. Ambas as fitas podem ser transcritas e podem dar origem a uma gama diversa de RNAs (RNAm, ncRNA, miRNA, entre outras). Uma molécula de RNA transcrita, através do processamento da mesma, pode dar origem a diversos RNAs diferentes (como, por exemplo, RNAm e miRNA).

Esta se refere ao DNA complexado com proteínas, forma como ele é encontrado dentro das células. Não se pode tentar entender transcrição de DNA pensando nele como uma molécula solta no núcleo celular. Proteínas (como histonas) interagem com o DNA de maneira que este não fique isolado no núcleo. Essa interação influencia se um gene vai ser expresso ou não. Alterações epigenéticas (como modificação de histonas e metilação de DNA) são decisivas nesse sentido e estas podem ser herdadas ou passadas para células-filha sem que ocorra uma mudança na sequência de bases do DNA.

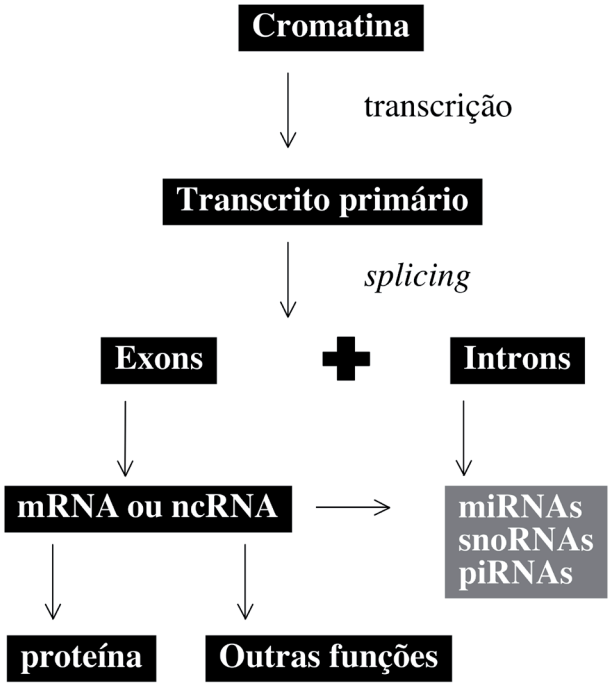


Figura 3.5. Fluxo da informação genética dentro de uma célula eucariota. O DNA contido na cromatina é transcrito para RNA e, por meio de *splicing*, dá origem ao RNAm maduro e ncRNAs. Estes podem ser oriundos de exons ou introns. Após sofrer processamento, pode-se obter ncRNAs diversos (como miRNAs, snoRNAs e piRNAs—representados no quadrado cinza). Enquanto as proteínas seguem para desempenhar papéis estruturais, catalíticos ou transdução de sinal e regulação gênica, os ncRNAs gerados atuam sobre todos os componentes da via, assim como regulando todos os passos (transcrição, *splicing*, processamento de RNA). Adaptado do trabalho de Mattick (2007).

3.4. Iguais ou Diferentes?

Como mencionado na introdução, dois projetos de sequenciamento do genoma humano se destacaram. Números como 99,9% de identidade entre todos os seres humanos do planeta ou 98,8% de identidade com o genoma do chimpanzé logo se espalharam e ficaram conhecidos. Justificavam que somos todos iguais entre nós e nossos primos próximos. No entanto, em outubro de 2007, um trabalho pioneiro de um pesquisador controverso (J. Craig Venter) trouxe um pouco mais de clareza a essa questão (Levy *et al.*, 2007). Como explicado, ambos os projetos genoma citados consistiam do consenso de alguns indivíduos para se montar o genoma completo. Os números de identidade eram baseados nesses dados. O manuscrito de 2007 continha o genoma completo de um único individuo (como não poderia deixar de ser, do próprio Dr. Venter). Mais que suas predisposições genéticas (como doença de Alzheimer, preferência pelo entardecer, comportamento anti-social e cera de ouvido do tipo úmido), foi feita uma distinção entre os cromossomos herdados da mãe e os do pai. Com isso, foi possível ver a identidade genética entre cromossomos homólogos de um mesmo individuo. Para surpresa em relação ao que vinha sendo divulgado, a identidade entre o par de cromossomos homólogos de um único individuo foi igual a 99,5%. Ou seja, a diferença genética de 0,5% para um único indivíduo era cinco vezes maior que a diferença esperada para todos os seres humanos. Diversos projetos estão sendo feitos com o sequenciamento individual de genomas e em breve poderemos saber mais sobre o quanto somos iguais ou diferentes.

Vale ressaltar que essas diferenças residem grandemente na variação estrutural que genomas de diferentes indivíduos apresentam. Variações estruturais, diferentemente da troca única de uma base nucleotídea por outra, consiste de rearranjos de 2 ou mais nucleotídeos (usualmente muitos). Dentre esses rearranjos, temos: deleção, inserção, inversão, variação no número de cópias e duplicação segmental (Fig. 3.6). Recentemente, constatou-se uma diferença da ordem de aproximadamente 1.300 variações estruturais entre dois indivíduos diferentes (Korbel *et al.*, 2007). Apesar de ocorrerem numa frequência muito menor que o polimorfismo num único nucleotídeo, uma vez que a variação estrutural envolve grandes segmentos de DNA, o número de nucleotídeos afetados é muito maior. Essa variação pode contribuir decisivamente para a diversidade de fenótipos observados, mudando o padrão de transcrição dos genes envolvidos ou próximos a eles (aumentando ou diminuindo sua expressão). Evolutivamente falando, essa variação é benéfica. É graças a essas variações que evoluímos e podemos dizer que cada ser humano é único. Um exemplo de como essa variação é importante para nós está detalhada abaixo.

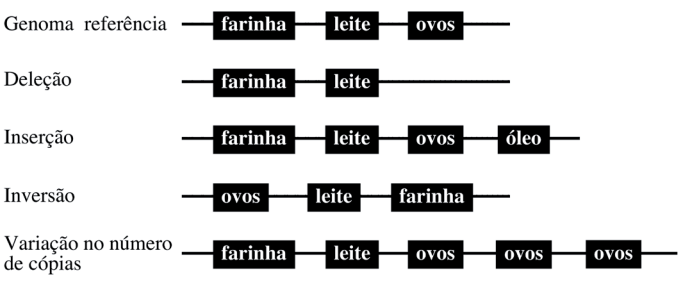


Figura 3.6. Variações estruturais comumente encontradas no genoma humano. Seguindo ainda a analogia da cozinha celular (com proteínas representadas como ingredientes culinários), podemos ter uma idéia das consequências para a receita quando alguma dessas variações acontece. Note que, mesmo não mostrado, esta variação estrutural também se aplica a regiões regulatórias (os cozinheiros).

3.5. Variar ou Não, Eis a Questão - o Cérebro e os Retrotransposons

Acredita-se que a complexidade do cérebro humano, com milhares de tipos de neurônios diferentes, tenha permitido o surgimento de sofisticados repertórios comportamentais, como a linguagem, uso de ferramentas, percepção do "eu", pensamento simbólico, aprendizado cultural e consciência. Dessa complexidade, emergiram obras de extraordinário conteúdo tecnológico e artístico numa história cultural relativamente curta de nossa espécie. Isso parece indicar que a complexidade cerebral tem um propósito criativo, ao contrário de outros sistemas amplamente mais complexos, porém brutos, como as galáxias e as milhares de estrelas que as compõem. Entender como a complexidade neuronal é moldada durante o desenvolvimento é mergulhar em questões fundamentais da origem da nossa espécie.

A formação do cérebro humano não é um processo otimizado. Pelo contrário, a maioria das células geradas será descartada e apenas uma pequena fração será usada. O mecanismo por trás dessa seleção é obscuro e existem evidências sugerindo que fatores extrínsecos e intrínsecos contribuam para a sobrevivência ou morte celular. Apenas as células precursoras com as propriedades corretas, no momento e local ideais, irão florescer e amadurecer em neurônios funcionais, contribuindo para a formação das redes nervosas. Nessa competição, forças de variação e seleção atuam para esculpir cada cérebro humano, cada rede neural, neurônio por neurônio, gerando a verdadeira individualidade na forma como cada um de nós recebe, processa e interage com o mundo exterior. Vale lembrar que a seleção natural precisa de variação para gerar os diferentes tipos neuronais no cérebro. Inicialmente, cogitou-se que a variação estaria contida nos "genes" codificantes para proteínas. No entanto, como discutido anteriormente, com menos de 2% de genes codificantes para proteínas no genoma, fica difícil gerar informação suficiente para os milhares de tipos celulares contidos no cérebro humano. Mesmo considerando eventos moleculares como o processamento alternativo do RNA ou modificações pós-traducionais, não existe variação suficiente. A variação deve residir em outro lugar.

A falta de uma função óbvia para os outros 98% do genoma inspirou o conceito de "DNA-lixo", ilustrando a idéia de que essas sequências seriam resquícios evolutivos, acumulados ao longo de milhares de anos no genoma. Como uma garagem cheia de tranqueira, o genoma parece lidar muito bem com o excesso de sequências, mas parece difícil compreender porque não se livra dele, economizando energia celular. Parte desse DNA-lixo é composto de elementos transponíveis, ou genes-saltadores, capazes de produzir cópias de si próprios, inserindo novas cópias no genoma e, eventualmente, alterando a expressão de genes próximos. A atividade desses elementos foi flagrada durante a evolução e esses parasitas genômicos ficaram conhecidos como genes-egoístas, com a única finalidade de se manterem vivos para as próximas gerações através da replicação em células germinativas dos indivíduos. A replicação em células não germinativas, somáticas, que não formarão um novo individuo não seria uma estratégia de sobrevivência, pelo menos até agora...

Estudando como os genes eram regulados durante a especialização neuronal a partir de células-tronco, descobriu-se que havia uma ativação dos elementos transponíveis tão logo a célula optasse pela diferenciação neuronal. Ao induzir a células-tronco a se diferenciarem em outros tipos celulares, nada era detectado, indicando que o fenômeno era específico dos neurônios (Muotri *et al.*, 2005). O achado confrontava tudo que sabíamos sobre o comportamento desses elementos e sua "vontade" de passar para as futuras gerações. Afinal, o que estariam fazendo ao proliferar no cérebro? Ao contrário do atraente conceito de que todas as células do corpo possuem o mesmo genoma e que as diferenças seriam meras consequências da regulação gênica, agora existem evidências fortes o suficiente para demonstrar que esse não é o caso do cérebro. Cada neurônio parecia ser único, cada um apresentava novas inserções no genoma, impactando genes vizinhos. Essa atividade amplificaria o efeito da regulação gênica, gerando uma enorme variação celular e aumentando o repertório de tipos celulares capazes de serem formados por um dado grupo de genes. Esse mecanismo de variação e flexibilidade parece contribuir para a originalidade de cada cérebro, explicando porque mesmo gêmeos geneticamente idênticos apresentam personalidades tão caracteristicamente distintas.

Filosoficamente, os dados estariam apontando para uma parcela de "acaso" na formação de cada personalidade. Especula-se que o aumento da variabilidade neuronal seria capaz de produzir indivíduos fora da curva normal, com qualidades diferentes. Organismos fora da curva teriam mais chances de se adaptarem a novos ambientes ou de reagir contra mudanças drásticas no ambiente. Além disso, existiriam eventuais indivíduos prodígios na população, com uma capacidade cognitiva superior. E talvez sejam indivíduos assim que aumentem a capacidade criativa da espécie humana, favorecendo a dominação de novos territórios, por exemplo. Nesse sentido, os elementos transponíveis continuariam sendo genes-egoístas, pois, ao manipular a mente humana, acabam por aumentar as chances reprodutivas da espécie. Curiosamente, durante a evolução dos primatas, observa-se uma impressionante correlação entre a adaptação humana e o surgimento de novas sequências transponíveis. Evidências de alterações climáticas globais sugerem que ambientes mais frios, secos e com maiores variações devem ter ocorrido cerca de três milhões de anos atrás. As alterações bruscas acabaram por diminuir o suprimento de comida e água, pressionando fortemente a adaptação de nossos ancestrais a novos ambientes. Interessantemente, novas famílias de elementos transponíveis no genoma surgem na mesma época em que os humanos adquirem o bipedalismo, apresentam um aumento da massa cerebral e apresentam as primeiras evidências de uso de ferramentas, consciência ou motivação artística.

Por outro lado, o fenômeno de inserções somáticas no cérebro pode não passar de um resquício evolutivo. Tanto o cérebro como o sistema reprodutivo passaram por grandes modificações durante a evolução. A expressão genética desses dois órgãos é relativamente parecida e os dois possuem diversas vias de sinalização em comum. Nesse contexto não parece novidade encontrar fenômenos moleculares presentes somente nesses órgãos. Se esse for realmente o caso, a atividade dos elementos transponíveis no sistema nervoso é descartável e não possui contribuição alguma para as redes neuronais, cognição ou comportamento. É plausível, mas fica faltando responder porque o genoma ficaria carregando uma carga que não tivesse utilidade. Qualquer que seja a função do mosaicismo genético dos neurônios, é preciso cautela no desenho dos experimentos que permitirão investigar o fenômeno. Atualmente é impossível usar técnicas clássicas de nocaute genético para eliminar os genes saltadores do genoma. São vários deles que estão ativos no genoma. Além disso, estão espalhados pelos cromossomos. Vai ser preciso bastante criatividade para buscar situações experimentais onde a hipótese possa ser testada. Qualquer que seja o resultado encontrado, só vai ser real se fizer sentido sob uma ótica evolutiva.

3.6. Considerações Finais

Como se pode perceber, muito do que conhecemos sobre o genoma humano está mudando. A visão centrada no DNA como

molécula volta-se agora para a cromatina como a detentora da informação. Essa informação é transcrita em proteína e também em ncRNA. A importância do RNA vem sendo colocada em seu devido lugar e já não é mais possível se falar no organismo sem se levar em consideração a intricada relação do trio: cromatina – RNA – proteínas. O "mundo de RNA" começa a despontar novamente, mas agora mostrando o oceano rico dessas moléculas que temos dentro de nossas células. Também o conceito de gene sofre evolução e, mais do que um produto, ele pode ser descrito como uma função (ou várias). Soma-se a esse cenário o fato de que o genoma não é estático e apresentar uma gama de variações estruturais entre indivíduos, assim como entre suas células, como no caso de neurônios. É devido a essas variações que estamos aqui hoje. Graças a essas diferenças, temos a variabilidade necessária para continuarmos evoluindo. O fato de que cada indivíduo apresenta um genoma peculiar reforça o quanto ele é único e precioso. Você é tão diferente de seu vizinho quanto de um japonês, alemão ou angolano. Os poucos genes que definem nossos fenótipos externos não podem ser usados para uma classificação humana em raças. Qualquer visão nesse sentido não se justifica perante as novas observações da ciência. Somam-se a isso as alterações epigenéticas e a influência do ambiente sobre cada um de nós, resultando no que definimos como humanidade.

Dado estarmos no início de novas descobertas, é difícil predizer o que encontraremos. Mas fica claro a importância do que antes era considerado lixo ou entulho. Thomas Carlyle afirmou, "Se não existe função no genoma não-codificante, nosso genoma é um grande desperdício de espaço.", que também é autor de outra famosa frase falando sobre vida extra-terrestre no universo: "Se não existe vida fora da terra, então universo é um grande desperdício de espaço". Carl Sagan fez referência a ela anos mais tarde.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C. *et al.* (2005). The transcriptional landscape of the mammalian genome. **Science 309**(5740):1559-63. Erratum in: (2006). **Science 311**(5768):1713.

Crick F.H. (1958). On protein synthesis. **Symp. Soc. Exp. Biol.** 12:138-163.

Crick F.H. (1970).Central dogma of molecular biology. **Nature 227**(5258):561-563.

ENCODE Project Consortium: Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigó, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., Thurman, R.E., *et al.* (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. **Nature 447**(7146):799-816.

Gerstein, M.B., Bruce, C., Rozowsky, J.S., Zheng, D., Du, J., Korbel, J.O., Emanuelsson, O., Zhang, Z.D., Weissman, S., Snyder, M. (2007). What is a gene, post-ENCODE? History and updated definition. **Genome Res. 17**(6):669-681.

Gilbert, W. (1986) The RNA world. **Nature 319**:618.

Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L *et al.* (2007). Paired-end mapping reveals extensive structural variation in the human genome. **Science 318**(5849):420-426.

Lander, E.S., Linton L.M., Birren, B., Nusbaum, C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W. *et al.* (International Human Genome Sequencing Consortium) (2001). Initial sequencing and analysis of the human genome, 2001). **Nature 409**(6822):860-921. Erratum in: 2001. **Nature 412**(6846):565; 2001. **Nature 411**(6838):720. Szustakowki, J [corrected to Szustakowski, J].

Licatalosi, D.D., Darnell, R.B. (2010). RNA processing and its regulation: global insights into biological networks. **Nat. Rev. Genet. 11**(1):75-87.

Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., Denisov, G. *et al.* (2007). The diploid genome sequence of an individual human. **PLoS Biol. 5**(10):e254.

Lodish, H., Berk, A., Zipursky, S.L., Matsudaira, P., Baltimore, D., Darnell, J.E. (1999). **Molecular cell Biology**. 4th ed., W H Freeman & Co, New York.

Mattick, J.S. (2007). A new paradigm for developmental biology. **J. Exp. Biol. 210**(9):1526-1547.

Müller U.F. (2006). Re-creating an RNA world. **Cell Mol. Life Sci. 63**(11):1278-1293.

Muotri A.R., Chu, V.T., Marchetto, M.C., Deng, W., Moran, J.V., Gage, F.H. (2005). Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. **Nature 435**(7044):903-910.

Popper, K.R. (1959). **The logic of scientific discovery**. Basic Books, New York.

Szymanski, M., Barciszewska, M.Z., Erdmann, V.A., Barciszewski, J. (2005). A new frontier for molecular medicine: noncoding RNAs. **Biochim. Biophys. Acta. 1756**(1):65-75.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A. *et al.* (2001). The sequence of the human genome. **Science 291**(5507):1304-1351.

Watson J.D. e Crick F.H. (1953). The structure of DNA. **Cold Spring Harb Symp Quant Biol. 18**:123-131.

Capítulo 4

O papel da interferência de RNA na célula eucariótica

Stephano Spanó Mello (stespano@gmail.com)
Depto. de Microbiologia
Instituto de Ciências Biomédicas
Universidade de São Paulo

Luciana Nogueira de Sousa Andrade (lnsa@usp.br)
Depto. de Microbiologia
Instituto de Ciências Biomédicas
Universidade de São Paulo

Carlos Frederico Martins Menck (cfmmenck@usp.br)
Depto. de Microbiologia
Instituto de Ciências Biomédicas
Universidade de São Paulo

"*...considering what we now know about RNA and RNA interference, it is perhaps a good time to reconsider the idea that genetic information is stored primarily in the nucleotide sequence of our DNA.*
.... if you get inspired and excited, please join the adventure and help explore the many unknowns that are still waiting to be addressed." (Craig C. Mello, 2006)

4.1. Redescobrindo as Funções do RNA nas Células

O dogma central da biologia molecular, envolvendo replicação e transcrição de ácidos nucléicos e a síntese de proteínas, foi estabelecido através de estudos de organismos simples, como a bactéria *Escherichia coli*. Nesses estudos, estabeleceu-se que o fluxo de informações hereditárias se dá do DNA para a proteína, sendo o RNA um intermediário entre a sequência do DNA e a proteína que será codificada. Além disso, pressupunha-se que a proteína era a única molécula capaz de realizar as diversas funções estruturais, catalíticas e regulatórias dentro de uma célula. Estudos em organismos eucariotos, no entanto, revelaram muitas surpresas. Por exemplo, verificou-se que os genes eucariotos são em geral descontínuos (com a presença de introns) e que o genoma eucarioto possui uma grande quantidade de sequências de DNA que não codificam proteínas (regiões intergênicas ou não codificantes). A essas regiões do genoma (que no caso humano se calcula em 90% de toda a sequência do DNA) às vezes se atribuiu o nome de "DNA lixo" (também chamado "DNA tralha", do inglês "junk DNA"). Também se atribuiu a essas regiões o status de "restos" do processo evolutivo, gerados a partir de genes antigos e pela inserção de elementos móveis ou parasitas gênicos (genes egoístas, do inglês "selfish DNA").

Recentemente se observou, no entanto, que muitas dessas sequências supostamente inertes são transcritas (Frith *et al*., 2005; The *et al*., 2005) e que o RNA por si só poderia desempenhar funções biológicas de grande importância (Eddy, 2001; John, 2003; Mattick, 2005; Mattick e Makunin, 2005). Desde sua descoberta, vários tipos de RNAs não-codificantes (ncRNA) foram descritos, mas a compreensão da maior parte de suas funções ainda é muito limitada (veja o Capítulo 3). Entre os tipos de ncRNAs, temos alguns conhecidos há bastante tempo, como o RNA transportador (tRNA) e o RNA ribossômico (rRNA). Esses ncRNAs possuem funções genéricas dentro da célula e estão envolvidos no processo de síntese de proteínas. Mais recentemente, muitos outros ncRNAs foram descobertos e novos tipos continuam a aparecer, ampliando o nosso conhecimento sobre o papel desses RNAs nas células (Eddy, 2001). Até o momento, os ncRNAs são divididos de acordo com seu tamanho, estrutura secundária e função (Tabela 4.1). Entretanto, ainda existe uma inconsistência na nomenclatura utilizada na literatura, que possivelmente será revisada em breve.

As propriedades catalíticas do RNA foram descobertas nos anos 1980, quando ficou evidente que os RNAs desempenhavam funções adicionais além das já descritas, incluindo a

Tabela 4.1. Os diferentes tipos de RNA não codificantes.

ncRNAs pequenos (<300 pares de base)

- **rRNA**
  RNA ribossômico
- **tRNA**
  RNA de transferência
- **miRNA**
  MicroRNA – reguladores da tradução protéica.
- **siRNA**
  Small interfering RNA – moléculas exógenas, ativadoras do mecanismo de interferência.
- **shRNA**
  Short hairpin RNA – moléculas exógenas em forma de grampo, ativadoras do mecanismo de interferência.
- **snRNA**
  Small nuclear RNA – pequenas moléculas de RNA achadas nos núcleos de células eucarióticas.
- **snoRNA**
  Small nucleolar RNA – pequenas moléculas de RNA capazes de guiar modificações químicas (ex. metilação) em outras moléculas de RNA (ex. rRNA).
- **piRNA**
  Piwi-interacting RNA – ativos nos ovários e testículos dos animais, e envolvidos no silenciamento através da interação com a proteína Piwi. São similares aos miRNAs.

Outros ncRNAs

- **ncRNAs grandes ou fRNAs (functional RNAs)**
  São transcritos pela RNA polimerase II, sofrem "splicing" e possuem cauda poli-A. Vários já foram caracterizados, como o gene XIST (envolvido na inativação do cromossomo X), AIR e H19 (envolvidos no "imprint" de genes paternos e maternos respectivamente), EVF e CTN.

possibilidade de moléculas de RNA desempenharem atividade catalítica (veja também o Capítulo 2 para detalhes sobre a descoberta das ribozimas). Recentemente foi identificada participação de moléculas de RNA dupla-fita em um potente mecanismo de regulação da expressão gênica. Esse mecanismo é conhecido como interferência de RNA (**RNAi**) e essa descoberta mudou a forma de como devemos enxergar o metabolismo do gene na célula eucariótica, além de abrir excelentes perspectivas para seu uso como ferramenta de estudo da célula e como possível agente terapêutico.

Os primeiros indícios da existência do mecanismo de RNAi vieram de experimentos com petúnias no início da década de 90 (Vanblokland *et al*., 1994), quando os pesquisadores induziram a superexpressão da proteína responsável pela síntese do pigmento violeta nas flores dessas plantas. Esperava-se obter flores com coloração violeta intensa, mas o fenótipo observado foi o oposto: as plantas produziram flores brancas ou rajadas de branco e violeta, indicando uma redução (e não aumento) da síntese do pigmento. Esse fenômeno foi conhecido como co-supressão, mas o mecanismo responsável por ele ficou desconhecido por vários anos. A ausência de síntese do pigmento foi o primeiro exemplo de silenciamento do transgene (e do gene endógeno) através do mecanismo de RNAi.

Esse fenômeno foi encontrado também em outros organismos eucariotos como na mosca drosófila e no verme nematóide *Caenorhabditis elegans*. O mecanismo de co-supressão só foi descrito anos mais tarde, quando os grupos dos pesquisadores americanos Andrew Z. Fire e Craig C. Mello publicaram em conjunto resultados de seus trabalhos com *C.elegans*, revelando que pequenas moléculas de RNA dupla-fita (dsRNA, de "double strand RNA", em inglês) poderiam silenciar a expressão de genes (Fire *et al*., 1998). Os experimentos apresentados chamam atenção pela simplicidade. Ao analisar os efeitos já conhecidos de inibição de expressão gênica por moléculas de RNA simples fita anti-senso e também senso (similar aos dados de co-supressão), os pesquisadores descobriram que o uso simultâneo das duas moléculas tinha um efeito sinergístico. Alguns experimentos e controles suplementares revelaram que a molécula efetora era, na verdade, o dsRNA. Pela habilidade do dsRNA de interferir na expressão genética, o mecanismo foi chamado de interferência por RNA, ou RNAi. Em 2006, apenas 8 anos após sua descoberta, Fire e Mello receberam o Prêmio Nobel de Medicina.

O mecanismo de RNAi é extremamente conservado em eucariotos, sendo encontrado em organismos tão diversos como protozoários, fungos, plantas, nematóides, insetos e até mesmo em mamíferos. O mecanismo de RNAi foi inicialmente associado à defesa contra patógenos virais, mas é provável que sua função seja muito mais ampla, uma vez que se descobriu que células eucarióticas codificam a partir de seu genoma pequenos RNAs chamados microRNAs (**miRNA**). Esses RNAs endógenos são capazes de "guiar" o silenciamento, alterando a expressão de genes codificadores de proteína e, portanto, constituem em importante maquinaria para controle e regulação gênica.

4.2. Interferência de RNA Endógena - O miRNA

Desde a descrição dos primeiros miRNAs em *C. elegans* (Lee *et al*., 1993), os miRNA passaram a ser reconhecidos como importantes moduladores da expressão gênica. Esses pequenos RNAs (21 a 26 nucleotídeos) são capazes de controlar a estabilidade do mRNA, assim como a sua tradução. Além disso, podem induzir modificações epigenéticas no genoma. O silenciamento realizado por esses pequenos RNAs estabeleceu um novo paradigma no entendimento da regulação gênica em eucariotos e revelou novas defesas contra vírus e transposons (Sijen e Plasterk, 2003).

A anotação da posição genômica dos miRNAs indicou que muitos deles se encontram em regiões intergênicas (a uma distância de pelo menos 1 Kbp de genes anotados/preditos), apesar de também existirem miRNAs em regiões intrônicas de genes, na orientação senso e antisenso em relação ao gene (Lagos-Quintana *et al*., 2001; Lagos-Quintana *et al*., 2003; Lau *et al*., 2001). Até o momento, foram identificados cerca de 500 miRNAs codificados a partir do genoma humano. Existem evidências de que genes de miRNAs formam unidades de transcrição policistrônicas, pois cerca de 50% dos miRNAs conhecidos são encontrados próximos de outros miRNAs (Lagos-Quintana *et al*., 2001; Lau *et al*., 2001) e podem ser transcritos a partir de um mesmo promotor (Cai *et al*., 2004; Lee, 2004), gerando transcritos primários policistrônicos (**pri-miRNAs**).

A síntese e ação dos miRNAs estão ilustradas na Figura 4.1. Os miRNAs são transcritos pela polimerase do RNA II (RNA pol II, (Cai *et al*., 2004; Lee, 2004). Inicialmente acreditava-se que a polimerase do RNA III fosse a responsável pela transcrição dos genes de miRNA, já que ela esta está envolvida na transcrição da grande maioria dos pequenos RNAs, como os tRNAs. Entretanto, os precursores pri-miRNAs não são pequenos, chegando algumas vezes a várias kilobases (milhares de bases) de extensão, um tamanho que é incompatível com a transcrição pela RNA pol III (Kato *et al*., 2005; Lee, 2004). Apesar dessas evidências, ainda não se descarta a possibilidade de que um pequeno número de miRNAs seja transcrito por outras RNA polimerases (Lee, 2004). Os precursores pri-miRNAs são protegidos pelo radical 7MGpppG e poliadenilados, sendo a proteína **DROSHA** (RNAse III) e seu cofator **PASHA** as responsáveis pelo processamento dos pri-miRNAs, convertendo-os em **pré-miRNAs**, com o tamanho de

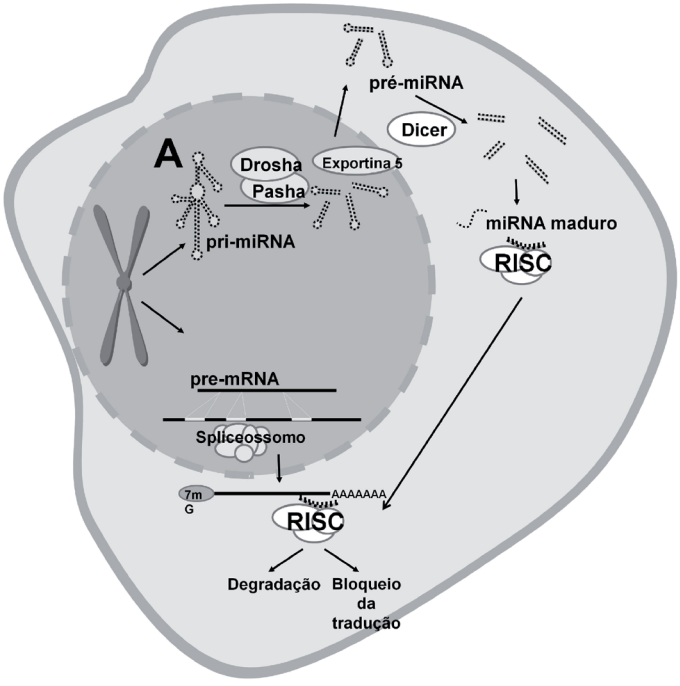


Figura 4.1. Esquema da biogênese de miRNAs e o seu mecanismo de ação em células eucarióticas. O pri-miRNA nascente (A) é primeiramente processado em um pré-miRNA com cerca de 70 nucleotídeos, pela enzima DROSHA. O pré-miRNA é então transportado para o citoplasma pela proteína Exportina 5 e processado pela enzima DICER em duplexes de miRNA. Apenas uma fita do duplex é incorporada no complexo de silenciamento RISC, que por complementaridade parcial ou total induz a repressão traducional ou a degradação do mRNA alvo.

cerca de 70 nucleotídeos, que se estruturam na forma de grampo ("hairpin"), formando o dsRNA (Nakahara e Carthew, 2004).

RNA-GTP e a Exportina-5 são as proteínas responsáveis pelo transporte do pré-miRNA para fora do núcleo, onde outra RNAse III, a proteína DICER (Bernstein *et al*., 2001), processa o precursor transformando-o em um duplex transitório de cerca de 22 nucleotídeos. O complexo protéico **RISC** (complexo de silenciamento induzido por RNA) inclui as proteínas da família Argonauta (Carmell *et al*., 2002) que se associam a uma das fitas do duplex (conhecida como RNA guia). Através da associação a sítios complementares na região 3´ dos RNAs mensageiros (mRNAs), o complexo RISC-miRNA regula negativamente a expressão gênica, podendo causar a degradação ou o impedimento da tradução do mensageiro. Acredita-se que um único miRNA pode controlar a expressão de centenas de mRNAs, modulando o metabolismo celular em vários aspectos. Além de interagir com o mRNA, também existem evidências de que miRNAS podem atuar diretamente no DNA, sendo responsáveis pela estruturação heterocromática do genoma (Fukagawa *et al*., 2004; Kanellopoulou *et al*., 2005), sendo também capazes de silenciar um gene através da metilação da região promotora (Morris *et al*., 2004).

4.3. Como a Maquinaria de RNAi Silencia Genes Específicos

Classicamente, as funções de diversos genes na célula foram elucidadas por mutações ou por meio de técnicas que empregam a deleção ou a redução nos níveis de expressão de um determinado gene, seguida da análise do fenótipo resultante. Entretanto, essas técnicas muitas vezes envolvem processos complexos que restringem sua aplicação em poucos modelos biológicos. Por exemplo, o emprego de camundongos com genes nocauteados (*knockout*, em inglês) depende de tecnologia elaborada, trabalhosa e que requer muito tempo (Thomas e Capecchi, 1987). Em células humanas esse tipo de abordagem é praticamente impossível. Nesse sentido, o conhecimento a respeito da biogênese dos miRNAs e do mecanismo de RNAi no controle da expressão gênica permitiu que esse processo celular fosse rapidamente utilizado em estudos de genômica funcional em diferentes organismos. Assim, a descoberta da existência desse mecanismo também em células humanas (Elbashir *et al*., 2001) revelou a importância evolutiva da via de RNAi, assim como abriu perspectivas de uso desse mecanismo como base para silenciamento gênico em mamíferos, com grande potencial tecnológico. Assim, esse mecanismo passou a ser largamente explorado tanto em células cultivadas *in vitro*, como diretamente em animais *in vivo*, através do uso de pequenas moléculas artificiais de dsRNA designadas **siRNA** (do inglês, "small interfering RNA").

Atualmente, muitas empresas de biotecnologia sintetizam moléculas de siRNA direcionadas para o silenciamento de qualquer gene, bastando que para isso a sequência deste seja conhecida. Em geral, as duplex de siRNA têm tamanho variando entre 19 a 30 pares de bases, cujas sequências são determinadas através de algoritmos específicos (em geral de acesso livre), que se baseiam na sequência do gene alvo para encontrar uma região complementar que possa atuar como RNAi. Na maioria das vezes, esses siRNA são complexados a macromoléculas de natureza química variável (lipossomos ou policátions) com a finalidade de maximizar a transfecção (ou seja, a incorporação) do siRNA por diferentes tipos celulares (Menck, 2006). A incorporação dessas moléculas pela célula resulta em um silenciamento robusto e específico do gene de sequência homóloga ao duplex de siRNA, o que implica a diminuição e/ou depleção no acúmulo da proteína codificada pelo mesmo, em processo conhecido, em inglês, como "knock-down". Para tal, as moléculas de siRNA utilizam parte da via de processamento dos miRNA para exercer sua ação final, como ilustrado na Figura 4.2. Em outras palavras, uma vez dentro da célula, os siRNAs mimetizam os substratos da enzima DICER (i.e, os pré-miRNAs), sendo clivados pela mesma até atingir o tamanho ideal para serem incorporados no complexo multienzimático denominado RISC que contém a proteína argonauta 2 (AGO2). Nesse complexo, uma das fitas do duplex de siRNA é clivada (fita senso), enquanto que a fita remanescente (fita antisenso ou fita guia) se liga à molécula de RNA mensageiro de sequência complementar. Nesse caso, o complexo RISC-AGO2 cliva a fita do mRNA entre os nucleotídeos complementares aos que se encontram na posição 10 e 11 da extremidade 5' da fita guia, causando a degradação deste mRNA. Com isso ocorre a redução da síntese da proteína codificada pelo gene alvo na célula transfectada (Filipowicz *et al*., 2005; Matzke e Birchler, 2005; Scott, 2005). Vale mencionar que muitas vezes os siRNA utilizados apresentam tamanhos inferiores a 21 nucleotídeos, sendo então dispensável a ação da enzima DICER nessas moléculas. Nesses casos, essas moléculas são diretamente incorporadas ao complexo RISC-AGO2, também provocando a degradação do mRNA do gene alvo.

Um aspecto que deve ser considerado nos estudos que empregam essa metodologia é a duração do silenciamento gênico. A diminuição nos níveis de expressão gênica induzida por siRNA é transitório e perdura por cerca de 3 a 5 dias após a adição do duplex em células em cultura. Tal fato pode ser um fator limitante quando se pretende estudar proteínas de meia vida longa, por exemplo. Uma alternativa eficaz para contornar tal limitação consiste na expressão estável de moléculas efetoras do processo de RNAi por meio do uso de vetores plasmidiais ou virais. Esses últimos ainda apresentam uma vantagem adicional por serem capazes de transduzirem células nas quais o processo de transfecção é ineficiente. A transcrição do transgene inserido nesses vetores produz moléculas de RNA palindrômicas que produzem uma estrutura secundária em forma de grampo, conhecida como shRNA (do inglês, "short hairpin RNA"). Essas moléculas são processadas pela maquinaria de RNAi e provocam o silenciamento a longo prazo do gene alvo (Mohammed *et al*., 2005).

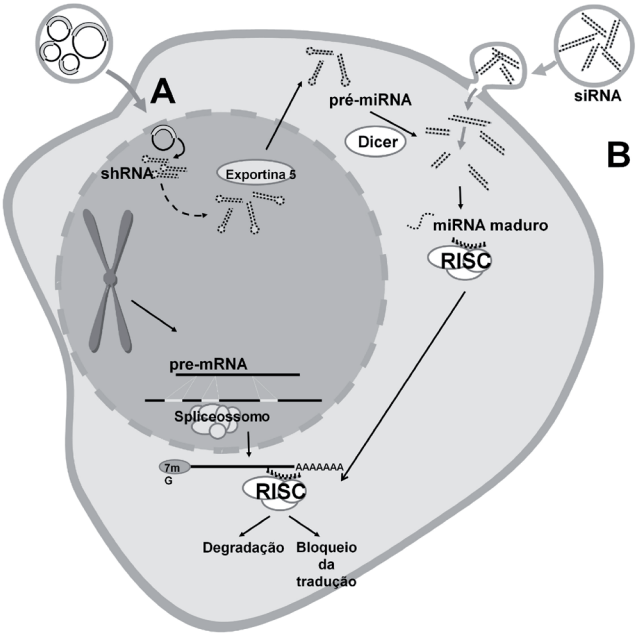


Figura 4.2. Esquema do mecanismo de RNAi induzido por moléculas exógenas em células eucarióticas. As moléculas de shRNA (A) e de siRNA (B) exógenas são desenhadas de maneira a simular o substrato da enzima DICER, gerando pequenas moléculas de RNAdf. A interação desses duplexes com a enzima DICER auxilia o acoplamento da fita guia no complexo RISC, promovendo a degradação do mRNA alvo.

Muitos grupos de pesquisa fazem uso de vetores que expressam shRNA com semelhança estrutural às moléculas precursoras dos miRNA (pri- e pré-miRNA), o que pode maximizar a eficiência do silenciamento (Chang *et al*., 2006). Além disso, esses vetores podem ser utilizados na criação de linhagens celulares imortalizadas, nas quais a supressão do gene alvo torna-se permanente, bem como na geração de animais transgênicos, com silenciamento de genes específicos (Paddison *et al*., 2002). Atualmente, várias bibliotecas plasmidiais contendo genes que participam de distintos processos biológicos como proliferação e morte celular, por exemplo, já se encontram disponíveis comercialmente.

4.4. Evolução do RNAi

As ribonucleases são proteínas capazes de degradar o RNA, controlando a expressão protéica. Essas proteínas estão amplamente distribuidas entre os seres vivos, sendo encontradas em organismos de todos os reinos. Como descrito acima, os eucariotos desenvolveram uma versão elaborada desse controle, onde uma série de proteínas é capaz de realizar a degradação dirigida por pequenos RNAs, regulando a expressão de transcritos específicos (Anantharaman *et al*., 2002; Cogoni e Macino, 2000). Esse fenômeno, que inclui a via de RNAi, é chamado de silenciamento gênico pós-transcricional (em inglês *post-transcriptional gene silencing* ou PTGS) nos eucariotos. A função ancestral exercida pelo mecanismo de RNAi é desconhecida. A ausência de proteínas importantes desse mecanismo em vários grupos taxonômicos, como *Saccharomyces cerevisiae*, *Trypanosoma cruzi*, *Leischmania major*, *Cyanidioschyzon merolae* e *Plasmodium falciparum*, sugere que essas não são essenciais. No entanto, a análise filogenética indica que o último ancestral comum eucarionte já possuía uma via simplificada do mecanismo de RNAi, sendo que alguns organismos perderam determinados genes envolvidos nesse mecanismo durante a evolução (Cerutti e Casas-Mollano, 2006).

Desde a sua descoberta, postulou-se que uma das primeiras funções da maquinaria de RNAi era defender as células contra parasitas genômicos, como elementos transponíveis e vírus (Buchon e Vaury, 2005; Grant, 1999; Waterhouse *et al*., 2001). Essa função manteve-se conservada entre os eucariotos (Cerutti e Casas-Mollano, 2006), sendo que novas funções começaram a ser associadas ao mecanismo de RNAi, como o controle de regiões heterocromáticas (Kanellopoulou *et al*., 2005; Murchison *et al*., 2005), ou mesmo a regulação de expressão gênica através de miRNAs (Rana, 2007). De fato, os miRNAs foram identificados apenas em plantas e animais multicelulares, mas estão aparentemente ausentes em vários eucariotos unicelulares, como *S. pombe* e *T. brucei* (Djikeng *et al*., 2001; Reinhart e Bartel, 2002).

Existem evidências de que o sistema de regulação por RNAs foi essencial para a evolução de organismos multicelulares complexos e para a expansão da complexidade fenotípica (Mattick, 1994; Mattick, 2004; Mattick e Gagen, 2001). Uma dessas evidências é a proporção de genes codificantes de proteínas que diminui em função da complexidade do organismo (Taft e Mattick, 2003). Além disso, os sinais regulatórios dos RNAs são processados em paralelo com os sinais protéicos, sendo que sinais provenientes dos RNAs regulatórios são capazes de interferir com a regulação de genes codificadores de proteínas independentemente das características bioquímicas da proteína-alvo (Mattick, 1994; Mattick e Gagen, 2001). Essa capacidade de regulação independente da característica bioquímica fez com que alguns pesquisadores se referissem à regulação via RNA como "regulação digital", enquanto as proteínas seriam as responsáveis pela "regulação analógica" (Mattick, 2004). O mecanismo de RNAi em especial possui uma alta flexibilidade regulatória, pois uma mesma base protéica (o complexo RISC) pode ser utilizada para a regulação de todo o genoma, sendo apenas necessário substituir o miRNA acoplado ao complexo.

Apesar de que o sistema de silenciamento por RNA seja considerado um mecanismo altamente conservado, ainda existem dúvidas sobre a sua presença nos organismos procariotos. Pequenos RNAs não codificantes são encontrados em bactérias há algum tempo (Gottesman, 2005; Wagner *et al*., 2002; Wagner e Brantl, 1998), entretanto não se sabe se o mecanismo pelo qual esses RNAs controlam a expressão de proteínas está relacionado com o silenciamento gênico tal como ele é observado nos eucariotos. Proteínas similares às Argonautas são encontradas em arqueobactérias e em *Aquifex* (Anantharaman *et al*., 2002; Cerutti *et al*., 2000). Essas versões ancestrais da proteína argonauta possuem apenas o domínio protéico PIWI, faltando o domínio PAZ, responsável pela ligação a pequenos RNAs (Anantharaman *et al*., 2002). Dessa maneira, acredita-se que elas seriam capazes de se ligar e quebrar moléculas de RNA, mas sem serem guiadas por pequenos RNAs.

Alguns pesquisadores sugerem que agrupamentos de pequenas sequências palindrômicas ("clustered regularly interspaced short palindromic repeats", ou CRISPRs) poderiam atuar como um mecanimo de RNAi em procariotos, conferindo resistência a infecções virais (Makarova *et al*., 2006; Sorek *et al*., 2008). Os CRISPRs estão presentes em várias bactérias e em quase todas as arqueobactérias (Sorek *et al*., 2008), sendo considerado um dos sistemas de defesa antiviral mais antigo do mundo microbiano (Makarova *et al*., 2006). Essas sequências seriam capazes de determinar a especificidade de uma resposta imune em procariotos, protegendo esses organismos contra bacteriófagos (Barrangou *et al*., 2007). Embora esteja voltado principalmente para a defesa antiviral, sugere-se que os CRISPRs seriam capazes de controlar genes endógenos, de maneira análoga ao sistema de RNAi em eucariotos. De fato, alguns autores demonstram que até 35% dos CRISPRs são complementares a genes cromossômicos bacterianos, reforçando a teoria de que os CRISPRs seriam capazes de atuar no controle da expressão de genes da própria bactéria (Horvath *et al*., 2008; Mojica *et al*., 2005). Muitos avanços foram obtidos na compreensão dessas sequências palidrômicas. Entretanto, ainda existem dúvidas se elas seriam capazes de funcionar como um sistema de silenciamento.

4.5. O RNAi na Terapia Gênica

Apesar da demonstração da existência de RNAi em células de mamíferos ter sido relatada recentemente (Elbashir *et al*., 2001), estratégias terapêuticas que exploram essa via se mostram muito promissoras, principalmente no caso de doenças humanas refratárias aos medicamentos existentes. Em 2003, foi apresentada a primeira evidência *in vivo* da eficácia do uso de duplexes de siRNA no tratamento de hepatite severa em camundongos (Song *et al*., 2003). O fato dessas moléculas sintéticas de siRNA serem incorporadas em uma via de sinalização intracelular com relativa facilidade propiciou a elaboração de estratégias que empregam a administração de siRNA para o silenciamento de genes cuja ativação relaciona-se de alguma forma ao desenvolvimento de uma enfermidade. Atualmente já foram publicados mais de 100 artigos científicos relatando o funcionamento de RNAi diretamente com experimentos em animais, alguns cujo gene alvo do silenciamento pode trazer benefícios terapêuticos. Várias empresas farmacêuticas já concluíram testes pré-clínicos em animais e já estão realizando protocolos clínicos de fase I (para avaliação de

segurança) em humanos empregando siRNA como molécula para a terapia. De fato, vários são os problemas de saúde que resultam de expressão aumentada de um ou mais genes, como as neoplasias malignas, hipercolesterolemia e infecções virais, dentre outros, e o uso de siRNA nesses casos é promissor (Menck, 2006).

Por outro lado, alguns fatores devem ser cuidadosamente analisados no uso racional de siRNAs como moléculas terapêuticas. Muitos trabalhos já relataram, por exemplo, a redução na expressão de genes diferentes do alvo do siRNA utilizado, efeito esse denominado de "off-target" (Jackson *et al*., 2003). Especula-se que esse efeito seja resultado do reconhecimento de sequências similares, mas não idênticas (i.e, complementação apenas parcial), ao alvo localizadas na região 3' não traduzida de diversos genes, através de mecanismo similar ao de miRNA. Ainda, a instabilidade estrutural da molécula na circulação sanguínea periférica, consequência da ação de nucleases, o alcance e penetrância do duplex nos tecidos alvos, além da possível ativação do sistema imune representam barreiras que devem ser contornadas para garantir o sucesso de terapias que exploram o mecanismo de RNAi *(*de Fougerolles *et al*., 2007).

Paradoxalmente, evidências de que pequenos RNAs regulam positivamente a ativação de genes cognatos foram recentemente descritas (Janowski *et al*., 2007; Kuwabara *et al*., 2004; Li *et al*., 2006) e cunhou-se o termo "RNA activators" (**RNAa**) para designar essas moléculas. Embora essa hipótese não seja facilmente aceita pela comunidade científica, é razoável supor que a complexidade do sistema de regulação gênica por pequenos RNAs vá além do que já conhecemos e, consequentemente, a investigação cautelosa de possíveis efeitos secundários por siRNAs devem ser considerados tanto nos estudos *in vitro* como *in vivo*.

4.6. Conclusões

Até poucos anos atrás, a molécula de RNA era vista basicamente como a molécula intermediária sintetizada no momento em que a informação biológica flui do DNA para proteína, sendo esse o cerne do funcionamento celular. De fato, foi relativamente fácil entre os biologistas assumir e generalizar que genes eram "sinônimos" à informação necessária para a síntese de proteínas, sendo essa premissa verdadeira para procariotos cujo genoma é formado quase que unicamente por sequências gênicas codificadoras de proteínas. Contudo, inúmeras são as evidências de que eucariotos, organismos nos quais o genoma apresenta poucas sequências codificadoras, apresentam mecanismos mais complexos que controlam a expressão gênica. Nesse sentido, em 1969, foi proposto que o RNA controla a expressão de certos genes em células eucarióticas (Britten e Davidson, 1969); entretanto, com a descoberta dos fatores de transcrição, a idéia de silenciamento gênico por RNA através da complementaridade entre DNA-RNA permaneceu no ostracismo por alguns anos até a década de 90, quando os RNA não codificantes apareceram como reguladores essenciais da expressão gênica em diferentes processos biológicos.

O mecanismo de RNAi foi descrito nos eucariotos como uma versão mais elaborada do processo de degradação das moléculas de RNA. Uma das primeiras funções desse mecanismo foi defender o genoma contra fagos e transposons, sendo que outras funções também foram atribuídas ao sistema de silenciamento, como a manutenção da estrutura heterocromática e o controle da expressão gênica via miRNAs. Existem evidências de que a regulação por RNAs foi essencial para o surgimento da complexidade dos organismos multicelulares. Apesar de altamente conservado, existem dúvidas de que o mecanismo de RNAi exista nos organismos procariotos, sendo possível, porém, que esses organismos possuam um sistema análogo ao encontrado nos eucariotos.

Atualmente, o fenômeno de RNAi começa a revolucionar a Biologia experimental desde os organismos unicelulares até os pluricelulares. A compreensão desse mecanismo permitiu que o mesmo fosse utilizado tanto na ciência básica, como estratégia alternativa de "knock-down" gênico, como na ciência aplicada, onde a administração de pequenas moléculas de siRNA surge como alternativa promissora no tratamento de doenças caracterizadas pela expressão aberrante de determinados genes. De fato, após a demonstração da participação de dsRNA no silenciamento de genes em petúnia e *C. elegans*, a molécula de RNA passou a ser vista com outros olhos - assim como o dogma central da biologia molecular.

Referências Bibliográficas

Anantharaman, V., Koonin, E.V. e Aravind, L. (2002). Comparative genomics and evolution of proteins involved in RNA metabolism. **Nucleic Acids Res. 30**: 1427-1464.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D.A. e Horvath, P. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. **Science 315**: 1709-1712.

Bernstein E., Caudy, A.A., Hammond, S.M. e Hannon, G.J. (2001). Role for a bidentate ribonuclease in the initiation step of RNA interference. **Nature 409**: 363.

Britten R.J. e Davidson, E.H. (1969). Gene regulation for higher cells: a theory. **Science 165**: 349-57.

Buchon N. e Vaury, C. (2005). RNAi: a defensive RNA-silencing against viruses and transposable elements. **Heredity 96**: 195-202.

Cai X., Hagedorn, C.H. e Cullen, B. R. (2004). Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. **RNA 10**: 1957.

Carmell M.A., Xuan, Z., Zhang, M.Q. e Hannon, G. J. (2002). The Argonaute family: tentacles that reach into RNAi, developmental control, stem cell maintenance, and tumorigenesis. **Genes Dev. 16**: 2733.

Cerutti H. e Casas-Mollano, J. (2006). On the origin and functions of RNA-mediated silencing: from protists to man. **Current Genetics 50**: 81-99.

Cerutti L., Mian, N. e Bateman, A. (2000). Domains in gene silencing and cell differentiation proteins: the novel PAZ domain and redefinition of the Piwi domain. **Trends Biochem. Sci. 25**: 481-482.

Chang K., Elledge, S.J. e Hannon, G.J. (2006). Lessons from Nature: microRNA-based shRNA libraries. **Nat. Meth. 3**: 707-714.

Cogoni C. e Macino, G. (2000). Post-transcriptional gene silencing across kingdoms. **Curr. Opin. Genet. Dev. 10**: 638.

de Fougerolles A., Vornlocher, H.-P., Maraganore, J. e Lieberman, J. (2007). Interfering with disease: a progress report on siRNA-based therapeutics. **Nat. Rev. Drug. Discov. 6**: 443-453.

Djikeng A., Shi, H., Tschudi, C. e Ullu, E. (2001). RNA interference in Trypanosoma brucei: cloning of small interfering RNAs provides evidence for retroposon-derived 24-26-nucleotide RNAs. **RNA 7**: 1522-1530.

Eddy S.R. (2001). Non-coding RNA genes and the modern RNA world. **Nat. Rev. Genet. 2**: 919-29.

Elbashir S. M., Harborth, J., Lendeckel, W., Yalcin, A., Weber, K. e Tuschl, T. (2001). Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. **Nature 411**: 494-498.

Filipowicz W., Jaskiewicz, L., Kolb, F.A. e Pillai, R.S. (2005). Post-transcriptional gene silencing by siRNAs and miRNAs. **Curr. Opin. Struc. Biol. 15**: 331-341.

Fire A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S. E. e Mello, C.C. (1998). Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. **Nature 391**: 806-811.

Frith M. C., Pheasant, M. e Mattick, J. S. (2005). Genomics: The amazing complexity of the human transcriptome. **Eur. J. Hum. Genet. 13**: 894-897.

Fukagawa T., Nogami, M., Yoshikawa, M., Ikeno, M., Okazaki, T., Takami, Y., Nakayama, T. e Oshimura, M. (2004). Dicer is essential for formation of the heterochromatin structure in vertebrate cells. **Nat. Cell Biol. 6**: 784.

Gottesman S. (2005). Micros for microbes: non-coding regulatory RNAs in bacteria. **Trends Genet. 21**:399-404.

Grant S.R. (1999). Dissecting the mechanisms of posttranscriptional gene

silencing: Divide and conquer. **Cell 96**: 303-306.
Horvath P., Romero, D.A., Coute-Monvoisin, A.-C., Richards, M., Deveau, H., Moineau, S., Boyaval, P., Fremaux, C. e Barrangou, R. (2008). Diversity, activity, and evolution of CRISPR loci in *Streptococcus thermophilus*. **J. Bacteriol. 190**: 1401-1412.
Jackson A.L., Bartz, S. R., Schelter, J., Kobayashi, S.V., Burchard, J., Mao, M., Li, B., Cavet, G. e Linsley, P.S. (2003). Expression profiling reveals off-target gene regulation by RNAi. **Nat. Biotech. 21**: 635.
Janowski B.A., Younger, S.T., Hardy, D.B., Ram, R., Huffman, K.E. e Corey, D.R. (2007). Activating gene expression in mammalian cells with promoter-targeted duplex RNAs. **Nat. Chem. Biol. 3**: 166-173.
John S. M. (2003). Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. **Bioessays 25**: 930-939.
Kanellopoulou C., Muljo, S.A., Kung, A.L., Ganesan, S., Drapkin, R., Jenuwein, T., Livingston, D.M. e Rajewsky, K. (2005). Dicer-deficient mouse embryonic stem cells are defective in differentiation and centromeric silencing. **Genes Dev. 19**: 489-501.
Kato H., Goto, D.B., Martienssen, R.A., Urano, T., Furukawa, K. e Murakami, Y. (2005). RNA Polymerase II Is Required for RNAi-Dependent Heterochromatin Assembly. **Science 309**: 467-469.
Kuwabara T., Hsieh, J., Nakashima, K., Taira, K. e Gage, F.H. (2004). A small modulatory dsRNA specifies the fate of adult neural stem cells. **Cell 116**: 779-793.
Lagos-Quintana M., Rauhut, R., Lendeckel, W. e Tuschl, T. (2001). Identification of novel genes coding for small expressed RNAs. **Science 294**: 853.
Lagos-Quintana M., Rauhut, R., Meyer, J., Borkhardt, A. e Tuschl, T. (2003). New microRNAs from mouse and human. **RNA 9**: 175.
Lau N.C., Lim, L.P., Weinstein, E.G. e Bartel, D.P. (2001). An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. **Science 294**: 858.
Lee R. C., Feinbaum, R.L. e Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. **Cell 75**: 843.
Lee Y. (2004). MicroRNA genes are transcribed by RNA polymerase II. **EMBO J. 23**: 4051.
Li L.-C., Okino, S.T., Zhao, H., Pookot, D., Place, R.F., Urakami, S., Enokida, H. e Dahiya, R. (2006). Small dsRNAs induce transcriptional activation in human cells. **Proc. Natl. Acad. Sci. USA 103**: 17337-17342.
Makarova K., Grishin, N., Shabalina, S., Wolf, Y. e Koonin, E. (2006). A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. **Biol. Direct 1**: 7.
Mattick J.S. (1994). Introns: evolution and function. **Curr. Opin. Genet. Dev. 4**: 823-831.
Mattick J.S. (2004). RNA regulation: a new genetics? **Nat. Rev. Genet. 5**: 316-323.
Mattick J.S. (2005). The Functional Genomics of Noncoding RNA. **Science 309**: 1527-1528.
Mattick J.S. e Gagen, M.J. (2001). The evolution of controlled multitasked gene networks: The role of introns and other noncoding RNAs in the development of complex organisms. **Mol. Biol. Evol. 18**: 1611-1630.
Mattick J.S. e Makunin, I.V. (2005). Small regulatory RNAs in mammals. **Hum. Mol. Genet. 14**: R121-132.
Matzke M. A. e Birchler, J. A. (2005). RNAi-mediated pathways in the nucleus. Nat Rev Genet 6: 24-35.
Menck C.F.M. (2006). RNA INTERFERÊNCIA - A nova corrida do ouro. **Microbiologia in foco 1**: 17-21.
Mohammed A., John, J.R. e Dongho, K. (2005). Approaches for chemically synthesized siRNA and vector-mediated RNAi. **FEBS Lett. 579**: 5974-5981.
Mojica F.J.M., Díez-Villaseñor, C. s., García-Martínez, J. e Soria, E. (2005). Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. **J. Mol. Evol. 60**: 174-182.
Morris K.V., Chan, S.W.L., Jacobsen, S.E. e Looney, D.J. (2004). Small interfering RNA-induced transcriptional gene silencing in human cells. **Science 305**: 1289-1292.
Murchison E.P., Partridge, J.F., Tam, O.H., Cheloufi, S. e Hannon, G.J. (2005). Characterization of dicer-deficient murine embryonic stem cells. **Proc. Natl. Acad. Sci. USA 102**: 12135-12140.
Nakahara K. e Carthew, R.W. (2004). Expanding roles for miRNAs and siRNAs in cell regulation. **Curr. Opin. in Cell Biol. 16**: 127.
Paddison P. J., Caudy, A.A., Bernstein, E., Hannon, G.J. e Conklin, D.S. (2002). Short hairpin RNAs (shRNAs) induce sequence-specific silencing in mammalian cells. **Genes Dev. 16**: 948-958.
Rana T.M. (2007). Illuminating the silence: understanding the structure and function of small RNAs. **Nat. Rev. Mol. Cell. Biol. 8**: 23-36.
Reinhart B.J. e Bartel, D.P. (2002). Small RNAs correspond to centromere heterochromatic repeats. **Science 297**: 1831.
Scott M.H. (2005). Dicing and slicing: The core machinery of the RNA interference pathway. **FEBS Lett. 579**: 5822-5829.
Sijen T. e Plasterk, R.H.A. (2003). Transposon silencing in the *Caenorhabditis elegans* germ line by natural RNAi. **Nature 426**: 310.
Song E., Lee, S.-K., Wang, J., Ince, N., Ouyang, N., Min, J., Chen, J., Shankar, P. e Lieberman, J. (2003). RNA interference targeting Fas protects mice from fulminant hepatitis. **Nat. Med. 9**: 347-351.
Sorek R., Kunin, V. e Hugenholtz, P. (2008). CRISPR [mdash] a widespread system that provides acquired resistance against phages in bacteria and archaea. **Nat. Rev. Microbiol. 6**: 181-186.
Taft R. e Mattick, J. (2003). Increasing biological complexity is positively correlated with the relative genome-wide expansion of non-protein-coding DNA sequences. **Bioessays 5**: P1.
The, F. C., Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., et al. (2005). The transcriptional landscape of the mammalian genome. **Science 309**: 1559-1563.
Thomas K. R. e Capecchi, M. R. (1987). Site-directed mutagenesis by gene targeting in mouse embryo-derived stem cells. **Cell 51**: 503-512.
Vanblokland R., Vandergeest, N., Mol, J.N.M. e Kooter, J.M. (1994). Transgene-mediated suppression of chalcone synthase expression in **Petunia hybrida**. Results from an increase in RNA turnover. **Plant J. 6**: 861-877.
Wagner E.G.H., Altuvia, S., Romby, P. e Jay, C. D. a. C. t. W. (2002). Antisense RNAs in bacteria and their genetic elements. **Adv. Genet. 46**:361-398.
Wagner E.G.H. e Brantl, S. (1998). Kissing and RNA stability in antisense control of plasmid replication. **Trends Biochem. Sci. 23**: 451-454.
Waterhouse P.M., Wang, M.-B. e Lough, T. (2001). Gene silencing as an adaptive defence against viruses. **Nature 411**: 834-842.

Capítulo 5

Estabilidade do material genético: mutagênese e reparo

Luis Eduardo Soares Netto (nettoles@ib.usp.br)
Departamento de Genética e Biologia evolutiva
Instituto de Biociências
Universidade de São Paulo

Carlos Frederico Martins Menck (cfmmenck@usp.br)
Departamento de Microbiologia
Instituto de Ciências Biomédicas
Universidade de São Paulo

"*Triste não é mudar de idéias. Triste é não ter idéias para mudar.*" (Barão de Itararé)

5.1. Histórico

Após a redescoberta das Leis de Mendel, no início século XX, o grande desafio no meio científico era a identificação da composição química dos genes. Ainda no início dos anos 1900, desconfiava-se que os genes estavam nos cromossomos: Sutton e Boveri demonstraram independentemente que havia um paralelismo entre os fatores hereditários de Mendel e o comportamento dos cromossomos na meiose e na fertilização (Sutton, 1902; Boveri, 1902). Devido ao fato de os cromossomos serem constituídos de ácidos nucléicos e proteínas, postulava-se que uma dessas macromoléculas poderia ser a matéria-prima dos genes. Como o DNA é uma molécula quimicamente homogênea e estável, enquanto as proteínas apresentam uma grande variedade estrutural, muitos pesquisadores naquele período acreditavam que eram as proteínas que carregavam a informação genética. Experimentos realizados por Griffith, Avery, McLeod e McCarty durante as décadas de 1940 e 1950 mostraram elegantemente que o DNA é o material genético das células. Em 1953, Watson e Crick apresentaram a estrutura de dupla hélice do DNA baseada em dados de difração de raios X. O modelo proposto era condizente com todas as propriedades físico-químicas do DNA. Além disso, através do modelo da dupla hélice, Watson e Crick previram que a replicação do DNA seria semi-conservativa (isto é, cada uma das duas fitas de uma molécula de DNA serve como molde para a síntese de uma nova cadeia), como realmente foi comprovado cinco anos mais tarde por Meselson e Stahl (1952). O modelo da dupla hélice também é coerente com a expressão do gene em proteínas.

Apesar da enorme relevância do modelo da dupla hélice, Watson e Crick não imaginaram que existisse nas células um grande gasto de energia metabólica para reparar lesões e, consequentemente, manter estáveis as informações contidas na sequência de bases do DNA (Watson e Crick, 1974). O início da descoberta de processos de reparo de DNA ocorreu no final da década de 1940, uma descoberta acidental em experimentos com outros propósitos, nos quais se verificou que alguns organismos (fungos e bactérias), quando mantidos na luz solar (próximo à janela), se mostravam menos sensíveis a irradiação ultravioleta, fenômeno conhecido como fotorreativação. No entanto, apenas na década de 1960 é que começou a ficar claro que a célula possui vários sistemas para proteção do genoma (para uma interessante leitura dos processos de descoberta, recomendamos a leitura do livro de Friedberg, 1997). Em parte por seu grande tamanho (na célula humana, temos um genoma com três bilhões de pares de base), o genoma não é muito estável, sofre alterações durante sua própria síntese ou é alvo da ação de vários agentes físico-químicos provenientes do ambiente externo ou mesmo endógenos, produtos do próprio metabolismo celular. Descreveremos, a seguir, a série de mecanismos, muitas vezes redundantes, que as células possuem e que reparam essas lesões. Na ausência desse processo de reparo, as lesões podem afetar o processo de replicação e transcrição, podendo levar à morte celular ou, então, fixar-se alterando a sequência do DNA original, ou seja, resultando em mutações. Se, por um lado, as mutações são em geral deletérias, elas geram diversidade genética. Logo, do delicado balanço entre correção de lesões e a formação das mutações resulta a manutenção da vida e o processo de evolução dos seres vivos.

5.2. Mutagênese

Mutações são mudanças permanentes que ocorrem nos genes. Os genes presentes nas células são compostos por polímeros de desoxirribonucleotídeos de timina, adenina, citosina e guanina e as mutações são alterações nas sequências de bases nitrogenadas dos genes. Mutações originam variações do mesmo gene, os alelos. As diferentes formas alélicas estão sujeitas à seleção natural, constituindo-se, portanto, no substrato para a evolução adaptativa. O alelo de um dado gene que confere maior vantagem adaptativa ao indivíduo sofre pressão seletiva positiva (Griffiths *et al.*, 1999). Mesmo que não haja diferenças em termos adaptativos, os alelos novos podem aumentar sua frequência através do processo conhecido como *deriva genética* (*genetic drift*, em inglês). Novos alelos podem ser introduzidos em uma população através de migração, mas sempre são originados em uma espécie através de mutação, que pode ser considerada como a única fonte de variabilidade genética.

As mutações podem ser classificadas como induzidas ou espontâneas (Friedberg *et al*., 2006). As mutações induzidas são produzidas por diferentes agentes físico-químicos, denominados mutagênicos. Por outro lado, as mutações ditas espontâneas podem resultar de erros independentes da ação de agentes físico-químicos ocorridos durante a replicação do DNA. Também podem ser consideradas como espontâneas as mutações que são causadas por agentes mutagênicos presentes no ambiente celular, incluindo aqueles provenientes do próprio metabolismo da célula. Operacionalmente, é impossível mostrar se uma dada mutação foi causada por um erro ocorrido durante a replicação ou se foi induzida por um agente físico-químico. Portanto, há dois processos necessários para que ocorra uma mutação: (a) erros de

incorporação dos nucleotídeos durante a replicação do DNA; e (b) lesão em um nucleotídeo que, após um ciclo replicativo, não foi reparada. Analisemos inicialmente o primeiro mecanismo.

A replicação é um processo extremamente eficiente. Em termos quantitativos, somente um a cada 10^9 a 10^{10} nucleotídeos incorporados pela polimerase do DNA ocorre de maneira incorreta (o significado do termo "incorporação incorreta", utilizado neste texto, refere-se a incorporações de nucleotídeos que não obedecem ao emparelhamento proposto por Watson e Crick). Na Figura 5.1, é esquematizada uma incorporação incorreta, ou seja, um nucleotídeo de citosina foi incorporado, quando o correto seria um de adenina.

Essa grande eficiência do processo replicativo pode ser explicada por três motivos. Em primeiro lugar, a subunidade catalítica da polimerase do DNA (representada na Figura 5.1 por um círculo cinza claro) é relativamente eficiente por si só, mas não é tão precisa quanto à replicação como um todo. A subunidade catalítica da polimerase do DNA comete um erro a cada 10^4 a 10^5 nucleotídeos adicionados. Além da subunidade catalítica, a polimerase do DNA possui uma subunidade revisora capaz de clivar ligações fosfodiéster no sentido 3'→ 5' (Figura 5.1, círculo cinza escuro). Quando um nucleotídeo é incorporado incorretamente (Figura 5.1), ocorre uma pequena distorção na estrutura do DNA, reconhecida pela subunidade catalítica que se move no sentido contrário à polimerização. A subunidade revisora cliva, então, o nucleotídeo incorporado incorretamente, dando "uma nova chance" para a polimerase do DNA. Em sistemas *in vitro*, foi possível estimar que a polimerase do DNA incorpora um nucleotídeo errado a cada 10^7 a 10^8, o que é um índice ainda inferior àquele observado a cada ciclo replicativo.

A polimerase do DNA pode incorporar erradamente nucleotídeos devido ao fato de as bases nitrogenadas poderem assumir diferentes configurações, algumas das quais poderem alterar o emparelhamento normal das bases. Essas diferentes formas das bases nitrogenadas são denominadas tautômeros, um tipo de isômero. As bases nitrogenadas descritas pelo modelo da dupla hélice do DNA de Watson e Crick estão na forma enólica. Quando as bases nitrogenadas estão na forma imino, elas tendem a emparelhar-se com bases diferentes, ocasionando uma incorporação errada pela polimerase do DNA. Esses tautômeros ocorrem em um equilíbrio dinâmico, em que as bases nitrogenadas permanecem a maior parte do tempo na forma enólica (em uma razão 10.000 para 1, o que explica a taxa de erro de 1 para 10^4).

Após cada ciclo replicativo, existe um período de tempo em que cada nucleotídeo incorporado incorretamente pode ser removido e substituído pelo correto por um sistema denominado reparo por emparelhamento errado de bases (*mismatch repair*, em inglês). Em *Escherichia coli*, esse sistema de reparo pode atuar graças à ação de uma metilase de DNA, que catalisa a metilação de adenina presente em determinadas sequências. A metilação de uma adenina permite à célula distinguir a cadeia molde (metilada) da cadeia recém-sintetizada e, dessa maneira, remover o nucleotídeo incorporado incorretamente da cadeia nova, e não da cadeia molde. Na Figura 5.2A, a cadeia molde é a inferior, pois está metilada na sequência palindrômica e a incorporação errada deu-se pela adição de uma guanina emparelhada com uma timina (emparelhamento descrito com fontes cinza claro).

Ao analisarmos a molécula de DNA da Figura 5.2A, poderíamos pensar que um nucleotídeo de timina foi incorporado erradamente. Sabemos que esse não é o caso, porque a cadeia molde é a inferior, já que está metilada. Uma série de enzimas desse sistema de reparo remove uma parte da cadeia recém-sintetizada (Figura 5.2B) permitindo que uma outra polimerase do DNA incorpore um oligonucleotídeo correto (Figura 5.2C). Após alguns instantes, a cadeia recém-sintetizada também é metilada e a célula não tem mais condições de distinguir a cadeia molde da cadeia recém-sintetizada. Portanto, o sistema de reparo por emparelhamento errado de bases tem alguns instantes após o término da replicação para atuar. Se o sistema de reparo não funcionar nesse curto espaço de tempo, a fita de DNA recém-sintetizada (representada pela cadeia superior na Figura 5.2A) será metilada resultando em mutação. Como dito anteriormente, esses erros ocorrem em uma frequência muito baixa e são, em parte, responsáveis pelas mutações espontâneas (Modrich, 1991).

Outro mecanismo que pode provocar uma mutação é o que envolve a formação de lesões no DNA. O esquema da Figura 5.3 serve como um exemplo para abordar vários conceitos comuns à mutagênese induzida por lesões, embora variações possam ocorrer. No exemplo da Figura 5.3, está representada uma lesão na base nitrogenada guanina, assinalada em destaque e com um asterisco. Nesse caso, a lesão na guanina faz com que essa base tenda a emparelhar-se com a timina, e não com a citosina, como ocorre com a guanina não modificada. Portanto, após o término de um ciclo replicativo, se a referida lesão não for reparada, um nucleotídeo de timina será incorporado à nova cadeia, em oposição

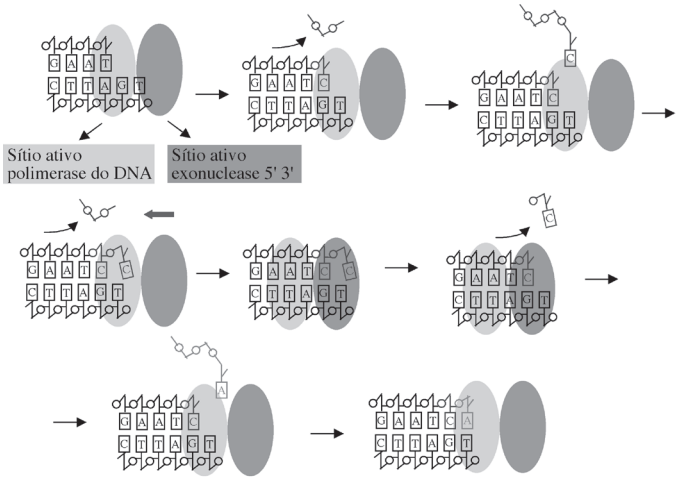


Figura 5.1. Atividades enzimáticas da polimerase III do DNA de *Escherichia coli*.

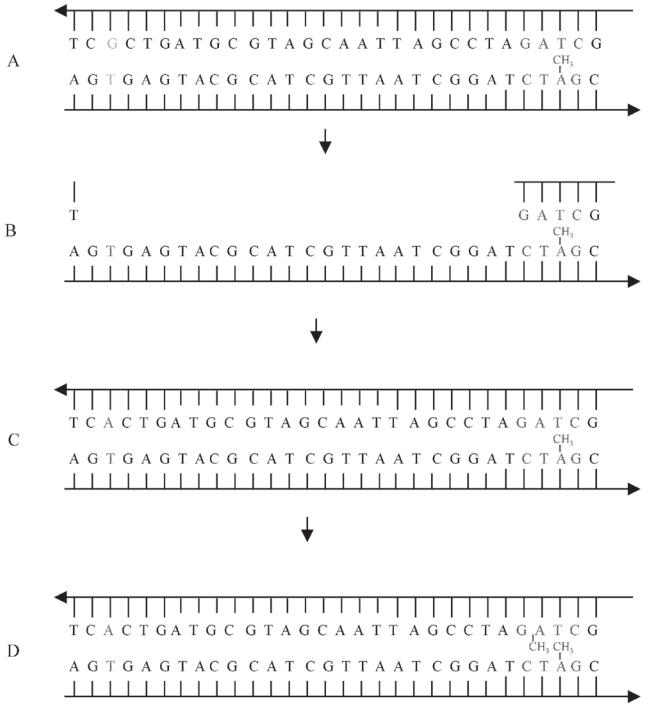


Figura 5.2. Reparo de emparelhamento errado.

à guanina lesionada (Figura 5.3), constituindo uma alteração na informação genética e, consequentemente, uma mutação. Após o término do segundo ciclo replicativo (Figura 5.3), é possível observar que o par GC foi alterado para um par AT. É importante observar que uma lesão, por si só, não se constitui em uma mutação. A lesão só induzirá uma mutação se, após um ciclo replicativo, ela não for reparada. Outro ponto que deve ser ressaltado é que nem toda lesão é mutagênica. Por exemplo: se uma lesão na guanina gerar um produto que ainda tende a emparelhar-se com a citosina, então a referida lesão não será mutagênica. Voltaremos a tratar esse assunto após analisar as lesões induzidas por diferentes agentes mutagênicos (Friedberg *et al*., 2006; Griffiths *et al.*, 1999).

Na mutação esquematizada na Figura 5.3, uma guanina, que é uma purina, foi substituída por uma timina—uma pirimidina. Substituições desse tipo (ou seja, de um derivado de purina por um derivado de pirimidina ou vice-versa) são denominadas transversões, enquanto substituições de uma purina por uma purina diferente ou de uma pirimidina por uma pirimidina diferente são denominadas transições (Tabela 5.1).

Mutações pontuais (transições ou transversões) podem alterar a sequência de nucleotídeos de um gene sem alterar a sequência de aminoácidos de uma proteína. Esse fato pode ser explicado pela degeneração do código genético. Isso significa que mais de um códon pode codificar para um mesmo aminoácido. Por exemplo, os códons AGA, AGG, CGG, CGA, CGU e CGC codificam para arginina no código genético padrão. Dessa forma, a ocorrência de uma única mutação em um gene com o códon AGG envolvendo a substituição de adenina por citidina não ocasionará a alteração na sequência de aminoácidos da proteína codificada pelo referido gene. Como a maioria dos casos de degeneração envolve a terceira posição do códon, na maior parte das vezes, a substituição de nucleotídeos nessa posição não leva à alteração dos aminoácidos da proteína. Uma substituição de nucleotídeo que não provoca uma alteração na sequência de aminoácidos da proteína correspondente é denominada "mutação silenciosa". Em contraposição às mutações silenciosas, mutações que alteram a sequência de aminoácidos de uma proteína são denominadas "mutações de sentido alterado" (*missense mutations*, em inglês; Griffiths *et al.*, 1999).

Os nucleotídeos presentes em posições de degeneração têm uma pressão seletiva relativamente baixa, dado que algumas substituições de bases não alteram a proteína expressa pelo referido gene. Dessa forma, quando comparamos as sequências de genes de diferentes organismos, existe uma grande variação nas posições de degeneração (ver Capítulo 7).

Algumas substituições de nucleotídeos podem provocar o aparecimento prematuro de um códon de terminação (códons "stop", UAA, UAG e UGA, no código genético padrão) em alguns genes. Esse tipo de mutação é chamada de "mutação sem sentido" (*nonsense mutation*, em inglês). O códon CAG, por exemplo, de um gene codifica para glutamina. A substituição do primeiro nucleotídeo por um de timidina gera um códon UAG, que codifica para o término da síntese protéica. Dessa forma, a proteína gerada por essa mutação será truncada e menor do que a original: todos os aminoácidos codificados por códons posteriores à mutação não serão incorporados à proteína mutante. Em geral, esse tipo de mutação gera proteínas com baixa ou nenhuma atividade.

Até aqui, analisamos diferentes tipos de substituições de bases, chamadas de "mutações pontuais". Outros tipos de mutações são as inserções e as deleções, que envolvem a adição ou a remoção de nucleotídeos em uma determinada sequência de DNA. As inserções ou deleções que não envolvem múltiplos de três nucleotídeos provocam uma alteração no quadro de leitura de um gene. Um exemplo é a inserção de quatro nucleotídeos (sublinhados) no gene cujo mRNA está representado abaixo:

mRNA original:	CGU	AUA	UCC	UAU	GCC	CCU	GAC
Proteína original:	Arg	Ile	Ser	Tyr	Gly	Pro	Asp
mRNA mutante:	CGU	AUA	UCU	AUC	CUA	UGC	CCC UGA C
Proteína mutante:	Arg	Ile	Ser	Ile	Leu	Cy	Pro PARADA

Essa inserção levou à alteração do quadro aberto de leitura (*open reading frame*, em inglês), tendo como uma das consequ-

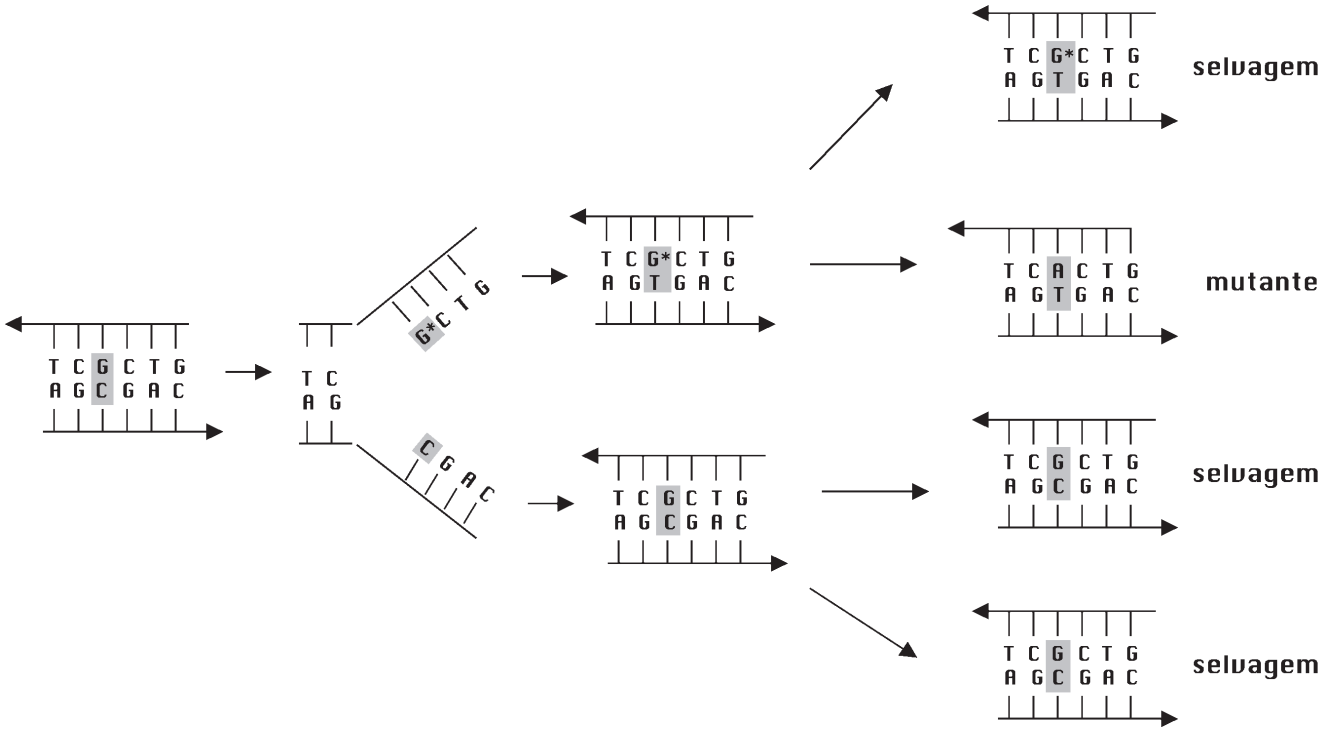


Figura 5.3. Mecanismo de mutagênese.

Tabela 5.1. Mutações pontuais.

Tipo de Mutação	Exemplos
Transversão	A·T → C·G
	A·T → T·A
	G·C → T·A
	G·C → C·G
	C·G → A·T
	C·G → G·C
	T·A → G·C
	T·A → A·T
Transição	A·T → G·C
	G·C → A·T
	C·G → T·A
	T·A → C·G

ências o surgimento de um códon de parada onde originalmente estava um códon que sinalizava a inserção de aspartato.

As deleções e inserções de nucleotídeos em números que não são múltiplos de três, em regiões codificadoras para polipeptídeos, levam a uma alteração de toda a proteína após a mutação, pois alteram o quadro aberto de leitura de um gene e podem provocar o aparecimento prematuro de um códon de terminação. Por esses motivos, inserções e deleções também são tipos de mutações que têm maiores probabilidades de resultar em proteínas com baixa ou nenhuma atividade.

No caso de organismos multicelulares, mutações podem ocorrer em células somáticas ou em células germinativas (células que sofrerão meiose originando os gametas do indivíduo). No primeiro caso, os efeitos da mutação serão sentidos somente pelo indivíduo e não por sua progênie. Como as células somáticas se dividem por mitose, todas as células originadas da célula mutada apresentarão o mesmo fenótipo, constituindo um *clone*. Como veremos adiante, mutações somáticas em alguns genes podem originar tumores.

Uma mutação em células da linhagem germinativa produz *gametas mutantes*. Se o gameta mutante participa da fertilização, então a mutação passará para todas as células do descendente. O efeito da mutação em células da linhagem germinativa não é, portanto, sentida pelos pais, mas pode afetar todas as células dos filhos. As mutações nas células somáticas, por sua vez, podem afetar os pais, mas não interferem na geração seguinte.

5.3. Reparo de Lesões na Molécula de DNA

As células dos seres vivos estão constantemente ameaçadas por uma imensa variedade de agentes físico-químicos e biológicos que podem provocar alterações permanentes nos genes, se não forem reparadas antes do término da replicação. Vamos agora analisar algumas dessas lesões para exemplificar seus mecanismos celulares de reparo (Friedberg *et al*., 2006; Sancar e Sancar, 1988).

5.3.1. Reparo por excisão de bases

Uma das lesões que ocorre espontaneamente nas células é a quebra da ligação que une as bases purínicas à desoxirribose (ligação N-glicosídica). Esse processo chama-se despurinação e gera o aparecimento de sítios apurínicos (sítios AP) no DNA (Figura 5.4A). Como exemplo, em uma célula humana mantida a temperatura de 37 ºC, há o aparecimento de cerca de 10.000 sítios AP no genoma a cada 24 horas (Lindahl e Nyberg, 1972). O aparecimento de sítios AP também pode ser induzido por agentes mutagênicos, como a aflatoxina B_1 (metabólito do fungo *Aspergillus flavus*, que infecta amendoim) que, ao se adicionar na posição 7 da guanina, fragiliza a respectiva ligação N-glicosídica. Sítios AP também podem ser gerados pela ação de um grupo de enzimas denominadas glicosilases do DNA, assim chamadas pelo fato de catalisarem a clivagem da ligação que une bases nitrogenadas à desoxirribose. Existem muitos tipos diferentes de glicosilases do DNA, com diferentes especificidades, como as glicosilases da N^1-metilguanina, da N^3-metilguanina e da N^7-metilguanina (Figura 5.5). Essas três enzimas são capazes de reconhecer diferenças, como metilações em posições diferentes da guanina.

Outro exemplo desse tipo de enzima é a glicosilase da uracila. Essa enzima reconhece uracilas formadas no DNA pela desaminação de citosina. A desaminação de bases nitrogenadas pode ocorrer espontaneamente, mas, como no caso da despurinação, pode ser acelerada por agentes mutagênicos, como as nitrosaminas, compostos encontrados na fumaça de cigarros.

Os sítios AP, gerados espontaneamente pela ação de agentes mutagênicos ou pela ação de glicosilases do DNA, são reconhecidos por outros tipos de enzimas (conhecidas como endonucleases de sítio AP) que clivam a ligação fosfodiéster adjacente ao sítio AP (Figura 5.4A, seta cinza). As enzimas glicosilase do DNA e endonuclease AP fazem parte de um sistema denominado reparo "por excisão de bases" (Figura 5.6). O primeiro passo desse sistema de reparo envolve o reconhecimento da lesão pela glicosilase do DNA, a qual cliva a ligação N-glicosídica, que une a base lesionada ao açúcar. O sítio AP gerado é então reconhecido pela endonuclease AP, a qual cliva uma ligação fosfodiéster. O resíduo 2'-desoxirribose-5-fosfato desprovido da base nitrogenada é removido pela enzima fosfodiesterase da 5'-desoxirribose (dRPase). O término do reparo é executado pelas enzimas polimerase do DNA e ligase do DNA (Figura 5.6). Em alguns casos, como no da enzima endonuclease III de *Escherichia coli*, uma única proteína pode ter as atividade de glicosilase do DNA e endonuclease AP. O fato de existir um repertório muito grande de diferentes tipos de glicosilases do DNA, cada uma específica para uma determinada lesão, indica que as células gastam muita energia para executar o reparo por excisão de bases. Esse é um dos casos que mostram o grande dispêndio energético realizado pelas células na manutenção da informação genética, representada pela sequência de nucleotídeos dos genes.

Outro exemplo notável desse enorme gasto de energia para manter a informação genética estável é o reparo da lesão 8-oxo-

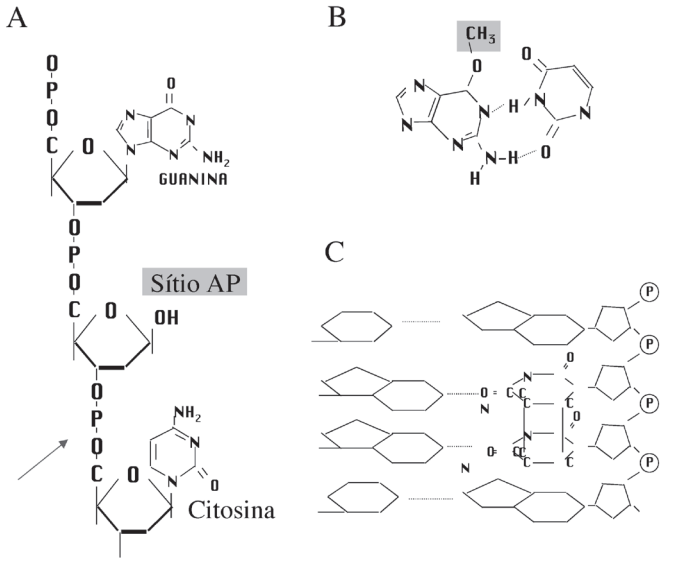


Figura 5.4. Exemplos de lesões. A seta indica o ponto em que a endonuclease AP atua.

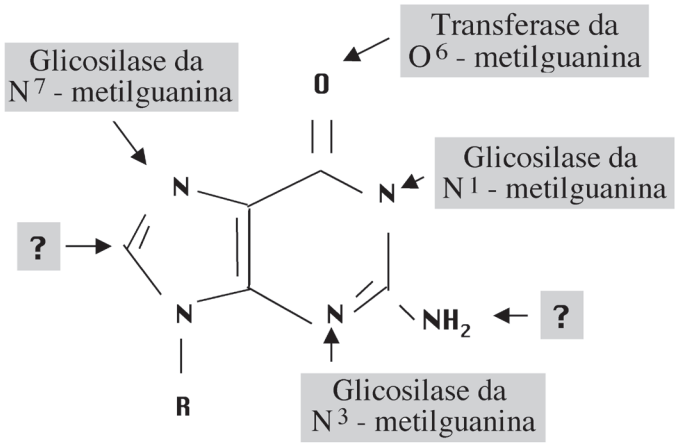


Figura 5.5. Pontos de adição de grupos metila à guanina e enzimas de reparo envolvidas.

guanina. Essa é uma das várias lesões induzidas por radicais livres de oxigênio. Essas substâncias são geradas como subprodutos secundários da cadeia respiratória mitocondrial. Um dos radicais livres de oxigênio, o radical hidroxila, é uma das substâncias mais reativas que se conhece, podendo provocar diferentes lesões em bases nitrogenadas, desoxirribose e fosfato. O ataque do radical livre hidroxila na posição 8 da guanina induz a formação do produto 8-oxoguanina, que tende a emparelhar-se com a adenina, e não com a citosina. Outros oxidantes produzidos durante o metabolismo celular, como o oxigênio singlete e o peroxinitrito, também podem induzir a formação de 8-oxoguanina. Em *Escherichia coli*, existem três vias que podem reparar essa lesão: (1) uma glicosilase que remove 8-oxoguanina; (2) uma glicosilase que remove a adenina emparelhada erradamente com 8-oxoguanina; e (3) uma enzima que degrada o nucleotídeo 8-oxoGTP, antes de ele ser incorporado a uma cadeia de DNA. Portanto, existem três vias envolvidas na eliminação do acúmulo da mesma lesão. Como mencionado anteriormente, a redundância no reparo de lesões é

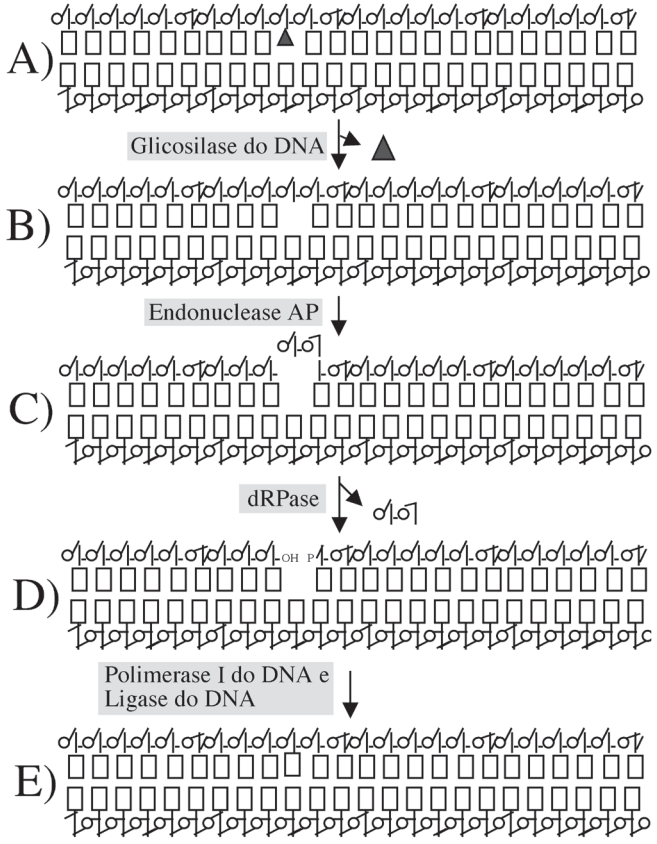


Figura 5.6. Reparo por excisão de base.

um fenômeno frequente e, mais uma vez, ilustra a enorme pressão de seleção que existe no sentido de manter a informação genética íntegra (Michaels *et al.*, 1992; Sakumi *et al.*, 1993).

Radicais livres também podem provocar lesões na desoxirribose dos nucleotídeos. Quando o radical hidroxila ataca o carbono 1' da desoxirribose, a ligação n-glicosídica enfraquece, podendo ocorrer a liberação da base nitrogenada correspondente. O sítio AP gerado pode ser então reparado pelo sistema de excisão de bases esquematizado na Figura 5.6. O radical livre hidroxila pode atacar outros carbonos da desoxirribose, como o carbono 4'. Nesse caso, há uma quebra da ligação fosfodiéster, provocando quebras simples no esqueleto do DNA (Demple e Harrison, 1994).

5.3.2. Reparo direto

Agentes alquilantes, como nitrosaminas, derivados de hidrazina e nitrosoguanidinas, aumentam a taxa de mutação ao induzirem a adição de grupos alquila (por exemplo, metila e etila) em bases nitrogenadas. Em geral, a guanina é a base mais susceptível à alquilação e, no caso da metilação, a adição pode dar-se em diferentes posições, originando diferentes produtos (Figura 5.5). A metilação no oxigênio da posição 6 da guanina é extremamente mutagênica, pois o produto O^6-metilguanina tende a emparelhar-se com timina (Figura 5.4B). O reparo dessa lesão não envolve a ação de uma glicosilase, como no caso da metilação nas posições 1, 3 e 7 da guanina. Nesse caso, uma proteína denominada transferase da O^6-metilguanina remove o grupo metila adicionado ao oxigênio da posição 6, restaurando a guanina não modificada. O reparo de O^6-metilguanina também demonstra o enorme gasto metabólico envolvido na manutenção da informação genética: para cada metilguanina reparada, uma transferase da O^6-metilguanina é gasta. O grupo metila removido pela transferase da O^6-metilguanina é transferido para uma cisteína da própria enzima e esta torna-se inativa. Como o reparo de O^6-metilguanina não envolve a remoção de qualquer componente dos nucleotídeos que fazem parte de uma molécula de DNA (base, 2'-desoxirribose e fosfato), esse sistema é denominado direto (Singer, 1975).

Outros produtos resultantes da metilação da guanina na posição 8 também foram descritos (Figura 5.5; Augusto *et al.*, 1992; Netto *et al.*, 1992). Recentemente, foi sugerido que a enzima AlkA (N-glicosilase II da 3-metiladenina) de *Escherichia coli* pode atuar sobre essa lesão, mas como a eficiência de reparo é muito menor do que o reparo de 3-metiladenina, esse é um

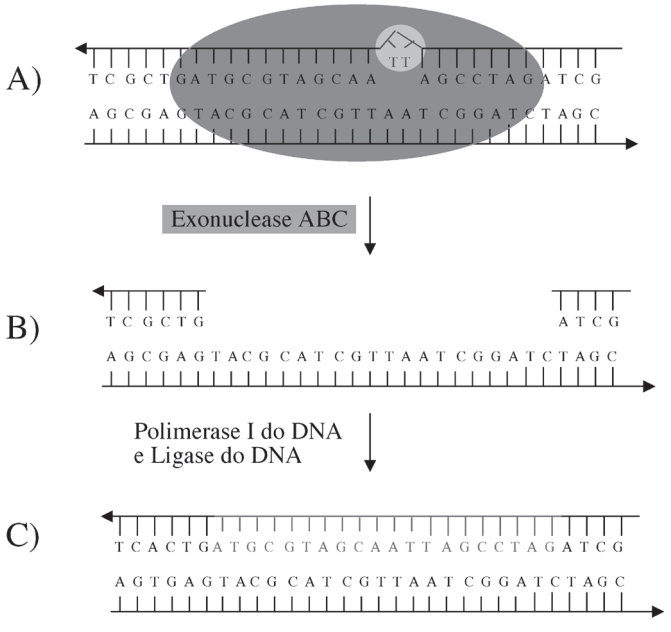


Figura 5.7. Reparo por excisão de nucleotídeo.

ponto ainda a ser melhor investigado (Gasparutto *et al*., 2002). De qualquer forma, é interessante ressaltar que a posição 8 da guanina parece ser particularmente suscetível ao ataque de radicais livres, como o radical hidroxila (gerando 8-oxoguanina), o radical metila (gerando 8-metilguanina) e o radical nitrosila (gerando 8-nitrosoguanina).

Recentemente foi descoberto um segundo tipo de reparo direto de bases metiladas (Trewick *et al*., 2002; Falnes *et al*., 2002). Esse sistema envolve uma única proteína (codificada pelo gene *alkB* em bactérias), cuja função se estudava, sem sucesso, há mais de 30 anos. Através do emprego de ferramentas da bioinformática, verificou-se que AlkB pertence a uma família de proteínas dependentes de alfa-cetoglutarato e ferro (II). Assim, com a adição desses compostos na reação de reparo *in vitro* foi demonstrado que essa enzima catalisa a eliminação do grupo metila das bases lesadas 1-metil-adenina e 3-metil-citosina, revertendo a adenina e citosina, respectivamente. Essa reação é dependente de alfa-cetoglutarato, ferro (II) e oxigênio, sendo que a catálise oxida o grupo metila a formaldeído, revertendo a lesão para a base não modificada. Como não há quebras de ligações químicas durante esse processo, este é mais um tipo de reparo direto. É importante ressaltar que o reparo oxidativo é muito conservado (há proteínas homólogas a AlkB em células humanas) e atua em moléculas de DNA e também de RNA contendo bases metiladas, em um primeiro caso conhecido de reparo de RNA (Falnes *et al*., 2007). Considerando que o RNA foi a molécula biológica primordial (ver Capítulo 2), esses dados podem ter importantes implicações evolutivas.

A luz solar possui componentes de luz dentro do espectro de comprimento de onda da radiação ultravioleta (UV), sendo que luz UVB (280-320 nm) e UVA (320 a 400 nm) atingem a superfície da Terra e podem causar lesões no DNA. As lesões induzidas por luz UV mais frequentemente citadas são os dímeros de pirimidina (Figura 5.4C), que são o produto de ligações covalentes entre duas bases pirimídinicas adjacentes no DNA. Os dímeros de pirimidina podem ser formados por ligações entre os carbonos 5 e 6 dos dois anéis pirimidínicos (Figura 5.4C) ou entre o carbono 4 de uma pirimidina e o 6 da pirimidina adjacente. Tanto citosina como timina podem fazer parte dos possíveis dímeros de pirimidina. O reparo dos dímeros de pirimidina pode envolver a ação da enzima fotoliase, que não remove componentes dos nucleotídeos (bases nitrogenadas, desoxirribose ou fosfato), também sendo considerado, por isso, um tipo de reparo direto. A fotoliase reconhece o dímero de timina e restaura os dois monômeros de pirimidina, utilizando energia obtida de luz visível. A luz visível ativa cofatores (flavinas) da fotoliase para um estado excitado (um radical livre), que atua, então, sobre o dímero, revertendo às bases originais. Esse tipo de reparo (também conhecido como fotorreativação) é encontrado em bactérias, eucariontes (incluindo mamíferos marsupiais) e mesmo alguns vírus, sendo que as enzimas fotoliases são em geral muito conservadas. Curiosamente, no entanto, por motivos desconhecidos, o gene correspondente à fotoliase foi perdido durante a evolução de mamíferos placentários, de modo que esse tipo de reparo está ausente em células humanas (Menck, 2002).

5.3.3. Reparo por excisão de nucleotídeo

Além do sistema catalisado pela fotoliase, existe outro sistema de reparo que atua sobre lesões formadas por UV. Nesse sistema uma série de enzimas reconhece lesões volumosas que causam distorções na dupla-hélice do DNA, como os dímeros de pirimidina, e também produtos da adição de agentes mutagênicos volumosos, como aflatoxina B_1. Como resultado desse reconhecimento inicial, endonucleases clivam em duas ligações fosfodiéster, próximas à lesão (Figura 5.6A-B), sendo que algumas enzimas conhecidas como helicases removem o oligonucleotídeo contendo a lesão, gerando uma lacuna. A lacuna originada é então preenchida pela ação das enzimas polimerase do DNA e ligase do DNA, concluindo o reparo (Figura 5.6C). Esse sistema é denominado reparo por excisão de nucleotídeo e é encontrado em todos domínios da vida (Arquéias, Bactérias e Eucariontes). Apesar da maquinaria encontrada nesses domínios ser diferente, o mecanismo de ação é muito similar, o que demonstra a alta eficiência e importância desse tipo de reparo de DNA (Costa *et al*., 2003). Em seres humanos, mutações que afetem o reparo por excisão de nucleotídeos podem resultar em síndromes bastante graves, levando a um aumento na suscetibilidade à luz solar, com alta frequência na formação de tumores, problemas de desenvolvimento e mesmo envelhecimento (ver abaixo). A ação da fotoliase e do reparo por excisão de nucleotídeos sobre um mesmo tipo de lesão é mais um caso da redundância nos sistemas que mantêm a integridade da informação genética.

5.3.4. Reparo recombinacional

A importância da recombinação do DNA tem sido relacionada com a geração de diversidade (através do processo de *crossing-over* durante a meiose) e com a segregação de cromossomos homólogos (pela formação de quiasmas também durante a meiose). Todavia, o envolvimento da recombinação do DNA no reparo de lesões parece ser o seu papel mais importante na célula. Mecanismos recombinatórios em princípio podem atuar em qualquer lesão, desde que exista uma cópia intacta da região afetada na mesma célula. Mas o sistema de reparo recombinacional parece ser ativado quando o aparato de replicação não é capaz de prosseguir na síntese de DNA. Isso ocorre quando a polimerase do DNA encontra uma lesão na fita molde, como um dímero de pirimidina, uma quebra de fita dupla ou uma ligação cruzada entre as duas fitas do DNA. A polimerase do DNA não adiciona nucleotídeos na região oposta à lesão, deixando um espaço em branco na cadeia recém-sintetizada (Figura 5.8A).

Essa descontinuidade é preenchida com DNA parental por um processo de recombinação (Figura 5.8B), originando uma descontinuidade no DNA parental, o qual é então reparado por polimerases do DNA e ligase do DNA (Figura 5.8C). Em *Escherichia coli*, uma das proteínas que faz parte desse sistema é a RecA (denominada assim a partir de um mutante que promove processos de recombinação homóloga). Em células humanas, especial atenção tem sido dedicada recentemente a estudos de sistemas de reparo recombinacionais. Basicamente, conhecem-se sistemas de recombinação de DNAs homólogos, que empregam

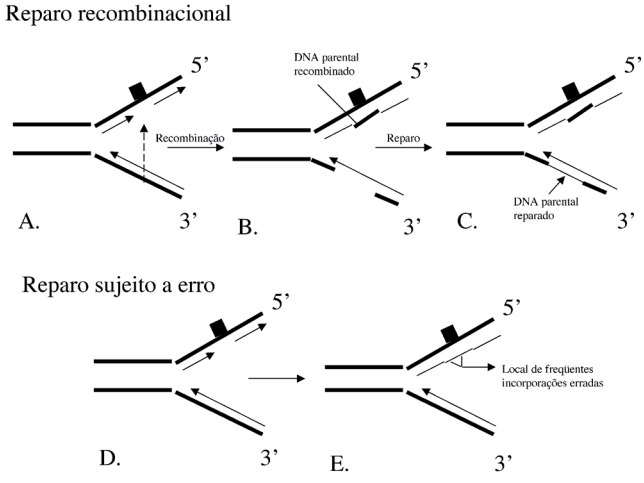


Figura 5.8. Reparo de lesões que induzem espaços em branco nas cadeias de DNA.

cromossomos homólogos para reparo, mas também existem sistemas de reparo de extremidades de DNA não homólogas (NHEJ, "non-homologous end joining"). Esses sistemas de reparo são bastante complexos e problemas genéticos que alteram seu funcionamento têm sido diretamente relacionados à alta frequência de tumores e mesmo ao envelhecimento (Shrivastav *et al*., 2008). Um exemplo de grande impacto foi a descoberta de que parte das famílias nas quais as mulheres apresentam tumor de mama com alta frequência e com menos de 40 anos apresentam mutações nos genes BRCA1 ou BRCA2 (*breast cancer*, em inglês), que estão envolvidos em reparo recombinacional homólogo(Zhang e Powell, 2005).

5.3.5. Reparo sujeito a erro

Quando as células são expostas a condições de estresse elevado, como exposição a radiações ou altas doses de um agente mutagênico, um sistema peculiar atua: o reparo sujeito a erro. A função desse sistema é impedir que a replicação seja interrompida pelo bloqueio que várias lesões localizadas na cadeia molde exercem à atividade catalítica da polimerase do DNA. Se isso ocorrer, há produção de cadeias com descontinuidades correspondentes a vários nucleotídeos e, se essas não forem preenchidas a tempo, a célula morre. Dessa forma, a alta precisão da replicação é deixada de lado e as enzimas envolvidas no sistema de reparo sujeito a erro adicionam, com baixa especificidade, nucleotídeos nas lacunas deixadas (Figura 5.8E). Ao contrário dos sistemas de reparo analisados anteriormente, nesse caso, as células que sobrevivem à situação de estresse apresentam uma alta taxa de mutação. Esse sistema está bem descrito em bactérias e faz parte das respostas SOS. Em situações de estresse elevado, bactérias induzem vários genes que garantem sua sobrevivência, mesmo com aumento da mutagênese. Entre os genes induzidos, destacam-se algumas polimerases do DNA, que têm capacidade de replicar o DNA na região da lesão, mesmo que isso acarrete em aumento de erro na leitura, o que resulta em mutações (Galhardo *et al*., 2005). Vários outros genes fazem parte do que é conhecido de regulon SOS, incluindo alguns envolvidos no reparo por excisão de nucleotídeos e na recombinação de DNA (RecA) (Rocha *et al*., 2008).

A presença de um sistema homólogo a esse em eucariontes ainda não foi demonstrada, no entanto se sabe que várias respostas são induzidas devido ao estresse causado por lesões no genoma humano. Além disso, foram descobertas várias polimerases do DNA com capacidade para replicar o DNA lesado, que são conhecidas genericamente como polimerases para síntese translesão, mas com fidelidade de replicação reduzida, sendo, portanto, sujeitas a erro (McCulloch e Kunkel, 2008).

5.4. Doenças Genéticas Humanas Relacionadas ao Reparo de DNA

A importância dos sistemas de reparo na manutenção da integridade da informação genética é evidenciada de forma dramática por diversas doenças humanas genéticas, com deficiências que afetam algum mecanismo de reparo de DNA celular. É o caso da síndrome conhecida como a xeroderma pigmentosum (XP). Indivíduos com essa síndrome apresentam naturalmente uma alta frequência de tumores de pele nas regiões do corpo expostas à luz solar. Em geral, a doença é percebida pela mãe da criança afetada em um primeiro passeio ao ar livre, mesmo se em dia nublado: depois de alguns minutos de exposição à claridade do dia, o bebê começa a chorar e a pele exposta "queima", como se houvesse passado horas. O paliativo para essas crianças é evitar completamente a exposição à luz do dia e, por isso, elas normalmente invertem o dia pela noite, sendo portanto conhecidas como "crianças da Lua". O defeito molecular em células de pacientes XP é a deficiência no sistema de reparo por excisão de nucleotídeos, o que resulta em uma incapacidade na remoção de lesões de DNA promovidas pela luz solar. Como resultado, as células da pele sofrem mais mutações, originando tumores. Alguns pacientes afetados também apresentam problemas no desenvolvimento, deficiência mental e mesmo envelhecimento precoce, o que também deve estar relacionado com o defeito em reparo de lesões no DNA. Foram identificados sete genes diferentes que podem estar afetados no defeito de reparo de DNA de pacientes XP, nomeados XPA a XPG. Por outro lado, alguns pacientes XP apresentam o reparo por excisão de nucleotídeos normal, mas apresentam defeito em uma polimerase do DNA que é capaz de sintetizar DNA através de lesões induzidas por luz UV. Essa enzima é conhecida como polimerase do DNA eta, codificada pelo gene XPV (XP variante). Além da síndrome XP, pelo menos duas outras síndromes que afetam o desenvolvimento do indivíduo estão diretamente ligadas a processo de reparo excisão de nucleotídeos: síndrome de Cockayne (CS, envolvendo os genes CSA e CSB) e tritiodistrofia (TTD, envolvendo o gene TTDA) (Andressoo *et al*., 2006). Curiosamente, a análise dos genes deficientes nessas síndromes revelou que alguns genes XP também podem estar envolvidos em mais de uma síndrome (por exemplo, genes XPB e XPD estão envolvidos com CS, XP e TTD). Essas observações feitas em pacientes com essas síndromes revelam a importância dos sistemas de reparo de DNA na manutenção da integridade do genoma celular, fundamental para garantir níveis de mutagênese suficientemente baixos para a sobrevivência da espécie.

Várias outras doenças genéticas têm sido associadas a deficiências de reparo de DNA, entre elas, a síndrome de Bloom (BS), anemia Fanconi (AF), tricotiodistrofia (TTD), síndrome de Cockayne (CS) e ataxia telagienctasia (AT). Devido à instabilidade genética provocada por deficiência em diferentes sistemas de reparo de DNA, observa-se uma alta frequência de aberrações cromossômicas em células de pacientes portadores da doença em várias dessas síndromes (BS, AF, CS, AT), acompanhada em algumas delas de alta frequência de tumores (XP, AT, AF, BS).

Algumas formas hereditárias de câncer têm sido associadas à deficiência em sistemas de reparo de DNA. É o caso de câncer de cólon não poliposo hereditário, identificado como sendo o resultado de deficiência em reparo de emparelhamento errado de bases (*Mismatch repair*, Figura 5.2). Além disso, dois genes cuja deficiência é responsável por câncer de mama de origem familiar foram clonados—BRCA1 e BRCA2 (Ingvarsson, 1999; Monteiro e Birge, 2000), que participam do processo de reparo recombinacional. A proteína supressora de tumor p53, cuja inativação está associada em cerca de 50% dos tumores humanos, também tem sido relacionada com reparo de DNA lesado (Albrechtsen *et al.*, 1999), além de ser responsável por vários processos de sinalização em resposta ao dano no DNA (Meulmeester e Jochemsen, 2008). Por sua importância e papel central no metabolismo e no ciclo celular, a proteína p53 já foi chamada de "guardiã do genoma" (Efeyan e Serrano, 2007).

5.5. Testes de Mutagênese

O fato dessas doenças associadas a defeitos no reparo a lesões no DNA estarem associadas ao aumento da probabilidade do paciente de sofrer um tumor indica que existe uma correlação forte entre a mutagênese e a carcinogênese. De fato, há muito tempo se sabe que agentes mutagênicos, como radiações ionizantes, luz ultravioleta, nitrosaminas e aflatoxina B_1, entre outros agentes

físico-químicos, também são carcinogênicos. Vários testes foram desenvolvidos para verificar se um determinado agente é carcinogênico, sendo que os testes clássicos são realizados em roedores. Esses estudos envolvem injetar ou alimentar roedores com as substâncias a serem analisadas e, depois, verificar a formação de tumores em uma população de indivíduos. Esses testes em geral são muito caros, pois envolvem a criação de um grande número de roedores. Além disso, esses testes têm baixa sensibilidade.

Pensando na correlação entre mutagenicidade e carcinogenicidade, Ames *et al.* (1973) desenvolveram um teste para medir mutagenicidade de agentes físico-químicos utilizando linhagens da bactéria *Salmonella typhimurium* (filogeneticamente próxima a *Escherichia coli*) que carregam mutações auxotróficas de vários tipos no gene responsável pela síntese de histidina (Umbuzeiro e Vargas, 2003). Esse teste é ainda amplamente empregado, por sua praticidade, e é conhecido como Teste de Ames. Assim, essas bactérias são incapazes de crescer em meios que não contenham esse aminoácido (auxotrofia). A reversão desses mutantes para o estado de prototrofia pode ser seguida facilmente por plaqueamentos em meios que não contenham histidina (Figura 5.9). Nesses meios, crescem somente bactérias que espontaneamente reverteram para o estado prototrófico (Figura 5.9). Esse caso aplica-se às bactérias que sofrem mutações espontâneas e serve como controle para as bactérias tratadas com substâncias a serem testadas. Paralelamente, estuda-se a indução à reversão por diferentes substâncias as quais se está testando sua atividade mutagênica. No exemplo da Figura 5.9, como a substância C foi a que induziu maior número de revertentes, essa substância é considerada altamente mutagênica. Como a mutagenicidade correlaciona-se com a carcinogenicidade, a substância C é considerada carcinogênica e seu uso deve ser muito bem controlado. Por outro lado, a substância A induz praticamente o mesmo número de revertentes que as colônias não tratadas (Figura 5.9), indicando que essa substância não é mutagênica e, consequentemente, não é carcinogênica. Nesse caso, devem ser feitos testes adicionais, como aqueles realizados em camundongos (Griffiths *et al.*, 1999).

Durante um período de vários anos, pesquisadores do mundo inteiro verificaram, através desse teste, que de fato existe uma correlação maior que 90% entre mutagenicidade e carcinogenicidade. Em alguns casos, é necessário adicionar frações microssomais de fígado para ativar o metabolismo de algumas substâncias, pois as propriedades mutagênicas só aparecem após metabolização. Como no organismo essa metabolização pode ocorrer, essas substâncias também são consideradas com potencial carcinogênico.

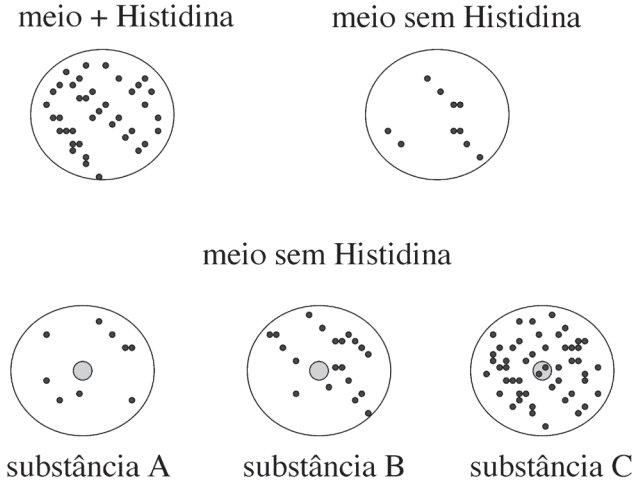


Figura 5.9. Teste de Ames. Os pequenos círculos pretos representam colônias isoladas geradas a partir de uma única célula.

A correlação entre carcinogenicidade e mutagenicidade é consistente com a teoria de que tumores são gerados por vários fatores que envolvem o acúmulo de mutações somáticas. O processo de carcinogênese envolve mutações em dois tipos de genes: proto-oncogenes e genes supressores de tumor. Em geral, proto-oncogenes e genes supressores estão envolvidos na regulação da divisão celular. Em células diferenciadas, os proto-oncogenes geralmente estão inativos. Quando sofrem determinadas mutações pontuais, os proto-oncogenes ficam ativos, sendo denominados oncogenes. Os oncogenes estimulam, então, a divisão celular em um momento em que as células deveriam permanecer em G_0, fase do ciclo celular de repouso com relação à divisão. A mutação nos genes supressores atua em sentido inverso: normalmente genes supressores estão ativos, inibindo a proliferação celular. Quando sofrem certas mutações, esses genes são inativados e deixam de inibir a proliferação de tumores. A maioria dos processos de carcinogênese envolve tanto a ativação de proto-oncogenes em oncogenes, como a inativação de genes supressores. O gene supressor p53 é um dos mais frequentemente relacionados a processos de carcinogênese em diferentes tecidos. Como já foi dito, a proteína p53 está envolvida no reparo de DNA lesado em mamíferos. Na verdade, hoje considera-se que genes envolvidos em reparo de DNA em geral são supressores de tumor, uma vez que eles atuam como uma espécie de guardiões do genoma e deficiências que afetam esses genes podem resultar na formação de tumores.

Apesar de a maior parte das mutações ser deletéria, podendo gerar, por exemplo, tumores, uma pequena parcela das mutações pode trazer vantagens seletivas a seus portadores. Mutações possibilitam o surgimento de novas formas alélicas dos genes, o que serve de matéria-prima para a seleção natural. Alelos que foram originados em determinadas situações podem não ser vantajosos para os indivíduos, mas podem ser positivamente selecionados em outros locais e/ou em outros momentos. Os seres humanos têm-se aproveitado do fato de mutações induzirem diversidade para selecionar artificialmente novas linhagens de micro-organismos, plantas e animais.

5.6. Evolução dos Sistemas de Reparo de DNA

Os estudos de reparo de DNA lesado revelaram o alto grau de conservação desses genes em espécies filogeneticamente distantes. Assim é que o produto do gene XP-B humano tem uma identidade de mais de 70% dos aminoácidos com a proteína de levedura e mesmo de plantas (Ribeiro *et al*., 1998). Estudos mais recentes confirmam que, de fato, o sistema de reparo por excisão de nucleotídeos é muito conservado entre os eucariontes, havendo várias proteínas homólogas encontradas em protozoários e humanos (Costa *et al*., 2003). No caso de genes de reparo de emparelhamento errado de bases (*mismatch repair*, em inglês), a alta conservação é ainda mais evidente, uma vez que os genes humanos (assim como os demais eucariontes, de modo geral) apresentam elevados valores de identidade em relação a genes correspondentes em eubactérias e arquéias. Isso sugere que esses genes têm ancestrais comuns anteriores à divergência desses três super-reinos, o que deve ter ocorrido há cerca de 3 bilhões de anos! A comparabilidade entre as moléculas evidencia que sua estrutura básica tem sido mantida até hoje. É, portanto, bastante provável que os sistemas de reparo de DNA existam desde os primórdios da vida na Terra, mantendo a integridade dessa molécula tão preciosa para a manutenção dos sistemas biológicos. Esses sistemas mantêm os níveis de mutagênese suficientemente baixos de modo a permitir a sobrevivência das espécies, mas sem

bloquear completamente o aparecimento de novas mutações, indispensáveis na manutenção da diversidade das espécies e para a própria evolução (Eisen e Hanawalt, 1999).

5.7. Mutagênese e Evolução

As condições ambientais podem variar de forma não previsível. Quanto maior o repertório de alelos diferentes em uma população, maior será a chance de indivíduos dessa população sobreviverem a essas variações. A sobrevivência de uma população depende de um grande repertório de alelos diferentes para vários genes, o que figurativamente equivaleria a comprar um grande número de bilhetes de loteria. A seleção natural pode ser comparada, assim, com o sorteio de uma loteria na qual um forte evento seletivo reduz a diversidade biológica (representada pelo número de alelos) porque somente uma pequena fração confere vantagem adaptativa, mas a diversidade biológica precisa continuar a ser restaurada para permitir a adaptação a novas variações ambientais que venham ocorrer no futuro (Radman, 1999).

Tem sido proposta a existência de vários mecanismos que aumentam a taxa de evolução de forma a gerar maior número de alelos (diversidade biológica). Esses processos nem sempre podem ser relacionados diretamente com o aumento da diversidade biológica (em termos de número de espécies), mas acabam levando à geração de novos alelos.

Um desses mecanismos são os "pontos quentes" mutacionais (*mutational hot spots*, em inglês), estudados desde a década de 1960 (Lewin, 1994). Em alguns genes, existem sítios que sofrem mutações muito mais frequentemente que em outros sítios. Mais recentemente, foi demonstrado que esses pontos quentes são compostos por citosinas que podem ser metiladas enzimaticamente. A metilação enzimática de citosinas está relacionada com a expressão do gene da qual essas bases fazem parte. Como descrito anteriormente (veja seção 5.3.1, no Capítulo 5), citosinas podem sofrer desaminações, gerando uracila, as quais podem ser reparadas no DNA, porque não fazem parte desse ácido nucléico. Por outro lado, a metil-citosina, quando desaminada, gera timina, uma base que faz parte do DNA e que, portanto, pode resultar em mutações do tipo C=>T. Consequentemente, mutações são geradas mais frequentemente nesses sítios.

Nas últimas décadas, foram descobertas polimerases do DNA em eucariontes, que podem ser chamadas de "mutases" (Radman, 1999). Tratam-se de polimerases que sintetizam novas moléculas de DNA, com maior taxa de erro do que a maquinaria sintética descrita acima (veja seção 5.2, no Capítulo 5). Dessa forma, ocorre um aumento na taxa de mutação. Essas polimerases do DNA atuam quando as células são submetidas a situações de estresse (por exemplo, radiação). Nesses casos, várias lesões são geradas no DNA, muitas das quais são volumosas, o que impede que a maquinaria de replicação prossiga na síntese de DNA. Essas "mutases" possibilitam que a síntese de DNA ocorra no sítio da lesão, mas, nesse caso, a incorporação de nucleotídeo é aleatória e a chance de erro é muito grande. Como descrito anteriormente (seção 5.3.5, no Capítulo 5), esse sistema de reparo sujeito a erro faz parte da resposta SOS em bactérias. Esse sistema, do qual as "mutases" fazem parte, foi selecionado, pois garante a sobrevivência em situações de estresse e, em contrapartida, gera maior número de erros e, portanto, de mutações. Dessa forma, em situações de estresse um número maior de alelos pode ser gerado e, se um novo alelo conferir vantagem adaptativa, pode ser selecionado positivamente.

Existem polimerases do DNA especializadas para a replicação do cromossomo e para atuar nos diferentes sistemas de reparo. As "mutases" são especializadas em sintetizar novas moléculas de DNA quando lesões volumosas (como dímeros de timina) representam obstáculos para a maquinaria de replicação. Em bactérias, foi demonstrada recentemente a existência de duas polimerases do DNA com essas características: a polimerase IV e a polimerase V do DNA. "Mutases" também foram descritas em eucariontes. Essas enzimas são capazes de replicar uma molécula de DNA contendo dímeros de timina com eficiência cerca de dez vezes maior do que a polimerase replicativa, porém com possibilidade maior de erro (Nelson *et al*., 1996). Genes homólogos a essas proteínas foram encontrados em células humanas (Gibbs *et al*.,1998; Xiao *et al*.,1998). Por exemplo, foi identificada a polimerase do DNA eta (em leveduras e células humanas), que é capaz de replicar DNA com lesões volumosas, com alta taxa de precisão na incorporação de nucleotídeos (Johnson *et al*., 1999; Masutani *et al*., 1999).

Como no caso dos pontos quentes, as DNA "mutases" não parecem ter evoluído diretamente a partir da seleção daqueles indivíduos que geram diversidade biológica. Provavelmente, essas enzimas foram selecionadas para a sobrevivência celular em situações de estresse, mas, como efeito lateral, essas enzimas geram maior diversidade de alelos. Em casos específicos, como a interação de agentes infecciosos e sistema imune, pode haver uma pressão seletiva no sentido de aumento da taxa de mutação (Radman, 1999). Existe uma pressão seletiva para vírus e bactérias sofrerem altas taxas de mutação, pois eles precisam escapar do sistema imune para sobreviver. Por outro lado, o sistema imune sofre mutações no sentido de gerar novos anticorpos capazes de reconhecer os novos variantes de vírus e bactérias.

Entre os mecanismos existentes que geram variabilidade genética, é importante ressaltar a importância das duplicações de genes. Como dito anteriormente, a maior parte das mutações ocorridas são deletérias para as células, uma vez que causam perda total ou parcial da função dos genes. Dessa forma, a duplicação de genes alivia a pressão de seleção negativa causada pela mutação. De fato, foi mostrado recentemente, por análises em escala genômica de sequências de sete espécies diferentes de eucariontes, que a ocorrência de duplicações é muito mais frequente do que se supunha: uma média de 0,01 duplicações por gene por milhão de anos (Lynch e Conery, 2000). A maioria das duplicatas dos genes torna-se silenciosa em alguns milhões de anos, enquanto o outro gene volta a sofrer forte pressão seletiva. Apesar disso, a duplicação é considerada uma etapa muito importante para a geração de diversidade biológica (Lynch e Conery, 2000).

Referências Bibliográficas

Albrechtsen, N., Dornreiter, I., Grosse, F., Kim, E., Wiesmuller, L. e Deppert, W. (1999). Maintenance of genomic integrity by p53: complementary roles for activated and non-activated p53. **Oncogene 53**:7706-7717.

Ames, B.N., Lee, F.D. e Durston, W.E. (1973). An Improved Bacterial Test System for the Detection and Classification of Mutagens and Carcinogens. **Proc. Natl. Acad. Sci. USA 70**: 782–786.

Andressoo, J.O., Hoeijmakers, J.H. e Mitchell, J.R. (2006) Nucleotide excision repair disorders and the balance between cancer and aging. **Cell Cycle 5**: 2886-2888.

Augusto, O., Netto, L.E.S. e Gomes, L.F. (1992). DNA Alkylation and carbon-centered radicals. **Brazilian J. Med. Biol. Res. 25**:1171-1183.

Boveri, T. (1902). Uber mehrpolige Mitosen als Mittel zur Analyse des Zellkerns. **Verh. Phys. Med. Ges. Wurzburg 35**:67–90.

Costa, R.M., Chiganças, V., Galhardo, R.S., Carvalho, H. e Menck, C.F. (2003) The eukaryotic nucleotide excision repair pathway. **Biochimie 85**:1083-1099.

Demple, B. e Harrison, L. (1994). Repair of Oxidative Damage to DNA: Enzymology and Biology. **Annu. Rev. Biochem. 63**: 915-948.

Efeyan, A., Serrano, M. (2007) p53: guardian of the genome and policeman of the oncogenes. **Cell Cycle 6**: 1006-1010.

Eisen, J.A. e Hanawalt, P.C. (1999). A phylogenomic study of DNA repair genes, proteins, and processes. **Mutat. Res**. **DNA Repair 435**: 171-213.

Falnes, P.O., Johansen, R.F. e Seeberg, E.(2002) AlkB-mediated oxidative demethylation reverses DNA damage in *Escherichia coli*. **Nature 419**: 178-182.

Falnes, P.O., Klungland A. e Alseth I. (2007) Repair of methyl lesions in DNA and RNA by oxidative demethylation. **Neuroscience 145**:1222-1232.

Friedberg, E.C. (1997). **Correcting the blueprint of life: An historical account of the discovery of DNA repair mechanisms**. Plainview, N.Y.: Cold Spring Harbor Laboratory Press. 210 pp.

Friedberg, E.C., Walker, G.C., Siede, W., Wood, R.D., Shultz, R.A., Ellenberger, T. (2006),**DNA Repair And Mutagenesis**. ASM Press, Washington, DC, EUA, 1118 pp.

Galhardo, R.S., Rocha R.P., Marques, M.V. e Menck, C.F. (2005) An SOS-regulated operon involved in damage-inducible mutagenesis in *Caulobacter crescentus*. **Nucleic Acids Res. 33**: 2603-2614.

Gasparutto, D.; Dhérin, C.; Boiteux, S. e Cadet, J. (2002) "Excision of 8-methylguanine site-specifically incorporated into oligonucleotide substrates by the AlkA protein of *Escherichia coli*" **DNA Repair 1**: 437-447.

Gibbs, P.E.M., McGregor, W.G., Maher, V.M., Nisson, P. *et al.* (1998) A human homolog of *Saccharomyces cerevisiae* REV3 gene, which encodes the catalytic subunit of DNA polymerase ζ. **Proc. Natl. Acad. Sci. USA 95**: 6876-6880.

Griffiths, A.J.F., Gelbart, W.M., Miller, J.H. e Lewontin, R.C. (1999). 7 Gene Mutations. **In Modern Genetics Analysis**. W.H. Freeman and Company, New York, pp. 197-234.

Ingvarsson, S. 1999. The Brca1 and Brca2 proteins and tumor pathogenesis **Anticancer Res. 4B** 2853-2861.

Johnson, R.E., Prakash, S. and Prakash, L. (1999) Efficient bypass of a thymine-thymine dimer by yeast DNA polymerase, Pol η. **Science 283**: 1001-1004.

Lewin, B. 1994. **Genes V**. Oxford University Press, Oxford.

Lindahl, T. e Nyberg, B. (1972) Rate of depurination of native deoxyribonucleic acid. **Biochemistry 11**:3610-3618.

Lynch, M. e Conery, J.S. (2000). The Evolutionary fate and consequences of duplicate genes. **Science 290**: 1151-1155.

Masutani, C., Kusumoto, R., Yamada, A., Dohmae, N., Yokoi, M., Yuasa, M., Araki, M., Iwai, S., Takio, K. and Hanaoka, F. (1999) The *XPV* (xeroderma pigmentosum variant) gene encodes human DNA polymerase η. **Nature 399**: 700-704.

McCulloch, S.D. e Kunkel, T.A. (2008) The fidelity of DNA synthesis by eukaryotic replicative and translesion synthesis polymerases. **Cell Res. 18**: 148-161.

Menck, C.F. (2002) Shining a light on photolyases. **Nat. Genet. 32**: 338-339.

Meselson, M. e Stahl, F.W. (1958). The replication of DNA in *Escherichia coli*. **Proc. Natl. Acad. Sci. USA** 44: 671–82.

Meulmeester, E. e Jochemsen, A.G. (2008) P53: a guide to apoptosis. **Curr. Cancer Drug Targets 8**: 87-97.

Michaels, M.L., Cruz, C., Grollman, A.P. e Miller, J.H. (1992). Evidence that MutY and MutM combine to prevent mutations by an oxidatively damaged form of guanine in DNA. **Proc. Natl. Acad. Sci. USA 89**: 7022-7025.

Modrich, P. (1991). Mechanisms and biological effects of mismatch repair. **Annu. Rev. Genet. 25**: 229-253.

Monteiro A.N. e Birge R.B. (2000). A nuclear function for the tumor suppressor BRCA1 **Histol. Histopathol. 15**: 299-307.

Nelson, J.R., Lawrence, C.W. e Hinkle, D.C. (1996). Thymine-thymine bypass by yeast DNA polymerase ζ **Science 272**: 1646-1649.

Netto, L.E.S., RamaKrishna, N.V.S., Kolar C., Cavalieri, E.L., Rogan, E.G., Lawson, T.A. e Augusto O. (1992). Identification of C8-Methylguanine in the Hydrolysates of DNA from Rats Administered 1,2-Dimethylhydrazine: Evidence for in vivo DNA Alkylation by Methyl Radicals. **J. Biol. Chem. 267**: 21524-21527.

Radman, M. (1999). Enzymes of evolutionary change. **Nature 401**:866-869.

Ribeiro, D.T., Machado C.R., Costa R.M., Praekelt U.M., Van Sluys M.A. e Menck, C.F. (1998) Cloning of a cDNA from Arabidopsis thaliana homologous to the human XPB gene. **Gene 208**: 207-213

Rocha, R.P., Paquola, A.C., Marques, M.V. Menck, C.F. e Galhardo, R.S. (2008) Characterization of the SOS regulon of *Caulobacter crescentus*. **J. Bacteriol. 190**: 1209-1218.

Sakumi, K., Furuichi, M., Tsuzuki, T., Kakuma, T., Kawabata, S., Maki, H. e Sekiguchi, M. (1993). Cloning and Expression of cDNA for a human enzyme that hydrolyzes 8-oxo-dGTP, a mutagenic substrate for DNA synthesis. **J. Biol. Chem. 268**: 23524-23530.

Sancar, A. e Sancar, G.B. (1988). DNA repair enzymes. **Annu. Rev. Biochem. 57**: 29-67.

Shrivastav, M., De Haro, L.P. e Nickoloff, J.A. (2008) Regulation of DNA double-strand break repair pathway choice. **Cell Res. 18**: 134-147.

Singer, B. (1975). The chemical effects of nucleic acid alkylation and their relation to mutagenesis and carcinogenesis. **Prog. Nucl. Acid Res. Biol. Med. 15**: 219-284.

Sutton, W.S. (1902). On the morphology of the chromosome group in *Brachystola magna*. **Biol. Bull. 4**:24-39

Trewick, S.C., Henshaw, T.F., Hausinger, R.P., Lindahl T. e Sedgwick B. (2002) Oxidative demethylation by Escherichia coli AlkB directly reverts DNA base damage. **Nature 419**: 174-178.

Umbuzeiro, G.A. e Vargas, V.M.F. (2003) Teste de mutagenicidade com Salmonella typhimurium (Teste de Ames) como indicador de carcinogenicidade em potencial para mamíferos. In: Ribeiro, L.R., Salvadori, D.M.F. e Marques, E.K. (ed) **Mutagênese Ambiental.** Editora da ULBRA, Canoas, RS, pp 81-104.

Xiao, W., Lechler, T., Chou, B.L., Fontaine, T. *et al.* (1998). Identification, chromosomal mapping and tissue-specific expression of hREV3 encoding a putative human DNA polymerase ζ. **Carcinogenesis, 19**: 945-949.

Watson J.D. e Crick F.H. 1953. A structure for deoxyribose nucleic acid. **Nature 171**: 737-738.

Watson J.D. e Crick F.H. 1974. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. **Nature 248**: 765.

Zhang, J. e Powell, S.N. (2005) The role of the BRCA1 tumor suppressor in DNA double-strand break repair. **Mol. Cancer Res. 3**: 531-539.

Capítulo 6

Sexo, por quê?

Sergio Russo Matioli (srmatiol@ib.usp.br)
Departamento de Genética e Biologia evolutiva
Instituto de Biociências
Universidade de São Paulo

Anita Wajntal (aniwa@usp.br)
Departamento de Genética e Biologia evolutiva
Instituto de Biociências
Universidade de São Paulo

"The big question now is not so much why sex evolved, but how." (Lane, 2009)

6.1. Reprodução sem Sexo e Sexo sem Reprodução

Embora o tema "sexo" pareça, em princípio, não ter relação alguma com os demais temas tratados neste livro, existem algumas teorias que relacionam sexo como decorrência necessária de processos que ocorrem na evolução no nível molecular. Isso será abordado no decorrer do capítulo, mas alguns conceitos básicos precisam ser inicialmente revistos. Sexo e reprodução são, para a espécie humana, conceitos intimamente relacionados, dado que somos uma espécie na qual a reprodução é obrigatoriamente sexuada e com sexos separados. No entanto, sexo e reprodução podem ser tratados como processos distintos e separáveis. Sexo envolve a combinação de genes de dois indivíduos em um novo indivíduo, enquanto que reprodução está relacionada à produção de novos indivíduos a partir daqueles de uma geração anterior. A reprodução na ausência de sexo é característica de organismos que produzem novos indivíduos por fissão, como ocorre em amebas ou em hidras, que produzem propágulos que originam novas colônias. O sexo sem reprodução também é característico entre organismos unicelulares, como nas bactérias, que podem transferir material genético através de mecanismos especiais, tais como a conjugação e a transformação. No processo de transformação, bactérias podem incorporar DNA que existe no meio em que vivem, uma das propriedades utilizadas para a demonstração que o DNA era o material genético, na década de 1940, e que é usada até hoje rotineiramente em laboratórios que empregam a tecnologia do DNA recombinante. Na conjugação bacteriana, duas bactérias unem fisicamente seus citoplasmas por uma estrutura tubular, o "pilus", por onde ocorre a transferência de DNA de uma bactéria doadora para uma receptora. Nesses processos, não há, no momento da transferência de material genético, a produção de descendentes. A transferência de material genético também pode ocorrer em protistas, como em *Paramecium*, que se reproduz por fissão, mas pode transferir material genético para outro indivíduo por conjugação. A união desses dois processos, sexo e reprodução, ocorre em alguns eucariontes unicelulares.

6.2. Sexos Iguais ou Diferentes

Em *Chlamydomonas*, dois indivíduos haplóides com apenas uma cópia de cada gene, a semelhança de gametas de mamíferos—sendo um da linhagem (+) e o outro a linhagem (-)—fundem-se e formam um zigoto diplóide que, por meiose, forma quatro novos indivíduos, com rearranjo dos genomas parentais (Figura 6.1). Apesar de não haver uma diferença morfológica entre os indivíduos que se fundiram, componente específicos para sexo, dispostos nos flagelos de tipos opostos, causam uma aglutinação permitindo o contato físico entre regiões específicas da membrana. O sexo feminino é definido como aquele que produz o gameta maior, o óvulo, e o sexo masculino é definido como aquele que produz o gameta menor, geralmente móvel, o espermatozóide. Assim, no caso de *Chlamydomonas* e de outras espécies que se reproduzem sexuadamente por isogamia (gametas iguais), a nomenclatura "macho" e "fêmea" não faz sentido. A anisogamia (produção de gametas de tamanhos diferentes) é amplamente majoritária nos seres vivos e deve ter surgido várias vezes independentemente, ou no mínimo duas, entre os metazoários e entre as plantas. De acordo com Iyer e Roughgarden (2008), em uma revisão, Kalmus, em 1932, foi o primeiro que sugeriu um modelo para a evolução da anisogamia, baseado na suposição de um compromisso entre o número de gametas produzidos e sua massa individual. Parker *et al.* (1972) e Maynard-Smith (1978) sugeriram que a condição isogâmica seria evolutivamente instável, pois, se alguns indivíduos mutantes produzissem gametas menores e se os mesmos se reproduzissem preferencialmente com indivíduos que produzem gametas maiores, a situação convergiria estavel-

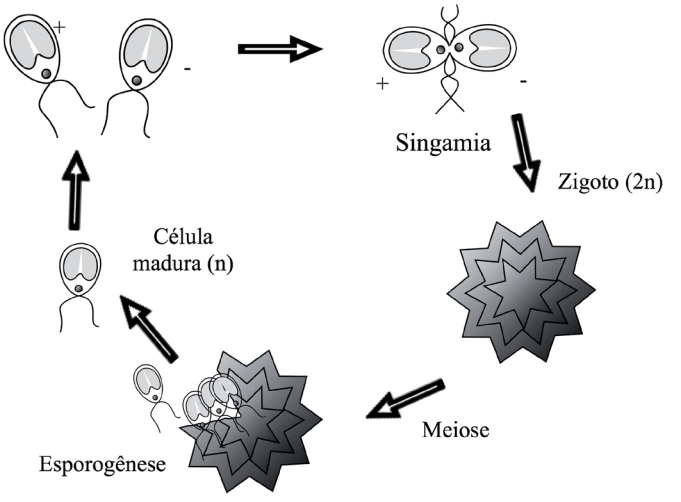


Figura 6.1. Esquema com o ciclo de vida da alga *Chlamydomonas* com isogamia, fecundação de gametas com tamanhos iguais.

mente para produção de gametas com um tamanho mínimo que se associariam a gametas com tamanho suficiente para a nutrição do embrião, o que ocorre no caso da maioria dos óvulos e dos espermatozóides. Como consequência imediata da anisogamia, temos dois sexos que investem de forma muito diferente na produção de gametas, sendo que o sexo feminino investe muito (em termos de massa e energia) em cada um dos gametas mas produz menos gametas e o sexo masculino investe pouco em cada um deles, mas produz uma quantidade muito maior de gametas. Existe uma ampla literatura que trata das consequências secundárias da anisogamia, tais como as diferenças genéticas, morfológicas, fisiológicas, de comportamento sexual e de cuidado com a prole, mas essas consequências fogem do escopo do presente capítulo. Para revisões recentes sobre essas importantes consequências da reprodução sexuada, consulte Blanckenhorn (2005), Wedell *et al.* (2006), Hokko e Jennions (2008).

6.3. Mais que Dois Sexos

Do ponto de vista estritamente teórico, seria possível a existência de três sexos de um organismo diplóide, onde cada um deles se reproduziria com a conjugação de seu gameta com outro gameta de outro indivíduo de um sexo diferente. No caso da determinação do sexo pelos sistemas XX/XY ou ZZ/ZW, há dois tipos de sexo, sendo um deles o homogamético (aqueles que possuem os dois cromossomos sexuais do mesmo tipo, XX sendo fêmeas e ZZ sendo machos) e outro heterogamético (com cromossomos sexuais diferentes, XY e ZW), em que há cruzamentos preferenciais totalmente negativos (por exemplo, no caso do sistema XX/XY, não são permitidos cruzamentos XX/XX e XY/XY). Quando há um cruzamento entre indivíduos de dois sexos diferentes XX e XY, há a produção de proporções semelhantes entre ambos os sexos. Não existe descrito na Natureza um sistema de cruzamentos com três sexos, algo como XX, XY e YY, para cada um dos sexos. Nesse caso, os cruzamentos XX/XX, XY/XY e YY/YY não seriam permitidos (um sistema de cruzamentos preferenciais totalmente negativo). A razão mais provável da inexistência de um sistema assim, com três sexos, é que esse seria um sistema bastante instável do ponto de vista da genética de populações. Nesse sistema, a estabilidade somente ocorreria se as frequências dos cromossomos X e Y permanecessem exatamente idênticas, o que somente ocorre em simulações computacionais. Qualquer desvio, por mínimo que fosse, causaria a extinção de um dos genótipos homozigotos.

A determinação do sexo por mecanismos cromossômicos é feita por uma grande diversidade de mecanismos. Além dos sistemas XX/XY e ZZ/ZW, existe ainda o sistema XX/X0—onde o macho é determinado por uma aneuploidia do cromossomo X—, e sistemas mais complexos, tais como X1X2/Y, X1X2/Y1Y2 ou até mesmo X1X1X2X2/X1X2Y (Solferini e Morgante, 1990). Em todos esses sistemas, no entanto, existem somente os dois sexos, macho e fêmea.

6.4. Reprodução Sexuada

A reprodução sexuada ocorre na grande maioria dos seres vivos multicelulares. Sua origem, sua evolução e sua manutenção são temas que têm desafiado biólogos e geneticistas, já que se pressupõe que sua ocorrência quase universal se deve a benefícios maiores que os custos associados a esse tipo de reprodução.

Um dos principais custos da reprodução sexuada foi considerado por Maynard Smith (1978), que imaginou, como modelo, o aparecimento de um mutante com reprodução assexuada em uma população sexuada com 50% dos indivíduos do gênero masculino e feminino. Os indivíduos de reprodução assexuada dobrariam seu número a cada geração em relação àqueles de reprodução sexuada. Os indivíduos de reprodução sexuada necessitam de dois indivíduos (um macho e uma fêmea) para produzir um descendente, enquanto que aqueles de reprodução assexuada poderiam produzir descendentes com um único progenitor, no caso definido sempre como uma fêmea. Esse raciocínio implicaria a existência dos machos como meros coadjuvantes na reprodução. Além do custo ocasionado pelos machos, a reprodução sexuada envolve outros custos, tais como aqueles relativos ao desenvolvimento de mecanismos relacionados à produção de gametas e órgãos sexuais diferenciados, atração entre sexos opostos e até a energia gasta para encontrar parceiros e o decorrente risco de não haver reprodução caso a busca de parceiro seja infrutífera. Assim, há uma série de vantagens teóricas para a reprodução assexuada e, em uma análise preliminar, a reprodução sexuada parece tratar-se de um paradoxo. Por esse motivo, imagina-se que haja razões muito fortes para que a reprodução sexuada prevaleça entre os eucariotos e a busca dessas razões pode levar ao entendimento das suas causas.

Diversas teorias foram propostas para explicar as eventuais vantagens da reprodução sexuada. Em 1889, August Weismann já havia proposto que a reprodução sexuada deve conferir uma vantagem em relação à assexuada, já que os descendentes de pais com prole mais variada teriam maior chance de transmitir seus genes, uma vez que estes apresentavam maiores chances de sobreviver diante de mudanças ou heterogeneidades ambientais. A hipótese levantada por Weismann foi amplamente aceita até a primeira metade do século XX, mas tem sido criticada desde então e retomada mais recentemente. Segundo Burt (2000), diversas hipóteses baseadas em considerações da genética de populações podem ser agrupadas dentro da hipótese geral feita por Weismann: a "catraca de Muller", a hipótese de Fisher-Muller (Figura 6.2), a hipótese determinística de mutação, as hipóteses de seleção flutuante e suas combinações.

6.5. Consequências da Reprodução Sexuada

Muller propôs em 1932 que a reprodução sexuada poderia impedir que as populações acumulassem mutações deletérias, que viriam a ser fixadas por flutuações casuais. Cada vez que uma mutação, mesmo que levemente deletéria, se fixasse em uma população equivaleria a um ponto sem retorno, pois a informação referente ao alelo "selvagem", presente nos genótipos de indivíduos mais adaptados, seria perdida. Nesse caso, a população como um todo teria o seu valor adaptativo médio diminuído. Embora em seu artigo de 1932 Muller não tenha mencionado a analogia desse mecanismo com aquele de uma catraca, ele fez isso mais tarde (Muller, 1964) e a hipótese chamada de "catraca de Muller" ("*Mulller's ratchet*", em inglês) ficou reconhecida como tal desde então.

Uma outra proposta para a evolução da reprodução sexuada é que o sexo é vantajoso pelo fato de permitir a recombinação de mutações benéficas que ocorrem em diferentes linhagens, diminuindo a competição entre estas diferentes mutações numa mesma população e aumentando a velocidade da adaptação às diferentes condições ambientais. Esta formulação, conhecida como o modelo de Fisher-Muller, apresenta suporte teórico, porém é difícil de ser testada experimentalmente. Por outro lado, não é muito evidente se a variação herdável da adaptabilidade é significativamente aumentada pelo sexo, pois o efeito mais óbvio da recombinação é a

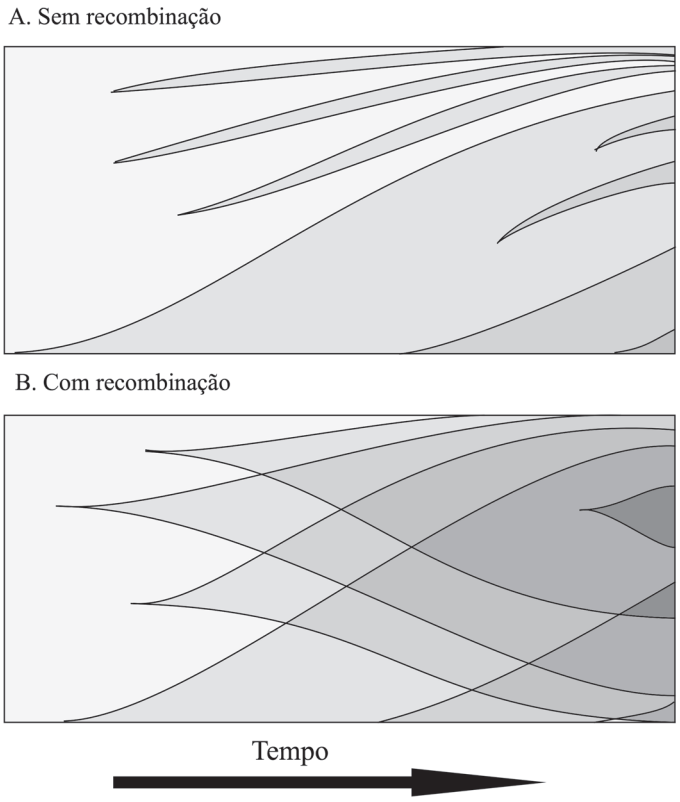


Figura 6.2. Consequências da recombinação, modificado de Muller (1932). As áreas mais escuras correspondem às proporções de indivíduos que apresentam simultaneamente mais mutantes. A. Na ausência de recombinação. B. Na presença de recombinação.

quebra de associações favoráveis de genes que foram acumulados por seleção, levando a uma carga genética recombinacional. A carga genética de uma população é definida como a parcela dos indivíduos que deixam de se reproduzir ou que deixam de sobreviver devido a uma causa genética. A segregação entre alelos do mesmo loco também tende a eliminar heterozigotos que poderiam ser favorecidos por seleção, já que, quando há cruzamentos entre dois indivíduos heterozigotos, que seriam aqueles mais adaptados, metade dos descendentes seria homozigoto, os menos adaptados. Essas forças poderiam resultar em eliminação do sexo e da recombinação genética de uma população em equilíbrio apenas sob o efeito da seleção. Recentemente, Cooper (2007), utilizando *E. coli* e plasmídeos F para mediar a recombinação e linhagens que apresentavam taxas altas ou baixas de mutação, obteve taxas de adaptação três vezes maiores na presença de recombinação quando mutações favoráveis tinham que competir para sua fixação, fornecendo uma evidência experimental para o modelo proposto por Fisher (1930) e Muller (1932).

Além dos dados experimentais, a observação da distribuição filogenética dos casos de reversão da reprodução sexuada com dois sexos para a reprodução com apenas um sexo, a partenogênese parece fornecer apoio empírico para as hipóteses baseadas nas idéias de Weismann. Existem espécies de organismos que podem se reproduzir por partenogênese tanto de maneira facultativa como obrigatória. Na maioria dos casos de partenogênese, aquela chamada de telítoca, fêmeas que não são fecundadas por machos produzem somente fêmeas, realizando aquilo que seria "o sonho das Amazonas"—tribo lendária originária da mitologia grega composta por mulheres guerreiras hábeis no manejo de cavalos e que geravam apenas descendentes do sexo feminino. Na partenogênese arrenótoca, comum no caso dos himenópteros, se as fêmeas não são fecundadas, produzem machos haplóides. Nesse tipo de partenogênese, não há ausência de sexo e tampouco de recombinação genética. Maynard-Smith (1978) chamou a atenção para a distribuição de espécies que reverteram para o estado de reprodução vegetativa ao longo da história evolutiva, em que se nota que a distribuição de espécies com reprodução por partenogênese telítoca não é concentrada em termos filogenéticos, mas ocorre aqui e ali. A interpretação desse padrão é que a "escolha" do modo assexuado de reprodução seria um beco evolutivo sem saída, onde os benefícios da assexualidade teriam uma duração limitada. Entretanto, há exceções. Trata-se de grupos de organismos que parecem estar há um longo período sem a reprodução sexuada. Judson e Normark (1996), em uma revisão sobre grupos inteiros de organismos que mantinham a reprodução assexuada, mostraram que um exame mais detalhado desses organismos permitiu constatar que, em vários deles, havia espécies com reprodução sexuada, o que evidencia que o abandono desse tipo de reprodução não havia sido tão antigo como se pensava anteriormente. De qualquer modo, esses autores ressaltam que uma análise das consequências da perda da reprodução assexuada em grupos de organismos que permaneceram por longos períodos de tempo sem sexo pode esclarecer as vantagens da própria reprodução sexuada. Uma das consequências observadas em espécies com histórico longo apenas com reprodução assexuada ficou conhecida como o "Efeito Meselson". Mark-Welsh e Meselson (2000) mostraram que, em rotíferos bdelóides, haplótipos amostrados em seus genomas apresentam uma divergência muito maior entre si que as formas alélicas que originaram tais haplótipos. Schön e Martens (2003) apontaram que o efeito Meselson não se aplica à espécie de ostrácode *Darwinula stevensoni* postulando, no entanto, que essa espécie poderia possuir um mecanismo mais eficiente de reparo de DNA, o que diminuiria a diversificação haplotípica nessa espécie.

Entre os diferentes modelos propostos para a manutenção da reprodução sexuada, um dos mais populares é a hipótese da Rainha Vermelha ("Red Queen", em inglês), exemplificado através da coevolução entre parasitas e hospedeiros. Nesse modelo, o parasita infecta os indivíduos que apresentam o genótipo mais abundante. Alguns poucos indivíduos são mutantes resistentes ao parasita e, por esse motivo, acabam substituindo os indivíduos sensíveis, eliminados pelo parasita. Inicialmente, também ocorre uma diminuição da população de parasitas, sobrevivendo apenas os mutantes que conseguem parasitar a nova linhagem de hospedeiros. Dessa forma, há uma coevolução rápida dos parasitas e dos hospedeiros. Esse modelo foi denominado de hipótese da Rainha Vermelha, inspirado na obra "Através do espelho e o que Alice encontrou por lá", de Lewis Carroll (assim nomeada por Van Valen, 1973). Nessa história, há um trecho onde a Rainha Vermelha declara que há necessidade de correr muito para não sair do lugar. No contexto de hipóteses da evolução da reprodução sexuada, esse modelo foi bastante criticado, uma vez que poderia ser importante apenas nos casos em que a evolução está relacionada a modificações rápidas, favorecendo uma minoria de indivíduos com associações de genes raros, gerados por segregação e recombinação durante a reprodução sexuada.

Kondrashov (1993, 1994) propôs que a reprodução sexuada seja mantida por sua capacidade de reduzir a carga genética de mutações deletérias que interagem sinergisticamente e concluiu que, apesar de existirem mais de 20 hipóteses que tentam explicar a evolução do sexo, há necessidade de mais dados experimentais para que se possa avançar no esclarecimento da origem e evolução do sexo. Os dados analisados por Keightley e Eyre-Walker (2000) parecem indicar que essa hipótese também não explica, por si só, a evolução do sexo.

Inúmeras hipóteses têm sido debatidas amplamente de 1930 até o presente por pesquisadores defendendo pontos de vista

extremamente divergentes, como a seleção dos genes favoráveis ou a eliminação dos genes desfavoráveis como responsáveis pelo saldo maior de vantagens conferidos pela reprodução sexual que seu custo. Há uma falta generalizada de evidências convincentes a favor de qualquer das hipóteses formuladas. Talvez o fator mais importante na permanência das discussões sobre a origem e a manutenção da reprodução sexuada é que vários dos fatores não são mutuamente exclusivos. Nesse caso, a importância relativa de cada um deles precisa ser determinada e também pode ser que essas importâncias relativas também não sejam homogêneas. Assim, em algumas situações, pode haver fatores preponderantes que não sejam os mesmos para outras situações.

6.6 Sexo e Reparo do DNA

Muito antes da origem do sexo dos eucariontes, os procariontes já tinham adquirido a capacidade de recombinação genética (emparelhamento, quebra e ligação entre regiões homólogas do DNA). A função original desses eventos provavelmente não está ligada a novidades evolutivas, mas ao reparo de danos normalmente causados por diversos tipos de agentes físicos e químicos, muitos dos quais presentes constantemente nas células (veja Capítulo 5). Uma das hipóteses relacionadas à evolução da reprodução sexuada diz muito mais respeito à manutenção da integridade dos genomas que ao provimento de variabilidade genética para as populações. Trata-se da hipótese de reparo do DNA, que ocorre pela verificação mútua de genomas que estavam em indivíduos diferentes mas que passam a compartilhar o mesmo núcleo celular (veja Capítulo 5, a respeito do reparo recombinacional homólogo). Essa hipótese foi inicialmente proposta por Dougherty (1955) e, depois, mais elaborada por Bernstein *et al.* (1981, 1985, 1987), que propuseram que o aspecto mais relevante na evolução do sexo é a capacidade de reparo de danos sofridos pela molécula de DNA, principalmente dos danos que afetam ambos os filamentos da dupla hélice. Conforme esses autores, esse tipo de dano é bastante frequente e radicais altamente reativos resultantes de respiração celular podem ser responsáveis por mais de 600 bases de DNA modificadas por dia no caso de cada uma das células humanas, valor calculado pela multiplicação das taxas espontâneas estimadas por nucleotídeo pelo tamanho de dois genomas. Como a proporção de danos de origem oxidativa para danos afetando os dois filamentos foi estimada em 1:60, cerca de 10 alterações afetariam ambos os filamentos. Mesmo considerando que 90% dos danos ocorrem no DNA mitocondrial e apenas 10% no DNA nuclear, bastaria um único dano afetando ambos os filamentos para levar a célula à morte. Os autores baseiam sua hipótese de evolução do sexo em diversas evidências:

a) muitas das enzimas relacionadas ao reparo estão envolvidas na recombinação meiótica (Marcon e Moens, 2005; veja também Capítulo 5);
b) em vírus, bactérias e fungos, após exposição a substâncias relacionadas a danos de DNA, há evidências de que na presença de dois genomas, as chances de recuperação aumentam substancialmente em relação aos casos em que o vírus ou a bactéria tem que contar apenas com seu próprio genoma, mostrando que o reparo é mais eficiente na presença de recombinação;
c) os organismos diplóides, além da vantagem que apresentam pela possibilidade de reparo de danos que afetam ambos os filamentos, também apresentam uma vantagem inicial, mascarando as novas mutações deletérias por complementação;
d) essas mutações acumulam-se em heterozigose, dificultando o retorno à haploidia ou autofecundação, que exporia os genes recessivos.

Portanto, conforme esse modelo, uma vez estabelecida, a reprodução sexuada tende a perpetuar-se. Essa hipótese, entretanto, tem sido questionada por diversos autores, que apontaram várias fragilidades nessa formulação.

Mais recentemente, Yahara *et al.* (2007) realizaram uma série de simulações para examinar a validade da hipótese de reparo, utilizando como modelo o reparo por conversão gênica de quebras induzidas nos dois filamentos do DNA. Na modelagem realizada, os autores utilizaram uma série de fatores envolvidos nas interações entre bacteriófago e uma bactéria que contém uma enzima de restrição. Assumindo que o genoma do bacteriófago contém um alelo sensível ou resistente à enzima e um outro loco que determina a proficiência (Rec^+) ou deficiência (R^-) na função de recombinação/reparo, as vantagens conferidas pelo alelo para proficiência de reparo (alelo sexual) só puderam ser detectadas quando a razão entre a quantidade de bacteriófagos liberados pela célula hospedeira contendo danos por duplas quebras (b_1) e a quantidade de bacteriófagos liberados por células hospedeiras não danificada (b_o) foram considerados. As simulações mostraram que a evolução do alelo Rec^+ só ocorre quando b_1/b_o é pequeno (< 1) e o custo envolvido é baixo. Portanto, a validade da hipótese de reparo para a origem do sexo foi confirmada apenas sob determinadas condições.

6.7. Origem da Meiose

Na reprodução sexuada da maioria dos organismos eucarióticos, a formação de gametas dá-se por um tipo de divisão celular conhecido como meiose. Nesse tipo de divisão celular, assim como ocorre na divisão mitótica, ocorre a replicação de cada um dos cromossomos na fase inicial. Ao contrário do que ocorre na mitose, os cromossomos duplicados emparelham-se após a replicação, com a formação de um complexo protéico que une as cromátides homólogas, denominado complexo sinaptonêmico. Historicamente, a observação da formação de quiasmas, estruturas citológicas cruciformes que ocorrem durante a divisão meiótica, associada ao fenômeno da recombinação genética, foi uma das evidências levantadas a favor da teoria cromossômica da herança, conforme verificado por Bridges, quando estudou anormalidades na segregação de marcadores genéticos de *Drosophila*, em 1916. O complexo sinaptonêmico foi descrito bem mais tarde por Moses (1956) e por Fawcett (1956). O fenômeno da recombinação é anterior temporalmente à formação do complexo sinaptonêmico entre as cromátides homólogas (Padmore *et al.*, 1991).

Algum tempo após se iniciarem os estudos de recombinação genética, notou-se o fenômeno da interferência. Quando se estudavam recombinantes entre três marcadores genéticos, observou-se que a taxa de recombinação observada não era a mesma que se esperaria se a distribuição dos eventos de recombinação fosse completamente aleatória, ou seja, o fato de haver um evento de recombinação interferia na chance de haver outro evento nas vizinhanças, diminuindo-a. Isso pode ser explicado se houvesse um numero limitado de quiasmas para cada cromossomo. De fato, atualmente se sabe que a quantidade de recombinação é proporcional não ao tamanho do genoma, mas à quantidade de braços cromossômicos (de Villena e Sapienza, 2001). De acordo com Wilkins e Hollyday (2009), o emparelhamento dos cromossomos na divisão celular seria o fator chave para explicar a evolução da divisão meiótica a partir da mitose, pois é a única

etapa realmente diferente entre os dois tipos de divisão celular. Segundo esses autores, o emparelhamento entre cromossomos teria sido favorecido pela seleção natural por limitar a quantidade de recombinação, uma afirmação surpreendente quando se adota a postura de enxergar a evolução da reprodução sexuada como um processo onde o fator do aumento de variabilidade é o que normalmente é levado em conta. Entretanto, dentro do ponto de vista de que a reprodução sexuada teria evoluído como um processo que ajuda a implementar o reparo do material genético, tal afirmação não seria tão surpreendente. Entretanto, as etapas para a evolução da meiose a partir da mitose não estão esclarecidas.

6.8. Conclusões

Entre as diversas teorias propostas para a origem, evolução e manutenção da reprodução sexuada, há algumas observações, evidências experimentais e simulações que oferecem apoio a cada uma das teorias propostas sob condições particulares. Em uma análise crítica sobre as diferentes hipóteses, West *et al.* (1999) concluíram que diferentes mecanismos ambientais e genéticos podem estar interagindo ou serem importantes em determinadas condições e em diferentes espécies. Esses autores propõem uma visão pluralista para explicar a evolução do sexo e da recombinação. Apesar de que uma teoria científica idealmente deva ser testável através da possibilidade de sua refutação, em Biologia é comum haver hipóteses que não são mutuamente exclusivas, o que torna a teoria correspondente uma teoria de natureza quantitativa, onde o que importa é a contribuição proporcional de cada um dos processos que atuam em um mesmo fenômeno. A reprodução sexual, afinal, propicia a eliminação de mutantes deletérios, a combinação de mutantes favoráveis, a possibilidade de verificação recíproca de genomas e também o aumento na velocidade da evolução. Tudo ao mesmo tempo. Esse tipo de pluralidade é bastante comum dentro da evolução. De qualquer maneira, devemos ter sempre em mente a analogia que François Jacob fez da evolução biológica com o trabalho de um funileiro (Jacob, 1977): ao contrário de um engenheiro, que se preocupa em fazer uma combinação de elementos especialmente desenhados com um propósito, o funileiro trabalha com os elementos que estão à mão.

Referências Bibliográficas

Bernstein, H., Byerly, H.C. e Michod, R.E. (1981). Evolution of sexual reproduction: importance of DNA repair, complementation, and variation. **Am. Nat. 117**(4):537-549.

Bernstein, H., Byerly, H.C., Hopf, F.A. e Michod, R.E. (1985). The evolutionary role of recombinational repair and sex. **Int. Rev. Cytol. 96**:1-28.

Blanckenhorn, W. U. (2005). Behavioral causes and consequences of sexual size dimorphism. **Ethology 111**:977-1016.

Burt, A. (2000). Sex, recombination, and the efficacy of selection - Was Weismann right? **Evolution 54**(2):337-351.

Cooper, T.F. (2007). Recombination speeds adaptation by reducing competition between beneficial mutations in populations of *Escherichia coli*. **PLoS Biol. 5**(9):e55

De Villena, F.P.M. e Sapienza. C. (2001). Recombination is proportional to the number of chromosome arms in mammals. **Mamm. Genome 12**:318–322.

Dougherty, E.C. (1955). Comparative evolution and the origin of sexuality. **Syst. Zool**. **4**(4):145-190.

Fawcett, D.W. (1956). The fine structure of chromosomes in the meiotic prophase of vertebrate spermatocytes. **J. Biophys. Biochem. Cytol. 2**:403–406.

Fisher, R.A. (1930) **The genetical theory of natural selection**. Clarendon Press, Oxford.

Hokko, H. e Jennions, M. D. (2008). Parental investment, sexual selection and sex ratios. **J . Evol. Biol. 21**:919–948.

Iyer, P. e Roughgarden, J. (2008) Gametic conflict versus contact in the evolution of anisogamy. **Theoret. Pop. Biol. 73**(4):461-472.

Jacob, F. (1977). Evolution and thinkering. **Science 196**:1161-1166.

Judson, O.P. e Normark, B.B. (1996). Ancient asexual scandals. **Trends Ecol. Evol. 11**(2):41-46.

Kalmus, H. (1932) Ueber den erhaltungswert den phaenotypischen anisogamie und die entstehung der ersten geschlectsunterschiede. **Biol. Zentral. 52**:716.

Keightley, P.D. e Eyre-Walker, A. (2000). Deleterious mutations and the evolution of sex. **Science 290**:331-333.

Kondrashov, A.S. (1993). Classification of hypotheses on the advantage of amphimixis. **J. Hered.** 84:372-387.

Marcon, E. E Moens, P.B. (2005). The evolution of meiosis: recruitment and modification of somatic DNA-repair proteins. **BioEssays 27**:795–808

Maynard-Smith, J. (1978). **The evolution of sex**. Cambridge University Press.

Moses, M.J. (1956). Chromosomal structures in crayfish spermatocytes. **J. Biophys. Biochem. Cytol. 2**:215–218.

Muller, H. G. (1932) Some genetic aspects of sex. **Am. Nat. 66**:118-138.

Muller, H. G. (1964) The relation of recombination to mutational advance. **Mutat. Res. 1**: 2–9.

Padmore, R., Cao, L. e Kleckner, N. (1991). Temporal comparison of recombination and synaptonemal complex formation during meiosis in *S. cerevisiae*. **Cell 66**(6): 1239-1256.

Parker, G.A., Baker, R.R. and Smith, V.G.F. (1972). The origin and evolution of gamete dimorphism and the male–female phemomenon. **J. Theor. Biol. 36**:529–533.

Schön, I. e Martens, K. (2003). No slave to sex. **Proc. R. Soc. Lond. B 270**:827-833.

Solferini, V.N. e Morgante, J.S. (1990). X1X1X2X2:X1X2Y mechanism of sex determination in *Anastrepha bistrigata* and *A. serpentina* (Diptera: Tephritidae). **Rev. Brasil. Genet. 13**(2):201-208.

Van Valen, L.M. (1973). A new evolutionary law. **Evol. Theory 1**:1–30.

Wedell, N., Kvarnemo, C, Lessells, C. M. e Tregenza, T. (2006). Sexual conflict and life histories. **Anim. Behav. 71**:999-1011.

Weismann, A. (1889). **Essays upon Heredity**. Vol. 1 e 2. Claredon Press, Oxford. Edição online: http://www.esp.org/books/weismann/essays/facsimile/

West, S.A. Lively, C. M. e Read, A. F. (1999). A pluralist approach to sex and recombination. **J . Evol. Biol. 12**:1003-1012.

Wilkins, A.S. e Holliday, R. (2009). The evolution of meiosis from mitosis. **Genetics 181**:3–12.

Yahara, K, Horie, R., Kobayashi, I. e Sasaki, A.(2007). Evolution of DNA double-strand break repair by gene conversion: coevolution between a phage and a restriction-modification system. **Genetics 176**:513–526.

Página deixada em branco

Capítulo 7

Taxas de evolução e relógios moleculares

Daniela Calcagnotto (dcalcag@terra.com.br)
Departamento de Genética e Biologia Evolutiva
Instituto de Biociências
Universidade de São Paulo

"*Todos se sucedem, todos se lembram uns dos outros. Todos estão ali à espera dos que chegam.*" (Cecília Meireles)

Uma das grandes contribuições de Charles Darwin para a Biologia foi propor um mecanismo de evolução no qual as espécies se modificariam e se diversificariam através de um fenômeno de natureza populacional. A esse fenômeno, Darwin denominou seleção natural, onde as variações herdáveis existentes se propagam nas populações conforme as vantagens adaptativas conferidas por ambientes determinados. O efeito cumulativo dessas variações origina a diversidade biológica.

Atualmente sabemos que a variabilidade genética existe no nível dos ácidos nucléicos e, como visto no Capítulo 5, sua única origem é a mutação. Sabemos ainda que, além da seleção natural, no processo de deriva genética, um alelo pode ter sua frequência alterada por uma questão de amostragem aleatória entre gerações.

O processo de substituição alélica, no qual um alelo é substituído por outro na evolução de uma população, combina o processo de mutação com um ou ambos os fenômenos responsáveis pela mudança das frequências alélicas, nominalmente, seleção natural e deriva genética.

A idéia de que as substituições alélicas ocorrem em intervalos mais ou menos regulares de tempo esteve presente desde o surgimento dos primeiros modelos criados para explicar a evolução molecular. A origem de novos alelos pode dar-se pela substituição de um nucleotídeo por outro, denominadas mutações pontuais, ou pela inserção ou deleção de nucleotídeos e inversões de sequências de nucleotídeos. Quando duas ou mais sequências homólogas de nucleotídeos apresentam tamanhos diferentes, pode-se inferir que, durante a história evolutiva dessas sequências, ocorreram inserções ou deleções de nucleotídeos. Nesse caso, as substituições são denominadas coletivamente de "indels" (inserção ou deleção), pois muitas vezes não se pode inferir com certeza o estado ancestral, isto é, se ocorreu uma inserção ou uma deleção. As indels podem ter efeitos deletérios muito graves se houver alteração do quadro de leitura do DNA, isto é, se provocarem mudança na sequência correspondente de aminoácidos. Mas, para efeito do estudo da evolução das proteínas, vamos nos deter às substituições de nucleotídeos.

Na primeira parte desse capítulo, serão discutidas as diferenças entre porções diferentes de um mesmo gene, entre genes e entre as regiões codificadoras e não codificadoras do genoma. Na segunda parte, serão apresentadas as diferenças observadas entre linhagens à luz das evidências contra e a favor da hipótese do relógio molecular. No final, serão introduzidas novas idéias sobre os fatores determinantes da taxa de evolução molecular e a descrição de modelos que combinam princípios de alometria e cinética bioquímica com a teoria neutralista de evolução. Além disso, serão apresentadas uma breve revisão sobre os métodos de análise mais recentes, os diferentes métodos de calibragem do relógio molecular, bem como suas aplicações em estudos filogenéticos e abordagens ecológicas (processos intra-específicos).

7.1. Bases para o Relógio Molecular

As mutações pontuais ocorrem quando uma base é erroneamente incorporada durante a replicação do DNA (ver Capítulo 5). As mutações que geram as substituições podem ser classificadas em: (1) *transições*, quando uma purina (A e G) é incorporada no lugar de outra purina ou quando uma pirimidina (C e T) é incorporada no lugar de outra pirimidina; e (2) *transversões*, nos casos de mutações que envolvem uma purina e uma pirimidina—por exemplo, quando A é trocada por T. Quando as mutações ocorrem em regiões codificadoras do genoma, elas podem ser classificadas em *sinônimas* e *não sinônimas*. Essa segunda classificação diz respeito ao efeito que a mudança provocará na proteína. Quando um nucleotídeo é substituído por outro sem causar alteração no aminoácido resultante, diz-se que a substituição é sinônima (Figura 7.1A). Quando há alteração no aminoácido resultante, a substituição é considerada não sinônima (Figura 7.1 B). Um último tipo pode ser classificado ainda como *mutação sem sentido*, quando a alteração transforma um códon que originalmente codificava para um aminoácido em um códon de parada (Figura 7.1C).

De uma maneira bem simples, a distância entre duas sequências pode ser estimada contando-se o número de bases diferentes. A taxa de substituição de nucleotídeos (r), ou seja, o

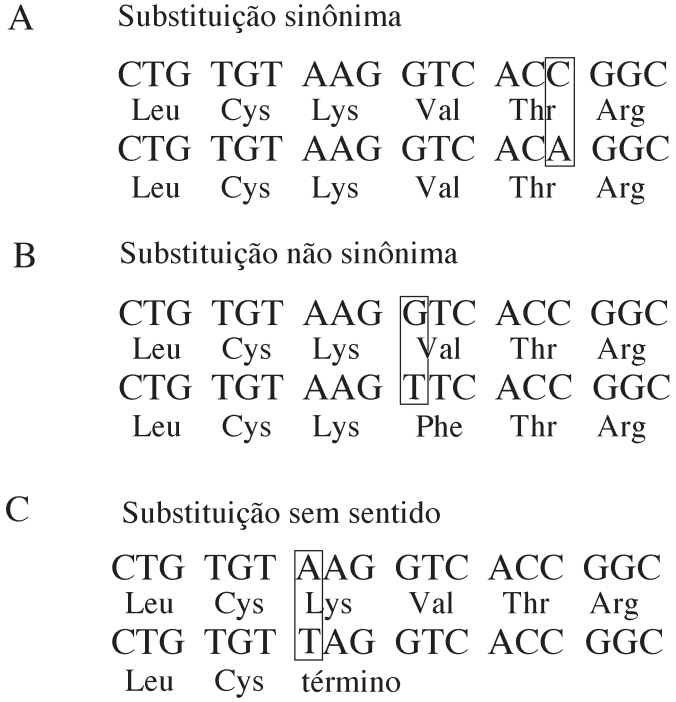


Figura 7.1. Tipos de mutação com relação aos seus efeitos na síntese protéica.

número de substituições que ocorrem por sítio por unidade de tempo, é uma medida mais precisa e pode ser calculada utilizando-se a expressão proposta por Li (1997),

$$r = k/2T,$$

onde k é o número de substituições por sítio entre duas sequências homólogas e T é o tempo de divergência entre essas duas sequências (Figura 7.2). O tempo de divergência T, normalmente inferido a partir de dados paleontológicos, é considerado igual para as duas espécies cujas sequências estão sendo analisadas, pois, de fato, o tempo transcorrido desde a espécie ancestral para cada uma delas até o presente foi o mesmo.

A taxa de substituição de nucleotídeos apresenta-se bastante variável, dependendo de qual parte do genoma está sendo analisada. Como se sabe, a maioria dos eucariotos possui uma quantidade de DNA muito superior àquela que se supõe necessária para produzir ou regular a produção das proteínas. Desse modo, é possível dividir o genoma em regiões que contêm *genes* ou suas *sequências regulatórias*, e regiões aparentemente sem função, denominadas *regiões não codificadoras*.

Quando comparamos as taxas de substituição entre esses dois tipos de regiões, percebemos que há uma clara diferença entre as regiões codificadoras e as não codificadoras do genoma. O DNA não codificador apresenta taxas de modificação superiores ao DNA que codifica para proteínas. Outra característica do DNA não codificador é a ocorrência de indels com maior frequência. A taxa de substituição de um gene é um bom indicativo de sua importância funcional e a comparação das taxas entre diferentes genes fornece informações valiosas sobre fatores históricos e demográficos (Muse, 2000). Por essas razões, regiões codificadoras recebem uma atenção especial e muitos estudos comparativos têm sido publicados (Matsubara e Yamanaka, 1978; Fambrough, 1994; Bargelloni *et al.*, 1999; Kocher *et al.*, 1995). A despeito da divisão do DNA entre regiões codificadoras e não codificadoras, é importante ressaltar que as substituições sinônimas ocorrem invariavelmente com mais frequência do que as não sinônimas.

Para facilitar o entendimento, optamos por tratar as regiões codificadoras e as não codificadoras separadamente e, em cada uma delas, analisar os efeitos das substituições sinônimas e não sinônimas.

7.1.1. Regiões Codificadoras

Na Figura 7.3, são apresentadas as taxas de substituição não sinônimas e sinônimas em alguns genes. Estes dados foram

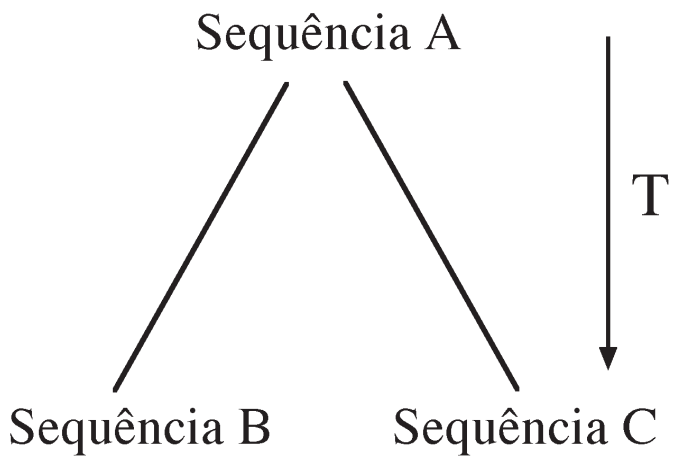


Figura 7.2. Divergência entre duas sequências (B e C) a partir de uma sequência ancestral (A), depois de transcorrido o tempo T. Note que, com relação às sequências B e C, transcorreu o tempo 2T.

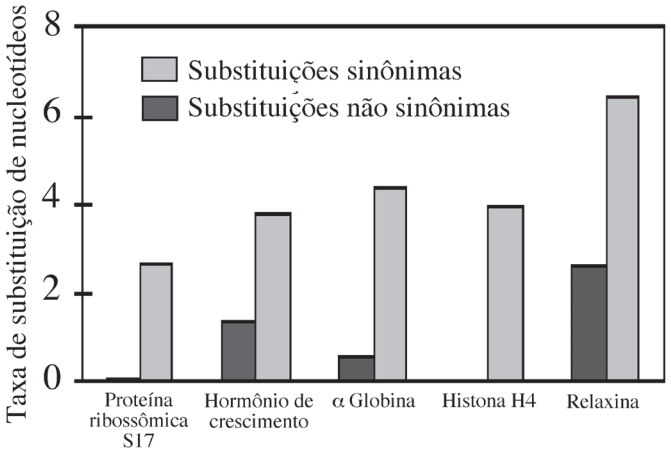


Figura 7.3. Taxas de substituições sinônimas e não sinônimas entre algumas proteínas de roedores e humanos (em número de substituições por sítio por 10^9 anos). Dados compilados por Li (1997).

compilados por Li (1997), que assumiu um tempo de divergência entre humanos e roedores de 80 milhões de anos. Examinando a Figura 7.3, podemos perceber que as taxas de substituições não sinônimas são bastante variáveis entre diferentes genes, com taxas que vão desde zero, em histonas, até 3×10^{-9}, no interferon γ. Esse tipo de substituição tem efeito direto na evolução das proteínas, pois produz alteração nos aminoácidos. As proteínas estruturais, como as histonas, que interagem diretamente com o DNA, são extremamente conservadas evolutivamente, ou seja, apresentam poucas substituições quando comparadas entre táxons muito diferentes. Proteínas que interagem com outras proteínas—por exemplo, o glucagon, hormônio que age na permeabilização de membranas celulares com relação à glicose, com efeito contrário ao da insulina—são mais flexíveis em termos evolutivos e apresentam altas taxas de substituição. Existem proteínas que apresentam taxas de substituição intermediárias, como no caso do hormônio de crescimento. Na maioria das enzimas, as taxas de substituições não sinônimas são baixas (Li, 1997).

Mutações em diferentes posições do códon apresentam probabilidades diferentes de causar alterações no aminoácido. Os sítios em regiões codificadoras são classificados como *não degenerados*, se todas as substituições que ocorrem nesses sítios forem não sinônimas, *duas vezes degenerados*, se uma das três substituições possíveis for sinônima e *quatro vezes degenerados*, se todas as substituições possíveis forem sinônimas (Li *et al.*, 1985). Isso ocorre por causa da natureza do código genético, onde existem certos aminoácidos que dependem apenas da sequência das duas primeiras bases, não importando qual seja a terceira. As taxas de substituição de nucleotídeos são mais baixas em sítios não degenerados, medianas nos sítios duas vezes degenerados e muito altas nos sítios quatro vezes degenerados, como pode ser visto na Tabela 7.1.

Como podemos ver na Tabela 7.1, o número de transições nos sítios de códons duas vezes degenerados é maior do que o nú-

Tabela 7.1. Taxas de transições e transversões em códons não degenerados, 2 vezes degenerados e 4 vezes degenerados expressas em unidades de substituição/sítio/109 anos.

Substituição	Não degenerado	2 vezes degenerado	4 vezes degenerado
Transição	0,40	1,86	2,24
Transversão	0,38	0,38	1,47
Total	0,78	2,24	3,71

Modificada a partir de Li (1997).

mero de transversões, apesar de existirem dois tipos de transversão e apenas um de transição. O fato de as transições ocorrerem com maior frequência que as transversões poderia ser uma explicação. Nos sítios duas vezes degenerados, as taxas de transição são um pouco menores do que nos sítios de códons quatro vezes degenerados, enquanto que a taxa de transversão é bem menor nos sítios de códons duas vezes degenerados que nos sítios de códons quatro vezes degenerados. Isso pode ser devido ao fato de que nos sítios duas vezes degenerados todas as transversões são não sinônimas. Nos sítios não degenerados, as taxas de transição e transversão são quase iguais e ambas são menores do que as observadas em sítios de códons quatro vezes degenerados. Isso ocorre porque todas as mutações nesses sítios são não sinônimas (Li, 1997).

Quando diferentes regiões do mesmo gene são comparadas, os sítios sinônimos novamente apresentam taxas evolutivas mais elevadas que os não sinônimos (Li, 1997; Page e Holmes, 1998). As mutações que resultam em mudança de aminoácido, isto é, as não sinônimas, têm uma chance maior de serem deletérias e grande parte será eliminada da população, resultando, assim, numa taxa menor de substituição nos sítios não sinônimos. Quando as taxas de substituição não sinônimas encontradas são superiores às sinônimas, como em algumas porções dos genes para imunoglobulinas (Tanaka e Nei, 1989) e do complexo de histocompatibilidade (Hughes e Nei, 1989), atribuem-se as altas taxas de substituição não sinônimas à seleção positiva; e, quando os genes são examinados como um todo, as taxas de substituição sinônimas serão maiores que as não sinônimas. Se, por um lado, as mutações não sinônimas têm grande chance de serem deletérias, as mutações sinônimas, por outro lado, têm uma maior probabilidade de serem neutras e, portanto, de serem fixadas na população (Li e Graur, 1991). Pelo menos em mamíferos, acredita-se que as mutações sinônimas sejam neutras e apresentam uma pequena probabilidade de fixação. Permanece aberta a questão de os sítios não sinônimos em mamíferos serem neutros ou apresentarem vantagens seletivas. Aparentemente, esses sítios poderiam estar reduzindo ou aumentando a adaptabilidade de maneira sutil, acarretando mudanças muito pequenas na função ou na estrutura da proteína (Page e Holmes, 1998).

As diferenças entre as taxas de substituições sinônimas e não sinônimas parecem indicar que, quanto maiores forem as restrições funcionais, menores serão as taxas evolutivas, ou seja, maior é a probabilidade de a mutação ser deletéria e não neutra. Fisher (1930) já argumentava que, quanto maior o efeito de uma mutação, menor é a chance de que seja vantajosa. O modelo formulado por Kimura (1977) considera que uma fração das mutações é neutra e as restantes são deletérias. Ainda, conforme esse modelo, a taxa de mutação/sítio/tempo é igual nos sítios sinônimos e não sinônimos. Entretanto, como a fração de substituições neutras é maior nos sítios sinônimos, estes apresentam taxas evolutivas mais elevadas. Em outras palavras, a maior taxa de substituição deve ser esperada nas sequências cuja alteração interfira pouco na sua função.

Para ilustrar os efeitos de restrições seletivas em genes diferentes, examinemos duas proteínas que possuem taxas de substituição não sinônimas bastante diferentes, as apolipoproteínas e a histona 3. As apolipoproteínas são moléculas transportadoras de lipídios que atuam na corrente sanguínea. Os domínios onde ocorrem as ligações com lipídios são compostos, em sua maioria, por resíduos hidrofóbicos. Análises comparativas de sequências de apolipoproteínas de vários mamíferos sugerem que a substituição de um aminoácido hidrofóbico por outro é possível em muitos sítios (Luo *et al.*, 1989). Já no caso das histonas, onde os aminoácidos interagem diretamente com o DNA ou com outras histonas, poucas substituições são possíveis, o que faz dessa uma das proteínas conhecidas com taxa de evolução mais baixa (Li, 1997).

Um aspecto importante que deve ser levado em consideração é que as restrições funcionais não são "entidades fixas" (Page e Holmes, 1998). Em um determinado ambiente, um aminoácido pode ser seletivamente importante, enquanto que, em outro, esse mesmo aminoácido pode ser neutro. Alterações na restritividade funcional podem ser medidas pela razão entre substituições não sinônimas e sinônimas (d_N/d_S). Se as substituições sinônimas, como as não sinônimas, forem neutras, teremos $d_N=d_S$ e a relação teria um valor próximo a 1. Se as substituições não sinônimas enfrentam restrições funcionais ($d_N<d_S$), a relação seria menor que 1. Na presença de seleção positiva, ocorrerão mais substituições não sinônimas e $d_N > d_S$, resultando numa relação maior que 1. Isso ocorreria porque a taxa com que as mutações são fixadas é maior que a taxa de mutação (Page e Holmes, 1998). Os vírus que causam imunodeficiência em humanos (HIV-1) e em símios (SIV_{MAC}) são exemplos de alteração de restrições funcionais. Acredita-se que esses dois vírus tenham surgido através de eventos de transmissão interespecífica a partir de outros primatas, onde eram menos patogênicos (SVI_{AGM} e SVI_{SMM}). Tanto o HIV-1 como o SIV_{MAC} provocam uma forte resposta imunológica nos novos hospedeiros. Nesse caso, haverá seleção positiva para aqueles vírus cuja capa protéica constitua-se de proteínas com mutações que não sejam detectadas pelo sistema imunológico do hospedeiro. Vírus com essa característica terão sua frequência aumentada rapidamente. A razão d_N/d_S das sequências da proteína g120 do envelope protéico é mais alta no HIV-1 e SIV_{MAC} do que no SIV_{AGM} e SIV_{SMM} (Shpaer e Mullins, 1993; Page e Holmes, 1998).

Além das restrições funcionais, a estrutura terciária das proteínas é outro fator que pode influenciar nas taxas de substituição de nucleotídeos. Há indícios de que diferentes regiões da estrutura terciária de uma proteína apresentam taxas evolutivas diferentes. O caso das hemoglobinas de vertebrados é um dos mais bem estudados. A hemoglobina é um tetrâmero composto de duas cadeias α e duas cadeias β. O interior da molécula contém principalmente aminoácidos hidrofóbicos e é bastante conservado. Nessa região, a taxa de substituição de nucleotídeos foi estimada em $0{,}17 \times 10^{-9}$ e $0{,}24 \times 10^{-9}$ substituições por ano nas cadeias α e β, respectivamente (Kimura e Ohta, 1973). Nos sítios que estão na superfície da molécula, as taxas de substituição são cerca de 10 vezes superiores. A existência de aminoácidos hidrofóbicos no interior da molécula e aminoácidos hidrofílicos no exterior é bastante comum em proteínas globulares e esses aminoácidos determinam sua estabilidade e dobramento correto durante a síntese protéica. A diferença nas taxas de substituição entre o interior com relação ao exterior pode ser explicada pela restrição funcional do interior da molécula, onde se liga o grupo heme, que contém um íon de ferro, o responsável último pela função da molécula no transporte de oxigênio.

O motivo da variação na taxa de substituição sinônima entre genes não é tão bem compreendido. Podem ser duas as razões envolvidas. Primeiro, a taxa de mutação pode variar entre regiões diferentes do genoma. Em outras palavras, a variação nas taxas de mutações sinônimas poderia estar refletindo a posição do gene no cromossomo. Segundo, é possível que substituições sinônimas estejam sujeitas a restrições seletivas distintas, isto é, enquanto algumas são selecionadas a favor, outras são selecionadas contra, uma vez que nem todos os códons sinônimos apresentam a mesma adaptabilidade (Li, 1997). Essa seleção poderia criar um viés na utilização de códons entre genes, provocando variação nas taxa de substituições sinônimas.

O fato de o código genético ser degenerado significa que a maioria dos 20 aminoácidos é codificada por mais de um códon. Se

todas as mutações sinônimas fossem realmente neutras, os códons sinônimos deveriam ser utilizados em frequências similares. Entretanto, mutações em diferentes posições do códon são diferentes em termos da probabilidade de provocar alteração de aminoácido, fazendo com que as três posições do códon sejam diferentes quanto às restrições funcionais a que estão sujeitas. À medida que um maior número de sequências se tornou disponível em bancos de dados, ficou evidente que, tanto em procariotos como em eucariotos, a utilização dos códons não é ao acaso (Grantham *et al.*, 1980; veja também o Capítulo 10, para uma discussão mais detalhada). Por exemplo, o ácido glutâmico é especificado por dois códons GAA e GAG. Nos genes nucleares humanos, GAG é utilizado na proporção de 40 para cada 1000 códons, enquanto que, no caso do códon GAA, a proporção é de 28 para 1000. Na bactéria *Escherichia coli*, a situação é inversa (Page e Holmes, 1998). A proposição de que a seleção natural favorecesse códons particulares, os quais, por sua vez, aumentariam a adaptabilidade do organismo e o modelo de coevolução códon-tRNA, corrobora a premissa neutralista de que existe uma relação entre restrição funcional e taxa de substituição de nucleotídeos.

Grantham *et al.* (1980) observaram que genes em organismos ou em espécies proximamente relacionados apresentam o mesmo padrão de utilização de códons sinônimos e propôs a hipótese de que o viés na utilização de códons é espécie-específico. Essa hipótese parece ser verdadeira de modo geral, mas existe uma grande variação na utilização de códons por diferentes genes. A Utilização Relativa de Códons Sinônimos (RSCU, sigla do inglês *Relative Synonymous Codon Usage*), proposta por Sharp e Li (1986), pode ser calculada dividindo-se o número observado de ocorrências de um determinado códon pelo número de ocorrências esperado de todos os códons para aquele aminoácido.

A Tabela 7.2 apresenta parte de uma extensa compilação de dados sobre a utilização não ao acaso de códons sinônimos feita por Sharp *et al.* (1988). Se a utilização de códons fosse homogênea, a frequência de cada um deveria ser igual a 1.

Observando-se a Tabela 7.2, percebe-se que a utilização é diferente de 1,0 na maioria dos casos. Em *E. coli* e em *S. cerevisiae*, percebe-se que o desvio é maior em genes com alta taxa de expressão. Isso pode ser devido ao efeito da seleção natural sobre a eficiência, no sentido de síntese rápida e minimização de erros durante a tradução (Tabela 7.3). Essa eficiência maior, quando existe desvio na utilização de códons, pode ser entendida da seguinte maneira: suponha a existência de um aminoácido que pode ser codificado por quatro códons diferentes. A velocidade de reação da síntese protéica depende, como qualquer outra reação química, da concentração dos substratos. Um dos substratos da reação é o tRNA específico para o códon em questão. Se houver quantidades iguais de cada um dos quatro tRNAs, qualquer códon seria sintetizado na mesma velocidade. No entanto, como acontece nos diferentes organismos, a abundância relativa de cada um dos tRNAs não é a mesma, existindo um ou mais deles em maior concentração. Nesse caso, a cadeia polipeptídica é sintetizada mais rapidamente se os códons utilizados corresponderem aos tRNAs mais abundantes. Nos genes com baixa expressão, as restrições seletivas são mais fracas e a utilização de códons seria, então, afetada pela pressão de mutação e deriva genética (Sharp e Li, 1986).

Em *Drosophila*, o desvio na utilização de códons está associado ao conteúdo de G + C na terceira posição do códon. Esse fato foi positivamente correlacionado com o conteúdo de G + C nos introns próximos (Kliman e Hey, 1994). A pressão de mutação foi a explicação sugerida para a ocorrência de mais G + C, uma vez que, tanto em introns quanto na terceira posição dos códons, as restrições funcionais são mais fracas. Nos códons degenerados, a variação do conteúdo G + C parece estar mais correlacionada com a variação na seleção natural, mas o efeito da pressão de mutação não é desprezível. O conteúdo de G + C é menor em pseudogenes que em seus similares funcionais (Moriyama e Gojobori, 1992) e o desvio na utilização de códons, menor em *D. melanogaster* do que em *D. simulans*, é explicado por uma diminuição populacional em *D. melanogaster*. Esses dois fatos sugerem que a seleção para eficiência na tradução tem um importante papel na escolha dos códons em *Drosophila* (Li, 1997).

Os padrões de utilização de códons em mamíferos podem ser explicados de diversas maneiras. A princípio, a pressão mutacional estaria determinando o viés na utilização de códons. O fato de a composição de bases de códons sinônimos não ser significativamente diferente da do DNA não codificante e dos introns ao seu

Tabela 7.2. Utilização relativa de códons sinônimos (RSCU) em três espécies de organismos.

Aminoácido	Códon	*Escherichia coli*		*Saccharomyces cerevisae*		Humano	
		↑*	↓*	↑*	↓*	G+C**	A+T**
Val	GUU	2,41	1,09	2,07	1,13	0,09	1,32
	GUC	0,08	0,99	1,91	0,76	1,03	0,69
	GUA	1,12	0,63	0,00	1,18	0,11	0,80
	GUG	0,40	1,29	0,02	0,93	2,78	1,19
Ile	AUU	0,48	1,38	1,26	1,29	0,45	1,60
	AUC	2,51	1,12	1,74	0,66	2,43	0,76
	AUA	0,01	0,50	0,00	1,05	0,12	0,64
Phe	UUU	0,34	1,33	0,19	1,38	0,27	1,20
	UUC	1,66	0,67	1,81	0,62	1,73	0,80
Met	AUG	1,00	1,00	1,00	1,00	1,00	1,00

Modificada a partir de Li e Graur (1991)

* ↑ = genes com alta taxa de expressão; ↓ = genes com baixa taxa de expressão.

** G+C = maior conteúdo de guaninas e citosinas; A+T = maior conteúdo de adeninas e timinas.

Tabela 7.3. Papel da expressão gênica na determinação do viés na utilização de códons e das taxas de substituições sinônimas em alguns organismos unicelulares.

	Genes	
	Altos níveis de expressão	Baixos níveis de expressão
Seleção para maior eficiência	Alta	Baixa
Quantidade de tRNAs	Restrita	Sem restrição
Viés na utilização de códons	Forte	Fraco
Taxa de substituição sinônima	Baixa	Alta

redor é uma evidência da ação da pressão mutacional. Além disso, os padrões de utilização de códons em mamíferos parecem não estar associados aos níveis ou à velocidade da expressão gênica, nem ao tipo de tecido, como seria de se esperar se a seleção natural estivesse operando (Page e Holmes, 1998). A maioria dos genes humanos possui um alto conteúdo de G + C na terceira posição, mas aparentemente existe apenas uma baixa correlação entre o conteúdo de G + C e o nível de expressão gênica. Entretanto, o conteúdo G + C na região onde se localiza o gene parece ser um dos fatores que determina o desvio na utilização de códons em humanos, ao invés da expressão gênica. Como o conteúdo G + C é maior nas terceiras posições, é possível que o padrão de utilização de códons em humanos esteja de alguma maneira sendo afetado pela seleção natural (Aota e Ikemura, 1986; Li e Graur, 1991). A utilização preferencial de códons encontra-se também discutida no Capítulo 10.

De acordo com Page e Holmes (1998), são três os mecanismos que poderiam estar envolvidos, isoladamente ou combinados, na variação do padrão de ocorrência de mutações no genoma: replicação de DNA, reparo de DNA e recombinação. A hipótese da replicação propõe uma correlação entre o tempo de replicação e a composição de bases. Genes que replicam primeiro apresentam um conteúdo maior de G + C do que os genes de replicação tardia, pois se acredita que os erros ocorridos no início da replicação acarretem uma maior inserção de G + C. A hipótese do reparo é baseada na premissa de que a eficiência do reparo varia ao longo do genoma e que alguns erros são "consertados" de modo mais eficiente que outros (ver detalhes no Capítulo 5). Segundo a hipótese da recombinação, regiões com baixo conteúdo de G + C estão associadas à baixa taxa de recombinação. Análises de sequências de DNA de genes de mamíferos revelaram pouca correlação entre o conteúdo de G + C e taxas de mutação sinônima (Bernardi *et al.*, 1993).

7.1.2. Regiões não codificadoras

Os dados a respeito de substituições de nucleotídeos nas regiões não codificadoras são bem menos abundantes. Em relação ao genoma nuclear, as sequências utilizadas com maior frequência são aquelas não traduzidas localizadas a 5' ou a 3' dos genes. Comparando-se as taxas de substituições de nucleotídeos entre essas sequências e os sítios quatro vezes degenerados, onde todas as substituições são sinônimas, os valores são cerca de 50% maiores nos últimos (Li, 1997).

Os pseudogenes são sequências derivadas—na maioria das vezes por duplicação—de genes funcionais em que uma das cópias perdeu a função devido à ocorrência de mutações que a inativaram. Por exemplo, foram detectados pseudogenes nas famílias gênicas α e β-globina, cuja homologia com os genes funcionais correspondentes variam de 75 a 80% (Ratner *et al.*, 1996). Como os pseudogenes não exercem qualquer função, não estão sujeitos às restrições funcionais, razão pela qual se espera que apresentem altas taxas evolutivas. Por exemplo, a taxa de substituição no pseudogene da globina ψα1 é maior do que nas três posições do gene funcional da α-globina, do qual é derivado (Nei, 1987). Outra região não codificadora, desta feita pertencente ao genoma mitocondrial, que tem sido muito estudada é a alça D ("*D-loop*", em inglês) ou região controladora (Lee *et al.*, 1995; Randi e Lucchini, 1998). A região controladora será discutida com mais detalhe no item a respeito de taxas de substituição de nucleotídeos em organelas.

7.1.3. Genoma de Organelas

Nem todo o DNA dos eucariontes está armazenado no núcleo. Há outras estruturas celulares, chamadas organelas, que contêm DNA próprio. A mitocôndria é uma organela encontrada na maioria dos eucariontes e sua função nas células envolve a degradação de açúcares e gorduras importantes na respiração celular. O cloroplasto, organela responsável pela fotossíntese, presente em algas e plantas terrestres, também apresenta seu próprio genoma (Capítulo 2).

O genoma de organelas constitui-se de genes que normalmente estão presentes em cópia única. Tanto os genes das mitocôndrias como os dos cloroplastos somente codificam as proteínas necessárias para a manutenção de suas funções e para a expressão gênica. Nessas organelas, as regiões não codificadoras são raras.

Características peculiares, como a herança uniparental, geralmente materna, a ausência de recombinação e as altas taxas evolutivas, quando comparado ao genoma nuclear, fizeram do DNA mitocondrial (mtDNA) uma ferramenta importante no estudo das relações evolutivas entre indivíduos, espécies e populações. Até o momento, vários genomas mitocondriais completos foram sequenciados e um grande número de sequências parciais, principalmente do gene do citocromo *b* e da região controladora, estão disponíveis em bancos de dados. Essas sequências permitiram a determinação mais precisa das taxas de substituição de nucleotídeos nos genomas de organelas.

A taxa de substituições sinônimas no genoma mitocondrial de vertebrados foi estimada em 5,7 x 10^{-8}/sítio/ano (Brown *et al.*, 1982). Esse valor é cerca de dez vezes maior que o encontrado no genoma nuclear. Em relação às substituições não sinônimas, há grande variação entre os 13 genes que codificam para proteínas, dependendo das restrições funcionais, a exemplo dos genes nucleares. A alta taxa de substituição de nucleotídeos pode ser devida a uma alta taxa de mutação que, por sua vez, poderia estar sendo causada pelo excesso de resíduos metabólicos, pela baixa fidelidade na replicação das mitocôndrias e pela ausência de um mecanismo de reparo (Li e Graur, 1991; Li, 1997).

Pesole *et al.* (1999), analisando a variabilidade da taxa de substituição de nucleotídeos em diferentes regiões funcionais do mtDNA de mamíferos, verificaram que existe uma grande variabilidade em termos da dinâmica evolutiva da molécula. A taxa de substituição de nucleotídeos varia dependendo da região considerada. Os sítios não sinônimos, o domínio central da região controladora e os genes que codificam para tRNAs e rRNAs apresentam taxas mais baixas do que os sítios sinônimos e do

que as duas regiões periféricas da região controladora. As taxas de substituições nos sítios sinônimos podem ser consideradas uniformes em todos os genes mitocondriais (Pesole *et al.*, 1999).

Para comparar adequadamente as taxas de substituição nucleotídicas entre genes nucleares e mitocondriais, os mesmos pares de espécies devem ser utilizados. Além disso, nos genes com taxas de substituição muito elevadas, é necessário utilizar pares de espécies que divergiram recentemente para evitar que os sítios estejam saturados. Na Tabela 7.4, são apresentadas as porcentagens de substituição por sítio no mtDNA e no DNA nuclear (nDNA).

Para os genes que codificam proteínas, os sítios sinônimos apresentam taxas de substituição 22 vezes superiores no mtDNA, se comparadas às do nDNA. Nos sítios não sinônimos, observa-se uma grande variabilidade em ambos os genomas. No genoma nuclear, as taxas variam de 0,0 a 4,3%, enquanto que, no mtDNA, a variação é de 0,6 a 4,5%. A alta taxa de substituição de nucleotídeos nos tRNAs e nos sítios sinônimos no mtDNA poderia ser explicada pelas baixas restrições funcionais ocasionadas pela instabilidade do emparelhamento códon/anticódon que ocorre no genoma mitocondrial (Pesole *et al.*, 1999). As baixas taxas verificadas no DNA nuclear, quando comparadas às do mtDNA, podem estar correlacionadas a restrições devidas à presença de *isócoros* no genoma nuclear (Alvarez-Valin *et al.*, 1998). Isócoros são sequências com mais de 300 kb que possuem um conteúdo de G + C relativamente uniforme, que estão espalhadas pelo genoma (veja o Capítulo 10 para mais detalhes).

Wolfe e colaboradores (1987, 1989), estudando várias dicotiledôneas e monocotiledôneas, estimaram a taxa de substituição de nucleotídeos em 31 genes representando os três genomas (mitocondrial, nuclear e do cloroplasto) presentes em plantas superiores. A taxa de substituição em genes presentes nos cloroplastos (1,1-2,9 $x10^{-9}$) são mais altas do que nos mitocondriais (0,2-1,1 $x10^{-9}$). Gaut *et al.* (1992) obtiveram os mesmos resultados, desta feita comparando taxas de substituição entre arroz e milho. É interessante ressaltar que as mono e seus grupos mais aparentados entre as dicotiledôneas divergiram há 200 milhões de anos, enquanto que o tempo de divergência entre o arroz e o milho é de cerca de`50 milhões de anos. A similaridade entre os dois estudos sugere que o padrão de substituição de nucleotídeos em plantas tem sido mantido. Em uma revisão recente, Muse (2000) sugere que vários fatores devem ser levados em consideração para explicar o padrão de substituição de nucleotídeos em plantas. Entre eles, podemos citar: (1) os três genomas, assim como diferentes genes em cada um deles, apresentam taxas bastante variáveis de substituições sinônimas e não sinônimas; (2) taxas relativas de substituições não sinônimas são influenciadas por fatores específicos para cada loco gênico; (3) taxas relativas de substituições sinônimas podem ser conservadas na maiorias dos genes de cloroplastos e podem ser mantidas nos três genomas; (4) susbtituições sinônimas e não sinônimas possuem dinâmicas geralmente independentes.

Em conclusão, pelo que foi apresentado, existe uma grande variação nas taxas de mutações sinônimas e não sinônimas. Nas regiões codificadoras, a variação nas taxas evolutivas está inversamente relacionada às restrições funcionais. Nas regiões não codificadoras, as taxas evolutivas de maneira geral são mais elevadas, pois não existe qualquer tipo de restrição funcional. O aumento na taxa de mutação é uma das explicações sugeridas para a taxa de substituição de nucleotídeos ser cerca de dez vezes maior no genoma mitocondrial que no genoma nuclear.

7.2. O Relógio Molecular

Comparando-se a quantidade de substituições de nucleotídeos em algumas proteínas, como, por exemplo, a hemoglobina, e o tempo de divergência de algumas espécies, percebe-se que, no exame de táxons cada vez mais distantes, o número de substituições aumenta de forma quase constante. A partir dessas observações, algumas proteínas passaram a ser utilizadas como uma espécie de "cronômetro" para estimar-se o tempo de divergência entre os táxons. As inúmeras tentativas de entender como esse cronômetro funciona e o quão bem calibrado ele é vêm proporcionando um melhor entendimento a respeito da evolução das sequências de nucleotídeos e aminoácidos.

Nessa segunda parte do capítulo, trataremos especificamente de como se dá a evolução das sequências de proteínas entre os táxons. No início dos anos 60, Zuckerkandl e Pauling (1962, 1965) e Margoliash (1963), em estudos comparativos com a hemoglobina e com o citocromo *c*, respectivamente, perceberam que a taxa de substituição de aminoácidos era quase a mesma entre diversas linhagens de mamíferos. A existência de um relógio molecular, isto é, taxas de evolução molecular aproximadamente constantes através do tempo em todas as linhagens para uma dada proteína, foi proposta por Zuckerkandl e Pauling em 1965. Essa talvez seja uma das idéias mais controversas na área de evolução molecular. É importante frisar que o conceito do relógio molecular não significa que as taxas evolutivas em todos os genes e proteínas sejam as mesmas. Como já foi discutido, as taxas de substituição de nucleotídeos variam grandemente entre genes e entre porções do mesmo gene, devido principalmente à alteração nas restrições funcionais. Além disso, diferentes genomas apresentam taxas evolutivas diferentes. No genoma mitocondrial de vertebrados, por exemplo, as substituições de nucleotídeos ocorrem cerca de dez vezes mais rápido do que no genoma nuclear.

De qualquer maneira, a hipótese do relógio molecular deu um grande impulso na utilização de macromoléculas em estudos evolutivos, pois, se existe realmente uma taxa de evolução com alguma constância, as macromoléculas podem ser utilizadas para datar o tempo de divergência entre espécies e grupos de espécies.

Tabela 7.4. Comparação entre a taxa de substituições por sítio no DNA mitocondrial (mtDNA) e no DNA nuclear (nDNA) entre as espécies: humano -chimpanzé (*Hsa* x *Ptr*) e rato-camundongo (*Rno* x *Mmu*).

	Sítio	mtDNA	nDNA	mtDNA/nDNA
Hsa x *Ptr*	Sinônimo	34,6 ± 3,9	1,6 ± 0,9	22
	Não-sinônimo			
	média	2,6 ± 0,4	0,8 ± 0,2	3
	mínima	0,6 ± 0,3	0,0	--
	máxima	4,5 ± 1,6	4,3 ± 2,2	1
Rno x *Mmu*	12s rRNA	7,7 ± 2,4	0,4 ± 0,3	19
	16s rRNA	17,2 ± 3,8	4,1 ± 0,8	4
	tRNAs	9,7 ± 2,4	0,1 ± 0,1	97

Modificada a partir de Pesole *et al* (1999).

O tempo de divergência entre duas espécies é justamente um dos principais focos da controvérsia em relação à hipótese do relógio molecular.

Uma das maneiras mais simples de testar o relógio molecular é utilizar o teste de taxa relativa (*relative rate test*, em inglês), desenvolvido por Sarich e Wilson (1973). O teste estima o número de substituições entre dois táxons proximamente relacionados e os compara a um terceiro, filogeneticamente mais distante. A principal vantagem dessa abordagem é que não se faz necessário saber os tempos de divergência entre as espécies em questão. Examinando a Figura 5.4, é possível perceber que o número de substituições entre as espécies 1 e 2 (K_{12}) é igual à soma das substituições que ocorreram do ponto 0 até 1 (K_{01}) e de 0 até 2 (K_{02}). Portanto,

$$K_{12} = K_{01} + K_{02}.$$

O mesmo pode ser calculado em relação aos outros dois pares de espécies. Como K_{12}, K_{13} e K_{13} podem ser estimados a partir da frequência de nucleotídeos, os valores de K_{01}, K_{02} e K_{03} podem ser obtidos através das equações

$$K_{01} = (K_{13} + K_{12} + K_{23})/2,$$
$$K_{02} = (K_{12} + K_{23} + K_{13})/2,$$
$$K_{03} = (K_{13} + K_{23} + K_{12})/2.$$

Como o tempo desde a separação de 1 e 2 é igual para as duas linhagens, de acordo com a hipótese do relógio molecular, K_{01} e K_{02} devem ser iguais, ou seja,

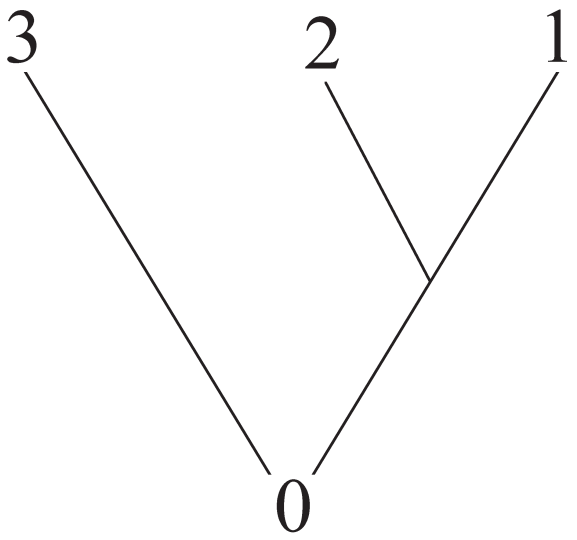
$$d = K_{01} - K_{02},$$

e *d* não dever ser estatisticamente diferente de 0. Para testar se um valor observado de *d* é ou não significativamente diferente de 0, basta compará-lo com o desvio padrão. Se, por exemplo, o valor obtido para *d* for duas vezes maior que o desvio padrão, pode ser considerado significativo no nível de 5% (Li e Graur, 1991; Li, 1997).

Esse é um teste simples que poder ser utilizado mesmo quando não se tem certeza da filogenia das espécies. Nesse caso, as equações acima podem ser utilizadas para estimar o comprimento dos ramos e considera-se a espécie com o maior comprimento de ramo como o grupo externo (Li, 1997). Entretanto, é importante ter o cuidado de não utilizar como grupo externo táxons muito distantes, porque isso aumenta as chances de mais de uma mutação ter ocorrido no mesmo sítio (Page e Holmes, 1998).

Jermiin (1996) desenvolveu o programa K2WuLi Version 1.0, que executa o teste de taxa relativa e estima a distância evolutiva e o desvio padrão entre sequências de DNA utilizando o modelo Kimura 2 parâmetros. Outros autores propuseram diferentes testes para a hipótese do relógio molecular (Tajima, 1993; Muse e Weir, 1992; Rambaut e Bromham, 1998).

Figura 7.4. O teste de taxa relativa. Para comparar as taxas de substituição que deram origem às espécies 1 e 2 a partir de um ancestral comum 0, um grupo externo 3 foi utilizado.



7.2.1. Relógios locais

Uma vez que a hipótese do relógio molecular é tão controversa, outra questão interessante é saber se as taxas de substituição de nucleotídeos se comportam de acordo com a hipótese do relógio molecular em relação a um grupo de organismos. Os ratos e camundongos são apropriados para testar essa hipótese, pois possuem aspectos fisiológicos semelhantes. Existem duas linhas de evidências que apontam para a existência de um relógio local nesse grupo de indivíduos. Dados obtidos a partir de experimentos de hibridação DNA-DNA revelaram uma taxa constante entre táxons das famílias Muridae (que compreende ratos e camundongos) e Microtidae (que compreende hamsters e lemingues –Brownell, 1983; Catzeflis *et al.*, 1987). Li *et al.* (1987), aplicando o teste de taxa relativa, encontraram taxas muito semelhantes entre ratos e camundongos. Examinando a Tabela 7.5, percebemos que o número de substituições sinônimas (K_S) entre camundongo e hamster e entre rato e hamster são quase iguais e a diferença entre as duas não é estatisticamente significativa (0,8 < desvio padrão). Já em relação aos sítios não sinônimos (K_A), a diferença pode ser considerada significativa, pois é cerca de duas vezes maior que o desvio padrão. Quando comparamos as linhagens utilizando a espécie humana como referência, as diferenças encontradas não são estatisticamente significativas. Essa é uma evidência de que esse grupo de roedores apresenta taxas de substituição de nucleotídeos quase iguais, pelo menos em sítios sinônimos, corroborando a hipótese de um relógio molecular local (O'hUigin e Li, 1992; Li, 1997).

7.2.2. Hipóteses sobre a diminuição da taxa evolutiva em hominídeos

Goodman (1961) e Goodman *et al.* (1971) postularam a hipótese de que a taxa de evolução molecular sofreu uma diminui-

Tabela 7.5. Porcentagem de substituições sinônimas (K_S) e não sinônimas (K_A) entre espécies de roedores.

	Sítios sinônimos		Sítios não sinônimos	
Espécies	K_S	L^b	K_A	L^b
Rato-camundongo	18,0 ± 0,7	4.229	1,8 ± 0,1	15.217
Camundongo-hamster	30,3 ± 1,0	4.229	2,9 ± 0,1	15.217
Rato-hamster	31,3 ± 1,0	4.229	2,7 ± 0,1	15.217
Camundongo-homem	53,4 ± 1,5	4.229	5,2 ± 0,2	15.217
Rato-homem	51,6 ± 1,5	4.229	5,0 ± 0,2	15.217
Hamster-homem	52,3 ± 1,5	4.229	5,1 ± 0,1	15.217

Modificada a partir de O'hUigin e Li (1992).
L^b = número de sítios comparados.

ção em hominídeos (humanos e símios) após sua separação dos macacos do velho mundo. Essa hipótese foi baseada em dados de distâncias imunológicas e sequências de proteínas. Sarich e Wilson (1967) e Wilson *et al.* (1977), aplicando o teste de taxa relativa, não encontraram evidências dessa diminuição e atribuíram os resultados encontrados por Goodman à utilização de tempos de divergência errôneos entre símios e humanos. Kohne *et al.* (1972) e Sibley e Ahlquist (1984) chegaram a conclusões similares às de Sarich e Wilson, mas, em trabalhos mais recentes, os resultados são conflitantes (Sibley e Ahlquist, 1987).

Resultados obtidos através de sequenciamento de DNA apóiam a hipótese da diminuição das taxas evolutivas em hominídeos (Koop *et al.*, 1986; Li e Tanimura, 1987; Li *et al.*, 1987). Mais recentemente, Li *et al.* (1996) obtiveram dados de sequenciamento de introns, regiões flanqueadoras e do pseudogene η da globina, indicando que a taxa de substituição de nucleotídeos é muito superior em macacos do velho mundo do que em hominídeos após a separação das duas linhagens.

Gilloly e colaboradores (2005) sugerem que as diferenças no tamanho corporal poderiam explicar a redução da evolução molecular em hominídeos. Segundo esses autores, a diferença de massa corpórea média entre hominóides (50 kg) e macacos do velho mundo (7 kg) resulta numa diminuição de 0,6 vezes nas taxas de substituição de nucleotídeos, que é muito próxima do valor estimado de 0,7 (Seino *et al.,* 1992).

7.2.3. Causas das diferenças nas taxas de substituição de nucleotídeos entre linhagens

A variação nas taxas de substituição de nucleotídeos pode estar relacionada a diversos fatores, como diferenças nos tempos de geração (Laird *et al*., 1969), nas taxas metabólicas (Martin & Palumbi, 1993), na eficiência dos mecanismos de reparo verificados entre linhagens e na temperatura do ambiente (Bleiweiss, 1998; Wright *et al.,* 2003). A correlação entre as taxas evolutivas em linhagens próximas é biologicamente possível, pois os fatores mencionados acima podem ser similares nas mesmas linhagens e alterações graduais nesses fatores promoveriam a divergência gradual nas taxas de substituição de nucleotídeos (Thorne *et al.*, 1998).

A hipótese do tempo de geração postula simplesmente que a taxa de mutação deve ser medida em termos de gerações, e não de tempo cronológico, assumindo que os erros ocorridos durante a replicação do DNA que origina os gametas seriam a principal fonte de mutação. No nível molecular, o tempo de geração é definido como o tempo que existe entre cada uma das replicações do DNA que originam gametas. Se a maioria das mutações é resultado de erros nesse processo e se espécies relacionadas possuem aproximadamente o mesmo número de divisões celulares, espécies com tempos de geração mais longos acumulariam menos substituições que as com tempos mais curtos. Gu e Li (1992) aplicaram o teste de taxa relativa para comparar sequências de proteínas de roedores e humanos utilizando galinhas como referência. Das 54 proteínas comparadas, 35 apresentaram taxas mais altas nos roedores, 12 em humanos e sete apresentaram taxas semelhantes. Eastel e Collet (1994), comparando as taxas de substituições de nucleotídeos em humanos e em roedores, usando marsupiais como referência, concluíram que as taxas de substituições sinônimas são iguais nas duas linhagens, enquanto que as taxas de substituições não sinônimas são significativamente maiores em roedores que em humanos (Tabela 7.5). Li *et al.* (1996) argumentam que, como as distâncias entre os genes de marsupiais e humanos ou roedores é grande, esses números devem ser considerados com cuidado. Ao reexaminarem os dados de Eastel e Collet, desta feita levando em consideração a taxa de transversões, ao invés da taxa de transições, Li e colaboradores (1996) concluíram que os resultados apresentados na Tabela 7.6 indicam uma taxa maior em roedores tanto para substituições sinônimas como não sinônimas. Ohta (1995) também encontrou taxas de mutações sinônimas 2,6 vezes maior em roedores do que em primatas.

Outras evidências a favor da hipótese do tempo de geração foram obtidas em plantas. Análises realizadas com o gene *rbc*L de cloroplastos mostrou que as taxas de substituição em gramíneas são cinco vezes maiores do que em palmeiras. Isso é compatível com as idades quando da primeira floração, utilizadas como medida do tempo de geração (Gaut *et al.*, 1992).

Nem sempre a correlação entre o tempo de geração e a taxa de substituição de nucleotídeos é fácil. Por exemplo, se a quantidade de divisões celulares for diferente em táxons com tempos de geração semelhantes, as taxas de substituição de nucleotídeos também serão diferentes. Isso explicaria porque as taxas de mutação não são tão diferentes quanto seria de se esperar quando comparadas aos tempos de geração. Os roedores, por exemplo, não apresentam taxas 40 vezes superiores àquelas dos primatas.

Como a hipótese do tempo de geração assume que os erros durante o processo de replicação do DNA são a principal fonte de mutações, ela pode ser testada comparando-se a razão da taxa de mutação em machos e fêmeas (α) e a razão entre a quantidade de divisões celulares na linhagem germinativa de machos e fêmeas (c). Li *et al.* (1996) estimaram o valor de $c = n_m/n_f$, onde n_m é o número de divisões na linhagem germinativa de machos e n_f nas fêmeas. As células da linhagem germinativa de fêmeas de ratos e de mulheres sofrem aproximadamente o mesmo número de divisões, 28 e 33, respectivamente. Já em homens e ratos, os números aumentam para 205 e 57, respectivamente. Esse aumento no número de divisões poderia estar contribuindo para diluir as diferenças nas taxas de mutação entre as duas espécies. A hipótese do tempo de geração também prevê que a taxa de mutação deveria ser maior nos machos que nas fêmeas, porque o número de divisões celulares durante a espermatogênese é maior que durante a oogênese. Shimmin *et al.* (1993), para testar a influência dos sexos nas taxas de evolução, estudaram uma proteína que apresentava o *motivo* "dedo de zinco" (*zinc finger*, em inglês). Motivo é um termo emprestado da música que significa uma pequena frase musical repetida com variações no transcorrer de uma peça. No estudo de sequências macromoleculares, esse termo tem um significado parecido, aplicado a trechos de sequências semelhantes, porém não idênticos, que ocorrem diversas vezes nos genomas. O "dedo de zinco" trata-se de uma sequência de aminoácidos

Tabela 7.6. Comparação da quantidade de transições (A) e transversões (B) em sítios não degenerados (ND), duas vezes degenerados (2X) e quatro vezes degenerados (4X) entre humanos, roedores e marsupiais.

	ND		2X		4X	
Táxons	A	B	A	B	A	B
Homem-roedor	4,4 ± 0,3	4,3 ± 0,3	27,4 ± 1,6	5,2 ± 0,5	25,8 ± 1,8	20,6 ± 1,4
Homem-marsupial	7,3 ± 0,4	8,7 ± 0,4	48,9 ± 2,8	8,5 ± 0,7	44,2 ± 3,6	45,7 ± 2,8
Roedor-marsupial	7,6 ± 0,4	9,7 ± 0,4	54,4 ± 3,2	9,3 ± 0,7	40,6 ± 3,6	52,4 ± 3,2

Dados de Eastel e Collet (1994), tabela modificada a partir de Li *et al* (1996).

presente em diversas proteínas, que se liga a um átomo de zinco e que apresenta alta afinidade com ácidos nucléicos. Esses autores compararam os últimos introns dos genes para essa proteína nos cromossomos Y e X (genes ZFY e ZFX) em algumas espécies de primatas e verificaram que a razão das taxas de substituição de nucleotídeos entre machos e fêmeas (definida como α) era igual a seis. Outros autores estimaram os valores de α para outras espécies de macacos do velho e do novo mundos e encontraram $\alpha = 4,2$. Portanto, α deve variar de 3 a 6 em primatas. Chang *et al.* (1994) sequenciaram os últimos introns dos mesmos genes de ratos e camundongos e encontraram $\alpha = 1,8$ e $\alpha = 2,0$ para os genes e pseudogenes *Ube1* ligados aos cromossomos X e Y (Chang e Li, 1995). Os valores de *c* são similares aos de α tanto para ratos e camundongos como para primatas, o que sugere que os erros na replicação de DNA são realmente a principal fonte de mutação.

Um problema com a hipótese do tempo de geração é que nem sempre existe correlação entre o tempo de geração e a taxa de substituição de nucleotídeos. Em primatas, por exemplo, a taxa de substituições não sinônimas apresenta uma melhor correlação com o tempo de geração do que a taxa de substituições sinônimas (Ohta, 1995).

Quando examinamos o mtDNA de alguns mamíferos, répteis e peixes que apresentam tempos de geração relativamente curtos e taxas de substituição baixas, percebemos desvios ainda maiores em relação à hipótese do relógio molecular. Esses desvios impulsionaram o surgimento de hipóteses para explicar o papel de outras variáveis no andamento das substituições de nucleotídeos. É importante salientar que organismos que apresentam diferenças em tempo de geração atuais podem não ter tido essa mesma diferença ao longo de todo o tempo evolutivo nas linhagens que as originaram. Evidentemente, o ancestral comum das espécies consideradas poderia ter um tempo de geração mais curto ou mais longo e as linhagens resultantes poderiam ter apresentado quase nenhuma diferença com relação a esse caráter por um longo período de tempo.

É possível encontrar na literatura muitos trabalhos testando se o relógio molecular segue o tempo cronológico ou de geração. Uma conclusão aparentemente geral é que, pelo menos em sítios sinônimos e nas regiões não codificadoras, o relógio parece funcionar melhor em termos de número de gerações.

De acordo com a hipótese das taxas metabólicas, as diferenças nas taxas metabólicas explicam melhor do que as diferenças nos tempos de geração a variabilidade encontrada nas taxas de evolução molecular. Organismos com alta taxa metabólica apresentariam taxas de evolução superiores aos organismos com baixas taxas metabólicas. Martin *et al.* (1992), examinando as taxas de substituições de nucleotídeos em sequências do citocromo *b* e da subunidade 1 da oxidase do citocromo *c* em 13 espécies de tubarões pertencentes a duas ordens, verificaram que elas são de 7 a 8 vezes menores que em primatas ou ungulados. Os autores sugerem que a baixa taxa metabólica dos tubarões estaria relacionada com a baixa taxa de mutações. Martin e Palumbi (1993), comparando os resultados obtidos com a análise do pseudogene da globina com as medidas corpóreas médias dos exemplares, verificaram que, quanto maior o tamanho dos indivíduos adultos de diversas espécies de vertebrados, menor era a taxa de substituição de nucleotídeos. Análises do gene do citocromo *b* do mtDNA de mamíferos e das taxas de substituições sinônimas obtidas com calibrações a partir de fósseis mostraram a mesma tendência. Ao examinar dados provenientes de análises com enzimas de restrição, os resultados obtidos foram semelhantes e dois grandes grupos foram evidenciados: pecilotérmicos e homeotérmicos. De maneira geral, a taxa de divergência do mtDNA é maior nos homeotérmicos do que nos pecilotérmicos. Aparentemente, a taxa metabólica também poderia explicar algumas exceções da hipótese do tempo de geração. Por exemplo, as baleias apresentam uma taxa de substituição de nucleotídeos mais baixa que os primatas, apesar do menor tempo de geração (Baker *et al.*, 1993). A Tabela 7.7 apresenta uma comparação entre as taxas de substituição sinônima, taxas metabólicas e tempo de geração em algumas espécies de primatas.

Examinando a Tabela 7.7, percebe-se que as taxas evolutivas estão indiretamente relacionadas com o tamanho e diretamente relacionadas com a taxa metabólica. Por exemplo, macacos-aranha, que pesam cerca de 7 Kg, apresentam taxas mais elevadas de substituição de nucleotídeos, enquanto que, nas espécies de maior porte, as taxas de substituição são menores, como em gorilas e no homem. No macaco-aranha e no gênero *Macaca*, as taxas metabólicas são elevadas, 415 e 450 ml de O_2/kg/hora, respectivamente, e as taxas de substituição de nucleotídeos também são altas.

A hipótese da taxa metabólica baseia-se no fato de que organismos com altas taxas metabólicas apresentam altas taxas de síntese de DNA e, portanto, teriam taxas de mutação superiores às dos organismos com baixas taxas metabólicas. As duas fontes de mutação—erros durante a replicação e danos produzidos por radicais de oxigênio—estão correlacionadas. O oxigênio produzido pelo metabolismo é altamente reativo e pode provocar danos no DNA, que, por sua vez, aumentam a taxa de reparo e a quantidade de erro por replicação. Essa hipótese é plausível em termos de mtDNA, pois cerca de 95% do oxigênio molecular metabolizado é processado nas mitocôndrias (Richter *et al.*, 1988). Esse fato poderia ser utilizado também para explicar a maior taxa de substituição de nucleotídeos no mtDNA, se comparado ao DNA nuclear (Martin e Palumbi, 1993). Entretanto, ainda não foi mostrado se o dano oxidativo exerce alguma influência no DNA nuclear, uma vez que o consumo de oxigênio verificado no núcleo é baixo.

Existem algumas exceções para a hipótese da taxa metabólica. As aves aparentemente apresentam uma taxa de substituição de nucleotídeos mais baixa que os mamíferos, apesar de possuírem taxas metabólicas mais altas (Mindell *et al.*, 1996). Vários fatores poderiam estar influenciando as taxas de evolução

Tabela 7.7. Taxas de substituição sinônima (subs/sítio/10^9 anos) no DNA nuclear, tamanho médio (kg), taxas metabólicas específicas (ml de O_2/kg/hora) e tempo de geração (anos) em algumas espécies de primatas.

Espécies	Taxa de substituição	Tamanho	Taxa metabólica	Tempo de geração
Chimpanzé	1,2	62	220	9
Gorila	1,2	155	200	9,4
Orangotango	1,2	55	230	12
Gibão	1,7	10,5	370	9,3
Macaco-aranha	1,9	7	415	4,6
Macaca	1,8	4	430	3
Homem	1,1	65	210	17

Modificada a partir de Martin e Palumbi (1993).

molecular em aves, entre elas o baixo conteúdo de DNA celular e a alta temperatura corporal, ambas características únicas das aves. Stanley e Harrison (1999), comparando sequências do citocromo *b* de aves e mamíferos, mostraram que a razão d_n/d_s varia tanto em aves como em mamíferos, mas é sempre mais baixa em aves. Essa seria a primeira evidência de restrição seletiva operando em proteínas de aves. Alguns autores sugerem que enzimas que funcionam a temperaturas mais altas toleram menos substituições de nucleotídeos. Stanley e Harrison (1999) sugerem que, uma vez que as taxas evolutivas são determinadas pela taxa de mutação e pelas taxas de substituição (fixação da mutação), é possível que as taxas de mutação sejam realmente mais altas em aves que em mamíferos, mas o aumento da restrição funcional diminuiria a taxa de substituição, diminuindo, portanto, de maneira geral, as taxas de evolução molecular em aves.

Segundo a hipótese do tempo de geração, os erros cometidos durante o processo de replicação são a principal fonte de mutação. Entretanto, é possível que as substituições de nucleotídeos sejam devidas a falhas nos mecanismos de reparo. Se a eficiência de reparo variar entre as espécies, aquelas que apresentam mecanismos mais eficientes apresentarão taxas de substituição de nucleotídeos mais baixas (Page e Holmes, 1998; veja também o Capítulo 5).

Atualmente tem sido muito discutido o papel da eficiência dos mecanismos de reparo nas taxas e padrões de substituição de nucleotídeos. A hipótese da eficiência dos mecanismos de reparo foi proposta por Britten (1986) para explicar as diferenças observadas entre linhagens. Alguns dados corroboram essa hipótese. Primeiro, há evidências a favor de uma conexão entre a maquinaria de transcrição de DNA e aquela responsável pelo reparo, de maneira que os genes que são muito transcritos são reparados de modo mais eficiente. Segundo, a composição de bases e a taxa de substituição em sítios sinônimos em genes de mamíferos tendem a ser específicas para genes, ao invés de específica para espécies. Isso talvez signifique que genes homólogos são transcritos e reparados de modo similar (Page e Holmes, 1998). Acredita-se que os mecanismos de reparo tenham uma influência limitada nas taxas de substituição de nucleotídeos, pois é improvável que espécies proximamente relacionadas apresentem taxas diferentes de substituição devido a diferenças nos mecanismos de reparo.

Durante a década de 1970, com o acúmulo de dados comparativos entre sequências de proteínas, começou a ficar evidente que a variação nas taxas de substituição de aminoácidos em várias proteínas era inconsistente com a distribuição de Poisson. Essa distribuição resulta de fenômenos raros, tais como o que ocorre nas substituições. No caso de substituições, é esperado que vários nucleotídeos não apresentem substituições, alguns apresentem uma única, poucos com substituições duplas, menos com três e assim por diante. Uma das características dessa distribuição é a existência de uma média pequena e de uma variância com um valor próximo ao da média. O índice de dispersão (R_t), que mede se a quantidade de variação existente entre linhagens, é maior que a esperada segundo a distribuição de Poisson, pode ser calculado dividindo-se a variância pela média de substituições entre linhagens. Quando $R_t = 1$, a taxa de substituição de nucleotídeos é consistente com a distribuição de Poisson, caso contrário, quando $R_t > 1$, ocorre um desvio com relação à distribuição de Poisson.

Gillespie (1989) e Ohta (1995) calcularam o valor de R para substituições sinônimas e não sinônimas em cerca de 50 genes. Os resultados obtidos sugerem que a variação nas taxas de substituição sinônimas entre linhagens é devida a diferenças nos denominados *efeitos de linhagem*, ou seja, na eficiência dos mecanismos de reparo, nas taxas metabólicas e nos tempos de geração. As substituições não sinônimas, além dos fatores mencionados acima, estão sujeitas a efeitos episódicos.

Ohta (1993, 1995) propôs que a maioria das substituições não sinônimas não é perfeitamente neutra, mas levemente deletéria e sugere a utilização da razão entre as taxas de substituições sinônimas e não sinônimas para testar a hipótese de mutação quase nula. Se a maioria dos mutantes quase neutros forem levemente deletérios, a hipótese prevê que a probabilidade de fixação dos alelos vai depender do tamanho da população, devido à atuação de dois fatores concomitantemente: a seleção natural negativa e a deriva genética. Em populações grandes, a probabilidade de fixação é menor, pois o efeito da seleção negativa é maior do que o da deriva genética. O contrário é esperado em populações pequenas. Isso pode explicar por que o relógio molecular em sítios não sinônimos é mais coerente com o tempo real do que com tempos de geração. Espécies que vivem em grandes populações tendem a ter tempos de geração e tamanho corporal menores do que aquelas que vivem em pequenas populações.

É importante ter em mente que uma distinção entre as hipóteses do tempo de geração e das taxas metabólicas é muito complicada, pois tanto a produção de radicais livres como o tempo de geração variam com a taxa metabólica, que por sua vez varia com o tamanho e temperatura corporais (Gillooly *et al*., 2005).

7.2.4 Um novo modelo: efeitos da temperatura e tamanho corporais

Gillooly e colaboradores (2005) propõem um novo modelo de substituição nucleotídica aliando a teoria a respeito da taxa metabólica com a teoria neutralista da evolução molecular. A taxa metabólica representa a taxa com que a energia e a matéria são retiradas do ambiente por um organismo e utilizadas para seu crescimento manutenção e reprodução. Acredita-se que a taxa metabólica influencia grande parte dos processos biológicos, inclusive os geradores de mutação: produção de radicais livres e o tempo de geração. De acordo com o modelo, a taxa metabólica de massa específica (B) varia com o tamanho corporal e a temperatura e pode ser descrita como:

$$B = b_o M^{-1/4} e^{-E/kT},$$

onde b_o é um coeficiente independente de tamanho e temperatura corporal, $M^{-1/4}$ é termo relacionado ao tamanho e $e^{-E/kT}$ é o fator de Boltzmann ou Arrhenius, relacionado à dependência da taxa metabólica na temperatura (Gillooly *et al*., 2001).

Segundo os autores, quando combinada com premissas da teoria neutralista (Kimura, 1968), essa equação pode ser usada para caracterizar as taxas de evolução molecular. A primeira premissa é que as mutações neutras, que são fixadas ao acaso nas populações, resultando, portanto, em substituições de nucleotídeos, seriam a principal causa da evolução molecular. De acordo com essa premissa, a taxa de substituição de nucleotídeos é igual à taxa de mutações neutras por geração, independentemente do tamanho populacional. A outra premissa é que a taxa de substituição de nucleotídeos é proporcional a B, assumindo que a maioria das mutações seja resultado de uma combinação de processos relacionados ao metabolismo. Gillooly e colaboradores (2005) defendem a existência de um único relógio molecular, mas que funciona em uma taxa constante de substituição nucleotídica por unidade de energia metabólica de massa específica, ao invés de por unidade de tempo. Por exemplo, os autores sugerem que a diferença na temperatura—15°C para calibragem do relógio de pecilotérmicos, enquanto que, nos peixes nototenióides da Antártica, seria 0°C— poderia ser responsável pela grande discrepância entre estimativas geológica e molecular da idade desses peixes, 38 milhões de anos (Ma) x 11 Ma, respectivamente.

5.2.5 Ritmos da evolução molecular

A constância relativa do relógio molecular foi um dos ícones do estudo da evolução molecular por cerca de 40 anos. Recentemente, surgiram indícios de que a taxa de evolução molecular acelera quando é medida em escalas de tempo curtas (Penny, 2005). Ho e colaboradores (2005), em um estudo analisando dados de primatas e aves, demonstraram que essa aceleração realmente ocorre. A taxa é mais alta entre gerações e decresce em um contínuo entre populações locais, depois em populações dispersas, até atingir um platô nos tempos evolutivos longos. A explicação sugerida para esse fenômeno tem por base o quão deletérias são as mutações: pouquíssimo deletérias, pouco deletérias, ou deletérias, mas não letais. Não se espera que as mutações deletérias sejam fixadas nas populações, mas elas podem persistir por longo tempo, dependendo do quão deletérias elas possam ser. Portanto, à medida que o tempo de observação diminui, a proporção de mutações pouco deletérias, que ainda serão eliminadas, é maior. As mutações deletérias (mas não letais) são observadas apenas nos estudos de pedigree (mais curtos) (Ho *et al*., 2005; Penny, 2005). Recentemente, estudos detectaram que as taxas moleculares são cerca de uma ordem de magnitude maior em escalas de tempo genealógica inferiores a 1 Ma que em escalas geológicas superiores a 1 Ma (Ho e Larson, 2006). Por exemplo, taxas de mutação de aproximadamente 30% por Ma foram estimadas para a região controladora do mtDNA de espécies do gênero *Bison* com divergência recente (Shapiro *et al*., 2004). É extremamente importante entender que as taxas moleculares não são constantes ao longo do tempo e que existe uma transição mensurável entre a taxa de mutação (taxa instantânea de ocorrência de alterações nucleotídicas), que é elevada em tempos curtos da taxa de substituição (taxa pela qual as mutações são fixadas no genoma), que são mais baixas e persistem por longos prazos (Ho e Larson, 2005). Segundo esses autores, a falta de distinção entre esses dois fatores e ausência de consideração de sua relação talvez sejam as principais causas dos resultados discrepantes entre datas moleculares e paleontológicas e/ou arqueológicas.

5.2.6 Calibrar é possível???

Apesar das controvérsias, o relógio molecular é uma ferramenta indispensável em estudos de biologia evolutiva. Violações na hipótese fazem evidentes seus efeitos tanto em estudos com espécies distantemente relacionadas, resultando em estimativas de tempos de divergência flagrantemente incorretas (Rannala e Yang, 2007), como quando o relógio, utilizado para datar eventos recentes e já datados com base em dados paleontológicos, arqueológicos e biogeográficos, fornece datas conflitantes (Ho e Larson, 2006).

Em vista da crescente quantidade de evidências de violação da hipótese de taxas constantes e de que o DNA, mesmo de espécies proximamente relacionadas, pode evoluir com taxas diferentes, como podemos tratar essa heterogeneidade de modo a minimizar os erros que ocorrem nas estimativas de tempo de divergências?

Uma abordagem inicial seria remover da análise as linhagens ou sequências, genes ou sítios com taxas anômalas. Existem alguns testes que permitem a identificação de linhagens com taxas variáveis, como o Teste de Taxas Relativas (Wu e Li, 1985), o Teste de Tajima (Tajima, 1993) e o *Likelihood Ratio Test* (Felsenstein, 1981). Essa abordagem tem duas deficiências principais: a primeira é que esses testes normalmente só detectam taxas muito discrepantes; segundo, ela só é válida quando as diferenças entre as taxas são a exceção e não a regra (Welch e Bromham, 2005; Rannala e Yang, 2007).

Um modo mais promissor de resolver o problema é considerar explicitamente a variação entre linhagens quando se estimam os tempos de divergência. O desenvolvimento de modelos com taxas variáveis tem sido o foco de estudos recentes, tanto utilizando metodologia de verossimilhança como Bayesiana. Na análise de verossimilhança, linhagens pré-determinadas recebem parâmetros de taxa independentes, estimados a partir dos dados (Rambaut e Bromham, 1998; Yoder e Yang, 2000). Yang e Yoder (2003) modificaram essa metodologia para permitir o uso de vários pontos de calibragem e a análise simultânea de mais de um gene, levando em consideração diferenças na taxa de substituição (Rannala e Yang, 2007).

A abordagem Bayesiana usa modelos estocásticos de mudanças das taxas evolutivas para especificar a distribuição *prior* (anterior) e, com a prior, calcula a distribuição posterior dos tempos e taxas (para maior detalhes ver Huelsenbeck, 2000). Esse método também foi aprimorado para permitir a análise de mais de um gene (Thorne e Kishino, 2002). A abordagem Baysiana vem sendo utilizada em um grande número de grupo de espécies, como mamíferos (Springer *et al*., 2003), aves (Pereira e Baker, 2006) e em plantas (Bell e Donoghue, 2005). Para uma revisão recente dos métodos que incorporam variação nas taxas de evolução, ver Welch e Bromham (2005).

A escolha de uma calibragem apropriada também é fundamental em ecologia molecular, onde a escala temporal evolutiva, isto é, aquela calculada a partir de dados moleculares, tem grande importância em estudos de biogeografia, especiação e biologia da conservação. As informações para a calibragem podem ser incorporadas na análise de muitas maneiras. Entre elas, podemos citar: (1) usando o registro fóssilífero para datar um evento de divergência; (2) usar taxas de substituição obtidas independentemente para outros táxons; (3) inclusão de sequências heterocrônicas de idade conhecida (Ho *et al*., 2008).

O uso de fósseis em estudos ecológicos não é apropriado, pois, como os processos evolutivos estudados são aqueles intraespecíficos (migração e extinção, por exemplo) e que ocorrem em escalas genealógicas, e não filogenéticas, diferentes estágios do processo de substituição nucleotídica estão sendo observados (ver item 5.2.5). Ho e colaboradores (2008) demonstraram, utilizando três estudos—sobre especiação de aves no final do Pleistoceno, sobre historia demográfica da baleia da Groenlândia (*Balaena mysticetus*) e sobre a biogeografia de ursos marrons (*Ursos arctos*) no Pleistoceno—que as estimativas das datas de divergência podem variar em até uma ordem de magnitude quando pontos de calibragens internos e relativamente recentes são usados. Existem várias maneiras de se obterem pontos de calibragem intraespecíficos. Uma delas seria usar amostras antigas datadas pelo método de rádiocarbono. Outra fonte de pontos internos de calibragem seriam dados de biogeografia, entretanto, esse tipo de dados pode ser difícil de interpretar e especificar corretamente. A última opção seria usar taxas de substituição de outras espécies, que sejam o mais proximamente relacionadas possível (Ho *et al.,* 2008). Uma das vantagens dos pontos de calibragem intraespecíficos é que elas produzem estimativas com algum grau de incerteza, pois a calibragem externa é colocada na raiz da árvore e também porque, ao se retirar o grupo externo, ocorre uma diminuição no conteúdo informativo da matriz de dados (Ho *et al.,* 2008).

Referências Bibliográficas

Alvarez-Valin, F., Jabbari, K. e Bernardi, G. (1998). Synonymous and nonsynonymous substitutions in mammalian genes: intragenic correlations. **J. Mol. Evol. 46**: 312-322.

Aota, S.-I., e Ikemura, T. (1986). Diversity in G + C content at the third position of codons in vertebrates and its causes. **Nucleic Acids Res.**

14: 6345-6355.

Baker, C.S., Perry, A., Bannister, J.L., Weinrich, M.T., Abernethy, R.B., Calambokidis, J., Lien, J., Lambertsen, R.H., Ramirez, J.U., Vasquez, O., Clapham, P.J., Alling, A., O'Brien, S.J. e Palumbi, S.R. (1993). Abundant Mitochondrial DNA Variation and World-Wide Population Structure in Humpback Whales. **Proc. Natl. Acad. Sci USA** 90: 8239-8243.

Bargelloni, L., Scudiero, R., Parisi, E., Carginale, V., Capasso, C. e Paternello, T. (1999). Metallothioneins in Antarctic fish: Evidence for independent duplication and gene conversion. **Mol. Biol. Evol. 16**: 885-897.

Bell, C.D. e Donoghue, M.J. (2007). Dating the Dipsacales: comparing models genes and evolutionary implications. **Am. J. Bot. 92**: 284-296.

Bernardi, G., Mouchiroud, D. e Gautier, C. (1993). Silent substitutions in mammalian genomes and their evolutionary implications. **J. Mol. Evol. 37**: 583-589.

Bleiweiss, R. (1998). Slow rate of molecular evolution in high elevation hummingbirds. **Proc. Natl. Acad. Sci. USA. 95**: 612-616.

Britten, R.J. (1986). Rates of DNA sequence evolution differ between taxonomic groups. **Science 231**: 1393-1398.

Brown, W.M., Prager, E.M., Wang, A. e Wilson, A.C. (1982). Mitochondrial DNA sequences of primates. Tempo and mode of evolution. **J. Mol. Evol. 18**: 225-239.

Brownell, E. (1983). DNA/DNA hybridization studies of muroid rodents: symmetry and rates of molecular evolution. **Evolution 37**: 1034-1051.

Catzeflis, F.M., Sheldon, F.H., Ahlquist, J.A. e Sibley, C.G. (1987). DNA/DNA hybridization studies evidence for the rapid rate o muroide rodent DNA evolution. **Mol. Biol. Evol. 4**: 242-253.

Chang, B.H.-J. e Li, W.-H. (1995). Estimating the intensity of male-driven evolution in rodents by using X-linked and Y-linked *Ube1* genes and pseudogenes. **J. Mol. Evol. 40**: 70-77.

Chang, B.H.-J., Shimmin, L.C., Shyue, S.-K, Hewett-Hemmett, D. e Li, W.-H. (1994). Weak male-driven molecular evolution in rodents. **Proc. Natl. Acad. Sci. USA 91**: 827-831.

Eastel, S. e Collet, C. (1994). Consistent variation in amino-acid substitution rate, despite uniformity of mutation rate: Protein evolution in mammals is not neutral. **Mol. Biol. Evol. 11**: 643-647.

Fambrough, D.M. (ed) (1994). **Molecular Evolution of Physiological Processes**. The Rockefeller University Press, New York, 297pp.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. **J. Mol. Evol. 17**: 368-376.

Fisher, R.A. (1930). **The Genetical Theory of Natural Selection**. Clarendon Press, Oxford.

Gaut, B.S., Muse, V.S., Clark, W.D. e Clegg, M.T. (1992). Relative rates of nucleotide substitution at the *rbc*L locus of monocotyledonous plants. **J. Mol. Evol. 11**: 961-963.

Gaut, B.S (1998). Molecular clocks and nucleotide substitution rates in higher plants. **Evol. Biol. 30**: 93-120.

Gillespie, J.H. (1989). Lineage effects and their index of dispersion of molecular evolution. **Mol. Biol. Evol. 6**: 636-647.

Gillooly, J.F., Brown, J.H., West, G.B., Savage, V.M., Charnov, E.L. (2001). **Science 293**: 2248-2251.

Gillooly, J.F., Allen, A.P., West, G.B., Brown, J.H. (2005). The rate of DNA evolution: effects of body size and temperature on the molecular clock. **Proc. Natl. Acad. Sci. USA. 102**: 140-145.

Goodman, M. (1961). The role of immunochemical differences in the phyletic development of human behavior. **Hum. Biol. 33**: 131-161.

Goodman, M., Barnabas, J., Matsuda, G. e Moore, G.W. (1971). Molecular evolution in the descent of man. **Nature 233**: 604-613.

Grantham, R., Gautier, C., Gouy, R., Mercier, R. e Pavé, A. (1980). Codon catalog usage and the genome hypothesis. **Nucleic Acids Res. 8**: 49-62.

Gu, X. e Li, W.-H. (1992). Higher rates of amino acid substitution in rodents than in humans. **Mol. Phylogenet. Evol. 1**: 211-214.

Ho, S.Y.W., Phillips, M.J., Cooper, A., Drummond, A.J. (2005). Time dependency of molecular rates estimates and systematic overestimation of recent divergence times. **Mol. Biol. Evol. 22**: 1561-1568.

Ho, S.Y.W e Larson, G. (2006). Molecular clocks: when times are a-changin'. **Trends in Genetics 22(2)**: 79-83.

Ho, S.Y.W., Saarma, U., Barnett, R., Haile, J., Shapiro, B. (2008). The effect of inappropriate calibration: three case studies in molecular ecology. **PloS One 3**: 1-8.

Huelsenbeck, J.P., Larget, B., Swofford, D. (2000). A compound Poisson process for relaxing the molecular clock. **Genetics 154**: 1879-1892.

Hughes, A.L. e Nei, M. (1989). Nucleotide substitution at major histocompatibility complex class II loci: Evidence for overdominant selection. **Proc. Natl. Acad. Sci. USA. 86**: 958-962.

Irwin, D.M., Kocher, T.D. e Wilson, A.C. (1991). Evolution of the cytochrome *b* gene of mammals. **J. Mol. Evol. 32**: 128-144.

Jermiin, L.S. (1996). **K2WuLi Version 1.0**. *http: //jcsmr.anu.edu.au/dmm/humgen.html/lars/k2wulisub.html*.

Kimura, M. (1968). Evolutionary rate at the molecular level. **Nature 217**: 624-626.

Kimura, M. (1977). Preponderance of synonymous changes as evidence for neutral theory of molecular evolution. **Nature 267**: 275-276.

Kimura, M. e Ohta, T. (1973). Mutation and evolution at the molecular level. **Genet. Suppl. 73**: 19-35.

Kliman, R.M. e Hey, J. (1994). The effects of mutation and natural selection on codon bias in the genes of *Drosophila*. **Genetics 137**: 1049-1056.

Kocher, T.D., Conroy, J.A., McKaye, K.R., Stauffer, J.R. e Lockwood, S.F. (1995). Evolution of the NADH dehydrogenase subunit 2 in east African cichlid fish. **Mol. Phyl. Evol. 4**: 420-432.

Kohne, D.E., Chiscon, J.A. e Hoyer, B.H. (1972). Evolution of primate DNA sequences. **J. Hum. Evol. 1**: 627-644.

Koop, B.F., Miyamoto, M.M., Embury, J.E., Goodman, M., Czelusniak, J. e Slightom, J.L. (1986). Nucleotide sequence and evolution of the orangutan ε-globin gene region and surrounding Alu repeats. **J. Mol. Evol. 24**: 94-102.

Laird, C.D., McConaughty, B.L., McCarthy, B.J. (1969). Rates of fixation of nucleotide substitution in evolution. **Nature 224**: 149-154.

Lee, W-J, Conroy, J., Howell, W. H. e Kocher, T.D. (1995). Structure and evolution of teleost mitochondrial control regions. **J. Mol. Evol.** 41: 54-66.

Li, W.-H. (1997). **Molecular Evolution**. Sinauer Associates, Sunderland Massachusetts, 487pp.

Li, W.-H. e Graur, D. (1991). **Fundamentals of Molecular Evolution**. Sinauer Associates, Sunderland, Massachusetts, 284pp.

Li, W.-H. e Tanimura, M. (1987). The molecular clock runs more slowly in man than in apes and monkeys. **Nature 326**: 93-96.

Li, W.-H, Wu, C.-I e Luo, C.-C. (1985). A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. **Mol. Biol. Evol. 2**: 150-174.

Li, W.-H., Tanimura, M. e Sharp, P.M. (1987). An evaluation of the molecular clock hypothesis using mammalian DNA sequences. **J. Mol. Evol. 25**: 330-342.

Li, W.-H., Ellsworth, D.L., Krushkal, J., Chang, B.H.-J. e Hewett-Emmett, D. (1996). Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis. **Mol. Phyl. Evol. 5**: 182-187.

Luo, C.-C., Li, W.-H e Chan, L. (1989). Structure and expression of dog apolipoprotein A-I, E, and C-I mRNAs: implications for the evolution and functional constrains of apolipoprotein structure. **J. Lipid Res. 30**: 1735-1746.

Margoliash, E. (1963). Primary structure and evolution of Cytochrome C. **Proc. Natl. Acad. Sci. USA. 50**: 672–679.

Martin, A.P. e Palumbi, S.R. (1993). Body size, metabolic rate, generation time, and the molecular clock. **Proc. Natl. Acad. Sci. USA 90**: 4087-4091.

Martin, A.P., Naylor, G.J.P. e Palumbi, S.R. (1992). Rates of mitochondrial DNA evolution in sharks are slow compared with mammals. **Nature 357**: 153-155.

Matsubara, H. e Yamanaka, T. (1978). **Evolution of Protein Molecules**. Japan Scientific Societies Press, 412pp.

Mindell, D.P., Knight, A., Baer, C. e Huddleston, C.J. (1996). Slow rates of molecular evolution in birds and the metabolic rate and body temperature hipothesis. **Mol. Biol. Evol. 13**: 422-426.

Moriyama, E.N. e Gojobori, T. (1992). Rates of synonymous substitution and base composition of nuclear genes in *Drosophila*. **Genetics 130**: 855-864.

Muse, S.V. (2000). Examining rates and patterns of nucleotide substitution in plants. **Plant Mol. Biol. 42**: 25-43.

Muse, S.V. e Weir, B. (1992). Testing for equality of evolutionary rates. **Genetics 132**: 269-276.

Nei, M. (1987). **Molecular evolutionary Genetics**. Columbia Univ. Press, New York.

Ohta, T. (1995). Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. **J. Mol. Evol. 40**: 56-63.

O'hUigin, C. e Li, W.-H. (1992). The molecular clock ticks regularly in muroid rodents and hamsters. **J. Mol. Evol. 35**: 377-384.

Page, R.D.M. e Holmes, E.C. (1998). **Molecular Evolution**: **A Phylo-**

genetic Approach. Blackwell Science. 346 pp.
Penny, D. (2005). Relativity for molecular clocks. **Nature 436**: 183-184.
Pereira, S.L. e Baker, A.J. (2006). A mitogenomic timescale for birds detects variable phylogenetic rates of molecular evolution and refutes the standard molecular clock. **Mol. Biol. Evol. 23**: 1731-1740.
Pesole, G., Gissi, C., De Chirico, A. e Saccone, C. (1999). Nucleotide substitution rate of mammalian mitochondrial genomes. **J. Mol. Evol. 48**: 427-434.
Rambaut, A. e Bromham, L. (1998). Estimating divergence dates from molecular sequences. **Mol. Biol. Evol. 15**: 442-448.
Rannala, B. e Yang, Z. (2007). Infering speciation times under an episodic molecular clock. **Syst. Biol. 56(3)**: 453-466.
Randi, E. e Lucchini, V. (1998). Organization and evolution of the mitochondrial DNA control region in the avian genus *Alectoris*. **J. Mol. Evol. 47**: 449-462.
Ratner, V.A., Zharkikh, A.A., Kolchanov, N. Rodin, S.N., Solovyov, V.V. e Antonov, A.S. (1996). **Molecular Evolution. Biomathematics** vol. 24 Springer-Verlag Berlin Heidelberg, pp. 93-145.
Richter, C., Park, J.-W. e Ames, N.B. (1988). Normal oxidative damage to mitochondrial and nuclear DNA is extensive. **Proc. Natl. Acad. Sci. USA 85**: 6465-6467.
Sarich, V.M. e Wilson, A. (1967). Immunological time scale for hominoid evolution. **Science 158**: 1200-1203.
Sarich, V.M. e Wilson, A.C. (1973). Generation time and genomic evolution in primates. **Science 179**: 1144-1147.
Seino, S., Bell, G.I., Li, W.R. (1992). Sequences of primate insulin genes support the hypothesis of a slower rate of molecular evolution in humans and apes than monkeys. **Mol. Biol. Evol. 9**: 193-203.
Shapiro, B., Drummond, A.J., Rambaut, A., Wilson, M.C., Matheus, P.E. *et al.* (2004). Rise and fall of the Beringian steppe bison. **Science 306**: 1561-1565.
Sharp, P.M. e Li, W.-H. (1986). An evolutionary perspective on synonymous codon usage in unicellular organisms. **J. Mol. Evol. 24**: 28-38.
Sharp, P.M., Cowe, E., Higgins, D.G., Shields, D., Wolfe, K.H. e Wright, F. (1988). Codon usage patterns in *Scherichia coli*, *Bacillus subtilis*, *Saccharomyces cerevisae, Schizosaccharomyces pombe*, *Drosophila melanogaster* and *Homo sapiens*: A review of the considerable within-species diversity. **Nucleic Acids Res. 16**: 8207-8211.
Shimmin, L.C., Chang, H.-J. e Li, W.-H. (1993). Male-driven evolution of DNA sequences. **Nature 362**: 745-747.
Shpaer, E.G. e Mullins, J.I. (1993). Rates of amino acid change in the envelope protein correlate with pathogenicity of primate lentiviruses. **J. Mol. Evol. 37**: 57-65.
Sibley, C.G. e Ahlquist, J.E. (1984). The phylogeny of the hominoid primates, as indicated by DNA-DNA hybridization. **J. Mol. Evol. 20**: 2-15.
Sibley, C.G. e Ahlquist, J.E. (1987). DNA hybridization evidence of hominoid phylogeny: results from an expanded data set. **J. Mol. Evol. 26**: 99-121.
Springer, M.S., Murphy, W.J., Eizirik, E., O'Brien S.J. (2003). Placental mammal diversification and the Cretaceous-Terciary boundary. **Proc. Natl. Acad. Sci. USA 100**: 1056-1061.
Stanley, S.E. e Harrison, R.G. (1999). Cytochrome *b* evolution in birds and mammals: an evaluation of the avian constraint hypothesis. **Mol. Biol. Evol. 16**: 1575-1585.
Tajima, F. (1993). A simple method for testing the molecular evolutionary clock hypothesis. **Genetics 135**: 599-607.
Tanaka, T. e Nei, M. (1989). Positive Darwinian selection observed at the variable region genes of immunoglobulins. **Mol. Biol. Evol. 6**: 447-459.
Thorne, J.L., Kishino, H. e Painter, I.S. (1998). Estimating the rate of evolution of the rate of molecular evolution. **Mol. Biol. Evol. 15**: 1647-1657.
Thorne, J.L. e Kishino, H. (2002). Divergence time and evolutionary rate estimation with multilocus data. **Syst. Biol. 51**: 689-702.
Welsh, J.W. e Bromham, L. (2005). Molecular dating when rates vary. **Trends Ecol. Evol. 20**: 320-327.
Wilson, A.C., Carlson, S.S. e White, T.J. (1977). Biochemical evolution. **Ann. Rev. Biochem. 46**: 573-639.
Wolfe, K.H., Li, W.-H. e Sharp, P.M. (1987). Rates of nucleotide substitution vary greatly among plant mitochondrial, chloropast, and nuclear DNAs. **Proc. Natl. Acad. Sci. USA 84**: 9054-9058.
Wolfe, K.H., Sharp, P.M. e Li, W.-H. (1989). Rates of synonymous substitution in plant nuclear genes. **J. Mol. Evol. 29**: 208-211.
Wright, S.D., Gray, R.D., Gardner, R.C. (2003). Energy and rate of evolution: inferences from plant rDNA substitution rates in western Pacific. **Evolution 57**: 2893-2898.
Wu, C.I. e Li, W.H. (1985). Evidence for higher ratesof nucleotide substitutions in rodents than in man. **Proc. Natl. Acad. Sci. USA 82**: 1741-1745.
Yang, Z. e Yoder, A.D. (2003). Comparison of likelihood and Bayesian methods for estimating divergence times using multiple gene loci and calibration points, with application to a radiation of cute-looking mouse and lemur species. **Syst. Biol. 52**: 705-716.
Yoder, A.D. e Yang, Z. (2000). Estimation of primate speciation dates using local molecular clocks. **Mol. Biol. Evol. 17**: 1081-1090.
Zuckerkandl, E. e Pauling, L. (1962). Molecular disease, evolution and genic heterogeneity. **In** Kasha, M. and Pullman B. (eds.). **Horizons in Biochemistry**. Academic Press, New York, pp. 189-225.
Zuckerkandl, E. e Pauling, L. (1965). Evolutionary divergence and convergence of proteins. **In** Bryson V. and Vogel H.J. (eds.). **Evolving Genes and Proteins.** Academic Press, New York, pp. 97-166.

Página deixada em branco

Capítulo 8

Evolução dos genes nucleares de RNA ribossômico

Eduardo Gorab (egorab@usp.br)
Departamento de Genética e Biologia Evolutiva
Instituto de Biociências
Universidade de São Paulo

"*O mundo do RNA ribossômico caindo aos pedaços*." (Gabriel Dover)

8.1. Introdução

Conforme visto no Capítulo 7, existe um grande arcabouço teórico que permite explicar vários aspectos do padrão de evolução das sequências de DNA que codificam para proteínas. Em geral, a teoria que diz respeito a taxas de evolução, restrições funcionais, entre outros aspectos moleculares, está relacionada com aquilo que se conhece a respeito da estrutura das proteínas. No genoma dos eucariotos, entretanto, a fração codificadora para proteínas é apenas uma parcela da potencialidade informativa que existe, conforme a grande quantidade de informação a respeito de genomas que se obtém atualmente.

Neste capítulo, será apresentado um modo com o qual se pode abordar evolutivamente o que acontece com as sequências de DNA responsáveis pela síntese de uma importante classe de moléculas, os RNAs ribossômicos (rRNAs), que, conforme visto no Capítulo 2, é uma das moléculas remanescentes dos primórdios da vida na Terra.

Os RNAs ribossômicos são componentes essenciais na fisiologia celular. Esses componentes interagem de modo específico com as proteínas ribossômicas para formar as subunidades dos ribossomos que atuam na síntese de proteínas. Os rRNAs são o principal produto da transcrição em qualquer célula, constituindo geralmente de 80 a 90% da massa de RNA total dos procariontes e eucariontes.

As sequências que codificam para rRNA são reiteradas, ocorrendo em número variável nos diversos organismos estudados. Nos eucariontes, essas cópias estão organizadas *in tandem* (isto é, em sequência) e agrupadas em uma ou mais regiões cromossômicas, as chamadas regiões organizadoras do nucléolo. Os RNAs ribossômicos são classificados conforme seu coeficiente de sedimentação sob campo centrífugo, que depende tanto do tamanho, quanto da densidade da molécula, em unidades Svedberg (de símbolo S). De forma geral, cada unidade de repetição do rDNA eucarionte possui uma organização conservada (Figura 8.1), consistindo de: (1) um espaçador externo (ETS, do inglês *external transcribed spacer*), transcrito em uma sequência que contém a extremidade 5' da molécula precursora do rRNA, esta com alto coeficiente de sedimentação (até 45S); (2) uma região que codifica para rRNA 18S; (3) um espaçador interno que é transcrito (ITS, do inglês *internal transcribed spacer*) e que contém a região que codifica para rRNA 5,8S; e (4) uma região que codifica para rRNA 25-28S e um espaçador externo (NTS, de *non transcribed spacer*), que não é transcrito. Maiores detalhes sobre a estrutura do rDNA podem ser vistos em Lewin (1994).

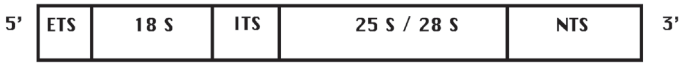


Figura 8.1. Representação esquemática da unidade de repetição do rDNA dos eucariontes. A proporção entre as diversas regiões dentro da unidade não está necessariamente representada na figura. ETS: espaçador externo; 18S: região que codifica para rRNA 18S; ITS: espaçador interno; 25S/28S: região que codifica para rRNA 25/28S; NTS: espaçador externo ou espaçador intergênico.

Além dos estudos funcionais dos produtos da transcrição do rDNA ou mesmo da possível implicação funcional de regiões que geralmente não são transcritas (NTS), vários componentes da unidade de repetição dos genes de rRNA têm sido estudados com um enfoque evolutivo, como é o caso dos genes para o rRNA 18S, 25-28S, a região do NTS e, mais recentemente, a região que contém o ITS. Neste capítulo, serão comentados alguns aspectos da evolução dos genes cujo rRNA está presente na subunidade maior do ribossomo, isto é, dos genes cujo coeficiente de sedimentação de seus produtos de transcrição varia de 23S, nos procariontes, a 28S, nos eucariontes superiores.

8.2. O rDNA da Subunidade Maior do Ribossomo nos Procariontes e Eucariontes

Um aspecto interessante na história da pesquisa nesse campo é que praticamente todas as etapas da síntese dos rRNAs, desde a transcrição de uma molécula precursora até a formação de espécies de rRNA através de processamento, assim como a existência de diferenças de tamanho entre genes de rRNA (Tabela 8.1) em virtude da variação observada no coeficiente de sedimentação do rRNA da subunidade maior do ribossomo dos procariontes (23S) e dos eucariontes (variável de 25S a 28S), foram evidenciadas muito antes do sequenciamento do rDNA. Contudo, somente com o emprego desse método foi possível uma análise comparativa dos genes de rRNA de várias espécies. A análise da estrutura primária dos genes de rRNA 28S do sapo do gênero *Xenopus*, 26S de levedura e 23S da bactéria *Escherichia coli*, obtida com os dados de sequenciamento no início da década de 1980 (Ware *et al.*, 1983), revelou uma série

Tabela 8.1. Variação observada no comprimento da molécula de rRNA da subunidade maior do ribossomo em procariontes e eucariontes e, consequentemente, nos seus coeficientes de sedimentação (dados de Lewin 1994).

Organismo	Comprimento em bases do rRNA da subunidade maior
Escherichia coli	2904
Saccharomyces cerevisiae	3750
Drosophila melanogaster	4100
Xenopus laevis	4475
Nicotiana tabacum	3700
Gallus domesticus	4625
Mus musculus	4712

de aspectos que não haviam sido ainda mostrados com estudos de hibridação heteróloga de ácidos nucléicos.

Demonstrou-se, então, que certas regiões apresentavam uma conservação significativa de sequências, tanto em procariotos como nos eucariotos analisados. Essas regiões conservadas em organismos filogeticamente distantes apareciam dispersas entre regiões com baixa porcentagem de similaridade entre suas sequências (Figura 8.2). Além disso, o rDNA 25/28S dos eucariontes, quando comparado ao de *E. coli*, apresentava inserções de sequências em posições definidas ao longo do gene; essas inserções eram de comprimento variável e não havia similaridade significativa entre suas sequências (estruturas primárias) em organismos distintos. Essas regiões eram as responsáveis pela diferença nos coeficientes de sedimentação de moléculas de rRNA de procariontes e eucariontes, ou mesmo entre moléculas funcionalmente equivalentes de rRNA dos eucariontes.

8.3. Os Domínios Divergentes dos Genes de rRNA 25S-28S

Os resultados de sequenciamento do rDNA demonstraram que os genes de rRNA da subunidade maior do ribossomo dos eucariontes estão estruturados como um mosaico de domínios conservados e variáveis. Esses domínios variáveis, ou inserções não conservadas, que não estão presentes nos genes de rRNA 23S dos procariontes, são denominadas *divergentes* (D), *variáveis* (V), ou ainda *segmentos de expansão*. A referência a um determinado segmento divergente faz-se sempre com uma numeração crescente relativamente a sua ordem de aparecimento a partir da extremidade 5' da molécula de rRNA. Assim, o segmento de expansão D1 será o primeiro domínio divergente próximo ao extremo 5' da molécula de rRNA 25S-28S, enquanto que o segmento de expansão D12 estará próximo ao extremo 3' da molécula (um exemplo pode ser visto em Hancock *et al.*, 1988). Com relação à sequência de bases, ou estrutura primária, os domínios divergentes apresentam baixa porcentagem de similaridade entre as diferentes espécies, diferentemente do grau de similaridade de sequências observado em regiões conservadas. Nesse último caso, a similaridade em certos trechos pode chegar a quase 100% em organismos evolutivamente distantes, como leveduras e mamíferos. Os segmentos de expansão também podem variar em número de bases nas diferentes espécies (Figura 8.3). No entanto, a posição desses segmentos no gene está conservada. Além disso, outro aspecto conservativo dos domínios divergentes refere-se a sua estrutura secundária dentro da molécula de rRNA 25-28S.

A representação da forma como uma determinada porção da molécula do RNA emparelha consigo mesma denomina-se estrutura secundária (*folding*, em inglês). Essas dobras moleculares possibilitarão interações entre estruturas secundárias e, consequentemente, em estruturas moleculares de complexidade espacial crescente. Embora existam programas de computador capazes de inferir estruturas secundárias de RNA (através de parâmetros termodinâmicos—para maiores detalhes, consultar Zucker e Stiegler, 1981), eles não são suficientes para prever a estrutura secundária de uma determinada sequência de bases do

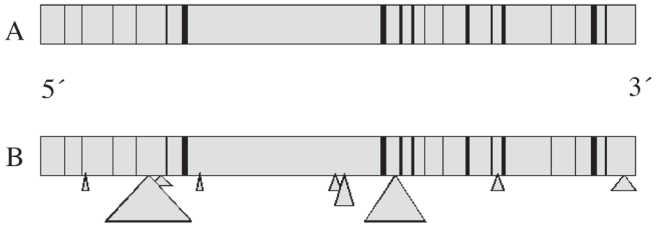


Figura 8.2. Representação esquemática (modificada de Ware *et al.*, 1983) dos genes de rRNA da subunidade maior (23S-28S) do ribossomo dos procariontes (A) e eucariontes (B). As linhas verticais e áreas em negrito mostram a posição dos domínios conservados (> 70% de similaridade) em procariontes e eucariontes. Áreas em cinza correspondem a regiões com baixa similaridade de sequências em procariontes e eucariontes. Os triângulos em B representam os maiores segmentos de expansão dos eucariontes (3 segmentos de expansão com menos de 25 bases não estão representados).

D5

Chironomus thummi *:TGAACGTAGCACGTAAGATTGTATCGCG
ATGGGTATGTAAGTCAACATGCTTGAT-TGATGTTGACCATTAGCTTACGTTCTTA*CAGTGGCCAAGTTT

ATGGGTAAGAACCTTAACTTTCTTGATATGAAGATCAAGGTTATGATATAAATGTCCCAGTGGGCCACTTT

Drosophila melanogaster

D6

Chironomus thummi *:TTC
GCTCAAGTCGTTTGCCGATACATGTCGTTAAGATAAAATCAGTGTGTGTTCATTGTTGGACAT*ATTTTGAAA

GCTTAAGTTGTATACCTATACATTACCGCTAAAGTAGATGATTTATATTAC-TTGTGATATAA ATTTTGAAA
Drosophila melanogaster

Figura 8.3. Alinhamento de sequências dos domínios divergentes D5 e D6 do rDNA de *Chironomus thummi* e de *Drosophila melanogaster* (modificado de Gorab *et al.*, 1995). A estimativa de divergência entre as duas espécies é da ordem de 150 milhões de anos. As sequências em negrito que flanqueiam os segmentos D5 e D6 pertencem a domínios conservados do rDNA eucarionte. As bases e motivos sublinhados são comuns às duas espécies. Os "gaps" (-) foram introduzidos para otimizar o alinhamento. Note a continuação do segmentos de expansão de *Chironomus* (asterisco), cujas sequências não podem ser alinhadas com os respectivos segmentos de *Drosophila*.

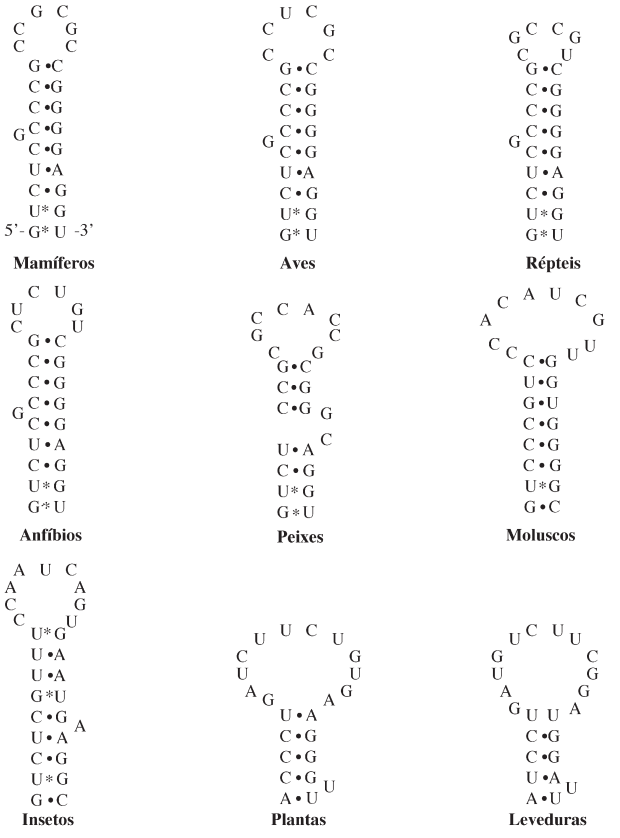


Figura 8.4. Exemplos de estruturas secundárias inferidas para parte do domínio variável D3 de alguns eucariontes (modificadas de Michot *et al.*, 1990). As demais regiões do segmento que não estão representadas também apresentam conservação em suas estruturas secundárias e, por isso, foram omitidas. Todas as estruturas secundárias da figura estão orientadas 5'-3', como mostradas na estrutura de mamíferos. Os emparelhamentos que não obedecem aos estabelecidos por Watson e Crick ("não canônicos") estão representados com asteriscos.

RNA. Particularmente em relação aos rRNAs, existem modelos de estruturas secundárias disponíveis na literatura que devem ser levados em conta antes da aplicação de um programa de computador. Isso porque esses modelos apóiam-se não somente em aspectos termodinâmicos, mas principalmente em dados bioquímicos (como digestão controlada com enzimas que degradam o RNA), que podem desvendar interações intramoleculares não previsíveis com a simples aplicação de um programa de computador.

Essas características dos domínios variáveis, posição conservada em moléculas de rRNA e conservação da estrutura secundária em espécies filogeneticamente distantes (Figura 8.4) têm gerado um debate em torno do possível papel dessas regiões. Por um lado, existem aqueles que acreditam que os domínios divergentes não contribuem funcionalmente no ribossomo. Um apoio a essa hipótese vem do fato de que a estrutura primária dessas regiões é variável; havendo regiões com alto grau de conservação de sequências em espécies evolutivamente distantes, essas últimas seriam as candidatas a um papel funcional relevante (e não os segmentos de expansão). Nesse caso, não havendo pressão seletiva, os domínios variáveis acumulariam mutações e, portanto, teriam estrutura primária com baixa similaridade de sequências nas diferentes espécies (detalhes em Gerbi *et al.*, 1985).

Por outro lado, outros acreditam que a manutenção da estrutura secundária dos domínios divergentes ao longo da evolução tenha necessariamente implicações funcionais e que eles devem cumprir algum papel, seja na função do ribossomo ou em algum outro processo. Os adeptos dessa visão têm proposto processos evolutivos através dos quais os domínios variáveis mantêm sua estrutura secundária. Por exemplo, no processo denominado "mutação compensatória", uma mutação em um ponto do segmento de expansão é compensada por outra mutação em outro ponto do mesmo segmento, de forma a manter o emparelhamento de bases que contribui para a manutenção da estrutura secundária da região, particularmente nos troncos (*stems*, em inglês) dos segmentos de expansão (Figura 8.5; ver também Hancock *et al.*, 1988). Outro exemplo de processo compensatório em domínios divergentes pode ser visto em Hancock e Dover (1990).

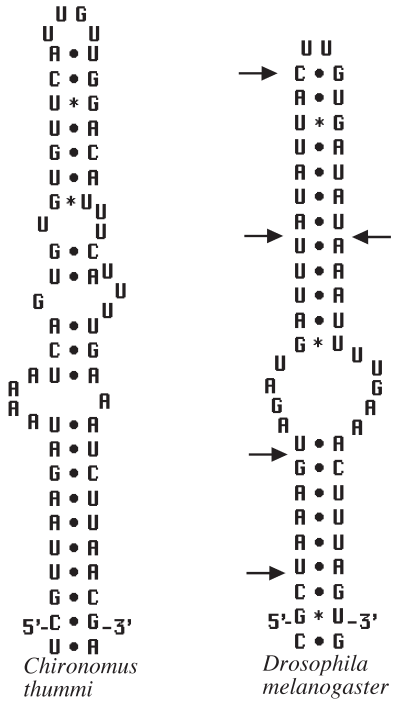


Figura 8.5. Previsão de estrutura secundária do segmento de expansão D6 de *Chironomus thummi* e *Drosophila melanogaster*. As setas indicam alguns exemplos de mutações compensatórias (ver Figura 6.3) no segmento de expansão que evoluiu mais recentemente, assumindo, nesse caso, que o segmento de expansão de um díptero aquático (*Chironomus*) deva ser considerado o mais primitivo. Os dados da figura foram obtidos em Hancock *et al.* (1988) e Gorab *et al.* (1995).

8.4. Papel Funcional dos Domínios Variáveis?

Embora se acredite que a conservação da estrutura secundária dos segmentos de expansão esteja relacionada à funcionalidade dessas regiões, seu papel ainda é essencialmente desconhecido. Por outro lado, estudos sobre o processamento do rRNA de alguns organismos—que será comentado adiante— mostram que envolve um domínio divergente particular, sendo este o único caso conhecido na literatura em que um mecanismo bioquímico se relaciona a um domínio variável.

Algumas espécies apresentam uma quebra específica, aproximadamente na metade da molécula de rRNA 26S, dividindo-a em metades designadas como α e β. Em condições não desnaturantes, essas metades continuam unidas por pontes de hidrogênio. Em géis desnaturantes, essa ruptura é facilmente visualizada, já que esse rRNA migrará mais rapidamente do que o esperado. No início, pensou-se que essa fragmentação fosse um artefato, consequência do procedimento de extração do RNA, naquele momento ainda objeto de cuidados extremos para evitar problemas com degradação. Mais tarde, observou-se que essa ruptura era um padrão comum, tendo sido documentada em protozoários, moluscos, anelídeos e artrópodes. Esse processo foi, então, estudado em detalhe em alguns insetos, como *Drosophila*, *Sciara* (Diptera) e *Bombyx* (Lepidoptera), quando se descobriu que essa ruptura não se tratava de uma simples quebra na molécula do rRNA. Na realidade, de 19 a 60 bases (dependendo da espécie) do rRNA são removidas num processo análogo ao de remoção dos introns dos RNAs mensageiros. Essas bases sempre fazem parte do domínio divergente D7a (Figura 8.6; para maiores detalhes, consultar os trabalhos referidos na figura). Pensou-se inicialmente que o motivo UAAU, comum a todas as alças (*loops*, em inglês) do RNA presentes no segmento, seria o sinal de reconhecimento para que uma enzima realizasse o processo de remoção de bases. Uma análise posterior de domínios divergentes D7a de outros insetos, cujo rRNA sofre esse processamento, demonstrou que esse motivo está ausente e, portanto, não participa desse mecanismo (Figura 8.7).

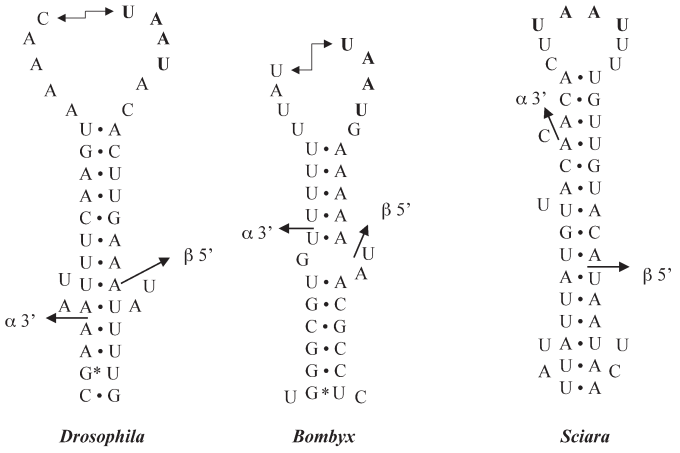


Figura 8.6. Estruturas secundárias dos domínios divergentes D7a de insetos cujo rRNA 26S é processado nesta região, resultando em metades α e β (modificadas de Fujiwara e Ishikawa, 1986). Os conectores colocados nas alças (*loops*) dos segmentos de expansão de *Drosophila* e *Bombyx* substituem motivos de 33 e 12 bases, respectivamente, ricos em A/U. As bases em negrito representam motivos de RNA conservados nos *loops* das três espécies. As setas assinalam as regiões de corte na molécula de rRNA, indicando a porção do segmento removido com o processamento.

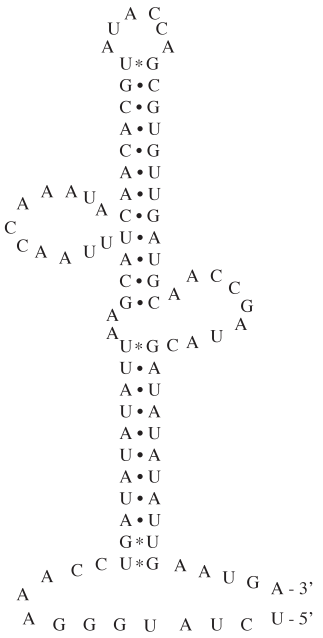


Figura 8.7. Estrutura secundária inferida para o domínio variável D7a de *Aedes albopictus* (modificada de Gorab *et al.*, 1995). Note a ausência do motivo UAAU (ver Figura 6.6) no segmento de expansão desse díptero, cujo rRNA 26S sofre o processamento ilustrado na Figura 6.6.

O que parece haver, nesse caso, é uma combinação de fatores que incluem, além da conservação de estrutura secundária, a própria composição de bases do segmento de expansão D7a. Um apoio a essa hipótese foi obtido com os dados de sequenciamento do rDNA do afídeo (pulgão) *Acyrthosiphon pisum*. Nesse inseto, a composição de bases do domínio divergente D7a, diferentemente daquela de insetos cujo rRNA 26S sofre o processamento nesta região (riqueza em AU), é rico em GC. Essa mudança na composição de bases do segmento de expansão D7a dessa espécie é acompanhada pela perda do processamento nesse segmento de expansão (Tabela 8.2). Maiores detalhes sobre a estrutura do rDNA dessa espécie podem ser vistos em Amako *et al.* (1996).

Os estudos feitos até o momento nesse campo sugerem que a seleção tem atuado positivamente no sentido de manter a estrutura secundária dos domínios divergentes. Isso também pode ser dito em relação ao processo de quebra da molécula de rRNA 26S em alguns organismos. Embora se saiba que o rRNA de espécies desprovidas desse processamento está perfeitamente adaptado a cumprir seu papel nos ribossomos, os estudos sugerem que deva haver alguma vantagem adicional para os organismos cujo rRNA seja processado no segmento D7a, o que explicaria a manutenção desse processo ao longo da evolução. Dados interessantes nesse sentido surgiram poucos anos atrás em estudos de síntese e processamento do rRNA de *Sciara coprophila* (Diptera: Sciaridae) realizado em *Xenopus* (detalhes em Basile-Borgia *et al*., 2005). Nesse sistema heterólogo, observou-se o processamento do rRNA no segmento de expansão D7a tal como ele ocorre no inseto. Embora esse processo não ocorra normalmente no rRNA de *Xenopus*, os resultados sugerem que a maquinaria responsável por ele esteja conservada em organismos evolutivamente distantes, como dípteros e anfíbios. Contudo, não se pode descartar uma possibilidade alternativa e atraente para explicar este fenômeno,

Tabela 8.2. A tabela mostra a riqueza em AU do segmento de expansão D7a de cinco espécies de insetos cujo rRNA sofre um processamento dividindo a molécula de rRNA 26S em metades α e β. *Acyrthosiphon pisum* é a única espécie da tabela cujo rRNA não sofre processamento. Os dados da tabela foram publicados em Gorab *et al.* (1995) e Amako *et al.* (1996).

Inseto	Porcentagem de AU e comprimento em bases do segmento de expansão D7a
Chironomus thummi	62 (131)
Aedes albopictus	67 (93)
Drosophila melanogaster	75 (92)
Bombyx mori	69 (77)
Sciara coprophila	72 (70)
Acyrthosiphon pisum	34 (60)

a de autocatálise no domínio divergente D7a.

Os genes de rRNA 23S/28S são exemplos de que a evolução dos genes ribossômicos—e provavelmente de outros genes—ocorre em outros níveis além daquele observável apenas com o alinhamento de sequências de diferentes espécies. Sendo apenas uma das possibilidades de análise molecular, a estrutura primária do RNA, analogamente à estrutura primária protéica, contém informações relevantes para o estudo da evolução estrutural dos transcritos e sua implicação em sua funcionalidade em organismos distintos.

Referências Bibliográficas

Amako, D., Kwon, O.-Y. e Ishikawa, H. (1996). Nucleotide sequence and presumed secondary structure of the 28S rRNA of pea aphid: implication for diversification of insect rRNA. **J. Mol. Evol. 43**: 489-475.

Basile-Borgia, A.E., Dunbar, D.A. e Ware, V.C (2005). Heterologous rRNA expression: internal fragmentation of *Sciara coprophila* 28S rRNA within microinjected *Xenopus laevis* embryos. **Insect Mol. Biol. 14**: 523-536.

Fujiwara, H. e Ishikawa, H. (1986). Molecular mechanism of introduction of the hidden break into the 28S rRNA of insects: implication based on structural studies. **Nucleic Acids Res. 14**: 6393-6401.

Gerbi, S.A., Jeppesen, C., Stebbins-Boaz, B. e Ares Jr, M. (1985). Evolution of eukaryotic rRNA: constraints imposed by RNA-RNA interactions. **Cold Spring Harbor Symp. Quant. Biol. 52**: 709-719.

Gorab, E., de Lacoba, M.G. e Botella, L.M. (1995). Structural constraints in expansion segments from a midge 26s rDNA. **J. Mol. Evol. 41**: 1016-1021.

Hancock, J.M., Tautz, D. e Dover, G.A. (1988). Evolution of the secondary structures and compensatory mutations of the ribosomal RNAs of *Drosophila melanogaster*. **Mol. Biol. Evol. 5**: 393-414.

Hancock, J.M. e Dover, G.A. (1990). 'Compensatory slippage' in the evolution of ribosomal RNA genes. **Nucleic Acids Res. 18**: 5949-5954.

Lewin, B. (1994) Gene numbers: Repetition and redundancy. **In Genes V**, Oxford University Press Inc. New York, 726 pp.

Michot, B., Qu L.-H. e Bachellerie, J.P. (1990) Evolution of large-subunit rRNA structure. **Eur. J. Biochem. 188**: 219-229.

Ware, V.C., Tague, B.W., Clark, C.G., Gourse, R.L., Brand, R.C. e Gerbi, A.S. (1983) Sequence analysis of 28 S ribosomal DNA from the amphibian *Xenopus laevis*. **Nucleic Acids Res. 11**: 7795-7817.

Ware, V.C., Renkawitz, R. e Gerbi, S.A. (1985). rRNA processing: removal of only nineteen bases at the gap between 28S and 28S rRNAs in *Sciara coprophila*. **Nucleic Acids Res. 13**: 3581-3597.

Zucker, M. e Stiegler, P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. **Nucleic Acids Res. 9**: 133-148.

Capítulo 9

O genoma instável, sequências genéticas móveis

Marie-Anne Van Sluys (mavsluys@usp.br)
Departamento de Botânica
Instituto de Biociências
Universidade de São Paulo

Nathalia de Setta (nsetta@hotmail.com)
Departamento de Botânica
Instituto de Biociências
Universidade de São Paulo

Katia C. Scortecci (kacscort@yahoo.com)
Depto Biologia Celular e Genética
Centro de Biociências, UFRN

Ana Paula Pimentel Costa (appcosta@mackenzie.br)
Centro de Saúde e Biologia
Univ. Presbiteriana Mackenzie

"*...e no entanto, se move.*" (Galileu Galillei)

"*Um experimento conduzido em torno de meados da década de quarenta (século XX) me preparou para aceitar respostas inesperadas do genoma frente a situações de choque para as quais o genoma não está preparado para enfrentar de forma ordenada e programada.*" (Barbara McClintock, abertura do discurso ao receber o Prêmio Nobel de Medicina, 8/12/1983)

9.1. Introdução

O genoma pode ser instável sem que isso seja deletério? É provável que T.H. Morgan, um dos fundadores da genética moderna, e Goethe, o pensador alemão, dissessem "Impossível!". Isso se daria devido ao fato que a instabilidade gera os mutantes que, *a priori*, foram reconhecidos como organismos não adaptados ao ambiente. Os genes, por conter as informações necessárias ao desenvolvimento de um organismo vivo, deveriam estar protegidos de processos que causam mudanças no genoma. Esse conceito de estabilidade foi a base do pensamento a partir da redescoberta dos trabalhos de Mendel. No entanto, conceitos recentes admitem que parte do genoma seja instável e é justamente parte dessa instabilidade que proporciona uma adaptação rápida ao ambiente quando o organismo está sujeito a condições que desafiam sua capacidade de sobrevivência. Associa-se à instabilidade a presença de sequências genéticas móveis, que podem estar representadas por elementos de transposição, foco deste capítulo, plasmídeos, vírus (incluem-se aqui os bacteriófagos) e ainda a região T dos plasmídeos de *Agrobacterium*. O uso do termo "elemento de transposição" em substituição a "elemento transponível" dá-se pela característica de mudança de lugar associada ao primeiro termo que abrange também a noção de que outros genes podem ser capturados no processo de mudança de lugar.

O que define e diferencia um elemento de transposição (TE, de *transposable element*, em Inglês) dos outros elementos mencionados acima é sua capacidade de mobilização no genoma de um organismo, induzindo mutações instáveis ao se inserirem em genes. As mutações geradas pela atividade dos elementos de transposição são geralmente recessivas no indivíduo selvagem, pois causam uma inativação parcial ou total do gene onde ocorre a inserção (Fedoroff, 1989). Associado à própria indução dessas mutações, verifica-se também que elas são instáveis, uma vez que pode haver a reversão dessa mutação no próprio organismo mutado. A Figura 9.1 ilustra a atividade desses elementos em flores e grãos. O fenótipo observado nessa figura é denominado de *variegação* e foi cunhado por Emerson em 1914, que o definiu como uma instabilidade fenotípica durante o desenvolvimento somático (Peterson, 1995). A observação das mutações instáveis produzidas pela mobilidade dos elementos de transposição no genoma é bem antiga. Darwin já havia descrito esse fenômeno em inflorescências de *Anthrrihinus majus* (boca-de-leão).

Na Figura 9.1D, é ilustrada de modo esquemático a inserção de um TE (representado como um triângulo) em um gene hipotético (retângulo cinza), cuja consequência é a inativação do gene. Quando o TE sai (excisa) do gene em questão, a atividade gênica é recuperada, observando-se, assim, setores compostos por células onde a atividade do gene volta ao fenótipo selvagem sobre um fundo mutado (Grandbastien, 1987; Fedoroff, 1989). A esse padrão de setores de fenótipo selvagem em fundos mutados dá-se o nome de fenótipo variegado.

A caracterização genética dos elementos de transposição iniciou-se na década de 1940, com os trabalhos de citogenética em milho desenvolvidos por Barbara McClintock. Ela denominou os TEs de "elementos controladores do gene" pelo fato de modularem a expressão dos genes aos quais estavam associados (McClintock, 1956). McClintock estudou rearranjos cromossômicos que promoviam quebras no braço curto do cromossomo 9 de milho (Figura 9.2A), uma vez que aí estavam localizados vários marcadores genéticos visíveis em grãos dessa espécie: o

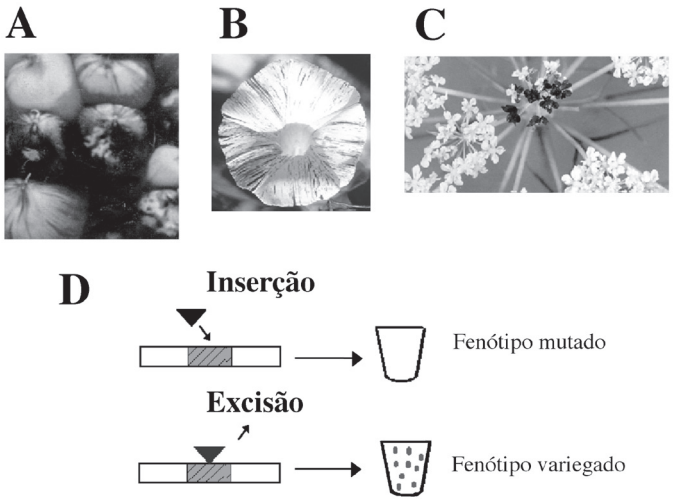


Figura 9.1. Fenótipos variegados. Representação esquemática do fenótipo variegado devido à atividade de um elemento de transposição. Em D, é representada a inserção de um elemento de transposição (triângulo cinza escuro) em um gene hipotético X (cinza claro), promovendo a inativação gênica, produzindo um fenótipo mutado. A excisão do elemento de transposição (triângulo cinza escuro) do gene X em cinza claro leva à expressão gênica normal em determinadas áreas, produzindo assim o fenótipo variegado.

loco *C1*, que define a coloração do pericarpo, o loco *Shrunken* (*Sh*), que confere o caráter rugoso, o loco *Bronze* (*Bz*), que afeta a coloração no aleurona, e o loco *Waxy* (*Wx*), responsável pela síntese de amido no endosperma.

As quebras cromossômicas observadas por McClintock ocorriam aleatoriamente, devido aos ciclos de quebra-fusão-ponte. A Figura 9.2B ilustra um ciclo onde se observa a formação de um cromossomo dicêntrico e outro acêntrico logo após a quebra da cromátide. Entretanto, em um dos casos analisados, essa quebra cromossômica ocorria sempre na posição proximal do cromossomo 9, próximo ao loco *Wx*. McClintock denominou essa posição de sítio *Ds* (*Dissociator*, desassociador, em inglês), devido à desassociação da cromátide. Ela verificou que a ocorrência desse evento segregava de acordo com as leis de Mendel. Reconstruindo os cruzamentos efetuados entre as diferentes linhagens, ela observou que, para que essa quebra proximal no cromossomo 9 ocorresse, eram necessários dois fatores: o sítio *Ds* e o sítio *Ac* (*Activator*, ativador, em inglês). Este último funcionava como um ativador da quebra. Em cruzamentos com outras linhagens,

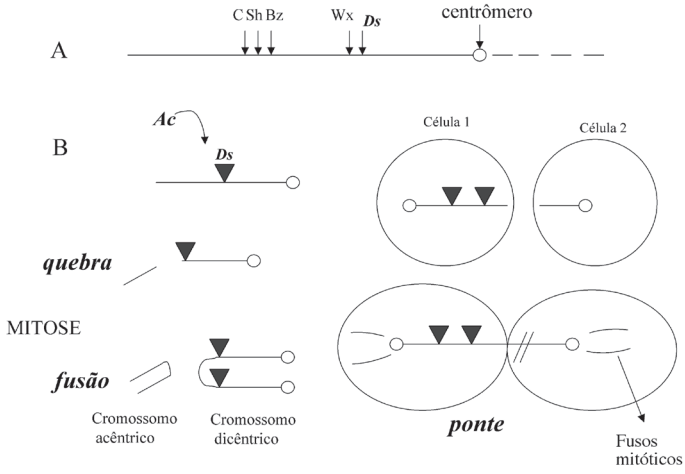


Figura 9.2. Detalhe do braço pequeno do cromossomo 9 de milho e ciclo de quebra-fusão-ponte. Em A, podem ser observadas as posições relativas dos locos C, Sh, Bz e Wx, relativos ao centrômero e a inserção do elemento de transposição Ds (*Dissociation*). Em B, é representado um ciclo de quebra-fusão-ponte causado pela mobilidade de Ds (triângulo cinza) induzido pela atividade de Ac.

ela observou novamente essas quebras cromossômicas próximas ao loco *Wx*. Esses resultados confirmaram a necessidade desses dois fatores: *Ac* e *Ds* para a ocorrência das quebras. Intrigada com o fenômeno, novos cruzamentos foram efetuados a partir de linhagens distintas daquelas já usadas e McClintock questionou a possibilidade de o sítio *Ds* se "mover" no genoma, mudando de posição entre os marcadores genéticos analisados. Desse modo, ela propôs que existiriam sequências no genoma que "saltariam", alterando a expressão de genes aos quais estivessem associadas (McClintock, 1956; McClintock 1984).

Os trabalhos de McClintock, entretanto, embora publicados em periódicos de renome, não foram aceitos na época de sua publicação, pois questionavam a idéia existente de que o material genético fosse estável. Somente após a caracterização da existência de sequências genéticas móveis em bactérias e leveduras, na década de 1960, em *Drosophila*, na década de 1970, e no homem, na década de 1980, que seu trabalho foi reconhecido pela comunidade científica. O reconhecimento definitivo ocorreu em 1983, quando McClintock ganhou o prêmio Nobel pela caracterização dos elementos de transposição no genoma do milho.

Desde a década de 1990, têm-se verificado que os TEs estão presentes em virtualmente todos os seres vivos, de modo que são considerados atualmente como elementos antigos associados ao genoma. As únicas exceções identificadas até o momento são algumas espécies de eubactérias, arqueobactérias e em *Plasmodium falciparum*, bem como, provavelmente, em espécies proximamente relacionadas a esse parasita. Por esse motivo, atribui-se a esses elementos uma importância evolutiva, uma vez que podem estar envolvidos na geração de variabilidade genética devido a sua mobilidade. Duas hipóteses são propostas para explicar a presença de elementos de transposição nos diferentes organismos (Starlinger, 1993). A primeira delas seria que esses elementos atuariam como genes egoístas, sequências presentes no genoma sem função determinada e que se manteriam ativos somente para se duplicar. Nesse caso, esses elementos seriam considerados "parasitas moleculares". A segunda hipótese postula o envolvimento desses elementos diretamente na evolução dos organismos que os portam, uma vez que rearranjos finos ou grosseiros podem ser produzidos com a inserção/excisão dos TEs. Essas hipóteses, no entanto, não são mutuamente exclusivas.

Consideram-se modificações finas aquelas que envolvem apenas um pequeno número de pares de base, muitas vezes associadas à cicatriz deixada pelo TE no processo de inserção/excisão (Saedler e Nevers, 1985). Modificações grosseiras são aquelas oriundas de rearranjos cromossômicos, onde podem ser geradas duplicações, deleções e transversões, entre outros processos (Peterson, 1993; Schwarz-Sommer *et al.*, 1985). Ambos podem ser considerados matéria-prima para a atuação de processos evolutivos, como agentes mutagênicos intrínsecos. Além disso, McClintock propõe que, em determinadas condições, esses elementos poderiam ser ativados em resposta a um estresse genômico e, com isso, permitiriam uma reestruturação do genoma hospedeiro. A resposta de ativação desses elementos em células estressadas, tecidos ou organismos poderia assegurar a sobrevivência em condições adversas. Em seu discurso de 1984, McClintock propôs que os elementos de transposição sejam componentes normais do genoma e que, por algum estresse biótico e/ou abiótico—como os causados por radiação ionizante ou ultravioleta (UV), infecção por vírus ou cultura *in vitro*—passam para uma fase ativa, caracterizada por sua mobilização no genoma hospedeiro.

Outro aspecto importante a ser considerado quanto a esses elementos é o fato de serem sequências repetidas no genoma. Em algumas plantas, os TEs chegam a perfazer até 50% do genoma total (Kumar e Bennetzen, 1999). Dependendo das características

dos TEs, eles podem ser moderadamente repetitivos, quando são mais de dez cópias por genoma, ou altamente repetitivos, quando esse número é maior que 250.000 cópias por genoma. Em alguns casos, os elementos altamente repetidos são encontrados associados especificamente aos centrômeros ou aos telômeros. Sugere-se, nesses casos, que esses TEs exerçam um papel estrutural (Pardue *et al.*, 2005, Wong e Choo, 2004, Lippman e Martiessen 2004). Na maioria dos casos, entretanto, os elementos de transposição estão dispersos no genoma hospedeiro sem uma função aparente (Bennetzen, 2000).

9.2. Estrutura dos Elementos de Transposição

Os elementos de transposição podem mover-se no genoma de um organismo por meio de intermediários de RNA ou de DNA. A classificação mais ampla desses elementos é feita de acordo com o tipo de intermediário, sendo definidos como pertencentes à Classe I aqueles que possuem intermediários de RNA e como pertencentes à Classe II, aqueles cujos intermediários são moléculas de DNA. Com o aumento de genomas sequenciados e a percepção de que esses elementos estão por toda parte no genoma dos seres vivos, um esforço de normatização de nomenclatura foi proposta recentemente (Wicker *et al*., 2007). Há um debate atual na literatura quanto a sua adequação, mas esta segue como proposta.

Os elementos pertecentes à Classe I também são conhecidos como retroelementos. A família dos retroelementos é composta de retrovírus e retrotransposons. A organização geral dos retrovírus e retrotransposons é muito similar, com a diferença que os retrotransposons não são infecciosos. Só os retrovírus geram partículas infecciosas, que podem se mover entre células. Essa característica deve-se à presença do domínio **ENV** funcional, que codifica uma glicoproteína que faz parte do envelope protéico da partícula viral, resultando em seu caráter infeccioso.

Os retrotransposons já foram descritos em todos os organismos eucariotos vivos. Os representantes mais conhecidos são os elementos *copia* e *gypsy* de *Drosophila* (Bingham e Zachar, 1989) e os elementos *Ty* de levedura (Boeke e Corces, 1989). Nos vegetais, os retrotransposons são muito representativos. Na maioria dos casos, eles estão presentes em um alto número de cópias, representando um dos maiores componentes do genoma vegetal (Kumar e Bennetzen, 1999).

Os retrotransposons podem ser divididos em dois tipos, considerando a presença ou não de longas repetições terminais (**LTR**). Os retrotransposons com LTR são flanqueados por repetições terminais longas e codificam todas as proteínas necessárias para sua transposição. Podem ser separados em dois grupos, com base na organização dos domínios protéicos semelhantes aos elementos considerados como plesiomórficos dentro do grupo—os elementos *copia* e *gypsy*, encontrados em *Drosophila*, e os elementos *Ty1* e *Ty3*, em levedura.

Os elementos *Ty1* e *copia* carregam dois quadros abertos de leitura (ORFs, de *open reading frames*, em inglês). O primeiro codifica uma proteína (**GAG**) semelhante ao capsídio protéico viral, que contém um ou mais motivos ricos em cisteína associados a sítios de ligação de ácidos nucléicos. O segundo codifica para uma poliproteína (**POL**) com atividade de protease (**PROT**), envolvida na maturação de diferentes proteínas: transcriptase reversa (**RT**), codificando a enzima responsável pela criação da cópia de DNA da fita molde de RNA; integrase (**INT**), codificando proteínas necessárias para a integração da cópia de DNA no genoma do hospedeiro; e RNAseH (**RH**), também envolvida na transcrição reversa. A superfamília *copia* difere da superfamília *gypsy* por ter o domínio INT anterior ao domínio RT, enquanto que a última apresenta o domínio INT posterior a RT (Figura 9.3).

As LTRs podem ser funcionalmente definidas como elementos promotores e terminadores da transcrição, pois começa na LTR 5' e termina na LTR 3'. Essa porção é dividida em três regiões: **U3** (unique 3'), **R** (repeat), **U5** (unique 5'). Internamente após cada LTR, existem sequências correspondentes a sítios iniciadores utilizados para a síntese de DNA pela transcriptase reversa. Como no caso da DNA polimerase, a RT necessita de pequenos oligonucleotídeos que forneçam pontas 3'OH passíveis de incorporar um nucleotídeo para iniciar a extensão (síntese) da nova molécula de DNA. Após a LTR 5', encontra-se o sítio de ligação do iniciador, ou PBS (*primer binding site*, em inglês), similar ao tRNA da metionina, que serve, portanto, de iniciador da síntese. A síntese da primeira fita de DNA estende-se até a ponta R do mensageiro. Antes da LTR 3', encontra-se uma sequência de polipurinas, ou PPT (*polypurine tract*, em inglês), responsável pelo início da síntese da segunda fita de DNA.

Como retrotransposons sem LTRs, temos dois grupos: o grupo LINE (*long interspersed nuclear element*, em inglês) apresenta os mesmos domínios protéicos que os retrotransposons com LTR, tendo no lugar da LTR 3' uma cadeia de adenosinas. O segundo grupo é chamado de SINE (*short interspersed nuclear element*, em inglês). Esses elementos não codificam sua própria transcriptase reversa, tendo uma grande similaridade de sequência com o genoma do hospedeiro (Smyth, 1993; Figura 9.3).

Os elementos propagados diretamente como moléculas de DNA (Classe II) podem ser divididos em duas subclasses: Subclasse I e II. Os elementos da Subclasse I exibem algumas características estruturais comuns: apresentam, nas suas extremidades, sequências inversamente repetidas. Essas sequências são denominadas de **TIR** (*terminal inverted repeat*, em inglês) e podem variar de 10 a 50 pb (pares de base). Esses elementos de transposição, quando se inserem no genoma, promovem duplicação de alguns pares de base no local de inserção (Figura 9.3) e, com sua excisão desse local, podem deixar uma cicatriz, isto é, alguns pares de bases resultantes da duplicação produzida no momento de inserção da sequência. Assim, o número de pares de base das IRs e da duplicação promovida no local de inserção são características de cada elemento de transposição via DNA (Feschotte, 2002). Quanto a sua distribuição, os elementos de Classe II já foram identificados tanto em procariotos, como em eucariotos. Dentre os elementos mais estudados, podemos citar o *P* e o *mariner* de *Drosophila*, bem como o *Ac* e o *Mu* de milho.

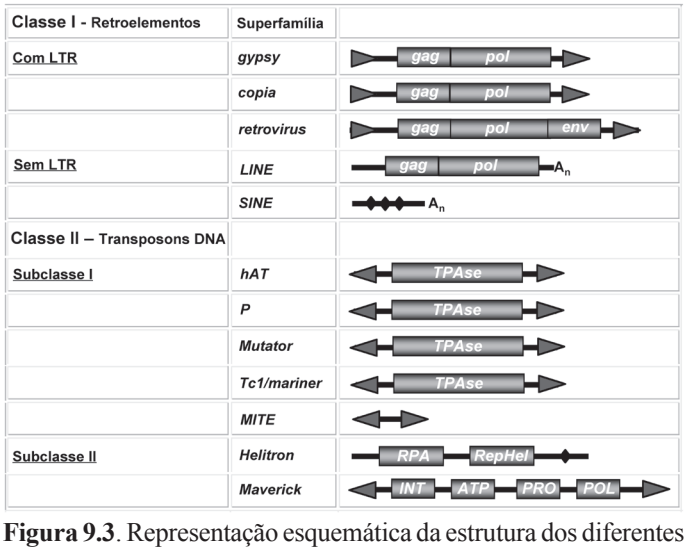


Figura 9.3. Representação esquemática da estrutura dos diferentes elementos de transposição das Classes I e II. Para detalhes das estruturas, dirija-se ao texto.

Acredita-se que os diferentes elementos de transposição da Subclasse I mobilizam-se por um mecanismo muito semelhante. O RNA mensageiro transcrito codifica para uma proteína com função de transposase (**TPAse**). Essa proteína é responsável pelo processamento da excisão e inserção do elemento de transposição no genoma hospedeiro. Inicialmente, a TPAse liga-se às extremidades do elemento, principalmente na região das IRs, aproximando as duas extremidades e, então, produz um corte na dupla fita de DNA, liberando esse elemento de transposição. A TPAse é responsável também pela inserção desse elemento em uma nova posição no genoma (Haren *et al*., 1999).

Recentemente foram identificados dois novos tipos de elementos de Classe II: os *Mavericks* e os *Helitrons*. Esses TEs foram classificados em uma nova subclasse de elementos (Subclasse II), pois, embora o mecanismo de mobilização ocorra com moléculas de DNA, eles não utilizam uma TPAse para mediar a transposição. Os *Mavericks* codificam diversas proteínas, variando entre espécies e inserções. Como padrão, são observados genes para INT, proteínas empacotadoras dependentes de ATP (**ATP**), **PROT** e DNA polimerase B (**POL**). No caso dos *Helitrons*, dois tipos de proteínas são codificadas: uma similar à proteína de replicação A (**RPA**) e, outra, com os domínios Helicase e iniciador de Replicação (**RepHel**). Uma característica diferencial dos *Helitrons* é a falta de TIRs, o que dificulta sua identificação pelos métodos convencionais de estudo de elementos de transposição por bioinformática. Esses dois tipos de elementos também são amplamente distribuídos, tendo sido identificados em diversas espécies de vertebrados, invertebrados, plantas e fungos.

9.3. A Grande Diversidade dos Elementos de Transposição

A ampla distribuição dos elementos de transposição nos seres vivos é acompanhada por sua alta representatividade em seus genomas. Apesar de ser sempre significativa, essa representatividade é variável de acordo com o genoma nos quais estão inseridos e com o tipo de elemento considerado. No primeiro caso, de modo geral, espécies com grandes genomas tendem a acumular altas proporções de TEs e espécies com genomas menores abrigam menores de concentrações dessas sequências. Por exemplo, a Figura 9.4 demonstra a proporção de elementos de transposição nos genomas de alguns eucariotos. Enquanto o genoma de *Caenorhabditis elegans* possui cerca de 12 Mpb e apenas 3% de elementos de transposição, o genoma de milho, que é cerca de 200 vezes maior (~2500 Mpb), pode ter até 70% de seu conteúdo composto por TEs, dependendo do cultivar analisado. Adicionalmente, é estimado que cerca de 44% do genoma humano é representado por elementos de transposição. Quando se avalia a diversidade de TEs em cada genoma, é observado que existem tanto elementos representados por apenas algumas poucas cópias—como é o caso de alguns retrotransposons de *Drosophila melanogaster* (Mugnier *et al.*, 2008) —, como outros acumulando entre centenas e milhares de inserções—notadamente os elementos *L1* e *Alu*, que juntos perfazem cerca de 25% do genoma humano (Deininger e Batzer, 2002). É imprescindível ressaltar que, nas espécies com altas proporções de TEs, grande parte desses elementos permanecem inativos, tanto por apresentarem degradação de sequências codificantes, como por mecanismos regulatórios dos genomas hospedeiros. Dessa forma, a diversidade dos elementos de transposição apresenta-se como uma intrigante questão, que tem motivado diversos estudos focando sua variabilidade, distribuição e evolução nos mais diversos seres vivos. Abaixo são descritos alguns dos grupos de TEs já caracterizados, suas principais características e espécies hospedeiras.

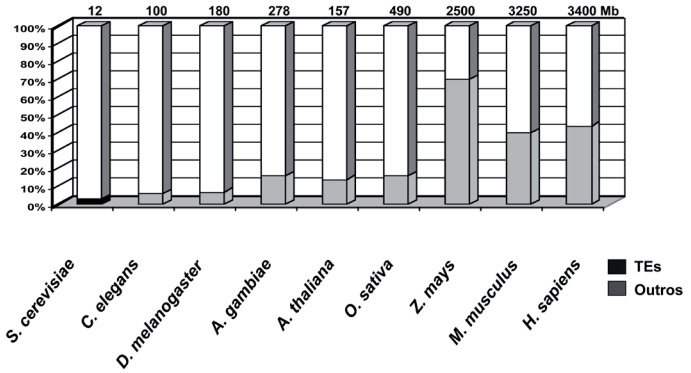


Figura 9.4. Representatividade (em %) de elementos de transposição nos genomas de algumas espécies-modelo. As barras listradas e lisas correspondem à porcentagem de TEs e de outro tipos de sequências, respectivamente. Acima de cada barra, está indicado o tamanho (em Mb) de cada um dos genomas analisados. Dados baseados nos estudos de Kim *et al*. (1998), Waterston e Sulston (1995), International Human Genome Sequencing Consortium (2001), *Arabidopsis* Genome Initiative (2000), Smith *et al.*, 2007, Holt *et al*. (2002), Yu *et al*. (2002), SanMiguel *et al*. (1996), Meyers *et al*. (2001), Mouse Genome Sequencing Consortium (2002).

9.3.1. Retrotransposons de plantas

Os retrotransposons só começaram a ser descritos em plantas após sua caracterização em *Drosophila* e levedura. A partir daí, vários retrotransposons têm sido descritos em diversas espécies vegetais. A maioria possui LTRs e assemelha-se a uma ou outra superfamília de retrotransposons: *copia* e *gypsy* (Tabela 9.1).

Vários elementos já foram completamente sequenciados, como o elemento *Ta1* de *Arabidopsis thaliana* (Voytas e Ausubel, 1988), *Tnt*1 do fumo (Grandbastien *et al.*, 1989) e *Bare*-1 de cevada (Mannienen e Schulman, 1993). Todos pertencem à superfamília *copia* e foram descobertos como fontes de polimorfismo do DNA (elemento *Ta1*) ou como inserções gênicas (*Tnt1*, *Bare-1*) causando mutações.

Dois grupos de elementos, pertencentes à superfamília *gypsy*, foram identificados por sua abundância no genoma. O elemento *Del*-1 (Smyth *et al.*, 1989) ocorre em mais de 13 mil cópias em *Lilium henry*, o equivalente a 4% de seu genoma. Esse elemento também aparece abundante em outras espécies do gênero. O elemento *IFG* de *Pinus* (Kossack e Kinlaw, 1999) também aparece com no mínimo dez mil cópias.

Existem elementos descritos cujas relações ainda não estão claras. O elemento *Cin-1* no milho (Shepherd *et al.*, 1984), por exemplo, apareceu abundantemente, mas como um resquício reduzido de seu ancestral. O elemento *Bs1* (Jin e Bennetzen, 1989), também do milho, descoberto por sua inserção no gene da desidrogenase alcoólica durante uma infecção virótica, apresentou-se tão rearranjado e mutado que dificilmente seria ativo de maneira autônoma. No trigo, o elemento *WIS-2* (Harberd *et al.*, 1987) foi descrito a partir de sua inserção no gene de uma das subunidades da gluteína. Esse elemento estava presente apenas em alguns cultivares do trigo, sugerindo que sua ocorrência no genoma do trigo seja recente. Uma nova família de elementos, com forte homologia com o elemento *Tnt1* de *Nicotiana*, foi descrita para o gênero *Lycopersicon*. O elemento *Retrolyc1* foi isolado originalmente em uma espécie selvagem de tomate (*L. peruvianum*), mas demonstra ser bem representativo nas demais espécies do gênero (Costa *et al.*, 1999). Foram encontradas evidências de uma possível atividade (Araújo et al, 2000), mas novas análises estão sendo realizadas para confirmação.

Tabela 9.1. Retrotransposons vegetais mais bem caracterizados (baseado em Bennetzen, 1996).

NOME	ESPÉCIES	TOTAL (Kpb)	Nº CÓPIAS	SUPERFAMÍLIA*	REF.
COM LTRs					
Bs1	*Zea mays*	3,2	0-5	———	Jin e Bennetzen, 1989
Zeon-1	*Zea mays*	8,6	250-1250	———	Hu *et al.*, 1995
Hopsoctch	*Zea mays*	4,8	nd	*copia*	White *et al.*, 1994
Magellan	*Zea mays*	5,7	nd	*gypsy*	Purugganan e Wessler, 1994
Prem-2	*Zea mays*	9,4	30000	*copia*	Turcich *et al.*, 1996
WIS-2	*Triticum aestivum*	8,6	200	*copia*	Moore *et al.*, 1991
Tst-1	*Solanum tuberosum*	5,1	1	*copia*	Camirand *et al.*, 1990
PDR-1	*Pisum sativum*	5,0	50	nd	Lee *et al.*, 1990
IFG7	*Pinus radiata*	5,9	10000	*gypsy*	Bennetzen, 1996
Tos 3-1	*Oryza sativa*	5,2	> 30	nd	Hirochika *et al.*, 1992
Tto-1	*Oryza sativa*	5,5	30 - 300	nd	Hirochika, 1993
Tnt1	*Nicotiana tabacum*	5,3	> 100	*copia*	Grandbastien *et al.*, 1989
del-1	*Lilium henry*	9,3	> 13000	*gypsy*	Smyth *et al.*, 1989
BARE-1	*Hordeum vulgare*	12,1	5000 - 50000	*copia*	Mannienen e Schulman, 1993
Athila	*Arabidopsis thaliana*	10,5	> 30	———	Pelissier *et al.*, 1995
Ta1	*Arabidopsis thaliana*	5,2	1 -3	*copia*	Voytas *et al.*, 1990
Retrolyc1	*Lycopersicon peruvianum*	> 4,0	nd	*copia*	Costa *et al.*, 1999
SEM LTR					
Cin4	*Zea mays*	1-6,8	50-100	———	Schwarz-Sommer *et al.*, 1988
Ts	*Nicotiana tabacum*	0,111- 0,258	50 000	———	Yoshioka *et al.*, 1993
del-2	*Lilium speciosum*	4,45	250000	———	Leeton e Smyth, 1993

*nd, não determinado.

9.3.2. Superfamílias dos elementos via DNA (Classe II)

Superfamília *hAT*

São classificados nessa família os elementos *Ac*/*Ds* (*Zea mays*), *Tam3* (*Antirrhinum majus*), *hobo* (*D. melanogaster*). O elemento *Ac* corresponde àquele caracterizado geneticamente por McClintock. Esse elemento apresenta 4.565 pb, TIRs com 11 pb e produz uma duplicação no local de inserção de 8 pb. Até o momento, foi caracterizado um único mRNA de 3,5 kb, cuja função está relacionada à proteína TPAse (Kunze, 1996). O elemento *Ac* corresponde a um tipo de elemento de transposição denominado **autônomo**, isto é, TEs que contêm em sua sequência toda a informação necessária para sua mobilização. Esses elementos autônomos são capazes de induzir a mobilização em *trans* de uma classe de elementos não autônomos, ou defectivos. Os elementos não autônomos só são capazes de mobilizar-se quando na presença em *trans* dos elementos autônomos, que fornecem toda a maquinaria protéica necessária para sua mobilização. Os elementos *Ds* correspondem à classe dos elementos **não autônomos**.

O elemento *Tam3*, isolado de boca-de-leão, apresenta características estruturais semelhantes ao elemento *Ac*, também possuindo TIRs com 11 pb e e produzindo uma duplicação de 8 pb durante sua inserção. Esse elemento corresponde ao grupo dos elementos autônomos, possui toda a informação necessária para sua mobilização e é capaz de mobilizar em *trans* os elementos não autônomos.

Superfamília *mariner-Tc1*

Essa família está amplamente distribuída entre os organismos, de protozoários a vertebrados, tendo sido caracterizada em invertebrados, vertebrados e fungos. Os elementos dessa família produzem uma duplicação no sítio de inserção de 2 pb.

O elemento *mariner* foi descrito em *Drosophila mauritiana* através de mutações instáveis no gene *white*. Esse elemento apresenta 1.300 pb, com TIRs de 28 pb. O elemento *Tc1* foi caracterizado no nemátodo *C. elegans*. Seu tamanho é de 1.600 pb e apresenta TIRs de 54 pb. Os elementos dessa família têm sido utilizados como ferramentas para análises genéticas, como a identificação e clonagem de genes e mapeamento gênico (Plasterk, 1996).

Superfamília *P*

O elemento *P* foi caracterizado pela primeira vez em *D. melanogaster* e seu tamanho é de 2.907 pb nessa espécie, com TIRs de 31 pb. Esses correspondem a elementos autônomos, com toda a informação necessária em sua sequência para sua mobilidade. Existem vários elementos *P* não autônomos ou defectivos originados principalmente por deleções internas naturais. Esses elementos não autônomos necessitam da presença em *trans* da cópia autônoma fornecendo as proteínas para a sua mobilização.

O número de cópias desse elemento no genoma de *D. melanogaster* varia de nenhuma a 60 cópias. Esse elemento está associado ao sistema P-M de disgenesia híbrida, descrita no item 9.4. (Engels, 1996).

Superfamília *Mutator*

Dentre os TEs dessa superfamília, o elemento *Mu* foi identificado em *Z. mays* por Robertson em 1978 em linhagens que apresentavam uma alta frequência de mutação. Muitas dessas novas mutações eram recessivas e estavam associadas à instabilidade somática, sugerindo que essa alta frequência de mutações pudesse estar associada a mutações instáveis descritas por Barbara McClintock. Estudos moleculares reforçaram essa idéia e permitiram comprová-la através de sua clonagem (Bennetzen, 1996).

A superfamília *Mutator* pode ser subdividida em seis subfamílias, todas com TIRs semelhantes, com 200 pb, produzindo 9 pb de duplicação no local de inserção. O elemento *MuDr* corresponde ao elemento autônomo dessa superfamília.

Superfamília *En/Spm* ou CACTA

Essa superfamília inclui os elementos *En/Spm*, de *Z. mays*, *Tam1/Tam2*, de *A. majus*, *Tgm1*, de boca-de-leão, e *Tnr3*, de arroz. Todos esses elementos de transposição apresentam TIRs com 13 pb, que começam com a sequência CACTA e produzem uma duplicação de 3 pb no local de inserção no genoma hospedeiro. Existem outros elementos que se mobilizam via DNA, que podem ser subdivididos em *Foldback* e *MITEs*. Os elementos da família *Foldback* apresentam longas TIRs, que podem formar uma estrutura semelhante a um grampo. Esses elementos foram primeiro descritos em *Drosophila*, mas já foram caracterizados também em ouriço-do-mar (elementos *TU*) e no nemátodo *C.*

elegans (elemento *Tc4*). Os *MITEs* correspondem a uma família bem menos caracterizada. A estrutura desses elementos sugere que eles pertençam à Classe II e sejam TEs não autônomos. Entre os *MITEs* já caracterizados, estão os elementos *Tourist* e *Stowaway*. São elementos pequenos, com 100 a 350 pb, altamente repetitivos, que não apresentam sequências codificantes e possuem um alto potencial de formação de estruturas secundárias.

Elementos de bactérias

Os elementos de transposição em bactérias são denominados como ***IS*** (sequências de inserção). Eles têm sido caracterizados em diferentes bactérias, tendo sido encontrados inclusive em arqueobactérias. Seu tamanho varia de 800 pb a 2500 pb e estão subdivididos em diferentes famílias com base em sua sequência e similaridades funcionais. Os elementos *IS* que carregam, além das funções necessárias para sua mobilidade, genes de resistência a antibióticos são denominados de *Tn*. O *Tn3* é o elemento de transposição que apresenta resistência à ampicilina, o *Tn10*, a resistência à tetraciclina e o *Tn5*, a resistência à canamicina.

9.4. Dinâmica e Controle da Mobilização

A variabilidade na proporção de elementos de transposição nos genomas hospedeiros está associada ao tipo de mecanismo de transposição utilizado por esses elementos—por meio de RNA ou de DNA. Os elementos que se mobilizam através do DNA apresentam dois mecanismos de transposição: transposição replicativa e transposição não replicativa. O mecanismo de transposição replicativo normalmente é encontrado nos elementos de bactérias e consiste da duplicação do elemento no momento de transposição. Essa cópia duplicada insere-se numa nova posição no genoma. A transposição não replicativa em geral é encontrada nos elementos transponíveis em plantas. Quando o elemento se mobiliza, esse elemento excisa-se do sítio doador, deixando uma cicatriz, e esse mesmo elemento integra-se em um novo local no genoma, sem que ocorra a duplicação, formando uma nova cópia. Essa inserção do elemento de transposição pode ocorrer em locais onde já ocorreu ou não a replicação cromossômica (Saedler e Nevers, 1985).

A mobilidade dos elementos de transposição necessita ser controlada ativamente para evitar um rápido acúmulo de mutações que podem ser letais para o organismo hospedeiro (Labrador e Corces, 1997). Dessa forma, os TEs, juntamente com seus hospedeiros, encontraram diferentes mecanismos regulatórios para controlar sua mobilidade, tanto em nível transcricional como em nível pós-transcricional. Em geral, o principal ponto de controle é a expressão das proteínas de mobilização.

Um mecanismo de controle de transposição menos complexo é encontrado na família *Tc1* em *C. elegans*. A TPAse é a única proteína necessária para sua mobilização e sua regulação controla a transposição. Tem sido observado que o aumento de transposição desse elemento está correlacionado ao nível de expressão da TPAse, assim como a fatores genéticos do hospedeiro (Labrador e Corces, 1997).

O elemento de transposição *mariner* também é regulado pelo nível de expressão da proteína TPAse. A alta produção dessa proteína reduz significativamente sua mobilidade (Capy *et al.*, 1998).

No caso do elemento *Activator* (*Ac*) em *Z. mays*, sua atividade pode ser regulada de duas maneiras: associada ao número de cópias ativas no genoma (denominada "efeito de dose") ou relacionada à presença ou ausência de metilação (presença de grupamentos metil na molécula de citosina) em determinadas sequências do elemento.

No primeiro caso, o fenômeno ainda não está totalmente esclarecido, sabendo-se apenas que, com o aumento de cópias ativas de *Ac* no genoma, observa-se um efeito negativo, uma diminuição na frequência e atraso no momento em que a transposição ocorre durante o desenvolvimento da planta (Fedoroff *et al.*, 1983). Propõe-se que esse efeito negativo de dose seria em nível pós-traducional, pois a molécula da TPAse deve interagir com ambas as extremidades do elemento para poder iniciar sua excisão/transposição (Fusswinkel *et al.*, 1991). Assim, com o aumento da concentração de proteínas, haveria uma competição entre as moléculas da TPAse pelo sítio de ligação (Schwartz, 1984, 1989). Heinlein *et al.* (1994) observaram que, aumentando a expressão da TPAse, havia a formação de agregados no citoplasma, que poderiam sequestrar a TPAse ativa, mantendo uma baixa frequência de transposição, apesar da alta produção dessa proteína.

O segundo modo de controle do elemento de transposição de *Ac* seria pela presença de metil-citosina em determinadas sequências dentro do elemento. A variação no padrão de metilação promove uma oscilação entre uma fase ativa e uma fase inativa (Chandler e Walbot, 1986; Chomet *et al.*, 1987). A região mais sensível do elemento *Ac* à metilação corresponde à sequência líder não traduzida, que se localiza na posição 5' do RNA mensageiro (5' ULR, 5' *untranslated leader region*, em inglês) (Kunze e Schwartz, 1988). O aumento da metilação nessa região está associado à inibição em nível de transcrição do elemento, uma vez que não se detecta o RNA mensageiro de 3,5 kb e, portanto, a proteína TPAse não é traduzida (Kunze e Schwartz, 1988).

Estudos em plantas transgênicas de *Nicotiana tabacum* e *Arabidopsis thaliana* utilizando a região promotora do elemento *Ac* mais a região 5' ULR fusionadas ao gene repórter LUC (gene da luciferase de vaga-lume) mostraram que essa região exerce um efeito negativo sobre o promotor *Ac*. O efeito negativo foi observado também sobre promotores fortes, como o promotor do vírus da couve-flor, CaMV 35S, utilizando sistemas de expressão transiente (Scortecci *et al.*, 1999). Esse efeito inibitório sobre uma região promotora provavelmente está associado à presença de uma estrutura secundária estável, formada pela presença da região 5' ULR no RNA mensageiro. A presença dessa estrutura reduz a eficiência de tradução desse RNA mensageiro, sendo um controle em nível pós-transcricional. Esses resultados sugerem que a região 5' ULR é uma sequência importante no controle da atividade desse elemento de transposição, seja em nível transcricional (pela presença de metilação) ou pós-transcricional (pela presença de uma estrutura secundária no RNA mensageiro; Scortecci *et al*., 1999).

O elemento *Spm* de *Z. mays* apresenta um único transcrito identificado com 2,4 kb. Esse transcrito é processado originando quatro diferentes transcritos, *tnpA*, *tnpB*, *tnpC* e *tnpD*. Esses são traduzidos nas proteínas TnpA, TnpB, TnpC e TnpD, respectivamente. Foi verificado que as proteínas TnpA e TnpD são necessárias para a transposição desse elemento. *Spm* também apresenta um mecanismo de controle de sua atividade pela oscilação entre uma fase ativa e outra inativa, através da variação no padrão de metilação, tanto em sua região promotora, como em sua região líder não traduzida (Fedoroff e Chandler, 1994). Tem sido observado um papel regulatório importante da atividade do elemento *Spm* pela proteína TnpA. Essa proteína pode ativar não somente a região promotora previamente metilada, promovendo sua desmetilação e permitindo, dessa maneira, uma expressão constitutiva do elemento *Spm*, como também pode ligar-se à região promotora, inibindo sua expressão e promovendo a metilação dessa região. Dessa forma, a regulação do elemento *Spm* está intimamente correlacionada ao nível existente de TnpA na

células, podendo ativar ou inativar a expressão desse TE (Fedoroff e Chandler, 1994).

Os elementos de transposição podem ter sua atividade controlada não apenas por fatores genéticos, mas também por fatores ambientais. Por exemplo, o elemento *Tam3* tem sua atividade controlada pela temperatura. Uma diminuição da temperatura de 25 ºC para 15 ºC aumenta sua frequência de transposição em torno de 1.000 vezes; para os elementos *Tam1* e *Tam2*, em 10 vezes (Coen, 1989).

Um outro elemento de transposição que apresenta mobilidade controlada por fatores genéticos e ambientais é o elemento *P* de *D. melanogaster*. Esse elemento produz uma síndrome denominada disgenesia híbrida, que se manifesta pela má formação das gônadas em função da temperatura, resultando em esterilidade da progênie, altas taxas de mutação, rearranjos e recombinações cromossômicas. Em geral, essa síndrome é observada quando ocorre a transposição do elemento *P* na progênie de linhagens derivadas de machos da linhagem P (que apresentam elementos *P* autônomos) e fêmeas da linhagem M (sem esses elementos). A progênie F1 tem um citoplasma tipo M e cromossomos com cópias do elemento *P*, derivado dos machos. O cruzamento reverso, isto é, entre fêmeas da linhagem P com machos da linhagem M, origina descendentes com traços de disgenesia híbrida bem reduzidos. Esse padrão é observado porque os ovos das fêmeas da linhagem P contêm altas concentrações de proteínas repressoras (produzidas pelo próprio elemento), que inibem a transcrição da TPAse e, consequentemente, a mobilização. Por outro lado, nos ovos das fêmeas da linhagem M, que não contém o repressor, pode ocorrer a transposição dos elementos herdados dos machos, levando à síndrome. Esse fenômeno é mais pronunciado em fêmeas e em temperaturas acima de 25 ºC (Engels, 1996). Estudos recentes têm demonstrado que o silenciamento do elemento *P* é mediado pelo mecanismo de RNA de interferência e que moléculas de RNAi específicas oriundas do próprio elemento *P*—presente nas linhagens P—inibiriam a produção de sua TPAse, silenciando o elemento. No entanto, a ausência dessas moléculas de RNAi nas progênies de fêmeas M poderia ser um importante fator no desencadeamento da disgenesia híbrida (Brenneck *et al.*, 2008).

Nos elementos da Classe I, o mecanismo de transposição é replicativo e basicamente o mesmo em retrotransposons de plantas, insetos e fungos. Muitos aspectos desse mecanismo também são comuns aos retrovírus. As etapas de transposição envolvem inicialmente a transcrição do elemento integrado no genoma do hospedeiro em fitas de RNA menssageiro (mRNA) pela polimerase do RNA celular (Figura 9.2). Essa transcrição é seguida pela tradução de parte das moléculas do mRNA em proteínas estruturais e funcionais necessárias para a transposição, incluindo a protease, a transcriptase reversa e a integrase. A protease, codificada pelo próprio elemento, está envolvida no processamento dos produtos primários da tradução. O empacotamento do transcrito ocorre em uma partícula semelhante ao vírus ou vírion (codificado pela GAG), juntamente com a transcriptase reversa e a integrase. A transcriptase reversa converte o RNA em uma fita linear de DNA (fita +) e, a partir desta, a síntese da segunda fita (fita -), ocorre pela ação de polimerases celulares. Esse DNA extracromossomal direciona-se para o núcleo da célula, a enzima integrase sendo responsável por sua inserção no genoma do hospedeiro. O elemento integrado normalmente será transcrito pela maquinaria celular (Fosket, 1994; Lewin, 1994; Boeke e Corces, 1989; Figura 9.2).

A transposição não é um processo aleatório. Em todos os sistemas eucarióticos, a retrotransposição é controlada pelo próprio elemento, por sinais dependentes do organismo hospedeiro e por fatores externos (Grandbastien *et al.*, 1997). Em *Drosophila* e leveduras, a transcrição e a transposição estão sob o controle de fatores hormonais, dependentes do desenvolvimento, podendo ser ativados por vários estresses e mudanças ambientais (Boeke e Corces, 1989).

Em plantas, de modo geral, a expressão dos retrotransposons é fortemente controlada. Observa-se a ausência da transcrição durante o desenvolvimento na maioria dos tecidos vegetais, os transcritos sendo detectados somente em tecidos específicos. A presença de transcritos para o elemento *Tnt1* é detectada apenas em raízes (Pouteau *et al.*, 1991), no endosperma para *Zeon-1* (Hu *et al.*, 1995) e em micrósporos jovens para o elemento *Prem-2* do milho (Turcich *et al.*, 1996). A expressão dos elementos *Tnt1* e *Tto1* do fumo e do elemento *Tos* do arroz é fortemente aumentada por estresses como o isolamento de protoplastos, cultura de células ou ataque de patógenos (Grandbastien *et al.*, 1997). A expressão do *Tnt1* em fumo foi estudada através da análise transcricional e pela análise da expressão de um gene repórter, colocado sob controle de uma das LTRs. Essa região é conhecida por conter o promotor e as sequências reguladoras dos retrotransposons (Casacuberta e Grandbastien, 1993). Os estudos demonstraram que a expressão do *Tnt1* está fortemente controlada quanto à especificidade dos tecidos e à etapa do desenvolvimento, indicando que o controle transcricional deve ser o passo principal para a regulação da transposição (Grandbastien *et al.*, 1994; Moureau-Mhiri *et al.*, 1996; Grandbastien *et al.*, 1997). Foi observado que a região das LTRs estaria envolvida nessa regulação (Casacuberta e Grandbastien, 1993). No fumo, nenhuma expressão foi detectada na maioria dos tecidos de plantas maduras, exceto nas raízes (Grandbastien *et al.*, 1994).

Esse elemento já foi transferido com sucesso para plantas transgênicas de *A. thaliana* (Lucas *et al.*, 1995) e de tomate (Moureau-Mhiri *et al.*, 1996). Estudos sobre a expressão do *Tnt1* nessa planta indicam a existência de uma regulação similar à do fumo. O elemento é fracamente expresso em tecidos foliares jovens, mas claramente expresso em raízes e tecidos foliares maduros (Moureau-Mhiri *et al.*, 1996). Em plantas de fumo, nenhuma expressão de *Tnt1* foi detectada em órgãos reprodutivos, tanto pela análise de transcritos quanto pela análise do gene repórter *GUS*. No entanto, a expressão de *Tnt1* foi claramente detectada em órgãos florais de *Arabidopsis*. Em tomate, onde existem sequências relacionadas ao elemento *Tnt1*, um padrão similar ao de *Arabidopsis* também foi encontrado (Lucas *et al.*, 1995; Moureau-Mhiri *et al.*, 1996). Essa diferença entre a expressão de *Tnt1* em seu hospedeiro natural e em espécies heterólogas sugere que apenas o fumo expressa fatores regulatórios específicos. Esses fatores preveniriam a expressão de *Tnt1* em órgãos florais, ou seja, evitariam a transmissão de possíveis eventos de transposição deletérios para a progênie (Grandbastien *et al.*, 1997).

Foi observado que a expressão de *Tnt1* é fortemente induzida em protoplastos recém-isolados de mesófilo de fumo, mas esse efeito se deve principalmente à presença de extratos brutos do fungo *T. viridae*, utilizado como celulase, que à cultura de células propriamente dita (Pouteau *et al.*, 1991). A expressão do *Tnt1* também é induzida por outros fatores de origem microbiana, chamados de "eliciadores bióticos", como, por exemplo, elicitinas fúngicas e sobrenadantes de cultura da bactéria *Erwinia chrysanthemi* (Pouteau *et al.*, 1994). Esses eliciadores têm a capacidade de induzir respostas associadas à indução de genes envolvidos nas respostas das defesas vegetais. A expressão de *Tnt1* foi detectada em tomate após a sua infeccção por vírus (Moureau-Mhiri *et al.*, 1996), em *Arabidopsis*, após infecção bacteriana (Moureau-Mhiri *et al.*, 1996), e em fumo, também após infecção por vírus (Grandbastien *et al.*, 1997). As defesas vegetais também são ativadas por estresse e outros fatores não patogênicos. A expressão de *Tnt1* foi induzida por injúria mecânica dos tecidos foliares e

por tratamentos químicos com sais de metais pesados, tanto em fumo quanto em *Arabidopsis* (Vernhettes *et al.*, 1997). Esses resultados indicam que a indução da expressão de *Tnt1* ocorre por uma grande variedade de estímulos, muitos dos quais também induzem respostas de defesa vegetais. A indução da expressão de *Tnt1* correlaciona-se à resposta biológica da planta ao estímulo, tanto em termos de modificações fisiológicas (como os sintomas de necrose), como em termos da expressão de genes de defesa vegetais (Grandbastien, 1998).

Muitos fatores estão envolvidos no controle da transposição. A transcrição controla tanto a produção de uma população de RNA, que em parte será usada como molde para a transcriptase reversa, quanto na produção de mRNA envolvida na síntese das proteínas necessárias para a transposição. Sinais para início e término da transcrição presentes nas LTRs devem ser reconhecidos pela maquinaria celular. Os transcritos devem ser traduzidos em produtos funcionais. Os sítios iniciadores das fitas (+) e (-) devem ser funcionais, permitindo a transcrição reversa. A endonuclease deve ser capaz de processar e integrar as terminações da LTR no genoma do hospedeiro (Suoniemi *et al.*, 1997).

Foram descobertas várias sequências regulatórias que controlam a expressão gênica, tanto em *Drosophila*, como em levedura (Boeke e Corces, 1989) ou nos vegetais (Casacuberta e Gransbastien, 1993; Grandbastien *et al.*, 1997; McDonald *et al.*, 1997; Suoniemi *et al.*, 1997). Por exemplo, estudos com o elemento *Tnt1* demonstraram que essas sequências regulatórias se encontram nas LTRs, mais precisamente em sua região U3 (Casacuberta e Grandbastien, 1993; Grandbastien *et al.*, 1997).

Quando um elemento é ativo, a cada ciclo de transposição há o aumento do número de cópias desse elemento no genoma do hospedeiro. Isso porque, diferentemente dos elementos que se transpõem via DNA, não há a excisão do elemento e sua reinserção em outro ponto do genoma, mas sim a inserção de sua cópia (Kumar, 1996). Isso significa que a sequência inserida permanecerá no genoma do hospedeiro e o fenômeno da transposição do elemento ocorre pela inserção de sua cópia. Esse fato confere aos retrotransposons um papel importante no aumento da variabilidade genética e na plasticidade do genoma. Contudo, o aumento do número de cópias desses elementos pode ter-se tornado, em algum ponto do processo evolutivo, potencialmente deletério. Atribui-se a esse fato o desenvolvimento de mecanismos específicos de controle sobre a mobilidade desses elementos.

Finalmente, como já foi mencionado para o elemento *P*, o mecanismo de RNA de interferência apresenta-se como um dos principais tipos de defesa do hospedeiro contra os efeitos deletérios da presença de vírus e elementos de transposição, participando do silenciamento gênico pós-transcricional (degradação de mRNA), alterações epigenéticas e formação de heterocromatina. Em *Drosophila*, o mecanismo de RNAi parece controlar a transposição por duas vias diferenciadas, uma via para as células somáticas e outra para as células germinativas, ambas no citoplasma. As proteínas das famílias *Dicer* e *Argonauta* são fundamentais nesses processos. As proteínas *Dicer* têm como função a maturação dos RNAs pequenos de fita dupla que atuam no silenciamento e as proteínas *Argonauta* se ligam às fitas de RNA picotadas por *Dicer* desempenhando atividade endonucleolítica supressora degradando os RNAs alvo ou, ainda, funcionam como uma plataforma para a reunião de proteínas do complexo de silenciamento (Hammond, 2005).

Na linhagem germinativa, as proteínas *Aub*, *Piwi* e *Ago3* (família *Argonauta*, subfamília *PIWI*) são responsáveis pela biogênese dos piRNAs (do Inglês, *PIWI interacting RNAs*), derivados de mRNAs das fitas senso e antisenso de retrotransposons (Kim *et al.*, 2009). Os mRNAs usados para a produção de piRNAs primários são derivados de locos cromossômicos heterocromáticos específicos com alta densidade de TEs, que servem como uma "memória genética" dos elementos de transposição à qual o genoma têm sido exposto. Os piRNAs récem gerados, associados às proteínas *Aub*, *Piwi* e *Ago3*, irão mediar a localização e degradação de outros mRNAs derivados de TEs (eucromáticos e heterocromáticos), que por sua vez promoverão a retroalimentação do processo. Esse mecanismo tem sido denominado piRNA *ping-pong*, pois piRNAs antisenso associados à *Aub* servirão como molde para a degradação de mRNAs senso e geração de piRNAs senso e, estes, associados a *Ago3*, irão se ligar a fitas de mRNAs antisenso, levando à sua degradação e biogênese de novos piRNAs antisenso, reiniciando o processo. Adicionalmente, a proteína *Piwi* tem sido relacionada ao processo de heterocromatinização e mudanças epigenéticas no núcleo de células germinativas e somáticas presentes em ovários.

O silenciamento de elementos de transposição nas células somáticas é dependente de RNAs pequenos de fita simples denominados esiRNAs (do inglês*, endogenous small interfering RNAs*) e das proteínas *Ago2* e *Dicer2*. Os esiRNAs são derivados de RNAs dupla-fita longos provenientes de TEs. Nesse mecanismo, basicamente, ocorre a ligação e picotamento dos RNAs de dupla-fita por *Dicer2*, produzindo os esiRNAs. Os esiRNAs serão associados à *Ago2*, servindo como molde para a degradação de mRNAs de fita simples complementares (Kim *et al.*, 2009).

Mesmo de forma resumida, os dados apresentados acima demonstram a função do mecanismo de RNA de interferência para proteger o genoma de ácidos nucléicos "parasitas", entre eles, os elementos de transposição. Embora o mecanismo apresentado acima seja o proposto para *Drosophila*, o controle da transposição por RNAi está ativo, com variações, em ciliados, plantas, outros invertebrados e vertebrados.

9.5. Evolução dos Elementos de Transposição

A presença de domínios com funções semelhantes nos elementos de transposição levantam a questão sobre a origem desses elementos, sendo que três hipóteses têm sido propostas: (1) os elementos poderiam estar presentes no genoma do ancestral comum de todos os seres vivos atuais, (2) eles poderiam ter evoluído convergentemente; ou (3) eles seriam agentes capazes de dispersão por transferência horizontal. Uma vez que esses elementos sejam introduzidos num genoma hospedeiro, como se tratam de unidades replicantes, existe uma tendência de aumento do número de suas cópias. Assim, o número final de inserções depende de um equilíbrio entre vários fatores, como a frequência de transposição e fixação dos elementos, o genoma hospedeiro, fatores ambientais e interações entre esses fatores. A interação entre elemento de transposição e hospedeiro é muito importante e pode ser considerada como sendo um processo de coevolução. Nesse caso, deveria existir uma "queda-de-braço" entre a capacidade de replicação do TE, tendendo ao aumento do número de cópias e à ocorrência de mutações, e o controle da transposição pelo hospedeiro, levando à minimização dos efeitos deletérios de sua mobilização. Existe aqui uma analogia entre as interações ecológicas de parasitismo (onde haveria apenas prejuízo por parte do hospedeiro), comensalismo (situação onde o hospedeiro não é prejudicado) ou até mesmo o mutualismo (quando há vantagens também em relação ao hospedeiro).

Os retrotransposons estão presentes na grande maioria dos organismos. Esse fato é um indicador de que sua presença nos organismos vivos é antiga (Kumar, 1996). Formas ativas desses elementos devem ter persistido nos genomas de hospedeiros por

longos períodos, através de transmissão vertical, e por transmissão horizontal, como a maioria das sequências gênicas (Capy *et al.*, 1998).

Análises comparativas da estrutura e dos domínios presentes nos retrotransposons sugerem que os retrotransposons com LTR (especialmente o grupo *gypsy*) seriam os ancestrais dos retrovírus (Bucheton, 1995). Acredita-se também que os próprios retrotransposons com LTR teriam origem num ancestral comum, com uma estrutura bem mais simples, semelhante aos retrotransposons sem LTR (Xiong e Eickbush, 1990). O aumento da complexidade da estrutura desses elementos teria ocorrido pela aquisição de genes endógenos de seus hospedeiros. Desse modo, no decorrer de sua evolução, vários domínios, como a transcriptase reversa, endonucleases, LTRs e sequências *gag*, teriam sido adquiridos em diferentes momentos, produzindo as várias categorias de elementos conhecidos hoje (Capy *et al.*, 1998).

A grande heterogeneidade de sequência entre os retrotransposons pode ser explicada por seu mecanismo de propagação através da transcriptase reversa. Uma das características da transcriptase reversa é sua propensão em fazer substituições de bases através da cópia com erros. Foi estimada a ocorrência de substituições em uma taxa de no mínimo 1 em 7000 a 1 em 50000 resíduos por replicação em retrovírus (Smyth, 1993). Desse modo, poucos ciclos de integração-transcrição-transcrição reversa-reintegração poderiam levar a uma rápida divergência na sequência de um retrotransposon. Em uma taxa muito menor, os retrotransposons também estão sujeitos a mutações, como as demais sequências no genoma de um organismo.

Essa heterogeneidade reflete uma associação muito antiga entre as plantas e os retrotransposons. Muitos estudos realizados com elementos do tipo *copia* corroboram esse fato. Buscas em banco de dados por sequências similares a esses elementos resultaram na descoberta de inserções de retrotransposons tipo *copia* nas regiões 5' e 3' de mais de 30 genes vegetais (White *et al.*, 1994). A inserção dessas sequências próximo a genes, junto com a natureza degenerada dessas sequências, é interpretada como uma evidência adicional de uma associação antiga entre os retrotransposons tipo *copia* e os genomas vegetais. Outro dado que indica essa associação antiga é a presença dos retrotransposons em todos os grupos vegetais estudados. Em muitos casos, sugere-se que a inserção de um retrotransposon deva ter ocorrido anteriormente ao processo de especiação que deu origem a cada uma dessas espécies (Voytas *et al.*, 1992, Flavell *et al.*, 1992).

É característico dos retrotransposons que o mesmo elemento seja encontrado em espécies relacionadas, às vezes em gêneros, mas não em táxons muito distantes (Flavell *et al.*, 1994). O elemento *Tnt1* (Grandbastien *et al.*, 1989), descoberto em fumo, também foi encontrado em três outras espécies próximas (tomate, petúnia e batata), mas não em oito outras famílias testadas. O elemento *Del*1 (Smyth *et al.*, 1989) foi detectado em 14 espécies de *Lilium* examinadas, mas não em várias outras monocotiledôneas. O elemento *Bs1* do milho (Jin e Bennetzen, 1989) foi encontrado em quatro outras espécies do gênero *Zea*, bem como em outros gêneros relacionados, mas não em outras gramíneas mais distantes.

O padrão de distribuição entre as espécies não reflete necessariamente sua filogenia. No entanto, em muitos casos, o grau de divergência na sequência de elementos *Ty1/copia* entre quaisquer espécies geralmente é proporcional à distância evolutiva entre elas (Voytas *et al.*, 1992). Uma possível interpretação para esses padrões seria a presença do retrotransposons em um ancestral comum na linhagem, que foi sendo submetido a ciclos esporádicos de amplificação nos diferentes ramos. Outra hipótese, que poderia ter contribuído para a dispersão dos retrotransposons e que poderia explicar a similariadade de sequência entre espécies não relacionadas, seria a transferência horizontal. Embora poucos estudos tenham focado esse fenômeno em plantas, dois retrotransposons—*Rider* e *Route66*—foram relacionados a eventos de transferência em tomate e entre espécies de gramíneas, respectivamente (Roulin *et al.*, 2008; Cheng *et al.*, 2009).

O fenômeno de transferência horizontal corresponde a um processo de transferência de material genético, sem que ocorra uma transmissão sexual entre espécies não relacionadas. O exemplo clássico de transferência horizontal é o do elemento *P* em *D. melanogaster* (Kidwell *et al.*, 1977, Bingham *et al.*, 1982). Esse elemento foi isolado e caracterizado após sua transposição no gene *whit*e, que confere alteração no padrão de pigmentação dos olhos das moscas. A presença desse elemento de transposição é considerada recente em *D. melanogaster*, menos de 100 anos, já que linhagens de laboratório coletadas na década de 50 não apresentam esse TE. Atualmente todas as populações naturais de *D. melanogaster* apresentam o elemento *P*. É sugerido que essa transferência horizontal deva ter ocorrido quando *D. melanogaster* e *D. willistoni* se tornaram simpátricas e o elemento *P* tenha se espalhado nas populações de *D. melanogaster*. Esse elemento é altamente conservado nessas duas espécies, sendo que, dos seus cerca de 2.900 pb, eles diferem em apenas em 1 pb. Detalhes do mecanismo de transferência horizontal são desconhecidos até o momento, mas se especula sobre a possibilidade de ter ocorrido por meio de vetores, como vírus, parasitas endossimbiontes e/ou ácaros.

Nos últimos anos, vários exemplos de transferência horizontal têm sido propostos. Em Drosophilidae, análises evolutivas de 21 TEs indicam a ocorrência de pelo menos 101 casos de transferência horizontal (para uma revisão, ver Loreto *et al.*, 2008). Dessas proposições, 52,4% envolvem transposons de DNA, 42,6% retrotransposons com LTR e, 5% retrotransposons sem LTR. Esses eventos são propostos com base em três fatores: alta similaridade, incongruências filogenéticas e distribuição descontínua entre espécies. No entanto, questões sobre quando e como ocorreram esses eventos de transferência horizontal em Drosophilidade continuam em aberto e constituem interessantes questões para estudos futuros.

Em vertebrados, as proposições de transferência horizontal são mais escassas que em *Drosophila*. O caso mais bem caracterizado envolve o elemento *SPIN*, que parece ter sido envolvido em uma onda de transferência horizontal ocorrida entre 15 e 46 milhões de anos (Pace *et al.*, 2008). Esses eventos devem ter envolvido diversas espécies de vertebrados, entre elas, *Mus musculus* (camundongo), *Otolemur garnettii* (gálago), *Tenrec ecaudatus* (tanreque), *Anolis carolinensis* (lagarto anole verde), *Xenopus tropicalis* (anuro) e *Monodelphis domestica* (marsupial). No entanto, este elemento está ausente pelo menos em 19 outros genomas de mamíferos, como por exemplo, no homem e no morcego jamaicano *Artibeus jamaicensis*.

9.6. Impacto no Genoma

Elementos de transposição são fonte de mudanças genéticas, incluindo a alteração na expressão gênica durante o desenvolvimento e a origem de rearranjos genômicos. Alterações de sequência causadas pelos elementos de transposição podem variar de poucos pares de bases a grandes rearranjos cromossômicos, como aqueles gerados por deleções, inversões e duplicações. Essas mutações podem gerar efeitos deletérios, como já exemplificado com a síndrome de disgenesia híbrida em *D. melanogaster*. Cerca de 0,27% de todas as mutações que acarretam doença em humanos são causadas por retrotransposons (Callinan e Batzer,

2006), entre elas, a distrofia muscular Duchene, a profiria e as hemofilias A e B. Algumas mutações, com mudança de fenótipo em plantas, também estão associadas a elementos de transposição. Por exemplo, a inserção do *Helitron Hel-It1* no gene responsável pela pigmentação *DFR-B* em campainha (*Ipomoea tricolor*) leva à ausência de pigmentação nas flores, púrpuras nas linhagens com fenótipo selvagem (Choi *et al.*, 2007).

Eventualmente, essas inserções—ou fragmentos delas—podem ser fixadas pela seleção natural e passar a desempenhar uma função no genoma hospedeiro. Esse processo tem sido denominado domesticação ou exaptação de elementos de transposição (Gotea e Makalowski, 2006). Tal processo tem sido exemplificado em diversos estudos, indicando que os elementos de transposição tiveram e ainda devem ter um papel importante na evolução dos genomas. Mais especificamente, os elementos de transposição são capazes de produzir mutações em regiões gênicas que podem afetar a estrutura e as funções de proteínas, bem como gerar novas unidades regulatórias (Figura 9.5).

A inserção de um TE na região regulatória de um gene pode levar a sua sub ou superexpressão devido à presença de fatores regulatórios em suas repetições terminais (Figura 9.5a). Schlenke e Begun (2004) demonstraram que a inserção do retrotransposon sem LTR *Doc* na região regulatória do gene *Cyp6g1* de *D. simulans* está correlacionada ao aumento de sua expressão, fenótipo que tem sido relacionado à resistência a inseticidas em *Drosophila*. É interessante notar que essa característica está presente em populações da Califórnia (USA), ausente nas populações da África, isto é, está presente nas populações sob processo de adaptação a ambientes recém-colonizados. Em outro estudo, foi demonstrado que cerca de 25% dos promotores humanos caracterizados experimentalmente possuem sequências derivadas de TEs (Jordan *et al.*, 2003).

A domesticação de um TE inserido dentro de um gene pode levar a alterações nas sequências protéicas. Um TE inserido diretamente dentro de um éxon se tornaria parte do mRNA transcrito (Figura 9.5b). No entanto, devido a altas frequências de *stop codons* nos TEs, o mRNA resultante provavelmente seria defectivo, isto é, não teria a capacidade de produzir uma proteína funcional. Por outro lado, a inserção de um TE em um íntron poderia levar a domesticação de um trecho codificante do elemento, resultando na criação de um novo éxon (Figura 9.5c). Um exemplo desses processos foi observado em humanos, pois, enquanto 4% dos mRNAs apresentam fragmentos de TEs, apenas 0,1% das proteínas humanas têm fragmentos codificados por TEs (Gotea e Makalowski, 2006). Entre esses exemplos, o gene PTPN1, uma proteína que catalisa desfosforilação, apresenta 101 nucleotídeos derivados do LINE *L3*, incluindo um sítio doador de *splicing*. Essa região provavelmente domesticada, determinada experimentalmente, corresponde a um fragmento do domínio da transcriptase reversa.

Em relação ao genoma de plantas, devido à presença de retrotransposons potencialmente ativos, as espécies vegetais poderiam ser capazes de responder mais rapidamente a futuras mudanças nas condições ambientais, devido a uma rápida variação genética. As formas mais adaptadas seriam, então, selecionadas entre todos os variantes. Nos estudos dos genes vegetais e de sua atividade, a presença de retrotransposons dentro ou próximo aos genes deve ser considerada como uma fonte de variação genômica (Smyth, 1993).

A seleção natural tende a minimizar qualquer impacto negativo que os retrotransposons venham a ter, enquanto procura utilizar os elementos de uma maneira positiva. Observa-se que muitos elementos estão metilados ou em domínios de heterocromatina (Bennetzen *et al.*, 1994).

No genoma de *Arabidopsis*, é interessante notar que o número de retrotransposons diferentes presentes é alto, mas o número de cópias de cada um é relativamente baixo. Apesar de a planta ter sido exposta a um grande número de elementos, o número de cópias foi minimizado, possibilitando associar os benefícios potenciais de um genoma pequeno com múltiplos tipos de retrotransposons (Bennetzen, 1996).

Em milho, as regiões entre genes apresentam retrotransposons com inserções de outros retrotransposons, que, juntos, perfazem cerca de 50% de seu DNA nuclear (SanMiguel *et al.*, 1996). Grandes blocos de retrotransposons (com mais de 50kpb) foram encontrados entre cópias simples de sequências gênicas do gene *adhI*. Esses blocos são compostos de no mínimo 10 famílias de retrotransposons, com número variado de cópias. Esses estudos realizados em milho sugerem que esses elementos foram selecionados negativamente em regiões codificadoras ou próximas a genes. Regiões intergênicas são hipermetiladas em relação às sequências codificadoras e, com base nos estudos com leveduras, poder-se-ia dizer que essas regiões da cromatina representariam um sinal, direcionando a integração dos retrotransposons (Voytas,

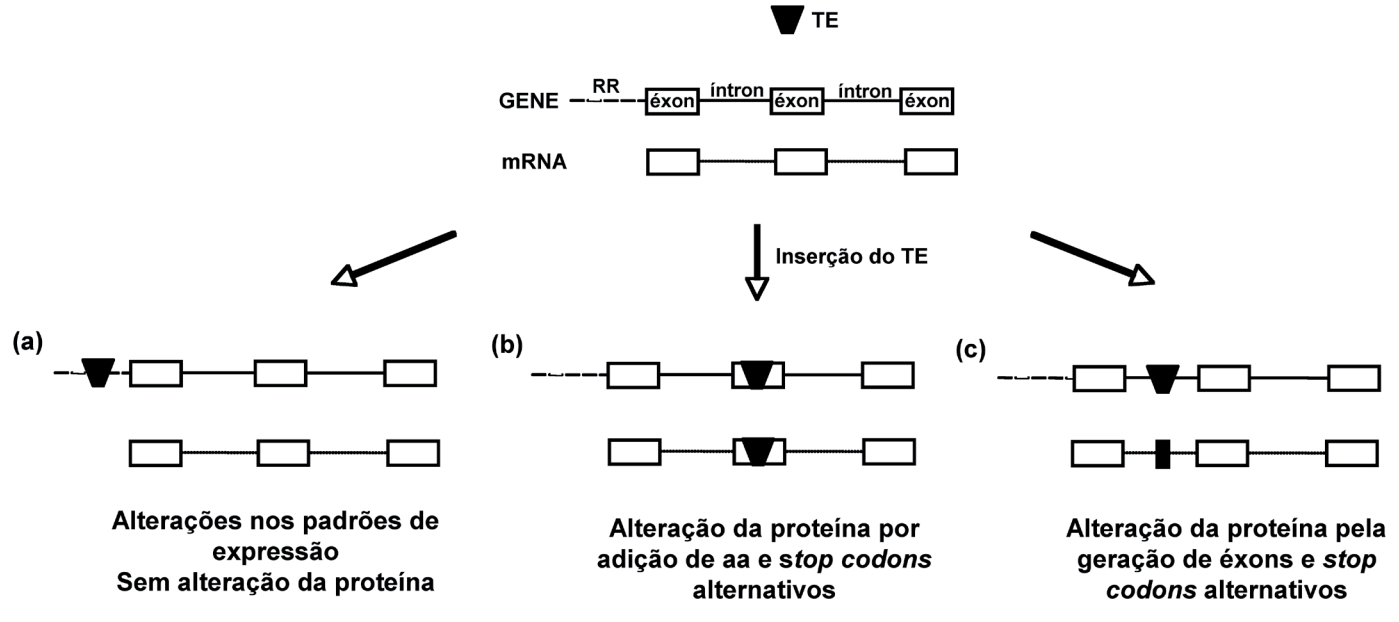


Figura 9.5. Possíveis impactos da inserção de elementos de transposição na produção de proteínas. (a) Inserção do TE na região regulatória (RR) de um gene, (b) dentro de um éxon, e (c) dentro de um íntron.

1996). Grupos hipermetilados de retrotransposons com inserções de outros retrotransposons já foram observados no fungo *Physarium polycephalum* (Bennetzen *et al.*, 1994), sugerindo que a integração dirigida pode ser uma estratégia largamente adotada pelos retrotransposons para a proliferação dentro do genoma do hospedeiro. Seria possível afirmar que o milho e seus retrotransposons coevoluíram em um mecanismo eficiente que permite a amplificação dos retrotransposons em níveis muito elevados dentro do genoma vegetal e, provavelmente, os retrotransposons contribuem para a variabilidade genética, podendo beneficiar seu hospedeiro durante o processo evolutivo (Voytas, 1996).

Todos esses exemplos demonstram o impacto da presença e mobilização dos elementos de transposição nos genomas hospedeiros e sua importância na geração de variabilidade genética. A disponibilidade da sequência genômica de diversas espécies tem proporcionado um aumento das possibilidades de estudos comparativos dessas sequências, mas diversas questões mais amplas sobre origem e possíveis funções ainda necessitam de respostas. As características que permitiram que os TEs fossem classificados com "DNA egoísta" também devem ter permitido a essas sequências participarem da geração de plasticidade genômica e o desenvolvimento de diversidade genética. Assim, podemos concluir que os elementos de transposição têm desempenhado um importante papel na evolução do genoma (Kidwell, 2005).

Referências Bibliográficas

Araújo P.G., J.M. Casacuberta, A.P.P. Costa, R.Y. Hashimoto, M.A. Grandbastien and M.A. Van Sluys. (2001) Retrolyc1 subfamilies defined by different U3 LTR regulatory regions in the *Lycopersicon* genus., **Mol. Genet. Genomics 266**:1, 35-41.

Bennetzen, J.L. (1996). The Mutator transposable element system in maize. **In** Saedler H. e Gierl A. (ed) **Transposable elements**, Spring Verlag, pp. 195-230.

Bennetzen, J.L. (2000). Transposable elements contributions to plant gene and genome evolution. **Plant Mol. Biol. 42**:251-269.

Bennetzen, J.L., Schrick, K., Springer, P.S., Brown, W.E. e Sanmiguel, P. (1994). Active maize genes are unmodified and flanked by diverse classes of modified highly repetitive DNA. **Genome 37**:565-576.

Bingham, J. e Zachar, M. (1989). Retrotransposon and the Fb transposon from Drosophila melanogaster. **In** Berg D.E. e Howe M.M. (ed.) **Mobile DNA**., American Society for Microbiology, Washington pp. 485-502.

Bingham, P.M., Kidwell, M.G. e Rubin, G.M. (1982). The molecular basis of P-M hybrid dysgenesis: the role of the *P* element, a *P*-strain-specific transposon family. **Cell 29**:995-1004.

Boeke, J.D. e Corces, V.D. (1989). Transcriptional and reverse transcription of retrotransposons. **Ann. Rev. Microbiol. 43**:403-34.

Brennecke, J., Malone, C.D., Aravin, A.A., Sachidanandam, R., Stark, A. e Hannon, G.J. (2008). An epigenetic role for maternally inherited piRNAs in transposon silencing. **Science 322**:1387-1392.

Bucheton, A. (1995). The relationship between the flamenco gene and gypsy in Drosophila; how to tame a retrovirus. **Trend. Genet. 11**:34-35.

Callinan, P.A. e Batzer, M.A. (2006). Retrotransposable elements and human disease. **Genome Dyn. 1**:104-115.

Camirand, A., St Pierre, B., Martineau, C. e Brisson, N. (1990). Ocurrence of copia-like transposable element in one of the introns of the potato starch phosphorylase gene. **Mol. Gen. Genet. 224**:33-39.

Capy, P., Bazin, C., Higuet, D. e Langin, T. (1998). Classification of transposable elements. **In** Capy, P., Bazin, C., Higuet, D. e Langin, T. (ed.) **Dynamics and evolution of transposable elements**. Spring Verlag,

Casacuberta, J.M. e Grandbastien, M.A. (1993). Characterisation of LTR sequences involved in the protoplast specific expression of the tobacco Tnt1 retrotransposon. **Nucleic Acids Res. 21**:2087-2093.

Chandler V.L. e Walbot V. (1986). DNA modification of a maize transposable element correlates with loss of activity. **Proc. Natl. Acad. Sci. USA. 83**:1767-1771.

Cheng, X., Zhang, D., Cheng, Z., Keller, B. e Ling, H.Q. (2009). A new family of *Ty1-copia*-like retrotransposons originated in the tomato genome by a recent horizontal transfer event. **Genetics 181**:1183-1193.

Choi, J.D., Hoshino, A., Park, K.I., Park, I.S. e Iida, S. (2007). Spontaneous mutations caused by a *Helitron* transposon, *Hel-It1*, in morning glory, *Ipomoea tricolor*. **Plant J. 49**:924-934.

Chomet, P.S., Wessler, S. e Dellaporta, S.L. (1987). Inactivation of the maize transposable element *Activator* (*Ac*) is associated with its DNA modification. **EMBO J**. 6:295-3O2.

Coen, E.E. (1989).Consequences and mechanisms of transposition in *Antirrinum majus*. **In** Berg D.E. e Howe M.M. (ed.) **Mobile DNA**. Whashington, American Society for Microbiology, pp. 413-436.

Costa A.P.P., Scortecci, K.C., Hashimoto, R.Y., Araujo, P.G., Granbastien, M.A., Van Sluys, M.A. (1999). Retrolyc1-1, a member of the Tnt1 retrotransposon super-family in the *Lycopersicon peruvianum* genome. **Genetica 107**:65-72.

Deininger, P.L. e Batzer, M.A. (2002). Mammalian retroelements. **Genome Res. 12**:1455-1465.

Emerson, R. A. (1914). The inheritance of a recurring somatic variation in variegated ears of maize. **Am. Nat. 48**:87-115.

Engels, W.R. (1996). P elements in *Drosophila*. **In** Saedler H. e Gierl A. (ed.); **Transposable elements**, Spring-Verlag, pp. 103-123.

Fedoroff, N.V. (1989). Maize transposable elements. **In** Berg, D.E e Howe, M.M. (ed.) **Mobile DNA**. Whashington, American Society for Microbiology, pp. 376-409.

Fedoroff, N.V. e Chandler, V. (1994). Inactivation of maize transposable elements. **In** Paszkowski J. (ed.), **Homologous recombination in plants**. Kluwer Academic, Dordrecht, pp. 335-348.

Fedoroff, N.V., Wessler, S. e Shure, M. (1983). Isolation of the transposable maize controlling elements *Ac* and *Ds*. **Cell 35**:235-242.

Feschotte, C., Jiang, N. e Wessler, S.R. (2002). Plant transposable elements: where genetics meets genomics. **Nat. Rev. Genet. 3**:329-341.

Flavell, A.J., Smith, D.B. e Kumar, A. (1992). Extreme heterogeneity of Ty1-copia group of retrotransposons in plants. **Mol. Gen. Genet. 231**:233-242.

Flavell, A.J., Pearce, S.R. e Kumar, A. (1994). Plant transposable elements and the genome. **Curr. Op. Gen. Dev. 4**:834-844.

Fosket, D.E. (1994). **The size and complexity of plant genomes In plant growth and Development – a molecular approach**. Cap 3. 1 ed. Academic Press. New York.

Fusswinkel, H., Schein, S., Courage, U., Starlinger, P. e Kunze, R. (1991). Detection and abundance of *mRNA* and protein encoded by transposable element *Activator* (*Ac*) in maize. **Mol**. **Gen**. **Genet**. **225**:186-192.

Gotea, V. e Makałowski, W. (2006). Do transposable elements really contribute to proteomes? **Trends Genet. 22**:260-267.

Grandbastien, M.A. (1987). Transposable Elements of plants. **In Nestle Research News 1986/1987**. Switzerland, Nestec Ltda, pp. 31-45.

Grandbastien, M.A. (1998). Activation of plant retrotransposons under stress conditions. **Trends Plant Sci. 3**:181-187.

Grandbastien, M.A., Audeon, C., Casacuberta, J., Grappin, P., Lucas, H., Moreau, C. e Pouteau, S. (1994). Functional analysis of the tobacco Tnt1 retrotransposon. **Genetica 93**:181-189.

Grandbastien, M.A., Lucas, H., Morel, J. B., Mhiri, C., Vernhettes, S. e Casacuberta, J. (1997). The expression of the tobacco Tnt1 retrotransposon is linked to plant defense response. **Genetica 100**:241-252.

Grandbastien, M.A., Spielmann, A. e Caboche, M. (1989). Tnt1, a mobile retroviral-like retrotransposable element of tobacco isolated by plant cell genetics. **Nature 337**:376-380.

Hammond, S.M. (2005). Dicing and slicing: the core machinery of the RNA interference pathway. **FEBS Lett. 579**:5822-5829.

Harberd, N.P., Flavell, R.B. e Thompson, R.D. (1987). Identification of a transposons-like insertion in a *Glu-1* allele of wheat. **Mol. Gen. Genet. 209**:326-332.

Haren, L., Ton-Hoang, B. e Chandler, M. (1999). Integrating DNA: transposases and retroviral integrases. **Annu. Rev. Microbiol. 53**:245-281.

Heinlein, M., Brattig, T. e Kunze, R. (1994). *In vivo* aggregation of maize *Activator* (*Ac*) transposase in nuclei of maize endosperm and *Petunia* protoplasts. **Plant J**. **5**:705-714.

Hirochika, H. (1993). Activation of tobacco retrotransposon during tissue culture. **EMBO J. 12**:25-31.

Hirochika, H., Fukuchi, A. e Kikuchi, F. (1992). Retrotransposons families in rice. **Mol. Gen. Genet. 233**:209-216.

Hu, W., Das, O e Messing, J. (1995). *Zeon-1*, a member of new maize retrotransposon family. **Mol. Gen. Genet. 248**:471-480.

Jin, Y.K. e Bennetzen, J.L. (1989). Structure and coding properties of Bs1, a maize retrovirus-like transposon. Proc. Natl. Acad. Sci. USA 86:6235-6239.

Jordan, I.K., Rogozin, I.B., Glazko, G.V. e Koonin, E.V. (2003). Origin of a substantial fraction of human regulatory sequences from transposable elements. **Trends Genet. 19**:68-72.

Kidwell, M.G. (2005). Transposable elements. **In** Gregory, T.R. (ed) **The evolution of the genome**, Elsevier, pp. 165-221.

Kidwell, M.G., Kidwell. J.F. e Sved, J.A. (1977). Hybrid dysgenesis in

Drosophila melanogaster: a syndrome of aberrant traits including mutation, sterility and male recombination. **Genetics 86**:813-833.

Kim, V.N., Han, J. e Siomi, M.C. (2009) Biogenesis of small RNAs in animals. **Nat. Rev. Mol. Cell. Biol. 10**:126-139.

Kossack, D.S. e Kinlaw, C.S. (1999). IFG, a gypsy-like retrotransposon in *Pinus* (Pinaceae), has an extensive history in pines. Plant Mol. Biol. 39:417-426.

Kumar, A. (1996). The adventures of the Ty1-copia group of retrotransposons in plants. **Trends Genet. 12**:41-43.

Kumar, A. e Bennetzen, J.L. (1999). Plant Retrotransposons. **Annu. Rev. Genet**. **33**:479-532.

Kunze, R. (1996). The maize transposable element *Activator* (*Ac*). **In** Saedler H. e Gierl A. (ed) **Transposable elements**, Spring-Verlag, pp. 161-194.

Kunze, R. e Schwartz, D. (1988). DNA methylation of the maize transposable element *Ac* interferes with its transcription. **Mol**. **Gen**. **Genet**. **214**:325-327.

Labrador, M. e Corces, V.G. (1997). Transposable element-host interactions: Regulation of insertion and excision. **Annu**. **Rev**. **Genet**. **31**:381-404 .

Lee, D., Ellis, T.H.N., Turner, L., Hellens, R.P. e Cleary, W.G. (1990). A copia-like element in *Pisuim* demonstrates the uses of dispersed repeated sequences in genetic analysis. **Plant Mol Biol 15**:707-722.

Leeton, P.J. e Smyth, D.R. (1993). An abundant LINE like element amplified in the genome of *Lillium speciosum*. **Mol. Gen. Genet. 237**:97-104.

Lewin, B. (1994). Retroviruses and retrotransposons **In** Lewin, B. **Genes V**. Cap 35. 2 ed. Oxford Press. London.

Lippman, Z. e Martienssen, R. (2004). The role of RNA interference in heterochromatic silencing. **Nature 431**(7006):364-370.

Loreto, E.L., Carareto, C.M. e Capy, P. (2008). Revisiting horizontal transfer of transposable elements in *Drosophila*. **Heredity 100**:545-554.

Lucas, H., Feuerbach, F., Kunert, K., Grandbastien, M.A. e Caboche, M. (1995). RNA-mediated transposition of the tobacco retrotransposon Tnt1 in *Arabidopsis thaliana*. **EMBO J. 14**:2364-2373.

McClintock, B. (1956). Controlling elements and the gene. **Cold Spring Harbor Symp**. **Quant**. **Biol**. **21**:197-216.

McClintock, B. (1984). The significance of responses of the genome to challenge. **Science 226**:792-801.

McDonald, J.F., Matyunina, L.V., Wilson, S., Jordan, I.K., Bownen, N.J. e Miller, W.J. (1997). LTR retrotransposons and the evolution of eukaryotic enhancers. **Genetica 100**:3-13.

Mannienen, I. e Schulman, A.H. (1993). Bare-1, a copia-like retroelement in barley (*Hordeum vulgare*). **Plant Mol. Biol. 22**:829-864.

Moore, G., Cheug, W., Scwarzacher, T. e Flavell, R.B. (1991). Bs-1, a major component of the cereal genome and a tool for studing genomic organization. **Genomics 10**:469-476.

Moureau-Mhiri, C., Morel, J.B., Audeon, C., Ferault, M., Grandbastien, M.A e Lucas, H. (1996). Regulation of expression of the tobacco Tnt1 retrotransposon in heterologous species following pathogen-related stresses. **Plant J. 9**:409-419.

Mugnier, N., Gueguen, L., Vieira, C. e Biémont, C. (2008). The heterochromatic copies of the LTR retrotransposons as a record of the genomic events that have shaped the *Drosophila melanogaster* genome. **Gene 411**:87-93.

Pace, J.K. II, Gilbert, C., Clark, M.S. e Feschotte, C. (2008). Repeated horizontal transfer of a DNA transposon in mammals and other tetrapods. **Proc. Natl. Acad. Sci. USA 105**:17023-17028.

Pardue, M.L., Rashkova, S., Casacuberta, E., DeBaryshe, P.G., George, J.A. e Traverse K.L. (2005). Two retrotransposons maintain telomeres in *Drosophila*. **Chromosome Res. 13**(5):443-53.

Pelissier, T., Tutois, S., Deragon, J.M., Tourmente S., Genestiers, S. e Picard, G. (1995). Athila, a new retroelement from *Arabidopsis thaliana*. **Genetica 97**:141-151.

Peterson, P.A. (1993). Transposable elements in maize: their role in creating plant genetic variability. **Adv. Agron. 51**:79-124.

Peterson, P.A. (1995). The development of transposon biology: from variegation to molecular confirmation. **Maydica 40**:117-124.

Plasterk, R.H.A. (1996). The *Tc1/mariner* Transposon Family. **In** Saedler, H. e Gierl A. (ed.). **Transposable elements**, Springer Verlag, pp. 125-143.

Pouteau, S., Huttener, E., Grandbastien, M.A. e Caboche, M. (1991). Specific expression of the tobacco Tnt1 retrotransposon in protoplasts. **EMBO J. 10**:1911-1918.

Pouteau, S., Grandbastien, M.A. e Boccara, M. (1994). Microbial elicitors of plant defense responses activate transcription of a retrotransposon. **Plant J. 5**:535-542.

Purugganan, M.D. e Wessler, S.R. (1994). Molecular evolution of magellan, a maize ty3/gypsy-like retrotransposon. **Proc. Natl. Acad. Sci. USA 91**:11674-11678.

Robertson, D.S. (1978). Characterization of a mutator system in maize. **Mut. Res. 51**:21-28.

Roulin, A., Piegu, B., Wing, R.A. e Panaud, O. (2008) Evidence of multiple horizontal transfers of the long terminal repeat retrotransposon *RIRE1* within the genus *Oryza*. **Plant J. 53**:950-959.

Saedler, H. e Nevers, P. (1985). Transposition in plants: a molecular model. **EMBO J. 4**:585-590.

SanMiguel, P., Tikhonov, A., Jin, Y.K., Motchoulskaia, N., Zakharov, D., Melakebehan, A., Springer, P.S., Edwards, K.J., Lee, M., Avramova, Z. e Bennetzen, J.L. (1996). Nested retrotransposons in intergenic regions of the maize genome. **Science 273**:765-769.

Schlenke, T.A. e Begun, D.J. (2004). Strong selective sweep associated with a transposon insertion in *Drosophila simulans*. **Proc. Natl. Acad. Sci. USA 101**:1626-1631.

Schwartz, D. (1984). Analysis of the *Ac* transposable element dosage effect in maize. **Mol**. **Gen. Genetics 196**:81-84.

Schwartz, D. (1989). Pattern of *Ac* transposition in maize. **Genetics 121**:125-128.

Schwarz-Sommer, Z., Gierl, A., Cuypers, H., Peterson, P.A. e Saedler, H. (1985). Plant transposable elements generate the DNA sequence diversity needed in evolution. **EMBO J**. **4**:591-597.

Schwarz-Sommer, Z. e Saedler, H. (1988). Transposons and retrotransposons in plants: analysis and biological relevance. **The Society for General Microbiology Symposium 43**:334-354.

Scortecci, K.C., Raina, R., Fedoroff, N.V. e Van Sluys, M.A. (1999). Negative effect of the 5'untranslated leader sequence on the Ac transpososn promoter expression. **Plant Mol. Biol. 40**:935-944.

Shepherd, N.S., Schwarz-Sommer, Z., Blumerg, V.E., Spalve, J., Gupla, M., Wienand, U. e Saedler, H. (1984). Similarity of the Cin-1 repetitive family of Zea mays to eukaryotic transposable elements. **Nature 397**:185-187.

Smyth, D.R. (1993). Plant Retrotransposons. **In** Verna, P. (ed). **Control of gene expression**, cap. 1. CRC Press, New York.

Smyth, D.R., Kallistis, R., Joseph, J.L. e Sentry, W. (1989). Plant retrotransposons from *Lilium henryi* is related to TY3 of yeast and gypsy group of *Drosophila*. **Proc. Natl. Acad. Sci. USA. 86**:5015-5019.

Starlinger, P. (1993). What do we need to know about transposable element *Ac*? **Gene 135**:251-255.

Suoniemi, A., Schimidt, D. e Schulmam, A. (1997). Bare-1 insertion site preferences of RNA and evolutionary conservation of RNA and cDNA processing sites. **Genetica 100**:219-230.

Turcich, M.T., Bokari-Riza, A., Hamilton, D.A., He, C., Messier, W., Stewart, C.B. e Mascarenhas, J.P. (1996). *Prem-2*, a copia-like type retroelement in maize is expressed preferentially in early microspores. **Sex. Plant Reprod. 9**:65-74.

Vernhettes, S., Grandbastien, M.A. e Casacuberta, J.M. (1997). *In vivo* characterization of transcriptional regulatory sequences involved in the defence-associated expression of the tobacco retrotransposon *Tnt1*. **Plant. Mol. Biol. 35**:673-679.

Voytas, D.F. (1996). Retroelements in genome organisation. **Science 274**:737-738.

Voytas, D.F. e Ausubel, F.M. (1988). A copia-like transposable element family in *Arabidopsis thaliana*. **Nature 336**:242-244.

Voytas, D.F., Cumming, M.P., Konieczny, A., Ausubel, F.M. e Rodermel, S.R.. (1992). Copia-like retrotransposons are ubiquitous among plants. **Proc. Natl. Acad. Sci. USA. 89**:7124-7128.

Voytas D.F., Konieczny, A., Cummings, M.P. e Ausubel, F.M. (1990). The structure, distribution and evolution of the Ta1 retrotransposable element family of *Arabidopsis thaliana*. **Genetics 126**:713-21.

Wicker, T., Sabot, F, Hua-Van, A., Bennetzen; J.L., Capy, P., Chalhoub, B., Flavell, A. Leroy, P., Morgante, M. Panaud, O. Paux, E. SanMiguel, P. e Schulman, A.H. (2007). A unified classification system for eukaryotic transposable elements. **Nature Rev. Genet. 8**:973-982.

White, S.E., Habera, L.F. e Wessler, S.R. (1994). Retrotransposons in the flanking regions of normal plant genes: A role for copia-like elements in the evolution of gene structure and expression. **Proc. Natl. Acad. Sci. USA. 91**:11792-11796.

Wong, L.H. e Choo, K.H. (2004). Evolutionary dynamics of transposable elements at the centromere. **Trends Genet. 20(**12):611-616.

Xiong, Y. e Eickbush, T. (1990). Origin and evolution of retroelements based upon their reverse transcriptase sequences. **Embo J. 9**:353-362.

Yoshioka, Y., Matsumoto, S., Kojima, S., Ohshima, K., Okada, N. e Machida, Y. (1993). Molecular characterization of a short interspersed repetitive element from tobacco that exhibits sequence homology to specific tRNAs. **Proc. Natl. Acad. Sci. USA**. **90**:6562-6566.

Capítulo 10

Evolução dos genomas

Dra. Laila Alves Nahum (laila@nahum.com.br)
Centro de Pesquisas René Rachou – CPqRR
Fundação Oswaldo Cruz – Fiocruz Minas

"*Evolução não é uma força, mas um processo: não uma causa, mas uma lei.*" (John Morley, *On Compromise*, 1874)

10.1. Introdução

A definição de genoma (do inglês, *genome*, "***gen***es + chromos***ome***s") foi proposta originalmente por Hans Winkler (1920) para designar o somatório dos genes de uma célula haplóide de um organismo. As sequências de DNA não codificantes foram identificadas posteriormente, sendo então incluídas nessa definição. Em 1986, Thomas Roderick propôs o termo genômica (*genomics*) para descrever os estudos de mapeamento, sequenciamento e análise dos genomas.

Os estudos moleculares desenvolveram-se de forma extraordinária nas últimas décadas como reflexo dos avanços nas áreas da biologia molecular, biologia computacional e bioinformática. Atualmente, o genoma é tratado como um sistema dinâmico, no qual se consideram seus aspectos estruturais, funcionais e evolutivos. A Figura 10.1 indica algumas descobertas e iniciativas que contribuíram para o estudo dos genomas desde a definição de gene, proposta por Wilhelm Johannsen em 1909.

Os avanços nessa área do conhecimento fazem-se notar, por exemplo, pela mudança de nossa perspectiva de classificação dos organismos, passando a distribuí-los em três reinos: Bacteria, Archaea e Eukarya. Outro reflexo desses avanços é o número crescente de publicações científicas, que constituem uma excelente fonte de consulta. Dentre os livros-texto que abordam os diferentes aspectos da evolução dos genomas, citam-se os de Page e Holmes (1998), Graur e Li (2000), Nei e Kumar (2000), Meyer e van de Peer (2003), Gregory (2005), Lynch (2007) e Barton *et al.* (2007).

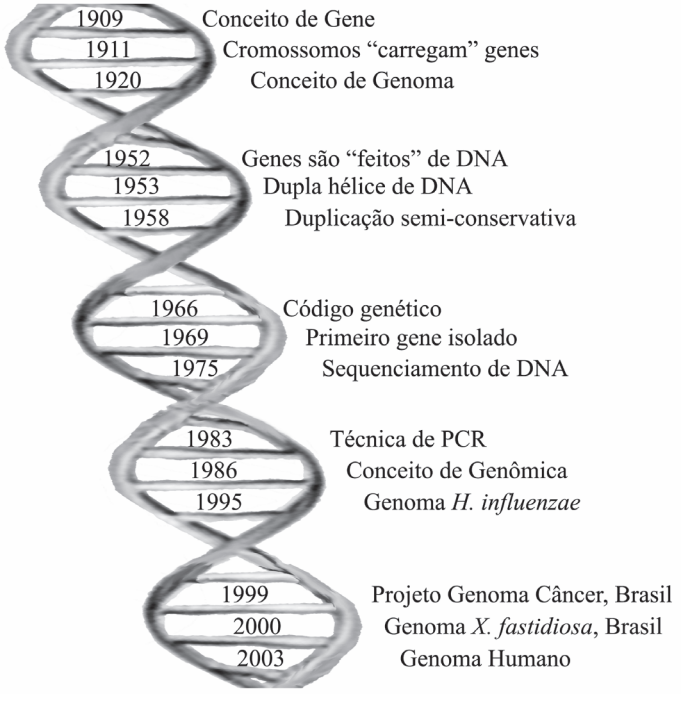


Figura 10.1. Perspectiva temporal de algumas das principais descobertas e iniciativas que contibuíram para os estudos de evolução dos genomas.

Neste capítulo, vamos considerar os dados gerados pelo sequenciamento de DNA, o tamanho, organização e composição dos genomas, os mecanismos de evolução dos genes e genomas, as perspectivas e desafios da genômica e algumas considerações finais.

10.2. Sequenciamento de DNA

10.2.1. A tecnologia de sequenciamento de DNA

Várias descobertas e avanços científicos impulsionaram a Genética e a Biologia Molecular a partir da segunda metade do século passado. Em 1952, Rosalind Franklin obteve dados de difração por raios-X, cruciais para a elucidação da estrutura da dupla hélice do DNA, descrita por Francis Crick e James Watson, em 1953. Em 1955, Arthur Kornberg e colaboradores isolaram a enzima polimerase do DNA, a qual seria usada para o sequenciamento de DNA duas décadas depois. Em 1977, Frederick Sanger e colaboradores, e também Maxan e Walter Gilbert, trabalhando independentemente, desenvolveram os métodos de sequenciamento de DNA, sendo o primeiro método mais popular (Sanger *et al*., 1977; Maxan e Gilbert, 1977). Mais recentemente, outras plataformas de sequenciamento foram desenvolvidas usando-se novas tecnologias, como, por exemplo, pirosequenciamento (Ronaghi *et al.*, 1998). O desenvolvimento da técnica de reação em cadeia da polimerase (do inglês, *polymerase chain reaction* – PCR) em 1983 revolucionou esse campo de forma extraordinária (veja também o Capítulo 19).

O impacto causado pela aplicação dessas metodologias faz-se notar na quantidade extraordinária de dados moleculares disponíveis em bancos de dados (Tabela 10.1). A integração entre biologia e informática surgiu da necessidade de criação de tais bancos de dados e das ferramentas computacionais para o armazenamento, análise e gerenciamento da quantidade maciça de dados gerada pelos projetos de sequenciamento de DNA, da obtenção da estrutura tridimensional de proteínas e outras metodologias analíticas empregadas em larga escala.

Na Tabela 10.1, estão indicados alguns bancos de dados e ferramentas computacionais e, na Tabela 10.2, algumas instituições de pesquisa relacionadas aos estudos dos genes e genomas de organismos diversos. Além disso, na tentativa de se formar um repositório de dados, a revista *Nucleic Acids Research* publica um grande número desses bancos de dados e ferramentas computacionais sempre no mês de janeiro (*Database Issue*), servindo como excelente fonte de consulta.

Tabela 10.1. Alguns bancos de dados e ferramentas de análise dos genomas e proteomas.

Acrônimo	Bancos e Ferramentas
BLAST	Basic Local Alignment Search Tool
EGenBio	Evolutionary Genomics and Biodiversity
ENCODE	ENCyclopedia Of DNA Elements Project
Ensembl	Ensembl Genome Browser
ExPASy	Expert Protein Analysis System Proteomics Server
GenomeAtlas	CBS Genome Atlas Database
Genomes	Genomic Biology at NCBI
GO	Gene Ontology Consortium
GOBASE	The Organelle Genome Database
GOLD	Genomes OnLine Database
MIPS	Munich Information Center for Protein Sequences
Pfam	Protein families database
PhyloFacts	Phylogenomic encyclopedias across the Tree of Life
UCSC Genome	University of California Santa Cruz Genome Browser
UniProt	Universal Protein Resource

Tabela 10.2. Algumas instituições envolvidas na análise dos genomas e proteomas.

Acrônimo	Instituições
Baylor HGSC	Baylor College of Medicine Human Genome Sequencing Center
BGI	Beijing Genomics Institute Shenzhen
BRGene	Projeto Genoma Brasileiro
Broad	Broad Institute of MIT and Harvard
Genomes	Genome Sequencing Consortiums & Centers
GENOSCOPE	Centre National de Séquençage
JCVI	J. Craig Venter Institute
JGI	DOE Joint Genome Institute
KDRI	Kazusa DNA Research Institute
LNCC	Laboratório Nacional de Computação Científica
NCBI	National Center for Biotechnology Information
NHGRI	National Human Genome Research Institute
ONSA	Organization for Nucleotide Sequencing and Analysis
RGMG	Rede Genoma de Minas Gerais
Sanger	Wellcome Trust Sanger Institute

10.2.2. Os genomas completamente sequenciados

O primeiro genoma a ser sequenciado e publicado foi o da bactéria de vida livre *Haemophilus influenzae* (Fleischmann *et al.*, 1995). A sequência completa do genoma da bactéria do trato intestinal, *Escherichia coli*, foi publicada dois anos depois (Blattner *et al.*, 1997). A obtenção da sequência completa desses genomas, juntamente com os de *Methanococcus jannaschii* (Bult *et al.*, 1996) e *Saccharomyces cerevisiae* (Goffeau *et al.*, 1997), representantes dos reinos Archaea e Eukarya, respectivamente, revolucionou o campo da evolução molecular e da genômica.

O Brasil inaugurou sua participação nesse cenário com a publicação do genoma da bactéria *Xylella fastidiosa*, o primeiro organismo fitopatógeno a ser sequenciado (Simpson *et al.*, 2000). A *Xylella fastidiosa*, frequentemente conhecida por "praga do amarelinho", causa a clorose variegada dos citros (CVC), atingindo um importante setor da produção agrícola e na economia nacional. O projeto de sequenciamento do genoma dessa bactéria foi lançado em outubro de 1997 pela FAPESP (Fundação de Amparo à Pesquisa no estado de São Paulo) com apoio da Fundecitrus (Fundo Paulista de Defesa da Citricultura). A realização desse projeto significou um novo modelo de trabalho cooperativo na área de pesquisa no Brasil, envolvendo um grande número de laboratórios e cientistas, tendo resultado na publicação da primeira sequência completa de um genoma fora do âmbito dos EUA, Europa e Japão. Desde então, as sequências genômicas de vários outros organismos foram analisadas pela comunidade científica brasileira, incluindo o das bactérias *Corynebacterium pseudotuberculosis*, *Chromobacterium violaceum*, *Leifsonia xyli*, *Mycoplasma synoviae* e *Xanthomonas axonopodis*, além de outros organismos como *Anopheles darlingi* e da cana-de-açúcar, dentre outros programas gerenciados por distintas instituições no Brasil.

O primeiro projeto de sequenciamento de um genoma de planta foi o de *Arabidopsis thaliana*. Contribuíram para a escolha dessa espécie, o tamanho reduzido do seu genoma (125 Mb) e o fato de que *Arabidopsis thaliana* se tornou um modelo de estudo em uma série de abordagens diferentes. Os estudos mostraram que a evolução desse organismo envolveu a duplicação de todo o genoma, seguida por perda de genes e duplicações gênicas, além de eventos de transferência lateral de um ancestral semelhante a uma cianobactéria. Outros genomas de eucariotos multicelulares completamente sequenciados incluem o de *Drosophila melanogaster* (Adams *et al.*, 2000) e *Caenorhabditis elegans* (The *C. elegans* Sequencing Consortium, 1998).

O projeto genoma humano permitiu produzir o mapa genético, o mapa físico e a sequência de DNA do conjunto completo dos cromossomos humanos. As duas mais importantes revistas científicas, *Nature* e *Science*, publicaram em fevereiro de 2001 a primeira versão do genoma humano (Lander *et al.*, 2001; Venter *et al.*, 2001), concluído em 2003. O projeto "1.000 Genomes", do Instituto Wellcome Trust Sanger, no Reino Unido (http://www.1000genomes.org/), pretende produzir um mapa das variações nas sequências de DNA de 1.000 genomas humanos. Essa iniciativa pretende contribuir para uma melhor compreensão da função dos genes em doenças humanas, permitindo um rápido avanço na pesquisa biomédica. Nesse contexto, destaca-se o Projeto Genoma do Câncer, criado no Brasil em março de 1999.

Em abril de 2011, o número de genomas completamente sequenciados e publicados e o número de projetos de metagenômica (do inglês, *metagenomics* ou *environmental genomics*) correspondiam a 1679 e 307, respectivamente (Genomes OnLine Database – GOLD). Note que os projetos de metagenômica surgiram posteriormente e se referem ao sequenciamento de amostras do ambiente (como o tubo digestivo de animais, ambientes aquáticos etc.). A velocidade com que esses números se alteram confirma o crescimento extraordinário dessa área. Para acompanhamento desses projetos, consulte os dados disponíveis no GOLD, National Center for Biotechnology Information (NCBI) ou em outra referência listada nas Tabelas 10.1 e 10.2.

10.3. Tamanho, Organização e Composição do Genoma

10.3.1. O tamanho e organização dos genomas

A notável flexibilidade do genoma pode ser facilmente evidenciada pela grande variação quanto a seu tamanho, organização

e composição dentre os diferentes organismos (Cavalier-Smith, 2005; Gregory *et al.*, 2007). Essa variação reflete a dinâmica da evolução dos genomas, sendo relevante em estudos da biodiversidade molecular.

O tamanho do genoma é definido pelo conteúdo haplóide (ou gamético) de DNA presente em um organismo, também conhecido por valor C. Frequentemente, o valor C é expresso em unidades de massa [1] ou em número de bases do DNA [2].

[1] 1 picograma (pg) = 10^{-12} grama (g)
[2] 1 megabase (Mb) = 10^{6} pares de bases (pb)

Para converter o número de pares de bases ou megabases de uma sequência de DNA em picogramas [3] e vice-versa [4], pode-se usar as seguintes fórmulas, que não são exatas, pois utilizam médias de massas moleculares de nucleotídeos:

[3] 1 Mb = 1,022 x 10^{-3} pg
[4] 1 pg de DNA = 0,978 × 10^{9} pb (ou 978 Mb)

Muito antes de os métodos de sequenciamento de DNA (seção 10.2.1) e de amplificação por PCR (Capítulo 19) estarem disponíveis, o tamanho dos genomas já era estimado utilizando-se uma série de técnicas, como fluometria, densitometria de Feulgen, citometria de fluxo e cinética de reassociação de DNA, dentre outras.

A Tabela 10.3 relaciona o tamanho do genoma de vários organismos pertencentes a unidades taxonômicas diversas. O genoma da *Amoeba dubia* (670.000.000 kb) é extraordinariamente grande, sendo o maior dentre os descritos até o momento. Em contraste, tem-se o genoma de *Mycoplasma genitalium* (580.000 kb), considerado o organismo com o menor genoma capaz de duplicar-se independentemente.

Tabela 10.3. Tamanho do genoma nuclear de alguns organismos.

Espécie	Grupo	Genoma (pb)
Amoeba dubia	Lobosea	670.000.000.000
Pinus resinosa	Embryophyta	68.000.000.000
Allium cepa	Embryophyta	18.000.000.000
Bufo bufo	Chordata	6.900.000.000
Mus musculus	Chordata	3.454.200.000
Homo sapiens	Chordata	3.400.000.000
Xenopus laevis	Chordata	3.100.000.000
Camelus dromedarius	Chordata	2.926.200.000
Limulus polyphemus	Arthropoda	2.700.000.000
Danio rerio	Chordata	1.900.000.000
Cyprinus carpio	Chordata	1.700.000.000
Gallus gallus	Chordata	1.200.000.000
Musca domestica	Arthropoda	900.000.000
Schistosoma mansoni	Platyhelminthes	270.000.000
Drosophila melanogaster	Arthropoda	180.000.000
Caenorhabditis elegans	Nematoda	100.000.000
Dictyostelium discoideum	Dictyosteliida	34.000.000
Saccharomyces cerevisiae	Ascomycota	12.067.280
Escherichia coli	Proteobacteria	4.639.221
Mycoplasma genitalium	Firmicutes	580

Fonte: "The NCBI Taxonomy" (http://www.ncbi.nlm.nih.gov/taxonomy); DOGS - "Database Of Genome Sizes" (http://www.cbs.dtu.dk/databases/DOGS/).

Quanto a seu tamanho, o genoma das plantas superiores varia de 125 Mb (*Arabidopsis thaliana*) a 50.000 Mb (algumas espécies de lírio). Em *Arabidopsis thaliana*, os genes estão distribuídos de forma homogênea, em contraste com a maioria dos genes do milho, arroz e cevada, que se agrupam em trechos longos de DNA (Barakat *et al.*, 1997). Mais de 50% do genoma do milho correspondem a sequências de DNA repetitivo dispersas no genoma (do inglês, *interspersed repeated DNA*), refletindo principalmente a presença de retrotransposons inseridos entre os genes (Bennetzen, 1998) (veja também o Capítulo 9).

A variação no tamanho dos genomas não pode ser explicada pelas diferenças de complexidade genotípica ou fenotípica dos organismos, conhecido como o paradoxo do valor C (veja Cavalier-Smith, 2005). Alguns bancos de dados relacionando o tamanho dos genomas de vários organismos estão disponíveis (veja Gregory *et al.*, 2007).

Além do tamanho, a organização dos genomas também é surpreendente. Tomemos, por exemplo, o genoma dos diversos tipos de vírus. O genoma pode ser constituído por uma molécula de DNA fita simples ou fita dupla, de RNA fita simples ou dupla, linear ou circular, e ainda pode ser fragmentado.

A variabilidade encontrada quanto à organização do genoma também é significativa em bactérias. Um exemplo extremo é o genoma da bactéria espiroqueta *Borrelia burgdorferi*, que consiste de um cromossomo linear de 910 kb e 21 elementos extracromossômicos lineares e circulares (Fraser *et al.*, 1997).

O genoma de organismos dos domínios Bacteria e Archaea é composto majoritariamente por sequências de DNA codificante, enquanto que o genoma de representantes de Eukarya está organizado em sequências de DNA codificante e não codificante. As sequências de DNA codificante incluem os genes de cópia única e as famílias (multi)gênicas (seção 10.4). As sequências de DNA não codificante, por sua vez, correspondem aos introns, às sequências intergênicas, ao DNA repetitivo em série etc. (seção 10.7). Nos maiores genomas, a maior parte do DNA corresponde a sequências não codificantes.

Como mencionado anteriormente, o tamanho e organização dos genomas variam enormemente entre os organismos (seção 10.3.1 e Tabela 10.3). Dentre os mecanismos que levam à expansão do tamanho do genoma, citam-se a duplicação gênica, a duplicação cromossômica parcial ou completa, a transposição, o *crossing over* desigual e a amplificação de DNA, dentre outros.

10.3.2. O conteúdo G+C e a origem dos isócoros

Um aspecto importante a ser considerado quando se trata do estudo dos genomas é a variação das sequências de DNA quanto à composição dos nucleotídeos entre regiões distintas. Essa heterogeneidade pode refletir, por exemplo, a presença de isócoros, sequências centroméricas e teloméricas (seção 10.7.2) e a presença de elementos de transposição integrados ao genoma (Capítulo 9). Além da frequência e da distribuição dos nucleotídeos nas sequências de DNA, a utilização dos códons é outro aspecto importante na investigação dos mecanismos de evolução dos genomas (seção 10.3.3).

As diferenças quanto à composição dos nucleotídeos é frequentemente expressa pelo conteúdo G+C, definido como a porcentagem média de guaninas e citosinas, conjuntamente, presentes nas sequências de DNA. O conteúdo G+C pode refletir a composição do genoma como um todo, de genes específicos ou mesmo de cada posição do códon. Tal variação pode ser observada ainda entre as duas fitas do DNA

O conteúdo G+C varia significativamente entre as diferentes regiões do genoma de um organismo e também entre genomas de distintos organismos. Os exemplos indicados na Tabela 10.4

Tabela 10.4. O conteúdo G+C no genoma de alguns organismos.

Organismo	Reino	G+C total	G+C3	Códons
Methanococcus jannaschii	Archaea	31,84	24,74	504035
Pyrococcus horikoshii	Archaea	42,32	42,97	570239
Archaeoglobus fulgidus	Archaea	49,37	58,42	669080
Aeropyrum pernix	Archaea	57,49	66,39	645004
Mycoplasma genitalium	Bacteria	31,64	23,01	191684
Haemophilus influenzae	Bacteria	37,50	27,00	142945
Thermogota maritima	Bacteria	46,45	52,64	627709
Escherichia coli	Bacteria	51,27	54,70	4641517
Treponema pallidum	Bacteria	52,54	54,20	392385
Deinococcus radiodurans	Bacteria	67,24	84,01	972472
Saccharomyces cerevisiae	Eukarya	39,72	37,98	5623302
Caenorhabditis elegans	Eukarya	42,58	39,67	8739594
Homo sapiens	Eukarya	52,53	59,26	11310862
Mus musculus	Eukarya	52,79	59,79	5184349

Conteúdo G+C do genoma (G+C total) e da terceira posição dos códons (G+C3) expressos em porcentagem e número total de códons no genoma de alguns organismos. Fonte: "Codon Usage Database" (http://www.kazusa.or.jp/codon/); "Codon Frequency Tables for Database Organisms & Others" (http://www.msu.edu/~jhjacksn/Tools/codefreq.htm).

não pretendem ser representativos de cada reino (Bacteria, Archaea e Eukarya), considerando-se a grande diversidade inter e intraclasse, mas são ilustrativos quanto à variação do conteúdo G+C entre diferentes organismos.

Entre os genomas das bactérias, o conteúdo G+C varia enormemente. O conteúdo G+C em *Mycoplasma sualvi*, por exemplo, equivale a 23,7%; em contraste, no genoma de *Corynebacterium insidiosum* esse valor alcança 77,1% (Galtier e Lobry, 1997). Em contrapartida, a variação intragenômica do conteúdo G+C é pequena nessas espécies. A variação do conteúdo G+C entre os genomas dos vertebrados é muito menor que aquela observada em bactérias e os valores encontrados estão na faixa entre 35% e 45% (Sueoka, 1964). No caso dos vertebrados, observa-se que a variação no conteúdo G+C entre as regiões distintas do genoma é muito maior. Essa variação é causada pela presença de isócoros.

Os isócoros foram descobertos por Macaya *et al.* (1976), recebendo essa denominação posteriormente (Cuny *et al.*, 1981). Os isócoros são segmentos de DNA iguais ou superiores a 300 kb, que apresentam homogeneidade quanto à composição das sequências, refletida pelo seu conteúdo G+C elevado. Os fragmentos de DNA de alto peso molecular (50-100 kb), resultantes da quebra física ou enzimática dos isócoros durante os processos de extração de DNA, são observados nas preparações de rotina em laboratório. Esse é um exemplo prático de como se pode evidenciar experimentalmente aspectos relacionados à organização do genoma. Outro exemplo é o ensaio de digestão do DNA com a nuclease micrococica para verificação do empacotamento ou condensação da cromatina.

O genoma dos vertebrados é um mosaico de isócoros, agrupados em um número pequeno de famílias. No genoma humano, por exemplo, são reconhecidas cinco famílias de isócoros—a saber: L1, L2, H1, H2 e H3—, que parecem ser bastante conservadas entre os vertebrados (Costantini *et al.*, 2007).

Quais seriam os mecanismos capazes de gerar e manter as diferenças locais quanto à composição dos nucleotídeos no genoma dos diversos organismos? Duas hipóteses foram propostas para explicar a presença dos isócoros no genoma. A hipótese selecionista propõe que os isócoros (ricos em G+C) representariam uma forma de adaptação do DNA às altas temperaturas. De fato, o genoma das aves e dos mamíferos apresenta um conteúdo G+C elevado, o que poderia representar uma adaptação à temperatura corporal elevada desses homeotérmicos, se comparada à de organismos pecilotérmicos. Entretanto, a ocorrência de sequências ricas em A+T em bactérias termófilas faz com que tal hipótese seja questionável.

A hipótese mutacionista ou neutralista, por sua vez, sugere que a origem dos isócoros seja o resultado de um enviezamento nas taxas de substituições dos nucleotídeos ou dos padrões de mutação ao longo do genoma (Sueoka, 1988; Sharp *et al.*, 1995). Em outras palavras, as diferenças na composição de bases seriam causadas por variações regionais nos padrões de mutação. Sem dúvida, esse é um tema ainda bastante controvertido e os mecanismos de evolução dos isócoros permanecem em debate (Bernardi, 2007; Costantini e Bernardi, 2008).

Em contraste, alguns genomas são extremamente ricos em A+T. Exemplos incluem o genoma humano e do parasita causador da malária, *Plasmodium falciparum*, cujos genomas contém 60% e 80% de A+T, respectivamente. Em *P. falciparum*, DNAs minissatélites ricos em A+T têm sido usados como alvos terapêuticos, numa abordagem farmacogenômica, com o desenvolvimento de drogas contra a malária (Woynarowski *et al.*, 2007).

10.3.3. O enviezamento na utilização dos códons

A grande maioria dos aminoácidos é codificada por mais de um códon, caracterizando a chamada *degeneração* do código genético. Os códons sinônimos são aqueles que codificam um mesmo aminoácido. Por exemplo, os seis códons que codificam a leucina (TTA, TTG, CTT, CTC, CTA e CTG) são considerados sinônimos. Uma substituição entre códons não-sinônimos leva à substituição do resíduo de aminoácido na sequência da proteína (veja também o Capítulo 7).

Curiosamente, os códons sinônimos, denominados também códons alternativos ou "redundantes", não são utilizados na mesma frequência pela maquinaria de tradução de proteínas de um dado organismo. Este enviezamento na utilização dos códons é observado no genoma de um largo espectro de organismos, incluindo bactérias, fungos, plantas, aves e mamíferos, além de vírus. Sabe-se ainda que a utilização dos códons sinônimos é um fenômeno que não acontece ao acaso, conforme foi demonstrado por diferentes estudos (Santos *et al.*, 2004; Stoletzki e Eyre-Walker, 2007; Coleman *et al.*, 2008).

A pergunta principal nesses trabalhos é: por que os organismos usam preferencialmente um subconjunto dos códons sinônimos disponíveis? Acredita-se que o enviezamento na utilização dos códons possa resultar de um desvio nas taxas de substituição e/ou da ação da seleção atuando sobre as trocas "silenciosas" no DNA, ou seja, substituições de nucleotídeos que não acarretam a substituição de aminoácidos na sequência de proteínas. A utilização dos códons sinônimos reflete a variação na composição dos nucleotídeos, observada nos genomas distintos. Em organismos cujos genomas são ricos em G+C, os códons usados preferencialmente são aqueles terminados em G ou C, em contraposição ao uso dos códons terminados em A e T nos genomas onde o conteúdo G+C é pequeno (Muto e Osawa, 1987).

Provavelmente, as bactérias são os organismos nos quais o fenômeno de enviezamento na utilização dos códons tenha sido mais bem estudado (*e.g.*, Eyre-Walker, 1999). A utilização dos códons em *Bacillus subtilis*, por exemplo, reflete a composição dos nucleotídeos no genoma dessa bactéria, que é rico em A+T. Por outro lado, a utilização dos códons em mamíferos parece ser determinada principalmente pelo arranjo espacial do conteúdo G+C, ou seja, pela estrutura do isócoro (Galtier e Mouchiroud, 1998).

Em bactérias, além da composição dos nucleotídeos, pelo menos outro fator está atuando no uso dos códons sinônimos: a ligação de tRNAs. Em *Escherichia coli*, foi demonstrado que alguns códons são reconhecidos preferencialmente pelas espécies de tRNA mais abundantes nesse organismo. Tais códons são escolhidos preferencialmente conferindo uma vantagem sob o ponto de vista da tradução (Sharp e Matassi, 1994). Esses códons "ótimos" são escolhidos no caso dos genes altamente expressos, em contraposição aos genes com níveis baixos de expressão. No segundo caso, observa-se que a utilização dos códons é mais uniforme. A vantagem em se usarem códons "ótimos" seria tornar a tradução mais eficiente. Resultados similares têm sido descritos para o genoma dos eucariotos (*e.g.*, Chiapello *et al.*, 1998; Duret e Mouchiroud, 1999).

Atualmente, a seleção para a eficiência no processo de tradução das proteínas é a hipótese mais aceita para explicar o enviezamento na utilização dos códons. Postula-se que as substituições sinônimas em um gene altamente expresso, que resultam em códons raros no conjunto dos tRNAs, sejam eliminadas por seleção natural. Os agentes seletivos, nesse caso, seriam a velocidade e a eficiência da tradução, fundamentais em se tratando de proteínas com elevada taxa de reposição (do inglês, *turnover*).

Os estudos sobre a utilização dos códons ajudam a compreender os mecanismos que atuam na evolução dos genomas. Cabe salientar o avanço considerável no conhecimento acerca dos fatores que determinam a utilização dos códons, como resultado do crescimento significativo dos dados de sequências de DNA, gerados pelos projetos de sequenciamento.

10.4. Grupos de Genes e Famílias Multigênicas

10.4.1. O mapeamento dos genes

A análise comparativa dos genomas baseia-se em mapas físicos, mapas genéticos e sequências moleculares (*i.e.*, sequências de DNA, RNA e proteínas). Os mapas físicos podem ser obtidos por tecnologias empregando YACs (do inglês, *yeast artificial chromosomes*), BACs (*bacterial artificial chromosomes*) e PACs (*phage P1 artificial chromosomes*).

Os mapas genéticos, por sua vez, são obtidos a partir da informação gerada por marcadores genéticos diferentes, os quais têm sido gerados para um largo espectro de organismos. Juntos, os mapas físicos e genéticos definem as estratégias de análise dos genomas. Em conjunto, estudos de mapeamento constituem ferramentas importantes para a compreensão de como os genes e os genomas estão organizados e como evoluem entre as espécies empregando-se distintas abordagens, como a genômica funcional, estrutural e evolutiva. Além disso, tais mapas contribuem em estudos de cariotipagem e rearranjos cromossômicos.

Os dados de mapeamento e sequenciamento de DNA revelam uma extensa conservação da ordem gênica (sintenia) entre os genomas dos organismos pertencentes a uma mesma família ou a famílias relacionadas (Descorps-Declère *et al.*, 2008; Tang *et al.*, 2008).

10.4.2. Os grupos de genes

Quando se analisam a ocorrência e distribuição dos genes no genoma, observa-se que alguns deles estão dispostos em série (*tandem*), formando grupos. Esses grupos de genes apresentam sequências nucleotídicas semelhantes e seus produtos desempenham a mesma função biológica. Os grupos de genes compartilham características evolutivas peculiares entre si, incluindo os genes ribossômicos, genes que codificam as histonas e os genes homeóticos (veja Capítulo 11).

Os genes ribossômicos eucarióticos estão representados por várias repetições dispostas em série, cuja unidade básica contém, além dos genes 28S e 18S, cujas cadeias fazem parte do ribossomo, os espaçadores transcritos externo (ETS) e interno (ITS), e os espaçadores não transcritos (NTS), mostrados na Figura 10.2. O número elevado de cópias presentes nos grupos de genes pode refletir a necessidade do organismo de sintetizar certos produtos gênicos em grande quantidade. Para se ter uma idéia quanto a esse valor, basta se considerar que existem aproximadamente 400 cópias de genes ribossômicos em humanos e chimpanzés (Arnheim *et al.*, 1980). No Capítulo 8, foram abordados aspectos relacionados aos genes ribossômicos dos procariotos e eucariotos.

10.4.3. As famílias (multi)gênicas

Os membros das famílias gênicas e multigênicas apresentam sequências nucleotídicas semelhantes, mas que diferem em menor ou maior grau quanto à função de seus produtos. Por exemplo, os genes da família das globinas são distintos uns dos outros, embora compartilhem uma similaridade significativa no nível funcional e nas respectivas sequências de DNA.

Os membros de uma família multigênica podem estar localizados em uma região limitada de um único cromossomo ou dispersos ao longo do genoma. Sob o ponto de vista da evolução, o arranjo dos genes em famílias pode ser considerado uma aquisição importante, uma vez que permite uma regulação eficiente dos genes que codificam produtos com funções semelhantes. Por outro lado, um dos processos que causa a duplicação gênica, a recombinação desigual, produz copias em série e, portanto, esse arranjo pode ser um simples efeito colateral do fenômeno que originou as cópias.

Nem todos os membros das famílias multigênicas são funcionais. As cópias não funcionais dos genes que codificam proteínas, inativadas ao longo do processo evolutivo, correspondem aos pseudogenes, cuja sequência é semelhante a um ou

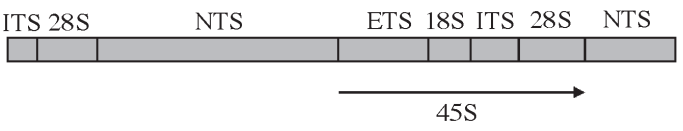


Figura 10.2. Organização dos genes ribossômicos em eucariotos. Além dos genes, estão indicados o espaçador transcrito interno (ITS), o espaçador transcrito externo (ETS) e o espaçador não transcrito (NTS).

mais genes parálogos da mesma família (seção 10.6.1). A perda funcional deve-se a falhas no processo de transcrição, tradução e/ou processamento de uma proteína. Os pseudogenes surgem, por exemplo, a partir da duplicação gênica via retrotransposição ou via duplicação do DNA genômico (Mighell *et al.*, 2000). Foram identificados 59 pseudogenes a partir da sequência completa do cromossomo humano 21 (Hattori *et al.*, 2000) e 134 no cromossomo 22 (Dunham *et al.*, 1999), correspondendo a 19,7% e 20,7% dos genes identificados nesses cromossomos, respectivamente.

A duplicação gênica (seção 10.5.1) parece ser o mecanismo mais plausível para se explicar a origem dos grupos de genes e famílias gênicas (Figura 10.3). Após o evento de duplicação gênica, as cópias novas podem ser distribuídas na população por deriva genética ou seleção. Ao longo do tempo, as cópias podem manter a mesma sequência de DNA, como no caso dos genes ribossômicos, ou divergirem em diferentes níveis, dando origem às famílias gênicas.

10.5. Mecanismos de Evolução dos Genes e Genomas

10.5.1. A duplicação gênica e divergência

A duplicação gênica (Figura 10.3) seguida pela divergência tem sido considerada o principal mecanismo da evolução molecular (Lewis, 1951; Ohno, 1970). O significado evolutivo da duplicação gênica foi reconhecido primeiramente por Haldane (1932) e Muller (1936), os quais sugeriram que a "duplicata redundante" de um gene poderia acumular substituições e eventualmente emergir como um gene novo. Entretanto, somente após o advento das técnicas de biologia molecular foi possível investigar mais profundamente esse aspecto.

Quando existe uma correspondência direta entre os éxons nos genes e os domínios estruturais ou funcionais na proteína, a duplicação de um ou mais éxons resultará na duplicação do domínio correspondente. Entretanto, os cenários evolutivos mais complexos são mais frequentemente encontrados quando a duplicação de um éxon leva à duplicação, por exemplo, de mais de um domínio ou mesmo de parte dele. A Figura 10.4 mostra as possíveis relações entre o arranjo dos éxons nos genes e os domínios estruturais nas proteínas.

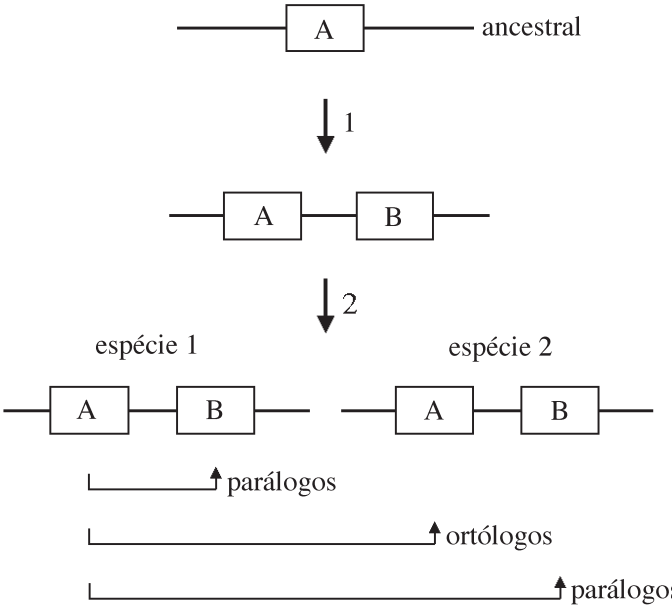


Figura 10.3. Modelo de duplicação gênica. Genes A e B, originados a partir de um gene ancestral. Duplicação gênica (1), especiação (2) e exemplos de genes parálogos e ortólogos estão indicados.

O domínio corresponde a uma região bem definida dentro de uma proteína, capaz de realizar uma função específica (domínio funcional), ou a uma unidade estrutural distinta das demais partes da proteína (domínio estrutural). Em vários casos, alguns aminoácidos, em diferentes posições da proteína, são responsáveis por sua função biológica, dificultando a identificação do domínio funcional. Em contrapartida, o domínio estrutural equivale a um segmento contínuo da sequência de aminoácidos. O conceito de domínio muitas vezes se confunde com o conceito de motivo (por exemplo, um sítio de ligação de ATP) ou de módulo, o qual corresponde a uma unidade evolutiva independente (Riley e Labedan, 1997).

O processo de duplicação envolve parte de um gene (duplicação gênica interna ou parcial), um gene (duplicação gênica completa), parte de um cromossomo (duplicação cromossômica parcial ou polissomia parcial), um cromossomo (duplicação cromossômica, aneuploidia ou polissomia) ou mesmo o genoma inteiro (duplicação genômica ou poliploidia). Conforme evidenciado por diversos estudos, a poliploidia é um mecanismo importante na evolução de anfíbios e plantas—como no caso já mencionado do genoma de *Arabidopsis thaliana* .

Cabe salientar que a similaridade no nível das sequências moleculares (DNA, RNA ou proteínas), no nível estrutural e/ou funcional pode ser um indicativo de ancestralidade compartilhada entre genes ou produtos gênicos estudados, porém outras trajetórias existem, como a evolução convergente (veja Nahum e Pereira, 2008).

10.5.2. O fenômeno da "evolução em concerto"

A partir de estudos das sequências repetitivas em eucariotos, observou-se que a similaridade entre as sequências de uma dada espécie é significativamente maior que a observada entre as sequências de espécies diferentes (Edelman e Gally, 1970). Essa observação estaria em desacordo com o esperado, caso a divergência entre as sequências de DNA

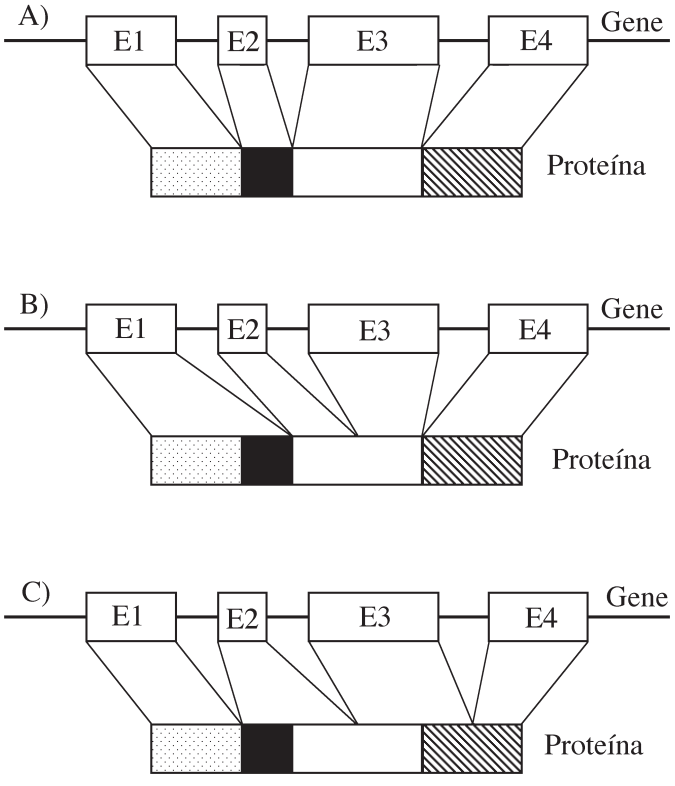


Figura 10.4. Diferentes relações entre os éxons de uma sequência codificante e os domínios de uma proteína hipotética. Modificado a partir de Li e Graur (1991).

fosse explicada somente por processos de mutação ao acaso e deriva genética.

Embora os estudos iniciais tenham se concentrado em eucariotos, outros, empregando bactérias, corroboram a hipótese de que as sequências repetitivas, bem como os membros das famílias multigênicas, não evoluem independentemente. Denominou-se "evolução em concerto" o processo molecular que leva à homogeneidade intraespecífica das sequências de DNA pertencentes a uma dada família (Zimmer *et al.*, 1980; Elder e Turner, 1995; Nei e Rooney, 2005; Eickbush e Eickbush, 2007). O termo homogeneidade, nesse caso, refere-se ao alto grau de similaridade observado entre as sequências de DNA. Outro termo, "evolução coincidente" (Hood *et al.*, 1975), tem sido empregado com menor frequência para designar o mesmo processo.

A primeira hipótese formulada para explicar os mecanismos associados à "evolução em concerto" foi a de replicação saltatória (Britten e Kohne, 1968; Buongiorno-Nardelli *et al.*, 1972; Amaldi *et al.*, 1973). Segundo essa hipótese, as sequências repetitivas seriam resultantes de duplicações recentes, não havendo, portanto, tempo suficiente para que nelas ocorresse acúmulo de substituições. Tal hipótese não se aplica à maioria das sequências repetitivas, nas quais os padrões de substituição distintos parecem ter sido fixados entre cada unidade de repetição (Wellauer *et al.*, 1976).

A hipótese de "mestre-escravo" (Callan, 1967; Thomas, 1970) considera essa observação, sugerindo que novas cópias "escravas" de uma sequência são replicadas a partir de uma sequência "mestre" em cada geração. Entretanto, a variação no tamanho e nos padrões de substituição em várias sequências repetitivas torna essa hipótese incompatível para vários locos.

Atualmente, acredita-se que esta "homogeneização" dos genes e famílias gênicas seja o resultado dos processos de *crossing over* desigual, conversão gênica, deslizamento das fitas de DNA durante a replicação e amplificação (Liao, 1999). Algumas considerações acerca desses mecanismos também podem ser vistas no Capítulo 7.

A Figura 10.5 mostra o modelo da "evolução em concerto", proposto para as famílias multigênicas cujos membros estão dispostos em série. Segundo esse modelo, uma substituição ocorrida em uma das unidades de repetição se propagaria rapidamente entre as demais pelo processo de homogeneização intracromossômica. Numa etapa subsequente, essa substituição "alcançaria" outro segmento de DNA através da conversão gênica intercromossômica. Um segundo evento de homogeneização intracromossômica fixaria a substituição no segundo segmento de DNA.

10.5.3. Outros mecanismos de evolução dos genomas

O *crossing over* desigual não explica a existência de DNA repetitivo localizado, uma vez que esse e outros mecanismos tendem a remover os arranjos em série, ao invés de aumentar seu tamanho e o número de cópias (Walsh, 1987). O mecanismo de amplificação gênica, que leva ao aumento considerável do número de cópias de uma sequência de DNA acima do nível normal em um organismo, foi proposto para explicar a origem de sequências repetitivas, tais como o DNA satélite.

O processo de amplificação envolve a formação de uma cópia circular e extracromossômica de uma sequência de DNA contendo um número variável de repetições, a qual pode ser replicada um grande número de vezes, segundo o modelo do círculo rolante (Bostok, 1986). As várias unidades extracromossômicas, contendo repetições em série da sequência original, são, então, reintegradas ao DNA, caracterizando a amplificação de um dado segmento.

Outros mecanismos de evolução dos genes e dos genomas incluem a fusão de genes (proteínas multimodulares), mutação, recombinação, inativação, perda de genes, transferência lateral, *lineage sorting* e poliploidia (Meyer e van de Peer, 2003; Barton *et al.*, 2007; Lynch 2007)

10.6. Evolução das Famílias de Genes e Proteínas

10.6.1. As famílias de genes parálogos

Segundo Fitch (1970), os produtos de um evento de duplicação gênica são considerados genes homólogos, podendo ser classificados como genes parálogos e genes ortólogos. Membros de uma família gênica que se originaram por duplicação e divergência são denominados genes parálogos. Em contrapartida, os genes homólogos que se originaram pelo processo de especiação são chamados ortólogos.

A Figura 10.3 ilustra a origem dos genes parálogos A e B a partir da duplicação do gene ancestral A. O gene A da espécie 1 e o mesmo gene da espécie 2 são considerados *ortólogos* (da mesma forma que o gene B das espécies 1 e 2). O gene A da espécie 1 e o gene B da espécie 2 são também chamados de *parálogos*.

Após a duplicação completa de um gene, ambas as cópias acumulam substituições, tornando-se progressivamente distintas do gene ancestral. As duplicatas podem manter a função original por certo período, produzindo um aumento no número de cópias do RNA e das proteínas. Ao longo do tempo evolutivo, porém, as cópias "reduntantes" de um gene podem seguir trajetórias distintas, resultando em neofuncionalização, subfuncionalização ou mesmo inativação do gene pelo acúmulo de mutações deletérias, originando um pseudogene (Descorps-Declère *et al.*, 2008; MacCarthy e Bergman, 2007; Nahum e Riley, 2001; Nahum *et al*., 2009; Roth *et al.*, 2007). Em conjunto, esses processos contribuem para a origem e evolução da biodiversidade molecular e de organismos.

Os genes parálogos podem ser agrupados em famílias a partir da identificação de sequências que apresentem um grau de

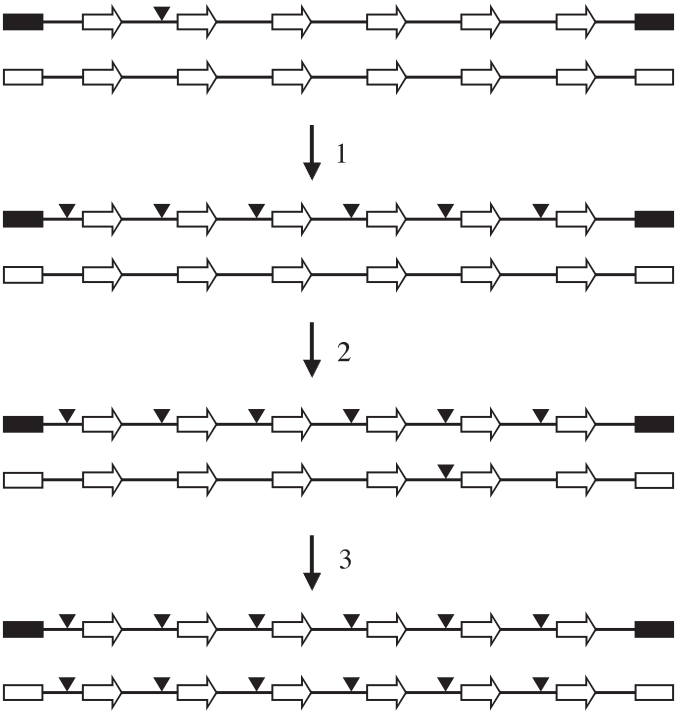


Figura 10.5. Modelo de "evolução em concerto" para famílias multigênicas. Regiões flanqueadoras (barras cheias, ■), cópias repetidas em série (setas), espaçadores intergênicos (linhas) e a ocorrência de mutações (barras vazias, □). Homogeneização intra-cromossômica (1), conversão gênica intercromossômica (2) e homogeneização intracromossômica (3). Modificado a partir de Liao *et al.* (1997).

similaridade significativo (Koonin *et al.*, 1995; Riley e Labedan, 1997; de Rosa e Labedan, 1998). A partir dessa abordagem, pode-se reconstruir a evolução das moléculas contemporâneas, desde seus ancestrais, assumindo que os descendentes de uma sequência ainda retêm vestígios de similaridade detectáveis, ou seja, que sua homologia possa ser corretamente estabelecida. Na abordagem evolutiva baseada na análise de proteínas, procede-se à identificação de grupos de proteínas cuja similaridade de sequência possa indicar ancestralidade comum.

Em função das suas características, os genes parálogos constituem candidatos excelentes para o estudo da evolução das proteínas, enquanto que os genes ortólogos têm sido empregados extensivamente em análises das relações filogenéticas entre organismos (Page e Holmes, 1998; Nahum *et al.*, 2003; Pereira e Baker , 2006; Nahum e Pereira, 2008) (veja também o Capítulo 16). Alternativamente, uma das cópias pode acumular mutações e ser inativada ao longo de gerações sucessivas, originando um pseudogene, como mencionado anteriormente.

A partir de dados moleculares, é possível estimar o tempo em que ocorreram os eventos de duplicação gênica. Primeiramente, é necessário estimar a taxa de substituição a partir do número de substituições entre os genes ortólogos em adição ao tempo de divergência entre duas espécies. A premissa básica para essa estimativa é constância de taxa (veja Capítulo 7). Outra possibilidade para a datação dos eventos de duplicação considera a distribuição filogenética dos genes e dados paleontológicos sobre o tempo de divergência das espécies em questão. Em ambos os casos, a estimativa é somente aproximada.

10.6.2. A origem das funções novas

Dentre os mecanismos capazes de gerar novas funções biológicas, citam-se o *splicing* alternativo, o *trans-splicing* e a edição de RNA (Capítulo 3), a ocorrência de genes sobrepostos (do inglês, *overlapping genes*) e o compartilhamento de genes (*gene sharing*) (Miyata e Yasunaga, 1978; Piatigorsky, 2003; Shikanai *et al*., 2006).

O *splicing* alternativo do transcrito primário de RNA (pré-mRNA) pode resultar na produção de polipeptídeos diferentes a partir de um mesmo segmento de DNA. Esse processo tem sido descrito para vários genes eucarióticos, alguns transposons eucarióticos e vírus de animais (*e.g.*, Smith *et al.*, 1989). Adicionalmente, os processos de *trans-splicing* e edição de RNA, descritos originalmente em tripanosomatídeos e atualmente confirmados em uma variedade maior de organismos, contribuem significativamente para a biodiversidade molecular observada nos organismos contemporâneos.

Além disso, um mesmo segmento de DNA pode codificar para mais de um gene, usando diferentes quadros de leitura (do inglês, *open reading frames*—ORFs). Esses genes sobrepostos, como são chamados com frequência, podem emergir também pelo uso da fita complementar de uma sequência de DNA. Os genes que codificam o RNA transportador da isoleucina ($tRNA^{ile}$) e do glutamina ($tRNA^{gln}$) e também os genes das subunidades 6 e 8 da ATPase no genoma mitocondrial humano ilustram esse fato. Os genes sobrepostos ocorrem amplamente no genoma dos vírus, bactérias e organelas. Espera-se que a taxa de evolução nos segmentos de DNA contendo genes sobrepostos seja mais lenta que a encontrada nas sequências similares com somente uma fase de leitura. A razão é que a proporção de códons não degenerados é maior nos genes sobrepostos que nos genes não sobrepostos, reduzindo assim a proporção de substituições sinônimas em relação ao número total de substituições (Miyata e Yasunaga, 1978).

O compartilhamento de genes consiste no recrutamento de um produto gênico para uma função adicional sem que haja qualquer mudança em sua sequência de aminoácidos. No processo de compartilhamento de genes, um gene adquire e mantém uma função secundária sem a perda ou duplicação da função primária exibida pelo produto gênico. Esse fenômeno foi descrito primeiramente em cristalinas, que são proteínas responsáveis pela manutenção da transparência e pela difração da luz no cristalino (Piatigorsky *et al.*, 1988).

As cristalinas α, β e γ, proteínas abundantes no cristalino dos olhos, são exemplos de proteínas que evoluíram a partir da duplicação e subsequente divergência de genes ancestrais, codificando proteínas diferentes. Estudos mostraram que as sequências das cristalinas são idênticas às de outras proteínas, como a desidrogenase láctica, liase do argininosuccinato e enolase α, cujas funções são completamente distintas (Estey *et al.*, 2007; Piatigorsky, 2003). A descoberta de que um mesmo polipeptídeo pode funcionar como proteína estrutural e também como enzima torna discutível a distinção entre enzimas e proteínas não enzimáticas. Esse polipeptídeo está sob a ação de duas forças de seleção distintas. Esse fato pode restringir a aquisição de mudanças na sequência do polipeptídeo, responsáveis por sua função enquanto cristalina, quando essas modificações sejam desvantajosas para a função enzimática.

10.7. Evolução das Sequências de DNA não Codificante

10.7.1. As sequências de DNA não codificante

Em Bacteria e Archaea, a maior parte das sequências de DNA codifica proteínas e RNAs, em contraste com o genoma dos representantes de Eukarya, nos quais predominam as sequências de DNA não codificantes. De fato, vários estudos comprovaram que uma fração surpreendente do genoma dos vertebrados é composta por DNA repetitivo (superior a 50% do genoma dos mamíferos) e que apenas 2% correspondem à fração codificante.

A grande variação quanto ao tamanho do genoma nos diferentes organismos (Tabela 10.3) deve-se, em grande parte, à presença de sequências de DNA não codificante, que incluem íntrons (Capítulo 8), DNA espaçador, sequências promotoras, sequências regulatórias, elementos de transposição (Capítulo 9), centrômeros, telômeros, sequências de DNA repetitivo (DNA satélite, minissatélites e microssatélites) e sequências de retrovírus integradas ao genoma (Capítulo 9).

Apesar de existir uma correlação positiva entre o tamanho do genoma e a quantidade de DNA não codificante, o significado biológico e evolutivo dessas sequências ainda não foi completamente elucidado. Por muito tempo, as sequências de DNA não codificante foram consideradas não funcionais. Nei (1969) denominou tais sequências de *non-sense* DNA e foi o primeiro a sugerir sua importância. Posteriormente, essas sequências foram tratadas por DNA "lixo" (*junk DNA*, Ohno, 1970), DNA "egoísta" (*selfish DNA*, Doolittle e Sapienza, 1980; Orgel e Crick, 1980) e DNA parasita (*parasite DNA*, Hickey, 1982).

Ao contrário do que se acreditava, porém, as sequências repetitivas estão envolvidas em diferentes processos no genoma. Por exemplo, tais sequências representam "pontos quentes" de recombinação, agem como elementos reguladores da transcrição, sítios de poliadenilação, entre outras funções (Makalowski, 2000). Além disso, a maioria das sequências de DNA repetitivo parece evoluir com um padrão organizado, denominado "evolução em concerto". Em conjunto, essas observações modificam completamente o contexto das sequências não codificantes e sugerem que elas, na verdade, desempenhem um papel importante no processo dinâmico da evolução dos genomas.

10.7.2. As sequências centroméricas e teloméricas

Os centrômeros são constituídos por sequências curtas de DNA repetitivo, semelhantes aos telômeros, que fornecem sítios específicos para a ligação da maquinaria de segregação do cromossomo durante a divisão celular. Na maioria das espécies, o padrão de organização dos centrômeros parece ser a ausência de conservação das sequências e a grande quantidade de sequências repetitivas. Talvez, o modelo mais bem estudado seja o descrito para *Arabidopsis thaliana* (Tyler-Smith e Floridia, 2000).

A função dos centrômeros é conservada em quase todos os eucariotos, mas o perfil de DNA satélite (veja item 10.7.3) é quase espécie-específico. Estudos mostram que as repetições em série presentes nas regiões centroméricas desempenham um papel importante na plasticidade dos genomas. Postula-se que as repetições em série presentes nos centrômeros podem ser vantajosas durante o processo de "evolução em concerto" (Plohl *et al.*, 2008).

Os telômeros são os terminais especializados dos cromossomos eucarióticos, constituídos por repetições em série de DNA, às quais se ligam proteínas. Os telômeros evitam a fusão entre extremidades dos cromossomos e lhes conferem proteção contra a degradação nucleolítica. Além disso, os telômeros servem de guia na localização dos cromossomos no núcleo e garantem a duplicação completa das suas extremidades (Kierszenbaum, 2000; McEachern *et al.*, 2000). Diversos estudos mostram a participação dos telômeros em processos celulares, tais como a regulação da expressão gênica, divisão celular, senescência e câncer (Shay e Wright, 2005; Garcia *et al.*, 2007).

Um dos traços mais marcantes dos telômeros eucarióticos é sua conservação sob o ponto de vista evolutivo. Os cromossomos da maioria das espécies apresentam telômeros compostos por repetições em série (5-8 pb) e grupos de guaninas na fita de DNA, cuja extremidade 3'-OH está exposta (Blackburn, 1994). Frequentemente, a sequência de DNA da unidade de repetição do telômero é compartilhada entre espécies distantes, sugerindo que o repertório das sequências teloméricas é relativamente limitado.

A organização dos telômeros, sua duplicação, o controle do tamanho dos telômeros e proteínas teloméricas e todos os avanços nessa área foram revisados recentemente (Garcia *et al.*, 2007; Lira *et al.*, 2007; Mason *et al.*, 2008).

10.7.3. O DNA satélite, minissatélite e microssatélite

Os DNAs satélite, minissatélite e microssatélite formam uma família de sequências repetitivas em série e correspondem a uma fração significativa do genoma. As unidades de repetição são relativamente curtas e variam entre o DNA satélite (10-110 pb), minissatélite (2-40 pb) e microssatélite (2-6 pb), divididos por classes de tamanho arbitrariamente. Provavelmente, o *crossing over* desigual seja um dos processos responsáveis pela geração de um número extremamente grande de cópias de DNA satélite.

A quantidade de DNA microssatélite parece estar diretamente relacionada ao tamanho do genoma, como mostrado por diversos estudos. Os microssatélites são ubíquos e altamente polimórficos no genoma dos eucariotos. Certas repetições ocorrem mais frequentemente que outras. Por exemplo, (A)n e (CA)n são mais comuns em humanos, (AT)n em plantas e (CT)n em algumas espécies de insetos (Primmer *et al.*, 1997). O genoma das aves contém menor quantidade de sequências não codificantes que a maioria dos mamíferos. Além disso, os microssatélites não estão associados a sequências dispersas curtas (do inglês, *short interspersed sequences* – SINEs). Estudos revelaram uma baixa frequência de microssatélites no genoma das aves, em particular nos minicromossomos, o que interfere diretamente na definição de marcadores para os mapas genéticos desses organismos.

Acreditava-se que essas sequências repetitivas fossem desprovidas de função e frequentemente eram referidas como DNA-lixo, conforme mencionado. No entanto, vários estudos sugerem um cenário diferente. O DNA satélite localiza-se principalmente em regiões de heterocromatina e acredita-se que ele esteja envolvido na estrutura e na função dos centrômeros. Alelos raros dos minissatélites estão associados ao oncogene *ras*, aumentando o risco de o indivíduo desenvolver alguns tipos de câncer. Alterações nos microssatélites, por sua vez, estão associadas a doenças neurodegenerativas humanas, como a síndrome do X frágil e a doença de Huntington (Orr e Zoghbi, 2007; Lukusa e Fryns, 2008).

Os microssatélites ocorrem tanto em procariotos quanto em eucariotos. Em procariotos, os microssatélites parecem estar associados à regulação da expressão gênica e a outras funções celulares, enquanto que o papel dessas sequências em eucariotos ainda necessita ser elucidado. Os microssatélites revelaram-se muito úteis como marcadores na construção de mapas genéticos, em testes de paternidade, análise forense, genética de populações etc. (Capítulo 19).

10.8. Evolução dos Genomas das Organelas

10.8.1. Considerações gerais

O conhecimento acerca do conteúdo e da organização do genoma das organelas é o ponto de partida para compreender os processos evolutivos peculiares a esses sistemas (Saccone *et al.*, 2000; Barbrook *et al.*, 2006; Woodson e Chory, 2008). Alguns aspectos sobre a origem endossimbionte das organelas e de seu genoma foram discutidos anteriormente (Capítulo 2). De acordo com a teoria endossimbionte, a mitocôndria e outros plastídeos seriam descendentes de uma eubactéria, cujos genes foram transferidos para o DNA nuclear. Esse processo, conhecido como transferência lateral, não foi interrompido, como comprovam alguns estudos recentes. Nesta seção, serão tratados alguns dos aspectos evolutivos do genoma da mitocôndria e do cloroplasto.

Uma característica comum entre o genoma mitocondrial dos metazoários e o do genoma do cloroplasto de plantas vasculares é que ambos são compostos por uma única molécula de DNA circular, que codifica genes essenciais às funções de respiração (mitocôndria) e fotossíntese (cloroplasto). Geralmente, uma única cópia de cada gene está presente em tais genomas. O genoma mitocondrial em plantas pode ser linear ou circular e, em muitos casos, a informação genética está dividida em duas moléculas de DNA chamadas de círculos subgenômicos (Li e Graur, 1991).

Nem todos os genes nucleares que codificam proteínas mitocondriais se originaram a partir do procarioto exógeno colonizando a célula eucariótica primitiva. No curso da evolução, novos genes podem ter sido adquiridos pelo núcleo. Nesse contexto, foi demonstrado que a forma mitocondrial da glutamina sintetase de *Drosophila* divergiu a partir da forma citoplasmática. A quantidade de variação acumulada nessas proteínas ao longo da evolução sugere que o tempo de divergência entre elas seja cerca de 600 milhões de anos, correspondente aproximadamente ao tempo de divergência entre vertebrados e invertebrados.

Como mencionado anteriormente, os genomas das organelas diferem em relação ao genoma nuclear quanto às taxas e padrões de evolução (Lynch *et al.*, 2006). Os estudos bioquímicos e os projetos de sequenciamento do genoma das organelas de vários organismos têm gerado uma quantidade extraordinária de dados, contribuindo de forma significativa para o conhecimento da organização e da função da organela. Os endereços listados na Tabela 10.2 contêm informações quanto aos genomas nucleares e das organelas.

10.8.2. O genoma mitocondrial

O tamanho do genoma mitocondrial pode variar de 6 kb a 2000 kb. A Tabela 10.5 apresenta exemplos dos genomas mitocondriais completamente sequenciados. Com relação à estrutura e organização do genoma mitocondrial, observam-se dois padrões distintos. O genoma das plantas pode ser significativamente grande (200-250 kb) e complexo, sendo sujeito à recombinação e rearranjos rápidos. Em contrapartida, o genoma dos metazoários apresenta tamanhos menores (~14-17 kb).

Com poucas exceções, o genoma mitocondrial codifica duas espécies de rRNAs (três, em plantas), um conjunto mais ou menos completo de tRNAs e um número limitado de mRNAs. O produto dos genes mitocondriais corresponde a vários complexos enzimáticos, juntamente com alguns produtos gênicos codificados pelo núcleo, que se localizam na membrana mitocondrial interna. A informação genética não é "redundante", exceto em plantas, nas quais o tamanho do genoma mitocondrial é muito maior que nos metazoários pela presença de cópias múltiplas dos genes.

De um modo geral, as principais diferenças encontradas entre o genoma mitocondrial dos vários organismos dizem respeito à presença e ausência de genes codificantes, que podem estar na mitocôndria ou no núcleo, no caso dos metazoários (Saccone *et al.*, 1999; Pereira, 2000). Com relação ao genoma das plantas, as diferenças observadas refletem a presença de sequências de DNA não codificante, além da migração de genes para o genoma mitocondrial.

Quanto à organização, a grande diversidade encontrada no genoma mitocondrial dos organismos pode ser atribuída principalmente à história evolutiva das várias linhagens (Saccone *et al.*, 1999; Boore, 1999; Pereira e Baker, 2006). A organização gênica do genoma mitocondrial dos vertebrados, por exemplo, parece ser extremamente conservada em grupos taxonômicos distintos, como em mamíferos placentários, peixes ósseos e cartilaginosos, anfíbios e outros vertebrados. Em contrapartida, grupos como aves, alguns répteis e marsupiais apresentam variação quanto ao número de genes e na organização do genoma mitocondrial.

10.8.3. O genoma do cloroplasto

O genoma do cloroplasto corresponde a uma molécula de DNA circular e geralmente apresenta um tamanho superior ao encontrado para o genoma mitocondrial, podendo variar de 120 kb a 220 kb. O tamanho do genoma do cloroplasto de alguns organismos é apresentado na Tabela 10.5. A variação no tamanho do genoma deve-se principalmente à presença de repetições invertidas (do inglês, *inverted repeats*—IRs), as quais separam regiões de cópia única pequena (do inglês, *small single copy*—SSC) e grande (*large single copy*—LSC).

O conteúdo gênico do genoma do cloroplasto é basicamente o mesmo entre os vários organismos, podendo apresentar variações espécie-específicas resultantes do processo de migração dos genes do cloroplasto para o genoma nuclear. Na maioria dos casos, o genoma do cloroplasto codifica quatro espécies de rRNAs, 30 tRNAs e cerca de 100 mRNAs, cujos produtos estão envolvidos na síntese protéica ou na fotossíntese (Hörmann *et al.*, 2007). A presença de íntrons foi descrita para alguns genomas.

O genoma do cloroplasto de *Arabidopsis thaliana* (154.478 pb) foi completamente sequenciado, revelando a presença de um par de repetições invertidas de 26.264 pb, separadas por regiões SSC (17.780 pb) e LSC (84.170 pb). Foram identificados quatro rRNAs, 37 tRNAs e um total de 87 potenciais genes codificantes de proteínas nesse genoma. A análise das sequências traduzidas de aminoácidos mostrou similaridade significativa com o genoma de *Nicotiana tabacum* (Sato *et al.*, 1999).

10.9. Perspectivas para o Estudo dos Genomas

10.9.1. A Genômica e outras "ômicas"

A genômica comparada consiste em analisar dois ou mais genomas em particular, cruzando informações sobre os genes relacionados no genoma de outros organismos. A genômica funcional, um neologismo que indica biologia baseada em genes ou genomas, procura analisar a expressão gênica e a função dos seus produtos em abordagens de transcriptômica e proteômica, respectivamente. A genômica evolutiva ou filogenômica, usa árvores evolutivas

Tabela 10.5. O genoma mitocondrial e do cloroplasto de alguns organismos.

Espécie	Grupo	Origem	Genoma	GenBank
Plasmodium falciparum	Alveolata	DNAmit	5.966	M99416
Caenorhabditis elegans	Nematoda	DNAmit	13.794	X54252
Anopheles gambiae	Arthropoda	DNAmit	15.363	L20934
Pagurus longicarpus	Arthropoda	DNAmit	15.63	AF150756
Homo sapiens	Chordata	DNAmit	16.569	V00662
Oryza sativa	Embryophyta	DNAmit	27.588	D32052
Paramecium aurelia	Alveolata	DNAmit	40.469	X15917
Podospora anserina	Fungi	DNAmit	100.314	X55026
Toxoplasma gondii	Alveolata	DNAcp	34.996	U87145
Epifagus virginiana	Embryophyta	DNAcp	70.028	M81884
Pinus thunbergii	Embryophyta	DNAcp	119.707	D17510
Oryza sativa	Streptophyta	DNAcp	134.525	X15901
Euglena gracilis	Euglenozoa	DNAcp	143.172	X70810
Chlorella vulgaris	Chlorophyta	DNAcp	150.613	AB001684
Cyanidium caldarium	Rhodophyta	DNAcp	164.921	AF022186
Porphyra purpurea	Rhodophyta	DNAcp	191.028	U38804

Tamanho do genoma em pares de bases (pb), número de acesso no Genbank, genoma mitocondiral (DNAmit) e genoma do cloroplasto (DNAcp). Fonte: Gobase (O'Brien *et al*., 2006); MitBASE (Attimonelli *et al*., 2000).

para inferir processos de evolução molecular de genes, famílias de genes ou de genomas com aplicações diversas (*e.g.*, Camargo e Nahum, 2005; Nahum *et al.*, 2006; Huerta-Cepas *et al.*, 2007; Nahum e Pereira, 2008).

A tecnologia de *microarrays* de DNA (do inglês, *oligonucleotide arrays*, ou simplesmente DNA *chips*) para a análise dos perfis da expressão gênica em larga escala no nível do mRNA tem contribuído significativamente para a análise genômica (Lockhart e Winzeler, 2000). O monitoramento da expressão gênica, detectando a presença e abundância dos mRNAs, é uma das aplicações mais importantes da técnica de *microarrays* de DNA. O termo transcriptoma tem sido usado para designar o conjunto de transcritos de um dado organismo, sendo altamente dinâmico e se modificando completamente de acordo com as diferentes condições celulares e ambientais. Em contraste com os *microarrays* de DNA, que buscam traçar um perfil global da expressão gênica de uma célula ou organismo, técnicas como *Northern blot*, *fingerprinting* de cDNA, PCR e outras analisam alvos específicos.

O termo proteômica refere-se ao estudo em larga escala das proteínas, sendo definido pelo estudo do conjunto de proteínas expressas por um organismo, tecido ou célula, incluindo as modificações no padrão de expressão das proteínas em condições diferentes. Tradicionalmente, o termo proteômica esteve associado aos géis bidimensionais de poliacrilamida (Anderson e Anderson, 1996), mas tomou um âmbito maior com o desenvolvimento da espectrometria de massa, que revolucionou a análise de proteínas a partir de 1990, superando a maioria das limitações impostas por outros métodos (Pandey e Mann, 2000). Consequentemente, a proteômica abrange a maior parte da genômica funcional, incluindo os estudos de identificação de proteínas, localização celular e interação. Modificações pós-traducionais, tais como glicosilação e fosforilação de proteínas, podem ser analisadas por essa metodologia, o que lhe confere uma enorme vantagem.

O vasto vocabulário nesta área (ORFeoma, quinoma, vacinoma etc.) reflete os avanços e os novos paradigmas nos campos da biologia molecular, genômica comparada e evolução molecular, desde que o termo "genômica" foi proposto por Thomas Roderick em 1986. Diante de tão pronunciada mudança no cenário das descobertas cientificas e tecnológicas, torna-se fundamental acompanhar os seus conceitos modernos, ferramentas e processos. Estão disponíveis na literatura eletrônica alguns livros e glossários com a terminologia atualizada, que podem funcionar como um auxiliar importante para todos os interessados na área. Nesse contexto, é necessário refletir sobre as mudanças curriculares e os instrumentos de formação de recursos humanos no cenário nacional e internacional para atender a demanda cada vez mais multidisciplinar e globalizada.

10.9.2. Os desafios e perspectivas

Um dos grandes desafios que surgem a partir do sequenciamento completo dos genomas é a organização da informação e sua disponibilização à comunidade científica. Esse desafio requer o desenvolvimento e aperfeiçoamento de ferramentas computacionais e suas aplicações em novas abordagens experimentais. Grupos dedicados à bioinformática reservam atenção especial à elaboração e construção de grandes bancos de dados. Alguns deles estão organizados como coleções de sequências genômicas com acesso a metodologias diversas (Nahum *et al.*, 2006, por exemplo).

Um desafio ainda maior continua sendo a inferência ou atribuição de função a partir das sequências moleculares (Bork e Koonin, 1998). O sequenciamento completo dos genomas é seguido pela etapa de anotação dos genes e proteínas, e análise da informação, usando-se métodos computacionais e estatísticos. Primeiramente, busca-se identificar o repertório gênico e os seus produtos. Em seguida, são atribuídas funções hipotéticas aos produtos gênicos ainda não caracterizados experimentalmente. Esse procedimento oferece desafios mesmo para organismos modelo, como a *Escherichia coli*, talvez o mais bem estudado genética e bioquimicamente (Serres *et al.*, 2001).

A disponibilidade das sequências dos genomas completos cria a oportunidade de se explorarem qualitativamente os aspectos estruturais, funcionais e evolutivos entre diferentes genomas. Uma das propostas é o estabelecimento de um conjunto ou repertório mínimo de genes, suficiente para sustentar as funções celulares e permitir a reconstrução do genoma do ancestral comum mais antigo (do inglês, *last common ancestor*—LCA) de Bacteria, Archaea e Eukarya (Penny e Poole, 1999; Glansdorff, 2000). Tais análises têm sido levadas a efeito analisando-se o genoma de *Haemophilus influenzae*, *Mycoplasma genitalium* e outros organismos. Embora esse tema seja bastante controvertido, alguns pesquisadores acreditam que seria possível definir esse conjunto mínimo de genes necessários para o funcionamento de qualquer célula existente e reconstruir a composição gênica dos organismos ancestrais.

O impacto da análise dos genomas completos pode ser avaliado considerando-se algumas das suas possíveis aplicações. Dentre elas, citam-se a identificação de novos caminhos metabólicos, enzimas com potencial na indústria, agentes antimicrobianos, alvos para diagnóstico de algumas doenças, candidatos a possíveis vacinas, bem como o auxílio no desenho de drogas (modelagem molecular). Essa abordagem contribui significativamente para o conhecimento sobre a evolução dos genes, dos genomas e dos organismos e da origem da vida (Meyer e van de Peer, 2003; Barton *et al.*, 2007; Lynch 2007).

Por causa da velocidade de obtenção e publicação dos dados nessa área, foram incluídos aqui alguns recursos de informação com relação aos bancos de dados, ferramentas computacionais, além de algumas instituições de pesquisa envolvidos no estudo dos genomas (Tabela 10.1 e 10.2).

10.10. Considerações Finais

Absolutamente extraordinária é a perspectiva de um estudo multidisciplinar, integrado e com potencial de resolução de questões das mais diversas. Sonhos e projetos daqueles que nos precederam nos séculos passados tomam hoje parte do cenário contemporâneo. Conhecer, respeitar, valorizar, explorar, modificar, melhorar. O tempo futuro, tão idealizado, não é senão o momento em que vivemos hoje. Estamos fascinados com a velocidade de geração de conhecimento e compreensão do sistema a nossa volta; com as possibilidades de aplicação do conhecimento e tecnologias modernas para a melhoria da vida do homem e outros seres em todos os ambientes do planeta que habitamos. Vale questionar o uso responsável e ético dessas ferramentas e o poder que a conquista desse conhecimento nos coloca nas mãos. Vale a confiança de que colheremos os frutos de um trabalho dinâmico e globalizado.

Agradecimentos

Ao Dr. Sergio Russo Matioli, pela oportunidade em contribuir para este livro de Biologia Molecular e Evolução, trazendo considerações sobre a Evolução dos Genomas. À Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e ao National Institutes of Health (NIH), pelo financia-

mento parcial da preparação deste capítulo. À Dra. Monica Riley e Dra. Gretta Serres (Marine Biological Laboratory, USA) e ao Dr. David Pollock (University of Colorado School of Medicine, USA), pelo apoio e estímulo em meus pós-doutoramentos no exterior, durante os quais expandi meu conhecimento acerca da genômica comparada e evolução molecular, enriquecendo enormemente a preparação deste capítulo. Agradeço especialmente ao Dr. Alastair Kerr (University of Edinburgh, Scotland), pelas sugestões sobre os temas de enviezamento na utilização dos códons e "evolução em concerto", e ao artista gráfico, José Adriano de Sousa (Universidade Federal de Minas Gerais), pela ilustração da Figura 10.1. Aos colegas e amigos Dra. Flora Maria de Campos Fernandes (Universidade Federal da Bahia) e Dr. Sérgio Luiz Pereira (The Hospital for Sick Children, Canadá), pelas críticas, sugestões e revisões deste trabalho.

Referências Bibliográficas

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., George, R.A., Lewis, S.E., Richards, S., Ashburner, M., Henderson, S.N., Sutton, G.G., Wortman, J.R., Yandell, M.D., Zhang, Q., Chen, L.X., Brandon, R.C., Rogers, Y.H., Blazej, R.G., Champe, M., Pfeiffer, B.D., Wan, K.H., Doyle, C., Baxter, *E.G.*, Helt, G., Nelson, C.R., Gabor, G.L., Abril, J.F., Agbayani, A., An, HJ., Andrews-Pfannkoch, C., Baldwin, D., Ballew, R.M., Basu, A., Baxendale, J., Bayraktaroglu, L., Beasley, E.M., Beeson, K.Y., Benos, P.V., Berman, B.P., Bhandari, D., Bolshakov, S., Borkova, D., Botchan, M.R., Bouck, J. *et al.* (2000). The genome sequence of *Drosophila melanogaster*. **Science. 287**: 2185-2195.

Amaldi, F., Lava-Sanchez, P.A. e Buongiorno-Nardelli, M. (1973). Nuclear DNA content variability in *Xenopus laevis*: redundancy regulation common to all gene families. **Nature 242**: 615-617.

Anderson, N.G. e Anderson, N.L. (1996). Twenty years of two-dimension electrophoresis: past, present and future. **Electrophoresis 17**: 443-453.

Arnheim, N., Krystal, M., Schmickel, R., Wilson, G., Ryder, O. e Zimmer, E. (1980). Molecular evidence for genetic exchanges among ribosomal genes on nonhomologous chromosomes in man and apes. **Proc. Natl. Acad. Sci. USA 77**: 7323-7327.

Attimonelli, M., Altamura, N., Benne, R., Brennicke, A., Cooper, J.M., D'Elia, D., Montalvo, A., Pinto, B., De Robertis, M., Golik, P., Knoop, V., Lanave, C., Lazowska, J., Licciulli, F., Malladi, B.S., Memeo, F., Monnerot, M., Pasimeni, R., Pilbout, S., Schapira, A.H., Sloof, P. e Saccone, C. (2000) MitBASE : a comprehensive and integrated mitochondrial DNA database. The present status. **Nucleic Acids Res. 28**: 148-152.

Barakat, A., Carels, N. e Bernardi, G. (1997). The distribution of genes in the genomes of Gramineae. **Proc. Natl. Acad. Sci. USA. 94**: 6857-6861.

Barbrook, A.C., Howe, C.J. e Purton, S. (2006) Why are plastid genomes retained in non-photosynthetic organisms? **Trends Plant Sci. 11**: 101-108.

Barton, N.H., Briggs, D.E.G., Eisen, J.A., Goldstein, D.B. e Patel, N.H. (2007) **Evolution.** Cold Spring Harbor Laboratory Press. 833 pp.

Bennetzen, J.L. (1998). The structure and evolution of angiosperm nuclear genomes. **Curr. Opin. Plant. Biol. 1**: 103-108.

Bernardi, G. (2007) The neoselectionist theory of genome evolution. **Proc Natl Acad Sci U S A. 104**: 8385-8390.

Blackburn, E.H. (1994). Telomeres: no end in sight. **Cell 77**: 621-623.

Blattner, F.R., Plunkett, G. 3rd, Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B. e Shao, Y. (1997). The complete genome sequence of *Escherichia coli* K-12. **Science 277**: 1453-1474.

Boore, J.L. (1999). Animal mitochondrial genomes. **Nucleic Acids Res. 27**: 1767-1780.

Bork, P. e Koonin, E.V. (1998). Predicting functions from protein sequences-where are the bottlenecks? **Nat. Genet. 18**: 313-318.

Bostock, C.J. (1986). Mechanisms of DNA sequence amplification and their evolutionary consequences. **Philos. Trans. R. Soc. Lond. B. Biol. Sci. 312**: 261-273.

Britten, R.J. e Kohne, D.E. (1968). Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. **Science 161**: 529-540.

Bult, C.J., White, O., Olsen, G.J., Zhou, L., Fleischmann, R.D., Sutton, G.G., Blake, J.A., FitzGerald, L.M., Clayton, R.A., Gocayne, J.D., Kerlavage, A.R., Dougherty, B.A., Tomb, J.F., Adams, M.D., Reich, C.I., Overbeek, R., Kirkness, E.F., Weinstock, K.G., Merrick, J.M., Glodek, A., Scott, J.L., Geoghagen, N.S. e Venter, J.C. (1996). Complete genome sequence of the methanogenic archaeon., *Methanococcus jannaschii*. **Science 273**: 1058-1073.

Buongiorno-Nardelli, M., Amaldi, F. e Lava-Sanchez, P.A. (1972). Amplification as a rectification mechanism for the redundant rRNA genes. **Nature New Biol. 238**: 134-137.

Callan, H.G. (1967). The organization of genetic units in chromosomes. **J. Cell Sci. 2**: 1-7.

Camargo, M.M. e Nahum, L.A. (2005) Adapting to a changing world: RAG genomics and evolution. **Human Genomics 2**: 133-138.

Cavalier-Smith T. (2005) Economy, speed and size matter: evolutionary forces driving nuclear genome miniaturization and expansion. **Ann Bot (Lond). 95**: 147-175.

Chiapello, H., Lisacek, F., Caboche, M. e Henaut, A. (1998). Codon usage and gene function are related in sequences of *Arabidopsis thaliana*. **Gene 209**: GC1-GC38.

Coleman, J.R., Papamichail, D., Skiena, S., Futcher, B., Wimmer, E. e Mueller, S. (2008) Vírus attenuation by genome-scale changes in codon pair bias. **Science. 320**: 1784-1787.

Costantini, M. e Bernardi, G. (2008) Replication timing, chromosomal bands, and isochores. **Proc Natl Acad Sci U S A. 105**: 3433-3437.

Costantini, M., Filippo, M.D., Auletta, F. e Bernardi, G. (2007) Isochore pattern and gene distribution in the chicken genome. **Gene. 400**: 9-15.

Cuny, G., Soriano, P., Macaya, G. e Bernardi, G. (1981). The major components of the mouse and human genomes. 1. Preparation, basic properties and compositional heterogeneity. **Eur. J. Biochem. 115**: 227-233.

de Rosa, R. e Labedan, B. (1998). The evolutionary relationships between the two bacteria *Escherichia coli* and *Haemophilus influenzae* and their putative last common ancestor. **Mol. Biol. Evol. 15**: 17-27.

Descorps-Declère, S., Lemoine, F., Sculo, Q., Lespinet, O. e Labedan, B. (2008) The multiple facets of homology and their use in comparative genomics to study the evolution of genes, genomes, and species. **Biochimie. 90**: 595-608.

Doolittle, W.F. e Sapienza, C. (1980). Selfish genes, the phenotype paradigm and genome evolution. **Nature 284**: 601-603.

Dunham, I., Shimizu, N., Roe, B.A., Chissoe, S., Hunt, A.R., Collins, J.E., Bruskiewich, R., Beare, D.M., Clamp, M., Smink, L.J., Ainscough, R., Almeida, J.P., Babbage, A., Bagguley, C., Bailey, J., Barlow, K., Bates, K.N., Beasley, O., Bird, C.P., Blakey, S., Bridgeman, A.M., Buck, D., Burgess, J., Burrill, W.D., O'Brien, K.P. *et al.* (1999). The DNA sequence of human chromosome 22. **Nature 402**: 489-495.

Duret, L. e Mouchiroud, D. (1999). Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. **Proc. Natl. Acad. Sci. USA 96**: 4482-4487.

Edelman, G.M. e Gally, J.A (1970). Arrangement and evolution of eukaryotic genes. **In** Schmitt, F.O. (ed) **The Neurosciences**: **Second Study Program**. Rockefeller University Press, New York.

Eickbush, T.H. e Eickbush, D.G. (2007) Finely orchestrated movements: evolution of the ribosomal RNA genes. **Genetics. 175**: 477-485.

Elder, J.F. Jr e Turner, B.J. (1995). Concerted evolution of repetitive DNA sequences in eukaryotes. **Q. Rev. Biol. 70**: 297-320.

Estey, T., Piatigorsky, J., Lassen, N. e Vasiliou, V. (2007) ALDH3A1: a corneal crystallin with diverse functions. **Exp Eye Res. 84**: 3-12.

Eyre-Walker, A. (1999). Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA. **Genetics 152**: 675-683.

Fitch, W.D. (1970). Distinguishing homologous from analogous proteins. **Systematic Zool. 19**: 99-113.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M. *et al.* (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. **Science 269**: 496-512.

Fraser, C.M., Casjens, S., Huang, W.M., Sutton, G.G., Clayton, R., Lathigra, R., White, O., Ketchum, K.A., Dodson, R., Hickey, E.K., Gwinn, M., Dougherty. B., Tomb, J.F., Fleischmann, R.D., Richardson, D., Peterson, J., Kerlavage, A.R., Quackenbush, J., Salzberg, S., Hanson, M., van Vugt, R., Palmer, N., Adams, M.D., Gocayne, J., Venter, J.C. *et al.* (1997). Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. **Nature 390**: 580-586.

Galtier, N. e Lobry, J.R. (1997). Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature

in prokaryotes. **J. Mol. Evol. 44**: 632-636.
Galtier, N. e Mouchiroud, D. (1998). Isochore evolution in mammals: a human-like ancestral structure. **Genetics 150**: 1577-1584.
Garcia, C.K., Wright, W.E. e Shay, J.W. (2007) Human diseases of telomerase dysfunction: insights into tissue aging**. Nucleic Acids Res. 35**: 7406-7416.
Glansdorff, N. (2000). MicroReview: about the last common ancestor, the universal life-tree and lateral gene transfer: a reappraisal. **Mol. Microbiol. 38**: 177-185.
Goffeau *et al.* (1997). The yeast genome directory. **Nature 387 (Suppl)**: 5-105.
Graur, D. e Li, W-H. (2000) **Fundamentals of Molecular Evolution**. Sinauer Associates, Inc. New York. 481 pp.
Gregory, T.R. (2005) **The Evolution of the Genome.** Elsevier/Academic Press. 740 pp.
Gregory, T.R., Nicol, J.A., Tamm, H., Kullman, B., Kullman, K., Leitch, I.J., Murray, B.G., Kapraun, D.F., Greilhuber, J. e Bennett, M.D. (2007) Eukaryotic genome size databases. Nucleic Acids Res. 35(Database issue): D332-8.
Haldane, J.B.S. (1932). **The causes of evolution**. Longmans, Green, New York.
Hattori, M., Fujiyama, A., Taylor, T.D., Watanabe, T., Yada, T., Park, H.S., Toyoda, A., Ishii, K., Totoki, Y., Choi, D.K., Soeda, E., Ohki, M., Takagi, T., Sakaki, Y., Taudien, S., Blechschmidt, K., Polley, A., Menzel, U., Delabar, J., Kumpf, K., Lehmann, R., Patterson, D., Reichwald, K., Rump, A., Schillhabel, M. e Schudy, A. (2000). The DNA sequence of human chromosome 21. The chromosome 21 mapping and sequencing consortium. **Nature 405**: 311-319.
Hickey, D.A. (1982). Selfish DNA: a sexually-transmitted nuclear parasite. **Genetics 101**: 519-531.
Hood, L., Campbell, J.H. e Elgin, S.C. (1975). The organization, expression, and evolution of antibody genes and other multigene families. **Annu. Rev. Genet. 9**: 305-353.
Hörmann, F., Soll, J. e Bölter, B. (2007) The chloroplast protein import machinery: a review. **Methods Mol Biol. 390**: 179-193.
Huerta-Cepas, J., Dopazo, H., Dopazo, J. e Gabaldón, T. (2007) The human phylome. **Genome Biol. 8**: R109.
Kierszenbaum, A.L. (2000). Telomeres: more than chromosomal non-sticking ends. **Mol. Reprod. Dev. 57**: 2-3.
Koonin, E.V., Tatusov, R.L. e Rudd, K.E. (1995). Sequence similarity analysis of *Escherichia coli* proteins: functional and evolutionary implications. **Proc. Natl. Acad. Sci. USA 92**: 11921-11925.
Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., Levine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C. *et al.* (2001) Initial sequencing and analysis of the human genome. **Nature 409**: 860-921.
Lewis, E.B. (1951). Pseudoallelism and gene evolution. **Cold Spring Harb. Symp. Quant. Biol. 16**: 159-174.
Li, W.-H e Graur, D. (1991). **Fundamentals of Molecular Evolution**. Sinauer Associates, Sunderland, Massachusetts.
Liao, D. (1999). Concerted evolution: molecular mechanism and biological implications. **Am. J. Hum. Genet. 64**: 24-30.
Liao, D., Pavelitz, T., Kidd, J.R., Kidd, K.K. e Weiner, A.M. (1997). Concerted evolution of the tandemly repeated genes encoding human U2 snRNA (the RNU2 locus) involves rapid intrachromosomal homogenization and rare interchromosomal gene conversion. **EMBO J. 16**: 588-598.
Lira, C.B., Giardini, M.A., Neto, J.L., Conte, F.F. e Cano, M.I. (2007) Telomere biology of trypanosomatids: beginning to answer some questions. **Trends Parasitol. 23**: 357-362.
Lockhart, D.J. e Winzeler, E.A. (2000). Genomics, gene expression and DNA arrays. **Nature 405**: 827-836.
Lukusa, T. e Fryns, J.P. (2008) Human chromosome fragility. **Biochim Biophys Acta.** 1779: 3-16.
Lynch, M. (2007) **The Origins of Genome Architecture.** Sinauer Associates, Inc. 389 pp.
Lynch, M., Koskella, B. e Schaack, S. (2006) Mutation pressure and the evolution of organelle genomic architecture. **Science. 311**: 1727-1730.
Macaya, G., Thiery, J.P. e Bernardi, G. (1976). An approach to the organization of eukaryotic genomes at a macromolecular level. **J. Mol. Biol. 108**: 237-254.
MacCarthy, T. e Bergman, A. (2007) The limits of subfunctionalization. **BMC Evol Biol. 7**: 213.
Makalowski, W. (2000). Genomic scrap yard: how genomes utilize all that junk. **Gene 259**: 61-67.
Mason, J.M., Frydrychova, R.C. e Biessmann, H. (2008) *Drosophila* telomeres: an exception providing new insights. **Bioessays. 30**: 25-37.
McEachern, M.J., Krauskopf, A. e Blackburn, E.H. (2000). Telomeres and their control. **Annu. Rev. Genet. 34**: 331-358.
Maxam, A.M. e Gilbert, W, (1977). A new method for sequencing DNA. **Proc. Natl. Acad. Sci. USA 74**:560–564.
Meyer, A e van de Peer, Y. (2003) **Genome Evolution: Gene and Genome Duplications and the Origin of Novel Gene Functions.** Springer. 256 pp.
Mighell, A.J., Smith, N.R., Robinson, P.A. e Markham, A.F. (2000). Vertebrate pseudogenes. **FEBS Lett. 468**: 109-114.
Miyata, T. e Yasunaga, T. (1978). Evolution of overlapping genes. **Nature 272**: 532-535.
Muller, H.J. (1936). Bar duplication. **Science 83**: 528-530.
Muto, A. e Osawa, S. (1987). The guanine and cytosine content of genomic DNA and bacterial evolution. **Proc. Natl. Acad. Sci. USA 84**: 166-169.
Nahum, L.A. e Pereira, S.L. (2008) Phylogenomics, Protein Family Evolution, and the Tree of Life. In: Smolinski TG; Milanova MG; Hassanien A-E.. **Computational Intelligence in Biomedicine and Bioinformatics: Current Trends and Applications.** 1 ed.: Springer, v. 122, p. 259-279.
Nahum, L.A. e Riley, M. (2001) Divergence of function in sequence-related groups of *Escherichia coli* proteins. **Genome Research 11**: 1375-1381.
Nahum, L.A., Goswami, S., Serres, M.H. (2009) Protein families reflect the metabolic diversity of organisms and provide support for functional prediction. **Physiol Genomics. 38**:250-60.
Nahum, L.A., Pereira, S.L., Fernandes, F.M.C., Matioli, S.R. e Wajntal, A. (2003) Diversification of Ramphastinae (Aves, Ramphastidae) prior to the Cretaceous/Tertiary boundary as shown by molecular clock of mtDNA sequences. **Genetics and Molecular Biology 26**: 411-418.
Nahum, L.A., Reynolds, M.T., Wang, Z.O., Faith, J.J., Jonna, R., Jiang, Z.O., Meyer, T.J., e Pollock, D.D. (2006) *EGenBio*: A data management system for evolutionary genomics and biodiversity. **BMC Bioinformatics 7**: S7. doi: 10.1186/1471-2105-7-S2-S7
Nei, M. (1969). Gene duplication and nucleotide substitution in evolution. **Nature** 221: 40-44.
Nei, M. e Kumar, S. (2000). **Molecular Evolution and Phylogenetics**. Oxford University Press, New York.
Nei, M. e Rooney, A.P. (2005) Concerted and birth-and-death evolution of multigene families. **Annu Rev Genet. 39**: 121-152.
O'Brien, E.A., Zhang, Y., Yang, L., Wang, E., Marie, V., Lang, B.F. e Burger, G. (2006) GOBASE--a database of organelle and bacterial genome information. Nucleic Acids Res.34(Database issue):D697-699.
Ohno, S. (1970). **Evolution by gene duplication**. Springer-Verlag, New York.
Orgel, L.E. e Crick, F.H. (1980). Selfish DNA: the ultimate parasite. **Nature 284**: 604-607.
Orr, H.T. e Zoghbi, H.Y. (2007) Trinucleotide repeat disorders. **Annu Rev Neurosci.** 30: 575-621.
Page, R.D.M e Holmes, E.C. (1998). **Molecular Evolution**: **A Phylogenetic Approach**. Blackwell Science. Malden, MA, USA. 346 pp.
Pandey, A. e Mann, M. (2000). Proteomics to study genes and genomes. **Nature 405**: 837-846.
Penny, D. e Poole, A. (1999). The nature of the last universal common ancestor. **Curr. Opin. Genet. Dev. 9**: 672-677.
Pereira, S.L. 2000. Mitochondrial genome organization and vertebrate phylogenetics. **Genet. Mol. Biol. 23**: 745-752.
Pereira, S.L. e Baker, A.J. (2006) A mitogenomic timescale for birds detects variable phylogenetic rates of molecular evolution and refutes the standard molecular clock. **Mol Biol Evol. 23**: 1731-1740.
Piatigorsky J. (2003) Crystallin genes: specialization by changes in gene regulation may precede gene duplication. **J Struct Funct Genomics. 3**: 131-137.
Piatigorsky, J., O'Brien, W.E., Norman, B.L., Kalumuck, K., Wistow, G.J., Borras, T., Nickerson, J.M. e Wawrousek, E.F. (1988). Gene sharing by delta-crystallin and argininosuccinate lyase. **Proc. Natl. Acad. Sci. USA. 85**: 3479-3483.
Plohl, M., Luchetti, A., Mestrović, N. e Mantovani, B. (2008) Satellite

DNAs between selfishness and functionality: structure, genomics and evolution of tandem repeats in centromeric (hetero)chromatin. **Gene. 409**: 72-82.

Primmer, C.R., Raudsepp, T., Chowdhary, B.P., Moller, A.P. e Ellegren, H. (1997). Low frequency of microsatellites in the avian genome. **Genome Res. 7**: 471-482.

Riley, M. e Labedan, B. (1997). Protein evolution viewed through *Escherichia coli* protein sequences: introducing the notion of a structural segment of homology, the module. **J. Mol. Biol. 268**: 857-868.

Ronaghi, M., Uhlén, M. e Nyrén, P. (1998) A sequencing method based on real-time pyrophosphate. **Science. 281**: 363-365.

Roth, C., Rastogi, S., Arvestad, L., Dittmar, K., Light, S., Ekman, D. e Liberles DA. (2007) Evolution after gene duplication: models, mechanisms, sequences, systems, and organisms. **J Exp Zoolog B Mol Dev Evol. 308**: 58-73.

Saccone, C., De Giorgi, C., Gissi, C., Pesole, G. e Reyes, A. (1999). Evolutionary genomics in Metazoa: the mitochondrial DNA as a model system. **Gene 238**: 195-209.

Saccone, C., Gissi, C., Lanave, C., Larizza, A., Pesole, G. e Reyes, A. (2000). Evolution of the mitochondrial genetic system: an overview. **Gene 261**: 153-159.

Sanger, F., Nicklen, S. e Coulson A.R. (1977). DNA sequencing with chain-terminating inhibitors. **Proc. Natl. Acad. Sci. USA** 74: 5463–4467.

Santos, M.A., Moura, G., Massey, S.E. e Tuite, M.F. (2004) Driving change: the evolution of alternative genetic codes. **Trends Genet. 20**: 95-102.

Sato, S., Nakamura, Y., Kaneko, T., Asamizu, E. e Tabata, S. (1999). Complete structure of the chloroplast genome of *Arabidopsis thaliana*. **DNA Res. 6**: 283-290.

Serres, M.H., Gopal, S., Nahum, L.A., Liang, P., Gaasterland, T. e Riley, M. (2001) A functional update of the *E. coli* K-12 genome. **Genome Biology 2**: 0035.1-0035.7

Sharp, P.M. e Matassi, G. (1994). Codon usage and genome evolution. **Curr. Opin. Genet. Dev. 4**: 851-860.

Sharp, P.M., Averof, M., Lloyd, A.T., Matassi, G. e Peden, J.F. (1995). DNA sequence evolution: the sounds of silence. **Philos. Trans. R. Soc. Lond. B. Biol. Sci. 349**: 241-247.

Shay, J.W. e Wright, W.E. (2005) Senescence and immortalization: role of telomeres and telomerase. **Carcinogenesis. 26**: 867-874.

Shikanai, T. (2006) RNA editing in plant organelles: machinery, physiological function and evolution. **Cell Mol Life Sci. 63**: 698-708.

Simpson, A.J., Reinach, F.C., Arruda, P., Abreu, F.A., Acencio, M., Alvarenga, R., Alves, L.M., Araya, J.E., Baia, G.S., Baptista, C.S., Barros, M.H., Bonaccorsi, E.D., Bordin, S., Bove, J.M., Briones, M.R., Bueno, M.R., Camargo, A.A., Camargo, L.E., Carraro, D.M., Carrer, H., Colauto, N.B., Colombo, C., Costa, F.F., Costa, M.C., Costa-Neto, C.M., Coutinho, L.L., Cristofani, M., Dias-Neto, E., Docena, C., El-Dorry, H., Facincani, A.P., Ferreira, A.J., Ferreira, V.C., Ferro, J.A., Fraga, J.S., Franca, S.C., Franco, M.C., Frohme, M., Furlan, L.R., Garnier, M., Goldman, G.H., Goldman, M.H., Gomes, S.L., Gruber, A., Ho, P.L., Hoheisel, J.D., Junqueira, M.L., Kemper, E.L., Kitajima, J.P., Marino, C.L. *et al.* (2000). The genome sequence of the plant pathogen *Xylella fastidiosa*. The *Xylella fastidiosa* Consortium of the Organization for Nucleotide Sequencing and Analysis. **Nature 406**: 151-157.

Smith, C.W., Patton, J.G. e Nadal-Ginard, B. (1989). Alternative splicing in the control of gene expression. **Annu. Rev. Genet. 23**: 527-577.

Stoletzki, N. e Eyre-Walker, A. (2007) Synonymous codon usage in *Escherichia coli*: selection for translational accuracy. **Mol Biol Evol. 24**: 374-381.

Sueoka, N. (1964). On the evolution of the informational macromolecules. Pp 479-496. **In** Bryson, V. and H.J. Vogel (eds.). **Evolving Genes and Proteins**. Academic Press, New York.

Sueoka, N. (1988). Directional mutation pressure and neutral molecular evolution. **Proc. Natl. Acad. Sci. USA 85**: 2653-2657.

Tang, H., Bowers, J.E., Wang, X., Ming, R., Alam, M. e Paterson, A.H. (2008) Synteny and collinearity in plant genomes. **Science. 320**: 486-488.

The *C. elegans* Sequencing Consortium (1998). Genome Sequence of the Nematode *C. elegans*: A Platform for Investigating Biology. **Science 282**: 2012-2018.

Thomas, B.R. (1970). The origin of the genetic code. **Biochem. Biophys. Res. Commun. 40**: 1289-1296.

Tyler-Smith, C. e Floridia, G. (2000). Many paths to the top of the mountain: diverse evolutionary solutions to centromere structure. **Cell. 102**: 5-8.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., Gocayne, J.D., Amanatides, P., Ballew, R.M., Huson, D.H., Wortman, J.R., Zhang, Q., Kodira, C.D., Zheng, X.H., Chen, L., Skupski, M., Subramanian, G., Thomas, P.D., Zhang, J., Gabor Miklos, G.L., Nelson, C., Broder, S., Clark, A.G., Nadeau, J., McKusick, V.A., Zinder, N., Levine, A.J., Roberts, R.J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C. *et al.* (2001). The sequence of the human genome. **Science 291**: 1304-1351.

Walsh, J.B. (1987). Sequence-dependent gene conversion: can duplicated genes diverge fast enough to escape conversion? **Genetics 117**: 543-557.

Wellauer, P.K., Reeder, R.H., Dawid, I.B. e Brown, D.D. (1976). The arrangement of length heterogeneity in repeating units of amplified and chromosomal ribosomal DNA of *Xenopus laevis*. **J. Mol. Biol. 105**: 487-505.

Winkler, H. (1920). **Vererbung und Ursache der Parthenogenese im Pflanze?** Und Tierreich, Fisher, Jena.

Woodson JD, Chory J. (2008) Coordination of gene expression between organellar and nuclear genomes**. Nat Rev Genet. 9**: 383-395.

Woynarowski, J.M., Krugliak, M. e Ginsburg, H. (2007) Pharmacogenomic analyses of targeting the AT-rich malaria parasite genome with AT-specific alkylating drugs. **Mol Biochem Parasitol. 154**: 70-81.

Zimmer, E.A., Martin, S.L., Beverley, S.M., Kan, Y.W. e Wilson, A.C. (1980). Rapid duplication and loss of genes coding for the alpha chains of hemoglobin. **Proc. Natl. Acad. Sci. USA 77**: 2158-2162.

Capítulo 11

Biologia evolutiva do desenvolvimento

Luiz Paulo Moura Andrioli (lpma@usp.br)
Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

"*Os biólogos são capazes de descrever em detalhe a composição, digamos, de um rato. São capazes de dizer como é que o rato se desloca, como respira, como digere. Mas não sabem absolutamente nada acerca da forma como ele se constrói a partir da célula ovo.*" (François Jacob, 1981)

11.1. Introdução

A Biologia do desenvolvimento engloba a tradicional área de Embriologia, mas é uma disciplina abrangente, que não se limita aos estágios iniciais do desenvolvimento. Em formas adultas, ocorre diariamente a reposição contínua de tipos celulares a partir de células-tronco nos animais ou de células meristemáticas nas plantas. Além disso, alguns animais são capazes de regenerar partes inteiras do organismo, outros passam por metamorfose durante seu ciclo de vida, enquanto algumas plantas são capazes de regenerar um organismo completo a partir de um pedaço de tecido somático. Todos esses processos biológicos fazem parte do escopo dessa disciplina. A abrangência de assuntos também é acompanhada pelo número de áreas que integram a moderna biologia do desenvolvimento, como a biologia molecular, genética, bioquímica, citologia, histologia, fisiologia e ecologia. Uma das interfaces da biologia do desenvolvimento que mais tem despertado atenção é com a biologia evolutiva, constituindo a biologia evolutiva do desenvolvimento, também conhecida pelo jargão "Evo-devo" (mnemônico do inglês "evolutionary developmental biology"). A Biologia evolutiva do desenvolvimento tem como proposta compreender a diversidade morfológica, buscando uma síntese entre os processos do desenvolvimento de um indivíduo (ontogenia) e os processos da história evolutiva de uma espécie ou de um grupo de espécies (filogenia). Este capítulo apresenta alguns aspectos abordados por essa disciplina e as dificuldades de integração entre as duas áreas formadoras. Inicialmente, veremos que, apesar de apontada como uma síntese recente, as duas áreas formadoras já apresentaram ligações no passado.

11.2. Primórdios da disciplina

Antes mesmo da publicação de "A Origem das Espécies" em 1859, por Charles Darwin, os assuntos que hoje compõem a biologia do desenvolvimento já eram abordados nos debates entre os cientistas que defendiam e aqueles que rejeitavam a evolução das espécies. Naquela época, eram muitas as lacunas no registro fóssil, assim como eram desconhecidos muitos princípios bioquímicos, celulares, genéticos e mesmo dos processos evolutivos que levam às modificações nas espécies. Portanto, as áreas de anatomia comparada e embriologia, que dispunham de um conjunto maior de dados sistematizados, fomentaram as discussões de cunho evolutivo.

O estudo comparativo realizado com embriões animais no início do século XIX, como o observado no transcorrer da embriologia de espécies de diferentes classes de vertebrados, levou à sugestão de que formas *superiores* recapitulam estágios adultos de fósseis ou de espécies *inferiores*, usando a terminologia da época. Em 1846, o influente embriologista Karl von Baer criticou os extremos da teoria da recapitulação, afirmando que não existe uma recapitulação plena no desenvolvimento de espécies *superiores* necessariamente passando pelos estágios adultos de formas *primitivas*. Já o famoso biólogo Ernst Haeckel propôs, em 1856, um mecanismo para a recapitulação, onde novas formas biológicas teriam surgido pela adição de um estágio terminal no desenvolvimento embrionário das formas ancestrais. É interessante notar que, no caso desses dois cientistas, eles não estavam convencidos da transformação das espécies (von Baer) ou tinham uma motivação lamarquista (Haeckel).

Com Darwin, surgiu uma reinterpretação da *Scala Naturae*. A partir de então, as espécies de um grupo taxonômico deixaram de ser consideradas entidades imutáveis, criadas por uma entidade à parte, e sim espécies historicamente relacionadas, que têm uma ancestralidade comum entre elas. Um corolário dessa afirmação é que as evidências da ancestralidade comum devem estar presentes em estruturas compartilhadas por diferentes espécies de um táxon. Alguns exemplos nesse sentido já eram conhecidos de longa data, como a semelhança observada entre os ossos dos vertebrados. Por outro lado, a asa de uma ave e o braço de um humano, por exemplo, apresentam grandes diferenças entre si. Um dos conceitos que surgiu para explicar essa observação conflitante foi o de homologia (Moczek, 2008). Esse termo foi cunhado, em sua acepção mais moderna, em 1843, pelo zoólogo inglês Richard Owen, contemporâneo de Darwin. Segundo Owen, homologia é a propriedade que o mesmo órgão tem em diferentes animais sob toda variedade de forma e função. Ele utilizava-se da noção de arquétipo, que pode ser definido como um padrão estrutural básico de um grupo de organismos afins. Owen, entretanto, não aceitou as idéias de evolução de Darwin e Wallace, e seu arquétipo não seria uma forma ancestral, mas um projeto básico seguido por um criador divino. Após o advento da teoria da evolução, o termo "homologia" passou a ter um sentido de equivalência de estruturas que já estavam presentes em seu ancestral comum mais recente, que herdaram essa estrutura com ou sem algum grau de modificação. Por exemplo, a asa e o braço são homólogos, o que implica que essas estruturas não são idênticas, mas formas adaptadas para funções diferentes, surgidas a partir da mesma estrutura, no caso, um par de apêndices anteriores presente em um ancestral vertebrado comum. Darwin também conhecia exemplos de homologia entre embriões e formas larvais de diferentes filos, que apontou como excelentes evidências da descendência com

modificação. Um desses exemplos eram as cracas. Esses organismos são sésseis e eram considerados moluscos pela presença da concha que envolve as formas adultas. No entanto, as larvas desses organismos são típicas de crustáceos e não de moluscos—um exemplo concreto de como a ontogenia recupera a filogenia.

Já no final do século XIX e início do século XX, ocorreu uma progressiva dicotomia entre a evolutiva e a biologia do desenvolvimento, que motivou uma ruptura entre as duas áreas até recentemente. Os aspectos evolutivos da recapitulação, sobretudo apoiados na premissa lamarquista concebida por Haeckel, estavam desacreditados. Os embriologistas passaram a se dedicar integralmente a um programa de estudos para compreender os mecanismos do desenvolvimento valendo-se da intervenção experimental e/ou de apenas descrição cada vez mais detalhada do início do desenvolvimento de um número crescente de espécies disponíveis nos centros de pesquisa. Por outro lado, com a redescoberta das leis de Mendel e a formulação da teoria cromossômica da herança, os evolucionistas concentram seus esforços nos genes, ao final, as unidades hereditárias passíveis de transformação evolutiva.

Richard Goldschmidt, um pesquisador alemão que emigrou para os Estados Unidos em 1935, publicou, em 1940, um livro que desafiou as ideias vigentes sobre o papel dos genes durante o desenvolvimento e na evolução (Godschmidt, 1940). Segundo Goldschmidt, a microevolução ocorria por mudanças de frequências de genes que sofrem micromutações, mas a evolução acima do nível de espécie, a macroevolução, deveria ocorrer por mecanismos distintos. Um desses mecanismos propostos por Goldschmidt, que ele denominou por "mutações sistêmicas", seria resultado do efeito de posição dos genes envolvidos em rearranjos cromossômicos. Goldschmidt admitiu outra possibilidade para explicar a macroevolução, que denominou por "macromutações do desenvolvimento". Neste caso, sua hipótese era que uma única mutação em um gene do desenvolvimento poderia efetivar potencialidades ontogenéticas no embrião e desencadear mudanças rápidas e drásticas no fenótipo. Cabe ressaltar que Goldschmidt definia gene como uma unidade do desenvolvimento e não um loco gênico ou um alelo. Em seus experimentos, Goldschmidt verificou o aparecimento de fenocópias, ou seja, de alterações nos fenótipos de *Drosophila melanogaster* que se assemelhavam a mutações ou conjunto de mutações conhecidas, mas que são induzidas por alterações físico-químicas durante o desenvolvimento dos insetos. As fenocópias, ao contrário das mutações genéticas, são reversíveis e, portanto, não se mantêm nas gerações subsequentes. Certos tipos de fenocópias às quais Goldschmidt dedicou suas pesquisas eram relacionadas às mutações homeóticas, ou seja, mutações que alteram o padrão de aparecimento de estruturas no padrão de segmentação dos insetos. A mutação *Ultrabithorax* (*Ubx*), por exemplo, faz com que as moscas drosófilas passem a ter dois pares de asas, ao contrario do padrão selvagem, de apenas um par de asas no segundo segmento toráxico, como na maioria dos demais dípteros (veja Figura 11.1).

A obra de Goldschmidt provocou grande reação negativa por parte da comunidade dedicada ao estudo da evolução justamente por contradizer o paradigma corrente na época, quando se imaginava que a macroevolução nada mais era do que o acúmulo de processos microevolutivos. Além disso, o fato de mostrar a importância do ambiente na variação fenotípica e não apenas do genótipo, foi interpretado como sendo uma tentativa de reaproximação com as idéias de Lamarck. Além de Goldschmidt, o embriologista britânico Conrad Hal Waddington, que também trabalhou nos Estados Unidos, colaborando com Dobzhansky e Sturtevant a partir de 1939, lançou ideias inovadoras, hoje consideradas importantes sobre o papel de genes no desenvolvimento com consequências evolutivas. Waddington foi um dos

Figura 11.1. Vista lateral de um indivíduo de *Drosophila melanogaster* portador da mutação *bithorax*. Notar o terceiro segmento torácico, onde há um par de asas no lugar dos halteres (Foto: cortesia do Dr. Paulo A. Otto).

principais precursores da biologia do desenvolvimento, criando vários conceitos na área, como o de "paisagem epigenética" (*epigenetic landscape*, em inglês), uma concepção visual das possíveis "escolhas" que uma célula poderia ter durante o desenvolvimento; o de homeostase genética, um conceito no qual previu o desenvolvimento embrionário como possuidor de mecanismos que minimizassem as perturbações genéticas e ambientais; e o conceito de canalização, onde propôs que a evolução poderia ocorrer no sentido de engessar as trajetórias do desenvolvimento (Slack, 2002). Também Waddington não teve o reconhecimento da importância de seu trabalho durante a época em que atuou, principalmente pelo fato que, até então, a embriologia não tinha uma explicação para o papel dos genes durante o desenvolvimento. Em contraposição, grandes progressos dessa época ocorreram na elucidação do papel dos genes repositórios de informações metabólicas. Somente depois da década de 1970, com a caracterização dos mecanismos de regulação da atividade gênica, as áreas da genética e da evolução estavam prontas para incorporar, em sua teoria, os mecanismos do desenvolvimento.

11.3. Mutações homeóticas, genes *Hox* e os *homeoboxes*

O surgimento das técnicas de biologia molecular nas décadas de 1970 e 1980, aliadas à genética clássica, possibilitaram revelar a natureza e a forma de atuação de moléculas reguladoras, fundamentais do início do desenvolvimento da *Drosophila*, responsáveis pelo estabelecimento do plano corporal nessa espécie. Essas moléculas, fatores de transcrição ou moléculas de sinalização atuam em uma cascata geradora de informação posicional no embrião, a cascata de segmentação, que determina a divisão do corpo da larva em segmentos correspondentes aos futuros segmentos do corpo do adulto. Imediatamente após a cascata de segmentação, atuam os genes da família *Hox*, codificadores para fatores de transcrição, que possuem uma sequência conservada, denominada *homeobox*, codificadora para uma sequência de 60 aminoácidos, o homeodomínio, responsável pela ligação da proteína na região reguladora de genes-alvo. Mutações em genes da família *Hox* podem ser homeóticas, ou seja, podem causar

mudanças na estrutura dos segmentos do corpo. Esses estudos mostraram pela primeira vez como o programa genético instrui para formação do plano corporal de uma espécie (Figura 11.2). Além disso, investigações com os genes *Hox* iniciaram a moderna biologia evolutiva do desenvolvimento, como veremos a seguir.

No final da década de 1940, Edward Lewis começou a estudar mutantes homeóticos a partir de cruzamentos controlados de drosófilas com esse tipo de mutações. Lewis formulou uma teoria de como genes homeóticos regulariam genes relacionados ao desenvolvimento e vislumbrou a possibilidade de que mutações do tipo *bithorax* indicariam a existência de genes reguladores capazes de controlar a formação de macro-regiões do corpo. Ele iniciou, então, um programa de estudos nos quais obteve mutantes com diversas deleções que envolviam genes homeóticos de *Drosophila* e, após décadas de estudos, essas deleções possibilitaram que Lewis mapeasse um grupo de genes homeóticos nesses insetos (Lewis, 1978). Lewis recebeu um prêmio Nobel em 1995 por suas descobertas que haviam sido então amplamente confirmadas.

A partir dos estudos de Lewis e de outros tipos de mapeamento, que hoje chegam à resolução no nível de nucleotídeos, sabe-se que os genes *Hox* de *Drosophila* estão agrupados em dois complexos gênicos. O complexo *Antennapedia* é composto por cinco genes ligados próximos a mais um complexo, *Ultrabithorax*, formado por outros três genes (sendo um deles, o gene *Ultrabithorax*, responsável pela mutação *bithorax*, Figura 11.3). Esses genes são expressos simultaneamente no início da embriogênese em regiões parcialmente coincidentes ao longo do eixo ântero-posterior. A combinação da expressão desses genes gera um código de sinalização espacial que determina a formação de segmentos do corpo com identidades diferentes entre si (Figura 11.3). Curiosamente, a expressão espacial dos genes *Hox* reflete sua ordem linear nos cromossomos. Assim, genes adjacentes no cromossomo são expressos em regiões vizinhas do corpo.

Genes ortólogos foram isolados de outras espécies animais, como, por exemplo, de vertebrados. Os genes *Hox* de vertebrados apresentam uma conservação estrutural, ou seja, mantêm a mesma sequência linear no complexo em relação aos respectivos ortólogos das moscas. Técnicas de hibridação *in situ* revelaram que a conservação também é funcional, pois esses genes são expressos sequencialmente ao longo do eixo ântero-posterior (também guardando correspondência com a posição linear nos cromossomos) e especificam diferencialmente estruturas repetidas em série que guardam semelhança entre si, como as costelas e as vértebras.

Os fatores de transcrição *Hox* pertencem a uma grande família multigênica caracterizada por variações no homeodomínio entre diferentes subfamílias, presentes em todos os animais, incluindo-se aqueles que não possuem o corpo segmentado e até mesmo em plantas, demonstrando grande conservação evolutiva

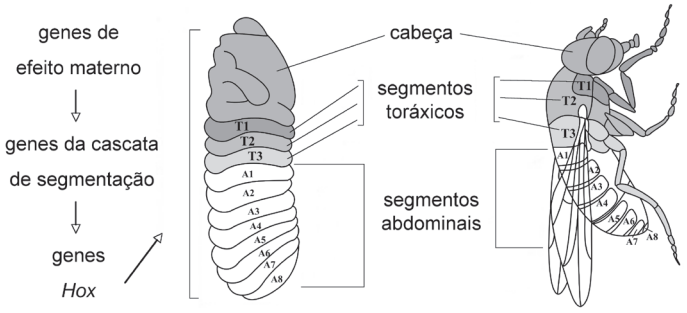


Figura 11.2. À esquerda, esquema da sequência temporal dos eventos moleculares (de cima para baixo) responsáveis pela especificação das divisões da gástrula (centro) em regiões correspondentes aos futuros segmentos do corpo do adulto (direita). Modificado de Gilbert, 1994.

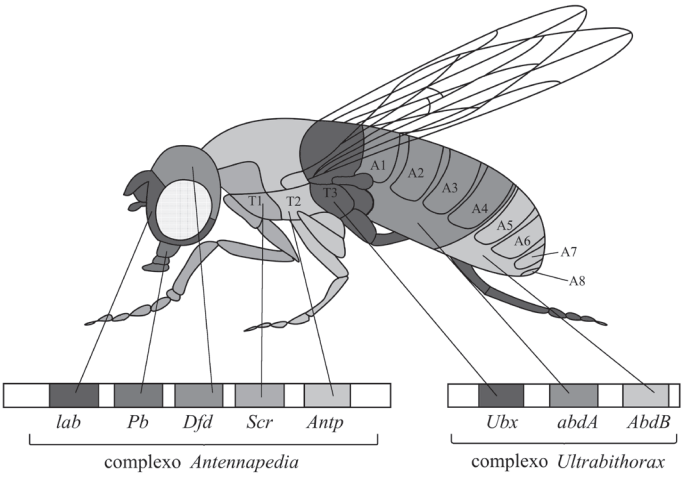


Figura 11.3. Genes *Hox* de *Drosophila*. Na parte inferior, estão representadas as posições relativas dos genes *Hox* no cromossomo 3 de *Drosophila*. Também estão indicadas as contribuições majoritárias de cada um dos genes *Hox* para a formação dos diferentes segmentos do corpo do adulto. Modificado de Gilbert, 1994.

(Holland e Garcia-Fernàndez, 1996). Com relação à subfamília *Hox*, especula-se que ao menos um gene já existisse no ancestral dos animais eumetazoários com base nas espécies atuais investigadas. Essa hipótese é consistente com a ausência de genes *Hox* nas esponjas e com a presença de dois desses genes nas hidras (Cnidaria) e em todos os outros táxons de animais triploblásticos—como os oito genes existentes em *Drosophila*, entre os artrópodes (Figura 11.4). Nos camundongos, assim como em humanos e demais tetrápodes analisados, estão presentes quatro complexos de genes *Hox*, cada qual contendo em torno de 13 genes. Portanto, esses dados demonstram a participação dos processos de duplicação e de divergência gênica dos genes *Hox* ao longo da filogenia dos animais e, em dado momento, a duplicação

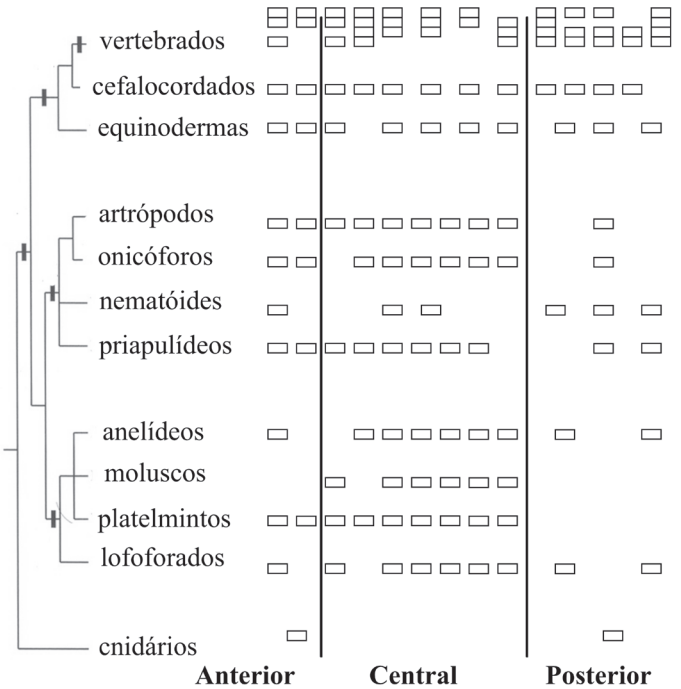


Figura 11.4. Evolução dos genes *Hox* nos metazoários. No cladograma, estão assinalados com pequenos traços os principais eventos de duplicação gênica ou ploidização, no caso dos vertebrados (notar os quatro complexos). Os genes *Hox* foram separados com barras, conforme expressão no eixo ântero-posterior do corpo. Notar que os cnidários possuem apenas dois genes, correspondentes ao grupo *Hox* anterior e posterior. Modificado de Carrol *et al.*, 2001.

de todo o complexo ao menos duas vezes para os vertebrados (Figura 11.4). A conservação estrutural e funcional dos genes *Hox* é de tal forma surpreendente que levou à sugestão de que essa característica seria um dos traços compartilhados por todos animais metazoários (Slack *et al.*, 1993).

11.4 Regulação da expressão gênica e modularidade

Conforme visto no Capítulo 10, as alterações numéricas das regiões codificadoras do genoma são consequências dos eventos de duplicação e divergência gênicas. O aumento do número de genes, por sua vez, é o principal mecanismo responsável pelo aumento da complexidade estrutural dos seres vivos (Carrol *et al.*, 2001). Define-se a complexidade estrutural como o grau da organização e a quantidade de partes, ao se compararem diferentes planos corporais (Valentine, 2003). Os diferentes tipos celulares e seus arranjos em tecidos e órgãos, por exemplo, podem ser usados como medida do grau de complexidade. No entanto, se o aumento de regiões codificadoras contempla o aumento da complexidade, certamente não é o principal mecanismo responsável pelo incremento da diversidade biológica (Carrol *et al.*, 2001). Os artrópodes, por exemplo, compõem o táxon animal de maior sucesso evolutivo baseado no número de espécies atuais. Esses animais são caracterizados pela presença de vários segmentos no corpo, de onde, no plano básico do grupo, partem pares de apêndices. A variação no número, tamanho e morfologia dos segmentos e de seus respectivos apêndices é responsável pela grande diversidade do grupo. Entretanto, os artrópodos compartilham o mesmo número de genes *Hox*, exemplos de genes reguladores chaves na formação dos planos corporais, conforme exposto no item anterior (Figura 11.4).

Pesquisas com artrópodes sinalizam que a diversidade morfológica do grupo possivelmente foi adquirida pela regulação da expressão gênica, especialmente no nível transcricional (Levine, 2002). Essa possibilidade parece confirmar-se pela natureza modular da região reguladora dos eucariotos (Figura 11.5, Small *et al.*, 1991; Small *et al.*, 1992). Esses módulos, denominados módulos reguladores em cis ou *enhancers*, são tipicamente compostos por poucas centenas de nucleotídeos, onde estão agrupados sítios de ligação (em média, sequências de 6 a 12 nucleotídeos), reconhecidos por fatores ativadores e repressores da transcrição. A proximidade de sítios de ligação possibilita a manifestação de propriedades importantes dos fatores de transcrição, como a cooperação entre moléculas de ativadores ou a competição entre ativadores e repressores (Figura 11.5, Cai *et al.*, 1996). Em suma, são as interações entre fatores de transcrição nos diferentes módulos reguladores de um gene que definem o padrão de expressão temporal e espacial da região codificadora.

O caso do gene *Ubx*, já mencionado, ilustra várias situações que correlacionam regulação da expressão gênica a alterações da morfologia. Por exemplo, em diferentes clados de crustáceos, *Ubx* determina o tipo de apêndice dos segmentos torácicos (Figura 11.6; Averof e Patel, 1997). Em crustáceos como os branquiópodes, o limite anterior de expressão de *Ubx* coincide com o limite anterior do primeiro segmento torácico. Nesses animais, os apêndices torácicos são locomotores. Em outros grupos, como nos isópodes, nota-se uma progressiva retração do domínio de expressão de *Ubx* em direção à região posterior. Nos isópodes e em outros grupos, na ausência de *Ubx* nos segmentos torácicos, são formados os maxilípedes, que são estruturas alimentares, e não locomotoras.

Outro exemplo é o tipo de apêndice formado no segundo segmento torácico dos insetos (Figura 11.6, Weatherbee *et al.*, 1998; Weatherbee *et al.*, 1999). Em lepidópteros, por exemplo, entre os insetos, no terceiro segmento torácico são formados um par de asas normais, a semelhança do par de asas anteriores formados no segundo segmento torácico, enquanto que, nos dípteros, no terceiro segmento é formado um par de asas vestigiais (halteres; Figura 11.2). Nesses dois grupos, *Ubx* é expresso, igualmente, nos segundo e terceiro segmentos torácicos. Entretanto, a proteína *Ubx* provavelmente encontra sítios ligadores apenas nas regiões reguladoras de genes-alvo que levam ao crescimento e achatamato da asa nos dípteros. Como *Ubx* é um repressor nos insetos, genes que levam à formação de asas normais são reprimidos nesses organismos. Nesses dois exemplos, é possível verificar como alterações nas regiões reguladoras podem desencadear alterações morfológicas drásticas, no primeiro caso pela alteração do padrão de expressão na região reguladora do gene efetor e, no segundo, por alterações da região reguladora de genes-alvo.

Os módulos das regiões reguladoras podem evoluir através de diferentes mecanismos moleculares: por duplicação e divergência de sequências, conforme já abordado, rearranjos envolvendo unidades funcionasis pré-existentes ou, teoricamen-

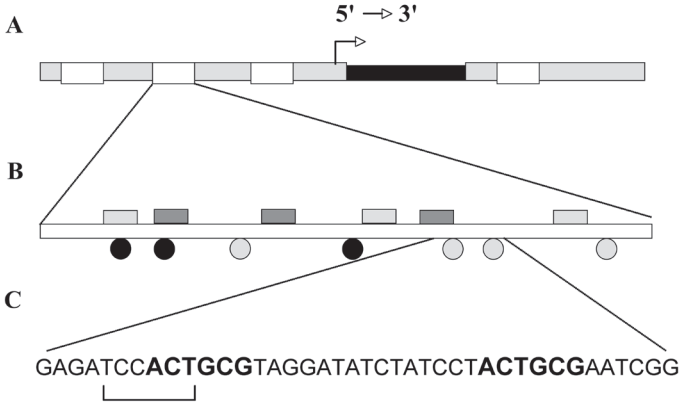


Figura 11.5. A. Esquema de um gene eucarioto contendo a região codificadora (barra escura) e a região reguladora (restante). Em destaque, as regiões cis-reguladoras (barras claras). B. Detalhe de uma região cis-reguladora, indicando a posição relativa de hipotéticos sítios ligadores de fatores ativadores (retângulos) e de fatores repressores da transcrição (círculos). C. Sequência de nucleotídeos de uma região cis-reguladora, indicando sítios hipotéticos reconhecidos por uma proteína repressora (em negrito) e de um sítio para uma proteína ativadora (colchete). Notar a coincidência parcial entre os sítios para as duas proteínas.

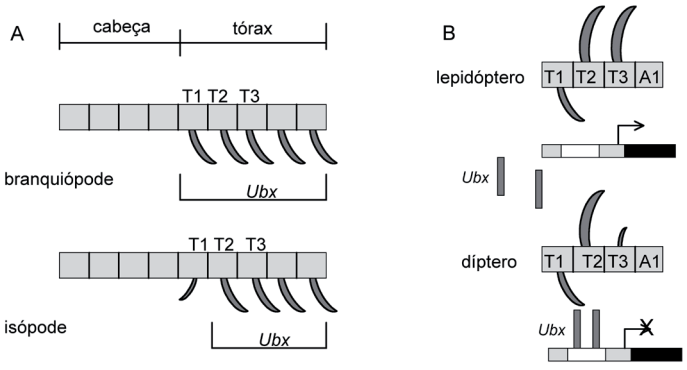


Figura 11.6. A. Correlação da expressão de *Ubx* nas regiões do corpo em dois grupos de crustáceos. Notar que, na ausência de *Ubx*, são formados maxilípedes no segmento T1 do isópode, e não mais uma estrutura locomotora, como no branquiópode. B. *Ubx* é expresso no segmento T3 de um lepidóptero e de um díptero. No entanto, a proteína repressora *Ubx* encontra sítios ligadores apenas nos genes responsáveis pelo desenvolvimento de asas normais nos dípteros, resultando na formação dos halteres. Modificado de Levine, 2002.

te, formação *de novo* a partir de DNA não funcional (Carrol *et al.*, 2001). As alterações podem representar ganhos ou perdas de sítios específicos, assim como de elementos inteiros. Acredita-se que as regiões não codificadoras dos genes estejam sujeitas a menores pressões seletivas que as regiões codificadoras, apresentando maior capacidade de gerar variação fenotípica herdável (*evolvability*, em inglês) (Ludwig *et al.*, 2005). Então, regiões reguladoras dos genes teriam uma capacidade facilitadora para a geração de variação fenotípica não letal a partir de mutações randômicas (Gerhart e Kirschner, 2003). No entanto, o estudo comparativo de genomas revela que regiões não codificadoras não são necessariamente mais sujeitas a alterações que regiões codificadoras (Kondrashov, 2005).

A natureza de módulo dos genes eucariotos é, na verdade, uma das facetas de uma propriedade mais genérica, que é a modularidade . Na biologia, estruturas anatômicas, genes, células, vias de sinalização ou fenômenos típicos do desenvolvimento podem ser considerados módulos (Gass e Bolker, 2003). Assim, os segmentos do corpo dos artrópodes podem ser considerados módulos, cada segmento sendo uma unidade funcional específica. Uma via de sinalização de apoptose também pode ser considerada um módulo, induzindo a morte celular em processos do desenvolvimento não relacionados, como remoção de tecido entre os dedos ou a destruição de células cancerosas (Penaloza *et al.*, 2006). Uma característica fundamental dos módulos é que eles não são unidades totalmente dependentes entre si, criando a possibilidade de que modificações em uma unidade não comprometam o todo. Assim, alterações nos módulos permitem o desacoplamento de um processo do desenvolvimento entre suas partes, criando uma perspectiva evolutiva.

Uma ressalva faz-se necessária. Apesar de as proteínas não serem consideradas as fontes primárias geradoras da diversidiade fenotípica, elas indubitavelmente contribuem com a variação morfológica. Nas sequências de aminoácidos de algumas proteínas, tais como os fatores de transcrição, estão presentes domínios e motivos que podem ser considerados módulos (Wagner e Lynch, 2008). Lembrando a possibilidade de regulação por *splicing* alternativo, a natureza modular das proteínas é ainda mais acentuada. De fato, experimentos com artrópodes confirmaram que a presença de domínios específicos na proteína *Ubx* pode ter sido responsável pela redução do número de apêndices nos segmentos do corpo dos artrópodes, quando comparados com os onicóforos, um grupo proximamente relacionado (Figura 11.5; Galant e Carrol, 2002). Os onicóforos possuem vários segmentos do corpo com apêndices. Esses segmentos expressam o gene *Ubx*, que atua como um fator ativador da transcrição nesses organismos. Ainda nesse estudo, os autores verificaram que o fato de *Ubx* ser um repressor constitutivo nos insetos deve-se à presença de um domínio regulador inexistente nos onicóforos. Uma situação intermediária é verificada nos crustáceos (Ronshaugen *et al.*, 2002). Nesses organismos, por exemplo em espécies de *Artemia*, o domínio repressor de *Ubx* é controlado pela presença de um segundo domínio na proteína, que possibilita a formação de apêndices em segmentos expressando *Ubx*.

11.5. Inovações e Homologias

A discussão de módulos na seção anterior, como os módulos anatômicos, pressupõe a existência de estruturas passíveis de alteração e surgimento de novas estruturas, ou seja, de inovações evolutivas. Essa percepção vai de encontro a uma definição clássica, intuitiva até, do que pode ser considerado "inovação", ou seja, o surgimento de uma nova estrutura ou propriedade que pressupõe a realização de uma nova função (Mayr, 1960). De acordo com essa definição, os maxilípedes formados nos segmentos toráxicos dos crustáceos são inovações em relação aos apêndices locomotores homólogos desses segmentos em espécies consideradas como relativamente mais "basais" no grupo. Entretanto, conforme já comentado, estudos moleculares recentes revelaram a existência de um mecanismo subjacente comum responsável pela formação de apêndices no corpo. Uma definição mais rigorosa consideraria, assim, uma "inovação morfológica" apenas as estruturas que não são homólogas a qualquer estrutura presente em um ancestral ou homônoma de qualquer estrutura presente no mesmo organismo (Muller e Wagner, 1991). Por homonomia, entende-se homologia seriada, ou seja, estruturas equivalentes que se repetem, como é o caso dos apêndices e dos segmentos do corpo dos artrópodos. Nesse caso, os maxilípedes não constituem uma inovação, mas uma modificação de uma inovação pré-existente. Um exemplo de inovação que passaria pelo crivo dessa definição é a presença de órgãos bioluminescentes de espécies de vaga-lumes emissores de luz. Embora a origem ontogenética desse órgão ainda seja pouco compreendida, essa característica restrita a uma família de besouros parece ser um exemplo incontestável de inovação (Moczek, 2008).

A definição mais rigorosa de inovação esbarra no conceito fundamental de homologia, um conceito bastante complexo em si mesmo. Um complicador tem sido a constatação, por meio dos estudos de genética molecular do desenvolvimento, de que caracteres às vezes considerados homólogos têm bases genéticas e moleculares diferentes (Wagner, 2007). Essa constatação é aparentemente paradoxal, pois se presume que a continuidade histórica de um caracter morfológico implica em sua continuidade genética. Por exemplo, o segundo molar presente na maioria dos carnívoros não está presente no ancestral dos felídeos, assim como na maioria das espécies atuais dessa família, com a exceção de uma única espécie, o lince (Raff, 1996). De acordo com o conceito filogenético de homologia, duas características são homólogas quando herdadas com maior ou menor grau de modificação a partir de um ancestral comum onde essa característica surgiu (Wagner, 1989). A interpretação literal, no caso do segundo molar dos linces, é que essa característica não constitui uma homologia entre os vários grupos que o apresentam. No entanto, de acordo com o conhecimento atual de genética do desenvolvimento, o reaparecimento do segundo molar nos linces possivelmente significa a existência de um mecanismo molecular do desenvolvimento latente, que estaria presente nos ancestrais e nos atuais felídeos, portanto caracterizando uma homologia subjacente.

Outra situação conhecida que tem sido reinterpretada é a formação de estruturas fotorreceptoras (Treisman, 2004). Nos animais, existem estruturas de fotorrecepção muito simples, formadas por um conjunto de células pigmentadas, como nos cnidários, até estrutras complexas, formadoras de imagens, como o olho humano, passando por várias formas de complexidade intermediárias, capazes de captar estímulos luminosos. Além da complexidade diferente dos vários tipos de estruturas fotorreceptoras, há variação fundamental na organização dessas estruturas, como, por exemplo, nos olhos compostos em insetos, formados por centenas de omatídeos, cada qual com uma lente própria. Um fato aparentemente surpreendente é que investigações em espécies com estruturas fotorreceptoras com os mais diferentes graus de complexidade e organização têm revelado a existência do mesmo tipo de molécula envolvida no início da formação dessas estruturas (*Pax 6*/ *eyeless*). Em *Drosophila*, a expressão ectópica de *eyeless* é capaz de induzir a formação de olhos compostos em outras regiões do corpo (Halder *et al.*, 1995), confirmando a natureza reguladora desse gene na iniciação do processo de formação

do olho. A expressão do ortólogo humano *Pax 6* em *Drosophila* também induz a expressão ectópica do olho, confirmando a conservação funcional dessa molécula (Xu *et al.*, 1997). Desse modo, uma situação por muito tempo interpretada como convergência, ou seja, a evolução independente de estruturas fotorreceptoras em vários grupos animais, agora é interpretada como resultado de evolução paralela baseada na história compartilhada de um mecanismo gerador de estruturas fotorreceptoras, estabelecido cedo na evolução animal, revelando uma homologia profunda (Shubin *et al.*, 2009). Constatações desse tipo (como também no caso do aparecimento do segundo molar nos linces) levaram à formulação do conceito de homologia biológica, que postula que a homologia verificada em um determinado nível organizacional não necessariamente está presente em outro nível (Wake, 2003). Por outro lado, ao aceitarmos de forma incontestável o conceito de homologia biológica, podemos mascarar um contínuo esperado da descendência com modificação por falta de maiores conhecimentos dos mecanismos evolutivos do desenvolvimento.

11.6. Conclusões

No início deste capítulo, foi mencionada a separação progressiva entre a biologia evolutiva e a biologia do desenvolvimento a partir do final do século XIX. Essa separação acirrou-se com a concretização da Síntese Evolutiva, entre as décadas de 1930 e 1940, que estabeleceu uma diretriz de pesquisas na biologia evolutiva centrada na genética de populações. Desde então, a diferença de abordagem entre a biologia evolutiva e a biologia do desenvolvimento tem dificultado a reaproximação entre as áreas. O modelo de genética de populações tem como uma de suas premissas a identificação de características genéticas nos adultos, responsáveis pelo sucesso reprodutivo. Por sua vez, a biologia do desenvolvimento prioriza os processos de regulação gênica nos embriões e nas larvas. A genética de populações foca variações intra-específicas, enquanto que a biologia do desenvolvimento foca variações inter-específicas. Contudo, é possível e desejável a convergência entre as áreas.

Um exemplo que ilustra essa possibilidade foi realizado com o peixe *Gasterosteus aculeatus* (Shapiro *et al.*, 2004). Essa espécie apresenta populações colonizadoras de lagos de água doce que evoluíram recentemente a partir de uma forma marinha. Apesar da grande semelhança morfológica, a redução ou inexistência do espinho na região pélvica das populações de água doce é uma diferença evidente. Os espinhos são supostamente estruturas de defesa contra peixes predadores marinhos. Por outro lado, as populações de água doce não estão sujeitas aos predadores marinhos e a presença de espinho na região pélvica, especula-se, favoreceria a captura dos peixes por parte de invertebrados presentes neste meio. Uma abordagem clássica de genética de populações possibilitou o isolamento de um gene regulador, o fator de transcrição *Pitx1*, cujo ortólogo em mamíferos é responsável pelo desenvolvimento de seus membros posteriores. Ao longo do desenvolvimento, o padrão de expressão de *Pitx1* verificado entre peixes marinhos e de água doce é o mesmo, com exceção da expressão na região pélvica, onde são formados os espinhos, ausente nas populações de água doce. Como a sequência de aminoácidos da proteína *Pitx1* na espécie é invariável, levantou-se a hipótese de que alterações de um *enhancer* responsável pela formação do espinho na região pélvica o tenha tornado inoperante nas formas de água doce. Essa hipótese foi confirmada em estudo onde o *enhancer* funcional responsável pela expressão do gene *Pitx1* em peixes marinhos foi isolado, enquanto, das populações de água doce, foram isolados apenas este *enhancer* não funcional, por apresentar mutações em sua sequência (Chan *et al.*, 2010). Esse exemplo mostra que a utilização de abordagens complementares focando genes reguladores que controlam o desenvolvimento pode convergir para uma melhor compreensão do processo evolutivo.

Referências Bibliográficas

Averof, M. e Patel, N. (1997). Crustacean appendage evolution associated with changes in *Hox* gene expression. **Nature 388**:682-686.

Cai, H. N., Arnosti, D. N., e Levine, M. (1996). Long-range repression in the *Drosophila* embryo. **Proc. Natl. Acad. Sci. USA 18**:9309-9314.

Carroll, S. B., Grenier, J. K., Weatherbee, S. D. (2001). From DNA to diversity. Blackwell Science.

Chan Y. F., Marks, M. E., Jones, F. C., Villarreal G., Shapiro, M. D., Brady, S. D., Southwick, A. M., Absher, D. M., Grimwood, J., Schmutz, J., Myers, R. M., Petrov, D., Jónsson, B., Schluter, D., Bell, M. A., Kingsley, D. M. (2010). Adaptative evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. **Science 327**:302-305.

Galant, R. e Carroll, S.B. (2002). Evolution of a transcriptional repression domain in an insect Hox protein. **Nature 415**:910-913.

Gass, G. L. e Bolker, J. A. (2003). Modularity. **In** Olson, H. and Olson, W. (eds) **Keywords & Concepts in Evolutionary Developmental Biology**, Harvard University Press, Cambridge, Massachusetts, USA.

Gerhart, J. e Kirschner, M. (2003). Evolvability. **In** Olson, H. and Olson, W. (eds) **Keywords & Concepts in Evolutionary Developmental Biology**, Harvard University Press, Cambridge, Massachusetts, USA.

Goldschmidt, R. (1940 The material basis of evolution. Yale University Press, New Haven.

Halder, G., Callaerts, P. Gehring, W. J. (1995). Induction of ectopic eyes by target expression of the *eyeless* gene in *Drosophila*. **Science 267**:1788-1792.

Holland, P. W. H. e Garcia-Fernàndez, J. (1996). *Hox* genes and chordate evolution. **Dev. Biol. 173**:382-395.

Kondrashov, A. S. (2005). Evolutionary biology: fruitfly genome is not junk. **Nature 7062**:1106.

Levine, M. (2002). How insects lose their limbs. **Nature 415**:848-849.

Lewis, E.B. (1978). A gene complex controlling segmentation in *Drosophila*. **Nature 276**:565-570.

Ludwig, M. Z., Palsson, A., Alekseeva, E., Bergaman, C.M., Nathan, J. e Kreitman, M. (2005). Functional evolution of a cis-regulatory module. **Plos Biology 3**:e93.

Mayr, E. (1960). The emergence of evolutionary novelties. **In** Tax, S. (ed) **Evolution after Darwin**. Chicago, University of Chicago Press.

Moczek, A.P. (2008). On the origins of novelty in development and evolution. **BioEssays 30**:432-447.

Muller, G.B. e Wagner, G.P. (1991). Novelty in evolution: restructuring the concept. **Ann. Rev. Ecol. Syst. 22**:229-256.

Penaloza, C., Lin, L., Lockshin, R.A. e Zakeri, Z. (2006). Cell death in development: shaping the embryo. **Histochem. Cell Biol. 126**:149-158.

Raff, R. A. (1996). The shape of life. The University of Chicago Press.

Ronshaugen, M., McGinnis, N. e McGinnis, W. (2002). *Hox* protein mutation and macroevolution of the insect body plan. **Nature 415**:914-917.

Shapiro, M.D., Marks, M.E., Peichel, C.L. Blackman, B.K., Nereng, K.S. Jónson, B., Schluter, D. e Kingsley, D.M. (2004). Genetic and developmental basis of evolutionary pelvic reduction in three-spine sticklebacks. **Nature 428**:713-723.

Shubin, N., Tabin, C. e Carroll, S. (2009). Deep homology and the origens of evolutionary novelty. **Nature 457**:818-823.

Slack, J. M. W., Holland, P. W. H. e Graham, C. F. (1993). The zootype and the phylotypic stage. **Nature 361**:490-492.

Slack, J.M.W. (2002). Conrad Hal Waddington: the last Renaissance biologist? **Nature Rev. Genet. 3**:889-895.

Small S., Kraut, R., Warrior, R. e Levine, M. (1991). Transcriptional regulation of a pair-rule stripe in *Drosophila*. **Genes & Development. 5:**827-839.

Small S., Blair, A., Levine, M. (1992). Regulation of *even-skipped* stripe 2 in the *Drosophila* embryo. **EMBO J. 11**:4047-4057.

Treisman, J. (2004). How to make an eye. **Development 131**:3823- 3827.

Valentine, J. W. (2003). Cell Types, Numbers and Body Plan Complexity **In In** Olson, H. and Olson, W. (eds) **Keywords & Concepts in Evolutionary Developmental Biology**, Harvard University Press, Cambridge, Massachusetts, USA.

Wagner, G.P. (1989). The origin of morphological characters and the biological basis of homology**. Evolution 43**:1157-1171.

Wagner, G.P. (2007). The developmental genetics of homology. **Nature**

Review of Genetics 8:473-479.

Wagner, G.P. e Lynch, V.J. (2008). The gene regulatory logic of transcription factor evolution. **Trends in Ecology and Evolution 7**:377-385.

Wake, D.B, (2003). Homology and Homoplasy. **In** Olson, H. and Olson, W. (eds) **Keywords & Concepts in Evolutionary Developmental Biology**, Harvard University Press, Cambridge, Massachusetts, USA.

Weatherbee, S.D., Halder, G., Kim, J., Hudson, A., Carrol, S. (1998). *Ultrabithorax* regulates genes at several levels of the wing-patterning hierarchy to shape the development. **Genes & Development 12**:1474-1482.

Weatherbee, S.D., Nijhout, H. F., Grunert, L. W., Halder, G., Selegue, J. e Carroll, S. (1999). *Ultrabithorax* function in butterfly wings and the evolution of insect wing patterns. **Curr. Biol. 9**:109-115.

Xu, Z.P., Woo, I., Her, H., Beier e Maas, R. L. (1997). Mouse *Eya* homologues of the *Drosophila eyes absent* gene require *Pax 6* for expression in lens and nasal placode. **Development 124**: 219-231.

Página deixada em branco

Capítulo 12

Reconstrução filogenética. Introdução e o método da máxima parcimônia

Cristina Yumi Miyaki (cymiyaki@ib.usp.br)
Departamento de Genética e Biologia Evolutiva
Instituto de Biociências
Universidade de São Paulo

Cláudia A. de Moraes Russo (claudia@biologia.ufrj.br)
Departamento de Genética
Instituto de Biologia
Universidade Federal do Rio de Janeiro

Dr. Sergio Luiz Pereira (sergiolp@gmail.com)
The Centre for Applied Genomics
The Hospital for Sick Children

"*A classificação por descendência não pode ser inventada por biólogos, ela pode apenas ser descoberta.*" (Theodosius Dobzhansky)

12.1. Hierarquias dos Sistemas Biológicos e a Classificação dos Organismos

Sistemas de classificação de organismos têm sido propostos desde a Antiguidade. Existem diferentes critérios que podem dar base a um sistema de classificação para quaisquer conjuntos de entidades. A princípio, toda tentativa de classificar objetos ou organismos tem por objetivo agilizar e padronizar a comunicação entre as pessoas a respeito dos objetos ou organismos classificados. Entretanto, classificações utilizando diferentes critérios de modo geral são conflitantes entre si e comprometem uma comunicação clara e objetiva entre os usuários do sistema. Por exemplo, na classificação de livros em uma estante, uma pessoa que classifique de acordo com a cor das capas não conseguirá estabelecer um sistema compatível com a coleção de outra pessoa que classifique os livros de acordo com o tamanho. Portanto, para a construção de um sistema universal de classificação de um determinado conjunto de objetos, é fundamental que haja um único critério organizador do sistema.

A classificação dos organismos também lida com essa mesma questão. A questão central, no entanto, é que critério escolher. A aceitação de uma ontologia evolutiva, faz com que a natureza das espécies, com suas relações de parentesco entre si, gera um critério natural para uma classificação hierárquica, consistente e única dos organismos. Um sistema de classificação baseado nesses princípios deveria, portanto, refletir os eventos que ocorreram durante a história da vida em nosso planeta. Aparentemente, o embrião do pensamento evolutivo nos naturalistas pré-darwinianos originou-se justamente da organização hierárquica da natureza biológica, que aflorava nos sistemas de classificações, em especial naquele apresentado por Lineu. Em outras palavras, o padrão hierárquico da natureza biológica parece ter ajudado a percepção de que o processo evolutivo tem a mesma estrutura.

Cerca de 100 anos após a proposição de Darwin, o entomólogo alemão Willi Hennig criou um método de reconstrução das relações de parentesco entre grupos de organismos, bem como uma escola de classificação em que se advoga que o melhor sistema de classificação biológica é aquele que reflete a filogenia—em seu conjunto chamada "sistemática filogenética" (Hennig, 1950). Segundo Hennig, o conjunto de organismos que compartilham a condição derivada de uma determinada característica poderia ser hipotetizado como sendo descendente da espécie ancestral na qual a condição primitiva ou pré-existente (ou plesiomórfica) dessa característica passou, por mutação, à condição derivada (ou apomórfica). Posteriormente, foi proposto um sistema, desenvolvido pelo próprio Hennig, que permitisse que a classificação dos organismos refletisse de modo inequívoco as hipóteses de relações de parentesco entre os grupos classificados.

Um exemplo poderá clarear a importância da sistemática filogenética. Vamos supor que um determinado pesquisador decida propor um grupo chamado "Verdata". Neste grupo, o pesquisador inclui todos os organismos verdes: algas verdes, plantas terrestres, pererecas, algumas cobras, invertebrados marinhos, entre outros. O grupo Verdata, naturalmente, não satisfaz as condições da sistemática filogenética de Hennig, pois não é um grupo natural, unindo espécies que compartilham um ancestral comum exclusivo. Dessa forma, nada mais sabemos sobre as espécies que compõem este grupo, além do fato de as espécies de Verdata são verdes. O poder de previsibilidade é nulo, neste caso. Mais que isso, para um leitor desatento, inferências evolutivas extremamente equivocadas podem ser feitas com base na suposição de que esse grupo, de fato, exista na natureza como uma entidade real.

Por outro lado, tomemos como exemplo a classe Mammalia, um grupo natural, possuidor de um ancestral comum exclusivo. Repare que nesse ancestral surgiram as características hoje utilizadas como diagnósticas da classe: pêlos, glândulas mamárias, dentes diferenciados. Tais características estavam presentes na espécie ancestral comum de todos os mamíferos, ou seja, presentes na espécie que deu origem a toda a diversidade do grupo.

Na realidade, não apenas essas características que chamamos de diagnósticas, mas todas as modificações surgidas (centenas ou milhares) na evolução dessa primeira espécie tam-

bém foram passadas para as primeiras linhagens descendentes, e portanto, potencialmente, estariam em todas as espécies de mamíferos (ou em condições delas derivadas). Repare que, com esse detalhe, sabemos que todas as espécies de mamíferos são mais aparentadas entre si do que com espécies de qualquer outro grupo. Isso vale tanto para características óbvias, quanto para características menos óbvias—como o ponto isoelétrico (p.I.) de uma enzima—ou para qualquer característica herdável ainda a ser descoberta. Assim, o uso de uma classificação com fundamentação filogenética, com grupos naturais nomeados, fornece um poder de previsão impressionante sobre a distribuição de todas as características herdáveis em toda a diversidade biológica. Esse é um dos aspectos mais poderosos de um sistema filogenético.

A sistemática filogenética foi pouco difundida no mundo científico na sua primeira década de existência e ainda em um contexto que a sistemática estava separada da genética de populações. Os conceitos originais de Hennig, publicados em alemão em 1950, permaneceram pouco conhecidos até a publicação, em inglês, em 1966, de um novo livro que ampliava muito sua apresentação original. Os trabalhos de reconstrução filogenética passaram, então, a contar com uma base metodológica bem definida. Desde então, uma série de adições foram feitas à estrutura conceitual e metodológica original de inferência filogenética proposta por Hennig, gerando um grande aperfeiçoamento nesse campo.

O papel da sistemática filogenética, portanto, ademais de realizar o trabalho tradicional da taxonomia, de descrever a diversidade biológica, também é de organizar o conhecimento sobre essa diversidade com base no conhecimento das relações de parentesco entre os grupos e do conhecimento da evolução das características morfológicas, comportamentais, ecológicas, fisiológicas, citogenéticas e moleculares dos grupos. Dessa forma, a sistemática e seus métodos de inferência passaram a estar associados a várias outras disciplinas, como zoologia, botânica, ecologia e genética, analisando e interpretando os padrões e processos evolutivos. De fato, a questão da análise filogenética deixou de ser uma questão restrita à taxonomia e passou a ser organizadora de todo o conhecimento biológico, ou seja, das características geradas em todos os campos da biologia. Os resultados dos estudos sistemáticos são representados graficamente na forma de filogenias ou árvores filogenéticas, indicando a relação histórica entre os organismos.

Esta introdução pretende fornecer conceitos básicos relacionados aos métodos de reconstrução filogenética discutidos neste capítulo, bem como nos Capítulos 13 e 14. Para uma visão mais detalhada da sistemática filogenética, sugerimos a consulta de Amorim (2002).

12.2. Reconstrução Filogenética

O conceito de filogenia surgiu com Darwin e é um corolário do próprio conceito de que as espécies se originaram de espécies ancestrais. É de sua autoria o primeiro diagrama publicado representando relações filogenéticas. Assim, as filogenias como diagramas nada mais são que gráficos indicando hipóteses de relações de ancestralidade para um conjunto de espécies. Na prática, o problema fundamental é que as espécies ancestrais existiram no passado da evolução dos organismos e, portanto, não podem ser observadas diretamente. Assim, é necessário buscar mecanismos para, analisando os organismos atuais (ou fósseis de que dispomos hoje), recuperar da informação a respeito das relações de parentesco entre os grupos.

Uma árvore filogenética representa a história evolutiva dos organismos nela incluídos. Graficamente, consiste de pontos (ou nós) ligados por linhas (ou ramos). As terminações de uma filogenia podem ser táxons colocados ao nível de família (ou grupos em níveis ainda mais altos), gênero, espécies ou populações, ou seja, táxons de qualquer nível na hierarquia sobre os quais se esteja inferindo a história evolutiva. Com frequência, as terminações da filogenia são chamadas "unidades taxônomicas operacionais", ou simplesmente OTUs (do inglês, *operational taxonomic units*). Árvores filogenéticas têm dois tipos de nós, os internos e os terminais. Os nós terminais representam as OTUs estudadas, unidas por ramos cujo nó interno representa o ancestral comum mais recente desses táxons (Figura 12.1). Em outras palavras, a topologia (ou estrutura particular) de uma árvore filogenética pode ser definida como uma representação gráfica unindo as OTUs através de ramos e nós. As árvores filogenéticas podem ser estritamente dicotômicas—onde a cada nó interno chega um único ramo (que vem da espécie ancestral) e do qual partem dois ramos (que vão para os descendentes)—ou podem conter politomias—em que, além do ramo que liga à ancestral, há três ou mais ramos descendentes. Politomias não necessariamente representam especiações politômicas. Elas podem indicar que a hipótese para aquele nível naquela filogenia não é capaz de discriminar, por falta de informação, entre três ou mais táxons, quais pares estão mais aparentadas entre si. Assim, uma politomia pode, ulteriormente, com um novo aporte de informação, ser melhor resolvida e resultar em duas ou mais dicotomias, ou, de fato, corresponder a uma especiação politômica.

Uma árvore filogenética pode ser representada de maneira não enraizada, ou seja, sem apontar onde está a espécie ancestral de todo o grupo (a "raiz" da filogenia). Nesse caso, o diagrama não tem uma estrutura temporal e, portanto, não indica precisamente quais são as relações de parentesco dentro do grupo. A Figura 12.2 corresponde à árvore não enraizada da filogenia dos Tetrapoda, da Figura 12.1. Nesse caso, a árvore não enraizada mostra apenas as relações de topologia ou de proximidade relativa entre as OTUs a partir de semelhanças compartilhadas, sem, no entanto, indicar a posição em que se encaixa a espécie ancestral de todo o grupo. A

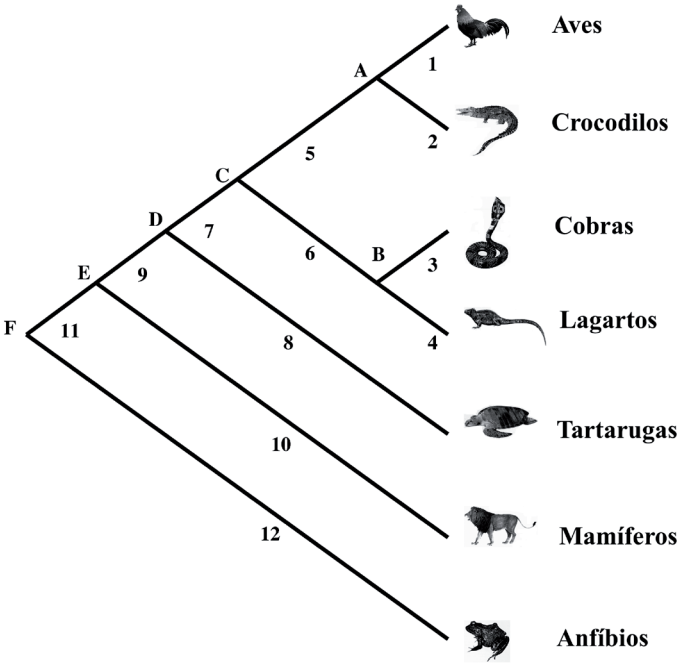


Figura 12.1. Representação de uma árvore dicotômica enraizada. Aves, crocodilos, cobras, lagartos, tartarugas, mamífeos e anfíbios são as OTUs representadas nos nós terminais. Essas OTUs são unidas aos seus ancestrais (representados pelos nós internos A, B, C, D, E e F) através de ramos (1-12).

presença de uma raiz, portanto, identifica o sentido da evolução, ou seja, a raiz dá o sentido temporal para a árvore. No caso da árvore não enraizada da Figura 12.2, se uma raiz fosse inserida no ramo mais longo, os Anfíbios passariam a ser interpretados erroneamente como grupo-irmão dos Mamíferos. Isso indica que se deve ter bastante cuidado na interpretação de árvores não enraizadas.

O número de árvores dicotômicas alternativas possíveis em um grupo varia com o número de OTUs e se a árvore é enraizada (*N*) ou não enraizada (*N**). Ou seja, há um número de soluções alternativas possíveis e distintas entre si de relações entre um determinado conjunto de táxons terminais, cada uma das quais correspondendo a uma hipótese diferente para a evolução do grupo. O número exato de possibilidades de relações (sem raiz) entre OTUs é dado pelas fórmulas:

$$N_u = \frac{(2t-5)!}{2^{(t-3)}.(t-3)!}$$

onde t corresponde ao número de táxons (Felsenstein, 1978). Conforme aumenta o número de OTUs, maior o número de possíveis árvores dicotômicas (Tabela 12.1)—e maior o tempo computacional para estimar as relações entre elas. Na realidade, o tempo computacional aumenta exponencialmente com o acréscimo de OTUs. Repare que, a partir de três OTUs, o número de árvores enraizadas é igual ao número de árvores não enraizadas com menos uma OTU.

Uma árvore, como a mostrada graficamente na Figura 12.1, indica que as Aves e os Crocodilos são grupos-irmãos, pois apresentam um ancestral comum mais próximo entre si que qualquer um desses dois grupos com os demais clados terminais, assim como o são Serpentes e Lagartos. Por sua vez, o clado "(Aves, Crocodilia)" é grupo-irmão de "(Serpentes, Lacertilia)" e assim por diante. Essa árvore também pode ter essa estrutura de hipóteses de parentesco representada linearmente: (Amphibia, (Mammalia, (Chelonia, ((Serpentes, Lacertilia), (Aves, Crocodilia))))). Ou seja, OTUs irmãs são representadas entre parênteses e separadas por vírgula. Outra maneira de representar a relação de grupos-irmãos é utilizar um "+", de maneira que essa mesma filogenia seria indicada como (Amphibia + (Mammalia + (Chelonia + ((Serpentes + Lacertilia), (Aves + Crocodilia))))).

Outro ponto importante da reconstrução filogenética é o conceito de monofilia. Grupos monofiléticos são agrupamentos de organismos que incluem um ancestral comum exclusivo e todos seus descendentes, ou seja, esse ancestral não é ancestral de nenhum outro membro externo ao grupo. Dito de outra maneira, partes dos descendentes dessa espécie ancestral não foram

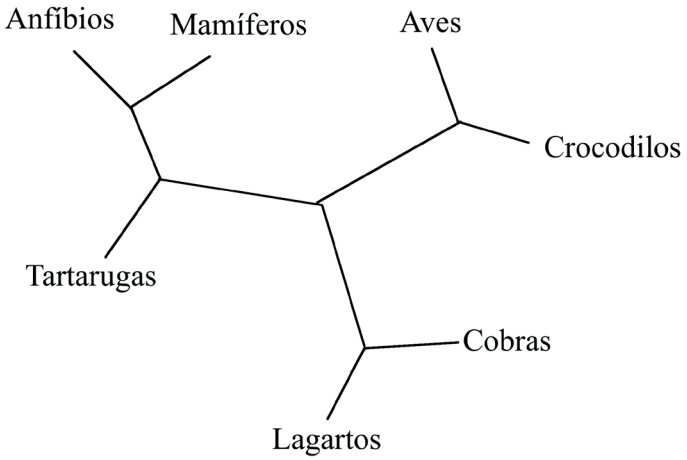


Figura 12.2. Árvore não enraizada com os mesmos táxons da Figura 12.1. Para tornar a árvore enraizada, basta escolher qualquer dos ramos e atribuir a raiz a esse ramo.

Tabela 12.1. Número de possíveis árvores dicotômicas, enraizadas e não enraizadas, para *n* OTUs.

Número de OTUs	Árvores enraizadas	Árvores não enraizadas
2	1	1
3	3	1
4	15	3
5	105	15
6	945	105
7	10.395	945
8	135.135	10.395
9	2.027.025	135.135
10	34.459.425	2.027.025

deslocadas para outros grupos taxonômicos. Muitas vezes, a monofilia das OTUs em uma análise é assumida implicitamente. Por exemplo, quando se infere as relações filogenéticas entre os gêneros dentro de uma mesma família usando apenas uma espécie representando cada gênero, as conclusões sobre parentesco referem-se apenas à relação entre os gêneros, mas não permitem inferir que cada gênero seja ou não monofilético.

Na Figura 12.1, por exemplo, podemos definir um grupo monofilético denominado Amniota, composto pelos clados dos mamíferos, tartarugas, lagartos, serpentes, crocodilos e aves. A espécie ancestral comum mais recente desse grupo é representada pelo nó interno **E**. Já o tradicional grupo "Reptilia" (que inclui apenas tartarugas, lagartos, serpentes e crocodilos) não constitui um grupo monofilético, uma vez que um dos descendentes da espécie ancestral de todo esse grupo (as aves) foi um grupo excluído dos répteis. O clado formado por crocodilos e aves, por sua vez, pode ser considerado monofilético e apresenta um ancestral representado pelo nó interno **A**. Esse grupo é denominado Archosauria.

A monofilia é um conceito crucial no contexto do estabelecimento de relações de parentesco e da compreensão da evolução de grupos em qualquer nível. Grupos não monofiléticos são denominados parafiléticos ou polifiléticos (Hennig, 1966). Um grupo é parafilético quando inclui uma espécie ancestral comum e apenas parte de seus descendentes, ou seja, uma ou mais das espécies descendentes dessa espécie ancestral foi excluída desse grupo (como o grupo dos Reptilia, de onde as Aves foram excluídas). Um grupo é polifilético quando inclui dois ou mais grupos monofiléticos sem que sua espécie ancestral faça parte do grupo, como seria o caso dos "vertebrados homeotérmicos", incluindo apenas Aves e Mammalia. Segundo Bernardi (1981), grupos não monofiléticos deveriam ser chamados de merofiléticos (*mero*, do grego, parte ou porção), uma vez que esse termo não provoca a confusão dos conceitos de parafilia e polifilia.

Outro conceito que a escola filogenética introduziu na sistemática é o conceito de grupos externos (Maddison *et al.*, 1984). Plesiomorfia e apomorfia são conceitos que indicam a idade relativa, em uma determinada série de transformação, de condições ou estruturas homólogas. Para afirmar qual das condições encontradas em um grupo é a apomórfica e qual é a plesiomórfica, é necessário saber determinar o eixo temporal, ou seja, determinar qual é a mais antiga. A compreensão da direção da mudança de um caráter em um grupo ou seja, seu eixo temporal, não pode ser determinada apenas considerando a variação dentro do grupo. Se o grupo no qual essa característica varia for monofilético, a condição mais antiga (plesiomórfica), de fato, precisa ter sido herdada de níveis ainda mais inferiores na filogenia. Assim, grupos externos a um determinado grupo monofilético devem apresentar essa condição plesiomórfica ou outra ainda mais plesiomórfica. Ou seja, é a comparação com grupos externos que acrescentam o sentido do tempo na análise de um grupo, obtido pela indicação da raiz da árvore

filogenética, que corresponde ao ancestral comum mais recente de todos os membros do grupo interno (Graur e Li, 2000). Em análises envolvendo o estabelecimento de relações de parentesco entre populações, espécies ou grupos supraespecíficos, o uso de grupos externos não é trivial. Eles devem ser escolhidos cuidadosamente, pois não podem ser testados como tal na mesma análise filogenética. Os grupos externos devem sempre ser assumidos com base em análises anteriores que indiquem que são OTUs próximas, mas que divergiram necessariamente antes do processo de diferenciação das OTUs do próprio grupo interno. Por exemplo, na Figura 12.1, os Anfíbios podem ser considerados um grupo externo em relação ao clado cujo ancestral comum mais recente é **E**.

O uso de informação sobre grupos externos permite, desse modo, estabelecer o sentido das modificações evolutivas. Dessa forma, os grupos externos indicam os estados primitivos ou derivados de um determinado caráter, isto é, ele permite a polarização (ou seja, a determinação dos "pólos" mais antigos ou recentes) dos estados do caráter. Na verdade, inferências sobre a história filogenética dos grupos (isto é, sobre relações de ancestralidade) só podem ser feitas quando é possível verificar o compartilhamento de características derivadas ou apomórficas por um conjunto de espécies, indicando a existência de uma espécie ancestral de todo um grupo na qual as mutações correspondentes surgiram.

A polarização dos estados de um caráter não é absoluta, mas depende do grupo taxonômico analisado. Por exemplo, a presença de pêlos é um estado derivado compartilhado pelos mamíferos dentro do grupo dos vertebrados. Por outro lado, a presença de pêlos é um estado ancestral dentro dos mamíferos e não pode ser usada como caráter para estabelecer relações filogenéticas entre as diferentes ordens de mamíferos. Como a polaridade dos caracteres pode ser influenciada pelos grupos externos escolhidos, alguns autores sugerem a inclusão de vários grupos externos em estudos filogenéticos, de modo a evitar que paralelismos em um dos grupos externos resultem em interpretações equivocadas sobre a evolução de caracteres no grupo interno. No caso de árvores não enraizadas, a polarização dos caracteres é feita *a posteriori*.

É necessário notar que o conceito de série de transformação não se aplica apenas à mudança entre dois estados, de uma condição original (plesiomórfica) para uma condição modificada (apomórfica). Ela se refere a qualquer sequência conhecida de modificação de uma condição original em uma ou mais condições derivadas. Algumas séries são restritas a apenas duas condições conhecidas e, nesse caso, a modificação é necessariamente linear. Há séries que conectam a condição original a duas ou mais condições derivadas conhecidas e, nesses casos, as condições derivadas podem ser sucessivamente apomórficas em uma sequência linear ou corresponder a condições que se originaram independentemente a partir da mesma condição plesiomórfica original. Por exemplo, o caráter "modo de reprodução" pode ter os estados "ovíparo", "ovovivíparo" ou "vivíparo". Nesse caso, podemos assumir que ovovivíparo é necessariamente um estado intermediário entre os outros dois estados. A análise dos dados de um grupo pode ser feita considerando que todas as sequências entre as condições conhecidas são necessariamente lineares, necessariamente ramificadas ou não estabelecer nenhuma hipótese a priori de ordenação (as análises "não ordenadas").

Finalmente, quanto à classificação dos grupos, observamos que diferentes critérios de construção de classificação coexistem, com variações quanto ao uso de categorias taxonômicas, nomeação dos vários grupos monofiléticos existentes etc.. No entanto, é preciso ficar claro que o problema de reconstrução filogenética dos organismos é diferente da questão do uso das filogenias para a classificação dos grupos. Grande parte das reconstruções filogenéticas não visa alterar as classificações existentes, mas de apenas conhecer as relações entre seus grupos e permitir compreender a evolução de características (moleculares, morfológicas, comportamentais etc.) ou de sua evolução espacial. Adiante, neste capítulo, são apresentados os conceitos básicos usados em métodos de reconstrução filogenética. Os algoritmos usados para obter as árvores filogenéticas serão discutidos a seguir, com ênfase especial na reconstrução de filogenias a partir de sequências de macromoléculas.

12.3. Algoritmos de Reconstrução Filogenética

A obtenção de uma árvore filogenética através de análises computacionais requer a escolha de algoritmos de reconstrução filogenética. Existem dois grupos básicos de algoritmos para esse fim. No primeiro grupo, o princípio do método está embutido no próprio algoritmo que resulta em uma árvore final. No segundo, o princípio do método é o critério usado para a escolha da melhor árvore dentre um conjunto delas. Tradicionalmente, o primeiro grupo é mais frequentemente aplicado em métodos de distância, enquanto os critérios de máxima parcimônia e máxima verossimilhança tradicionalmente utilizam algoritmos com base no segundo grupo de métodos.

No caso do segundo grupo, os algoritmos de reconstrução farão a busca das árvores de acordo com o método de otimização de caracteres escolhido. O conjunto de árvores pode conter todas as árvores possíveis (*algoritmos exatos*), caso em que o método necessariamente encontrará a árvore (ou as árvores) que satisfaça(m) o critério de otimização. Contudo, como foi mostrado anteriormente, para um número grande de táxons terminais, o tempo computacional requerido pode ser impraticável devido ao imenso conjunto de árvores possíveis. Uma alternativa é procurar uma árvore dentro de um subconjunto de árvores (*algoritmos heurísticos*), gastando menos tempo computacional para a estimativa de uma árvore. Há o risco, no entanto, de não ser encontrada a melhor árvore segundo aquele critério. Ou seja, aplicando algoritmos heurísticos, o pesquisador deve ter em mente que pode estar lidando com uma árvore subótima para aquele critério escolhido. Por outro lado, foi demonstrado que algoritmos heurísticos, mais simples e mais rápidos, podem ser tão ou ainda mais eficientes do que algoritmos mais complexos (Russo *et al.*, 1996; Takahashi e Nei, 2000; Criscuolo e Gascuel, 2008).

12.3.1. Algoritmos Exatos

Busca Exaustiva

A busca exaustiva consiste em enumerar todas as possíveis árvores estritamente bifurcadas existentes para um grupo de OTUs. O algoritmo consiste em obter uma árvore não enraizada para três OTUs da amostra. Nesse caso, há apenas uma árvore não enraizada. A seguir, uma quarta OTU deve ser adicionada em todas as posições possíveis nessa árvore. Nessa situação, há três possíveis árvores não enraizadas indicando as relações estritamente bifurcadas entre esses táxons. Repete-se esse procedimento até que todas as OTUs tenham sido incorporadas às árvores e, dessa maneira, todas as árvores possíveis para esse grupo tenham sido construídas (Swofford *et al.*, 1996). A seguir, avalia-se, perante o critério escolhido, qual delas melhor representa a melhor otimização do modelo evolutivo por um determinado critério. As chamadas árvores de máxima parcimônia são aquelas em que o número de passos é o critério para a escolha da árvore. Ou seja, a árvore escolhida será aquela que requer o menor número de substituições nucleotídicas, no caso de dados moleculares, para explicar perfeitamente o conjunto de dados. Uma dificuldade desse algoritmo é o número enorme de árvores não enraizadas potencialmente existentes para poucos táxons, aumentando imensamente o tempo computacional para completar a análise.

Branch-and-bound

Uma maneira de contornar o problema do tempo computacional foi apresentada com o desenvolvimento de um algoritmo, também exato, denominado *branch-and-bound* (Hendy e Penny, 1982). Em princípio, esse algoritmo é semelhante à busca exaustiva, uma vez que todas as topologias candidatas à melhor árvore são testadas. No entanto, a diferença está em eliminar partes do conjunto de árvore que representam soluções subótimas. No caso da máxima parcimônia, eliminar-se-iam árvores com um número maior de passos e, no caso dos métodos de distâncias aditivas, as árvores com maior somatório dos quadrados dos desvios seriam eliminadas. Para simplificar, apenas o algoritmo para parcimônia será descrito a seguir.

A eliminação de árvores subótimas é realizada da seguinte maneira: uma árvore inicial é construída com todos os táxons por um algoritmo heurístico. O número de passos dessa árvore é calculado e armazenado com o valor X. Em seguida, a partir de uma árvore-núcleo inicial com apenas três táxons, os demais táxons são incorporados a essa árvore-núcleo, um a um. A cada passo em que um táxon é incorporado à árvore, o número de passos é estimado e aquelas que apresentam valores maiores que X são descartadas. O processo descarta também todas as árvores resultantes daquela inicialmente rejeitada, porque a adição de mais táxons necessariamente aumenta o número de passos da árvore. A vantagem desse algoritmo é que ele permite ao pesquisador reduzir o tempo gasto para encontrar a árvore que melhor se ajuste ao seu critério de otimização, diminuindo o número de árvores a serem analisadas. Entretanto, no caso de sequências em que há muito ruído filogenético (isto é, homoplasias, resultados de convergências ou paralelismos), ocorre o aumento do tempo computacional gasto para se chegar a uma estimativa, muitas vezes se igualando à própria busca exaustiva.

12.3.2. Algoritmos heurísticos

"Decomposição da politomia" (ou "decomposição da estrela")

O algoritmo de decomposição da politomia consiste em unir inicialmente todos os táxons em um único nó interno, formando uma árvore totalmente politômica. O passo seguinte consiste em avaliar todas as árvores possíveis, onde cada par de táxons é considerado um táxon único. A árvore que apresentar a união entre dois táxons que otimize o critério escolhido será escolhida como a árvore inicial para o próximo passo. Assim, reduz-se em um o número de táxons a serem conectados ao nó interno central no passo seguinte. A seguir, a árvore eleita como a melhor no passo anterior é usada nesse processo de busca de formação de um novo grupo que otimize o critério adotado. Esses procedimentos são repetidos até que todos os táxons tenham sido conectados, o que gera uma árvore final completamente resolvida, ou seja, sem politomias (Swofford *et al.*, 1996).

"Stepwise Addition"

O algoritmo *stepwise addition* é bastante simples. Primeiramente, três táxons são escolhidos para compor uma árvore-núcleo. A seguir, um quarto táxon é unido a eles em todas as posições possíveis e aquela que se apresentar a melhor árvore será considerada no passo seguinte. Então, um quinto táxon é adicionado à árvore anterior e, mais uma vez, os escores de todas as possíveis árvores bifurcadas são avaliados, sendo a melhor delas escolhida para o passo seguinte. Isso continua até que todos os táxons tenham sido inseridos e uma árvore tenha sido escolhida entre as possíveis árvores finais. Uma desvantagem desse algoritmo é que as três primeiras OTUs escolhidas da árvore-núcleo irão influenciar o subconjunto de árvores pesquisadas. Em outras palavras, o trio de táxons escolhidos para iniciar o processo cria um viés e pode levar a uma escolha que, na realidade, representa uma árvore subótima, que não seria escolhida se outros trios tivessem iniciado o processo (Swofford *et al.*, 1996). Uma maneira de contornar esse problema é realizar a busca diversas vezes, escolhendo diferentes trios iniciais de OTUs a cada vez.

"Branch-swapping"

Esse grupo de algoritmos foi desenvolvido para casos em que se analisa um grande número de táxons. Basicamente, esse algoritmo promove rearranjos em uma dada árvore inicial e a melhor árvore entre todas as construídas é a escolhida. Existem três formas de rearranjar a árvore inicial (para maiores detalhes, ver Nei e Kumar, 2000). Na primeira forma, denominada "poda e enxerto de subárvores", uma parte da árvore é selecionada, cortada e inserida em diferentes posições da árvore restante. A cada inserção, o critério é usado e é armazenada a melhor árvore até então. Na segunda maneira, denominada "bissecção de árvore", a árvore é quebrada em duas subárvores. As duas subárvores são unidas por um par de ramos diferentes dos originais. O procedimento também é repetido, até que todos os pares possíveis de ramos dessas duas subárvores sejam unidos e a melhor árvore seja definida (Swofford *et al.*, 1996). Na terceira maneira, denominada "troca de vizinhos mais próximos" (do inglês, *nearest neighbor interchange*), OTUs que aparecem separadas por apenas um ramo interno são trocadas de lugar e, a cada troca, o critério do método é usado, escolhendo-se no final a melhor das árvores reconstruídas.

Essa última maneira pode apresentar uma pequena modificação, chamada de troca de vizinhos próximos (do inglês, *closest neighbor interchange*), quando as OTUs trocadas de lugar podem estar separadas por um ou dois ramos internos. Mais especificamente, o algoritmo busca, entre as árvores próximas da temporária, aquelas que apresentem uma distância topológica de 2 ou 4. A distância topológica é, para árvores completamente resolvidas (bifurcadas), duas vezes o número de ramos internos que separam o conjunto de OTUs em dois subconjuntos diferentes (Robinson e Foulds, 1981; Penny e Hendy, 1985).

"Busca Min-Mini"

O algoritmo heurístico de min-mini foi desenvolvido por Kumar *et al.* (1993; veja também Nei e Kumar, 2000) para a busca de árvores com máxima parcimônia. Nesse algoritmo, muitas árvores que possuem baixa probabilidade de serem as árvores de máxima parcimônia são eliminadas do cálculo do comprimento da árvore, diminuindo sensivelmente o tempo computacional gasto. Entretanto, a árvore final, como nos outros métodos heurísticos, pode não ser a árvore de máxima parcimônia. Uma outra vantagem desse critério de busca é que o pesquisador pode especificar um fator de busca que vai controlar sua extensão. Um valor mais alto nesse fator determina que um número maior de árvores será examinado e, portanto, que a busca irá demorar mais.

12.4. Árvores de Sequências Macromoleculares

O uso original do princípio da parcimônia na proposição de explicações para fatos data do século XVI, conhecido inicialmente como a "navalha de Ockham" (*Ockham's razor*, em inglês). Foi formulado pelo filósofo inglês William de Ockham. O princípio pode ser assim entendido: se existem diversas explicações para uma determinada observação, é adotada a mais simples, ou seja, a que requer o menor número de pressupostos. Esse princípio é amplamente empregado na ciência com o objetivo de evitar conjuntos infinitos de pressupostos adicionais para explicar um

determinado conjunto de observações.

O emprego do princípio de parcimônia para a inferência de filogenias com dados de sequências macromoleculares é geralmente encarado como uma consequência direta do uso da metodologia criada por Willi Hennig para esse tipo de dados. Historicamente, no entanto, o desenvolvimento não foi exatamente assim.

A primeira tentativa de usar dados citogenéticos para a reconstrução de uma filogenia com o emprego do conceito de parcimônia aconteceu em 1936, quando Sturtevant e Dobzhansky propuseram uma genealogia de populações baseada nas inversões presentes nos cromossomos 3 de *Drosophila* (Figura 12.3). Nesse caso, existem outras sequências de eventos de inversão cromossômica que poderiam transformar o padrão "Arrowhead" em "Chiricahua II". Formalmente, pode ser demonstrado que as maneiras possíveis seriam infinitas. Empregando-se o princípio de máxima parcimônia, no entanto, contorna-se essa armadilha lógica. Evidentemente, essa aplicação do critério de máxima parcimônia pode ser considerada como a primeira a ter sido utilizada com "dados moleculares" somente se considerarmos um cromossomo como uma molécula, algo aceitável no nível do DNA, mas tecnicamente incorreto, pois os cromossomos também incluem as proteínas associadas. Além disso, o método de reconstrução filogenética foi empregado por esses autores em populações de uma mesma espécie, onde é natural se imaginar que os ancestrais continuam presentes.

Um dos primeiros registros que se tem do emprego explícito do princípio de máxima parcimônia para a proposta de filogenias de sequências macromoleculares foi realizado por Eck e Dayhoff, em 1966. Esses autores conheciam o trabalho de Zuckerkandl e Pauling (1962), que propunha a reconstrução de filogenias a partir de sequências de aminoácidos com o emprego de uma matriz de distâncias (veja o Capítulo 7). Eck e Dayhoff utilizaram, além do emprego das matrizes de distâncias, um algoritmo semelhante ao *stepwise addition*, descrito acima, para a escolha da árvore que comportava o menor número de passos ("mutações", segundo os autores). O algoritmo incluía a inferência das sequências ancestrais, salientando que a busca exaustiva era inviável na época.

De forma independente, segundo Edwards (1996), Cavalli-Sforza e Edwards, em 1963, propuseram o método por eles denominado de evolução mínima, mas dentro de uma perspectiva probabilística. Camin e Sokal (1965), conhecendo o trabalho de Hennig, foram os primeiros que usaram o termo "parcimônia" para um método de reconstrução de filogenias para caracteres discretos. O termo "parcimônia" vem sendo, então, empregado com significados distintos pelas razões de natureza histórica comentadas acima. Parcimônia é um princípio filosófico, como destacado por Sober (1988), mas também um método de inferência filogenética que emprega um princípio homônimo (Camin e Sokal, 1965), que pode ser entendido como um método probabilístico através de uma simplificação de natureza estatística, que implica um modelo de evolução (Farris, 1973; Felsenstein, 1973).

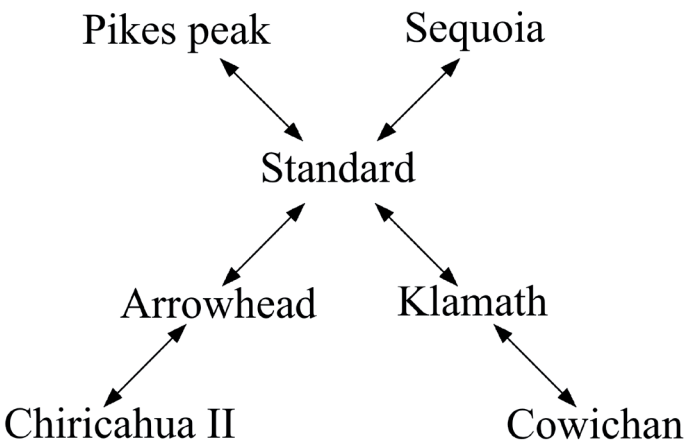


Figura 12.3. Relação entre inversões cromossômicas encontradas em *Drosophila pseudoobscura*. Modificado de Sturtevant e Dobzhansky (1936).

Em uma primeira abordagem, pode-se considerar que os dados relativos às sequências de macromoléculas são apenas mais um conjunto de caracteres, como quaisquer outros (caracteres morfológicos, ecológicos, comportamentais, fisiológicos etc.). Vários autores defendem esse ponto de vista na atualidade, enquanto outros consideram que é importante que um tratamento diferencial seja aplicado a dados que consistem de sequências de macromoléculas (para discussões sobre esse assunto, consulte de Queiroz *et al.*, 1995; Brower *et al.*, 1996). Apresentaremos aqui alguns pontos relativos a essa discussão.

Qualquer caráter precisa ser herdável para ter significado evolutivo. Sob esse prisma, devemos descartar, portanto, a variação que resulta da atuação do ambiente na formação do fenótipo. Isso significa que, de um ponto de vista reducionista, a variabilidade de um caráter estará sendo determinada por alguma variação no nível da sequência de nucleotídeos do genoma. No entanto, com raras exceções, essa relação não é tão simples. Tomando o caso de características morfológicas como um exemplo, a atuação do genoma dá-se pelos genes que se manifestam durante o desenvolvimento de acordo com o controle de transcrição de genes específicos para grupos de células ou tecidos. Atualmente, várias abordagens experimentais têm sido utilizadas para verificar a importância da atuação de sequências de DNA específicas no desenvolvimento de estruturas (veja Capítulo 11). Os resultados estão muito longe de apontar o papel de nucleotídeos específicos para os caracteres utilizados em abordagens não moleculares. Revisões críticas sobre as implicações do estudo da genética no desenvolvimento da evolução podem ser encontradas em Lewontin (2000), Oyama (2000) e Keller (2001), onde se alerta contra pressupostos simplistas ou excessivamente reducionistas.

Durante a década de 1970, a genética de populações passou a ser fortemente integrada com a biologia molecular na análise dos fenômenos de evolução, em especial com relação aos resultados obtidos na comparação de macromoléculas homólogas de vários grupos de organismos e a presença de polimorfismos moleculares. O geneticista de populações Motoo Kimura, junto com sua discípula Tomoko Ohta, propôs, em 1971, que esses dois fenômenos estão ligados pelos processos de mutações neutras e deriva genética, dentro da formulação do que foi chamado de "teoria neutralista da evolução molecular" (veja o Capítulo 7). Embora na biologia evolutiva seja comum o emprego de hipóteses com os pressupostos reducionistas embutidos (como a existência de um "gene para o altruísmo", premissa utilizada em alguns modelos de sociobiologia, como aventado por Dawkins, 1976), a maioria das características fenotípicas, não pode ser abordado da maneira como são, por exemplo, os polimorfismos de nucleotídeos simples (veja o Capítulo 19).

Conforme visto no Capítulo 10, apenas uma parcela reduzida do genoma de vários organismos tem função que pode ser atribuída diretamente a algum componente do fenótipo. Mesmo na fração do genoma que reconhecidamente contribui para as características fenotípicas, parte da variação ou das substituições observadas em sequências homólogas de diferentes organismos evolui sob as leis da evolução neutra (veja o Capítulo 7). Assim, se considerarmos que a grande maioria da variação observada no nível de sequência de macromoléculas resulta dos processos de mutações aleatórias (veja, no entanto, o Capítulo 7) e de deriva

genética, não estaremos incorrendo em risco de fazer uma premissa errada.

Se levarmos em conta, portanto, as considerações acima, tratando as filogenias moleculares, diferentemente do que tratam as filogenias obtidas através da análise de outros tipos de caracteres, alguns procedimentos deverão ser observados.

12.5. Definição do Método de Máxima Parcimônia

O método da máxima parcimônia é muito simples em sua concepção. Sob uma perspectiva probabilística, ele se baseia em um modelo de evolução (implícito) onde uma mudança é mais provável do que duas. Ou seja, trata substituições independentes gerando o mesmo resultado como um evento relativamente raro. Para ilustrar esse princípio, utilizaremos três organismos: rato, lagarto e peixe (Figura 12.4). Em cada uma das possíveis topologias (A, B e C), o aparecimento de pulmão está indicado por um traço nos ramos onde esse evento ocorreu. Nesse caso, seguindo o princípio de máxima parcimônia, a topologia **A** seria a escolhida, pois as demais topologias apresentam duas mudanças (aparecimento de pulmão), enquanto que a topologia **A** precisa assumir apenas uma mudança. Diz-se, então, que a topologia **A** representa a árvore mais parcimoniosa porque requer o menor número de mudanças para explicar o caráter "pulmão".

Na reconstrução filogenética pelo critério da máxima parcimônia, os sítios informativos merecem uma atenção especial. Isso porque nem toda a variabilidade das sequências é útil para o método de máxima parcimônia, pois aqueles que apresentarem variação em uma única sequência adicionarão um passo para qualquer uma das árvores consideradas. Uma regra básica é que a variabilidade das sequências deve dividir as OTUs em pelo menos dois grupos, de modo que cada um tenha no mínimo de dois representantes. Por exemplo, na Figura 12.5a, onde estão representadas quatro sequências de DNA hipotéticas, o sítio 2, que possui o padrão A, G, G e G em um grupo de quatro táxons (1, 2, 3 e 4), nessa ordem, não é informativo, já que todas as árvores possíveis requerem igualmente apenas uma substituição para explicá-lo perfeitamente (Fig. 12.5b). Da mesma forma, o padrão G, C, A e A, do sítio 3, também não é informativo, já que todas as árvores possíveis requerem duas substituições. Por outro lado, o sítio 5 (G, G, A e A) é informativo, já que a árvore (topologia I) que agrupa a OTU 1 com a 3 requer apenas uma substituição para explicar perfeitamente esses dados, enquanto que as outras duas árvores possíveis (agrupando 1 com 3 ou 1 com 4) requerem no mínimo duas substituições para explicar o mesmo conjunto de dados.

A análise inicia-se, portanto, com a seleção dos sítios informativos. Em seguida, o número de substituições em cada sítio informativo é inferido para cada uma das árvores possíveis e o total de substituições para cada uma das árvores possíveis é calculado. A árvore que requer o menor número de mudanças (menor comprimento) é selecionada como a mais parcimoniosa.

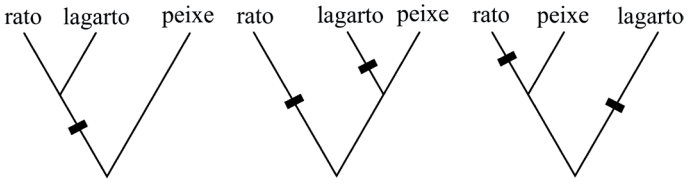


Figura 12.4A-C. Diferentes topologias possíveis relacionando rato, lagarto e peixe. Os traços perpendiculares aos ramos representam, em cada topologia, o nível em que teria aparecido o pulmão.

a)

	Sítios								
Táxons	1	2	3	4	5	6	7	8	9
1	A	A	G	A	G	T	G	C	A
2	A	G	C	C	G	T	G	C	G
3	A	G	A	T	A	T	C	C	A
4	A	G	A	G	A	T	C	C	G
					*		*		*

* - sítios informativos

b)

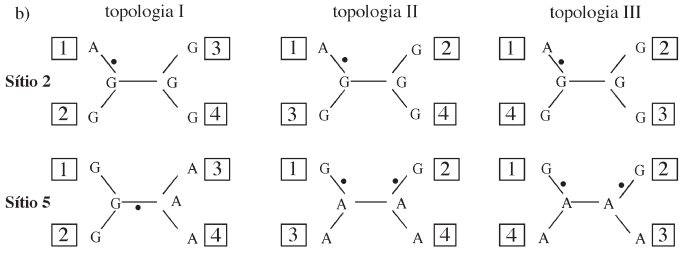


Figura 12.5. Sequências hipotéticas e suas reconstruções possíveis obtidas por máxima parcimônia: (a) matriz de dados; (b) possíveis árvores mais parcimoniosas não enraizadas obtidas da análise independente de dois sítios. ● - eventos de substituições ocorridas. Os dois nucleotídeos dos nós internos representam uma das possíveis reconstruções; por exemplo, na árvore III do sítio 5, poderia haver Gs no lugar de As nos nós internos e, mesmo assim, o número de substituições seria mantido. Modificado de Li (1997).

Alternativamente, mais de uma árvore pode apresentar o mesmo comprimento e todas as árvores mais parcimoniosas com comprimento igual deverão ser consideradas na interpretação final dos resultados.

Na inferência de filogenias moleculares, um grupo externo também pode ser utilizado para enraizar a árvore. No entanto, a polarização dos caracteres tem relativamente pouca importância nesse caso. Isso acontece porque, no caso do modelo neutro de evolução, a pressuposição de um estado plesiomórfico "G" em contraposição a um estado apomórfico "T", por exemplo, seria no mínimo arriscada, uma vez que isso implicaria a ausência de evolução nesse sítio a partir do ancestral comum até os representantes atuais dos grupos internos e externos. O estabelecimento de estados intermediários com relação a nucleotídeos de um determinado sítio (A para G, por exemplo) não abrange passagem obrigatória pelos outros estados (C ou T). Por outro lado, no caso de sequências que codificam para aminoácidos, é possível realizar a ordenação dos caracteres (veja Araújo, 1997). Por exemplo, três espécies de pererecas do gênero *Hyla* possuem em uma posição homóloga de um determinado gene os códons GAG (ácido glutâmico), GTG (valina) e CTG (leucina). Nesse caso, adotando o critério de máxima parcimônia, podemos assumir que a transformação tenha ocorrido em uma das seguintes direções: GAG → GTG → CTG ou CTG → GTG → GAG, já que ambas requerem apenas uma substituição nucleotídica a cada um dos passos.

Uma observação útil aqui é que, evidentemente, a evolução, em algumas situações, pode resultar em "árvores mais longas"— ou seja, a questão das árvores serem mais curtas é probabilística, não uma proibição à natureza.

12.6. Variantes do Método

Embora haja uma base lógica robusta subjacente ao conceito de máxima parcimônia, sua aplicação a diversas situações evolutivas não é imediata, e o método de máxima parcimônia tem sido ajustado para aplicar a questões particulares ligadas à evolução molecular. Nesta seção, discutiremos algumas das mo-

dificações mais utilizadas na literatura. Aqui vale a pena alertar o leitor de que a questão da pesagem de caracteres (assim como tantas outras) ainda se encontra em plena discussão, pois é complexa e tem sido criticada por alguns autores (como Lake, 1987; Jin e Nei, 1990; Rognes, 1999; Broughton *et al.*, 2000). O uso dos pesos em muitos casos pode ser arbitrário e, portanto, com pouca ou nenhuma base lógica ou biológica. Por esse motivo, o emprego de pesos deve ser feito com extrema cautela.

Uma das principais maneiras de pesar os caracteres é pesar transições e transversões diferencialmente. Os quatro nucleotídeos, unidades do DNA, podem ser classificados em duas classes: purinas ou pirimidinas. As purinas (A e G) possuem dois anéis de carbono-nitrogênio em sua estrutura e as pirimidinas (C e T) possuem apenas um anel. O emparelhamento das bases na dupla hélice é estruturado de forma que uma purina sempre está ligada por pontes de hidrogênio a uma pirimidina, tendo o DNA uma largura interna sempre equivalente a três anéis por toda sua extensão.

Um dos fenômenos encontrados no nível dos nucleotídeos envolve as diferenças nas frequências de substituições dos tipos transição e transversão. Geralmente, as transições (purina ↔ purina ou pirimidina ↔ pirimidina) são mais comuns que as transversões (purinas ↔ pirimidinas). Isso acontece porque uma transversão gera uma distorção de largura na molécula de DNA e a probabilidade de o sistema de reparo da célula atuar neste tipo de mutação é muito maior que no caso de uma transição.

As transições, com isso, podem chegar a um ponto no qual elas deixam de ser informativas devido à saturação, trazendo apenas ruído aos dados, especialmente com baixa amostragem taxonômica. Por esse motivo, alguns autores sugerem não considerar as transições em seus estudos. Essa situação é denominada de parcimônia de transversão, na qual se atribui peso zero para substituições dentro de cada uma das classes de nucleotídeos e, portanto, somente são levadas em conta as mudanças entre as classes (Swofford *et al*, 1996).

Outros pesquisadores não ignoram totalmente as transições, pois essas mudanças podem ser informativas nas comparações entre grupos próximos. Nesse caso, é dado mais peso às transversões em relação às transições. No caso da pesagem de transições e transversões, não há regras estabelecidas para atribuir os pesos para cada um desses tipos de evento. No entanto, como vários autores utilizam tal recurso, vamos mostrar como a pesagem destas substituições pode ser feita.

O peso a ser dado para as transições e transversões pode ser extraído a partir dos próprios dados a serem analisados. Por exemplo, dadas duas sequências (Figura 12.6a), podemos observar que a frequência de ocorrência de transições é cinco vezes maior do que a de transversões. Portanto, na matriz de pesos (Figura 12.6b), assume-se que as transversões são cinco vezes "mais importantes" do que as transições.

Outra maneira de diferenciar os tipos de caracteres é realizada pela pesagem sucessiva (no inglês, *weighted parsimony*). Por essa estratégia, uma análise de congruência entre os caracteres é realizada e aos dados mais congruentes entre si (ou seja, como evidenciado pela incongruência interna entre esse conjunto de caracteres, supostamente menos homoplásticos) são atribuídos pesos maiores. Esse procedimento é repetido sucessivamente até que os valores de pesos se tornem constantes (uma revisão do princípio pode ser encontrada em Meyer, 1997; veja também Rognes, 1999). Nesse caso, um risco do método é o efeito da análise inicial, que se propaga nas análises posteriores, que teoricamente implica a subrepresentação relativa de parte dos dados.

Outro fenômeno encontrado em estudos evolutivos moleculares é a heterogeneidade das taxas de evolução. Por exemplo, como já visto no Capítulo 7, o código genético é degenerado e,

a)

OTU	Sequência
1	AACGTGCTTGGCTAGTCATGCA
2	A***G***CGT***A***CT***C***GG<u>A</u>TA***A***TC***G***TGCA

b)

	A	C	G	T
A	0	5	1	5
C		0	5	1
G			0	5
T				0

Figura 12.6. (a) Exemplo de duas sequências hipotéticas mostrando transições em negrito itálico e transversão sublinhada. (b) Matriz de pesos para transições e transversões baseadas nas sequências hipotéticas.

com isso, as mudanças na terceira base do códon nem sempre têm como consequência a troca de aminoácido. Como a pressão seletiva nas regiões codificadoras do genoma é mais acentuada no nível da sequência de aminoácidos, a 3ª base do códon é a que apresenta menores restrições evolutivas, dado o padrão de degeneração de códons do código genético. Consequentemente, espera-se que a 3ª base será a mais saturada de substituições (considere, no entanto, o viés de utilização de códons, abordado no Capítulo 7, para certas sequências, especialmente aquelas mais expressas). Por isso, é possível dar mais peso às 1as e 2as bases (esse peso podendo ser maior para uma delas) ou simplesmente ignorar totalmente os dados da 3as base, embora, se há amostragem taxonômica suficiente para evitar a saturação, isso implique em perda de informação.

Além disso, pode ser dado peso maior a certas regiões da sequência que sejam mais informativas. Por exemplo, no caso de uma proteína estrutural transmembrânica, que possui domínios que atravessam uma membrana (mitocondrial, por exemplo) e outros domínios que formam alças para fora dela, as regiões intramembranosas serão mais conservadas, pois delas depende o ancoramento da proteína como um todo, ao contrário das regiões de alças, que teriam maior "flexibilidade" frente a mutações, sem prejudicar o bom funcionamento da proteína. As diferenças nas restrições evolutivas no nível da estrutura da proteína se refletem na quantidade de variação apresentada no nível da sequência de nucleotídeos que a codificam.

Finalmente, cabe lembrar que a parcimônia estrita em sua versão tradicional também pode ser encarada como uma forma de pesagem: todos os caracteres são *a priori* definidos como tendo peso 1, o que também não deixa de corresponder a uma decisão sobre pesagem. Isso significa que as discussão metodológica relativa a pesagem não são de simplesmente haver pesagem ou não, pois a parcimônia estrita também é uma forma de pesagem. A questão, portanto, é qual conceito de pesagem, uniforme ou diferencial, é mais adequado para tratar os dados de maneira a obter como resultados reconstruções confiáveis, dentre as milhões ou bilhões de topologias alternativas, para representar a filogenia verdadeira dos grupos.

Essas variações do método de parcimônia (pesagens de caracteres) podem ser utilizadas no programa PAUP* (Swofford, 1999). Esse programa permite introduzir uma matriz de pesos como a da Figura 12.6b ou ainda definir pesos específicos para posições do códon (no caso de genes codificadores) ou regiões específicas da sequência (como domínios protéicos transmembrânicos e domínios não transmembrânicos).

Alternativamente, a análise de topologias envolvendo quatro sequências de nucleotídeos pode ser realizada pelo

método de invariantes de Lake (*Lake's Method of Invariants*). Esse método foi sugerido por Lake (1987), que o denominou de parcimônia evolutiva, embora não seja baseado em princípios de parcimônia (Li, 1997). Tal método assume que as substituições em um determinado sítio são independentes, os tipos de transversões são igualmente prováveis (por exemplo, A pode sofrer mudança para T ou C com igual probabilidade) e inserções e deleções podem ser ignoradas. Nesse caso, transversões paralelas em dois ramos produzem igual número de nucleotídeos do tipo 1 e do tipo 2. Por exemplo, na Figura 12.7, as transversões paralelas do tipo 1 são as que geram o mesmo nucleotídeo (por exemplo, A → C na OTU 1, e C → A na OTU 4) e as do tipo 2 resultam em nucleotídeos diferentes (por exemplo, A → C na OTU 1, e C → G na OTU 4). Com isso, o efeito dessas transversões pode ser cancelado com a subtração dos eventos do tipo 2 dentre os do tipo 1, reduzindo as homoplasias. Entretanto, esse método mostrou-se ineficiente na recuperação da filogenia verdadeira quando existem diferenças nas taxas de transversões ou quando a taxa de substituição não varia com os ramos (Jin e Nei, 1990).

12.7. Vantagens e Desvantagens dos Métodos que Envolvem a Máxima Parcimônia

A grande vantagem do método de máxima parcimônia em relação a outros métodos de inferência filogenética é que ele tem um conceito subjacente muito simples, ou seja, sua maneira de interpretar os eventos pode ser facilmente compreendida. De um modo bastante simplificado, o método da máxima parcimônia

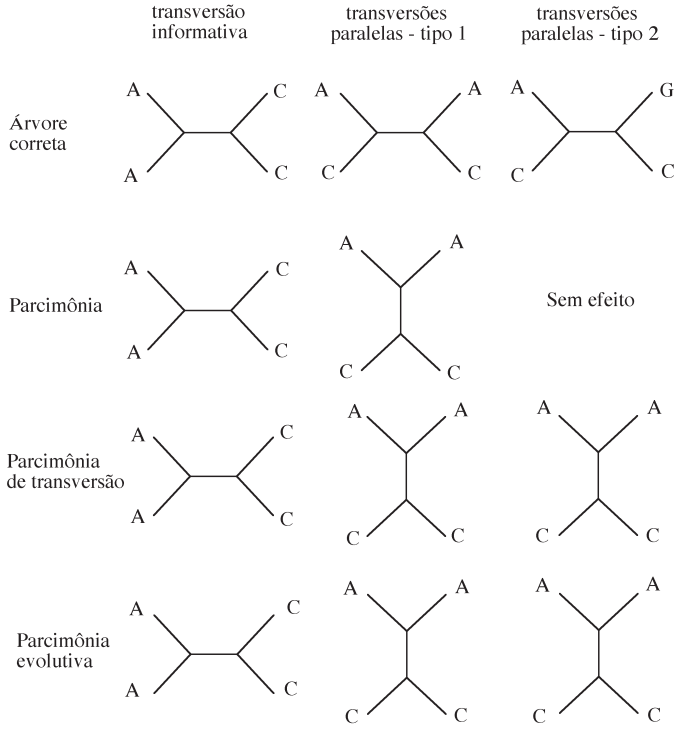


Figura 12.7. Efeitos de substituições de nucleotídeos em diferentes variantes do método de parcimônia. Uma transversão informativa no ramo interno resulta em árvores corretas na análise com três variantes de parcimônia. Duas transversões paralelas (tipo 1) resultam em árvores incorretas com as três variantes. Duas transversões gerando dois nucleotídeos diferentes (tipo 2) não são informativas na parcimônia tradicional; a parcimônia de transversão gera a árvore incorreta e a parcimônia evolutiva estima as substituições múltiplas nos ramos periféricos, diminuindo o suporte da árvore incorreta (modificado de Swofford *et al.*, 1996).

escolhe o caminho "menos tortuoso", ou seja, aquele que exige de um menor número de pressupostos sobre eventos independentes na evolução de um grupo. Seu princípio fundamental está na idéia de que os eventos de substituições independentes gerando condições finais iguais devem ser visto como menos prováveis e, portanto, devem ser minimizados.

Tanto em estudos com dados moleculares como com caracteres morfológicos, quando o conjunto de dados apresenta uma quantidade de homoplasias grande relativamente à quantidade de sítios informativos, a abordagem com máxima parcimônia pode incorrer em erros (Jin e Nei, 1990; Takezaki e Nei, 1994; Tateno *et al.*, 1994; Nei *et al.*, 2000).

Consideremos um exemplo no qual supomos *a priori* uma topologia decorrente de taxas de evolução heterogêneas de alguns ramos periféricos. Esse evento pode ser visualizado na Figura 12.8, onde a taxa de evolução nos ramos periféricos levando aos táxons 1 e 4 é muito mais elevada que nos demais ramos. Devido a essa grande diferença, os demais ramos são tão curtos a ponto de os táxons 2 e 3 quase não apresentarem diferenças. Por essa razão, os táxons 2 e 3 são incorretamente agrupados pelo critério de máxima parcimônia e os dois ramos longos tendem a permanecer juntos, independentemente da filogenia real. Nesse caso, quanto maior o comprimento da sequência usada, maior será o suporte estatístico equivocado desse grupamento inexistente na natureza. Esse problema é chamado de "atração de ramos longos" (em inglês, *long branch attraction*, LBA), mostrando uma situação onde o emprego do critério de máxima parcimônia produz consistentemente interpretações errôneas. Para contornar esse problema, é possível adicionar OTUs cujos ramos se liguem a esses ramos maiores, subdividindo-os (Swofford *et al.*, 1996), ou seja, aumentando a amostragem taxonômica. Nesse caso, é necessário que existam essas OTUs (há alguns clados com pouquíssimas espécies conhecidas) e que seus dados estejam disponíveis ou possam ser coletados para a análise.Podemos ainda limitar as análises aos caracteres mais conservados (menos homoplásticos) das sequências. Outra alternativa seria escolher um gene (ou segmento de gene) com menos variação para análise (ver Capítulo 16).

Quando os caracteres utilizados não apresentam muitas homoplasias (ou seja, quando a divergência entre os táxons estudados é pequena), no entanto, é possível obter filogenias corretas e com suporte estatístico alto (*e.g.*, Russo *et al.*, 1996; Takahashi e Nei, 2000). Nesses casos, os modelos estatísticos que lidam com homoplasias podem se ajustar melhor a dados de sequências de macromoléculas que a outros tipos de dados, uma vez que os primeiros podem ser mais previsíveis evolutivamente sob certas condições (sequências de cópias únicas com evolução neutra ou com restrição funcional reduzida, não relacionadas a transposição, com ausência de "pontos quentes" mutacionais em populações com tamanho relativamente estável).

O método de parcimônia estrita também não considera a heterogeneidade da taxa de substituição entre determinados sítios

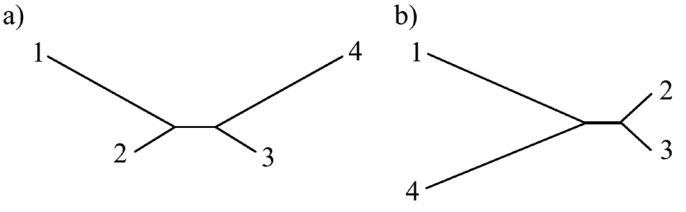


Figura 12.8. Topologia hipotética de quatro táxons contendo dois ramos periféricos muito longos e os demais ramos muito curtos: (a) topologia que corretamente representa a evolução do grupo; (b) topologia mais parcimoniosa, que representa incorretamente a evolução do grupo (modificado de Swofford *et al.*, 1996).

(mesmo dentro de uma mesma sequência, é possível ter regiões mais variáveis), de uma maneira semelhante ao que acontece nos nucleotídeos correspondentes às bases degeneradas dos códons, conforme visto anteriormente. Um problema de natureza prática que ocorre com o emprego do critério de máxima parcimônia é que, para muitas sequências, o número de filogenias possíveis (Tabela 12.1) pode ser muito grande, ou seja, pode ser um método com grande demanda computacional. Entretanto, esse problema diminui relativamente com a disponibilidade de computadores mais eficientes.

Agradecimentos

Este trabalho foi financiado em parte pela FAPESP, FAPERJ e CNPq. Os autores agradecem também as sugestões e comentários do Dr. Dalton S. Amorim.

Referências Bibliográficas

Amorim, D.S. (2002). **Fundamentos de Sistemática Filogenética**. Holos, Editora. Ribeirão Preto.

Araújo, M.E. (1997). Caracteres moleculares e morfológicos de *Phillodryas* e *Tropidodryas* (Serpentes: Xenodontinae): um estudo cladístico. **Tese de Doutorado em Genética**, Universidade Federal do Rio de Janeiro.

Bernardi, N. (1981). Phylogenetic relationships, monophyletic group and related concepts. **Rev. Bras. Entomol. 25**: 323-326.

Broughton, R.E., Stanley, S.E. e Durrett, R.T. (2000). Quantification of homoplasy for nucleotide transitions and transversions and a reexamination of assumptions in weighted phylogenetic analysis. **Syst. Biol. 49**: 617-627.

Brower, A.V.Z., DeSalle, R. e Vogler, A. (1996). Gene trees, species trees, and systematics: a cladistic perspective. **Ann. Rev. Ecol. Syst. 27**: 423-450.

Camin, J.H. e Sokal, R.R. (1965). A method for deducing branching sequences in phylogeny. **Evolution 19**: 311-326.

Cavalli-Sforza, L.L. e Edwards, A.W.F. (1963). The reconstruction of evolution. **Ann. Human Genet. 27**: 104-105.

Criscuolo, A. e Gascuel, O. (2008). Fast NJ-like algorithms to deal with incomplete distance matrices. **BMC Bioinformatics 9**:166.

Dawkins R. (1976). **The Selfish Gene**. Oxford University Press, New York.

de Queiroz, A., Donoghue, M. J. e Kim, J. (1995). Separate versus combined analysis of phylogenetic evidence. **Ann. Rev. Ecol. Syst. 26**: 657-681.

Eck, R.V. e Dayhoff, M.O. 1966. **Atlas of Protein Sequence and Structure**. Natl. Biomed. Res. Found., Silver Springs, Maryland.

Edwards, A.W.F. (1996). The origin and early development of the method of minimum evolution for the reconstruction of phylogenetic trees. **Syst. Biol. 45**: 79-91.

Farris, J.S. (1973). A probability model for inferring evolutionary trees. **Syst. Zool. 22**: 250-256.

Felsenstein, J. (1973). Maximum likelihood and minimum steps methods for estimating evolutionary trees from data on discrete characters. **Syst. Zool. 22**: 240-249.

Felsenstein, J. (1978). The number of evolutionary trees. **Syst. Zool. 27**: 27-33.

Graur, D. e Li, W. -H. (2000). **Fundamentals of molecular evolution**. 2a edição. Sinauer Press, Sunderland, Massachusetts.

Hendy, M. D. e Penny, D. (1982). Branch and bound algorithms to determine minimal evolutionary trees. **Math. Biosci. 59**: 277-290.

Hennig, W. (1950). **Grundzuge einer Theorie der phylogenetischen Systematik**. Deuscher Zentralverlag, Berlin.

Hennig, W. (1966). **Phylogenetic Systematics**. University of Illinois Press, Urbana, Ill.

Jin, L. e Nei, M. (1990). Limitations of the evolutionary parsimony method of phylogenetic analysis. **Mol. Biol. Evol. 7**: 82-102.

Keller, E.F. (2001). **The Century of the Gene**. Harvard University Press, Cambridge.

Kimura, M. e Ohta, T. (1971). Protein polymorphism as a phase of molecular evolution. **Nature 229**: 467-469.

Kumar, S., Tamura, K. e Nei, M. (1993). **MEGA: molecular evolutionary genetics analysis**. Pennsylvania State University, University Park, EUA.

Lake, J.A. (1987). Rate-independent technique for analysis of nucleic acid sequences: Evolutionary parsimony. **Mol. Biol. Evol. 4**: 167-191.

Lewontin, R.C. (2000). **The Triple Helix: Gene, Organism, and Environment**. Harvard University Press, Cambridge.

Li, W.-H. (1997). **Molecular Evolution**. Sinauer Associates, Inc. Sunderland, MA.

Maddison W.P., Donoghue M.J. e Maddison D.R. (1984). Outgroup analysis and parsimony. **Systematic Zoology 33**: 83-103.

Meyer, D. (1997). Análise filogenética de sequências de DNA. **In** Amorin, D.S. **Elementos Básicos de Sistemática Filogenética**. Holos, editora e Sociedade Brasileira de Entomologia. Ribeirão Preto, SP. pp. 187-212.

Nei, M. e Kumar, S. (2000). **Molecular Evolution and Phylogenetics**. Oxford University Press, New York, New York, USA.

Nei, M., Kumar, S. e Takahashi, K. (2000). The optimization principle in phylogenetic analysis tends to give incorrect topologies when the number of nucleotides or amino acids used is small. **Proc. Natl. Acad. Sci. USA 95**: 12390-12397.

Oyama, S. (2000). **The Ontogeny of Information: Developmental Systems and Evolution**. Duke University Press,

Penny, D. e Hendy, M.D. (1985). The use of tree comparison metrics. **Syst. Zool. 34**: 75-82.

Robinson, D. F. e Foulds, L.R. (1981). Comparison of phylogenetic trees. **Math. Biosci. 53**: 131-147.

Rognes, K. (1999). Successive weighting and polymorphic terminals – a warning to PAUP users. **Cladistics 15**: 69-72.

Russo, C.A.M., Takezaki, N. e Nei, M. (1996). Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny. **Mol. Biol. Evol. 13**: 525-536.

Sober, E. (1988). Reconstructing the past. Parsimony, evolution and inference. MIT Press, Cambridge

Sturtevant, A.H. e Dobzhansky, T. (1936). Inversions in the third chromosome of wild races of *Drosophila pseudoobscura*, and their use in the study of the history of the species. **Proc. Natl. Acad. Sci. USA 22**: 448-450.

Swofford, D.L. (1999). **PAUP*. Phylogenetic Analysis Using Parsimony (* and other methods)**. Versão 4.0, Sinauer Associates, Sunderland, Massachussets, USA.

Swofford, D.L., Olsen, G.J., Waddell, P.J. e Hillis, D.M. (1996). Phylogenetic inference. **In** Hillis, D.M. Moritz, C. e Mable, B.K.. **Molecular Systematics**. 2nd edition. Sinauer Associates, Inc. Sunderland, MA, pp: 407-514.

Takahashi, K. e Nei, M. (2000). Efficiencies of fast algorithms of phylogenetic inference under the criteria of maximum parsimony, minimum evolution, and maximum likelihood when a large number of sequences are used. **Mol. Biol. Evol. 17**: 1251-1258.

Takezaki, N. e Nei, M. (1994). Inconsistency of the maximum parsimony method when the rate of nucleotide substitution is constant. **J. Mol. Evol. 39**: 210-218.

Tateno, Y., Takezaki, N. e Nei, M. (1994). Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum parsimony methods when substitution rate varies with site. **Mol. Biol. Evol. 11**: 261-277.

Zuckerkandl, E. e Pauling, L. (1962). Molecular disease, evolution and genic heterogeneity. **In** Kasha, M. e Pullman B. (eds.). **Horizons in Biochemistry**. Academic Press, New York, pp. 189-225.

Capítulo 13

Reconstrução filogenética: Métodos geométricos

Cláudia A. de Moraes Russo (claudia@biologia.ufrj.br)
Departamento de Genética
Instituto de Biologia
Universidade Federal do Rio de Janeiro

Cristina Yumi Miyaki (cymiyaki@ib.usp.br)
Departamento de Genética e Biologia Evolutiva
Instituto de Biociências
Universidade de São Paulo

Dr. Sergio Luiz Pereira (sergiolp@gmail.com)
Department of Natural History
Royal Ontario Museum

"*E as dúbias sombras tomavam forma* (...) *e as linhas desenhavam-se nítidas, e tudo se ia esclarecendo e tudo se aclarava* (...)" (Aluísio de Azevedo, *O cortiço*)

13.1. Introdução

Sequências alinhadas de nucleotídeos ou de aminoácidos podem ser comparadas quantitativamente através de vários modelos, que refletem sua semelhança entre os resíduos comparados. Em sequências homólogas, ou seja, quando se trata de trechos de DNA herdados de um ancestral comum, as diferenças entre elas originaram-se a partir de mutações (substituições) que ocorreram em linhagens que se diversificaram a partir de seu ancestral comum exclusivo. A quantificação dessas diferenças—que pode ser expressa na forma de valores ou distâncias—pode ser utilizada para a inferência de filogenias. Os algoritmos de reconstrução filogenética que utilizam essas medidas são conhecidos coletivamente como métodos de distância. Enquanto os métodos de distância reduzem a variação entre cada duas sequências a uma única medida de distância entre elas e trabalham com essas distâncias na estimativa da árvore final, os métodos baseados em caracteres discretos analisam cada sítio separadamente e constroem a árvore final diretamente a partir dos próprios caracteres. Tais algoritmos não utilizam a matriz de distância, mas reconstroem a árvore filogenética diretamente a partir dos dados (Nei, 1987; Felsenstein, 2004), do mesmo modo que os métodos baseados em máxima parcimônia, verossimilhança e análise bayesiana (ver outros capítulos deste livro).

Neste capítulo, forneceremos uma breve descrição desses algoritmos com base em distância e a aplicação dos métodos de distância em reconstrução filogenética.

13.2. Modelos de Substituição

Nos métodos que utilizam distância para reconstrução filogenética, duas etapas são necessárias: o cálculo da distância propriamente dita e a construção da topologia. O cálculo da distância é feito como uma estimativa par a par do número de substituições ocorridas em cada um das sequências desde a divergência da sequência ancestral que deu origem às duas sequências comparadas. Assim, uma distância de 0,10 significa que 10% dos resíduos mudaram numa ou na outra sequência comparada. Vamos primeiro discutir os diferentes tipos de distância mais usados na literatura, para, em seguida, considerar os algoritmos de reconstrução filogenética usados na matriz de distância para a recuperação da topologia e a estimativa dos tamanhos de ramos.

Distância p

Uma das abordagens mais simples, a distância **p**, é meramente a proporção de posições em que as duas sequências diferem, ou seja,

$$p = n_p/n, \qquad (1)$$

onde n_p é o número de nucleotídeos ou aminoácidos diferentes entre duas sequências e n é o número total de nucleotídeos ou aminoácidos comparados, respectivamente. Sua variância é razoavelmente simples e pode ser estimada pela fórmula

$$V(p) = p(1 - p)/n. \qquad (2)$$

Apesar de essa distância não levar em consideração substituições múltiplas (ver a próxima seção), ela ainda é eficaz quando a distância média entre as espécies é menor do que 0,20 (Nei, 1991). Na verdade, na faixa de 0 a 0,20, não existe muita diferença na média da distância estimada pelos diferentes métodos para cálculo da distância. Portanto, a mais recomendada, nesse caso, seria a **p**, já que sua variância é menor (Nei, 1987, 1991; Russo, 1997).

Além da taxa de substituição, outro ponto que poderá influenciar a eficiência da distância **p** na recuperação da topologia correta é a constância das taxas de evolução entre as linhagens. Nesse momento, torna-se necessário definir a consistência de métodos de análise filogenética. Um método é chamado consistente se, baseado em um número infinito de dados (isto é, erro amostral equivalente a zero), ele recupera a árvore correta. A distância **p** é uma distância consistente, ou seja, eficiente na recuperação da topologia correta se as sequências estiverem evoluindo a uma

taxa constante (isto é, segundo o relógio molecular) e se a taxa de substituição for baixa. Nesse caso, estamos falando de topologias e não de árvores filogenéticas propriamente ditas (topologia e comprimento de ramos), já que os tamanhos de ramos estarão necessariamente sendo subestimados caso a taxa de substituição seja alta, independentemente do relógio molecular estar atuando ou não. Portanto, a distância **p** irá produzir árvores filogenéticas erradas com alto suporte estatístico se as taxas de substituição estiverem variando entre as linhagens (Tajima e Takezaki, 1994; Rzhetsky e Sitnikova, 1996).

Distância de Jukes-Cantor (Jukes e Cantor, 1969)

Quando comparamos genes que possuem taxa de substituição um pouco maior que a ideal para aquele determinado problema filogenético (ver Capítulo 16), a distância **p** torna-se uma subestimativa grave da distância real entre as sequências. Isso significa que a linearidade do gráfico distância p *versus* tempo só funciona para os estágios iniciais de divergência (Sullivan e Joyce, 2005). A partir desse estágio inicial de divergência, que varia de acordo com a molécula estudada, mutações reversas, paralelas e múltiplas se acumulam, mascarando o número real de substituições ocorridas desde o ancestral comum, subestimando o número de substituições.

Por exemplo, na Tabela 13.1, podemos observar a sequência hipotética ancestral (0) e duas sequências atuais derivadas da ancestral (1 e 2). Nos seis sítios ilustrados nessa tabela, podemos notar uma grande diferença entre o número observável de substituições (3, quando comparamos apenas as sequências recentes) e o número real de substituições (8, quando comparamos as duas sequências recentes com a ancestral) e notamos que alguns sítios sofreram mais de uma substituição. Portanto, geralmente temos uma subestimativa do número real de substituições. Verificando cada sítio exclusivamente, não é possível saber se o número observável de substituições está gravemente subestimado. Entretanto, quando comparamos a sequência inteira do gene, podemos, a partir do alinhamento, ter uma idéia da variabilidade da região sequenciada. Por exemplo, no caso das sequências apresentadas na tabela, 50% dos sítios apresentam substituições observáveis entre as duas sequências, porcentagem que, em si, já é um indicativo forte de que o número de substituições deve estar subestimado. De uma maneira geral, se a distância **p** para toda a região for muito grande (> 0,20), devemos necessariamente corrigi-la com um modelo de substituição que leve em consideração substituições múltiplas.

Para contornar esse problema, Jukes e Cantor desenvolveram um modelo baseado na distribuição de Poisson, que leva em consideração substituições múltiplas, e a distância, nesse caso, é dada pela fórmula

$$d = -\,3/4\,\log_e(1 - 4/3p), \qquad (3)$$

com variância

$$V(d) = p(1 - p)/[(1 - 4/3p)^2 n], \qquad (4)$$

onde **p** é estimado pela fórmula (1). A distância de Jukes-Cantor torna-se incalculável (infinita) quando **p** é maior do que 0,75, já que isto faz com que o termo logarítmico da fórmula (3) fique negativo. Tajima (1993) desenvolveu um algoritmo de correção que podemos aplicar ao modelo Jukes-Cantor e evitar que o termo logarítmico fique negativo. No entanto, note que um **p** de 0,75 significa que apenas 25% dos sítios não apresentaram substituições **observáveis**, ou seja, o número real de substituições seria seguramente maior do que 1 (mais de uma substituição por sítio, na média), o que nos levaria a questionar o próprio alinhamento das sequências. Em uma situação como essa, seria mais apropriado procurar regiões genômicas mais conservadas para estudar as relações filogenéticas do grupo em questão.

Para sequências de aminoácidos, essa distância é equivalente ao que chamamos de distância de Poisson. A distância, nesse caso, pode ser estimada pela fórmula

$$d = -\,(19/20)\log_e(1 - 20p/19),$$

ou simplesmente

$$d = -\,\log_e(1 - p), \qquad (5)$$

com variância

$$V(d) = p/[(1 - p)n]. \qquad (6)$$

Essa distância é baseada em um modelo de substituição onde cada base ou aminoácido tem a mesma probabilidade de substituir qualquer outra base ou aminoácido. Na prática, essa condição muitas vezes não é satisfeita, de modo que outros métodos foram propostos e serão descritos a seguir.

Distância de Kimura 2-parâmetros (Kimura, 1980)

A distância de Kimura 2-parâmetros leva em consideração um fator a mais que as substituições múltiplas de Jukes-Cantor, no caso, a razão de transições e transversões (P/Q). As transições são as substituições de uma base purínica por outra base purínica, bem como aquelas de uma base pirimídica por outra base pirimídica. Por outro lado, as transversões são as substituições de uma purina por uma pirimidina e vice-versa.

Se a substituição de nucleotídeos ocorresse ao acaso, essa taxa seria de 0,5, já que o número de transversões possíveis (A↔C, A↔T, G↔C, G↔T) é duas vezes maior que o número de transições (A↔G, C↔T). No entanto, em geral não é isso que observamos. Na maior parte das vezes, quando analisamos duas sequências homólogas, essa taxa é maior do que 0,5, podendo chegar a números expressivos e até ao infinito. Isso se deve ao fato de que as purinas têm dois anéis de carbono, enquanto que as pirimidinas têm apenas um. Assim, as substituições de purina por purina ou pirimidina por pirimidina são fisicamente mais prováveis, já que não afetam tanto a conformação da molécula de DNA e, portanto, passarão mais facilmente pelo crivo dos mecanismos de reparo da célula. Além disso, como a maior parte das transversões altera o aminoácido a ser codificado, elas provavelmente serão eliminadas pela seleção negativa (Dickerson, 1971).

Em geral, quando observamos pares de espécies próximas, a P/Q é alta ou até infinita, porque o número de transversões é muito baixo ou zero, respectivamente. Conforme as sequências vão divergindo, o número de transversões vai aumentando em

Tabela 13.1. Sequência ancestral 0 que originou duas sequências recentes 1 e 2. Note que as linhagens que deram origem às sequências recentes apresentam um número real de substituições (8) bem maior que aquele observável apenas comparando as sequências recentes (3).

Seq. 0 (ancestral)	A	T	C	C	G	A
	Conservado	Simples	Paralela	Múltipla	Simples	Reversa
Seq. 1 (recente)	A	T	G	G-T	T	C-A
Seq. 2 (recente)	A	C	G	C	G	A

relação ao número de transições. Isso ocorre porque uma transversão pode mascarar uma transição, de modo que o número de transições se reduz à medida que as sequências divergem.

O modelo de Kimura leva em consideração esse desvio na direção das transições, ou seja, se o nucleotídeo original for A, então a probabilidade de que ele seja substituído por um G é maior do que por C ou T, dependendo da taxa P/Q. Essa proporção pode ser estimada diretamente a partir dos dados, por exemplo empregando-se o programa MEGA (Kumar *et al.*, 1993), como Kimura originalmente formulou. A fórmula da distância é

$$d = -\log_e(1 - 2P - Q) - \log_e(1 - 2Q), \qquad (7)$$

com uma fórmula de variância mais complexa,

$$V(d) = [c_1{}^2P + c_3{}^2Q - (c_1P + c_3Q)^2]/n, \qquad (8)$$

onde

$$c_1 = 1/(1 - 2P - Q),\ c_2 = 1/(1 - 2Q) \text{ e } c_3 = (c_1 + c_2).$$

Note que, nesse caso, **d** se torna indefinida (isto é, o termo logarítmico fica negativo) quando **p** (fórmula 1) é maior do que 0,85, mas a correção de Tajima (1993) também pode ser aplicada nesse modelo para contornar o problema. Outra opção é usar o programa PHYLIP (Felsenstein, 1995). Nesse caso, a taxa P/Q é dada pelo usuário e o termo logarítmico raramente torna-se negativo (Nei *et al.*, 1995).

Distância de Tajima e Nei (Tajima e Nei, 1984)

Quando a taxa de composição de nucleotídeos G e C—com frequência chamada de conteúdo de GC—varia entre as sequências, pode haver uma distorção nas distâncias, já que a probabilidade de substituição entre as bases também não é igual. Portanto, duas sequências com proporção de bases nitrogenadas convergentemente semelhantes parecerão mais próximas do que elas o são na realidade. Por isso, Tajima e Nei desenvolveram uma fórmula para contornar esse problema. Nesse caso, a taxa GC também é estimada diretamente dos dados. A fórmula é

$$d = -b\log_e(1 - p/b), \qquad (9)$$

para o cálculo da distância, com variância

$$V(d) = p(1 - p)/[(1 - p/b)^2 n], \qquad (10)$$

onde

$$b = \tfrac{1}{2}\left(1 - \sum_{i=1}^{4} g_i^2 + p^2/c\right) \text{ e } c = \sum_{i=1}^{3}\sum_{j=i+1}^{4} x_{ij}/2g_ig_j .$$

Nesse caso, g_i e g_j são as frequências dos nucleotídeos i e j, respectivamente (i,j=A, T, C, G), e x_{ij} é a frequência relativa do par de nucleotídeos ij, na comparação entre duas sequências. O termo logarítmico da fórmula (9) pode ficar negativo quando **b** é maior que **p** e, aqui, a correção de Tajima (1993) também pode ser aplicada para evitar o problema.

Tamura 3-parâmetros (Tamura, 1992)

A distância de três parâmetros de Tamura agrupa todos os parâmetros previamente discutidos, ou seja, considera substituições múltiplas, conteúdo de GC e a taxa de transição/transversão. À primeira vista, pode parecer que essa seja a mais apropriada de todas as distâncias, por levar em consideração mais parâmetros, pois, quanto maior o número de parâmetros incluídos no cálculo da distância, mais próxima seria a distância média estimada da distância verdadeira. Entretanto, com o aumento do número de parâmetros, maior também será a variância associada a essa medida (Russo *et al.*, 1996, Capítulo 16; Sullivan e Joyce, 2005).

Portanto, essa distância é recomendada somente para casos em que o conteúdo de GC e P/Q são bem diferentes do esperado, 0,5. A fórmula para essa distância é

$$d = -2\Theta(1-\Theta)\log_e\{1 - P/[2\Theta(1-\Theta)] - Q\} - \_[1 - 2\Theta(1-\Theta)]\log_e(1-2Q), \quad (11)$$

onde Θ é o conteúdo de GC (em porcentagem) e P e Q são as frequências dos pares de transições e transversões, respectivamente, comparando cada duas sequências. A variância, nesse caso, é dada pela fórmula,

$$V(d) = [c_1{}^2P + c_3{}^2Q - (c_1P + c_3Q)^2]/n, \qquad (12)$$

onde

$$c_1 = 1/[1 - P/2\Theta(1-\Theta)] - Q,\ c_2 = 1/(1-2Q) \text{ e } c_3 = 2\Theta(1-\Theta)(c_1 - c_2) + c_2.$$

Distância de Tamura e Nei (1993)

O tipo de distância que leva em consideração o maior número de parâmetros é a distância de Tamura e Nei (1993). Nesse caso, considera-se que as taxas de transição entre purinas (A e G) e entre pirimidinas (C e T) podem ser diferentes e elas são incluídas separadamente na análise dos dados. Para a estimativa da distância, os autores desenvolveram a fórmula

$$\begin{aligned} d = &-(2g_Ag_G/g_R)\ln[(1 - g_RP1/2g_Ag_G) - Q/2g_R] \\ &-(2g_Tg_C/g_Y)\ln[(1 - g_YP2/2g_Tg_C) - Q/2g_Y] \\ &-2(g_Rg_Y - g_Ag_Gg_Y/g_R - g_Tg_Cg_R/g_Y)\ln(1 - Q/2g_Rg_Y), \end{aligned} \qquad (13)$$

onde P1 e P2 são as proporções de transições entre A e G e entre C e T, respectivamente, e Q é a proporção de transversões entre as sequências. Essa distância foi desenvolvida especificamente para ser usada na região controladora ("D-loop") do DNA mitocondrial (Tamura e Nei, 1993). Foi sugerido, entretanto, que a heterogeneidade da variabilidade entre os sítios também deve ser considerada quando trabalhamos com essa região (Excoffier e Yang, 1999). Nesse caso, a distância mais apropriada seria a Gama-Tamura-Nei, como veremos abaixo, o uso da correção gama (Felsenstein, 2004).

Distância Gama-Poisson

Todas as distâncias que discutimos acima são baseadas no pressuposto de que todos os sítios evoluem a uma mesma taxa, o que raramente acontece. Por exemplo, em regiões codificadoras, a terceira posição do códon evolui mais rápido que as outras duas. Quando cada sítio evolui de acordo com a distribuição de Poisson, mas a taxa (λ) varia de sítio para sítio, essas taxas seguem uma distribuição gama (Uzzell e Corbin, 1971). A distribuição gama é descrita por dois parâmetros, α e β. O primeiro parâmetro é a variação das taxas de acordo com o sítio (α = 0 para taxas completamente diferentes e α = ∞ para taxas idênticas) e o segundo é a calibração de acordo com a média. Esses dois parâmetros são relacionados pela fórmula

$$\alpha/\beta = \mu, \qquad (14)$$

onde

$$\alpha = (\mu^2)/c, \quad (15)$$

μ é a média e **c** é a variância das taxas por todos os sítios. Vários métodos foram desenvolvidos para estimar o número de substituições por sítio, a média e a variância ao longo dos sítios. Os métodos baseados em parcimônia em geral subestimam o número de substituições em um sítio, ou seja, superestimam o parâmetro α (Uzzel e Corbin, 1971; Tamura e Nei, 1993). Por outro lado, os métodos baseados em verossimilhança não têm esse problema, mas demandam um tempo computacional muito grande, que pode inviabilizar o cálculo para um número grande de sequências (Yang, 1993, 1994). Um método alternativo e computacionalmente rápido e simples para estimar esse parâmetro foi desenvolvido, evitando o desvio dos métodos de parcimônia (Nei e Kumar, 2000, Felsenstein 2004).

A correção gama apenas adiciona o parâmetro de variação de taxas no cálculo do número de substituições. Assim, podemos incorporar todos os modelos de evolução de nucleotídeos já vistos, como Jukes-Cantor, Kimura 2-parâmetros, Tamura 3-parâmetros etc., assim como o de Poisson, no caso de sequências de aminoácidos. Ou seja, todos os sítios evoluem de acordo com uma determinada distribuição (Poisson, Kimura 2-parâmetros etc.), mas as taxas de evolução variam de sítio para sítio. A fórmula da distância gama-Kimura é dada por

$$d= (\alpha/2)[(1 - 2P - Q)^{-1/\alpha}+1/2(1 - 2Q)^{-1/\alpha}-3/2], \quad (16)$$

com variância,

$$V(d) = [c_1^2P+c_3^2Q-(c_1P+c_3Q)^2]/n, \quad (17)$$

onde

$$c_1 = (1 - 2P - Q)^{-(1/\alpha+1)}, c_2 = (1 - 2Q)^{-(1/\alpha+1)}, c_3 = 1/2(c_1+c_2)$$

e P e Q são os mesmos parâmetros definidos para a fórmula (7). No caso de aminoácidos, podemos usar o modelo gama-Poisson, cuja fórmula é

$$d= \alpha[(1 - p)^{-1/\alpha}-1], \quad (18)$$

com variância

$$V(d) = p[(1-p)^{-(1+2/\alpha)}]/n. \quad (19)$$

Nos dois casos, α é dado pela fórmula (15). A correção para a heterogeneidade de sítios deve ser, em princípio, mais importante para sequências de nucleotídeos que de aminoácidos. No entanto, é importante verificar se o valor de α é baixo, ou seja, menor do que 0,65; nesse caso, a correção gama torna-se necessária (Nei e Kumar, 2000). Entretanto, um estudo mostra que uma correção gama é por vezes necessária mesmo quando o parâmetro gama é maior do que 1.0. Takezaki e Gojobori (1999), por exemplo, sugeriram que a correção gama é fundamental para "corrigir" a árvore filogenética "incorreta", que é reconstruída com todos os métodos quando toda porção codificadora do DNA mitocondrial é usada num grupo de vertebrados para o qual há uma filogenia robusta conhecida (Russo *et al.*, 1996; Takezaki e Gojobori, 1999).

Distância PAM e outras matrizes

No caso de aminoácidos, também existe uma certa tendência de um aminoácido substituir outro mais frequentemente que os demais. Estudos empíricos mostram que as substituições ocorrem mais frequentemente entre aminoácidos com propriedades bioquímicas semelhantes. A probabilidade de substituição de aminoácidos varia de gene para gene dentro de um certo intervalo de tempo (Dickerson, 1971). Esse intervalo não é medido em tempo real (anos, por exemplo), mas sim em PAMs (*Point Accepted Mutation*, em inglês; Dayhoff *et al.*, 1978). Um PAM é o intervalo de tempo necessário para um determinado gene sofrer uma substituição a cada 100 aminoácidos. Portanto, o tempo real vai variar de acordo com a sequência analisada, isto é, se usamos uma sequência que evolui lentamente, um PAM será maior, em termos de anos, do que um PAM de uma sequência que evolui mais rapidamente. Dayhoff *et al.,* (1978) propuseram uma matriz de transição entre aminoácidos onde são mostradas as probabilidades de transição entre os diferentes aminoácidos baseadas em dados empíricos na comparação entre sequências de citocromo **c** em mamíferos.

Por exemplo, a probabilidade de troca de alanina por serina é muito maior (ou seja, foi mais observada na comparação entre as sequências) que a de substituição de alanina por triptofano. Isso porque a substituição de apenas um nucleotídeo (de G para T) pode gerar a troca de alanina por serina, enquanto que a substituição de uma alanina por um triptofano requer a substituição de dois nucleotídeos (GCG tornando-se UGG). Outros parâmetros, como hidrofobicidade, tamanho do aminoácido e acidez/basicidade também são relevantes e devem ser levados em consideração.

Em 1992, Jones *et al.,* desenvolveram um programa de computador (MAKEPET) que permite a construção de uma matriz de transição específica para genes ou grupo de organismos e propuseram a matriz JTT. Ao contrário da matriz PAM, a matriz JTT foi construída usando um número grande de conjunto de dados, retirados dos bancos de dados disponíveis e tem uma aplicabilidade mais geral. A matriz WAG também foi construída com o mesmo princípio (Whelan e Goldman, 2001). Recentemente, um novo método de reconstrução dessas matrizes foi proposto e implementado na construção da matriz LG (Le Quang e Gascuel, 2008), baseada num conjunto maior de dados do que o usado no desenvolvimento das matrizes anteriores.

Por outro lado, a série de matrizes BLOSOM pode ser usada em conjuntos de dados de diversos níveis de substituição (Henikoff e Henikoff, 1992). Essa série é implementada em vários programas de alinhamento múltiplo de sequências codificantes de proteínas. Outras matrizes de substituição de aminoácidos foram propostas para trabalhar com conjuntos específicos de dados, por exemplo a matriz mtREV (genomas mitocondriais, Adachi e Hasegawa, 1996), a mtART (genomas mitocondriais de artrópodos, Abascal *et al.*, 2007), a MtMam (genomas mitocondriais de mamíferos, Yang *et al.*, 1998), a rtREV (proteínas retrovirais, Dimmic *et al.*, 2002) e cpREV (proteínas de cloroplasto Adachi *et al.*, 2000).

Afinal, que distância escolher?

Atualmente há um número muito grande de métodos disponíveis para o cálculo de distâncias entre sequências de DNA ou aminoácidos. Por razões práticas, apresentamos aqui aqueles mais frequentemente usados na reconstrução de filogenias e disponíveis na maioria dos programas mais usados em análise filogenética.

Descreveremos, a seguir, uma abordagem geral que pode servir como regra para a escolha do método de distância mais adequado a um conjunto específico de dados. Um ponto-chave é que deve estar claro *a priori* se o problema filogenético envolve apenas a recuperação de topologias (isto é, esclarecer a história filogenética dos organismos ou o padrão de duplicação entre as sequências) ou se o conhecimento do tamanho de ramos também

Tabela 13.2. Estimativa de erros-padrões da distância de Poisson para um grupo de vertebrados, métodos analíticos (abaixo da diagonal) e *bootstrap* (acima da diagonal).

	Humano	Cavalo	Vaca	Canguru	Salamandra	Carpa
Humano		0,031	0,031	0,039	0,078	0,083
Cavalo	0,031		0,030	0,043	0,083	0,081
Vaca	0,031	0,031		0,038	0,080	0,079
Canguru	0,040	0,043	0,039		0,081	0,084
Salamandra	0,074	0,080	0,076	0,080		0,090
Carpa	0,082	0,081	0,079	0,086	0,089	

é necessário (isto é, estimar o tempo de divergência entre as sequências). Na realidade, esses dois problemas podem ser abordados independentemente. Por exemplo, podemos usar um modelo de distância mais simples e com menor variância (por exemplo, distância **p** ou Jukes-Cantor) para reconstruir a topologia e um modelo mais complexo, ou seja, não enviesado para a estimativa dos tamanhos de ramos (por exemplo, Tamura e Nei, 1993). Entretanto, infelizmente, os programas de análise filogenética ainda não realizam esse tipo de análise em mosaico.

Para a estimativa de topologias, devemos ser bastante cuidadosos quanto ao número de parâmetros usados nos modelos, já que a variância associada à estimativa de distância é diretamente proporcional ao número de parâmetros do modelo (Kelchner e Thomas, 2006). Nesse caso, devemos empregar a distância **p** entre sequências quando a distância média entre elas for menor ou igual a 0,2. Caso a distância média seja maior que 0,2, devemos corrigir o número de substituições pelo modelo adequado, estimando agora outros parâmetros, como o conteúdo de GC, a razão entre transições e transversões, e a heterogeneidade da taxa de substituição entre os sítios. Da mesma forma, devemos usar a distância **p** se as taxas de substituição não variarem entre as linhagens. Isso pode ser testado através de uma árvore linearizada, como descrito adiante.

Por outro lado, se houver interesse na estimativa dos tempos de divergência entre as linhagens ou dos eventos de duplicação entre genes parálogos, devemos usar uma distância não viciada, ou seja, a média da distância estimada (não levando em consideração a variância da estimativa) deve ser bem próxima da distância real. Nesse caso, mesmo que **p** seja menor do que 0,2, devemos corrigir o número de substituições pelo modelo adequado ao padrão de evolução das sequências.

Uma maneira alternativa de estimar a variância associada às distâncias é o *bootstrap*, usando reamostragens com reposição dos dados para tal fim (Nei e Kumar 2000). A Tabela 13.2 mostra a quase-identidade entre as variâncias da distância de Poisson, estimadas através de *boostrap* e da fórmula (6), para um grupo de vertebrados. Esses dados indicam que esse é um bom método para estimar o erro-padrão de distâncias evolutivas.

Mais recentemente, outros modelos mais e mais complexos (por exemplo, GTR; Tavaré, 1986) estão sendo usados em análises filogenéticas de diversos grupos de organismos. Entretanto, tais modelos não são apropriados para serem usados em métodos de distância, considerando sua vulnerabilidade a altas variâncias (Nei & Kumar, 2000; Felsenstein, 2004). Por conta disso, na escolha do modelo de substituição para métodos de distância, os programas ModelTest (Posada e Crandall, 1998) e jModelTest (Posada, 2008) não são indicados (ver também Kelchner e Thomas, 2006).

13.3. Algoritmos para Reconstrução de Topologias

Após a construção da matriz de distâncias par a par utilizando qualquer dos modelos descritos acima, temos que escolher um algoritmo de reconstrução da árvore propriamente dita. Assim como no caso das distâncias, existem diversos algoritmos para a reconstrução de topologias. Nesta seção, iremos descrever apenas aqueles mais frequentemente usados na literatura. É importante notar que, se um determinado método foi utilizado para o cálculo das distâncias, este não tem influência no algoritmo a ser empregado posteriormente, mas certamente pode influenciar os tamanhos dos ramos dentro da topologia ou a própria topologia.

Os algoritmos para a construção da árvore procuram acomodar, na topologia, as distâncias calculadas entre as sequências macromoleculares constantes da matriz. Na Figura 13.1, por exemplo, estão representadas duas matrizes de distância e as topologias correspondentes, onde as distâncias da matriz estão perfeitamente acomodadas. Na Figura 13.1, a árvore do lado esquerdo (I) é ultramétrica, ou seja, existe um ponto equidistante de todas as pontas (terminais), apresentando um soma de distância dos ramos de 96 unidades arbitrárias. Nesse caso, tal ponto é inferido como a raiz da árvore. No caso da árvore do lado direito (II), apesar de todas as distâncias na matriz correspondente estarem perfeitamente acomodadas nos tamanhos dos ramos, não existe um ponto equidistante a todos os terminais. Essa árvore não é ultramétrica, mas aditiva, ou seja, as distâncias estão completamente acomodadas, assim como na árvore I, que também é aditiva.

Alguns algoritmos, tal como o UPGMA, descrito abaixo, somente produzem árvores ultramétricas. Se as distâncias na matriz estiverem perfeitamente acomodadas na árvore, não há qualquer problema. Para verificar isso, pode-se calcular o coeficiente da correlação cofenética (Sokal e Rohlf, 1962), que é a correlação entre as distâncias na matriz e as distâncias entre cada par de terminais na árvore. Valores altos, como 0,98, significam que a acomodação é satisfatória. Caso contrário, é necessário o emprego de outro algoritmo que não produza somente árvores ultramétricas.

O coeficiente de correlação cofenética (c) é calculado pela fórmula

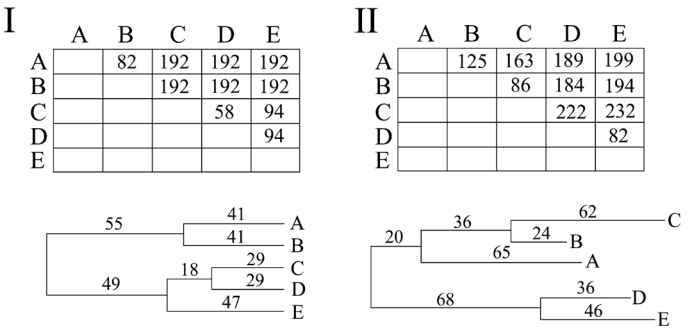


Figura 13.1. Matrizes de distâncias (arbitrárias) e suas respectivas árvores (dendrogramas). I. Árvore ultramétrica, onde existe um ponto equidistante dos terminais (a soma é sempre 96). II. Árvore não ultramétrica, onde não é possível localizar um ponto contido na árvore que seja equidistante dos terminais. Ambas árvores apresentam a propriedade de aditividade, ou seja, as distâncias na matriz correspondem exatamente à soma dos tamanhos de ramos nas árvores entre quaisquer dois terminais.

$$c = \frac{\sum_{i<j}(x_{i,j} - x) \times (t_{i,j} - t)}{\sqrt{\left[\sum_{i<j}(x_{i<j} - x)^2\right] \times \left[\sum_{i<j}(t_{i,j} - t)^2\right]}},$$

onde $x_{i,j}$ é a distância na matriz, calculada entre as sequências i e j, $t_{i,j}$ é a soma dos tamanhos dos ramos na árvore, que estão entre as sequências i e j, x é a média das distâncias $x_{i,j}$ e t é a média das distâncias $t_{i,j}$.

UPGMA (*unweighted pair group method with arithmetic means*, em inglês) (Sneath e Sokal, 1973)

De todos os algoritmos de reconstrução filogenética, UPGMA é o mais simples, de mais fácil entendimento e, por essas razões, foi amplamente utilizado. Apesar dessas vantagens, esse algoritmo assume o relógio molecular para a construção da topologia, ou seja, assume que todas as linhagens evoluem a uma taxa constante de evolução e produz sempre árvores ultramétricas. O algoritmo agrupa o par de OTUs com a menor distância (por exemplo, OTUs 1 e 2) e recalcula a matriz de distância assumindo 1 e 2 como uma única OTU composta 12. Note que, nesse caso, as células da nova matriz de distância relativas à OTU composta serão as médias das células individuais da matriz antiga—a distância de $d_{12\text{-}3}$ será $(d_{1\text{-}3}+d_{2\text{-}3})/2$. Em seguida, com base na nova matriz de dados, a célula com menor distância será incluída na árvore e assim consecutivamente até que todas as OTUs estejam agrupadas na árvore final. Isso significa que o agrupamento de UPGMA é feito com base em similaridade total, assumindo taxas regulares de substituição (ou seja, o relógio molecular). Com isso, a árvore final já está enraizada sem a necessidade de um grupo externo, diferentemente de todos os demais algoritmos de reconstrução filogenética, que descreveremos a seguir.

Existe muita resistência ao método UPGMA justamente por basear-se em similaridade total e, dessa forma, sua acurácia depende de que a taxa de substituição se mantenha constante por toda a filogenia. Entretanto, sua eficiência na reconstrução de árvores "verdadeiras" não é tão baixa assim. Na verdade, em casos onde as taxas de substituição não variam muito de linhagem para linhagem, esse método pode ter desempenho tão bom ou até melhor do que os outros, que não assumem taxas constantes (Nei e Takezaki, 1994; Pollock e Goldstein, 1995). A grande vantagem desse método é a facilidade e, portanto, a rapidez com que a árvore final é construída. Para construir uma árvore com 120 OTUs e com 400 pares de bases, por exemplo, um computador Pentium II 400Mhz leva apenas cerca de 5 segundos para processar essa análise.

O algoritmo UPGMA foi um dos primeiros a ser utilizado na construção de topologias e, como é um método de distância, algumas pessoas erroneamente assumem que todos os métodos de distância são dependentes do pressuposto de taxas constantes de substituição (Stewart, 1993), invocando o uso de parcimônia como alternativa. Na verdade, outros algoritmos de distância, como o agrupamento de vizinhos (descrito a seguir), são menos dependentes desse pressuposto para um bom desempenho do relógio molecular do que os métodos de parcimônia (*e.g.*, Saitou e Imanishi, 1989; Takezaki e Nei, 1994).

Agrupamento de vizinhos (*neighbor-joining*, NJ; Saitou e Nei, 1987)

Esse método faz parte do grupo de métodos de evolução mínima (Cavalli-Sforza e Edwards, 1967; Saitou e Imanishi, 1989), no qual se estima a árvore com a menor soma total de ramos. No entanto, o método de evolução mínima descrito originalmente, do ponto de vista do tempo computacional, era muito demorado, uma vez que todas as árvores possíveis tinham que ser construídas e a soma total de seus ramos estimada, para que a árvore com soma mínima fosse encontrada. Quando o número de OTUs é maior que 30, o número possível de árvores é muito grande (veja Tabela 12.1, no Capítulo 12) e, com isso, o tempo requerido é impraticável. Portanto, um algoritmo heurístico para a construção da árvore de evolução mínima foi proposto. Esse algoritmo heurístico foi chamado de agrupamento de vizinhos, onde a evolução mínima está implícita a cada passo do algoritmo, produzindo uma única árvore final.

Nesse caso, começamos com uma árvore em forma de estrela, ou seja, sem qualquer resolução. Em seguida, procuramos o par de vizinhos que minimize a soma total dos ramos da árvore. Para isso, temos que calcular a somatória dos ramos para árvores politômicas cujo único ramo interno (ligando os nós X e Y) é separado por um par de OTUs. Calculamos a somatória para todos os pares de OTUs (S_{12}, S_{13}, S_{14}, S_{15}, S_{16}, S_{17}, S_{18}, S_{23}, S_{24}...) e escolhemos o par cuja somatória dos ramos é a menor (no exemplo da Figura 13.1, S_{12}). Esse par (1,2) será a primeira bifurcação ou resolução da árvore e, daqui por diante, será tratado como uma única OTU para o cálculo da topologia [((1, 2), 3, 4, 5, 6, 7, 8)]. A fórmula para o cálculo do somatório do tamanho de ramos (S_{12}) é

$$S_{12} = L_{1Y} + L_{2Y} + L_{XY} + \sum_{i=3}^{N} L_{iY},$$

onde

$$L_{XY} = \frac{1}{2(N-2)}\left[\sum_{k=3}^{N}(D_{1k} + D_{2k}) - (N-2)(L_{1X} + L_{2X}) - 2\sum_{i=3}^{N} L\right.$$

$$L_{1X} + L_{2X} = D_{12} \quad \text{e}$$

$$\sum_{i=3}^{N} L_{iY} = \frac{1}{N-3}\sum_{3 \le i \le j} D_{ij}.$$

Depois disso, procuramos o segundo par que minimize a soma total dos ramos da árvore—((1, 2), (3, 4), 5, 6, 7, 8). Alternativamente, a menor soma total dos ramos pode ser encontrada quando unimos uma única OTU (3) ao primeiro par já resolvido—(((1, 2), 3), 4, 5, 6, 7, 8). Seguiremos esse processo de unir OTUs até que a árvore esteja totalmente resolvida—((((((1, 2), (3,4)), 5), 6), 7), 8) ou (((((((1, 2), 3), 4), 5), 6), 7), 8) etc.—, ou seja, a cada passo, procuram-se os vizinhos (*neighbors*) que minimizem a soma dos ramos da árvore. Quando duas somatórias de ramos se igualam, a escolha dos ramos que serão unidos depende do programa empregado. Alguns programas escolhem sempre a topologia onde a primeira OTU apresentada no arquivo de dados é separada da politomia, enquanto que outros programas escolhem ao acaso entre as duas ou mais topologias possíveis. Nesses casos, dois processamentos do programa para o mesmo conjunto de dados podem resultar em árvores distintas. A consequência imediata e óbvia desse procedimento é que uma árvore de *neighbor-joining* é sempre totalmente resolvida, ou seja, nunca vai apresentar politomias, por mais curtos que sejam os ramos, podendo até assumir valores negativos. O cálculo dos ramos internos para árvores de *neighbor-joining* é descrito a seguir.

Esse algoritmo é um pouco mais lento que UPGMA (para os mesmos dados e computador do exemplo anterior, a árvore é construída em 10 segundos, o dobro do tempo), porém tem a gran-

de vantagem de não depender de que as taxas de substituição sejam constantes (ver abaixo). Além disso, simulações computacionais mostram que esse algoritmo é bastante eficaz na reconstrução da topologia correta, como veremos mais adiante (Nei, 1991; Rzhetsky e Nei, 1992; Tateno *et al.*, 1994; Huelsenbeck, 1995).

Evolução mínima (*minimum evolution*, ME; Rzhetsky e Nei, 1992)

Como já mencionamos, o algoritmo de evolução mínima procura a árvore de menor soma dos ramos (Cavalli-Sforza e Edwards, 1967; Saitou e Imanishi, 1989). Na prática, no entanto, o cálculo de todas as árvores possíveis é muito laborioso e não é necessário, já que a árvore de ME geralmente é a árvore de NJ ou alguma muito próxima a ela (Rzhetsky e Nei, 1992). Esse método foi, então, modificado.

No novo algoritmo de evolução mínima, começamos com a árvore de NJ e, a partir dela, procuramos, nas árvores mais próximas (um ou dois rearranjos de ramos internos), aquelas que tenham soma total de ramos menor que a árvore de NJ. Simulações de computador mostram que, se usarmos distâncias não viciadas (ou seja, em que a média da distância estimada seja igual à distância verdadeira), a árvore ME é a árvore verdadeira (Rzhetsky e Nei, 1993). Portanto, esse algoritmo possui uma base teórica forte. Na prática, no entanto, esse método geralmente recupera a árvore de NJ (Rzhetsky e Nei, 1992).

Outros métodos de distância

O algoritmo de Fitch-Margoliash é semelhante àqueles do grupo de evolução mínima, mas, nesse caso, a árvore final é aquela com a menor soma dos quadrados das diferenças entre a distância estimada—por exemplo Poisson, Jukes, Cantor etc.—e o somatório dos comprimentos de ramos entre as respectivas OTUs (Fitch e Margoliash, 1967). Esse método caiu em desuso recentemente, já que agrupamento de vizinhos se tem mostrado superior em simulações de computador.

Recentemente, métodos de reconstrução filogenética como variações de agrupamento de vizinhos foram propostos, por exemplo o BIONJ, o agrupamento de vizinhos ponderado e o NJML (Gascuel, 1997; Bruno *et al.*, 2000; Ota e Li, 2000). Esses algoritmos são híbridos de outros já estabelecidos e de grande sucesso. No entanto, ainda não está claro quão eficientes esses algoritmos são em estudos de simulação mais abrangentes, em relação àquelas dos artigos originais.

Tamanhos de Ramo

O cálculo dos tamanhos de ramos (Fitch e Margoliash, 1967) pode ser feito com base nas distâncias entre as sequências. Consideraremos uma topologia com três OTUs para facilitar a ilustração—para o caso de mais de três OTUs, juntamos as restantes como C, em uma OTU composta, e a decompomos passo a passo. Nossos dados de distância são três valores conhecidos (d_{12}, d_{13} e d_{23}) e as três incógnitas são os tamanhos dos ramos dessa topologia (x, y e z), onde a relação entre eles é

$$d_{12} = x + y, \quad d_{13} = x + z, \quad d_{23} = y + z \qquad (20)$$

e, portanto,

$$x=(d_{12}+d_{13}-d_{23}/2;\ y=(d_{12}+d_{23}-d_{13})/2;\ z=(d_{13}+d_{23}-d_{12})/2. \quad (21)$$

Árvore linearizada (Takezaki *et al.*, 1995)

O algoritmo da árvore linearizada usa uma topologia previamente construída com algum método de reconstrução filogenética de distância que não assuma o relógio molecular (*neighbor-joining*, evolução mínima etc.) e testa se a diferença entre os ramos-irmãos de cada nó da árvore é significativamente diferente de zero. A vantagem de usarmos esse algoritmo é justamente a estimativa do tempo de divergência entre as sequências, sem a necessidade de assumirmos *a priori* o relógio molecular. Por exemplo, para a estimativa do tempo de divergência entre as sequências de Adh (desidrogenase alcoólica) e Adhr (relacionado à desidrogenase alcoólica) de drosofilídeos (Russo *et al.*, 1995), os autores usaram uma árvore linearizada de acordo com o método descrito por Takezaki e colaboradores (1995). Nesse método, iniciou-se a análise a partir de uma topologia conhecida ou bem confiável, já que não teria sentido a estimativa de tempo com base em uma topologia com pouco suporte. Essa topologia deve ser enraizada, ou seja, o grupo externo tem que ser conhecido. Naquele artigo, os autores estimaram a árvore filogenética de drosofilídeos pelo método de NJ com a distância Kimura 2-parâmetros para todas as posições do códon, já que a razão entre o número de transições e transversões era muito diferente do esperado ao acaso.

Para o teste do relógio molecular, precisamos descobrir se a diferença entre o comprimento dos ramos dos grupos ou sequências-irmãs é significativamente diferente de zero. Isto é feito pelo teste z com base na variância e covariância do tamanho dos ramos (Takezaki *et al.*, 1995), o que, por sua vez, têm como base o modelo de distância que foi assumido. Depois de eliminados os ramos que desviaram significativamente, o próximo passo é reestimar o tamanho dos ramos, com uma medida na qual os sítios evoluem de uma maneira razoavelmente linear, assumindo uma taxa constante de substituição. Nesse caso, vamos considerar que o comprimento do ramo de um determinado par de OTUs será a metade da distância entre essas duas OTUs. Isso pode ser feito continuamente, desde os ramos de divergência mais recente até aqueles com divergência maior. A única restrição ocorre quando o comprimento de um agrupamento mais interior (divergência mais antiga) é menor do que aquele imediatamente exterior (divergência mais recente). Nesses casos, assume-se que essa diferença é negligenciável (zero), resultando em uma politomia na árvore final.

Modelos de distância podem variar de caso para caso. Por exemplo, no caso da estimativa do tempo de divergência em drosofilídeos com Adh, Russo *et al.,* (1995) usaram somente a terceira posição do códon (uma vez que essa é menos influenciada pela seleção negativa) com a correção Tajima-Nei (uma vez que existe heterogeneidade das frequências de nucleotídeos na terceira posição do códon entre as sequências). Por outro lado, quando, nesse mesmo estudo, o tempo de divergência entre os dois genes foi estimado, os autores concluíram que os genes se originaram de uma duplicação muito anterior à divergência dos grandes grupos de drosofilídeos (Adh e Adhr). Nesse caso, eles usaram apenas as duas primeiras posições do códon, com a correção Kimura 2-parâmetros, já que a terceira posição do códon estava saturada de substituições. Concluindo, apesar de o gene Adhr não ter sido encontrado no subgênero *Drosophila*, provavelmente essas espécies perderam essa cópia do gene em algum momento de sua história evolutiva.

13.4. Testes de Confiança em Topologias

Bootstrap

O *bootstrap* é o teste de confiança em topologia mais usado tanto em árvores de distância, como de parcimônia e até mesmo de verossimilhança. A base do método consiste de uma simples reamostragem, com reposição pseudoaleatória

dos dados. Em cada reamostragem, o número total de dados amostrados mantém-se constante, que é o total de sítios do alinhamento original. Como cada posição de nucleotídeo tem a mesma probabilidade de ser amostrada, algumas posições podem ser consideradas mais de uma vez, enquanto outras podem não ser amostradas. A cada reamostragem, uma árvore-réplica é construída. Em geral, 500 ou 1000 árvores réplicas são produzidas. Na versão original de Felsenstein (1985), a árvore com os valores de *bootstrap* mostra a topologia de consenso de todas as (500 ou 1000) árvores-réplica (usada no programa PAUP), o qual, portanto, pode ser diferente da árvore original. Na segunda versão, a topologia da árvore final é exatamente a mesma da árvore original (com 100% dos dados), adicionando apenas as percentagens com que cada grupamento apareceu nas réplicas (implementada no programa MEGA).

Um teste de *bootstrap* no programa MEGA em uma sequência de 15 nucleotídeos (de 1 a 15), por exemplo, seria realizado da seguinte maneira. Em primeiro lugar, uma árvore usando as 15 posições seria construída—essa é a árvore original. Como estamos trabalhando com o programa MEGA, essa também é a árvore final. Em seguida, passamos ao teste de *bootstrap*, propriamente dito, para a construção da primeira árvore-réplica, quando o algoritmo irá selecionar 15 posições de nucleotídeos. No entanto, lembre-se que, como essa seleção é com reposição, alguns sítios serão amostrados mais de uma vez, enquanto que outros serão retirados dessa réplica. Esse procedimento é repetido por várias vezes (por exemplo, 1000 vezes), construindo uma árvore-réplica a cada ciclo, cada uma baseada em um conjunto de dados diferente. Por exemplo, assumindo que, na primeira réplica, as posições sorteadas foram 12, 6, 5, 6, 9, 12, 3, 2, 5, 3, 10, 13, 7, 2 e 13, a primeira árvore-réplica será baseada nesses dados. Digamos que, na segunda réplica, as posições sorteadas foram 5, 8, 10, 5, 15, 9, 5, 14, 3, 4, 4, 4, 9, 4 e 8, e a segunda árvore-réplica será construída. No final das 1000 réplicas, o teste compara cada uma das árvores-réplicas com a árvore original e a proporção de *bootstrap* é simplesmente a porcentagem de vezes que o mesmo grupamento original foi recuperado nas árvores-réplicas.

A técnica de *boostrap* revela a consistência interna dos dados, ou seja, se a topologia muda muito conforme a reamostragem dos dados, o valor do *bootstrap* será menor e, portanto, menor a segurança que teremos nessa topologia. Teoricamente, estudos de simulações mostram que essa confiança do *bootstrap* depende da consistência do método que usamos na própria construção da topologia. Portanto, se estivermos usando um determinado método que é inconsistente para uma dada faixa de parâmetros, os valores de *bootstrap* até podem ser altos, mas isso não é um reflexo real da sustentação da topologia (Hillis e Bull, 1993). Em casos de teste usando filogenias bem estabelecidas (Russo, 1997), o *bootstrap* também reflete bem a consistência dos ramos internos.

Em outros casos, quando usamos parâmetros em que o método é consistente, o valor da probabilidade do *bootstrap* geralmente é uma subestimativa da probabilidade daquele determinado ramo ser "verdadeiro" (Hillis e Bull, 1993; Sitnikova *et al.*, 1995). Isso se deve ao fato de que, mesmo sendo "verdadeiro" por acaso, pode haver sítios que endossem outro tipo de partição. A subestimativa aumenta com o aumento do número de OTUs na filogenia (Sanderson e Wojciechowsky, 2000).

Nível de confiança

Outra técnica desenvolvida para revelar a confiabilidade de um determinado nó, é o nível de confiança ou probabilidade de confiança (CP; Rzhetsky e Nei, 1992). Essa técnica utiliza o erro padrão do ramo anterior ao nó analisado e é definida como

$$CP = (1 - a),$$

onde **a** é o erro associado à hipótese de aquele ramo interno ter comprimento maior do que zero. Essa técnica, diferentemente do *bootstrap*, é eficiente quando usamos o modelo adequado para o cálculo da distância e não é tão dependente do método para a construção da topologia. Por outro lado, se usarmos o modelo de Jukes-Cantor, quando na verdade a evolução da sequência segue o modelo de Kimura-2 parâmetros, teremos uma subestimativa da variância e, consequentemente, uma superestimativa da confiança para aquele nó (Rzhetsky e Nei, 1992). Se usarmos uma estimativa de distância que tenha variância muito grande, como Gama-Tamura-Nei, quaisquer das árvores obtidas não serão significativamente diferentes de uma árvore totalmente politômica.

Um dado interessante é que, em estudos com filogenias bem estabelecidas (Russo, 1997), esse teste se mostra menos conservador que o *bootstrap* (i.e., baixo erro do tipo II), sem aumentar o erro do tipo I associado ao teste, indicando ser um teste poderoso e consistente.

Referências Bibliográficas

Abascal, F., Posada, D e Zardoya, R. (2006). MtArt: a new model of amino acid replacement for Arthropoda. **Mol. Biol. Evol. 24**:1-5.

Adachi, J. e Hasegawa, M. (1996). Model of amino acid substitution in proteins encoded by mitochondrial DNA. **J. Mol. Evol. 42**: 459-468.

Adachi, J., Waddell, P.J., Martin, W. e Hasegawa, M. (2000). Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. **J. Mol. Evol. 50**: 348-358.

Bruno, W.J., Socci, N.D. e Halpern, A.L. (2000). Weighted neighbor-joining: a likelihood based approach to distance-based phylogeny reconstruction. **Mol. Biol. Evol. 17**: 189-197.

Cavalli-Sforza, L.L. e Edwards, A.W.F. (1967). Phylogenetic analysis: models and estimation procedures. **Am. J. Hum. Genet. 19**: 233-257.

Dayhoff, M.O., Schwartz, R.M. e Court, B.C. (1978). A model of evolutionary change in proteins. **In Atlas of protein sequence and supplement**, Vol 5, Supl. 3.

Dickerson, R.E. (1971). The structure of cytochrome c and the rates of molecular evolution. **J. Mol. Evol. 1**: 26-45.

Excoffier, L. e Yang, Z.H. (1999). Substitution rates variation among sites in mitochondrial hypervariable region I of humans and chimpanzees. **Mol. Biol. Evol. 16**:1357-1368.

Felsenstein, J. (1985). Confidence limits on phylogenies: na approaching using the bootstrap. **Evolution 39**:783-791.

Felsenstein, J. (1995). **PHYLIP: phylogeny inference package.** Version 3.5c. Distributed by the author, Department of Genetics, University of Washington, Seattle.

Felsenstein, J. (2004). **Inferring phylogenies**. Sinauer Assoc., Sunderland.

Fitch, W.M. e Margoliash, E. (1967). Construction of phylogenetic trees. A method based on mutation distances as estimated from cytochrome c sequences is of general applicability. **Science 155**: 279-284.

Gascuel, O. (1997). BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. **Mol. Biol. Evol. 14**: 685-695.

Henikoff, S. e Henikoff, J.G. (1992). Amino acid substitution matrices from protein blocks. **Proc. Natl. Acad. Sci. USA 89**: 10915-10919.

Hillis, D.M. e Bull, J.J. (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. **Syst. Biol. 42**: 182-192.

Huelsenbeck, J.P. (1995). The performance of phylogenetic methods in the four-taxon case. **Syst. Biol. 44**: 17-48.

Jones, D.T., Taylor, W.R. e Thornton, J.M. (1992). The rapid generation of mutation data matrices from protein sequences. **CABIOS 8**: 275-282.

Jukes, T.H. e Cantor, C.R. (1969). Evolution of protein molecules. **In** Munro, H.N. (ed). **Mammalian protein metabolism.** Volume 3. Academic Press, New York, pp. 21-132.

Kelchner, S.A. e Thomas, M.A. (2006) Model use in phylogenetics: nine key questions. **Trends Ecol. Evol. 22**: 87-94.

Kimura, M. (1980). A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. **J. Mol. Evol. 16**: 111-120.

Kumar, S., Tamura, K. e Nei, M. (1993). **MEGA: molecular evolutionary genetics analysis, version 1.0.** Institute of Molecular Evolution-

ary Genetics, The Pennsylvania State University, University Park, Pennsylvania.

Le Quang, S. e Gascuel, O. (2008). An improved general amino acid replacement matrix. **Mol. Biol. Evol.** (on-line)

Nei, M. (1987). **Molecular evolutionary genetics**. Columbia University Press, New York.

Nei, M. (1991). Relative efficiencies of different tree-making methods for molecular data. **In** M. M. Miyamoto e J. Cracraft (eds.). **Phylogenetic analysis of DNA sequences**. Oxford University Press, New York, pp. 90-128.

Nei, M. e Kumar, S. (2000). **Molecular evolution and phylogenetics**. Oxford University Press, New Iorque.

Nei, M. e Takezaki, N. (1994). Estimation of genetic distances and phylogenetic trees from DNA analysis. **Proc. 5th World Cong. Genet. Appl. Livestock Prod. 21**: 405-412.

Nei, M., Takezaki, N. e Sitnikova, T. (1995). Assessing molecular phylogenies. **Science 267**: 253-255.

Ota, S. e Li, W.-H. (2000). NJML: a hybrid algorithm for the neighbor-joining and maximum-likelihood methods. **Mol. Biol. Evol. 17**: 1401-1409.

Pollock, D.D. e Goldstein, D.B. (1995). A comparison of two methods for constructing evolutionary distances from a weighted contribution of transition and transversion differences. **Mol. Biol. Evol. 12**: 713-717.

Posada, D. (2008). jModelTest: Phylogenetic model averaging. Mol. Biol. Evol. (on-line)

Posada, D. e Crandall, K. A. (1998). MODELTEST: testing the model of DNA substitution. Bioinformatics 14(9):817-818.

Russo, C.A.M. (1997). Efficiencies of different statistical tests in supporting a known vertebrate phylogeny. **Mol. Biol. Evol. 14**: 1078-1080.

Russo, C.A.M., Takezaki, N. e Nei, M. (1995). Molecular phylogeny and divergence times of drosophilid species. **Mol. Biol. Evol. 12**: 391-404.

Russo, C.A.M., Takezaki, N. e Nei, M. (1996). Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny. **Mol. Biol. Evol. 13**: 525-536.

Rzhetsky, A. e Nei, M. (1992). A simple method for estimating and testing minimum-evolution trees. **Mol. Biol. Evol. 9**: 945-967.

Rzhetsky, A. e Nei, M. (1993). Theoretical foudation of the minimum-evolution method of phylogenetic inference. **Mol. Biol. Evol. 10**: 1073-1095.

Rzhetsky, A. e Sitnikova, T. (1996). When is it safe to use an oversimplified substitution model in tree-making? **Mol. Biol. Evol. 13**: 1255-1265

Stewart, C.-B. (1993). The powers and pitfalls of parsimony. **Nature 361**: 603-607.

Saitou, N. e Imanishi, M. (1989). Relative efficiencies of the Fitch-Margoliash, maximum-parsimony, maximum-likelihood, minimum-evolution and neighbor-joining methods of phylogenetic tree construction in obtaining the correct tree. **Mol. Biol. Evol. 6**: 514-525.

Saitou, N. e Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. **Mol. Biol. Evol. 4**: 406-425.

Sanderson, M.J. e Wojciechowsky, M.F. (2000). Improved bootstrap confidence limits in large scale phylogenies, with an example from neo-astragalus (Leguminosae). **Syst. Biol. 49**: 671-685.

Sitnikova, T., Rhzetsky, A. e Nei, M. (1995). Interior-branch and bootstrap tests of phylogenetic trees. **Mol. Bio. Evol. 12**: 319-333.

Sneath, P.H.A. e Sokal, R.R. (1973). **Numerical taxonomy**. W.H. Freeman, San Francisco.

Sokal, R.R. e Rohlf, F.J. (1962). The comparison of dendrograms by objective methods. **Taxon 11**:33-40.

Sullivan, J e Joyce, P. (2005). Model selection in phylogenetics. **Annu. Rev. Ecol. Syst. 36**: 445-466.

Tajima, F. (1993). Simple methods for testing the molecular evolutionary clock hypothesis. **Genetics 135**: 599-607.

Tajima, F. e Nei, M. (1984). Estimation of evolutionary distance between nucleotide sequences. **Mol. Biol. Evol. 1**: 269-285.

Tajima, F. e Takezaki, N. (1994). Estimation of evolutionary distance between nucleotide sequences. **Mol. Biol. Evol. 10**: 677-688.

Takezaki, N. e Gojobori, T. (1999). Correct and incorrect vertebrate phylogenies obtained by the entire mitochondrial DNA sequences. **Mol. Biol. Evol. 16**: 590-601.

Takezaki, N. e Nei, M. (1994). Inconsistency of the maximum parsimony method when the rate of nucleotide substitution is constant. **J. Mol. Evol. 39**: 210-218.

Takezaki, N., Rzhetsky, A. e Nei, M. (1995). Phylogenetic test of the molecular clock and linearized trees. **Mol. Biol. Evol. 12**: 823-833.

Tamura, K. (1992). Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C content bias. **Mol. Biol. Evol. 9**: 678-687.

Tamura, K. e Nei, M. (1993). Estimation of the number of nucleotide substitution in the control region of mitochondrial DNA in humans and chimpanzees. **Mol. Biol. Evol. 10**: 512-526.

Tateno, Y., Takezaki, N. e Nei, M. (1994). Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. **Mol. Biol. Evol. 11**: 261-277.

Tavaré, S. (1986). Some probabilistic and statistical problems on the analysis of DNA sequences. Pp. 57-86 em Lectures in mathematics in the life sciences 17.

Uzzel, T. e Corbin, K.W. (1971). Fitting discrete probability distributions to evolutionary events. **Science 172**: 1089-1096.

Whelan, S. e Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. **Mol. Biol. Evol. 18**: 691-699.

Yang, Z.H. (1993). Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. **Mol. Biol. Evol. 10**:1396-1401.

Yang, Z.H. (1994). Estimating the pattern of nucleotide substitution. **J. Mol. Evol. 39**: 105-111.

Yang, Z.H., Nielsen, R. e Hasegawa, M. (1998); Models of amino acid substitution and applications to mitochondrial protein evolution. **Mol. Biol. Evol. 15**:1600–1611.

Página deixada em branco

Capítulo 14

Reconstrução filogenética: Métodos probabilísticos

Dr. Sergio Luiz Pereira (sergiolp@gmail.com)
The Centre for Applied Genomics
The Hospital for Sick Children

Cristina Yumi Miyaki (cymiyaki@ib.usp.br)
Departamento de Genética e Biologia Evolutiva
Instituto de Biociências
Universidade de São Paulo

Cláudia A. de Moraes Russo (claudia@biologia.ufrj.br)
Departamento de Genética
Instituto de Biologia
Universidade Federal do Rio de Janeiro

"*A sua inteligência (...) guiava-o nesse raciocínio (...); previa todas as hipóteses, combinava todas as probabilidades, e preparava-se para realizar o seu plano com a certeza e a energia de ação que ninguém possuía em grau tão elevado.*" (José de Alencar, *O Guarani*)

14.1. Breve Histórico

A máxima verossimilhança (em inglês, *maximum-likelihood*, ML) é o método de inferência filogenética com a base teórica mais estreitamente ligada à teoria de máxima verossimilhança de R. A. Fisher, o matemático inglês que impulsionou um grande desenvolvimento na estatística e na genética de populações no início do século XX. Quando aplicado à filogenia molecular, esse método requer um modelo probabilístico de evolução de caracteres e pode levar em consideração parâmetros como taxa de substituição entre pares de nucleotídeos em uma sequência de DNA, frequência de bases, proporção de sítios invariáveis e heterogeneidade de taxas de substituição entre sítios (Nei e Kumar, 2000). Os parâmetros podem ser estimados diretamente dos dados e usados no cálculo da probabilidade de um estado de caráter ser substituído por outro, de acordo com o modelo evolutivo escolhido. A árvore final é aquela com a maior verossimilhança (maior probabilidade de que aqueles resultados tenham sido originados de acordo com o modelo evolutivo selecionado), dado um determinado alinhamento (de nucleotídeos ou aminoácidos, por exemplo).

Os primeiros pesquisadores a utilizarem o método de máxima verossimilhança em reconstrução de árvores filogenéticas a partir de dados de frequências gênicas foram Cavalli-Sforza e Edwards, em 1967. Posteriormente, Felsenstein (1973, 1981) desenvolveu algoritmos para que sequências de ácidos nucléicos pudessem ser usadas para estimar histórias evolutivas. Na década de 90, Kishino e colaboradores (1990) aplicaram o método a sequências de proteínas. Os mesmos princípios de reconstrução filogenética aplicam-se a sequências de nucleotídeos dos ácidos nucléicos e a sequências de aminoácidos de uma proteína. Nesse caso, temos 20 diferentes caracteres, que correspondem aos 20 aminoácidos, ao invés de apenas quatro caracteres, correspondentes a cada uma das bases de uma sequência de ácido nucléico. Neste capítulo, consideraremos principalmente o uso de sequências de nucleotídeos para inferências de filogenias moleculares.

O método original descrito por Fisher era um pouco diferente daquele utilizado atualmente em reconstruções filogenéticas. O método original visava a estimativa de um único parâmetro e, para isso, estimava a probabilidade associada a cada valor do parâmetro, escolhendo, para um determinado conjunto de dados, o valor do parâmetro associado à verossimilhança máxima.

Se tivermos uma cesta com bolas azuis e vermelhas, por exemplo, e retirarmos com reposição quatro bolas azuis e seis vermelhas, o que podemos concluir sobre a proporção de bolas vermelhas na cesta? Dado o conjunto de caracteres (quatro bolas azuis e seis bolas vermelhas), se usarmos a teoria da verossimilhança, descobriremos que o ponto de máxima probabilidade é quando a proporção de bolas vermelhas é igual a 6/10. Nesse caso, estaremos estimando um único parâmetro (a proporção de bolas vermelhas), cuja probabilidade varia de acordo com seu valor. Note que a probabilidade de esse parâmetro ter qualquer outro valor entre 0 e 1 é diferente de zero, mas de acordo com nossos dados ela é máxima quando p=0,6. Esse valor máximo pode ser facilmente encontrado, derivando a fórmula da verossimilhança e igualando-a a zero, já que o coeficiente angular da tangente (isto é, a derivada) será igual a zero no topo da curva de probabilidade, onde a tangente é paralela ao eixo horizontal (isto é, os pontos onde ocorre a curvatura das distribuições das probabilidades na Figura 14.1).

No caso de filogenias, a questão não é tão simples. A verossimilhança (probabilidade) de observarmos um determinado conjunto de dados é maximizada para cada topologia e a topologia com a maior verossimilhança é a escolhida. No entanto, os parâmetros considerados não são as topologias propriamente ditas, mas os tamanhos de ramos para cada topologia. Dessa forma, a verossimilhança é maximizada variando os tamanhos de ramos em cada topologia (Nei *et al*., 1998). Entretanto, para escolher a árvore de máxima verossimilhança, é necessário calcular a probabilidade associada a diferentes topologias e cada uma delas com as variações nos tamanhos dos ramos e nos outros parâmetros do modelo evolutivo.

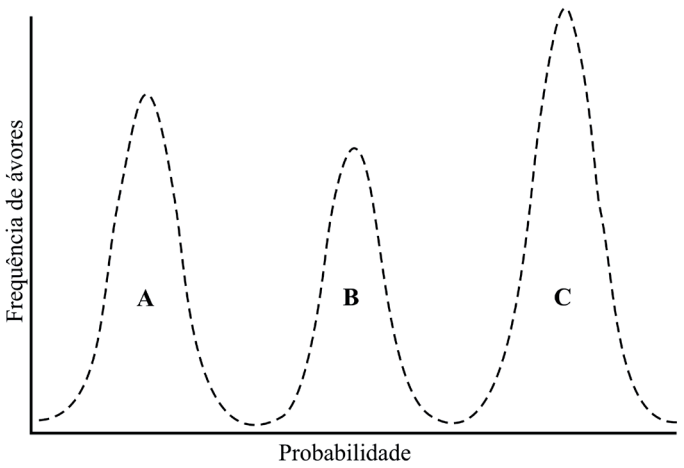


Figura 14.1. Curvas de distribuições das probabilidades associadas à: (A) árvore incorreta; (B) correta; e (C) verdadeira.

Isso não seria um problema tão grande se o método de verossimilhança fosse um método rápido, como o de UPGMA (veja o Capítulo 13), onde o processamento de uma árvore com 1000 táxons leva poucos segundos para ser completado. No entanto, na máxima verossimilhança, esse cálculo torna-se inviável à medida que o número de táxons aumenta (Nei *et al*., 1998, Ota e Li, 2000; Takahashi e Nei, 2000).

Apesar dessas diferenças nas bases do método em relação ao método original, o método de máxima verossimilhança tem sido muito usado (ver Page e Goodman, 2001; Rodriguez-Robles *et al*., 2001; Shi *et al*. 2005; Nishihara *et al*. 2007). Realmente, diversos estudos indicam a eficiência desse método na reconstrução de filogenias (Kuhner e Felsenstein, 1994; Huelsenbeck, 1995; Russo *et al*., 1996; Ota e Li, 2000). Por essas razões, abordaremos neste capítulo as bases teóricas do método de máxima verossimilhança aplicado a filogenias.

14.2. Princípios do Método de Máxima Verossimilhança

O princípio básico do método de verossimilhança consiste em estimar a verossimilhança de um conjunto de dados representar um processo que realmente ocorreu, com base em um determinado modelo evolutivo. No caso de sequências de DNA, o método irá calcular a probabilidade *L* de acordo com as premissas do modelo evolutivo θ, que uma dada topologia *T* e os comprimentos de seus ramos *B* expliquem a evolução das sequências *D.* Nesse caso, a topologia, o comprimento dos ramos e os parâmetros do modelo evolutivo são as variáveis a serem estimadas, dadas as sequências de DNA representadas nas extremidades dos ramos. Em termos matemáticos, o método de máxima verossimilhança maximiza a função

$$L(T, B, \theta \mid D) = f(D \mid T, B, \theta).$$

A probabilidade deve ser calculada para todas as topologias possíveis, estimando iterativamente o tamanho dos ramos para um grupo de unidades taxonômicas operacionais (sejam espécies, grupos monofiléticos acima do nível da espécie, populações, genes ou qualquer outra unidade evolutiva) e os parâmetros do modelo evolutivo. A árvore filogenética (isto é, a topologia mais o comprimento dos ramos) que apresentar a maior verossimilhança (probabilidade, dado o alinhamento) é considerada a melhor estimativa da história evolutiva das sequências *D*.

Calcular a verossimilhança de uma árvore envolve o cálculo das probabilidades de ocorrência de todos os possíveis estados ancestrais de caracteres nos nós internos da árvore; isto é, calcular a possível ocorrência de cada um dos nucleotídeos ter estado presente em um nó interno. A maioria dos modelos evolutivos para sequências de DNA admite reversão de caracteres ao longo do tempo, visto que a mudança de nucleotídeos, digamos de A para C é independente da mutação reversa, de C para A. Essa forma de mudança de um estado para outro sem a passagem por estados intermediários é conhecido como não ordenada. Devido a essa propriedade de reversão, a verossimilhança de uma árvore independe do posicionamento de sua raiz.

14.3. Modelos de Evolução em Verossimilhança

Alguns dos modelos evolutivos apresentados no Capítulo 13 para os métodos de distâncias também podem ser empregados pelo método de verossimilhança para calcular as probabilidades de substituição entre os nucleotídeos e calcular os comprimentos dos ramos da árvore. A seguir, descrevemos os modelos mais frequentemente usados em reconstrução filogenética usando máxima verossimilhança e como são calculadas as probabilidades de substituição de um nucleotídeo para outro. O desenvolvimento de modelos cada vez mais complexos deu-se graças ao aumento gradativo do conhecimento a respeito da evolução de sequências de ácidos nucléicos. Esses modelos envolvem a utilização de parâmetros que regem a evolução das sequências. Apesar de os valores dos parâmetros não serem conhecidos, podem ser inferidos a partir dos dados.

Modelo de um tipo de substituição (Jukes e Cantor, 1969, ou JC69)

Jukes e Cantor (1969) propuseram um modelo bastante simples, onde os nucleotídeos A, C, G e T, em uma sequência de DNA, ocorrem em frequências iguais e a probabilidade de substituição de um nucleotídeo i para um nucleotídeo j em um intervalo de tempo d*t* depende simplesmente da taxa de substituição *u*, onde $i \neq j$.

Matematicamente, podemos representar essa expressão de forma simplificada como

$$P_{ij}(\mathrm{d}t) = u\mathrm{d}t.$$

Modelo de dois parâmetros (Kimura, 1980, K2P ou K80)

Substituições do tipo transições (entre pirimidinas ou entre purinas) ocorrem mais frequentemente que as transversões (entre uma pirimidina e uma purina, ou vice-versa). Por exemplo, em uma comparação de 11 espécies de aves neotropicais (mutuns e jacus), o gene mitocondrial da subunidade II da oxidase do citocromo *c* apresentou entre 13 e 63 transições nas comparações par a par entre as espécies, e de zero a nove transversões (Tabela 14.1). Baseando-se nesse tipo de observação comum em certos genomas ou regiões genômicas, Kimura (1980) desenvolveu um modelo onde as taxas de transição α e as taxas de transversão β são consideradas separadamente. Nesse caso, a probabilidade é estimada como

$$P_{ij}(\mathrm{d}t) = \begin{cases} \alpha \mathrm{d}t \text{ para transição} \\ \beta \mathrm{d}t \text{ para transição} \end{cases}$$

e a matriz de substituição Q é representada por:

$$Q = \begin{pmatrix} * & \beta & \alpha & \beta \\ \beta & * & \beta & \alpha \\ \alpha & \beta & * & \beta \\ \beta & \alpha & \beta & * \end{pmatrix},$$

Tabela 14.1. Número de transições (acima da diagonal) e transversões (abaixo da diagonal) em sequências do gene da subunidade II da oxidase do citocromo *c* em espécies de 11 gêneros de Cracidae (Aves, Galliformes). Fonte, Pereira (2000).

	Abu	Cha	Cra	Mit	Not	Ore	Ort	Pau	Pen	Pel	Pip
Aburria	Abu	36	53	59	53	48	52	55	25	45	13
Chamaepetes	0	Cha	54	60	54	49	63	52	45	45	35
Crax	8	8	Cra	25	20	48	46	19	59	59	52
Mitu	7	7	1	Mit	25	48	56	21	65	55	58
Nothocrax	7	7	1	0	Not	40	48	18	57	59	50
Oreophasis	6	6	6	5	5	Ore	56	45	53	55	47
Ortalis	6	6	4	3	3	6	Ort	49	52	64	53
Pauxi	8	8	2	1	1	6	4	Pau	61	55	56
Penelope	1	1	9	8	8	7	7	9	Pen	49	24
Penelopina	1	1	9	8	8	7	7	9	2	Pel	48
Pipile	0	0	8	7	7	6	6	8	1	1	Pip

onde os elementos não diagonais representam as probabilidade de substituição de i por j, para $i \neq j$, multiplicada pela frequência de j, e os elementos diagonais * são definidos de maneira que a somatória de cada linha é 1. Os elementos da diagonal da matriz Q representam a probabilidade de que aquele nucleotídeo não seja substituído por nenhum dos outros três nucleotídeos alternativos, isto é,

$$* = 1 - \sum_{i \neq j} p_{ij}.$$

Modelo proporcional (Felsenstein, 1981 ou F81)

A frequência dos quatro nucleotídeos, contudo, nem sempre é similar. A cadeia leve do genoma mitocondrial de vertebrados, por exemplo, apresenta uma redução na proporção de G em relação às demais bases. Para exemplificar isso, a Tabela 14.2 mostra a composição de bases do gene mitocondrial da subunidade II da citocromo oxidase *c* de 11 mutuns e jacus. Observamos claramente uma maior frequência de A e C, e uma redução na frequência de G, considerando esse gene em sua totalidade. Ainda, ao analisar separadamente as três posições do códon, a diferença na proporção de cada uma das bases é ainda mais marcante na terceira posição, havendo 42% de C e 5% de G, nesse exemplo. Felsenstein (1981) elaborou o modelo para calcular a probabilidade de substituição, levando em consideração a frequência desigual de bases,

$$P_{ij}(\mathrm{d}t) = u\pi_j \mathrm{d}t,$$

onde π_j representa a frequência do nucleotídeo j que substitui o nucleotídeo i, onde $i \neq j$.

Modelos HKY85 e F84

Os modelos K80 e F81 foram combinados, concomitantemente, por Hasegawa *et al.* (1984, 1985) e por Felsenstein (em seu pacote de programas PHYLIP versão 2.6 e posteriores, Felsenstein, 1995) em outros dois modelos, denominados HKY85 e F84, respectivamente, com o objetivo de distinguir diferenças nas taxas de transição e transversão. Matematicamente, a probabilidade P e a matriz de substituição Q para o modelo HKY85 são representadas por

$$P_{ij}(\mathrm{d}t) = \begin{cases} \alpha\pi_j \mathrm{d}t \text{ para transição} \\ \beta\pi_j \mathrm{d}t \text{ para transição} \end{cases}$$

$$Q = \begin{pmatrix} * & \beta\pi_c & \alpha\pi_g & \beta\pi_t \\ \beta\pi_a & * & \beta\pi_c & \alpha\pi_t \\ \alpha\pi_a & \beta\pi_c & * & \beta\pi_t \\ \beta\pi_a & \alpha\pi_c & \beta\pi_c & * \end{pmatrix}.$$

Note que o modelo HKY85 também considera desigualdades na frequência de bases, dado pelo parâmetro π_j. Similarmente ao modelo K2P, visto acima, os elementos diagonais da matriz Q representam 1 menos a soma dos elementos da linha correspondente.

A diferença entre os modelos HKY85 e F84 é o emprego do parâmetro κ no último, que determina a razão transição/transversão. κ é calculado pela soma das frequências $\pi_C+\pi_T$, se *j* for uma pirimidina, ou $\pi_A+\pi_G$, se *j* for uma purina. A representação matemática simplificada do modelo F84 é:

$$P_{ij}(\mathrm{d}t) = \begin{cases} \left(\frac{\kappa}{\Pi_j} + 1\right) u\pi_j \mathrm{d}t & \text{para transição} \\ u\pi_j \mathrm{d}t & \text{para transversão} \end{cases}.$$

Modelo TN93

A diferença na composição de bases (Tabela 14.2) reflete diferenças não apenas na taxa de transversões e transições, mas também na taxa de transições entre pirimidinas e na taxa de transições entre purinas. Para adequar essa diferença a modelos de evolução de sequências, Tamura e Nei (1993) elaboraram o seguinte modelo:

$$P_{ij}(\mathrm{d}t) = \begin{cases} \alpha_R \pi_j \mathrm{d}t & \text{para transição entre purinas} \\ \alpha_Y \pi_j \mathrm{d}t & \text{para transição entre primidinas} \\ \beta\pi_j \mathrm{d}t & \text{para transversão} \end{cases},$$

$$Q = \begin{pmatrix} *\beta\pi_c\alpha_R\pi_g\beta\pi_t \\ \beta\pi_a *\beta\pi_c\alpha_Y\pi_t \\ \alpha_R\pi_a\beta\pi_c *\beta\pi_t \\ \beta\pi_a\alpha_Y\pi_c\beta\pi_c * \end{pmatrix},$$

onde α_R e α_Y representam a taxa de transição entre purinas e a taxa de transição entre pirimidinas, respectivamente, e os elementos diagonais * em Q são calculados por 1 menos a soma dos demais elementos de sua linha.

Modelo Geral de Reversão ao Longo do Tempo (GTR, do inglês, *General Time Reversible*)

Um modelo geral de substituição foi desenvolvido posteriormente (Rodriguez *et al.*, 1990) e considera frequência desigual entre os nucleotídeos e taxas de substituição diferentes entre os seis possíveis tipos reversíveis de substituição entre A, C, G e T. O modelo é representado pela matriz

$$Q = \begin{pmatrix} * & \mu_{ac}\pi_c & \mu_{ag}\pi_g & \mu_{at}\pi_t \\ \mu_{ca}\pi_a & * & \mu_{cg}\pi_c & \mu_{ct}\pi_t \\ \mu_{ga}\pi_a & \mu_{gc}\pi_c & * & \mu_{gt}\pi_t \\ \mu_{ta}\pi_a & \mu_{tc}\pi_c & \mu_{tg}\pi_c & * \end{pmatrix},$$

onde μ_{ac} corresponde à taxa de substituição de A por C, e assim por diante. Novamente, os elementos diagonais * representam 1 menos a soma dos demais elementos da linha correspondente.

Um modelo mais generalizado foi descrito posteriormente por Yang (1994), onde as probabilidades de cada um dos 12 tipos de substituição entre nucleotídeos são assimétricas. Isto é, a probabilidade de A ser substituído por T é diferente da probabilidade de T ser substituído por A, e assim por diante.

14. 4. Calculando a Probabilidade de uma Árvore Filogenética

A idéia fundamental do método de máxima verossimilhança é calcular a probabilidade de o conjunto de dados em mãos ter resultado em uma determinada árvore filogenética de acordo com o modelo evolutivo escolhido. Obviamente, isso deve ser calculado para todas as árvores possíveis para encontrar aquela que melhor explique a evolução das sequências. Sem dúvida, o tempo computacional para isso chega a ser demasiadamente longo, ou mesmo inviável, especialmente quando o número de parâmetros a ser estimado é grande.

Considere a árvore da Figura 14.2, onde Aves, Crocodilos, Lagartos e Tartarugas são as OTUs e *x*, *y* e *z* são os nós ancestrais internos. Imagine que A, C, G e T são os nucleotídeos observados em um determinado sítio para as Aves, Crocodilos, Lagartos e Tartarugas, respectivamente. Calcular a probabilidade desse sítio envolve calcular a probabilidade de o nucleotídeo observado na linhagem das Tartarugas ser T, considerando que ele poderia ter

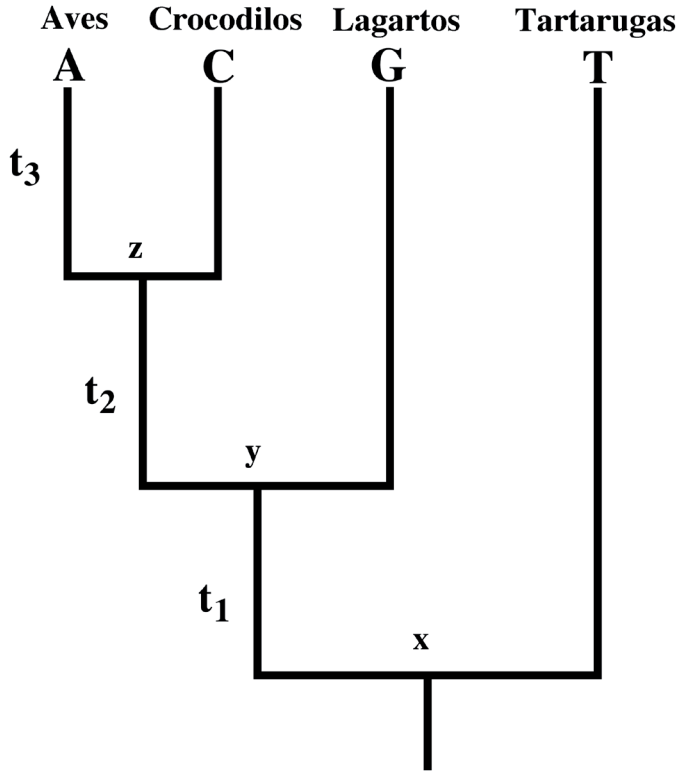


Figura 14.2. Árvore demonstrando as relações entre quatro táxons onde os estados terminais são A, C, G e T; x, y e z representam os nós internos e t_1, t_2 e t_3 representam os comprimentos dos ramos.

sido qualquer um dos quatros possíveis nucleotídeos no nó ancestral *x*, isto é, somar as probabilidades de o nucleotídeo em *x* ter sido A e mudado para T nas Tartarugas, ter sido G e mudado para T, ter sido C e mudado para T, e de não ter mudado (ou seja, era T em *x*). Essas probabilidades individuais são estimadas por meio de um modelo evolutivo, como os descritos anteriormente, e são representadas matematicamente como $\Sigma\, P_{xT}\,(t_1+t_2+t_3)$, onde t_1, t_2 e t_3 representam os tempos decorridos desde a separação da linhagem que originou as tartarugas dos demais ramos e P_{xt} é a probabilidade associada com a mudança de um nucleotídeo qualquer para T. Agora devemos calcular a probabilidade de G ser observado em Lagartos: $\Sigma\, P_{xy}(t_1)\, P_{yG}(t_2+t_3)$, onde P_{xy} é a probabilidade associada à mudança de nucleotídeo entre os ramos que ligam os ancestrais nos nós *x* e *y*, e $P_{yG}(t_2+t_3)$ é a probabilidade associada aos eventos de substituição entre o nó interno *y* e a linhagem dos Lagartos. Finalmente, a probabilidade associada à observação de A nas Aves e C nos Crocodilos é dada por: $\Sigma\, P_{yz}(t_2)\, P_{zA}(t_3)\, P_{zC}(t_{13})$. Assim, a probabilidade final da árvore exemplificada na Figura 14.2 é

$$\text{Prob(árvore)} = \Sigma\, \pi_x\, P_{xT}\,(t_1+t_2+t_3)\, \Sigma\, P_{xy}(t_1)\, P_{yG}(t_2+t_3)\, \Sigma\, P_{yz}(t_2)\, P_{zA}(t_3)\, P_{zC}(t_3).$$

Dado que, na prática, não sabemos qual é a probabilidade associada ao nucleotídeo ancestral no nó interno *x*, que representa, nesse caso, a raiz dessa árvore, devemos adicionar o parâmetro π_x, representando a frequência dos nucleotídeos.

Esse exemplo assume que as taxas evolutivas dos táxons são constantes, isto é, essas sequências evoluem de acordo com um relógio molecular. No caso de as sequências não seguirem um relógio, os parâmetros t_1, t_2 e t_3 passam a serem considerados os comprimentos dos ramos, refletindo, então, a quantidade de modificações observada em cada ramo.

Esse caso mostra o cálculo de probabilidade da ocorrência dos nucleotídeos em um único sítio. A probabilidade de uma árvore explicar um conjunto de sequências é calculada de maneira similar, isto é, para cada sítio calcula-se a probabilidade de maneira semelhante à realizada acima e, posteriormente, multiplicam-se as probabilidades calculadas para cada sítio da sequência. Em resumo, a probabilidade final de uma árvore é o produto das probabilidades de cada um dos sítios.

No exemplo acima foi considerado que conhecemos *a priori* o comprimento dos ramos. Contudo, o método de máxima verossimilhança estima o comprimento dos ramos de modo a maximizar a probabilidade de a árvore explicar a evolução das sequências examinadas.

14.5. Procurando a Árvore Filogenética mais Verossímil

Encontrar a árvore que apresente a maior probabilidade de explicar os dados obtidos de acordo com o modelo evolutivo escolhido nem sempre é uma tarefa fácil. Encontrá-la significa averiguar não só todas as topologias possíveis, mas também as variações de comprimento de ramos e valores dos parâmetros do modelo evolutivo para cada topologia. Na prática, isso é inviável de ser realizado para um grande número de táxons, uma vez que o número de possibilidades é astronômico (veja a Tabela 12.1) e encontrar a melhor árvore significaria gastar um tempo extremamente longo. Nesse caso, o emprego de algoritmos heurísticos pode auxiliar a contornar esse problema (ver Capítulo 12). Na realidade, estudos foram realizados para comparar diferentes algoritmos e indicam que métodos heurísticos são bastante eficientes ou, em alguns casos, até mais eficientes do que métodos exaustivos (Russo *et al*., 1996; Nei *et al*., 1998; Takahashi e Nei, 2000, Criscuolo e Gascuel, 2008).

Tabela 14.2. Composição de bases do gene mitocondrial da subunidade II da oxidase do citocromo *c* (em %) em espécies de 11 gêneros de Cracidae (Aves, Galliformes). Fonte, Pereira (2000).

Gênero	Gene inteiro				1ª posição do codon			
	A	T	C	G	A	T	C	G
Aburria	30,3	23,5	31,9	14,3	24,6	19,7	27,2	28,5
Chamaepetes	29,8	22,2	33,2	14,8	24,6	19,3	27,6	28,5
Crax	30,0	24,1	31,3	14,6	24,1	19,3	27,6	28,9
Mitu	29,5	23,7	31,9	14,9	24,6	19,7	27,2	28,5
Nothocrax	30,4	23,8	31,7	14,0	25,4	19,7	27,2	27,6
Oreophasis	29,4	23,4	32,0	15,2	24,6	19,3	27,6	28,5
Ortalis	29,4	24,0	31,7	14,9	24,1	18,0	28,9	28,9
Pauxi	30,0	23,4	32,0	14,6	25,0	19,3	27,6	28,1
Penelope	30,1	24,0	31,3	14,6	24,1	19,7	27,2	28,9
Penelopina	29,8	24,3	31,3	14,6	24,6	19,7	27,2	28,5
Pipile	30,1	23,8	31,6	14,5	24,6	20,2	26,8	28,5
Média	29,9	23,7	31,8	14,6	24,6	19,5	27,5	28,5
Gênero	2ª posição do codon				3ª posição do codon			
	A	T	C	G	A	T	C	G
Aburria	27,6	36,4	25,9	10,1	38,6	14,5	42,5	4,4
Chamaepetes	27,6	35,5	26,8	10,1	37,3	11,8	45,2	5,7
Crax	27,6	36,4	25,9	10,1	38,2	16,7	40,4	4,8
Mitu	27,6	36,4	25,9	10,1	36,4	14,9	42,5	6,1
Nothocrax	27,6	36,4	25,9	10,1	38,2	15,4	42,1	4,4
Oreophasis	27,6	36,4	25,9	10,1	36,0	14,5	42,5	7,0
Ortalis	27,6	36,4	25,9	10,1	36,4	17,5	40,4	5,7
Pauxi	27,6	36,4	25,9	10,1	37,3	14,5	42,5	5,7
Penelope	27,6	36,4	25,9	10,1	38,6	15,8	40,8	4,8
Penelopina	27,6	36,4	25,9	10,1	37,3	16,7	40,8	5,3
Pipile	27,6	36,4	25,9	10,1	38,2	14,9	42,1	4,8
Média	27,6	36,3	26,0	10,1	37,5	15,2	42,0	5,3

Devido ao grande tempo computacional envolvido na análise filogenética tradicional de máxima verossimilhança, uma estratégia denominada Quebra-Cabeça de Quartetos (*Quartet Puzzling*, em inglês) foi desenvolvida com o objetivo de reduzir o tempo gasto no cálculo de probabilidades das árvores (Strimmer e von Haeseler, 1996).

Por meio dessa estratégia, examinam-se as relações entre todos possíveis conjuntos de quatro táxons (quartetos) de nossa amostra. Para quatro táxons *a*, *b*, *c* e *d*, três topologias são possíveis, como demonstrado na Figura 14.3. A probabilidade de cada uma dessas árvores é examinada e a que apresentar a maior probabilidade será escolhida como a melhor árvore. Isso é realizado para todos os possíveis quartetos de nossa amostra. A seguir, as melhores árvores obtidas para cada um dos quartetos são usadas em um processo de "quebra-cabeça", no qual se pretende encaixar as topologias encontradas em uma topologia global.

Entretanto, Cao e colaboradores (1998) mostraram que esse tipo de estratégia pode não encontrar a árvore com a maior verossimilhança (probabilidade). Então, a utilidade dessa estratégia está em encontrar uma árvore em um curto período de tempo e usá-la como árvore inicial em procedimentos de rearranjos, na tentativa de buscar a árvore mais provável. Esse tipo de procedimento pode ser feito em pacotes computacionais filogenéticos, como o MOLPHY e PAML (ver Anexo), por meio de algoritmos que produzirão rearranjos em uma árvore inicial fornecida pelo usuário, nesse caso a árvore fornecida pelo "quebra-cabeça" de quartetos.

Mais recentemente, a automação e a redução de custos para obter sequências específicas de DNA e genomas completos reativaram o interesse no desenvolvimento de novos algoritmos para otimizar e acelerar as buscas de verossimilhança para conjuntos de dados com grande número de sequências. Os algoritmos genéticos mais frequentemente usados hoje em dia estão implementados nos programas PHYML (Guindon e Gascuel, 2003), RAxML-VI (Stamatakis, 2006) e GARLI (Zwickl, 2006). Em GARLI, por exemplo, o princípio fundamental do algoritmo genético é criar, em um dado intervalo de tempo, uma amostra de soluções; cada solução é composta de diversas variáveis (topologia, comprimento de ramos e parâmetros do modelo evolutivo). A verossimilhança de cada solução é avaliada e um escore é atribuído para cada solução. Uma proporção de soluções com os melhores escores é selecionada para a próxima etapa. Nessa etapa, uma variável é modificada e as demais são re-estimadas para maximizar a verossimilhança, de acordo com a váriavel inicial modificada, e o escore das novas soluções são estimados. O processo é repetido continuamente até que mudanças na topologia não alterem significativamente o escore após um determinado número de ciclos, a verossimilhança não aumente acima de um limite mínimo pré-estabelecido e um parâmetro que controla o grau de otimização de comprimento de ramos tenha atingido o seu valor mínimo.

PHYML e RAxML-VI são classificados como métodos de "subida de morro" (do inglês, "hill-climbing"). Subida de morro é uma analogia ao fato que, durante uma busca, o algortimo tentará achar o ponto mais alto nas curvas de distribuição das probabilidades, que maximiza a verossimilhança dos dados serem explicados por uma topologia, dado um modelo evolutivo (Figura 14.1). Em poucas palavras, com esses métodos, lança-se mão de uma abordagem rápida, como o Agrupamento de Vizinhos ("Neighbor-Joining", Capítulo 13), para obter uma árvore inicial (topologia) e estimar as variáveis de interesse (e.g., comprimento de ramos e parâmetros do modelo evolutivo). Um sub-ramo da árvore inicial é "podado" e re-inserido em uma posição alternativa,

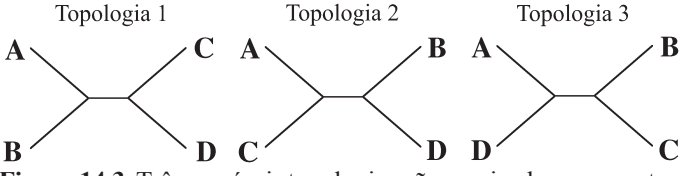


Figura 14.3. Três possíveis topologias não enraizadas para quatro táxons A, B, C e D.

e as variáveis da nova topologia são estimadas e comparadas com a topologia inicial. Se houver melhora significativa na verossimilhança da nova topologia, ela passará a ser considerada a melhor, e o procedimento se repete um determinado número de vezes. PHYML e RAxML diferem em alguns detalhes, como o fato de os sub-ramos serem escolhidos e bem como onde eles podem ser inseridos, como a verossimilhança é re-estimada e quais variáveis são mantidas na "memória" para acelerar a estimativa de todas as variáveis nos passos subsequentes.

A Tabela 14.3 compara os três programas acima mais usualmente empregados para análises de conjuntos de dados com grande número de sequências. Todas as análises foram realizadas usando o modelo GTR assumindo heterogeneidade de taxas de substituição entre os sítios (secção 14.6). O tempo de execução no RAxML-VI e no GARLI é similar para os seis conjuntos de dados menores, mas passa a ser consideravelmente maior no GARLI, acima de 4000 sequências. PHYML apresentou desempenho comparável ao RAxML-VI e GARLI, mas apresenta limitações técnicas mais severas, que o impedem de ser usado quando o número de sequências ultrapassa 2000. Além disso, o algoritmo de alteração de topologia ("nearest neighbor interchange", Capítulo 12) implementado no programa PHYML explora apenas uma fração pequena do espaço de parâmetros, comparado com os dois outros algoritmos.

Embora a comparação entre os melhores algoritmos de busca por verossimilhança tenha sido realizada em uma rede de computadores com múltiplos processadores, computadores pessoais também avançaram significativamente nos últimos anos, permitindo que buscas por verossimilhança sejam realizadas em poucas horas ou dias para conjuntos de dados com cerca de 1000 sequências entre 1000 e 10000 pares de bases.

14.6. Heterogeneidade de Taxas de Substituição Entre os Sítios

Além das probabilidades de substituição entre as bases, da frequência relativa entre elas e das taxas de transição e transversão, outros parâmetros podem ser acoplados ao modelo evolutivo, como mencionado no Capítulo 13. Um desses refere-se à heterogeneidade na taxa de substituição entre os sítios de uma mesma sequência de ácido nucléicos, cujo efeito em inferência filogenética é consideravelmente importante. Por exemplo, se os sítios em uma sequência de DNA apresentam a mesma probabilidade de sofrer uma substituição, dizemos que essa sequência apresenta taxa homogênea de substituição entre os sítios. Ao contrário, se alguns sítios em uma sequência de DNA apresentam maior taxa de substituição que outros sítios, dizemos haver heterogeneidade de taxa de substituição entre sítios nessa sequência.

Esse tipo de propriedade das sequências de nucleotídeos pode ser representada graficamente por meio da distribuição gama contínua, representada pela letra grega Γ. A Figura 14.4 ilustra essa distribuição, onde dois parâmetros são necessários. Um deles, denominado alfa (α), representa a forma da curva de distribuição gama e outro, denominado beta (β), representa a escala. Apenas α é incorporado ao modelo. Quanto menor o valor de α, maior a heterogeneidade de substituição entre os sítios (se α=zero, cada sítio possui uma taxa exclusiva de substituição; se α tende ao infinito, todos os sítios têm a mesma taxa de substituição).

De fato, as taxas de substituição ao longo de uma sequência não são iguais em muitos casos. Um exemplo comum em sequências codificadoras de proteínas é que a terceira posição do códon tem taxas maiores de substituição do que a primeira e a segunda posições. Já em sequências de rRNAs, as alças e hastes também apresentam taxas diferentes de substituição. Outros casos envolvendo determinados aminoácidos em uma sequência de proteínas apresentam maior pressão seletiva devido a algum papel importante no correto funcionamento da proteína, como proximidade em relação ao sítio ativo. A Figura 14.5 ilustra uma comparação entre sequências de aminoácidos de uma proteína reguladora da expressão gênica no homem e na galinha, denominada receptora de estrógeno. Dois dos domínios dessas sequências são extremamente conservados, com mais de 94% de similaridade entre essas duas espécies. Esses domínios sofrem uma forte ação de seleção negativa, uma vez que são domínios importantes de ligação entre o DNA (domínio C) e o hormônio (domínio E). Já os demais domínios estão sob menor restrição evolutiva e sua taxa de divergência é maior, com cerca de 60% de similaridade.

Os valores da heterogeneidade podem ser estimados diretamente a partir dos dados. A Tabela 14.4 traz vários exemplos de α calculados para diferentes tipos de dados em diferentes grupos de organismos. Note a variação, de 0,16 para sequências do gene mitocondrial 12S rDNA em roedores (heterogeneidade forte de substituição) até 0,95 para alguns genes mitocondriais codificadores de vertebrados (heterogeneidade mediana de substituição).

Levar em conta a heterogeneidade de substituição entre sítios em uma análise de verossimilhança tem mostrado resultados muito positivos. Muitas vezes, assumir homogeneidade entre os sítios pode levar a uma estimativa da filogenia que não reflete a verdadeira história evolutiva entre as unidades taxonômicas operacionais (Takezaki e Gojobori, 1999). Fazendo isso, subestimamos a taxa de substituição de sítios onde a variação é alta e superestimamos a taxa de substituição para sítios onde a variação é baixa ou inexistente.

Assumir a distribuição gama contínua, no entanto, pode levar ao aumento do tempo computacional para se estimar uma árvore. Uma maneira de agilizar o processo assumindo heteroge-

Tabela 14.3. Tempo de execução de buscas por máxima verossimilhança usando três estratégias implementadas em diferentes programas. O tempo é dado em horas:segundos. Análises foram realizadas em rede de computadores com múltiplo processadores de 2.4 GHz e 8 GB de memória RAM. GARLI e PHYML não permitem buscas acima de um certo número de sequências e são marcados com um hífen. Fonte: modificado do suplemento em Stamakis (2006).

Número de sequências	Pares de bases	RAxML-VI	GARLI	PHYML
1000	5547	09:24	07:13	10:29
1497	1241	05:33	09:12	10:09
1663	1577	07:40	10:44	06:04
1728	1276	06:20	05:44	04:35
2000	1251	12:47	11:52	02:44
2560	1232	08:04	14:25	-
4114	1263	24:37	50:32	-
6722	1122	43:10	115:33	-
7769	851	35:41	-	-
8780	1217	65:11	-	-

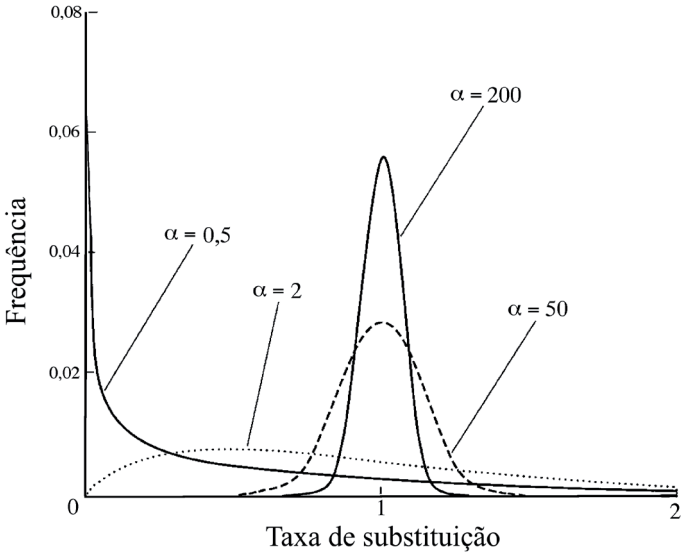


Figura 14.4. Distribuição gama para três valores de α. (modificado de Hillis *et al*. 1994).

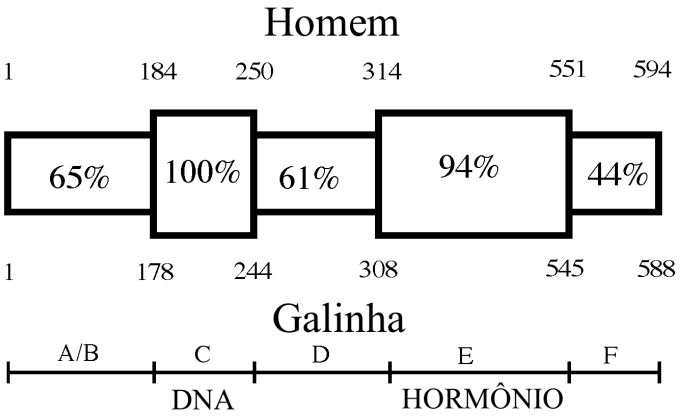


Figura 14.5. Comparação esquemática das sequências de aminoácidos de um receptor de estrógeno, do homem e da galinha, envolvido em regulação da transcrição gênica. Os números correspondem à posição dos resíduos de aminoácidos, excluindo a primeira metionina. As porcentagens representam a proporção de aminoácidos conservados entre os dois organismos. A divisão do esteróide em seis domínios (A-F) é indicada abaixo, na figura. Domínios C e E são, respectivamente, os domínios de ligação ao DNA e ao hormônio. Modificado de Li (1997), após Green e Chambom (1986).

neidade de taxas é distribuir os sítios em classes distintas de taxas de substituição. Por exemplo, quatro ou mais classes de taxa de substituição poderiam ser definidas, de forma que teríamos um grupo onde a taxa de substituição é nula ou quase nula, uma classe onde a taxa de substituição é alta e duas classes intermediárias. Obviamente, quanto maior o número de classes, mais realista será a análise, porém maior será o tempo necessário para estimar o valor de α. Porém, como apenas α é incorporado aos modelos evolutivos, o número de classes determinadas para a distribuição gama não interfere no tempo total do cálculo da probabilidade da árvore, uma vez conhecido o valor de α.

Outra maneira de incorporar heterogeneidade de substituição entre os sítios é dada pelo modelo de sítios invariáveis. Em um conjunto de sequências alinhadas, algumas posições são invariáveis, isto é, não são observadas substituições, ao passo que os demais sítios têm a mesma probabilidade de substituição. Esse modelo é mais simples do que o modelo de distribuição gama, no sentido de que apenas um parâmetro está envolvido, a proporção de sítios invariáveis.

Uma alternativa mais refinada de acomodação dessa heterogeneidade de substituições entre os sítios pode ser realizada por meio de um modelo misto, onde um parâmetro determina a proporção de sítios invariáveis e os sítios variáveis apresentam heterogeneidade de taxa de acordo com a distribuição gama.

Atualmente, alguns programas de análises filogenéticas, como PAML (Yang, 1997, 1998b) e GARLI (Zwickl, 2006), permitem que se partilhem os dados em classes descontínuas e os parâmetros para cada classe podem ser estimados independentemente das demais classes. Dessa maneira, para diferentes genes ou regiões gênicas, seus próprios parâmetros (taxa de transição/transversão, parâmetro α da distribuição Γ, frequência de bases e outros) podem ser usados no cálculo da probabilidade da árvore. Imagine o exemplo da Figura 14.6, onde é dada a taxa de substituição para diferentes regiões de uma sequência e para pseudogenes. Cada uma das regiões definidas nesse exemplo pode ser considerada independentemente das demais e seus parâmetros podem ser estimados e usados independentemente em uma análise.

De fato, estimar a forma da distribuição gama baseada em um conjunto concatenado de sequências pode não diminuir a forma real da distribuição para cada gene isoladamente. Nesse caso, a melhor opção seria usar cada gene separadamente para reconstruir a filogenia e, não havendo conflitos entre as estimativas dadas por diferentes genes, somar as probabilidades dessas árvores (Cao *et al.*, 1999, 2000; Adachi *et al.*, 2000).

A incorporação de dois ou mais parâmetros acima descritos pode ser feita concomitantemente em uma análise. Contudo, nem sempre a inclusão de parâmetros extras pode resultar em uma melhora significativa na probabilidade de uma árvore explicar a evolução das sequências em questão (Huelsenbeck e Crandall, 1997; Huelsenbeck e Rannala, 1997). Como veremos na secção 14.8, existem testes estatísticos para estabelecer até que ponto a inclusão de parâmetros ao modelo evolutivo melhora a verossimilhança de uma filogenia.

14.7. Relógio Molecular

Se os táxons analisados apresentam sequências que evoluem a uma taxa constante, podemos assumir que as substituições

Tabela 14.4. Valores do parâmetro alfa da distribuição gama, estimado para diversos conjuntos de dados de sequências de DNA.

Sequências	Táxons	Valor estimado	Fonte
1ª+2ª posição do codon dos genes de α- e β- globina,	5 mamíferos	0,36	Yang *et al*. (1994)
Lisozimas	24 primatas	0,66	Yang (1998a)
Genoma do vírus da Hepatite B	13 variantes	0,26	Yang, Lauder, e Lin (1995)
12S rRNA	9 roedores	0,16	Sullivan *et al*. (1996)
Domínios 1 e 2 da região controladora do DNA mitocondrial	25 humanos	0,17	Yang e Kumar (1996)
1ª+2ª posição do codon dos 13 genes mitocondriais codificantes	11 vertebrados	0,13-0,95	Kumar (1996a)
Citocromo b	10 aves	0,42	Miyaki *et al*. (1998)

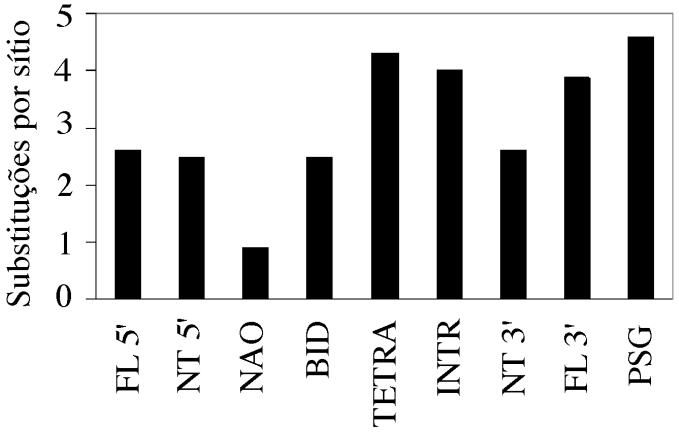


Figura 14.6. Taxa de substituição entre regiões gênicas e pseudogenes. FL 5' e FL 3'– região flanqueadora 5'e 3', respectivamente; NT 5' e NT 3' – região não traduzida 5'e 3', respectivamente; NAO – sítios não degenerados; BID – sítios bidegenerados; TETRA – sítios tetradegenerados; INTR – introns; PSG – pseudogene. Fonte, Li (1997).

ocorrem de acordo com um *relógio molecular*. Nessa situação, o comprimento dos ramos de uma árvore refletiria o número de substituições de cada linhagem, permitindo, com base nesse relógio molecular, estimar o tempo de divergência entre qualquer nó interno se pelo menos a idade de um determinado evento de cladogênese tiver sido previamente estimada (por exemplo, utilizando-se o registro fóssil, algum evento biogeográfico ou ainda outros dados moleculares).

Uma maneira simples de testar a hipótese do relógio molecular é por meio de um teste de razão de verossimilhanças, onde são comparados os logaritmos das verossimilhanças de uma mesma árvore sem e com a restrição do relógio molecular da seguinte maneira:

$$2 \times (\log \text{Árvore}_{\text{sem relógio}} - \log \text{Árvore}_{\text{com relógio}}).$$

Nesse caso, o número de graus de liberdade é igual ao número de OTUs menos 2 e a significância do teste pode ser avaliada por meio da distribuição de χ^2 (Felsenstein, 1988). Esse teste pode ser realizado em programas como o TREEPUZZLE (Strimmer e von Haeseler, 1996) e o PAUP* (Swofford, 1999), que estimam a verossimilhança de uma filogenia com e sem a restrição do relógio molecular.

Alternativamente, o teste de taxas relativas (*relative-rate test*; Sarich e Wilson, 1967, Takezaki *et al.*, 1995; veja também o Capítulo 7) pode ser aplicado. Suponha que queiramos saber se, na árvore apresentada na Figura 14.2, as Aves e os Lagartos apresentam a mesma taxa de substituição. Se a taxa de substituição é a mesma, então o número de substituições ocorridas no ramo das Aves desde sua separação de seu ancestral C não deve ser estatisticamente diferente do número de substituições ocorridas desde que os Lagartos se separaram de seu ancestral C, o qual é o ancestral comum mais recente compartilhado com as Aves. Para avaliar isso, é necessário incluir uma linhagem que tenha divergido anteriormente à separação das Aves e Lagartos. Nesse caso, poderíamos usar as Tartarugas como essa linhagem. Portanto, o número de substituições (K) de cada linhagem pode ser obtido diretamente das sequências, de modo que temos:

$$K_{\text{Aves/Tartarugas}} = K_{\text{Aves/Ancestral C}} + K_{\text{Tartarugas/Ancestral C}}$$
$$K_{\text{Lagartos/Tartarugas}} = K_{\text{Lagartos/Ancestral C}} + K_{\text{Tartarugas/Ancestral C}}$$
$$K_{\text{Aves/Lagartos}} = K_{\text{Aves/Ancestral C}} + K_{\text{Lagartos/Ancestral C}}.$$

Resolvendo essas três equações, temos

$$K_{\text{Aves/Ancestral C}} = (K_{\text{Aves/Tartarugas}} + K_{\text{Aves/Lagartos}} - K_{\text{Lagartos/Tartarugas}})/2$$
$$K_{\text{lagartos/Ancestral C}} = (K_{\text{Lagartos/Tartarugas}} + K_{\text{Aves/Lagartos}} - K_{\text{Aves/Tartarugas}})/2$$
$$K_{\text{Tartarugas/Ancestral C}} = (K_{\text{Aves/Tartarugas}} + K_{\text{Lagartos/Tartarugas}} - K_{\text{Aves/Lagartos}})/2.$$

De acordo com a teoria do relógio molecular (ver Capítulo 7), a diferença entre $K_{\text{Aves/Ancestral C}}$ e $K_{\text{Lagartos/Ancestral C}}$ não deve ser estatisticamente diferente de zero. Caso contrário, a hipótese do relógio molecular é rejeitada a 1% de significância.

Existem alguns programas que realizam o teste de taxas relativas. O PHYLTEST (Kumar, 1996b) e MEGA4 (Tamura *et al.*, 2007), por exemplo, permitem que o usuário defina dois grupos cujas taxas devem ser testadas e qual grupo é externo, indicam se a hipótese do relógio molecular é rejeitada ou não. Já o programa LINTREE (Takezaki *et al.*, 1995) vai mais além e detecta quais OTUs apresentam taxa de substituição significativamente mais rápidas ou mais lentas que as demais. OTUs com taxas muito diferente da média são retiradas do conjunto de dados e a análise é repetida até que o teste não encontre diferenças significativas nas taxas de substituição entre as OTUs restantes, e o relógio molecular pode ser aplicado.

Muitas vezes, contudo, as sequências utilizadas não apresentam taxas semelhantes de substituição e, em geral, a retirada de OTUs não é um procedimento desejável. Nesses casos, o uso de relógios locais ou "dispersos" é uma alternativa mais apropriada, que não requer a exclusão de sequências. Diversos modelos de relógios dispersos foram propostos. Eles serão brevemente citados aqui, uma vez que seria necessária uma introdução a métodos matemáticos e estatísticos nos quais esses métodos se baseiam, fugindo do escopo do presente capítulo.

De uma maneira geral, os métodos de relógios dispersos acomodam a variação nas taxas de substituição entre as sequências e os comprimentos dos ramos são estimados baseados em pontos de calibração que podem ser derivados de dados paleontológicos, biogeográficos, geológicos ou mesmo estimativas moleculares obtidas previamente. Basicamente, esses métodos podem ser agrupados em três categorias, de acordo com a maneira pela qual os comprimentos dos ramos são estimados.

Em uma dessas categorias, o comprimento de cada ramo é estimado de acordo com os procedimentos usados no método de máxima verossimilhança, isto é, o modelo e o parâmetro são escolhidos de modo a maximizar a probabilidade da árvore. Exemplos desse procedimento incluem os trabalhos de Cutler (2000) e Yoder e Yang (2000), baseados nos trabalhos pioneiros de Hasegawa *et al.* (1989) e Kishino e Hasegawa (1990), desenvolvidos para métodos de distância.

Cutler (2000) descreveu um modelo estacionário de verossimilhança onde a taxa de substituição em diferentes linhagens varia porém é fixa em uma linhagem. Ele usou o teorema do limite central para elaborá-lo, permitindo que a variância não seja igual à média, como é assumido pela distribuição de Poisson. Além de simulações, o autor usou dois conjuntos reais de dados para mostrar a aplicabilidade do modelo proposto: sequências de *rbc*L, um gene do cloroplasto de plantas terrestres e cinco genes mitocondriais e dois nucleares de metazoários. Em ambos os casos, o modelo estacionário não foi rejeitado, mas o modelo de Poisson, sim. Cutler (2000) concluiu que, sem o conhecimento da variância do tempo entre as substituições, dados moleculares podem ser incapazes de fornecer boas estimativas de tempo de divergência entre OTUs.

Yoder e Yang (2000) descrevem um modelo de relógios locais intermediário entre o modelo de Cutler, descrito acima, e o modelo global, onde todas as linhagens apresentam a mesma taxa

de substituição. Nesse caso, OTUs próximas apresentam taxas semelhantes de substituição e diferentes partes da árvore podem apresentar diferentes taxas. Esse modelo está implementado no programa PAML, de autoria de Yang (1998b), e sua utilização foi demonstrada para um conjunto de dados de genes mitocondriais de 31 espécies de mamíferos.

Uma segunda categoria de métodos baseia-se no pressuposto de que as taxas de substituição são autocorrelacionadas no tempo, isto é, ramos próximos na árvore devem apresentar taxas de substituição mais semelhantes entre si. Dessa maneira, as diferenças entre as taxas de substituição entre descendentes e ancestrais são minimizadas (Sanderson, 1997). Esse método não paramétrico foi aplicado também em um conjunto de dados de *rbc*L de plantas terrestres e os tempos de divergência estimados por esse método são congruentes com dados paleobotânicos. A conclusão de Sanderson (1997) é que o método se sobressai sobre o método de máxima verossimilhança quando o tamanho das sequências é suficientemente grande, quando as taxas de substituição entre as linhagens apresentam variação e quando as taxas são realmente correlacionadas no tempo. Mais recentemente, Sanderson (2002) desenvolveu um novo método, denominado verossimilhança penalizada semi-paramétrica (do inglês, "semi-parametric penalized likelihood"). O método aceita a existência de variação das taxas evolutivas entre as linhagens, mas uma penalidade é aplicada para minimizar variações de taxas entre um nó e seus descendentes, através de um parâmetro de suavização. Quanto menor o valor deste parâmetro de suavização, maior a variação de taxas observadas entre as sequências. Um teste de validação cruzada gerenciado pelos dados é aplicado para escolher o melhor valor para o parâmetro de suavização (Sanderson, 2002). Durante o procedimento de validação cruzada, o algoritmo escolhe aleatoriamente linhagens a serem retiradas da árvore, estima o parâmetro de suavização para as linhagens restantes e repetidamente tenta prever o valor do parâmetro de suavização para a linhagem retirada, até o melhor parâmetro ser encontrado (Sanderson, 2002). Esse método representa uma melhoria em relação ao método não-paramétrico de Sanderson (1997).

Finalmente, uma última categoria para datação de tempos de divergência baseia-se em uma estatística bayesiana. Nessa categoria, os parâmetros são estimados por inferência bayesiana e as taxas de substituição podem variar discreta (Huelsenbeck *et al.*, 2000) ou continuamente (Kishino *et al.*, 2001; Thorne e Kishino, 2002). Além disso, as taxas associadas a ramos descendentes podem ser correlacionadas (Thorne e Kishino, 2002) ou não correlacionadas (Drummond *et al.*, 2006) aos ramos ancestrais usando algoritmos Bayesianos mais realistas que relógios estritos descritos previamente. Entretanto, tais métodos envolvem algoritmos complexos, fugindo ao escopo do presente capítulo e, portanto, não serão descrito em detalhes. Vale a pena ressaltar que os relógios Bayesianos se tornaram muito populares, pois permitem que modelos de evolução diferentes sejam aplicados para diferentes genes e ainda levam em consideração incertezas que permeiam análises filogenéticas, tais como os pontos de calibração, as estimativas dos comprimentos de ramos e as taxas de substituição ao longo do tempo.

14.8. Teste de Comparação entre Modelos Evolutivos

Os modelos evolutivos apresentados no início deste capítulo são um caso especial um dos outros. Por exemplo, se não há diferenças nas taxas de transição entre purinas e transição entre pirimidinas, o modelo de TN93 se reduz ao modelo de HKY85. Essa propriedade dos modelos permite que eles sejam comparados por meio do teste de razão de verossimilhança (LRT, do inglês, *likelihood ratio test*; Goldman, 1993; Huelsenbeck e Crandall, 1997; Huelsenbeck e Rannala, 1997; Sullivan e Joyce, 2005)

O teste de LRT é realizado multiplicando a diferença entre os logaritmos das probabilidades por 2 e verificando a significância das diferenças de acordo com a distribuição de χ^2. Os graus de liberdade equivalem à diferença entre o número de parâmetros livres entre os modelos testados. Se a diferença for significativa, a adição de mais parâmetros melhora a probabilidade da árvore, isto é, o modelo com mais parâmetros reflete de maneira mais apropriada a evolução das sequências analisadas, aumentando a probabilidade de que aquela árvore melhor represente a história evolutiva daquele conjunto particular de sequências.

O critério de informação de Akaike (AIC, do inglês *Akaike Information Criterion*; Akaike, 1974) é outra alternativa para a escolha do modelo evolutivo, o qual penaliza a adição de parâmetros extras. Ele é estimado como AIC = -2 × log Probabilidade + 2 × (número de parâmetros). O modelo que minimizar AIC é considerado o mais adequado para ser utilizado na reconstrução filogenética. Entretanto, a eficiência desses testes (LRT e AIC) na seleção de um modelo apropriado foi questionada (Takahashi e Nei, 2000) e, em certos casos, modelos mais simples podem ser tão ou mais eficientes do que os mais complexos (Kelchner e Thomas, 2006).

Mais recentemente, o critério de informação Bayesiana (BIC, do inglês "Bayesian Information Criterion") tem sido usado com frequência na escolha de modelos evolutivos, principalmente quando se trata de conjuntos de dados grandes, onde LRT e AIC tendem a favorecer modelos mais complexos e rejeitar modelos mais simples. O BIC de cada modelo é estimado seguindo a fórmula

$$\mathrm{BIC} = -2\,L + p \log n,$$

onde n é o tamanho amostral (usualmente o número de caracteres no alinhamento). O modelo com o menor BIC representa o modelo onde a probabilidade posterior é máxima e, portanto, o modelo é mais adequado para explicar a evolução das sequências.

No caso de comparação onde os modelos não representam um caso especial um do outro, Felsenstein (1988) sugere admitir apenas 1 grau de liberdade entre eles e compará-los em uma tabela de χ^2.

14.9. Teste de Comparação Entre Topologias Alternativas

O teste de razão de verossimilhança também pode ser usado para a comparação entre duas topologias concorrentes. Nesse caso, o número de graus de liberdade será computado de acordo com a diferença do número de ramos entre as duas árvores. Contudo, o teste é inválido se uma dessas árvores não é um subconjunto da outra, ou seja, se as duas árvores forem totalmente bifurcadas (Swofford *et al.*, 1996). Nesses casos, a mesma sugestão dada por Felsenstein (1988) para modelos evolutivos que não são um caso especial um do outro é válida para árvores que diferem apenas por um rearranjo de seus ramos: considerar um grau de liberdade.

O teste desenvolvido por Kishino e Hasegawa (1989) é uma outra opção amplamente usada atualmente. A probabilidade de cada uma das árvores concorrentes é calculada e a diferença entre as probabilidades é avaliada estatisticamente pelo teste t. Pacotes de inferência filogenética como o PAUP* e o PUZZLE utilizam esse teste quando uma matriz de nucleotídeos e as árvores concorrentes são fornecidas.

Goldman *et al.* (2000), no entanto, contestaram a validade do teste de Kishino-Hasegawa para comparações entre topologias obtidas *a posteriori*, isto é, o teste não seria válido quando as árvores são obtidas para genes diferentes ou por métodos diferentes para um mesmo conjunto de sequências. A sugestão desses autores é usar testes paramétricos, como o teste SOWH (Swofford *et al.*, 1996), com poder estatístico maior do que os testes não paramétricos. No entanto, tal teste requer a comparação da verossimilhança de nossa árvore com uma distribuição nula para a diferença entre os valores de verossimilhança e, para obter essa distribuição nula, o pesquisador gasta um tempo computacional considerável, uma vez que é necessária a simulação de matrizes de sequências com parâmetros (por exemplo, frequências de bases, razão de transição/transversão, variação de taxas de substituição entre os sítios) semelhantes à da matriz original.

Para solucionar parte dos problemas associados com topologias obtidas *a posteriori* e reduzir o viés de incluir mais árvores no intervalo de confiança com o aumento no número de árvores sendo comparadas, Shimodaira (2002) propôs um teste denominado "Aproximadamente Não-enviesado" (AU do inglês, *Approximately Unbiased*). O teste AU usa um procedimento de *bootstrap* (processo de reamostragens sucessivas, com reposição dos sítios usados) e verossimilhanças de cada sítio do alinhamento para derivar valores P das topologias testadas. A distribuição das probabilidades posteriores de cada topologia alternativa é obtida por meio da contagem do número de vezes em que a hipótese é apoiada nas réplicas de i. No entanto, se as árvores sendo comparadas apresentam escores muito similares, a árvore "verdadeira" pode ser eliminada do intervalo de confiança (Shimodaira, 2002).

14.10. Vantagens e Desvantagens do Método de Verossimilhança

O método de verossimilhança tem-se mostrado mais útil que os métodos geométricos e de parcimônia (Hasegawa e Fujiwara, 1993, Kuhner e Felsenstein, 1994, Huelsenbeck, 1995; Swofford *et al.*, 1996) porque: (1) apresenta menor variância do que os métodos geométricos e de parcimônia, já que é menos afetado por erros de amostragem, mesmo com sequências curtas; e (2) tende a ser mais robusto a violações do modelo evolutivo, já que os processos evolutivos atuantes em um determinado sítio ocorrem de maneira semelhante em muitos outros.

Embora uma melhoria significativa tenha sido ocorrido nos últimos anos para acelerar o processo de estimativa das variáveis necessárias, o método de verossimilhança ainda pode ter a desvantagem de consumir grande tempo no cálculo das probabilidades, especialmente em se tratando de análises que envolvam um grande número de OTUs ou genomas completos. A adição de parâmetros extras ao modelo evolutivo também aumenta o tempo necessário para se calcular a probabilidade final. Nesses casos, a maior velocidade computacional dos métodos geométrico e de parcimônia faz com que eles sejam os únicos métodos possíveis de reconstrução filogenética para muitas sequências.

14.11. Exemplos

Um dentista da Flórida, Estados Unidos, portador do vírus HIV que causa a síndrome da imunodeficiência adquirida (AIDS), suspeitou que tinha contaminado alguns de seus pacientes. Alguns pesquisadores isolaram o vírus dos pacientes, do dentista e de outras pessoas infectadas que não eram pacientes do mesmo dentista, e sequenciaram alguns genes desses vírus. Após uma série de relatos controversos, Hillis *et al.* (1994, 1996) usaram o método de máxima verossimilhança para tentar estabelecer uma conclusão a respeito da possível contaminação pelo vírus da AIDS ter ocorrido no consultório dentário. A inferência filogenética obtida levou os pesquisadores a concluirem que alguns dos pacientes foram contaminados pelo dentista, conforme indicado pelos pacientes representados no retângulo da Figura 14.7. No entanto, as sequências obtidas do vírus de outros pacientes (F e D, na Figura 14.7) são mais relacionadas com sequências do vírus de outras pessoas da comunidade, que não eram pacientes do dentista, indicando que, nesse caso, a infecção não ocorreu no consultório dentário. Desde então, resultados de análises de sequências de DNA pelo método de máxima verossimilhança passaram a ser usados pelo sistema judicial americano como evidência criminológica.

Esse método também já foi extensivamente usado para recuperar relações filogenéticas questionáveis, como no caso das lampreias e das feiticeiras. Esses dois grupos são definidos, dentre outras características, pela ausência de mandíbulas, e são considerados os grupos mais basais de vertebrados. Durante alguns anos,

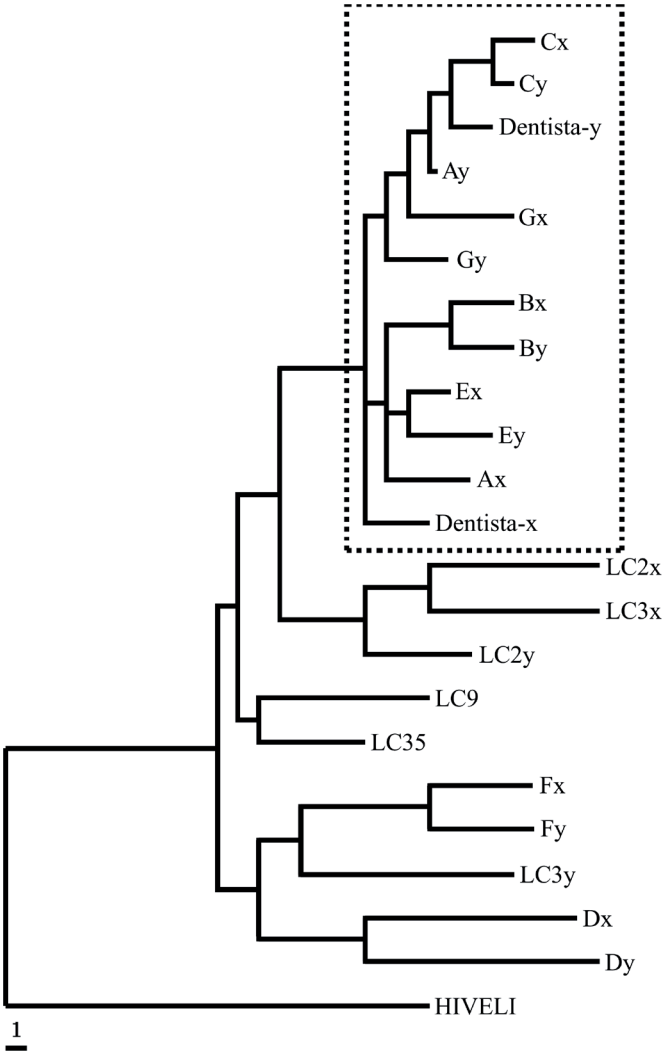


Figura 14.7. Uso de filogenias para mostrar um caso de transmissão do vírus da síndrome da imunodeficiência adquirida (HIV). As letras A-G representam diferentes pacientes do mesmo dentista, portador do HIV; x e y representam duas linhagens diferentes de vírus sequenciadas. LC representam indivíduos da mesma comunidade, mas que não são pacientes do dentista. HIVELI é uma sequência de uma linhagem africana e foi usada como grupo externo. Veja o texto para as conclusões do estudo. A barra corresponde ao número de substituição de nucleotídeos. Modificado de Hillis *et al.*, 1994.

houve um debate se eles representariam um grupo monofilético ou se as lampreias seriam grupo-irmão dos vertebrados com mandíbulas (gnatostomados) e as feiticeiras seriam a linhagem mais basal entre todos os vertebrados. Recentemente, Mallat e Sullivan (1998) mostraram, por meio de inferências filogenéticas usando o método de máxima verossimilhança, que as lampreias e as feiticeiras formam um grupo monofilético, irmão dos vertebrados gnatostomados. Esses pesquisadores concluíram ainda que outros vertebrados agnados extintos podem não pertencer a esse grupo monofilético de agnados atuais e poderiam representar os vertebrados mais basais.

Nem sempre, no entanto, os resultados são conclusivos em relações à posição filogenética de determinados táxons. Um exemplo de debate recente causado por diferentes estimativas filogenéticas baseadas em análises de verossimilhança está nas relações basais dos grupos de aves. Estudos recentes têm resultados em hipóteses, por meio do uso de sequências completas do DNA mitocondrial, em que as aves ratitas e os tinamídeos, consideradas como as aves mais primitivas (Cao *et al*., 2000; Cracraft, 1988; Stapel *et al*., 1984), poderiam ser um grupo-irmão de clados antigos, como o das aves galináceas, e que os Passeriformes representariam o grupo mais basal entre as Aves (Härlid e Arnason, 1998, Mindell *et al.*, 1999), uma visão até pouco tempo não imaginada. Contudo, a composição de bases é bastante heterogênea entre os diferentes grupos de aves, e métodos de distância e máxima verossimilhança, que corrigem diferenças na composição de bases, recuperam a filogenia tradicional, com as ratitas e tinamídeos em um clado irmão das demais aves, e os Passeriformes em uma posição mais alta na topologia (Haddrath e Baker, 2001).

Em casos onde as sequências usadas parecem evoluir a uma taxa constante entre os táxons, os métodos de máxima verossimilhança permitem datar quando as sequências desses táxons iniciaram sua divergência. Por exemplo, os psitaciformes neotropicais (papagaios, araras e afins) parecem ter se separado dos australasiáticos cerca de 76 milhões de anos atrás (Ma), coincidente com a transformação geológica ocorrida na Terra que levou à separação dos continentes da América do Sul, África, Índia, Austrália e Antártica. A partir daí, a separação entre os gêneros modernos de psitaciformes neotropicais parece ter ocorrido entre 27 e 20 Ma e pode ser relacionada com eventos paleogeográficos específicos ocorridos durante os períodos Oligoceno e Mioceno (Figura 14.8; Miyaki *et al.*, 1998).

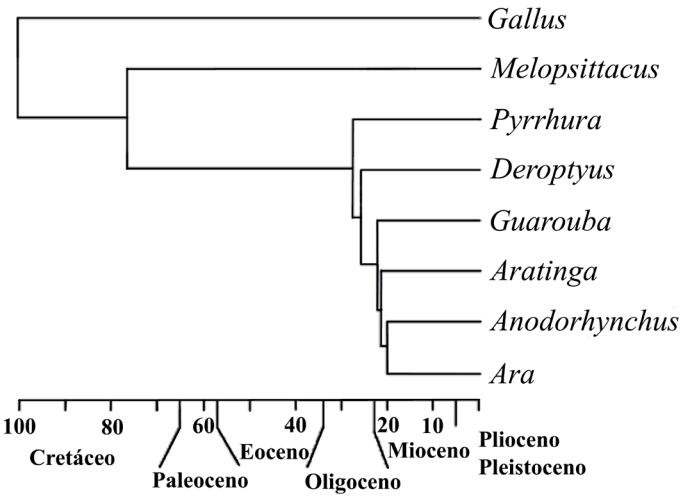


Figura 14.8. Árvore linearizada baseada em 1771 pb de genes mitocondriais com representantes de seis gêneros de psitacídeos neotropicais (*Pyrrhura*, *Deroptyus*, *Guarouba*, *Aratinga*, *Anodorhynchus* e *Ara*), de um gênero de psitacídeo australasiático (*Melopsittacus*) e grupo externo (*Gallus*).

14.12. Considerações Finais sobre Métodos de Reconstrução Filogenética

Depois de tantas fórmulas para o cálculo das distâncias entre sequências e tantos métodos para reconstrução de filogenias, fica a pergunta, o que usar? Essa, com certeza, não é uma pergunta para ser respondida *a priori*. Um aspecto que está se tornando mais claro é que simulações de computador e estudos com filogenias conhecidas devem ser considerados quando da escolha do método a ser usado para um determinado grupo de dados.

Como visto também nos Capítulos 12 e 13, cada tipo de método de reconstrução filogenética tem suas vantagens e suas desvantagens. Métodos de máxima parcimônia têm sido usados com frequência desde a década de 60 na análise de caracteres morfológicos e métodos de máxima verossimilhança possuem uma base teórica forte, enquanto que métodos de distância possuem bons testes de confiança em topologias.

Na realidade, a gama de valores dos parâmetros de evolução molecular é tão grande que pouco podemos afirmar com base em um ou dois estudos de simulações em computador. Tendo em vista tantos resultados conflitantes nas simulações (Hillis *et al.*, 1994; Huelsenbeck, 1995; Nei *et al.*, 1995) e estudos de filogenia conhecida (Russo *et al.*, 1996), devemos nos concentrar em estudos nos quais a gama de valores reflete aqueles biologicamente relevantes (Nei *et al.*, 1995). Por exemplo, um estudo onde a probabilidade de substituição por sítio é muito alta (por exemplo, $p > 0,7$) dificilmente será útil para testarmos a eficiência de métodos, já que um autor mais cuidadoso iria escolher outra região mais conservada do genoma para a análise filogenética.

Agradecimentos

Este trabalho foi financiado em parte com projetos que receberam recursos da FAPESP, FAPERJ e CNPq. Agradecemos as valiosas sugestões da Profa. Dra. Anita Wajntal e do Dr. Allan J. Baker, que contribuíram para a melhoria deste texto.

Referências Bibliográficas

Adachi, J., Wadell, P., Martin, W. e Hasegawa, M. (2000). Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. **J. Mol. Evol. 50**: 348-358.

Akaike, H. (1974). A new look at the statistical model identification. **IEEE Trans. Autom. Contr. AC-19**: 716-723.

Cao, Y., Sorenson, M.D., Kumazawa, Y., Mindell, D.P. e Hasegawa, M. (2000). Phylogenetic position of turtles among amniotes: evidence from mitochondrial and nuclear genes. **Gene 259**: 139-148.

Cao, Y., Adachi, J. e Hasegawa, M. (1998). Comment on the quartet puzzling method for finding maximum-likelihood tree topologies. **Mol. Biol. Evol. 15**: 87-89.

Cao, Y., Kim, K., Ha, J. e Hasegawa, M. (1999). Model dependence of phylogenetic inference: relationship among carnivores, perissodactyls and cetartiodactyls as inferred from mitochondrial genome sequences. **Genes, Genet. Syst. 74**: 211-217.

Cavalli- Sforza, L.L. e Edwards A.W.F. (1967). Phylogenetics analysis: models and estimation procedures. **Am. J. Hum. Genet. 19**: 233-257.

Cracraft, J. (1988). The major clades of birds. **In** Benton, M. J. (Ed). **The phylogeny and classification of tetrapods**. Systematics Assoc. Special Vol 35A. Claredon Press, Oxford, pp. 333-355.

Criscuolo, A. e Gascuel, O. (2008). Fast NJ-like algorithms to deal with incomplete distance matrices. **BMC Bioinformatics 9**:166.

Cutler, D.J. (2000). Estimating divergence times in the presence of overdispersed molecular clock. **Mol. Biol. Evol. 17**: 1647-1660.

Drummond, A.J., Ho, S.Y.W., Phillips, M.J. & Rambaut, A. (2006) Relaxed Phylogenetics and Dating with onfidence. **PLoS Biol** 4(5): 88. doi:10.1371/journal.pbio.0040088.

Felsenstein, J. (1973). Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. **Syst. Zool. 22**: 240-249.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. **J. Mol. Evol. 17**: 368-376.

Felsenstein, J. (1988). Phylogenies from molecular sequences: inference and reliability. **Annu. Rev. Genet. 22**: 521-565.

Felsenstein, J. (1995). **PHYLIP, Phylogenetic Inference Package. Versão 3.5c**. Washington University. Programa distribuído pelo autor.

Goldman, N. (1993). Statistical tests of model of DNA substitution. **J. Mol. Evol. 37**: 650-661.

Goldman, N., Anderson, J.P. e Rodrigo, A.G. (2000). Likelihood-based tests of topologies in phylogenetics. **Syst. Biol 49**: 652-670.

Green, S. e Chambom, P. (1986). A superfamily of potentiallyoncogenic hormone receptors. **Nature 324**: 615-617.

Guindon, S, e Gascuel, O. (2003). A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. **Syst. Biol. 52**: 696-704.

Haddrath, O. e Baker, A.J. (2001) Complete mitochondrial DNA genome sequences of extinct birds: ratite phylogenetics and the vicariance biogeography hypothesis. **Proc. R. Soc. Lond. B 268**: 939-945

Härlid, A. e Arnason, U. (1998). Analyses of mitochondrial DNA nest ratite birds within the Neognathae: supporting a neotenous origin of ratite morphological characters. **Proc. R. Soc. Lond. B 266**: 305-309.

Hasegawa, M. e Fujiwara, M. (1993). Relative efficiencies of the maximum likelihood, maximum parsimony, and neighbor-joining methods for estimating protein phylogeny. **Mol. Phylogen. Evol. 2**: 1-5.

Hasegawa, M., Kishino, H. e Yano T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. **J. Mol. Evol. 32**: 443-445.

Hasegawa, M., Kishino, H. e Yano, T. (1989). Estimation of branching dates among primates by molecular clocks of nuclear DNA which slowed down in Hominoidea. **J. Hum. Evol. 18**: 461-476.

Hasegawa, M., Yano, T. e Kishino, H. (1984). A new molecular clock of mitochondrial DNA and the evolution of hominoids. **Proc. Japan Acad. B60**: 95-98.

Hillis, D.M., Huelsenbeck, J.P. e Cunningham, C.W. (1994). Application and accuracy of molecular phylogenies. **Science 264**: 671-677.

Hillis, D.M., Mable, B.K. e Moritz, C. (1996). Applications of Molecular Systematics: the state of the field and a look to the future. **In** Hillis, D.M., Moritz C. e Mable B.K. (eds.). **Molecular Systematics**. 2ª edição, Sinauer Assoc., Sunderland, MA, pp. 515-534.

Huelsenbeck, J.P. (1995). The performance of phylogenetic methods in the four-taxon case. **Syst. Biol. 44**: 17-48.

Huelsenbeck, J.P. e Crandall, K.A. (1997). Phylogeny estimation and hypothesis testing using maximum likelihood. **Annu. Rev. Ecol. Syst. 28**: 437-466.

Huelsenbeck, J.P., Larget, B. e Swofford, D. (2000). A compound Poisson process for relaxing the molecular clock. **Genetics 154**: 1879-1892.

Huelsenbeck, J.P. e Rannala, B. (1997). Phylogenetic methods come of age: testing hypothesis in an evolutionary context. **Science 276**: 227-232.

Jukes, T.H. e Cantor, C.R. (1969). Evolution of protein molecules. **In** Munro, H.M. (ed), **Mammalian Protein Metabolism**. Academic Press, New York, pp. 21-132.

Kelchner, S.A. e Thomas, M.A. (2006). Model use in phylogenetics: nine key questions. **Trends Ecol. Evol.22**:2.

Kimura, M. (1980). A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucletide sequences. **J. Mol. Evol. 16**: 111-120.

Kishino, H. e Hasegawa, M. (1989). Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. **J. Mol. Evol. 29**: 170-179.

Kishino, H. e Hasegawa, M. (1990). Converting distance to time: application to human evolution. **Methods Enzymol. 183**: 550-570.

Kishino, H., Miyata, T. e Hasegawa, M. (1990). Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. **J. Mol. Evol. 31**: 151-160.

Kishino, H., Thorne, J.L. e Bruno, W.J. (2001). Performance of a divergence time estimation method under a probabilistic model of rate evolution. **Mol. Biol. Evol. 18**: 352-361.

Kuhner, M.K. e Felsenstein, J. (1994). A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. **Mol. Biol. Evol. 11**: 459-468.

Kumar, S. (1996a). Patterns of nucleotide substitution in mitochondrial protein coding genes of vertebrates. **Genetics 143**: 537-548.Kumar, S. (1996b). **PHYLTEST: a program for testing phylogenetic hypothesis, Version 2.0**. Institute of Molecular Evolutionary Genetics and Department of Biology. The Pennsylvania State University, University Park, Pennsylvania, USA.

Mallat, J. e Sullivan, J. (1998). 28S and 18S rDNA sequences support the monophyly of lampreys and hagfishes. **Mol. Biol. Evol. 15**: 1706-1718.

Mindell, D.P., Sorenson, M.D., Dimcheff, D.E., Hasegawa, M., Ast, J.C. e Yuri, T. (1999). Interordinal relationships of birds and other reptiles based on whole mitochondrial genome. **Syst. Biol. 48**: 138-152.

Miyaki, C.Y., Matioli, S.R., Burke, T. e Wajntal, A. (1998). Parrot evolution and paleogeographical event: mitochondrial DNA evidence. **Mol. Biol. Evol.15**: 544-551.

Nei, M. e Kumar, S. (2000). **Molecular evolution and phylogenetics**. Oxford University Press, New York.

Nei, M., Kumar, S. e Takahashi, K. (1998). The optimization principle in phylogenetic analysis tends to give incorrect results when the number of nucleotides or amino acids used is small. **Proc. Natl. Acad. Sci. USA 95**: 12390-12397.

Nei, M., Takezaki, N. e Sitnikova, T. (1995). Assessing molecular phylogenies. **Science 267**: 253-255.

Nishihara, H., Okada, N. e Hasegawa, M. (2007). Rooting the eutherian tree: the power and pitfalls of phylogenomics. **Genome Biol. 8**:R199.

Ota, S. e Li, W.-H. (2000). NJML: a hybrid algorithm for the neighbor-joining and maximum-likelihood methods. **Mol. Biol. Evol. 17**: 1401-1409.

Page, S.L. e Goodman, M. (2001). Catarrhine phylogeny: Noncoding DNA evidence for a diphyletic origin of the mangabeys and for a human-chimpanzee clade. **Mol. Phylogen. Evol. 18**: 14-25.

Pereira, S.L. (2000). **Filogenia e Evolução Molecular em Cracidae (Aves)**. Tese de Doutorado, Instituto de Biocências, USP, São Paulo.

Rodriguez, F., Oliver, J.L., Marín, A. e Medina, R. (1990). The general stochastic model of nucleotide substitution. **J. Theor. Biol. 142**: 485-501.

Rodriguez-Robles, J.A., Stewart, G.R. e Papenfuss, T.J. (2001). Mitochondrial DNA-based phylogeography of North American rubber boas, *Charina bottae* (Serpentes: Boidae). **Mol. Phylogen. Evol. 18**: 227-237.

Russo, C.A.M., Takezaki, N. e Nei, M. (1996). Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny. **Mol. Biol. Evol. 13**: 525-536.

Sanderson, M. J. (1997). A nonparametric approach to estimating divergence times in the absence of rate constancy. **Mol. Biol. Evol. 14**: 1218-1232.

Sanderson, M. J. (2002). Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. **Mol Biol Evol 19**, 101-9.

Sarich, V.M. e Wilson, A.C. (1967). Immunological time scale for hominid evolution. **Science 158**: 1200-1203.

Shi, X., Gu, H., Susko, E. e Field (2005). The comparison of the confidence regions in phylogeny. **Mol. Biol. Evol. 22**:2285-2296.

Shimodaira H (2002) An approximately unbiased test of phylogenetic tree selection. **Syst Biol 51**:492-508.

Stamakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22: 2688-2690.

Stapel, S.O., Leunissen J.A.M., Versteeg, M., Wattel, J. e de Jong, W.W. (1984). Ratite as oldest offshoot of avian stem – evidence from alpha-crystallin A sequences. **Nature 311**: 257-259.

Strimmer, K. e von Haeseler, A. (1996). Quartet puzzling: a quartet maximum-likelihood method for reconstructiong tree topologies. **Mol. Biol. Evol. 13**: 964-969.

Sullivan J., Holsinger, K.E. e Simon, C. (1996). The effect of topology on estimates of among-site rate variation. **J. Mol. Evol. 42**: 308-312.

Sullivan, J. Joyce, P. (2005). Model selection in phylogenetics. **Annu. Rev. Ecol. Syst. 36**: 445-466.

Swofford, D.L. (1999). **PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods), versão 4.0**. Sinauer Associates, Sunderland, Massachusetts, USA.

Swofford, D.L., Olsen, G.J., Waddell, P.J. e Hillis, D.M. (1996). Phylogenetic inference. **In** Hillis, D.M., Moritz C. e Mable B.K. (eds.). **Molecular systematics**. 2ª edição, Sinauer Assoc., Sunderland, MA, pp. 407-514.

Takahashi, K. e Nei, M. (2000). Efficiencies of fast algorithms of phylogenetic inference under the criteria of maximum parsimony, minimum evolution and maximum likelihood when a large number of sequences are used. **Mol. Biol. Evol. 17**: 1251-1258.

Takezaki N., Razhetsky, A. e Nei, M. 1995. Phylogenetic test of the molecular clock and linearized trees. **Mol. Biol. Evol. 12**: 823-833.

Takezaki, N. e Gojobori, T. (1999). Correct and incorrect vertebrate phylogenies obtained by the entire mitochondrial DNA sequences. **Mol. Biol. Evol. 16**: 590-601.

Tamura, K. e Nei, M. (1993). Estimation of the number of base nucleotide substitution in the control region of mitochondrial DNA in humans and chinpanzees. **Mol.Biol. Evol. 10**: 512-526.

Tamura, K, Dudley, J. Nei, M, e Kumar, S. (2007). MEGA4: Molecular Evolutionary Genetics Analysis (MEGA). **Mol. Biol. Evol. 24**: 1596-1599.

Thorne, J.L. and Kishino H (2002) Divergence time and evolutionary rate estimation with multilocus data. **Syst Biol 51**:689-702.

Yang, Z. (1994). Estimating the pattern of nucleotide substitution. **J. Mol. Evol. 39**: 105-111.

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. **CABIOS 13**: 555-556.

Yang, Z. (1998a). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. **Mol.Biol. Evol. 15**: 568-573.

Yang, Z. (1998b). **Phylogenetic Analysis by Maximum Likelihood (PAML), Version 1.4**. University College London. Programa distribuído pelo autor.

Yang, Z. e Kumar S. (1996). Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rates among sites. **Mol. Biol. Evol. 13**: 650-9.

Yang, Z., Goldman, N. e Friday, A.E. (1994). Comparison of models for nucleotide substitution used in maximum likelihood phylogenetic estimation. **Mol. Biol. Evol. 11**: 316-324

Yang, Z., Lauder, I.J. e Lin, H.J. (1995). Molecular evolution of the hepatitis B virus genome. **J. Mol. Evol. 41**: 587-596.

Yoder, A.D. e Yang, Z. (2000). Estimation of primate speciation dates using local molecular clocks. **Mol. Biol. Evol. 17**: 1081-1090.

Zwickl, D. (2006). **Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion.** Tese de Doutorado. University of Texas, Austin, TX.

Anexo. Programas de inferência filogenética e outros relacionados aos métodos probabilísticos.

GARLI – Implementa um algoritmo genético eficiente e rápido para buscar uma árvore filogenética sobre o critério de máxima verossimilhança. Apresenta também uma opção de realizar *bootstraps* de máxima verossimilhança.

LINTREE - Permite detectar OTUS que apresentam taxas de evolução estatisticamente diferentes da média; gera árvores sob vários modelos evolutivos, porém sob o critério de distâncias. Programa pouco popular hoje em dia devido ao desenvolvimento de métodos de relógios moleculares dispersos.

MEGA4 – Pacote de análises filogenéticas e evolutivas. Implementa teste de taxas relativas de substituição, estimativas de diversas estatísticas (composição de base, taxa de transição/transversão, e outras).

MODELTEST - Usado para testar quais diferentes modelos de evolução se adequam às sequências em questão; faz o teste de razão de verossimilhança (LRT), critério de informação de Akaike (AIC) e critério de informação Bayesiana (BIC); fornece algumas estatísticas básicas sobre a escolha do melhor método.

MOLPHY – Pacote que inclui programas para se obter número de transição e transversão entre os táxons e estimar sua filogenia; permite ainda fazer rearranjos em árvores fornecidas pelo usuário.

PAML – Pacote que inclui os métodos de parcimônia e de verossimilhança, incluindo modelos para reconstrução de filogenia baseado em sequência de aminoácidos; implementa modelos de relógio molecular global e local.

PAUP* – Multipacote que inclui todos os métodos de inferência filogenética, realiza testes de *bootstrap*, *jacknife*, Kishino-Hasegawa e outros; implementa diversos tipos de modelos de relógio molecular, que permitem acomodar variação das taxas de substituição entre as OTUs.

PHYLIP – Multipacote que inclui todos os métodos de inferência filogenética, realiza *bootstrap*, redesenha árvores. Permite usar dados binários, como os obtidos para fragmentos de restrição.

PHYML – Incorpora um algoritmo rápido de buscas de árvores filogenéticas e *bootstrap* por máxima verossimilhança para conjunto de dados com grande número de sequências.

PHYLTEST – Realiza teste de constância de taxa de substituição entre os sítios, teste de ramos interiores e outros.

RAxML-VI – Busca rápida e eficiente da árvore mais verossímil para um grande número de sequências.

TREEPUZZLE – Apresenta dois novos métodos. Um deles avalia o grau de sinal filogenético contido em um alinhamento (*likelihood mapping*). O outro realiza uma análise de todos os possíveis quartetos (*quartet puzzling*) para minimizar o tempo gasto no cômputo da probabilidade de uma árvore. Realiza ainda o teste de Kishino-Hasegawa entre árvores diferentes. Testa a hipótese de relógio molecular global.

Página deixada em branco

Capítulo 15

Reconstrução filogenética: Inferência bayesiana

Dr Sergio Luiz Pereira (sergiolp@gmail.com)
Department of Natural History
Royal Ontario Museum

"*Minhas idéias abstratas, de tanto as tocar, tornaram-se concretas: são rosas familiares que o tempo traz ao alcance da mão, rosas que assistem à inauguração de eras novas.*" (Murilo Mendes, *Poesias*)

15.1. Inferência Filogenética Bayesiana

O método bayesiano de inferência filogenética é uma alternativa aos métodos tradicionais vistos nos Capítulos 12 a 14, introduzido na área de inferência filogenética mais recentemente (Huelsenbeck *et al.*, 2001; Mau *et al.* 1999; Rannala e Yang, 1996; Yang e Rannala, 1997). Originalmente, no entanto, a estatística bayesiana surgiu em 1763, quando o reverendo Thomas Bayes propôs um método formal para incorporar a distribuição ou valores de uma variável obtidos *a priori* no cálculo da probabilidade de ocorrência de um dado evento. Em inferência filogenética bayesiana, por exemplo, informações *a priori* assumem várias formas de como atribuir maior probabilidade a uma dada topologia, assumir uma distribuição pré-delimitada para os possíveis valores que os parâmetros do modelo evolutivo possam assumir ou mesmo restringir quais OTUs (*Operational Taxonomic Units*, em inglês) devem ser mais relacionadas entre si em exclusão de outras. Como veremos mais adiante, a atribuição de informações *a priori* constitui um dos tópicos mais controversos em inferência filogenética bayesiana. O método bayesiano tornou-se bastante popular ultimamente graças ao desenvolvimento das simulações Monte Carlo via cadeias de Markov (MCMC, do inglês *Markov chain Monte Carlo*), que facilitam a obtenção de estimativas de algumas variáveis impossíveis de serem estimadas analiticamente e à distribuição gratuita de diversos programas de reconstrução filogenética, como MrBayes (Ronquist e Huelsenbeck, 2003), BEAST (Drummond e Rambaut, 2007) e BAMBE (Simon e Larget, 2000).

Os métodos bayesiano e de máxima verossimilhança (em inglês, *Maximum Likelihood*, ML) são matematicamente relacionados. No Capítulo 14, vimos que o método de verossimilhança estima a probabilidade L de que os dados observados D tenham sido gerados perante uma hipótese H, a qual inclui um modelo evolutivo θ, uma topologia T e os comprimentos de seus ramos B. O método bayesiano, por sua vez, inverte o problema e calcula a probabilidade posterior $L(H|D)$ de que H seja correta uma vez obtido D e assumindo previamente determinados valores e distribuições (probabilidades *a priori*). A probabilidade condicional é representada pelo teorema de Bayes,

$$L(H \mid D) = \frac{L(H) \times L(D \mid H)}{L(D)},$$

em que $L(H)$ é a probabilidade *a priori* atribuída à hipótese H antes que D tenha sido observado, $L(D \mid H)$ é a função que maximiza a verossimilhança e $L(D)$ é uma constante de normalização que limita a probabilidade posterior entre o intervalo 0 e 1. Em outras palavras, o método bayesiano procura a hipótese H em que a probabilidade posterior é máxima; a probabilidade posterior é proporcional à verossimilhança multiplicada pela probabilidade *a priori* de a hipótese ser correta.

O método bayesiano trata dados e parâmetros como variáveis aleatórias nas quais a incerteza sobre os valores dos parâmetros é medida pela distribuição da probabilidade posterior (Huelsenbeck e Ronquist, 2005). A probabilidade posterior dos parâmetros de interesse (por exemplo, topologia) é calculada pela função integral de todos os valores possíveis dos outros parâmetros, cada um pesado por sua probabilidade. Isso implica que a inferência filogenética não é influenciada por um valor específico de um parâmetro qualquer. Uma vez estimada a distribuição de probabilidade de um parâmetro, a média, a variância e a dispersão associadas à maior proporção da probabilidade posterior podem ser calculadas. Em contraste, o método de verossimilhança considera que as variáveis da hipótese H possuem valores não aleatórios. Dessa forma, as probabilidades não podem ser atribuídas diretamente; sob o metódo de máxima verossimilhança, para estimar-se a incerteza associada com valores dos parâmetros, deve-se, portanto, comparar a distribuição das estimativas de máxima verossimilhança obtidas para vários conjuntos de dados independentes e com o mesmo tamanho que o conjunto original (Huelsenbeck e Ronquist, 2005).

Uma outra diferença fundamental entre ML e o método bayesiano é como os parâmetros do modelo evolutivo são estimados, o que pode levar ambos os métodos a escolherem diferentes filogenias para um mesmo grupo de sequências e o mesmo modelo de substituição de DNA. Muitas vezes, esses parâmetros não são de interesse direto, mas devem ser estimados nas equações de verossimilhança. No método de verossimilhança, os parâmetros do modelo evolutivo são estimados conjuntamente para achar o ponto onde a verossimilhança é máxima. Por sua vez, o método bayesiano não estima a altura da probabilidade posterior, mas o volume (ou densidade) da superfície da probabilidade posterior, e os parâmetros são integrados para obter a probabilidade posterior marginal da árvore filogenética.

Consideremos apenas um parâmetro x para ilustrar este processo, o qual pode ser qualquer variável do modelo de substituição, como, por exemplo, a taxa de transição/transversão, composição de bases ou taxa de substituição do nucleotídeo i por j. A Fig. 15.1a representa as probabilidades marginais de duas árvores alternativas obtidas pelo método bayesiano. Embora a árvore 1 tenha a maior probabilidade posterior medida pela altura do pico, a área sob a superfície da probabilidade posterior é muito menor que aquela da árvore 2 (Fig. 15.1a). Portanto o método bayesiano favorece a árvore 2. A Fig. 15.1a mostra ainda que embora a árvore 1 apresenta o maior pico, a árvore 2 é apoiada por uma gama maior de valores para o parâmetro x. Esse fato é particularmente interessante, porque

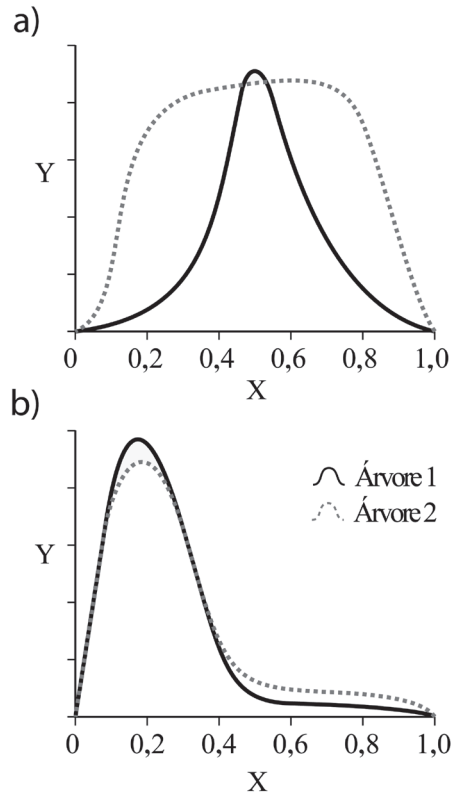


Figura 15.1. Estimativas de probabilidades marginal (a) e conjunta (b). O eixo da abscissa representa um parâmetro hipotético *x* de interesse não direto e o eixo da ordenada representa a verossimilhança ou a densidade da probabilidade posterior, nos métodos de máxima verossimilhança e bayesiano, respectivamente. Em ambos os casos, a estimativa conjunta de *x* e da árvore favorece a árvore 1, que apresenta o pico mais alto. A estimativa marginal, no entanto, favorece a árvore 2, que apresenta o maior volume sob a superfície da probabilidade posterior. Modificado de Holder e Lewis (2003).

ele ilustra que a estimativa de parâmetros contém incertezas e a "árvore verdadeira" pode ser próxima da árvore mais verossímil ou a com maior probabilidade posterior (Holder e Lewis, 2003). Embora pareça óbvio que a estimativa marginal possa ser superior à estimativa conjunta, a estimativa marginal requer que os parâmetros sejam integrados de acordo com sua probabilidade posterior e, portanto, requer o uso de probabilidades *a priori*. A escolha de probabilidades *a priori* pode ser subjetiva e enviesar a densidade das probabilidades posteriores (Alfaro e Holder, 2006), especialmente quando a inferência é baseada em um conjunto de dados pequeno.

Por outro lado, a estimativa conjunta de *x* e da filogenia favorece a árvore 1, que apresenta o maior pico na distribuição de verossimilhanças (Fig. 15.1b). Observe que a densidade de *x* na Fig. 15.1b favorece valores menores do que 0,4. O uso de estimativas conjuntas em inferência bayesiana, no entanto, iria favorecer também a árvore 2 neste caso, porque sua área de distribuição sob a superfície da verossimilhança é maior. A não ser que exista evidência contra valores de *x* maiores que 0,4, o método bayesiano pode resultar em estimativas enviesadas, especialmente se essas estimativas são grandemente influenciadas pelas probabilidades *a priori*.

Para fins de ilustração de como a inferência bayesiana difere da inferência estatística clássica (incluindo aqui o método de máxima verossimilhança), considere que em uma caixa existam 100 dados, sendo 90 deles perfeitos e 10 imperfeitos (material suplementar em Huelsenbeck *et al.*, 2001). A probabilidade de observar cada face do dado é idêntica para cada face nos dados perfeitos, mas difere para cada face nos dados imperfeitos. Considere que as probabilidades sejam:

Face do dado:	1	2	3	4	5	6
Dado perfeito:	$^1/_6$	$^1/_6$	$^1/_6$	$^1/_6$	$^1/_6$	$^1/_6$
Dado imperfeito:	$^1/_{21}$	$^2/_{21}$	$^3/_{21}$	$^4/_{21}$	$^5/_{21}$	$^6/_{21}$

Suponha que selecionamos aleatoriamente um dado da caixa e o jogamos duas vezes. Na primeira vez, obtivemos a face 4 e na segunda vez, 6. Qual é a probabilidade de que o dado escolhido seja imperfeito? Para a estatística clássica, devemos calcular simplesmente a probabilidade de obter 4 e de obter 6 para cada tipo de dado, assumindo que ambas as jogadas do dado são independentes, ou seja:

$$\Pr(4,6 \mid perfeito) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36},$$

e

$$\Pr(4,6 \mid imperfeito) = \frac{4}{21} \times \frac{6}{21} = \frac{24}{441}$$

A probabilidade de observar as faces 4 e 6 é 1,96 vezes maior, considerando a hipótese que o dado é imperfeito. Portanto, podemos concluir, com base nas probabilidades, que o dado é imperfeito.

Em inferências bayesianas, devemos considerar nosso conhecimento prévio sobre as probabilidades de obter cada face em um dado perfeito e um imperfeito. Sabemos que há 10 dados imperfeitos em um conjunto total de 100 dados, Portanto, podemos assumir *a priori* que a probabilidade de o dado ser imperfeito é 0,1. No entanto, se esta informação não nos fosse acessível, a atribuição de uma probabilidade *a priori* pode ser impraticável em termos objetivos. Portanto sob a óptica bayesiana temos que:

$$\Pr(imperfeito \mid 4,6) = \frac{\Pr(imperfeito) \times \Pr(4,6 \mid imperfeito)}{\Pr(4,6 \mid imperfeito) \times \Pr(imperfeito) + \Pr(4,6 \mid perfeito) \times \Pr(perfeito)}$$

onde Pr(*imperfeito*) e Pr(*perfeito*) são as probabilidades *a priori* de que o dado é imperfeito e perfeito, respectivamente. Podemos assumir como probabilidades *a priori* a proporção de dados imperfeitos e perfeitos na caixa. Assim, temos que:

$$\Pr(imperfeito \mid 4,6) = \frac{\frac{1}{10} \times \frac{24}{441}}{\frac{24}{441} \times \frac{1}{10} + \frac{1}{36} \times \frac{9}{10}} = 0,179.$$

Nossa opinião de que o dado usado foi um dado imperfeito passou, portanto, de 0,1 (a proporção da dados imperfeitos na caixa) para 0,18 (a probabilidade posterior obtida quando incorporamos informação prévia sobre a distribuição de dados perfeitos e imperfeitos na caixa).

15.2. Probabilidade *a priori*

A escolha de probabilidades *a priori*, ou simplesmente, *priori*, é causa de muita controvérsia, principalmente em discussões entre os adeptos da estatística clássica e os adeptos da estatística bayesiana. Probabilidades bayesianas *a priori* podem

ser originadas de expectativas teóricas, experiência prévia ou observações obtidas de dados anteriores. Por exemplo, no caso do jogo de dados, dificilmente temos informação sobre a proporção da dados imperfeitos na vida real. Portanto podemos esperar *a priori* que a probabilidade de tirar qualquer face é um sexto. Similarmente, em um jogo de cara ou coroa, esperamos obter cara em 50% das jogadas de uma moeda, se não sabemos nada a respeito da imperfeição desta moeda. No entanto, diferentes pesquisadores possuem percepções diferentes do mesmo problema. Por exemplo, no jogo de cara ou coroa o leitor pode assumir que a probabilidade de se obter cara em número finito de jogadas cai entre 0,4 e 0,6, enquanto o autor deste capítulo pode considerar que esta mesma probabilidade tem distribuição entre 0,45 e 0,55. A subjetividade na definição da probabilidade *a priori* dada por diferentes pesquisadores é a principal causa de objeção contra o uso da estatística bayesiana.

No entanto, a incorporação de probabilidades *a priori* tem uma base sólida e formal em estatística bayesiana. Embora *apriorismos* possam ser subjetivos, eles não são necessariamente arbitrários e não assumem valores específicos, mas uma distribuição de possíveis valores. Em se tratando de sequências macromoleculares, sabemos bastante a respeito dos processos que regem sua evolução e parte dessa informação pode ser útil para limitar a distribuição de valores de um parâmetro específico do modelo evolutivo. Por exemplo, se existe um viés em composição de bases em sequências mitocondriais (veja Tabela 14.2 no Capítulo 14), é perfeitamente justificável o uso *a priori* de uma distribuição não uniforme para cada uma das quatro bases de DNA. Similarmente, as taxas de transição e transversão podem ser facilmente estimadas das sequências e uma distribuição ao redor dessa estimativa pode ser usada como a possível gama de valores para as razão de transição/transversão. Uma grande vantagem do método bayesiano é que a probabilidade posterior passa a ser menos afetada pela probabilidade *a priori* com o aumento de observações realizadas. Por exemplo, os três gráficos na Figura 15.2 foram gerados assumindo que a probabilidade *a priori* de obtermos cara se limita ao intervalo [0,4; 0,6] em 95% dos casos, medido pela área sob a superfície da distribuição da probabilidade *a priori*. No entanto, obtivemos cara em 60% das jogadas nos três experimentos e a distribuição das probabilidades *a priori* e posterior passam a ser mais distintas entre si, com o aumento do número de observações.

Ainda nesse exemplo, o *priori* é dito informativo porque é dada uma maior probabilidade do valor de interesse cair em um sub-intervalo pré-definido dentro do intervalo [0, 1] de possíveis valores de probabilidade. Alternativamente, a distribuição da probabilidade *a priori* pode ser definida por *prioris* não informativos, também chamados de *prioris* vagos ou dispersos. *Prioris* dispersos refletem nossa ignorância sobre um determinado parâmetro. Compare os gráficos da Figura 15.3. A distribuição da probabilidade posterior é menos afetada pelo uso de *prioris* dispersos. Quando não temos conhecimento *a priori* da distribuição de possíveis valores para um parâmetro qualquer, o uso de *prioris* dispersos pode ser considerado uma vantagem do método bayesiano, cujo efeito na distribuição da probabilidade posterior, espera-se, é negligenciável. No entanto, em alguns casos, a probabilidade posterior tende a ser centrada no valor médio da distribuição *a priori*, especialmente com um número limitado de dados.

As distribuições ilustradas acima são válidas apenas para dados binomiais, como o jogo de moedas, ou no caso de sequências de DNA, para a razão de transição/transversão, que se adequam a uma distribuição beta da forma Beta(α, β) em um espaço bidimensional. Outros parâmetros do modelo evolutivo, como as taxas de substituição nucleotídica ou a proporção de bases, requerem um outro tipo de distribuição multimodal, como a distribuição de Dirichlet, com *n* dimensões, em que *n* é o número de parâmetros a serem estimados.

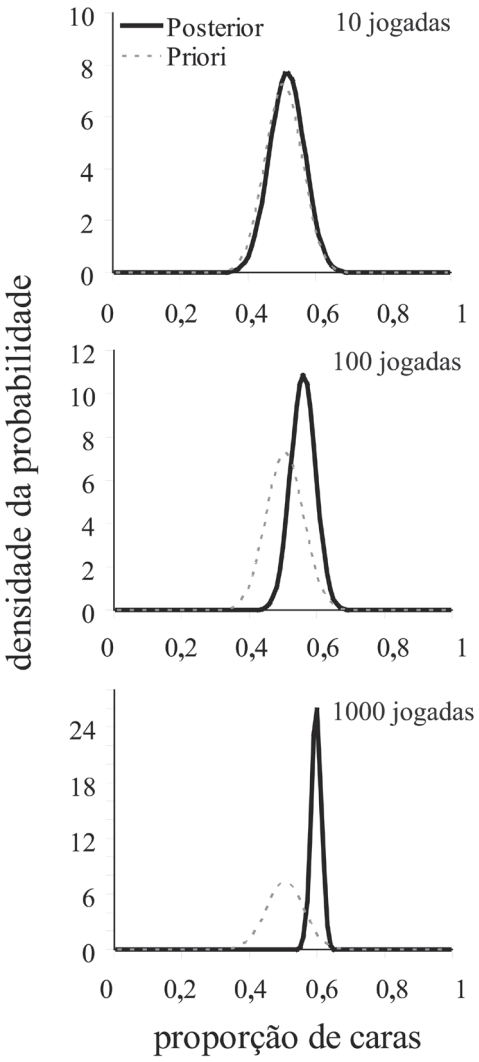


Figura 15.2. Distribuição da probabilidade *a priori* e da probabilidade posterior jogando uma moeda 10, 100 e 1000 vezes e obtendo cara em 60% das jogadas. A expectativa *a priori* de se obter cara em 95% dos casos é mantida constante no intervalo de 0,4 a 0,6 nos três gráficos. Note que, com o aumento do número de jogadas, a probabilidade *a priori* influencia menos a distribuição da probabilidade posterior.

15.3. Probabilidade Posterior

Uma diferença fundamental entre estatística clássica e bayesiana é a maneira como o apoio aos clados é gerado e interpretado. Métodos de distância, máxima parcimônia e máxima verossimilhança usualmente empregam o uso de *bootstrap* não paramétrico como uma medida da confidência dos nós internos de uma filogenia (Capítulo 13). A técnica de *bootstrap* re-amostra, como foi visto, a matriz de dado original e gera um número determinado de matrizes (pseudo-réplicas) com o mesmo número de caracteres existente na matriz original. A amostragem é feita com reposição. Para cada pseudo-réplica, uma árvore é gerada de acordo com o critério de otimização escolhido. A frequência com que os agrupamentos mais comuns são encontrados nas árvores estimadas pelas pseudo-réplicas são indicados na topologia. Por exemplo, um valor de 85% em um nó interno da árvore consenso foi observado em 85% das topologias geradas pelas pseudo-réplicas. Os valores de *bootstrap* são uma medida da repetibilidade

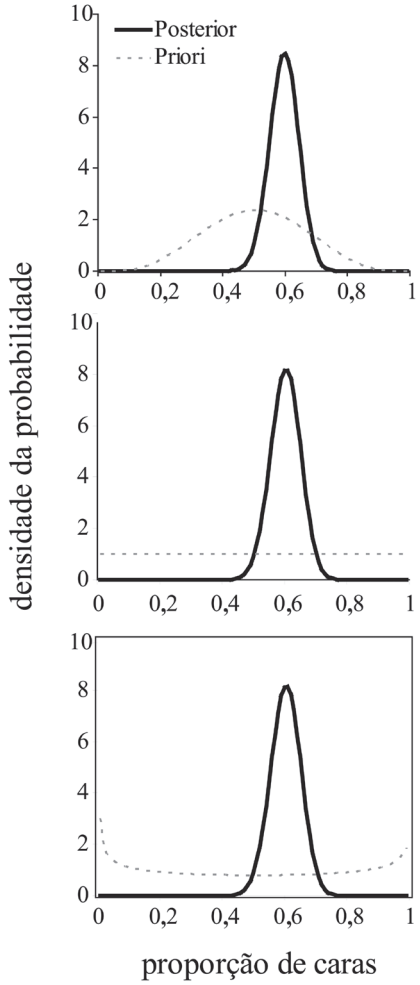


Figura 15.3. Distribuições alternativas da probabilidade *a priori*. A probabilidade posterior é centrada em 0,6. O gráfico superior corresponde a um *priori* bem delimitado no intervalo [0,1] de se obter cara em 95% dos casos em 100 jogadas do moeda. O gráfico central e o gráfico inferior representam *prioris* dispersos. No gráfico central, a probabilidade *a priori* é igualmente distribuída no intervalo [0, 1]. Já o gráfico inferior representa uma distribuição bimodal com grande variância.

dos resultados. Conjuntos de dados que apresentam sinais filogenéticos incompatíveis podem resultar em valores de *bootstrap* discordantes para um mesmo agrupamento de sequências.

Em análises bayesianas, resumimos uma amostra da distruibuição posterior de árvores filogenéticas em termos de frequência (probabilidade posterior) de clados individuais. Probabilidades posteriores são baseadas na distribuição dos parâmetros do modelo evolutivo, os quais são estimados conjuntamente com a topologia e interpretados com a medida direta da incerteza associada à estimativa. A probabilidade posterior de um clado é interpretada como a probabilidade de que aquele clado é verdadeiro, dado que a hipótese *H* é verdadeira.

Comparações diretas de valores de *bootstrap* e probabilidades posteriores levaram à conclusão de que probabilidades posteriores são infladas e colocam muita confidência nos clados (ver Suzuki *et al.*, 2002; Douady *et al.*, 2003). No entanto, *bootstrap* e probabilidades posteriores medem diferentes aspectos da reconstrução filogenética. Para avaliar a questão mais a fundo, Huelsenbeck e Rannala (2004) geraram um modelo filogenético onde a topologia, comprimento de ramos e parâmetros do modelo evolutivo foram amostrados aleatoriamente de uma distribuição de probabilidades *a priori*. Então, uma matriz de dados foi gerada com base no modelo filogenético e a inferência filogenética bayesiana foi realizada por meio de MCMC usando o mesmo modelo usado para gerar as matrizes. Similarmente, modelos mais simples ou mais complexos também foram usados para inferir a filogenia simulada.

Huelsenbeck e Rannala (2004) demonstraram por meio de simulações que a probabilidade posterior mede a probabilidade de que o clado seja verdadeiro. Por exemplo, a relação entre a probabilidade posterior e a probabilidade de recuperar a filogenia correta é maior quando as análises bayesianas foram realizadas usando modelos evolutivos mais próximos do modelo usado para gerar a matriz de dados (Figura 15.4). No entanto, quanto maior a violação das premissas do modelo usado para gerar as sequências, mais distante a relação entre a probabilidade posterior e probabilidade de a filogenia correta ser recuperada. Além disso, Huelsenbeck e Rannala (2004) também mostraram que a relação entre as probabilidades é menos afetada quando o modelo usado para gerar os dados são mais simples que o modelo usado na análise bayesiana.

Finalmente, as simulações de Huelsenbeck e Rannala (2004) também mostraram que, quando as premissas do modelo evolutivo que geraram as sequências são incorporadas na análise (JC69 *versus* JC69), valores de *bootstrap* e probabilidades posteriores são similares (Figura 15.5, gráficos à direita). No entanto, a análise de *bootstrap* parece ser menos influenciada pela especificação errônea do modelo de substituição (GTR+Γ versus JC69; Figura 15.5, gráficos à esquerda). Na prática, não sabemos como os dados foram gerados (isto é, não conhecemos os processos evolutivos que geraram as sequências) e os valores de *bootstrap* e de probabilidades posteriores podem diferir grandemente. Enquanto se considera que um clado que receba um valor de *bootstrap* maior ou igual a 75% seja relativamente bem sustentado pelos dados coletados, apenas valores de probabilidade posterior maior ou igual a 95% são consideradas sustentação forte para um clado.

15.4. Algoritmos de Monte Carlo via Cadeias de Markov

A probabilidade posterior de uma árvore filogenética é resultado da somatória de todas as váriaveis envolvidas no problema a ser estimado. Assim sendo, não há métodos analíticos

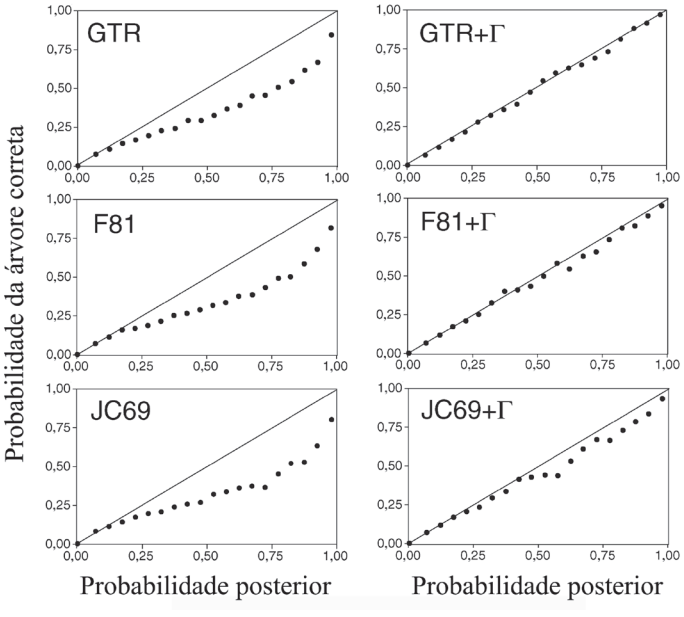


Figura 15.4. Relação entre probabilidades posteriores das filogenias e probabilidade de que a árvore estimada seja correta. O modelo evolutivo usado para gerar matrizes com 100 pares de bases foi GTR+Γ. O modelo usado durante a análise bayesiana é indicado no canto superior esquerdo para cada simulação. Modificado de Huelsenbeck e Rannala (2004).

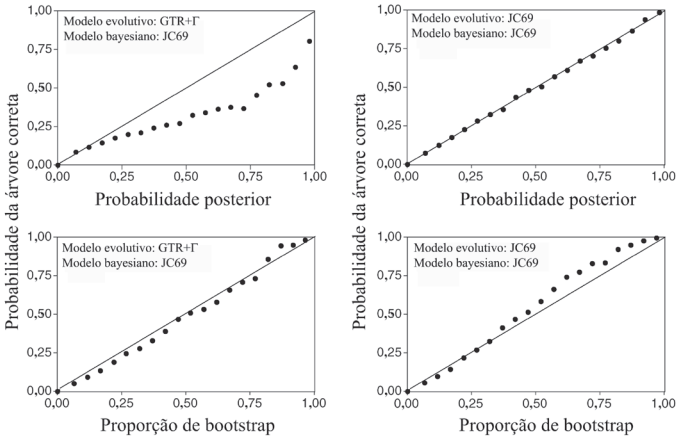


Figura 15.5. Relação entre probabilidades posteriores das filogenias (gráficos superiores) e valores de *bootstrap* (gráficos inferiores) com a probabilidade de que a árvore estimada seja correta. O modelo evolutivo usado para gerar matrizes com 100 pares de bases e o modelo usado na análise bayesiana são indicados no canto superior esquerdo para cada simulação. Modificado de Huelsenbeck e Rannala (2004).

que possam ser usados para se estimar a probabilidade posterior de problemas complexos, como no caso de inferência filogenética. No entanto, métodos de Monte Carlo via cadeias de Markov podem aproximar as estimativas da probabilidade posterior nesses casos, tornando tratáveis a inferência de parâmetros de interesse. Métodos MCMC (Hastings 1970; Metropolis *et al.,* 1953) são um grupo de algoritmos que calculam integrais multidimensionais e amostram uma distribuição de probabilidades baseadas em cadeias de Markov. Cadeias de Markov são processos estocásticos que possuem a propriedade de que dado um estado presente, estados futuros são independentes de estados passados. Em outras palavras, a descrição do estado presente reflete toda a informação que poderia influenciar estados futuros. No mais, a distribuição desejada de um grupo de parâmetros em MCMC é dada como a distribuição de equilíbrio. O estado da cadeia após um número grande de ciclos ou gerações é usado como uma amostra da distribuição de equilíbrio.

Em termos de MCMC, uma filogenia compreende um conjunto de variáveis incluindo a topologia, comprimento de ramos e parâmetros do modelo de evolução, os quais podemos denominar como um estado *p* da cadeia de Markov. O estado *p* representa um ponto no espaço multidimensional de parâmetros. Uma modificação de qualquer uma das variáveis em *p* leva a um novo ponto *p'* posicionado em outro local do espaço multidimensional. Por meio de sucessivas iterações, o espaço de variáveis é investigado e amostras do conjunto de variáveis em diferentes pontos do espaço amostral são registradas, construindo-se assim a cadeia de Markov.

A cadeia de Markov é construída da seguinte maneira. Se estamos no primeiro ciclo ou geração da cadeia, *p* recebe um valor entre 0 e 1. Um novo estado *p'* é proposto aleatoriamente, cuja distribuição uniforme é centrado no valor atual de *p*, ou seja, [$p - \varepsilon$, $p + \varepsilon$]. A seguir, calcula-se a probabilidade de aceitação de *p'* como

$$R = \min\left(1, \underbrace{\frac{p'^{x}(1-p')^{n-x}}{p^{x}(1-p)^{n-x}}}_{\text{razão de verossimilhança}} \times \underbrace{\frac{f(p')}{f(p)}}_{\text{razão de priori}} \times \underbrace{\frac{f(p \mid p')}{f(p' \mid p)}}_{\text{razão da proposta}}\right),$$

em que *n* é o número de vezes em que um parâmetro é observado, *x* é o número de observações (amostras) obtidas e $f(p'|p)$ e $f(p|p')$ são as probabilidades da proposta de uma mudança de *p* para *p'* e vice-versa, respectivamente. Em outras palavras, a equação avalia se o estado proposto tem probabilidade maior que o estado atual e então muda o estado atual de *p* para *p'*. Após calculado o valor de *R*, um número aleatório com distribuição uniforme [0, 1] é gerado. Se o número aleatório for menor do que *R*, o estado proposto é aceito e, então, $p = p'$. Se o número aleatório é maior do que *R*, o estado proposto *p'* não é aceito e o estado atual *p* permanece o mesmo. Esse processo de propostas é repetido várias vezes e os estados gerados ao longo da análise formam a cadeia de Markov. Esse procedimento permite que, em alguns casos, o estado atual mude para o novo estado proposto mesmo quando o estado proposto tenha probabilidade posterior ligeiramente inferior à do estado atual. A cadeia de Markov funciona de modo análogo a escalar um rochedo. Se um estado proposto implica em ocupar uma posição mais elevada no rochedo, ele será aceito. No entanto, em alguns casos será necessário recuar para uma posição menos elevada do que a atual para atingir posteriormente um posição mais elevada do que as já visitadas até o momento.

A Figura 15.6 mostra um exemplo em que, em um jogo de cara ou coroa, obtivemos 5 coroas a cada conjunto de 10 jogadas. Note que, com o aumento da cadeia, a aproximação usando MCMC converge para o valor verdadeiro (analítico) da probabilidade posterior. No entanto, em termos filogenéticos, não é possível comparar a aproximação de MCMC com o valor analítico, uma vez que este último não pode ser computado devido à complexidade do problema de inferência filogenética. Na prática, cabe ao pesquisador determinar quando interromper a cadeia de Markov e determinar se a distribuição da probabilidade posterior atingiu o equilíbrio desejado. Em princípio, esperamos que a cadeia de Markov seja iniciada em um ponto aleatório do espaço de variáveis e relativamente distante do pico da probabilidade posterior de interesse. As amostras obtidas antes de atingir o pico de interesse são grandemente influenciadas pelo valor

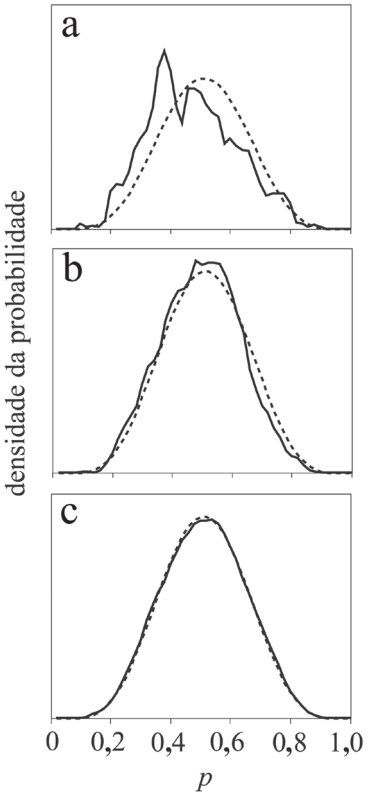


Figura 15.6. Probabilidade posterior *p* do jogo de cara ou coroa. A linha tracejada representa o cálculo analítico da probabilidade posterior e a linha contínua, a aproximação da probabilidade posterior usando MCMC. A cadeia foi gerada com 5.000 (a), 50.000 (b) e 50.0000 ciclos (c). Modificado de Huelsenbeck *et al.* (2000).

inicial da cadeia de Markov e, se esse valor não se encontrar na distribuição de equilíbrio, as amostras iniciais são enviesadas em direção ao valor inicial. Assim sendo, as amostras iniciais são descartadas e apenas as amostras tiradas da distribuição de equilíbrio são usadas para resumir os valores dos parâmetros de interesse (isto é, topologia, comprimento de ramos, parâmetros do modelo de substituição). A proporção de amostras descartada é denominada "queima" (em inglês, *burn-in*) da cadeia, ou simplesmente descarte. O valor do descarte pode ser avaliado graficamente após a cadeia de Markov ter sido interrompida. Nesse caso, observa-se o comportamento do logaritmo da verossimilhança ao longo da cadeia de Markov, como representado na Figura 15.7. Inicialmente observamos que a verossimilhança aumenta consideravelmente nos ciclos iniciais e, após um certo número de gerações, atinge um platô. As amostras obtidas no platô foram extraídas da distribuição de equilíbrio e, portanto, são uma aproximação da distribuição da probabilidade posterior. Embora as verossimilhanças das amostras apresentem distribuição limitada, as demais variáveis estimadas (isto é, proporção de sítios invariáveis, taxa de substituição entre nucleotídeos *i* e *j*, entre outras) não são necessariamente semelhantes entre as amostras.

Desde que a cadeia seja não periódica e irredutível, ela irá atingir o equilíbrio se o número de ciclos executados for grande o suficiente. Em certos casos, a cadeia de Markov requer um longo tempo para atingir o equilíbrio. Por exemplo, a escolha de valores apropriados para cada parâmetro do conjunto de variáveis é um fator muito importante porque eles influenciam a mobilidade da cadeia de Markov no espaço de variáveis. Valores muito pequenos implicam um tempo indesejavelmente maior para percorrer o espaço e atingir o equilíbrio da distribuição. Valores muito elevados, por sua vez, implicam em propostas muito distantes do estado atual que serão raramente aceitas, levando a uma amostragem local do espaço de variáveis.

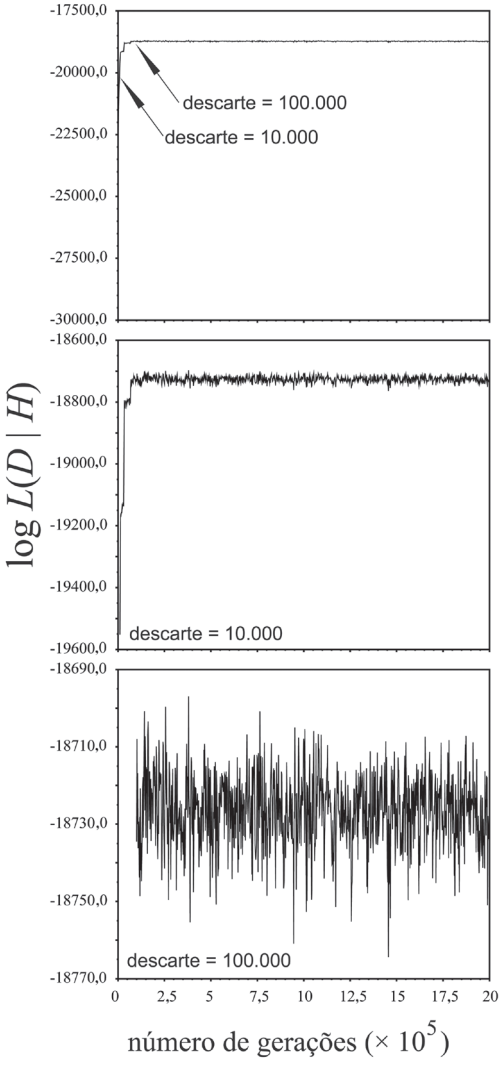


Figura 15.7. Variação do logaritmo da verossimilhança para sequências de ND2 e citocromo *b* de um grupo de aves passeriformes obtido em uma cadeia de Markov com 20 × 10^5 gerações (S. L. Pereira, não publicado). Note que, nos gráficos central e inferior, a escala no eixo do logaritmo da verossimilhança foi refinada com o aumento do número de amostras descartadas em relação ao gráfico superior.

Embora, em uma cadeia de Markov, o estado atual não dependA de estados passados, a natureza do sistema de amostragem da cadeia gera amostras que são auto-correlacionadas. O grau de auto-correlação depende em grande parte da maneira como a cadeia foi construída. Se a proposta de novos parâmetros é baseada em mudanças muito limitada dentro do possível intervalo de valores, a cadeia de Markov será representada por valores com alto grau de auto-correlação. Os estados da cadeia então representam diferentes valores de *p* e são amostras válidas, mas dependentes da probabilidade posterior (Altekar *et al.*, 2004). Para reduzir o grau de auto-correlação entre as amostras, a cadeia de Markov pode ser refinada de maneira que uma amostra seja registrada após um número determinado de ciclos. Em filogenias moleculares, geralmente as amostras são tomadas a cada 100 ou 1000 ciclos. O comprimento da cadeia é então dado por (1 + número de ciclos nos quais amostras não foram tomadas) × número de amostras registradas. A desvantagem de tal procedimento é que um número maior de ciclos deve ser gerado para que a coleta de amostras não seja enviesada em direção a determinados valores e que o espaço de variáveis tenha sido devidamente amostrado. A cadeia de Markov também sofre do problema de otimização local, isto é, os valores considerados como representativos da probabilidade posterior podem representar na realidade valores subótimos amostrados em um local limitado do espaço de variáveis. O problema é facilmente corrigido com o uso do algoritmo de Metropolis aplicado ao método de Monte Carlo via cadeias de Markov, conhecido como MCMCMC, ou $(MC)^3$, do inglês, *Metropolis-coupled Markov chain Monte Carlo*.

O método $(MC)^3$ é implementado nos programas mais frequentemente usados em inferência bayesiana como MrBayes, BEAST e BAMBE, e gera simultaneamente *n* cadeias de Markov, onde *n*-1 delas são "aquecidas" (Ronquist e Huelsenbeck, 2003). Cada cadeia é gerada da maneira descrita acima e a distribuição da probabilidade posterior é elevada à potência β (Altekar *et al.*, 2004). O parâmetro β, com valores no intervalo $0 < \beta < 1$, é o valor de aquecimento ou "temperatura" da cadeia, dado por:

$$\beta = \frac{1}{1 + T(i-1)},$$

em que *i* é cada cadeia de Markov (*i* = 0, 1, 2, ..., *n*-1) e *T* é a temperatura definida pelo pesquisador. Para *i* = 0, a cadeia é elevada à potência 1 e a distribuição da probabilidade posterior não é afetada pelo aquecimento. O aquecimento das cadeias diminui a altura dos picos locais no espaço de variáveis e permite uma melhor exploração do espaço de variáveis. Se β = 0, a cadeia atualmente é um estado plano onde a probabilidade posterior em qualquer ponto do espaço é sempre 1. A probabilidade de aceitação de novos estados é maior nas cadeias aquecidas. Portanto, cadeias aquecidas se movem mais rapidamente no espaço de variáveis do que a cadeia fria e estão menos sujeitas a ficarem presas em locais subótimos. No entanto, a cadeia de Markov é construída apenas com base nos estados observados na cadeia fria. A função das cadeias aquecidas é espalhar as buscas por todo o espaço de

variáveis de maneira mais eficiente. Quando as cadeias completam um ciclo, duas delas são comparadas aleatoriamente e uma probabilidade é estimada para estabelecer se as cadeias criadas por elas irão permutar o estado atual em que se encontram no espaço de variáveis. Se a cadeia fria é escolhida para a comparação com uma cadeia quente e a permutação é aceita, a cadeia de Markov passará a ser amostrada em uma região distinta do espaço de váriaveis e, portanto, a cadeia de Markov será constituída por estados que foram amostrados em diferentes locais do espaço de variáveis.

15.5. Escolha de Modelos e Teste de Hipóteses

O método bayesiano parece ser sensível ao modelo evolutivo usado nas análises (Fig. 15.5). No entanto, o problema é relacionado com os modelos evolutivos e não com o método bayesiano em si. A filogenia bayesiana com a maior probabilidade posterior é em geral idêntica ou muito similar à filogenia obtida por máxima verossimilhança porque os *prioris* deverão ter pouca influência na distribuição da probabilidade posterior quando temos um número suficiente de dados e a cadeia de Markov foi longa o suficiente para que as amostras representem a distribuição de equilíbrio. As simulações vistas nas Figuras 15.4 e 15.5 (Huelsenbeck e Rannala, 2004) e aquelas executadas em vários outros estudos (Huelsenbeck, 1995a, 1995b) mostram que é importante usar modelos evolutivos que sejam os mais próximos possíveis dos processos que geraram os dados. Infelizmente, mesmo os modelos evolutivos mais complexos não acomodam certos processos evolutivos importantes que geraram os dados (Galtier e Gouy, 1998, Foster, 2005). No entanto, assim como no método de máxima verossimilhança em que usamos o teste de razão de verossimilhança ou critério de Akaike para escolher o modelo que melhor representa a evolução das sequências (Capítulo 14), o teste de hipóteses em estatística bayesiana é realizado usando o fator Bayes.

O fator Bayes é definido como a razão entre o teorema de Bayes de duas hipóteses evolutivas alternativas. Considere que H_1 e H_2 sejam essas hipóteses. Para cada uma delas, o teorema de Bayes visto no início deste capítulo é usado para calcular sua probabilidade posterior. O fator Bayes de H_1 e H_2 é definido matematicamente como:

$$\boldsymbol{B} = \frac{L(H_1 \mid D)}{L(H_2 \mid D)} = \frac{L(H_1) \times L(D \mid H_1)}{L(H_2) \times L(D \mid H_2)}.$$

O fator Bayes é interpretado como a chance em favor de H_1 sobre H_2 (Lee, 1997). Em análises filogenéticas, considera-se que FB > 10 indica apoio decisivo em favor de H_1. Valores entre 10 e 5 e entre 5 e 2 são considerados evidência forte e positiva, respectivamente, para H_1. Valores entre 2 e 0 não são informativos e valores negativos favorecem H_2. As hipóteses comparadas pelo fator Bayes vão desde topologias alternativas até qualquer variação entre parâmetros específicos do modelo evolutivo. Ao contrário do teste de razão de verossimilhança, as hipóteses testadas pelo fator Bayes não precisam ser necessariamente um caso especial da outra.

15.6. Estatística clássica *versus* estatística bayesiana

Tanto a estatística clássica quanto a estatística bayesiana utilizam-se de probabilidade para avaliar a confidência estatística da ocorrência de um evento qualquer. No entanto, a interpretação da probabilidade é feita de modo diferente. A estatística clássica usa probabilidade como uma medida da repetibilidade ou frequência em que se obtém um valor pelo menos tão extremo quanto o valor observado, considerando-se verdadeira a hipótese nula, isto é, de não haver diferenças entre as réplicas de um experimento quando a maioria das variáveis são mantidas constantes. A probabilidade medida como um valor *P* discreto no intervalo [0,1] é usada para rejeitar a hipótese nula quando um valor igual ou menor do que o nível de significância é obtido. Por sua vez, a estatística bayesiana interpreta a probabilidade como uma medida direta da incerteza associada com um parâmetro. Apenas os dados observados são relevantes no cálculo da probabilidade e eventos tão ou mais extremos que os observados são irrelevantes. Contrariamente à probabilidade em estatística clássica, a probabilidade posterior bayesiana é interpretada como evidência em favor ao modelo. No mais, a probabilidade posterior bayesiana reflete a distribuição de valores do parâmetro de interesse e não a distribuição dos dados como em estatística clássica.

Para interpretar probabilidades em termos clássicos e bayesianos, consideremos, por exemplo, o jogo de uma moeda perfeita no qual observamos o número de coroas. A estatística clássica não rejeitaria a hipótese de que a moeda é perfeita quando o número de coroas é $p = 0{,}5$. No entanto, valores de p ligeiramente maiores ou menores do que 0,5, digamos, 0,48 ou 0,55, seriam suficientes para rejeitar a hipótese de que a moeda é perfeita. Portanto, o teste não verifica se a moeda é perfeita, mas se a moeda foi jogada um número suficiente de vezes para provar que ela não é perfeita. Por outro lado, a estatística bayesiana interpreta a probabilidade como uma distribuição contínua entre 0 e 1. Assim, um observador pode considerar que a moeda é perfeita *a priori* se $0{,}45 < p < 0{,}55$, por exemplo.

A confiabilidade das estimativas também é interpretada de maneira diferente. Em estatística clássica, o intervalo de confiança mede o erro de amostragem. Uma probabilidade de 95% representa uma chance de 95% de gerar um intervalo que contém o mesmo valor estimado para o parâmetro de interesse. Em estatística bayesiana, o intervalo é denominado intervalo de credibilidade e mede a incerteza associada com as estimativas dos valores dos parâmetros. A probabilidade de que o valor do parâmetro a ser estimado estará contido no intervalo de credibilidade é de 95%. As distribuições das probabilidades posteriores integram para 1 e, portanto, a área sob a distribuição de interesse é uma medida direta da variação de valores do parâmetro de interesse. Em estatística clássica, tal interpretação é impossível, porque a superfície da curva de verossimilhança não pode ser integrada para um valor definido.

15.7. Outras Aplicações de Inferência Bayesiana em Biologia Evolutiva

O método bayesiano tem sido aplicado em diversas áreas da biologia evolutiva. Similarmente à infererência de filogenias, o atrativo do método encontra-se na possibilidade de acomodar a incerteza associada aos parâmetros de interesse. A seguir, veremos alguns exemplos do uso conjunto da inferência bayesiana e filogenias moleculares para investigar aspectos evolutivos além das relações filogenéticas dos organismos.

15.7.1. Relógio molecular bayesiano

O método bayesiano de estimativas de tempos de divergência e taxas de substituição foram primariamente desenvolvidos por Thorne e seus colaboradores (Thorne *et al.*, 1998, Thorne e Kishino, 2002). As vantagens do método incluem o uso de in-

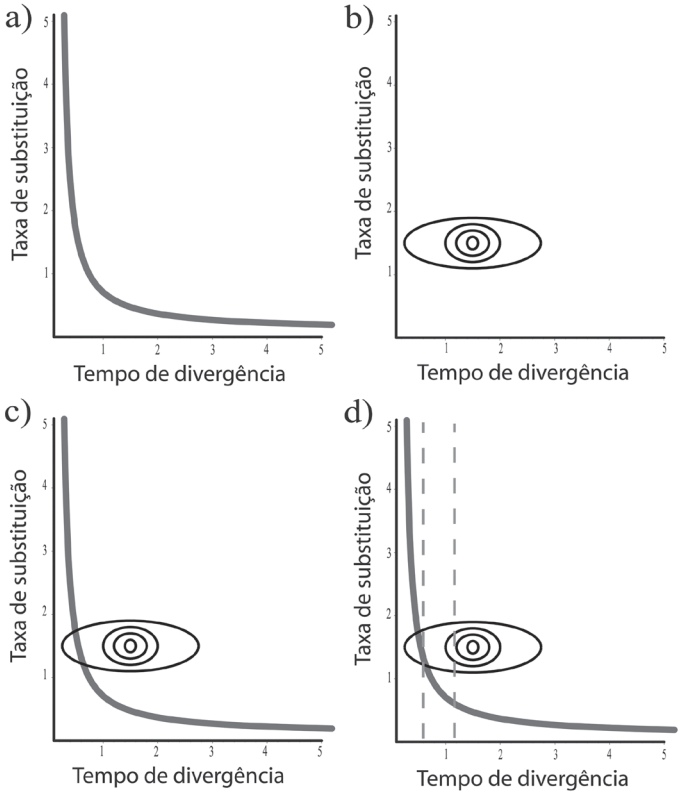


Figura 15.8. Incorporação de *prioris* de tempo de divergência e taxas de substituição na estimativa de distribuição bayesiana posterior. O tempo de divergência é dado em milhões de anos (Ma) e as taxas de substituição em 0,1 substituição/sítio/Ma. a. A linha contínua indica todas os pontos no espaço bidimensional onde a taxa de substituição equivale a 0,065 substituições por sítio. b. Distribuição *a priori* de taxas e tempos. Note que as elipses mais internas são mais restritas quanto à possível gama de valores assumidos *a priori* pelas taxas de substituição e pelos tempos de divergências. c. Combinação de *prioris* com o comprimento de ramos de 0,065 substituições por sítio. d. A linha vertical tracejada representa restrições temporais de divergência derivadas de fósseis, eventos geológicos ou datações moleculares obtidas de outras fontes que limitam os valores da distribuição posterior. Modificado de Thorne e Kishino (2005).

formação *a priori* sobre a divergência temporal entre sequências ou organismos ou suas taxas de evolução, a incorporação de incertezas associadas aos tempos de divergência e taxas de substituição, e o relaxamento do relógio molecular, permitindo que taxas de evolução variem não apenas entre diferentes linhagens mas também ao longo do tempo (Thorne *et al.,* 1998, Thorne e Kishino, 2002, Yang e Rannala, 2006). A Figura 15.8 ilustra o processo incorporado no método bayesiano. Os comprimentos de ramos são um produto das taxas de substituição e tempo de divergência, e ocupam uma linha unidimensional definida no espaço bidimensional entre taxas de substituição e tempo de divergência (Fig. 15.8a). Consideremos que duas sequências de DNA diferem entre si por 0,065 substituições por sítio. Se essas sequências se originaram em um ancestral comum há 6,5 milhões de anos (Ma), por exemplo, a taxa de substituição seria de 0,01 substituições/sítio/Ma. No entanto, não podemos discernir uma outra possibilidade, por exemplo de que essas sequências se separaram há 1 Ma, e, portanto, sua taxa de substituição equivale a 0,065 substituições/sítio/Ma ou ainda se a separação se deu há 10 Ma com uma taxa equivalente a 0,0065 substituições/sítio/Ma. Na ausência de informação temporal sobre a divergência dessas sequências, os comprimentos de ramos não podem ser decompostos em seus termos de tempo de divergência e taxa de substituição, mesmo com sequências infinitamente longas. Em termos práticos, o problema agrava-se ainda mais, uma vez que as taxas de substituição são heterogêneas entre espécies e ao longo do tempo, e a relação bidimensional entre taxas de substituição e tempos de divergência é menos clara (isto é, se a hipótese do relógio molecular não for válida, veja Capítulo 7).

No entanto, se tivermos *prioris* sobre tempos de divergência e/ou taxas de substituição (Figura 15.8b), eles podem ser incorporados com a informação contida nas sequências de DNA (Fig. 15.8c). Embora a combinação de *prioris* com informação retirada das próprias sequências de DNA ainda possam representar pontos não distintos no espaço unidimensional do universo bidimensional de taxas e tempos, algumas taxas de substituição e tempos de divergência possuem densidades posteriores maiores que outros (Fig. 15.8c) determinados pela distribuição *a priori* (Fig. 15.8b). Quando dados temporais como idades míminas ou máximas extraídas de dados fósseis, eventos geológicos ou mesmo datações moleculares prévias são incorporadas na análise, a distribuição posterior das taxas de substituição e dos tempos de divergência se restringe ainda mais a um número limitado de valores possíveis (Fig. 15.8d).

O método bayesiano de datação vem rapidamente substituindo o método de datação por verossimilhança. Embora ambos os métodos incorporem incerteza das idades das calibrações temporais (Thorne *et al.*, 1998, Thorne e Kishino, 2002, Yang e Yoder, 2003), no método de datação por verossimilhança, os comprimentos de ramos são tidos como parâmetros fixos, cuja incerteza não é considerada nas estimativas de taxas de substituição e tempos de divergência. Ainda, um dos maiores problemas para se obterem estimativas de tempo de divergências por máxima verossimilhança se encontra na decisão do pesquisador em dividir os ramos em grupos com taxas distintas. O método de relógios dispersos usando verossimilhança penalizada, visto no Capítulo 14, oferece um algoritmo no qual as taxas de cada ramo são inferidas dos próprios dados (Sanderson, 2002).

Pereira e Baker (2006) aplicaram o método bayesiano para elucidar questões controversas em ornitologia. Muito tem sido debatido sobre a origem das aves modernas. As hipóteses alternativas vão desde propostas em que os tempos de divergência e a diferenciação ecológica ocorreram após 65 Ma, com o extermínio dos dinossauros durante a transição entre os Períodos Cretáceo e Paleogeno, até hipóteses nas quais muitas linhagens modernas já eram ecologicamente diferenciadas no Cretáceo, antes da extinção em massa sofrida pelos dinossauros (revisão em Penny e Phillips, 2004). A fonte de controvérsias pode ser atribuída à fossilização pobre de aves em geral e à falta de conhecimento sobre a evolução das taxas de substituição no grupo. Usando sequências de genomas mitocondriais e impondo diversas restrições temporais quanto à diversificação de vertebrados, Pereira e Baker (2006) demonstraram que várias linhanges de aves modernas já se encontravam diferenciadas no Cretáceo e que as taxas de evolução são muito heterogêneas entre os diferentes grupos de aves, variando entre 0,09 e 1,2%/Ma nas espécies estudadas.

15.7.2. Seleção positiva

Adaptação gênica e genômica é o processo evolutivo responsável pela diferenciação morfológica, comportamental e fisiológica, e pela divergência entre espécies e aquisições de inovações evolutivas (Yang, 2006). Embora vários testes tenham sido desenvolvidos para detectar seleção positiva em nível molecular, Yang e colaboradores (Nielsen e Yang, 1998, Yang *et al.,* 2000. Yang *et al.,* 2005) desenvolveram uma estratégia mista entre máxima verossimilhança e análise bayesiana para determinar amino ácidos sofrendo seleção positiva, negativa ou neutra. Detalhes do método estão fora do escopo deste capítulo, mas, em resumo, o

método utiliza o teste de razão de verossimilhança para detectar a presença de sítios sob seleção positiva em um conjunto de sequências codificadoras de proteínas e usa uma estratégia bayesiana para detectar quais sítios nas sequências estão sob seleção positiva. Essa estratégia incorpora incertezas nas estimativas de verossimilhança, ao mesmo tempo que comprimentos de ramos são mantidos fixos durante as iterações do algoritmo, uma vez que, teoricamente, os comprimentos de ramos pouco afetariam a decisão sobre quais sítios estão sob seleção positiva (Yang, 2006).

Muitos estudos demostraram que o método de Yang é útil para detectar seleção positiva em proteínas envolvidas nos sistemas de defesa e imunológicos, proteínas virais e bacterianas que escapam da detecção pelos sistema imunológico, toxinas, proteínas envolvidas em digestão ou reprodução, proteínas duplicadas e proteínas com várias outras funções (revisão em Yang, 2006). Em geral, sítios sob seleção positiva são responsáveis por conferir uma função molecular específica. Por exemplo, proteínas do complexo major de histocompatibilidade (MHC) são responsáveis por identificar a invasão de patógenos em vertebrados e desencadear a resposta imune. Os genes MHC-1 e MHC-2 fazem parte de famílias gênicas e são os responsáveis direto pelo reconhecimento de proteínas exógenas e apresentação dos antígenos para as células responsáveis por sua destruição (isto é, macrófagos e células T). Diversos sítios das proteínas codificadas por MHC-1 e MHC-2 estão sob forte seleção positiva e são os responsáveis pela identificação de diferentes patógenos a que um organismo está exposto (Yang, 2006).

15.7.3. Reconstrução de estados ancestrais

O mapeamento filogenético é uma ferramenta importante para inferir estados ancestrais e entender a evolução de caracteres moleculares, morfológicos, comportamentais e ecológicos. Até recentemente, o método de parcimônia era o mais utilizado para tais fins. Estratégias estocásticas usando máxima verossimilhança e inferência bayesiana foram recentemente desenvolvidas (Lewis, 2001; Huelsenbeck *et al.*, 2003). A incerteza filogenética é considerada na inferência bayesiana por meio do mapeamento do caráter de interesse nas árvores contidas na distribuição posterior (por exemplo, as árvores amostradas no programa MrBayes ou BEAST). Dada uma topologia, comprimento de ramos e modelo de evolução, a história evolutiva de um caráter pode ser inferida por MCMC usando um algoritmo simples: primeiramente, calcula-se a probabilidade condicional de cada estado de caráter em cada nó interno, incluindo a raiz da árvore; a seguir, estados ancestrais para cada nó interno são simulados e amostrados da distribuição posterior; finalmente, uma simulação da história das mudanças de estados de caráter é realizada com base no passo anterior, os estados nos nós terminais são registrados (Bollback, 2006). Esse procedimento é implementado no software SIMMAP, de distribuição gratuita (Bollback, 2006). A análise no programa SIMMAP reporta o número esperado de mudanças, a direção da mudança em árvores enraizadas, o tipo e número esperado de mudanças em cada ramo, e a correlação entre os caracteres mapeados (Huelsenbeck *et al.*, 2003).

O método de mapeamento filogenético também foi aplicado mais recentemente para estabelecer a origem geográfica de pombos e aves relacionadas (Columbiformes), de maneira análoga ao mapeamento de caracteres morfológico (Pereira *et al.*, 2007). Inicialmente, os autores atribuíram áreas biogeográfica de importância de acordo com sua distribuição atual a todos os gêneros de columbiformes. A seguir, as áreas biogeográficas foram consideradas como caracteres observados nos nós terminais, e as áreas ancestrais em cada nó interno na filogenia foram inferidas usando o método bayesiano. Os autores mostraram que aves columbiformes modernas se originaram há cerca de 54 Ma, mais provavelmente na região Neotropical ou na Australasiática, que fizeram, até a metade do Cretáceo, parte de uma massa continental maior única chamada Gondwana. Ao longo do processo de separação entre partes da Gondwana e deriva dos continentes atuais, os columbiformes se dispersaram para outras regiões do mundo, com a colonização múltipla das regiões biogeográficas africana, australasiática e oriental (Pereira *et al.*, 2007).

Sob o ponto de vista molecular, o método bayesiano tem sido usado para inferir a sequência de resíduos de aminoácidos de proteínas ancestrais, reconstruí-las artificialmente em laboratório e testar as propriedades funcionais que alteram processos bioquímicos e produzem novos fenótipos (revisões em Thornton, 2004, Dean e Thornton, 2007). Yokoyama e seus colaboradores são os pioneiros nos estudos funcionais que levaram à elucidação da mudanças de amino ácidos responsáveis pela evolução da visão em cores em vertebrados (Shi *et al.* 2001, Shi e Yokoyama, 2003, Yokoyama *et al.*, 2006). Usando o método bayesiano de inferência de sequências ancestrais, eles foram capazes de reconstruir e expressar *in vitro* proteínas ancestrais e identificar duas mudanças de aminoácidos responsáveis pela adaptação visual de celacantos às profundidades marinhas em que eles vivem. Similarmente, eles demonstraram que a visão ultravioleta presente em algumas aves estava presente no ancestral de todas as aves modernas e foi adquirida pela substituição de quatro aminoácidos.

Referências Bibliográficas

Alfaro, M.E. e Holder, M.T. (2006). The posterior and the prior in Bayesian phylogenetics. **Ann. Rev. Ecol. Evol. Syt. 37**: 19-42.

Altekar, G., Dwarkadas, S., Huelsenbeck, J.P. e Ronquist, F. (2004). Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. **Bioinformatics 20**:407-415.

Bollback, J.P. (2006). SIMMAP: Stochastic character mapping of discrete traits on phylogenies. **BMC Bioinformatics 7**:88.

Dean, A.M. e Thornton, J.W. (2007). Mechanistic approaches to the study of evolution: the functional synthesis. **Nature Reviews Genetics 8**: 675-688.

Douady, C. J., Delsuc, F., Boucher, Y. Doolittle, W. F. e Douzery, E. J. P. (2003). Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. **Mol. Biol. Evol. 20**:248–254.

Drummond, A.J. e Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. **BMC Evol Biol 7**:214.

Foster, P. G. (2004). Modeling compositional heterogeneity. **Syst. Biol. 53**: 485-495.

Galtier, N. E Gouy, M. (1998). Inferring pattern and process: maximum likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. **Mol. Biol. Evol. 15**: 871-879.

Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. **Biometrika 57**:97-109.

Holder, M. e Lewis, P.O. (2003). Phylogeny estimation: traditional and Bayesian approaches. **Nat Rev Genet 4**:275-284.

Huelsenbeck, J.P. (2005a).Performance of phylogenetic methods in simulation. **Syst. Biol. 44**: 17-48.

Huelsenbeck, J.P. (2005b). The robustness of two phylogenetic methods: four taxon simulations reveal a slight superiority of maximum likelihood over neighbor joining. Mol. Biol. Evol. 12: 843-849.

Huelsenbeck, J.P., Nielsen, R. e Bollback, J.P. (2003). **Stochastic mapping of morphological characters. Syst. Biol. 52:** 131-158.

Huelsenbeck, J.P., Rannala, B. (2004). Frequentist propoerties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. **Syst. Biol. 53**: 904-913.

Huelsenbeck, J.P., Rannala, B. e Larget, B. (2000). A Bayesian framework for the analysis of cospeciation. **Evolution 54**: 352-364.

Huelsenbeck, J.P., Ronquist, F. (2005). Bayesian analysis of molecular evolution using MrBayes. **In** Nielsen, R. (Ed). **Statistical methods in molecular evolution**. Springer, pp. 183-232.

Huelsenbeck, J.P., Ronquist, F., Nielsen, R. e Bollback, J.P. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. **Science 294**: 2310-2314.

Lee, P.M. (1997). **Bayesian statistics: an introduction**. Arnold Publishers, Londres.

Lewis, P.O. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. **Syst. Biol. 50**: 913-925.

Mau, B., Newton, M.A. e Larget, B. (1999). Bayesian phylogenetic inference via Markov chain Monte Carlo methods. **Biometrics 55**:1-12.

Metropolis, N., Rosenbuth, A.W., Rosenbluth, M.N., Teller, A.H. e Teller, E. (1953). Equations of state calulations by fast computing machines. **J. Chem. Phys. 21**:1087-1091.

Nielsen, R. e Yang, Z. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. **Genetics 148**: 929-936.

Penny, D., e Phillips, M. J. (2004). The rise of birds and mammals: are microevolutionary processes sufficient for macroevolution? **Trends Ecol. Evol. 19**: 516-522.

Pereira, S. L. e Baker, A. J. (2006). A mitogenomic timescale for birds detects variable phylogenetic rates of molecular evolution and refutes the standard molecular clock. **Mol. Biol. Evol. 23**: 1731-1740.

Pereira, S. L., Johnson, K. P., Clayton, D. H. e Baker, A. J. (2007). Mitochondrial and nuclear DNA sequences support a Cretaceous origin of Columbiformes and a dispersal-driven radiation in the Paleogene. **Syst. Biol. 56**: 656-672.

Rannala, B. e Yang, Z. (1996). Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. **J Mol Evol 43**:304-311.

Ronquist, F. e Huelsenbeck, J.P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. **Bioinformatics 19**:1572-1574.

Sanderson, M. J. 2002 Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. **Mol Biol Evol 19**: 101-9.

Shi, Y. e Yokoyama, S. (2003). Molecular analysis of the evolutionary significance of ultraviolet vision in vertebrates. **Proc. Natl. Acad. Sci. USA 100**: 8308-8313.

Shi, Y., Radlwimmer, R.B. e Yokoyama, S. (2001). Moelcular genetics and the evolution of ultraviolet vision in vertebrates. **Proc. Natl. Acad. Sci. USA 98**: 11731-11736.

Simon, D. e Larget, B. (2000). **Bayesian analysis in molecular biology and evolution (BAMBE).** Department of Mathematics and Computer Science, Duquesne University.

Suzuki, Y., Glazko, G. V.e Nei, M. (2002). Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. **Proc. Natl. Acad. Sci. USA 99**:15138–16143.

Thorne, J.L. e Kishino, H. (2002). Divergence time and evolutionary rate estimation with multilocus data. **Syst. Biol. 51**: 689-702.

Thorne, J.L. e Kishino, H. (2005). Estimation of divergence times from molecular sequence data. **In** Nielsen, R. (Ed). **Statistical methods in molecular evolution**. Springer, pp. 234-256.

Thorne, J.L., Kishino, H. e Painter, I.S. (1998). Estimating the rate of evolution of the rate of molecular evolution. **Mol. Biol. Evol. 15**: 1647-1657.

Thornton, J.W. (2004). Resurrecting ancient genes: experimental analysis of extinct molecules. **Nature Reviews Genetics 5**: 366-375

Yang, Z. (2006). **Computational molecular evolution**. Oxford Series in Ecology and Evolution. Oxford University Press, Oxford, Reino Unido.

Yang, Z. e Rannala, B. (1997). Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. **Mol Biol Evol 14**:717-724.

Yang, Z. e Rannala, B. (2006). Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. **Mol. Biol. Evol. 23**: 212-226.

Yang, Z. E Yoder, A. D. (2003). Comparison of likelihood and Bayesian methods for estimating divergence times using multiple gene loci and calibration points, with application to a radiation of cute-looking mouse lemur species. **Syst. Biol. 52**: 705-716.

Yang, Z., Nielsen, R., Goldman, N. e Pedersen, A.-M. K. (2000). Codon substitution models for heterogeneous selection pressure at amino acid sites. **Genetics 155**: 431-449.

Yang, Z., Wong, W. S. W. e Nielsen, R. (2005). Bayes empirical Bayes inference of amino acid sites under positive selection. **Mol. Biol. Evol. 22**: 1107-1118.

Yokoyama, S. Starmer, W.T., Takahashi, Y. e Tada, T. (2006). Tertiary structure and spectral tuning of UV and violet pigments in vertebrates. **Gene 365**: 95-103.

Capítulo 16

Como escolher genes para problemas filogenéticos específicos

Claudia A. M. Russo (claudia@biologia.ufrj.br)
Departamento de Genética
Instituto de Biologia
Universidade Federal do Rio de Janeiro

Carolina Moreira Voloch (carolina@biologia.ufrj.br)
Departamento de Genética
Instituto de Biologia
Universidade Federal do Rio de Janeiro

Carlos G. Schrago (guerra@biologia.ufrj.br)
Departamento de Genética
Instituto de Biologia
Universidade Federal do Rio de Janeiro

"As coisas são semelhantes: isto faz a Ciência possível; as coisas são diferentes: isto faz a Ciência necessária." (Levins e Lewontin, 1985)

16.1. Introdução

Os inúmeros avanços dos métodos moleculares, nas últimas quatro décadas, impulsionaram a implantação de laboratórios especializados em técnicas moleculares em todo mundo. Bancos de dados genéticos, como o Genbank (www.ncbi.nlm.nih.gov) vêm acumulando sequências exponencialmente (Figura 16.1). De 1982 até o presente, o número de bases no GenBank, que hoje soma mais de 80 bilhões, está crescendo a uma taxa impressionante, dobrando de tamanho a cada 18 meses aproximadamente (Benson *et al*., 2008).

Como consequência da alta disponibilidade, dados moleculares tornaram-se uma fonte de informação biológica fundamental em todas as áreas, incluindo a medicina clínica e forense, bioconservação e sistemática (Burks, 1997; Sullivan e Joyce, 2005). Nesta última área, esse tipo de dados é usado, por exemplo, para resolver problemas de filogenia de grandes grupos (Cavalier-Smith, 2003; Lartillot *et al*., 2007; Woese *et al*., 1990), relações genealógicas das populações humanas, a filogenia dos grandes primatas e das ordens de mamíferos (Cavalli-Sforza *et al*., 2003; Kong *et al*., 2006; Murphy *et al*. 2001a, b; Satta *et al*., 2000). Uma questão especial envolve os estudos sobre a posição filogenética de grupos fósseis, estabelecendo pontes entre disciplinas tão tradicionalmente distintas, quanto a paleontologia e a biologia molecular (Debruyne *et al*., 2003; Green *et al*., 2006; Pääbo *et al*., 2004).

Para a reconstrução de uma árvore filogenética com base em caracteres moleculares, é necessário um algoritmo de reconstrução filogenética aplicado a um determinado conjunto de sequências. Os Capítulos 12 a 15 abordaram essa primeira parte da questão. O presente capítulo abordará o segundo problema, ou seja, como escolher um gene apropriado para um determinado problema filogenético. O conjunto de dados moleculares a ser usado em uma reconstrução filogenética molecular pode ser um conjunto de sequências de nucleotídeos ou de aminoácidos. Naturalmente, existem outros tipos de moléculas nos organismos, mas a filogenia molecular diz respeito apenas à evolução de nucleotídeos ou de aminoácidos. Isso porque apenas nucleotídeos e aminoácidos carreiam a informação entre gerações, revelando assim a história evolutiva das linhagens quando analisados.

A decisão por nucleotídeos ou aminoácidos, no entanto, deve levar em consideração a taxa de evolução do gene e o tempo de divergência das espécies a serem estudadas (Russo *et al.*, 1996). As sequências de aminoácidos têm uma taxa de evolução mais lenta que as de nucleotídeos (Nei e Kumar, 2000) e, portanto, podem ser usadas quando as sequências de nucleotídeos apresentam alguma saturação, como será discutido adiante. Com as opções virtualmente infinitas na escolha do conjunto de dados disponíveis (Benson *et al*., 2008), o pesquisador depara-se com o problema de qual conjunto será mais apropriado para abordar seu problema filogenético específico. A escolha dos dados é crucial, já que árvores construídas para um mesmo grupo de organismos com base

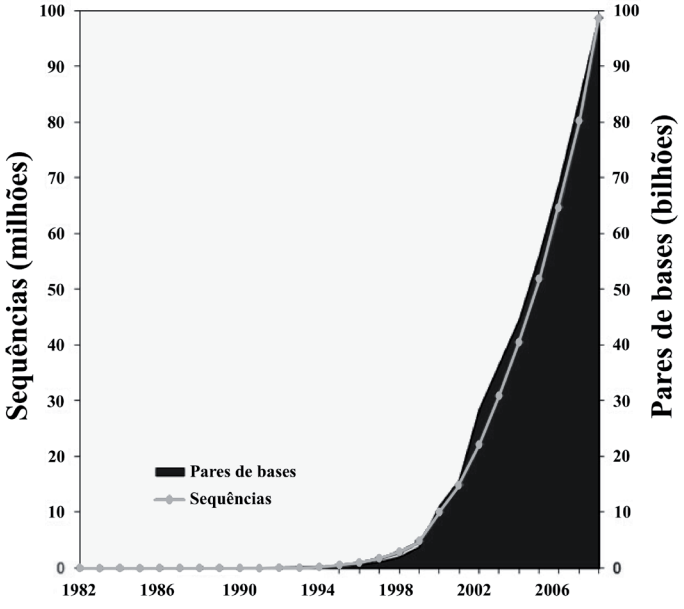


Figura 16.1. Gráfico mostrando o crescimento exponencial do Genbank (fonte: NCBI/HIH).

em conjuntos de dados diferentes podem resultar em filogenias diferentes (Ericson *et al.*, 2006; Kuma e Miyata, 1994; Morgan-Richards *et al*., 2008; Russo *et al.*, 1996; Satta *et al.*, 2000). Esse problema também acontece quando temos que escolher entre os diferentes métodos de reconstrução filogenética (Russo *et al.*, 1996; Takezaki e Gojobori, 1999; Kelsey *et al.*, 1999).

Simulações de computador (Nei, 1991; Huelsenbeck, 1995; Takezaki, 1998; Wiens e Servedio, 1998) ajudam a testar a influência dos principais parâmetros (e.g., frequência de bases, taxa de transição e transversão, parâmetro gama) e dos métodos de reconstrução filogenética na recuperação do sinal filogenético dos caracteres moleculares. Infelizmente, simulações não são úteis para testar o desempenho de genes específicos na reconstrução filogenética, mas fornecem apenas alguns valores determinados de seus parâmetros e, por vezes, de pequeno significado biológico (Nei *et al.*, 1995). Portanto, a extrapolação desses resultados para os genes a serem usados em análises filogenéticas não é simples. Uma alternativa às simulações em computador é o uso de grupos taxonômicos cujas relações filogenéticas estejam estabelecidas sem ambiguidades por métodos não moleculares. Dessa forma, o teste empírico da eficiência de genes (Russo *et al.*, 1996) ou de métodos na reconstrução filogenética (Cunningham, 1997) torna-se factível.

Russo e colaboradores (1996) usaram uma filogenia conhecida de vertebrados (duas baleias, rato, camundongo, marsupial, galinha, sapo e três peixes ósseos) para testar a eficiência de genes mitocondriais e dos métodos na reconstrução dessa topologia. Naquele trabalho, os autores concluíram que a escolha cuidadosa do gene é muito mais importante que a escolha do método de reconstrução filogenética. De uma maneira geral, a diferença de eficiência entre os genes na reconstrução da filogenia verdadeira foi muito maior do que a diferença entre métodos distintos de reconstrução filogenética. O gene *ND5*, por exemplo, foi o que obteve o melhor desempenho, reconstruindo a mesma árvore com todos os métodos de reconstrução filogenética, enquanto que o gene *ND4L* obteve o pior desempenho, reconstruindo invariavelmente uma árvore que agrupa anfíbios com mamíferos e excluindo as aves, com valores significativos de *bootstrap* e teste do ramo interno (ver também Russo, 1997). Por outro lado, a diferença de eficiência entre os métodos de reconstrução filogenética–teor de tantos debates acalorados em congressos e simpósios sobre evolução molecular–foi rigorosamente menor, indicando que, se a escolha do gene for cuidadosa, qualquer método de reconstrução filogenética tem boas chances de recuperar a árvore verdadeira.

Esta revisão, portanto, irá abordar os fatores que devem ser levados em consideração quando da escolha do segmento de nucleotídeos ou aminoácidos para resolver um determinado problema filogenético.

16.2. Homologia

Na reconstrução de qualquer tipo de árvore filogenética, a preocupação primordial do pesquisador deve ser a homologia (Hennig, 1966; Phillips *et al.*, 2000), isto é, devemos sempre comparar caracteres homólogos nas diferentes espécies. A homologia é um conceito fundamental em sistemática, como se percebe com a formalização do método comparativo (Pinna, 1991). O termo homologia, entretanto, foi cunhado por Owen, um anatomista inglês que não aceitava a idéia da transmutação das espécies, no meio do século XIX, antes do pensamento sistemático evolutivo. Owen considerava homólogas as estruturas que, mesmo quando diferentes em sua morfologia, faziam parte de uma mesma entidade no desenho geral dos corpos, especialmente quando se referia a arquétipos (Lewin, 1997).

Atualmente, entende-se por homologia uma propriedade relativa a entidades que tenham uma origem evolutiva comum, ou seja, que surgiram em um ancestral comum. Em síntese, duas estruturas (morfológicas, comportamentais ou moleculares, por exemplo) são homólogas se suas partes são semelhantes devido a uma origem filogenética comum (Patterson, 1988; Titus e Frost, 1996; Lewin, 1997; Graur e Li, 2000). Homologia é, fundamentalmente, um termo qualitativo e não quantitativo. Ainda assim, vários autores ainda utilizam erroneamente o conceito de homologia como sinônimo de similaridade. Por exemplo, a frase "duas sequências de nucleotídeos possuem 60% de homologia" não faz sentido. Tais sequências possuem 60% de similaridade. Naturalmente, a inferência da homologia é feita a partir da similaridade, mas os dois conceitos são diferentes em essência, pois a idéia evolutiva de homologia incorpora uma dimensão temporal, ausente na idéia de simples semelhança. Grau de similaridade pode ser mensurado, mas homologia corresponde a uma hipótese (ver também Patterson, 1988). Apesar de os pesquisadores reconhecerem a importância da homologia no estabelecimento de relações filogenéticas, na prática o problema agrava-se, já que a homologia não é é diretamente observável, permanecendo como uma hipótese.

Quaisquer duas sequências geradas ao acaso (ou seja, não homólogas) serão idênticas em 25% dos resíduos de nucleotídeos ou 5% dos de aminoácidos. Isso significa que qualquer porcentagem maior que esses números corresponde a um primeiro indicativo de homologia entre as sequências comparadas. Por outro lado, de acordo com o tamanho e a composição de bases, é possível sequências com uma alta similaridade ou mesmo idênticas não serem estritamente homólogas. Se a sequência for pequena (até 20 pares de bases, por exemplo), a probabilidade é relativamente alta de termos, por acaso, sequências idênticas por convergência (não homólogas) espalhadas pelo genoma. Duas sequências idênticas de 400 pares de bases são homólogas.

Em um artigo importante nessa discussão, Pinna (1991) faz uma proposta interessante e cunha os termos homologia primária e homologia secundária. Nessa proposta, Pinna sugere que a homologia primária seria a aceitação de que, em princípio, as estruturas podem ser comparadas, uma homologia conjectural baseada em similaridade ou de uma análise em um nível filogenético mais abrangente. Mas estruturas com homologia primária não necessariamente apresentam modificações homólogas—como o fato de que os membros anteriores de morcegos são homólogos aos membros anteriores de aves, mas o que se chama de asas em cada um desses grupos não é homólogo. Particularmente, no caso de evolução molecular, as hipóteses de homologia primária seriam explicitadas pelo alinhamento. Por outro lado, a homologia secundária é a legitimação da hipótese de homologia primária quando esta passou em testes de congruência entre hipóteses de homologia primária. Mais especificamente, quando existe uma concordância entre a proposição filogenética e a partição da posição do alinhamento, cada uma das hipóteses de homologia primária será transformada em uma (congruência total) ou mais hipóteses (congruência parcial) de homologia secundária.

Um exemplo irá ilustrar bem esses dois conceitos. Suponhamos que, após o alinhamento múltiplo de quatro sequências, uma determinada posição mostra a seguinte partição para as sequências de DNA das espécies *a*, *b*, *c, d* e *e* que, nesta posição do códon, possuem A, A, T, T, T, respectivamente. Dessa forma, o alinhamento reflete duas hipóteses de homologia primária que são 1) os dois A e 2) os três T desta posição. Entretanto, a homologia primária 1 só será legitimada em homologia secundária (única), se na árvore filogenética as espécies *a* e *b* compartilharem um ancestral comum exclusivo.

Em termos moleculares, a homologia dos genes comparados em uma filogenia é, obviamente, uma questão essencial, mas não é

suficiente. Existem vários processos que podem gerar genes homólogos: a duplicação gênica e a especiação (Li, 1997; veja também os Capítulos 7 e 10). Ou seja, dois genes são homólogos se eles descendem de um mesmo ancestral. No caso da origem de dois ou mais genes por eventos de duplicação, eles são chamados parálogos. Quando os genes passaram a ter histórias evolutivas independentes a partir de um evento de especiação, eles são chamados ortólogos. Existem ainda outros dois tipos de homologia molecular—a xenologia, quando existe transferência horizontal, e a plerologia, quando existe conversão gênica (Koonin, 2005). Genes xenólogos e plerólogos são pouco informativos ou mesmo desinformativos em análises filogenéticas, mascarando as relações ancestral-descente que buscamos em filogenias. No presente capítulo, vamos nos ater, portanto, aos genes parálogos e ortólogos. Essa não é uma discussão irrelevante. Se, trabalhando com genes ortólogos há dificuldades computacionais grandes de não ser recuperada a filogenia real de um grupo, a comparação de genes parálogos pode gerar um ruído na análise que certamente resulta em reconstruções incorretas.

A Figura 16.2 mostra a evolução de um dado gene ancestral α, que, a partir de um evento de duplicação gênica, passa a apresentar duas cópias, α e β. Nesse caso, as cópias α e β são chamadas *cópias parálogas*. Supondo que, ao longo do tempo, essa população passe por um evento de especiação, as duas cópias α e β irão evoluir independentemente nas duas espécies, acumulando substituições únicas e, por conseguinte, diferenciando-se. Nesse outro caso, as cópias α1 e α2 entre si (bem como β1 e β2) serão *cópias ortólogas*. Portanto, dependendo do problema em questão, deve-se usar genes parálogos ou ortólogos. Mais especificamente, para estudar eventos de duplicação gênica em famílias ou superfamílias de genes, cópias parálogas de uma única espécie devem ser estudadas. Alternativamente, genes ortólogos devem ser escolhidos para a reconstrução filogenética de grupos taxonômicos.

Assim, é necessário sempre trabalhar com as cópias ortólogas do gene em questão para construir a filogenia de espécies. Na prática, entretanto, a distinção entre as cópias α e β no cromossomo pode não ser tão fácil, já que não existem marcas ou etiquetas nos cromossomos. Isso, na realidade, pode ser um problema grave, principalmente quando trabalhamos com genes, famílias gênicas, genes de múltiplas cópias ou genes cujo padrão de duplicação ainda é desconhecido—algo comum em todos os genomas (Li *et al.*, 2001; Nei, 1969; Piontkivska e Nei, 2003). Nesse sentido, se usarmos a comparação da cópia α de uma espécie com a cópia β de outra espécie estaremos adicionalmente superestimando o tempo de divergência entre as duas espécies (uma vez que o evento de duplicação gênica é anterior ao de especiação).

Na maioria dos drosofilídeos, por exemplo, existem algumas cópias homólogas do gene que codifica a enzima desidrogenase alcoólica (*Adh)*. A Figura 16.3 é um diagrama de um possível caminho de duplicação. A alta divergência entre *Adh* e *Adhr* sugere que o evento de duplicação que produziu as cópias é antigo e anterior à divergência de todos os drosofilídeos conhecidos. No entanto, note que *Adhr* não é encontrado em todas as espécies ou porque elas perderam ou porque o gene ainda não foi encontrado (Russo *et al.*, 1995). Além disso, os eventos de duplicação antigos podem ser úteis para reconstrução filogenética quando empregados como grupos externos recíprocos. Um caso clássico dessa prática foi a utilização de cópias parálogas para enraizar a árvore da vida e determinar as relações filogenéticas entre os três domínios, Bacteria, Archaea e Eukarya (Woese *et al.*, 1990; Osawa e Honjo, 1991).

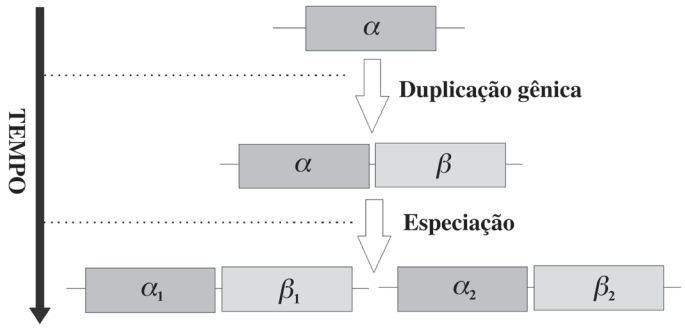


Figura 16.2. Processos de origem de genes homólogos: por especiação (genes ortólogos) e por duplicação gênica (genes parálogos).

Exemplos de genes de cópias múltiplas frequentemente usados em análises filogenéticas são os genes codificantes para o RNA ribossômico (veja os Capítulos 8 e 10). Existem, de modo geral, centenas de cópias desses genes nos genomas de vertebrados, o que torna provável que os eventos de duplicação gênica que geraram essas cópias já tivessem ocorrido no ancestral desse grupo. Entretanto, um fenômeno interessante foi observado em algumas espécies: cópias de genes ribossômicos num mesmo indivíduo são virtualmente idênticas, enquanto que estas diferem mesmo entre espécies relativamente próximas (Zimmer *et al.,* 1989). Esse fenômeno é conhecido como evolução em concerto, e as principais bases moleculares usadas para explicá-lo são a conversão gênica e *crossing over* desigual (Nei *et al.*, 1997; Li, 1997; Li, 2000; Wittzell *et al.*, 1999; Capítulo 10 deste livro). Por essa razão, a questão paralogia/ortologia não deve ser muito problemática pelo menos em genes ribossômicos. Entretanto, quanto menores forem os intervalos de tempo entre a divergência das linhagens que queremos estudar, mais perfeita deve ser a evolução em concerto, ou seja, mais próxima de 100% deve ser sua eficiência para não atrapalhar as filogenias feitas com esses genes.

Em alguns casos, como já mencionado, o padrão de duplicação é conhecido, mas é necessário ser extremamente cuidadoso ao inferir filogenias baseadas em genes pouco estudados. Uma solução simples para esse problema, pelo menos trabalhando com metazoários, é trabalharmos com DNA mitocondrial, pois o conteúdo e o número de genes do genoma mitocondrial é razoavelmente constante entre os diversos grupos de animais (Gillham, 1994), garantindo que estaremos trabalhando sempre com genes ortólogos (Capítulo 10). Um problema é a existência de cópias parálogas de genes mitocondriais encontrados no genoma nuclear. De fato, quando usamos um único gene mitocondrial para fazer uma filogenia, uma probabilidade não negligenciável é a de estarmos incluindo na análise cópias nucleares desses genes. É claro que, se as cópias forem recém-duplicadas, a diferença entre elas não será grande e as distorções na reconstrução da topologia devem ser pequenas ou nulas. Por outro lado, se essas cópias nucleares forem antigas, a possibilidade de terem se tornado pseudogenes aumenta, permitindo reconhecer com facilidade sua condição de cópia nuclear, o que também permite que não atrapalhe a análise filogenética.

Por outro lado, genomas mitocondriais de plantas não apresentam a estabilidade encontrada nos de metazoários, além de o conteúdo gênico e o tamanho do genoma variar muito entre as espécies de vegetais (Palmer, 1985). Os genomas de plastídeos também apresentam grandes disparidades em conteúdo e tamanho. Portanto, é necessário estar atentos ao usar genes de plastídeos em estudos filogenéticos devido ao problema de paralogia/ortologia.

16.3. Taxa e Modo de Evolução

16.3.1. O alinhamento

O objetivo do alinhamento é fazer com que, entre sequências consideradas homólogas, a posição (sítio) de cada base (ou aminoácido) comparada das várias espécies amostradas também seja homóloga. Por causa do problema de perdas ou ganhos de

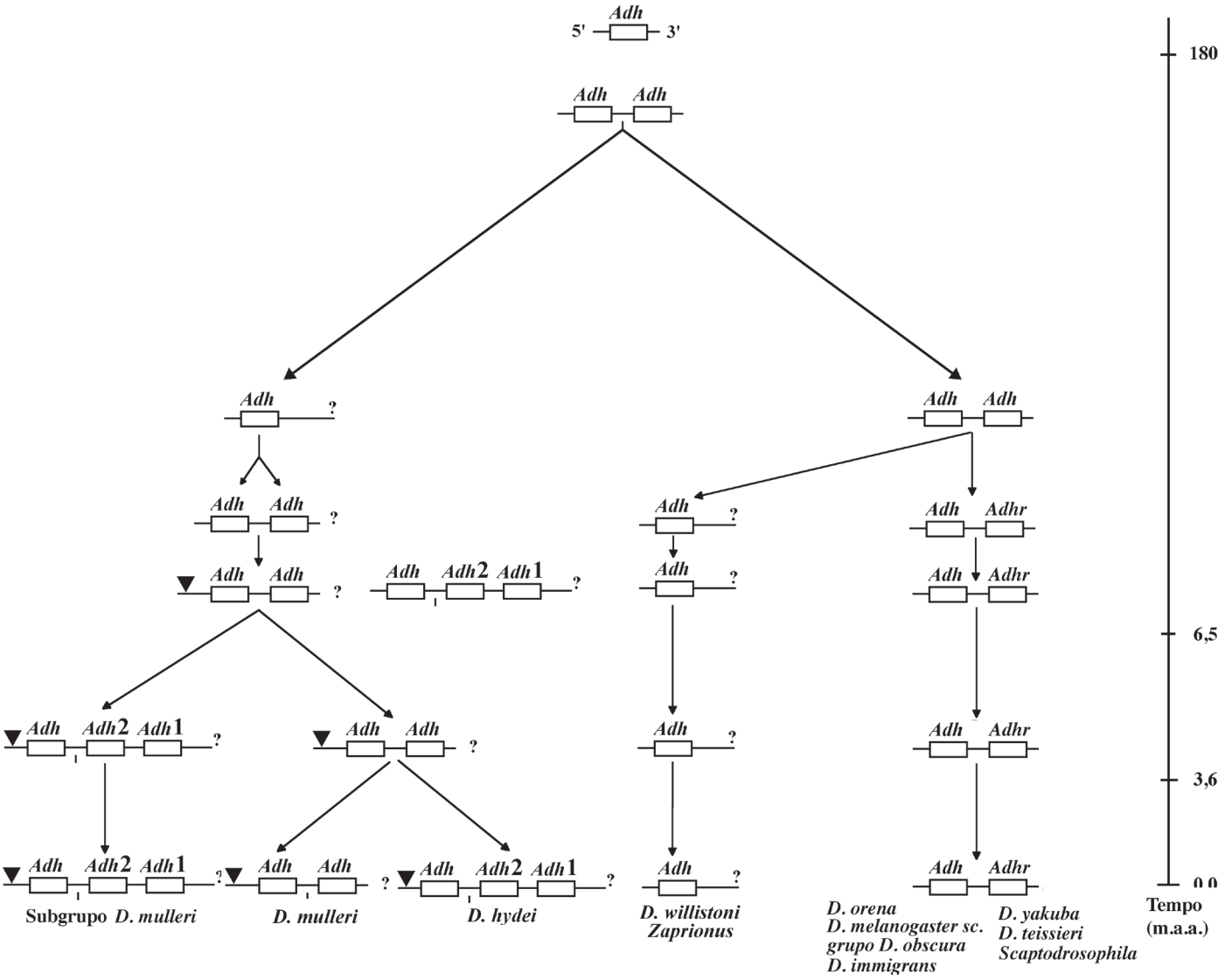


Figura 16.3. Evolução da família gênica *Adh* em drosofilídeos.

trechos nos vários genes (indels, ver Capítulo 7) é necessário que se "insiram" intervalos (com posições ocupadas por um hífen) nas sequências das espécies que perderam regiões ou daquelas que não os ganharam (Phillips *et al.*, 2000). Existem, hoje em dia, diversos programas de computador para alinhar sequências de aminoácidos e nucleotídeos (ClustalW é o mais popular.). Tais programas são eficientes (para uma comparação de desempenho entre os programas, ver Hickson *et al.*, 2000) e recomendáveis, mas alguns autores sugerem que, se um determinado conjunto de dados requer um computador para alinhar as sequências, essas sequências não deveriam estar sendo usadas! Obviamente esse comentário é exagerado, mas ilustra a importância crucial do alinhamento: se o alinhamento não é confiável, a homologia proposta para as relações entre as bases individualmente também não o será e, portanto, dificilmente a filogenia produzida será correta (Mitchinson, 1999; Phillips *et al.*, 2000). O problema torna-se mais complicado se considerarmos o número gigantesco de alinhamentos possíveis em uma determinada sequência (Morrison e Ellis, 1997; Slowinski, 1998).

Na construção de filogenias, estamos interessados no alinhamento de um conjunto de sequências ao invés de apenas um par e, portanto, devemos usar algoritmos que conduzem alinhamentos múltiplos. No programa ClustalW (Thompson *et al.*, 1994)—o alinhamento pode ser feito localmente ou em sítios www (ver Apêndice)—e na maior parte dos programas, o alinhamento múltiplo é feito em duas etapas. Na primeira, todas as sequências são comparadas par a par e uma medida da similaridade máxima entre cada duas sequências é calculada. Para calcular essa medida de similaridade máxima, um gráfico (ou uma matriz de pontos) é construído onde seus dois eixos representam as duas sequências comparadas. Para cada identidade entre as bases (independente da posição delas), um ponto é colocado no gráfico. Por exemplo, no caso de duas sequências idênticas, uma fileira de pontos será encontrada em toda a diagonal do gráfico, independentemente de sua composição de nucleotídeos. Com base nessa matriz de pontos, o programa maximiza a similaridade entre as sequências, usando para isso o que chamamos de penalidades. Dois tipos básicos de penalidades são usados nesse caso: a penalidade de intervalo, que é o número de bases idênticas entre as duas sequências que devemos ganhar para inserir um intervalo, e a penalidade de substituições. O valor da similaridade final entre cada duas sequências é calculado com o total de bases idênticas menos o número de substituições, menos o número de intervalos multiplicado pela penalidade de intervalos.

A partir dos dados de similaridade par a par, um dendrograma é construído e o alinhamento final é feito a partir dos nós mais externos. Essa etapa é denominada de alinhamento progressivo (Feng e Doolitle, 1987). Ou seja, o programa começa alinhando primeiro as sequências mais semelhantes, em seguida as que se conectam a essas e, assim por diante, até que todas as sequências estejam alinhadas. Essa segunda fase também é feita com base em um índice de similaridade total entre todas as sequências em processo de alinhamento.

O resultado final do alinhamento é, na realidade, um excelente indicador de quão adequado é o gene escolhido para o problema filogenético específico. Por exemplo, na Figura 16.4a,

A

```
seq1  A-CCTCACCGACCG-----TATTCTCTAC----AAACCACAAAGATATTG
seq2  ATAATCAC-G-C-GC-GACTATT-TCT-CCAGACAGACAGACCACAA--A
seq3  ATGATCAC--ACCGCTG-CTAT-CT-TACGGACCACAAAGATATTGGAAC
seq4  AT-TTCACCGACTGCTGACT-TT-TC-A---ACCAC----------AAAG
seq5  A--TTCACTTTCCGCTGAC-AT-CT--ACAAACCACA--GATATTGGAAC
```

B

```
seq1  ATGTTCACCGACCGCTGACTATTCTCTACAAACCACAAAGATATTGGAAC
seq2  ATGTTCACCGACCGCTGACTATTCTCTACAGACCACAAAGATATTGGAAC
seq3  ATGTTCACCGACCGCTGACTATTCTCTACAAACCACAAAGATATTGGAAC
seq4  ATGTTCACCGACTGCTGACTATTCTCTACAAACCACAAAGATATTGGAAC
seq5  ATGTTCACCGACCGCTGACTATTCTCTACAAACCACAAAGATATTGGAAC
```

Figura 16.4. Alinhamento de aminoácidos correspondentes ao gene mitocondrial *ND6* para um grupo de vertebrados (Bp = *Balaenoptera physalus* e Bm = *B. musculus*). A. Sequências muito variáveis geram alinhamentos inapropriados para reconstrução filogenética. B. Sequências muito conservadas são pouco informativas para reconstruções filogenéticas.

o alinhamento do gene mitocondrial *ND6* de espécies animais está ruim, com muitos indels (eventos de inserção/deleção), indicando que aquele gene está evoluindo muito rapidamente para a reconstrução de uma filogenia envolvendo essas espécies de vertebrados, que divergiram há 400 milhões de anos (Russo *et al.*, 1996). A lógica por trás dessa afirmativa é que, já que existem muitos eventos de indels nesse alinhamento, é provável que tenham ocorrido muito mais eventos de substituições do que aqueles que podemos observar. Ou seja, duas bases iguais—por exemplo, uma adenina—em uma mesma posição podem estar ali por convergência e não por origem comum, acarretando o questionamento da própria homologia das posições comparadas.

A solução, nesse caso, é a escolha de um gene mais conservado para inferir a filogenia nesse nível da evolução do grupo. Para esse exemplo, o gene mitocondrial *ND5* mostrou-se menos divergente e, por isso, mais eficiente para a reconstrução da filogenia entre essas espécies (Russo *et al.*, 1996). As sequências do gene *ND6* são tão divergentes entre as espécies que é realmente necessário um computador para alinhá-las; ainda assim, o alinhamento final será questionável e, igualmente, os resultados filogenéticos baseados nele. Na realidade, o pesquisador ainda pode tentar um alinhamento das sequências de proteínas (mais conservadas que as de nucleotídeos), mas como nesse caso, além das substituições, o problema são os inúmeros eventos de inserção e deleção, isso não faria muita diferença em termos de confiabilidade.

Hoje em dia, vários programas estão disponíveis que, ao alinhar sequências codificadoras, traduzem a sequência de nucleotídeos em aminoácidos, alinham tais sequências, e as transformam novamente em sequências de nucletídeos. O interessante desses algoritmos é que preservam a regra de inserções e deleções só acontecem de três em três nucleotídeos. Ou seja, se um alinhamento de uma região codificadora for encontrado, traduza para aminoácidos e verifique se o alinhamento está *in frame*, isto é, não existem códons de parada dentro da região codificadora do gene. Se você encontrar códons de parada numa região que deveria ser codificadora, cheque seu alinhamento e verifique se existem inserções ou deleções de um ou dois nucleotídeos. Neste caso, tais deleções ou inserções mudariam a *frame* de leitura dos aminoácidos e seriam inviáveis. Portanto, nesses casos, tais intervalos devem ser eliminados, rodando o alinhamento mais uma vez usando um programa que considera a *frame* de leitura (DAMBE, MEGA etc.)

Com isso em mente, é sempre interessante empregar um programa de computador para alinhar sequências de tamanhos diferentes. Slowinski (1998) mostrou que existem 1×10^{18} possibilidades de alinhamento englobando apenas cinco sequências de cinco nucleotídeos cada uma! Desse modo, o uso de um programa de computador é necessário para testar o maior número de possibilidades e escolher aquela mais parcimoniosa, ou seja, aquela que requer o menor número de substituições entre as sequências. No entanto, o alinhamento deve não somente ser revisto, retirando as regiões de alinhamento duvidoso ou pouco informativas, mas também colocado em questão a utilidade do gene. Obviamente, a utilidade de um gene é determinada diretamente pelo problema filogenético em questão. Por exemplo, mesmo o gene *ND6* pode ser adequado para a inferência filogenética de grupos cujas espécies começaram a divergir entre si há menos de 100 milhões de anos.

Um bom indicador da eficiência do alinhamento é o número de sítios conservados entre as sequências. Por outro lado, a Figura 16.4b mostra um alinhamento perfeito, tão perfeito que é pouco informativo para a reconstrução filogenética, uma vez que, nesse caso, não há informação para resolver muitos pontos da filogenia. Portanto, uma questão que surge neste momento é a das taxas de substituição evolutiva do gene em questão.

16.3.2. As taxas de substituição

Uma das principais vantagens de trabalhar com sequências moleculares é que podemos escolher genes que apresentam variabilidade (isto é, grau de divergência) compatível com o problema filogenético em questão (Russo *et al.*, 1996). A primeira regra básica para essa escolha é o alinhamento, como mencionado anteriormente. A segunda regra é que a proporção de diferenças máxima entre os pares de sequências alinhadas deve ser de 5% a 40% entre os pares de uma sequência de nucleotídeos. Chega-se a esse número já que a simples proporção de posições diferentes é, na realidade, uma subestimativa do número real de substituições que ocorreram entre as sequências (Capítulo 14). Como o número possível de bases é finito e pequeno (há apenas quatro possíveis), quanto menor essa proporção de diferenças, menor será a subestimativa. O limite mínimo é necessário para haver variação suficiente para resolver a filogenia. Por outro lado, 40% seria o limite de variabilidade para boa confiança no alinhamento (e, portanto, nos próprios resultados da análise filogenética) (ver Russo *et al.*, 1996).

Ou seja, idealmente deve-se trabalhar com sequências que não precisam de correção para substituições múltiplas, reversas ou paralelas, nas quais o número de substituições visíveis é praticamente idêntico ao número real de substituições (Grishin, 1999). Por outro lado, no caso de genes cuja taxa excede o limite de 25%, a solução é a utilização de modelos de evolução que corrijam substituições múltiplas, reversas e paralelas. Entretanto, alguns métodos de reconstrução filogenética (especialmente os baseados em matrizes de distância) são muito sensíveis a variância das estimativas, de maneira que modelos mais simples tendem a gerar resultados melhores que modelos multiparamétricos, particularmente quando o pesquisador está interessado somente na reconstrução da topologia (Nei e Kumar, 2000; Hoyle e Higgs, 2003).

Embora a proporção de diferenças seja um bom indicador da variabilidade das sequências, é importante ter em mente que esse valor assume que a probabilidade de substituição é constante ao longo dos sítios, o que nem sempre acontece, havendo regiões mais conservadas que outras. Sequências codificadoras, por exemplo, frequentemente apresentam grande variação, pois existe limitação funcional diferencial nas diversas regiões da proteína (Li, 2000). Dessa forma, ao comparar genes homólogos, alguns sítios tendem a acumular mudanças mais rapidamente e, em alguns casos, mesmo quando a proporção de diferenças observada é pequena, existe o risco de subestimarmos a quantidade de substituições se elas ocorreram apenas num grupo

restrito de sítios (Yang, 1996). Por exemplo, um valor de proporção de diferenças de 10% num gene de 100 sítios indica que, após a separação do ancestral, o par de sequências sofreu dez substituições. Entretanto, se essas dez substituições estiverem concentradas em apenas dois sítios, observaremos uma diferença máxima de 2% no par. Felizmente, existem modelos que incorporam a heterogeneidade de taxas ao longo dos sítios através da distribuição gama e corrigem tais discrepâncias (Uzzell e Corbin, 1971; Yang, 1993)

De certa maneira, essas conclusões parecem ser pouco informativas e razoavelmente frustrantes, já que tanto o alinhamento como a taxa de substituição são características que podem ser percebidas somente após o sequenciamento do gene. Assim, qual seria sua utilidade? A recomendação, nesse caso, é avaliar o conhecimento disponível sobre o gene em questão antes do início do sequenciamento. A pesquisa pode ser feita em bibliografia ou diretamente dos bancos de sequências, como o GenBank, fazendo uma avaliação inicial da variabilidade do gene para o seu grupo de trabalho ou para grupos próximos. Caso a variabilidade seja adequada para o nível da filogenia que precisamos estudar, segue-se em frente com o gene, sequenciando apenas as espécies para as quais não há informação, para resolver o problema filogenético em questão. Caso o gene seja muito variável ou muito conservado, outro gene mais adequado deve ser procurado.

Um ponto importante é que, exceto no nível de espécie—definidas como entidades biológicas reprodutivamente coesas—, há pouca informação sobre o tempo de divergência entre as sequências em outros níveis taxonômicos (Avise e Johns, 1999). Ou seja, dados sobre a variabilidade de determinado gene entre ordens de mamíferos (tempo de divergência médio de 80 milhões de anos) podem ser pouco informativos para o estudo da relação entre ordens de cnidários ou esponjas, cujo tempo de divergência provavelmente supera 700 milhões de anos. Além disso, fatores como tempo de geração e sistema de reparo podem influenciar quando da extrapolação de resultados de um grupo taxonômico para outro, embora de uma maneira mais branda (Bromham *et al.*, 1996). Lembre-se de que o objetivo de uma análise filogenética é esclarecer as relações filogenéticas entre diferentes grupos taxonômicos—e não de estudar as relações entre famílias de genes. Caso o gene de escolha seja inadequado, o pesquisador dificilmente alcançará seu objetivo.

16.3.3. Função do gene

Uma das aplicações mais interessantes da filogenia molecular é o estudo da evolução (origem única ou múltipla) de características morfológicas, bioquímicas, fisiológicas ou comportamentais (por exemplo, Block *et al.*, 1993; Meyer *et al.*, 1994; Sturmbauer *et al.*, 1996; Kitaura *et al.*, 1998; O' Foighil e Taylor, 2000) das espécies. Esse tipo de estudo está muito difundido hoje em dia, já que a possibilidade de uso de caracteres independentes (como caracteres moleculares) permite, sem o risco de raciocínio circular, a análise da origem dessas características. Por exemplo, O'Foighil e Taylor (2000) estudaram a origem e a evolução de características comportamentais em ostras através de sequências do gene ribossomal 28S. Dessa forma, o gene ribossômico representa um conjunto de dados independente do padrão de comportamento das ostras e, por essa razão, é mais adequado para um teste sobre a origem da característica testada.

Obviamente, a função de um gene influencia, por exemplo, sua variabilidade. O citocromo *c* está entre os genes mais conservados (menos variáveis) nos diversos grupos de organismos, certamente devido a sua função primordial na respiração celular. Esse viés conservador ocorre em todos os organismos de maneira equivalente, pois a função do citocromo c é basicamente a mesma em todos os organismos, de maneira que não irá influenciar na resolução da análise filogenética.

Por outro lado, é importante salientar que, nesse tipo de teste, não devemos usar genes ligados especificamente aos caracteres testados. Ou seja, não devemos usar um gene ligado à coloração de plumagem para testar filogeneticamente a evolução dessa coloração, já que isso resultaria em um viés ou uma circularidade em relação ao que queremos testar. Nessas circunstâncias, é possível que uma análise filogenética utilizando esse gene ligado à coloração apontasse para a origem única desses caracteres, distorcendo seu verdadeiro padrão evolutivo. Como a sequência de aminoácidos é determinante da função do gene, quaisquer testes envolvendo a origem dessa função, através dessas mesmas sequências, corresponderia a uma argumentação circular. Esse problema também pode ser aplicado a sequências de nucleotídeos, apesar de que, pelo fato de o código genético ser degenerado, convergências envolvendo aminoácidos (que determinam em última análise a função da proteína) não devem ser idênticas às de nucleotídeos.

16.3.4. Erros de amostragem

As técnicas de sequenciamento estão cada vez mais acessíveis para os laboratórios. Uma consequência imediata desse fato é a disponibilidade de um número maior de genes para a solução de problemas filogenéticos. Associado a esse aumento do volume de dados, a concatenação de vários genes em uma metassequência de alguns milhares de pares de bases tem se tornado comum (Murphy, *et al.* 2001a,b). Embora estatisticamente razoável (Nei *et al.*, 2001) e amplamente utilizado para diminuir o problema de amostragem limitada de dados, associado ao tamanho finito das sequências (Nei, 1986, Bucknam *et al*., 2006), a utilização de sequências concatenadas pode ser problemática. Nesta seção, discutiremos alguns aspectos importantes que precisam ser levados em conta antes da seleção de fragmentos.

Em primeiro lugar, filogenias baseadas em sequências concatenadas devem ter os genes alinhados individualmente antes da análise. Isso porque a concatenação de muitos genes inapropriados, como porções saturadas, podem mascarar o sinal filogenético dos genes apropriados.

Em segundo lugar, adicionar mais pares de bases às sequências não melhora necessariamente a qualidade da reconstrução filogenética. Não existe motivo para acreditar que uma filogenia baseada num número enorme de sequências, reunidas sem critério, seja mais acurada que uma baseada em um único gene com dados particularmente consistentes. De fato, o pesquisador deve buscar resultados robustos—qualidade, não quantidade é a melhor referência. Erros de amostragem de volume de dados é o erro derivado do volume pequeno dos dados. Há dois tipos principais de erros de amostragem: tamanho finito das sequências e amostragem taxonômica limitada. Ou seja, quanto maior for a sequência (número de nucleotídeos ou aminoácidos), menor é a probabilidade de a análise resultar em grupos devido a substituições paralelas por acaso, que não refletem a história filogenética dos organismos em questão

Por outro lado, quanto maior a sequência, mais provável será encontrar regiões com variabilidade diferente, ou seja, regiões que estão evoluindo a taxas distintas. Nesse caso, provavelmente poderemos incluir regiões com muito ruído (no caso daquelas que evoluem muito rapidamente) ou regiões pouco informativas (no caso daquelas muito conservadas) para aquele determinado problema filogenético. Esse tipo de erro pode causar problemas sérios, como dados moleculares incluindo todos os genes codificadores de proteínas do DNA mitocondrial levarem à reconstrução de uma filogenia profundamente equivocada, onde os mamíferos formam um grupo-irmão dos demais tetrápodes (Russo *et al.*, 1996; ver

Takezaki e Gojobori, 1999, para a solução do problema). Além disso, a análise de grandes quantidades de dados requer cautela do pesquisador com a ocorrência de erros resultantes de estatísticas tendenciosas. A razão disso é que, quando o tamanho amostral é muito grande, sua variância tende a zero e, dessa forma, o uso de métodos e modelos incorretos pode levar a sinais filogenéticos equivocados significativos (Delsuc *et al*., 2005).

16.4. Considerações Finais

O objetivo deste capítulo foi salientar a importância na seleção de genes e de trechos de genes para abordar a história filogenética de um grupo determinado, o que precisa ser levado em consideração antes de iniciarmos o sequenciamento. Dados úteis sobre a qualidade dos dados a serem usados nas análises filogenéticas necessitam ser avaliados antes do início da análise. O alinhamento é o primeiro filtro que deve ser usado para avaliar a adequação das sequências para reconstrução filogenética. Em seguida, metodologias simples, como o cálculo de uma matriz de distancia *p*, sempre auxilia na estimativa da variabilidade dos dados antes de realizar qualquer análise mais complexa. Além disso, deve-se prestar atenção à variação das taxas evolutivas dentro das sequências, pois sempre é mais fácil eliminar sítios não desejáveis que sequenciar outro gene. Obviamente, outros fatores, principalmente de ordem técnica, também devem ser levados em consideração, como disponibilidade de *primers*, facilidade de amplificação do fragmento etc. No entanto, é necessário ter claro que não adianta sequenciarmos 2 kb de cada espécie para a reconstrução de uma filogenia que, no final das contas, não apresenta resolução suficiente para estabelecer conclusões confiáveis. Uma análise prévia da variabilidade do fragmento a ser sequenciado é sempre recomendável.

Agradecimentos

Gostariamos de agradecer a Sergio Luiz Pereira e Cristina Yumi Miyaki por seus comentários em uma versão anterior do capítulo. Agradecemos também à FAPERJ (Fundação de Amparo a Pesquisa do Estado do Rio de Janeiro) e ao CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) pela concessão de bolsas e auxílios que permitiram a conclusão deste capítulo.

Referências Bibliográficas

Avise, J. e Johns, G.C. (1999). Proposal for a standardized temporal scheme of biological classification for extant species. **Proc. Natl. Acad. Sci. USA. 96**: 7358-7363.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. e Wheeler, D.L. (2008) GenBank. **Nucl. Acid Res. 36**: D25-D30.

Block, B.A., Finnerty, J.R., Stewart, A.F.R. e Kidd, J.(1993). Evolution of endothermy in fish-mapping physiological traits on a molecular phylogeny. **Science 260**: 210-214.

Bromham, L., Rambaut, A. e Harvey, P.H. (1996). Determinants of rate variation in mammalian DNA sequence evolution. **J. Mol. Evol. 43**: 610-621.

Bucknam, J., Boucher, Y. e Bapteste, E. (2006) Refuting phylogenetic relationships. **Biol. Direct 1**: 26.

Burks, C. (1997). Molecular Biology Databases. **In** Bishop, M.J. e Rawlings, C.J. (eds.). **DNA and protein sequence analysis - A practical approach**. IRL Press at Oxford University Press, New York, pp. 1-30.

Cavalier-Smith, T. (2003). Protist phylogeny and the high-level classification of Protozoa**. Eur J Protistology. 39**: 338-348.

Cavalli-Sforza, L.L. e Feldman, M.W. (2003). The application of molecular genetic approaches to the study of human evolution. **Nature Genet. 33**: 266-275.

Cunningham, C.W. (1997). Is congruence between data partitions a reliable predictor of phylogenetic accuracy? Empirically testing an interactive procedure for choosing among phylogenetic methods. **Syst. Biol. 46**: 464-478.

de Pinna, M.C.C. (1991). Concepts and tests of homology in the cladistic paradigm. **Cladistics 7**: 367-394.

Debruyne, R., Barriel, V. e Tassy, P. (2003). Mitochondrial cytochrome b of the Lyakhov mammoth (Proboscidea, Mammalia): new data and phylogenetic analyses of Elephantidae. **Mol. Phylogen. Evol. 26**: 421–434.

Delsuc, F., Brinkmann, H. e Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. **Nat. Rev. Genet. 6**: 361-375.

Ericson, P.G.P., Anderson, C.L., Britton, T., Elzanowski, A., Johansson, U.S., Kellersj, M., Ohlson, J.I., Parson, T.J., Zusson, D. e Mayr, G. (2006). Diversification of neoaves: integration of molecular sequence data and fossils. **Biol. Lett. 4**: 543-547.

Feng, D. e Doolitle, R.F. (1987). Progressive sequence alignment as a prerequisite to correct phylogenetic trees. **J. Mol. Evol. 60**: 351-360.

Gillham, N.W. (1994). **Organelle genes and genomes**. Oxford University Press.

Graur, D. e Li, W. -H. (2000). **Fundamentals of molecular evolution**. 2ª edição. Sinauer Press, Sunderland, Mass.

Green, R.E., Krause, J., Ptak, S.E., Briggs, A.W., Ronan, M.T., Simons, J.F., Du, L., Egholm, M., Rothberg, J.M., Paunovic, M. e Pääbo, S. (2006) Analysis of one million base pairs of Neanderthal DNA. **Nature 444**: 330-336.

Grishin, N.V. (1999). A novel approach to phylogeny reconstruction from protein sequences. **J. Mol. Evol. 48**: 264-273.

Hennig, W. (1966). **Phylogenetic systematics**. University of Illinois Press, Urbana.

Hickson, R.E., Simon, C. e Perrey, S.W. (2000). The performance of several multiple-sequence alignment programs in relation to secondary-structure features for an rRNA sequence. **Mol. Biol. Evol. 17**: 530-539.

Hoyle, D.C., Higgs P.G. (2003). Factors affecting the errors in the estimation of evolutionary distances between sequences. **Mol Biol Evol. 20**:1-9.

Huelsenbeck, J.P. (1995). Performance of phylogenetic methods in simulation. **Syst. Biol. 44**: 17-48.

Kelsey, C.R., Crandall, K.A. e Voevodin, A.F. (1999). Different models, different trees: the geographic origin of PTLV-1. **Mol. Phylogenet. Evol. 13**: 336-347.

Kitaura, J., Wada, K. e Nishida, M. (1998). Molecular phylogeny and evolution of unique mud-using territorial behavior in ocypodid crabs (Crustacea: Brachyura: Ocypodidae). **Mol. Biol. Evol. 15**: 626-637.

Kong, Q.P., Bandelt, H.J., Sun, C., Yao, Y.G., Salas, A., Achilli, A., Wang, C.Y., Zhong, L., Zhu, C.L., Wu, S.F., Torroni, A. e Zhang, Y.P. (2006). Updating the east asina mtDNA phylogeny: a prerequisite for the identification of pathogenic mutations. **Human Molecular Genetics 15**: 2076-2086.

Koonin, E.V. (2005). Orthologs, paralogs and evolutionary genomics. **Annu. Rev. Genet. 39**: 309-338.

Kuma, K. e Miyata, T. (1994). Mammalian phylogeny inferred from multiple protein data. **Jpn. J. Genet. 69**: 555-566.

Lartillot, N., Brinkmann, H. e Philippe, H. (2007). Suppression of long-branch attraction in the animal phylogeny using a site-heterogeneous model. **BMC Evol. Biol. 7**:S4.

Lewin, R. (1997). **Patterns in evolution: the new molecular view**. Scientific American Library.

Li, W.-H. (1997). **Molecular evolution**. Sinauer Press, Sunderland, Massachusetts.

Li, W-H. (2000). **Molecular Evolution**. Sinauer Associates, Sunderland, Massachusetts.

Li, W.H., Gu, Z.L., Wang, H.D., Nekrutenko, A. (2001). Evolutionary analyses of the human genome. **Nature 409**: 847-849.

Meyer, A., Morrissey, J.M. e Schartl, M. (1994). Recurrent origin of a sexually selected trait in xiphophorus fishes inferred from a molecular phylogeny. **Nature 368**: 539-542.

Mitchinson, G.J. (1999). A probabilistic treatment of phylogeny and sequence alignment. **J. Mol. Evol. 349**: 11-22.

Morgan-Richards, M., Trewick S.A., Bartosh-Härlid A., Kardailsky O., Phillips M.J., McLenachan P.A. e Penny, D. (2008). Bird evolution: testing the Metaves clade with six new mitochondrial genomes. **BMC Evol. Biol. 8**: 20-32.

Morrison, D.A. e Ellis, J.T. (1997). Effects of nucleotide sequence alignment on phylogeny estimation: a case study of 18S rDNAs of Apicomplexa. **Mol. Biol. Evol. 14**: 428-441.

Murphy, W.J., Eizirik, E., Johnson, W.E., Zhang, Y.P., Ryder, O.A.,

O'Brien, S.J. (2001a). Molecular phylogenetics and the origins of placental mammals. **Nature 409**: 614-8.

Murphy, W.J., Eizirik, E., O'Brien, S.J., Madsen, O., Scally, M,, Douady, C.J., Teeling, E., Ryder, O.A., Stanhope, M.J., de Jong, W.W., Springer, M.S. (2001b). Resolution of the early placental mammal radiation using Bayesian phylogenetics. **Science 294**: 2348-2351.

Nei, M. (1969). Gene duplication and nucleotide substitution in evolution. **Nature 221**: 40-42.

Nei, M. (1986). Stochastic errors in DNA evolution and molecular phylogeny. **In** Gershowitz, D.L. Rucknagel e R.E. Tashian (eds.). **Evolutionary perspectives and the new genetics**. Alan R. Inc. Press, pp. 133-147.

Nei, M. (1991). Relative efficiencies of different tree-making methods for molecular data. **In** M.M. Miyamoto e J. Cracraft (eds.). **Phylogenetic analysis of DNA sequences**. Oxford Press. New York, pp. 90-128.

Nei, M., Gu, X. e Sitnikova, T. (1997). Evolution by the birth-and-death process in multigene families of the vertebrate immune system. **Proc. Natl. Acad. Sci. USA 94**:7799-7806.

Nei, M. e Kumar,S. (2000). **Molecular Evolution and Phylogenetics**. Oxford University Press.

Nei, M. (2003). Genome evolution - Let's stick together. **Heredity 90**: 4511-412.

Nei, M., Takezaki, N. e Sitnikova, T. (1995). Assessing molecular phylogenies. **Science 267**: 253-256.

Nei, M, Xu P, Glazko G. (2001). Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms. Proc Natl Acad Sci U S A. 98:2497-2502.

O'Foighil, D. e Taylor, D.J. (2000). Evolution of parental care and ovulation behavior in oysters. **Mol. Phylogen. Evol. 15**: 301-313.

Osawa, S. e Honjo, T. (1991). **Evolution of life: fossils, molecules, and culture.** Springer-Verlag, New York; Tokyo.

Pääbo, S., Poinar, H., Serre, D., Jaenicke-Deprés, V., Hebler, J., Rohland, N., Kuch, M., Krause, J., Vigilant, L. e Hofreiter, M. (2004). Genetic Analyses from ancient DNA. **Annu. Rev. Genet. 38**: 645-679.

Palmer, J.D. (1985). Chloroplast dna and molecular phylogeny. **Bioessays 2**: 263-267.

Patterson, C. (1988). Homology in classical and molecular biology. **Mol. Biol. Evol. 5**: 603-625.

Phillips, A., Janies, D. e Wheeler, W. (2000). Multiple sequence alignment in phylogenetic analysis. **Mol. Phylogenet. Evol. 16**: 317-330.

Piontkivska, H and Nei, M. (2003). Birth-and-Death Evolution in Primate MHC Class I Genes: Divergence Time Estimates. **Mol Biol Evol. 20**: 424-434.

Russo, C.A.M. (1997). Efficiencies of different statistical tests in supporting a known vertebrate phylogeny. **Mol. Biol. Evol. 14**:1078-1080.

Russo, C.A.M., Takezaki, N. e Nei, M. (1995). Molecular phylogeny and divergence times of drosophilid species. **Mol. Biol. Evol. 12**: 391-404

Russo, C.A.M., Takezaki, N. e Nei, M. (1996). Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny. **Mol. Biol. Evol. 13**: 525-536.

Satta, Y., Klein, J. e Takahata, N. (2000). DNA archives and our nearest relative: the trichotomy problem revisited. **Mol. Phylogenet. Evol. 14**: 259-275.

Slowinski, J.B. (1998). The number of multiple alignments. **Mol. Phylogenet. Evol**. **10**: 264-266.

Sturmbauer, C., Levinton, J.S. e Christy, J. (1996). Molecular phylogeny analysis of fiddler crabs: Test of the hypothesis of increasing behavioral complexity in evolution. **Proc. Natl. Acad. Sci. USA 93**: 10855-10857.

Sullivan, J. e Joyce, P. (2005). Model selection in phylogenetics. **Annu. Rev. Ecol. Evol. Syst. 36**: 445-466.

Takezaki, N. (1998). Tie trees generated by distance methods of phylogenetic reconstruction. **Mol. Biol. Evol. 15**: 727-737.

Takezaki, N. e Gojobori, T. (1999). Correct and incorrect vertebrate phylogenies obtained by the entire mitochondrial DNA sequences. **Mol. Biol. Evol. 16**: 590-601.

Thompson, J.D, Higgins, D.G. e Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. **Nucleic Acid Res. 22**: 4673-4680.

Titus, T.A. e Frost, D.R. (1996). Molecular homology assessment and phylogeny in the lizard family Opluridae (Squamata: Iguania). **Mol. Phylogenet. Evol. 6**: 49-62.

Uzzell, T. e Corbin, K.W. (1971). Fitting discrete probability distribution to evolutionary events. **Science 172**: 1089-1096.

Wiens, J.J. e Servedio, M.R. (1998). Phylogenetic analysis and intraspecific variation: performance of parsimony, likelihood, and distance methods. **Syst. Biol. 47**: 228-253.

Wittzell, H., Bernot, A., Auffray, C. e Zoorob, R. (1999). Concerted evolution of two Mhc Class II B loci in pheasants and domestic chickens. **Mol. Biol. Evol. 16**: 479-490.

Woese, C.R., Kandler, O. e Wheelis, M.L. (1990). Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria and Eucarya. **Proc. Natl. Acad. Sci. USA 87**: 4576-4579.

Yang, Z. (1993). Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. **Mol. Biol. Evol. 10**:1396-1401

Yang, Z. (1996). Among-site rate variation and its impact on phylogenetic analyses. **Trends Ecol. Evol. 11**:367-372.

Zimmer, E.A., Hamby, R.K., Arnold, M.L.,. Leblanc, D.A e Theriot. E.C. (1989). Ribosomal RNA phylogenies and flowering plant evolution. In Fernholm, B., Bremer, K. e Jornvall, H. (eds.), **The hierarchy of life**, 205-214. Elsevier Science, Amsterdam.

Apêndice. Algumas páginas na rede relevantes em estudos de evolução molecular

Bancos de dados:

Nucleotídeos:

GenBank (http://www.ncbi.nlm.nih.gov/Entrez)
EMBL (http://www.ebi.ac.uk)
DDBJ (http://www.ddbj.nig.ac.jp)
Genome Sequence (http://www.ncgr.org/).

Proteínas:

UniProt (http://www.uniprot.org/)
Swiss-Prot (http://ca.expasy.org)

Alinhamento:

ClustalW

http://www.clustal.org/

Programas de Filogenia:

Revisão

(http://evolution.genetics.washington.edu/phylip/software.html)

MEGA

(http://www.megasoftware.net/)

Árvore da vida:

http://tolweb.org/tree/phylogeny.html

Capítulo 17

Polimorfismos de isozimas

Vera Nisaka Solferini (solferin@unicamp.br)
Departamento de Genética e Evolução
Instituto de Biologia
Universidade Estadual de Campinas

Denise Selivon (dselivon@ib.usp.br)
Departamento de Genética e Biologia Evolutiva
Instituto de Biociências
Universidade de São Paulo

"Tudo que existe no Universo é fruto do acaso e da necessidade." (Demócrito)

17.1. Histórico

Desde os primórdios da biologia evolutiva, uma das principais questões tem sido a caracterização e a explicação da variabilidade genética intra e interpopulacional. Atualmente existem muitas metodologias para a descrição da variabilidade genética, porém, até a década de 60, havia apenas alguns poucos organismos conhecidos adequados para esse tipo de estudo. Alguns tipos de variação fenotípica com herança mendeliana clássica podiam ser estudados, como o polimorfismo de grupos sanguíneos em humanos e algumas mutações em drosófila e milho. Além disso, havia os estudos com cromossomos politênicos em drosófila realizados pela escola de Dobzhansky, onde o principal objetivo era avaliar o efeito de grandes porções do genoma no valor adaptativo das populações. Nesses casos, embora os cromossomos politênicos oferecessem um alto grau de resolução na detecção de rearranjos estruturais, as interpretações eram limitadas no que se refere aos genes propriamente ditos.

As abordagens utilizadas limitavam, portanto, os organismos que podiam ser estudados: além de possuírem os marcadores adequados, precisam ser adaptados a condições de laboratório, pois havia a necessidade do acompanhamento por gerações consecutivas. Uma outra limitação era a quantificação de diferenças interespecíficas, até então impossível.

Quanto à interpretação das variações observadas, os geneticistas estavam praticamente divididos em dois pontos de vista conflitantes: Dobzhansky e seus colaboradores pertenciam ao que ele denominou de "escola balanceada", sustentando que, em uma população de reprodução sexuada, cada indivíduo seria heterozigoto para a maioria dos locos e que essa situação seria mantida por heterose (Dobzhansky, 1955). Por outro lado, os modelos matemáticos da denominada "escola clássica" de genética sustentava que praticamente todos os locos seriam homozigotos, com raras mutações deletérias que seriam responsáveis pela "carga genética" das populações (Muller, 1950). Um dos grandes problemas era que cada escola estava envolvida com metodologias e sistemas diferentes de estudo. Enquanto os seguidores da escola balanceada eram eminentemente drosofilistas, a escola de Muller contava com geneticistas que trabalhavam com mutações e que estavam preocupados com os efeitos da radiação e com uma legislação que limitasse sua utilização e regulamentasse os testes nucleares (a tecnologia nuclear era então relativamente recente e o impacto das explosões de Hiroshima e Nagasaki, ainda vívido—Crow, 1998).

Conceitualmente, a Genética de Populações estava limitada quanto a suas fontes de informação e dividida quanto à interpretação dos padrões de variabilidade encontrados. No início da década de 1960, já existia a metodologia de sequenciamento de aminoácidos em cadeias protéicas, porém sua aplicação para estudos populacionais ainda era impensável devido aos custos e à quantidade de trabalho envolvidos.

Em 1966, dois grupos publicaram independentemente seus resultados experimentais, introduzindo a técnica de eletroforese de proteínas (Harris, 1966; Hubby e Lewontin, 1966; Lewontin e Hubby, 1966). O método em si consistia da aplicação de dois conhecimentos correntes na época: sabia-se que a atividade e, portanto, a presença de algumas enzimas podia ser visualizada em extratos simples de organismos, através de coloração citoquímica. Sabia-se também que até mesmo a substituição de um único aminoácido em uma proteína podia alterar seu ponto isoelétrico, de forma que ela se movimentaria de forma diferente quando exposta a um campo elétrico (ou seja, a variação eletroforética em proteínas já era conhecida).

Do ponto de vista da aquisição de dados, pode-se dizer que a eletroforese de isozimas empregada em escala populacional representou uma revolução para a genética de populações: pela primeira vez, era possível ter acesso a um grande número de locos em qualquer organismo que se pretendesse estudar—bactérias, fungos, plantas e animais, dos mais diversos ambientes—e, em grande parte das vezes, utilizando apenas uma amostra de tecido do organismo. Por um certo tempo, houve uma verdadeira explosão no número de trabalhos descrevendo os padrões eletroforéticos dos mais diversos organismos. Pode-se dizer que foi uma época de exploração experimental da nova metodologia, antes de se começar a utilizá-la como uma ferramenta poderosa e aplicável a questões evolutivas.

Quanto à interpretação dos resultados, a grande quantidade de polimorfismos revelada com essa técnica mostrou que a manutenção da variabilidade observada não podia ser explicada pela heterose, como propunha o modelo balanceado. Tampouco estava de acordo com o modelo clássico, que postulava a existência de pouca variabilidade e alguns eventos raros de mutações deletérias.

Em 1971, foi proposto o modelo neutralista de evolução molecular, segundo o qual, ainda que alguns polimorfismos estivessem sendo mantidos por forças seletivas, a maior parte da variabilidade observada seria neutra; as novas mutações

poderiam ser fixadas por processos estocásticos e os polimorfismos observados seriam transitórios (Kimura e Ohta, 1971, Ohta, 1974).

A primeira metade da década de 1970 foi marcada por discussões apaixonadas entre os chamados selecionistas e os neutralistas, reavivando o antigo debate sobre o significado e a manutenção dos polimorfismos genéticos. Os selecionistas procuravam explicar a variabilidade enzimática como adaptativa, interpretando os padrões encontrados como resultado de processos seletivos. Já os neutralistas consideravam improvável tamanha quantidade de variação mantida por seleção e tratavam os padrões de variabilidade segundo modelos neutros.

Atualmente, alguns trabalhos foram capazes de demonstrar que diferentes genótipos alozímicos podem ter valores adaptativos diferentes, sendo inclusive possível correlacioná-los com fatores ambientais (Nevo, 1990; Powers *et al.*, 1991). Entretanto, são muito poucos os casos onde se conseguiu demonstrar diferenças no valor adaptativo de genótipos alozímicos. As isozimas (enzimas que atuam sobre o mesmo substrato) têm sido usadas para estudos de biologia evolutiva através de duas maneiras principais: o estudo detalhado de um loco e suas variantes alélicas (alozimas), e a utilização de muitos locos para estudos populacionais.

Quando se utilizam polimorfismos enzimáticos para a comparação entre populações ou espécies, diversos locos são amostrados em cada indivíduo analisado. Os métodos de análise tratam todos os locos da mesma maneira, ou seja, como adaptativamente neutros.

17.2. Fundamentos Metodológicos

O princípio básico da eletroforese é a migração diferencial de moléculas com cargas e tamanhos diferentes, quando submetidas a um campo elétrico. As proteínas são compostas por um ou mais polipeptídeos, que são cadeias de aminoácidos. A sequência de aminoácidos em um polipeptídeo é determinada pela sequência de nucleotídeos no gene que o codifica. Pequenas mudanças na sequência do DNA podem alterar a estrutura e, portanto, a mobilidade eletroforética de uma proteína. Uma das vantagens dessa técnica é que se pode amostrar vários indivíduos por vez e ter diversos locos analisados para cada um. A seguir, apresentamos um sumário das principais etapas dessa metodologia. Uma descrição detalhada pode ser encontrada em Alfenas (1998).

17.2.1. Suportes de Eletroforese

A eletroforese pode ser desenvolvida em vários tipos de suporte, sendo os mais utilizados para enzimas, o amido e a poliacrilamida, por permitir melhor separação que outros suportes, como agarose e acetato de celulose, por exemplo. Os géis podem ser preparados com tampões de composição e pH variados, de acordo com as enzimas que serão estudadas. A eletroforese pode ser feita em sistemas contínuos, onde os tampões do gel e do eletrodo são os mesmos, ou em sistemas descontínuos, onde o tampão do gel é distinto daquele dos eletrodos.

17.2.2. Extração das enzimas

A extração das enzimas é feita a partir da homogeneização de indivíduos inteiros (por exemplo, pequenos insetos) ou apenas órgãos ou tecidos, como fígado de vertebrados ou folhas, plântulas, sementes etc. Para a extração, é feito um macerado em solução tampão; para tecidos vegetais, é necessária a adição de alguns compostos ao tampão de extração para reduzir a formação de complexos fenólicos.

17.2.3. Aplicação das amostras no gel

O extrato de cada indivíduo pode ser aplicado diretamente nos géis em cavidades feitas durante a polimerização do gel ou ainda a solução de extração pode embeber pequenos retângulos de papel de filtro, aplicados em uma fenda feita no gel após a sua solidificação.

17.2.4. Corrida eletroforética

As condições, tanto de voltagem aplicada ao gel como de duração do tempo de corrida, variam de acordo com o sistema tampão utilizado. A corrida pode durar de 3 a 24 horas, com 2 a 12 v/cm. Como o princípio da técnica envolve a manutenção da atividade enzimática, o sistema como um todo, desde a extração do material individual até a corrida eletroforética, deve ser conduzido a baixas temperaturas. A Figura 17.1 ilustra um sistema de eletroforese horizontal.

17.2.5. Coloração

A distância de migração das enzimas é evidenciada por coloração histoquímica. O mecanismo de coloração inclui a presença do substrato específico da enzima, coenzimas, cofatores e um corante que precipita, oxida ou fluoresce em consequência da reação principal.

17.2.6. Interpretação dos géis

Para a interpretação do zimograma (padrão de bandas), é necessário o conhecimento da estrutura quaternária das enzimas. Assim, uma enzima monomérica (constituída por uma única cadeia polipeptídica) apresentará duas bandas no heterozigoto. Já uma enzima dimérica (formada pela união de duas cadeias polipeptídicas) apresentará três bandas no heterozigoto, e uma enzima tetramérica apresentará cinco, como ilustrado na Figura 17.2.

Os géis corados para os diferentes sistemas enzimáticos podem ser "lidos" interpretando-se geneticamente as bandas encontradas (Figura 17.3). Os genótipos individuais são anotados. Com isso, os dados podem ser utilizados para a obtenção de todos os parâmetros que levam em conta frequências gênicas e genotípicas, além de praticamente todas as análises desenvolvidas para variação qualitativa.

17.3. Polimorfismos Enzimáticos e Biologia Evolutiva

Os estudos sobre isozimas têm fornecido informações importantes para a biologia evolutiva. Esses dados têm permitido, entre outras coisas, a quantificação de níveis de variabilidade genética, estimativas de fluxo gênico, elucidação de limites interespecíficos e estabelecimento de relações evolutivas entre diferentes táxons.

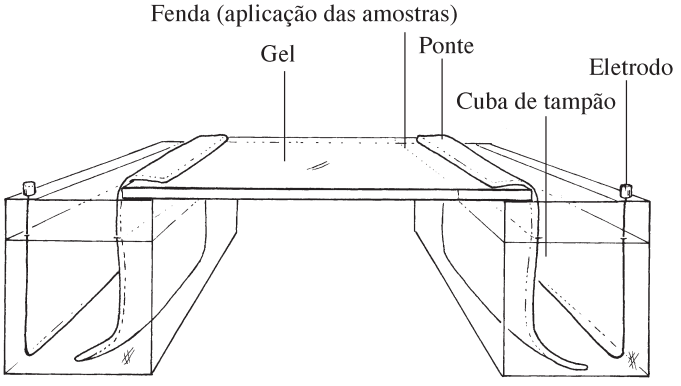


Figura 17.1. Esquema de um sistema para eletroforese horizontal em gel de amido.

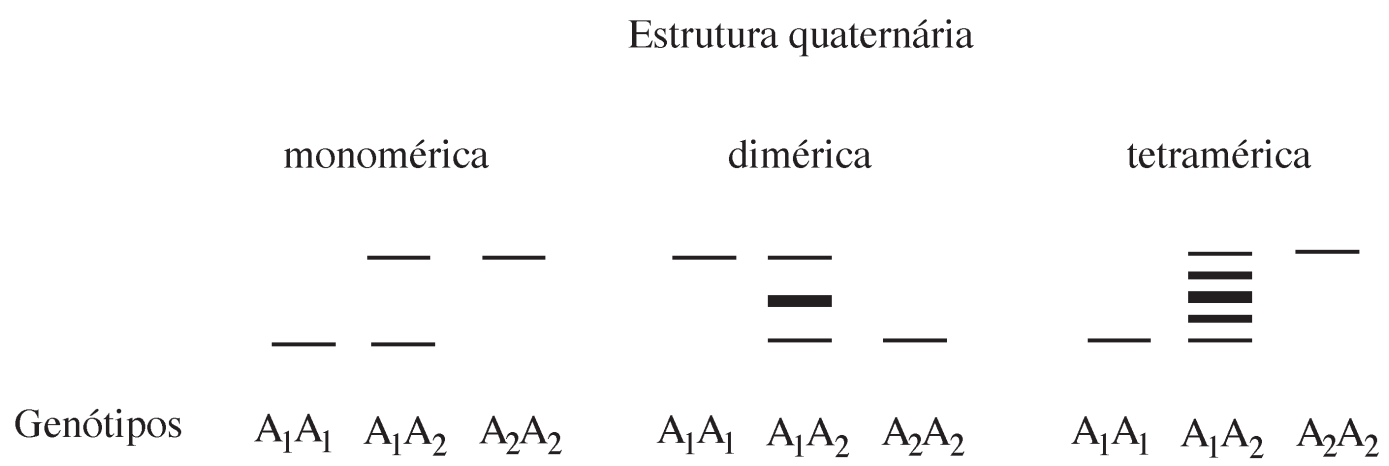


Figura 17.2. Zimogramas de sistemas com diferentes estruturas quaternárias.

É uma técnica de custo relativamente baixo, facilmente adaptável para qualquer grupo de organismos e que permite analisar um grande número de indivíduos e muitos locos de cada amostra. Existem, também, algumas dificuldades inerentes à metodologia e que devem ser consideradas quando de sua escolha—os níveis de variabilidade podem estar sendo subestimados, pois mutações no material genético nem sempre levam a alterações na estrutura protéica e nem toda alteração na sequência de aminoácidos provoca diferença na mobilidade eletroforética. Considerações sobre as aplicações e as limitações da eletroforese de isozimas podem ser encontradas em Murphy *et al.* (1996).

17.3.1 Medidas da variabilidade genética

A variabilidade genética de uma população pode ser quantificada através das frequências gênicas obtidas com a análise de isozimas. Pode-se considerar que os locos analisados representam uma amostra aleatória do genoma e, dessa maneira, seria representativa da população. Os parâmetros obtidos são interpretados como indicadores populacionais.

De um modo geral, a análise de isozimas é a maneira mais direta e rápida de avaliar genotipicamente muitos locos em um grande número de indivíduos. Por ser um marcador codominante (todos os alelos são evidenciados no gel, salvo quando da existência de alelos nulos), os genótipos individuais podem ser inferidos através do padrão de bandas. Pode-se calcular as frequências gênicas e genotípicas, a heterozigosidade observada na amostra e a esperada, testar se a população está em equilíbrio de Hardy-Weinberg (HW), a proporção de locos polimórficos e os coeficientes de endogamia. Pode-se, ainda, testar modelos de isolamento por distância.

Outro aspecto que se observa nos estudos de variação genética é que ela difere entre os táxons superiores. Inicialmente, considerou-se que os invertebrados teriam níveis de variação enzimática maiores que os vertebrados (Selander, 1976). Nevo (1978), no entanto, demonstrou que essa diferença não era significativa se os dados sobre as espécies de *Drosophila* fossem removidos dos cálculos. Posteriormente, verificou-se que os índices de variabilidade poderiam ser alterados, dependendo dos locos que eram considerados nos estudos comparativos, sendo alguns deles caracteristicamente mais polimórficos que outros (Koehn e Eanes, 1978). Entretanto, Ward *et al*. (1992) revisaram o assunto e concluíram que, de fato, é encontrada maior variabilidade entre os invertebrados, mesmo se levando em consideração essas restrições.

A comparação da variabilidade enzimática (em geral traduzida por valores de heterozigosidade média) entre várias espécies levou ao estabelecimento de correlações entre os níveis de variabilidade encontrados e componentes adaptativos e de ciclo de vida dos organismos. Uma das questões amplamente discutidas foi se a variabilidade encontrada nas proteínas estaria relacionada à heterogeneidade ambiental (Levene, 1953; Soulé e Stewart, 1970; Hedrik, 1986). Algumas associações foram estudadas, como, por exemplo, a correlação de padrões de variabilidade genética com distribuição geográfica (Bryant, 1974), a ocupação de diferentes microhábitats pela espécie (Powell e Taylor, 1979) e a estabilidade temporal de recursos tróficos (Ayala *et al.*, 1975; Valentine, 1976). Em 1978, Nevo apresentou uma revisão para 243 espécies de animais e vegetais, sugerindo que espécies generalistas, de ampla distribuição geográfica, grande mobilidade e que exploram um grande número de nichos ecológicos apresentam maior variabilidade genética do que espécies especialistas, com distribuição geográfica restrita e baixa mobilidade.

As estimativas de variabilidade genética frequentemente também são usadas para fazer inferências sobre eventos históricos que tenham influenciado a estrutura genética das populações. Por exemplo, baixos níveis de heterozigosidade nos locos enzimáticos podem ser interpretados como eventos recentes de afunilamento populacional, especialmente quando espécies relacionadas possuem níveis muito mais altos de variabilidade, embora, segundo alguns autores, a heterozigosidade nem sempre seja um bom indicador dos afunilamentos populacionais ocorridos no passado (Nei *et al*., 1975; Chakraborty e Nei, 1977; Leberg, 1992).

17.4. Estrutura Genética Populacional

A subdivisão populacional, por si só, pode afetar as frequências genotípicas. O modelo clássico de deriva mostra que

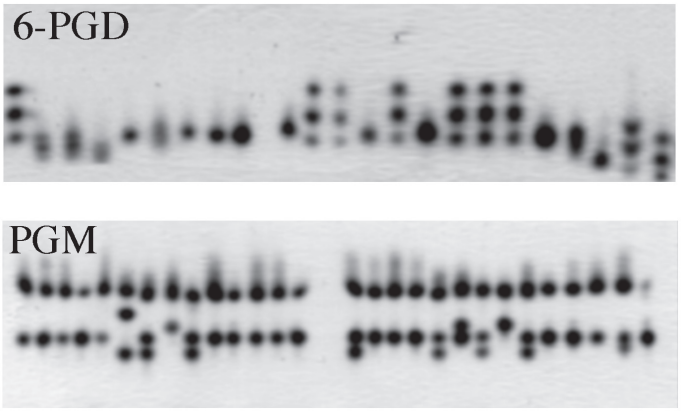


Figura 17.3. Exemplos de zimogramas em gel de amido de *Pleurothallis* (Orchidaceae). Desidrogenase do 6-fosfogluconato (6PGD); sistema com um loco, enzimas diméricas. Fosfoglicomutase (PGM); sistema com dois locos, com enzimas monoméricas (Borba *et al.*, 2000).

as frequências gênicas vão diferir entre as subpopulações e na espécie como um todo. Os genótipos homozigotos vão aumentar em frequência às expensas dos heterozigotos. Esse efeito pode ser quantificado através da estatística F (F*st* de Wright). Esse índice reflete a proporção da variabilidade genética encontrada entre populações, ou seja, devido à subdivisão populacional. Segundo Wright (1978), valores de F*st* entre 0,15 e 0,25 indicam estruturação populacional moderadamente alta, enquanto que valores acima de 0,25 refletem estruturação muito alta. A interpretação dos valores de F*st*, no entanto, depende do estudo das distribuições das frequências alélicas nas populações.

O polimorfismo protéico em populações naturais tem sido usado para descrever as mudanças de frequências alélicas, tanto no tempo quanto no espaço. Na escala espacial, a estrutura genética populacional é inferida a partir da distribuição geográfica das frequências alélicas. Em muitos casos, a variabilidade genética encontra-se distribuída em diferentes níveis hierárquicos, que se correlacionam com parâmetros da história de vida (Chesser, 1993).

A variabilidade alozímica também pode ser usada para inferir eventos históricos que influenciaram a estrutura genética das populações, tais como o "efeito gargalo" (*bottleneck*, em inglês). Níveis muito baixos de heterozigosidade em grandes extensões geográficas podem significar que um ou mais *bottlenecks* drásticos possam ter ocorrido num passado recente. No entanto, estimativas de heterozigosidade nem sempre são bons indicadores de gargalos populacionais (Nei *et al.*, 1975; Chakraborty e Nei, 1977; Leberg, 1992).

Estudos alozímicos, através da análise da distribuição geográfica de frequências alozímicas, podem fornecer informações úteis para a inferência de padrões de fluxo gênico e estruturação de acasalamentos (Slatkin, 1985, 1993). Outras abordagens também têm sido consideradas consistentes com esse tipo de análise. Por exemplo, a partir dos valores de F*st*, podem ser estimadas as taxas de migração entre as subpopulações analisadas (Borba *et al.*, 2000, 2001).

Por outro lado, muitas vezes, estudos genéticos mostram deficiência de heterozigotos na amostra estudada, o que pode ser consequência de endocruzamento ou ainda do efeito Wahlund (a inclusão de duas ou mais populações com frequências gênicas diferentes numa única amostra, mesmo que cada uma delas esteja em equilíbrio de HW individualmente).

Variação geográfica é encontrada em muitas espécies de insetos (Powell e Taylor, 1979; Slatkin, 1985; Futuyma e Peterson, 1985; Nevo, 1988) e surge por processos evolutivos locais, que envolvem interações de mecanismos genéticos e processos ambientais. A variação geográfica expressa-se em várias características biológicas (Baker e Tomas, 1987) e tem sido bem documentada em diversos grupos de organismos por diferentes abordagens, como análise alozímica, morfométrica, cromossômica e da variabilidade do DNA (White, 1973; Lewontin, 1974; Powell, 1975; Nevo, 1983a,b; Parsons, 1983, 1991; Lewontin, 1991; Daly, 1985; Nevo, 1988; Roderick, 1996).

Vários tipos de eventos podem levar à diferenciação genética entre populações: colonização de um novo hábitat (Bartlett e Richardson, 1986; Holt, 1987; Rice e Salt, 1988); colonização de um novo território ou região (efeito do fundador) (Barton e Charlesworth, 1984; Carson e Templeton, 1984; Bartlett e Richardson, 1986); afunilamento populacional (Nei *et al*., 1975; Hedrick, 1986); mudanças genéticas por eventos estocásticos, tais como deriva genética e mutação (Templeton, 1980; Bush, 1981), ou seleção natural (direcional e desruptiva) (Rice, 1987; Barker e Tomas, 1987). Embora esses processos sejam invocados para explicar a maioria dos modelos de especiação (para revisão, ver Barton e Charlesworth, 1984), a diferenciação genética das populações não conduz necessariamente à especiação, mas leva ao reconhecimento de categorias taxonômicas infra-específicas (Endler, 1977).

Além de as informações sobre a estrutura de populações naturais permitirem estudos de natureza evolutiva, também podem ser utilizadas como apoio em programas de manejo de espécies ameaçadas de extinção.

17.5. Medidas de Distâncias Genéticas

As distâncias genéticas são medidas da diferenciação entre populações utilizando dados de frequências genotípicas. É uma medida estatística, de quantificação das diferenças genéticas. As frequências gênicas e genotípicas, obtidas a partir da eletroforese de enzimas para indivíduos tomados ao acaso nas populações naturais, também podem fornecer uma estimativa da diferenciação intra e interespecífica. Essas frequências podem ser transformadas em uma série de índices que permitem estimar o grau de similaridade ou de distância genética entre diferentes espécies e populações. Além disso, variantes alélicas muitas vezes são encontradas isoladas em alguma população ou táxon e, nesse caso, podem ser usadas como um meio de identificação. Esse tipo de abordagem tem sido muito eficiente no reconhecimento de espécies de difícil identificação morfológica (crípticas) (Selander, 1976; Gould 1993). Entretanto, espécies de ampla distribuição geográfica podem apresentar diferenças locais para as variantes alélicas.

Embora as medidas de distância genética mais largamente usadas sejam aquelas propostas por Nei (1972, 1978), que não são estritamente geométricas, contamos ainda com outros algoritmos para o cálculo de distâncias, tais como os propostos por Farris (1981) e Rogers (1972). Uma exposição detalhada sobre os diferentes métodos de cálculo das distâncias genéticas é apresentada no Capítulo 13.

17.6. Perspectivas

O estudo de populações naturais por eletroforese de isozimas parece ter seguido uma tendência de crescimento expressivo durante as décadas de 1980 e 1990, seguida de um declínio a partir do início do século XXI (Figura 17.4). Várias são as possíveis razões para esse tipo de progressão. Em primeiro lugar, muitas espécies que vinham sendo alvo de estudos sistemáticos, em especial aquelas do hemisfério norte, já foram intensivamente caracterizadas com o emprego dessa metodologia. Além desse fator, houve a popularização da tecnologia da análise de DNA, pelos métodos de RFLP e de sequenciamento direto, especialmente com o emprego conjunto da técnica de PCR (veja Capítulos 18 e 19). Existem alguns componentes que podem ter contribuído para esse desvio de enfoque na pesquisa relacionada à medida de variabilidade genética e da quantidade de diferenciação populacional, aspectos que são mais frequentemente levados em consideração pelos estudos que envolvem alozimas. Em primeiro lugar, amostras que se destinam à análise por eletroforese de alozimas devem ser conservadas principalmente pela criopreservação, uma vez que a temperatura baixa é fundamental para a preservação da atividade enzimática, sem a qual seria impossível a detecção física das enzimas. As análises de DNA são muito menos rigorosas nesse sentido, pois se pode extrair DNA de amostras fixadas em solventes como álcool, simplesmente secas ou até mesmo de material muito antigo, com milhares ou até milhões

Figura 17.4. Curva que mostra a quantidade de artigos publicados com as palavras-chave "allozyme", "allozymes" ou "isozyme electrophoresis" no "Web of Science" (ISI, 2010), ao longo das últimas décadas.

de anos de idade (Anchordoquy e Molina, 2007). Outro fator que também pode ser importante é a redução de custo das metodologias para a análise de DNA, que foi muito significativa, ao passo que o custo da análise de amostras por eletroforese de isozimas praticamente não se alterou. Por exemplo, o grama de Beta-NAD custava, em 1975, a quantia, atualizada pela inflação acumulada, de 40 dólares, ao passo que, em 2010, a mesma substância do mesmo fornecedor, custa cerca de 60 dólares. O custo do sequenciamento de DNA em 1990 era orçado por cerca de 10 dólares por base, atualmente é uma fração de centavo de dólar. Finalmente, há o aspecto da tradição do laboratório que pode influenciar na aplicação da metodologia de alozimas, pois a curva de aprendizado para o emprego desta é mais lenta que aquela que existe naquelas relativas aos marcadores de DNA (experiência pessoal das autoras).

Apesar do declínio observado nas pesquisas que empregam eletroforese de isozima na caracterização genética de populações naturais, esse tipo de metodologia tem revelado padrões que são únicos, quando comparado com métodos baseados na tecnologia de DNA. Estudos recentes que abordaram justamente a comparação entre marcadores genéticos alozímicos e não alozímicos revelaram que o papel da seleção natural pode ser mais relevante nos marcadores alozímicos que em outros marcadores. Riginos *et al.* (2002) sugeriram que a introgressão do marisco *Mytilus edulis* em populações do Mar Báltico de *M. trossulis* deixou um efeito bem maior em marcadores de DNA nucleares que nos marcadores alozímicos devido ao efeito da seleção natural estabilizadora ou balanceada que atua nestes últimos. Em populações das ilhas britânicas de *M. edulis*, Silva e Skibinski (2009) também apontaram a ação de seleção natural balanceada em locos de alozimas para explicar os diferentes padrões de variabilidade genética observado nas populações desses animais. Os diferentes padrões de variabilidade genética revelados por alozimas, marcadores de DNA mitocondriais e nucleares também foram apontados como efeito de seleção estabilizadora ou balenceada que atuariam em alguns locos de alozimas em populações de ostras (Arnauld-Haond *et al.*, 2003). Por outro lado, uma comparação feita com relação à variabilidade genética de marcadores genéticos alozímicos, de DNA nuclear e de DNA mitocondrial, com base em milhares de estudos revelou diferenças marcantes na variabilidade do DNA nuclear, que é mais homogênea (Bazin *et al.*, 2006). Os autores apontam a ação da seleção natural positiva que atuaria no caso dos genoma mitocondriais

17.7. Considerações Gerais

O estudo dos padrões de variabilidade genética em populações naturais fornece um subsídio para hipóteses acerca das causas dos padrões observados, mas não necessariamente incluem a análise direta dos mecanismos. Esses podem ser inferidos a partir do estudo correlato de diferentes características biológicas, aliado a um desenho experimental cuidadoso, incluindo amostragens em escalas temporal e espacialmente significativas. Apenas o conjunto de dados poderá fornecer a oportunidade de interpretar os mecanismos responsáveis pelos processos demográficos, reprodutivos e geográficos das espécies em questão.

Um das grandes dificuldades está na escolha dos marcadores moleculares mais adequados a serem empregados. Essa escolha muitas vezes é influenciada pela disponibilidade de recursos, além do modismo e da experiência dos pesquisadores. O desenvolvimento das metodologias que permitiram a análise direta dos ácidos nucléicos deu a impressão de que a eletroforese de alozimas seria abandonada. De fato, esses marcadores apresentaram uma potencialidade maior para estudo de polimorfismos, dada sua abundância. Na maioria dos casos, todavia, marcadores moleculares anônimos, como é a maioria daqueles empregados com a metodologia de DNA, não têm funções fisiológica e bioquímica conhecidas, o que não ocorre com as isozimas. Além disso, quando há variabilidade alozímica, essa não só é uma abordagem para estimar a variabilidade genética, como ainda hoje fornece um grande número de informações sobre a variabilidade genética intra e interespecífica. Por esses motivos, a análise de isozimas ainda é uma das melhores abordagem para o estudo de espécies próximas, assim como o de populações conspecíficas (Avise, 1994).

Referências Bibliográficas

Alfenas, A. C. (1998). **Eletroforese de isoenzimas e proteínas afins**. Editora UFV, Universidade Federal de Viçosa, Brasil.

Anchosdoquy, T.J. e Molina, M.C. (2007). Preservation of DNA. **Cell Preserv. Technol. 5**(4):180-188.

Avise, J. C. (1994)**. Molecular markers, Natural History and Evolution**. Chapman and Hall. New York, NY.

Ayala, F.J., Valentine, J.W., Hedgecock, D. e Barr, L.G. (1975). Deep-sea asteroids: high genetic variability in a stable environment. **Evolution 19**: 203-212.

Barker, J.S.F. e Thomas, R.H. (1987). A quantitative genetic perspective on adaptive evolution. **In** Loeschcke, V. (ed.), **Genetic Constraints on Adaptive Evolution**. Berlin, Springer-Verlag, pp. 3-23.

Bartlett, D.C.H. e Richardson, B.J. (1986). Genetic attributes of invading species. **In** Groves, R.H. e Burdon, J.J. (eds.), **Ecology of Biological Invasions. Cambridge**, Cambridge Univ. Press, pp. 21-33.

Barton, N.H. e Charlesworth, B. (1984). Genetic revolutions, founder effects and speciation. **Annu. Rev. Ecol. Syst**. **15**: 133-164.

Bazin, E., Glémin, S. e Galtier, N. (2006). Population size does not influence mitochondrial genetic diversity in animals. **Science 312**:570-572.

Borba, E.L.; Felix, J.M.; Semir, J; Solferini, V.N. (2000). *Pleurothallis fabiobarrosii*, a new Brazilian species: morphological and genetic data, and notes on the taxonomy of Brazilian rupicolous *Pleurothallis*. **Lindleyana 15**: 2-9,

Borba, E.L.; Felix, J.M.; Solferini, V.N. e Semir, J. (2001). Fly pollinated *Pleurothallis* (Orchidaceae) have high genetic variability: evidence from isozyme markes. **Am. J. Bot**. **88**: 418-428.

Bryant, E.H. (1974). On the adaptive significance of enzyme polymorphisms in relation to the environment. **Am. Nat. 108**: 1-9.

Bush, G.L. (1981). Stasipatric speciation and rapid evolution in animals. **In** Atchley, W.R. e Woodruff, D.S. (eds.). **Evolution and Speciation in Honor of M.J.D. White**. Cambridge, Cambridge University Press, pp. 201-218.

Carson, H.L. e Templeton, A.R. (1984). Genetic revolutions in relation to speciation phenomena: the founding of new populations. **Ann. Rev. Ecol. Syst**. **15**: 97-131.

Chakraborty, R. e Nei, M. (1977). Bottleneck effects on average heterozigosity and genetic distance with the stepwise mutation model.

Evolution 31: 347-356.
Chesser, R.K. (1993). Genetic variability within and among populations of the black-tailed prairie dog. **Evolution**, **37**: 320-3333.
Crow, J.F. (1998). Overdominance. A half century later. **In** Hecht, M.K. *et al*. (eds.). **Evolutionary Biology**. Plenum Press, New York, pp. 1-13.
Daly, H.V. (1985). Insect morphometrics. **Annu. Rev. Entomol**. **30**: 415-438.
Dobzhansky, T. (1955). A review of some fundamental concepts and problems of population genetics. **Cold Spring Harbor Symp. Quant. Biol**. **20**: 1-15.
Endler, J.A. (1977). **Geographic variation, speciation and clines.** Princeton, NJ, Princeton Univ. Press.
Farris, J.S. (1981). Distance data in phylogenetic analysis. **In** Funk, V. e Brooks, D.R. (eds.). **Advances in Cladistics: Proceedings of the First Meeting of the Willi Hennig Society**. New York Botanical Garden, New York.
Futuyma, D.J. e Peterson, S.C. (1985). Genetic variation in the use of resources by insects. **Annu. Rev. Entomol**. **30**: 217-238.
Gould, F. (1993). The spatial scale of genetic variation in insect populations. **In** Kim, K. C. e McPheron, B. A. (eds.). **Evolution of Insect Pests: Patterns of Variation**. John Wiley e Sons, Inc. New York, pp. 67-85.
Harris, H. (1966). Enzyme polymorphism in man. **Proc. Royal Soc. London**, Ser.B **164**: 298-310.
Hedrick, P.W. (1986). Genetic polymorphism in heterogeneous environments: a decade later. **Ann. Rev. Ecol. Syst. 17**: 535-566.
Holt, D.R. (1987). Population dynamics and evolutionary process: The manifold roles of habitat selection. **Evol. Ecol. 1**: 331-47.
Hubby, J. L. e Lewontin, R.C. (1966). A molecular approach to the study of genetic heterozygosity in natural populations. I. The number of alleles at different loci in *Drosophila pseudoobscura*. **Genetics 54**: 577-594.
ISI - Institute of Scientific Information (2010). http://www.isiknowledge.com.
Kimura, M. e Ohta, T. (1971). **Theoretical Aspects of Population Genetics**. Princeton, Princeton Univ. Press, 1971.
Koehn, R.K. e Eanes, W.F. (1978). Molecular structure and protein variation within and among population. **Evol. Biol. 11**: 39-100.
Leberg, P.L. (1992). Effects of population bottlenecks on genetic diversity as measures by allozyme electrophoresis. **Evolution 46**: 477-494.
Levene, H. (1953). Genetic equilibrium when more than one ecological niche is available. **Amer. Natur. 87**: 331-333.
Lewontin, R.C. e Hubby, J.L. (1966). A molecular approach to the study of genetic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. **Genetics 54**: 595-609.
Lewontin, R.C. (1974). **The genetics basis of evolutionary change**. Columbia Univ. Press, New York.
Lewontin, R.C. (1991). Twenty-Five Years Ago in Genetics: Electrophoresis in the Development of Evolutionary Genetics: Milestone or Millstone? **Genetics 128**: 657-662.
Muller, H.J. (1950). Our load of mutations. **Am. J. Hum. Genet. 2**: 111-176.
Murphy, R.W., Sites, J.W., Buth, D.G. e Haufler, C.H. (1996). Proteins: Isozyme Electrophoresis. **In** Hillis, D.M., Moritz, C. e Mable, B.K. **Molecular Systematics**. Sinauer, Sunderland, MA, 2ª ed. 1996, pp. 51-120.
Nei, M. (1972) Genetic distance between populations. **Am. Nat**., **106**: 283-292.
Nei, M. (1978). Estimation of average heterozygosity and genetic distance from a small number of individuals. **Genetics 89**: 583-590.
Nei, M., Maruyama, T., Chakraborty, R. (1975). The bottleneck effect and genetic variability in populations. **Evolution**, **29**: 1-10.
Nevo, E. (1978) Genetic variation in natural populations: patterns and theory. **Theor. Popul. Biol**. **13**: 121-177.
Nevo, E. (1983a). Population genetics and ecology: The interface. **In** Bendall, D.S. (ed.) **Evolution from Molecules to Men.** Cambridge University Press, pp. 287-321.
Nevo, E. (1983b). Adaptive significance of protein variation. **In** Oxford, G.S., Rollinson, D. (eds.). **Polymorphism: Adaptive and Taxonomic Significance.** Academic Press, New York, pp. 239-282.
Nevo, E. (1988). Genetic Diversity in Nature: Patterns and theory. **Evol. Biol**., **23**: 217-246.
Nevo, E. (1990) Molecular evolutionay genetics of isozymes: pattern, theory, and application. **In Isozymes: Structure, Function, and Use in Biology and Medicine (**Proceedings ofthe Sixth International Congress on Isozymes Held at Toyama, Japan), pp 701-742.
Ohta, T. (1974). Mutational pressure as the main cause of molecular evolution and polymorphism. **Nature 252**: 351-354.
Parsons, P.A. (1983). **The Evolutinary Biology of Colonizing Species**. Cambridge: Cambridge Univ. Press.
Parsons, P.A. (1991). Evolutionary rates: Stress and species boundaries. **Annu. Rev. Ecol**. **Syst. 22**: 1-18.
Powell, J.R. (1975). Protein variation in natural population of animals. **In** Dobzhansky, T., Hecht, M.K., Steere, W.C. (ed). **Evol. Biol**. **8**: 79-119.
Powell, J.R., Taylor, C.E. (1979). Genetic variation in ecologically diverse environments. **Am. Sci**. **67**: 590-596.
Powers, D.A., Lauerman, T., Crawford, D. e DiMichele, L. (1991). Genetic Mechanisms for adapting to a changing environment. **Annu. Rev. Genet. 25**: 629-659.
Rice, W.R. (1987). Speciation via habitat specialization. **Evol. Ecol**. **1**: 301-314.
Rice, W.R. e Salt, G.W. (1988). Speciation via disruptive selection on habitat preference: Experimental evidence. **Am. Nat**. **131**: 911-917.
Riginos, C., Sukhdeo, K. e Cunningham, C. W. (2002). Evidence for selection at multiple allozyme loci across a mussel hybrid zone. **Mol. Biol. Evol. 19**(3):347–351.
Roderick, G.K. (1996). Geographic Structure of Insect Populations: Gene Flow, Phylogeography, and Their Uses. **Ann. Rev. Entomol**. **41**: 325-352.
Rogers, J.S. (1972). Measures of genetic similarity and genetic distance. **Univ. Texas Publ. 7213**:145-153.
Selander, R.K. (1976). Genetic variation in natural populations. **In** Ayala, F.J. (ed.). **Molecular Evolution**. Sunderland, Mass., Sinauer Associates, pp. 21-44.
Silva, E.P. e Skibinski, O.F. (2009). Allozymes and nDNA markers show different levels of population differentiation in the mussel *Mytilus edulis* on British coasts. **Hydrobiologia 620**:25-33.
Slatkin, M. (1985). Gene flow in natural populations. **Annu. Rev. Ecol. Syst. 16**: 393-430.
Slatkin, M. (1993). Isolation by distance in equilibrium and non-equilibrium populations. **Evolution 47**: 264-276.
Soulé, M. e Stewart, B.R. (1970) The "niche-variation" hypothesis: a test and alternatives. **Amer. Natur. 104**: 85-97.
Templeton, A.R. (1980) The theory of speciation via the founder principle. **Genetics 94**: 1011-1038.
Valentine, J.W. (1976). Genetic strategies of adaptation. **In** Ayala, F.J. (ed.), **Molecular Evolution**. Sinauer Sunderland, Massachussetts, pp. 78-94.
Ward, R.D., Skibinski, D.O.F. e Woodwark, M. (1992). Protein heterozygosity, protein structure and taxonomic differentiation. **Evol. Biol. 26**: 73-159.
White, M.J.D. (1973). **Animal cytology and evolution**. Cambridge Univ. Press, London.
Wright, S. (1978). **Evolution and the genetics of populations. Vol. 4: variability within and among natural populations**. University of Chicago Press, Chicago.

Capítulo 18

RFLP: O emprego de enzimas de restrição para detecção de polimorfismos no DNA

Maria Cristina Arias (mcarias@ib.usp.br)
Departamento de Genética e Biologia Evolutiva
Instituto de Biociências
Universidade de São Paulo

Maria Elena Infante-Malachias (marilen@usp.br)
Escola de Artes, Ciências e Humanidades
Universidade de São Paulo.

"*A razão é um apoio fundamental à minha crença, é um imenso suporte à minha fé. Como cientista, examino os fatos e, baseado neles, formulo raciocínios, alguns testáveis, outros não, pois nem tudo é Ciência.*" (Paulo Nogueira Neto)

18.1. Histórico e Impacto das Técnicas de Análise de DNA na Detecção de Variabilidade Genética

Desde o desenvolvimento da eletroforese em gel de amido (Smithies, 1955) e da visualização histoquímica das enzimas em géis (Hunter e Market, 1957), os estudos clássicos de Hubby e Lewontin (1966) e Lewontin e Hubby (1966) levaram à descoberta e utilização de uma nova classe de marcadores genéticos, como foi detalhado no Capítulo 17. A análise de enzimas representou um avanço frente às técnicas citológicas e morfológicas, já que esses marcadores genéticos, as enzimas, não são tão profundamente afetados pela interação com o ambiente e possibilitaram um grande avanço na compreensão dos processos micro e macroevolutivos, através de variadas aplicações. Apesar da extensiva utilização da técnica de eletroforese de proteínas, os estudos de estrutura populacional e outras aplicações para análises intraespecíficas requeriam níveis suficientes de variabilidade, de maneira que os resultados das análises de isoenzimas não se mostravam suficientemente variáveis em alguns organismos, fazendo com que, para esses casos, a eletroforese de proteínas não fosse a técnica mais apropriada.

No fim da década de 1960, Linn e Arber (1968) e Meselson e Yuan (1968) revolucionaram a nascente biologia molecular com a descoberta das enzimas de restrição ou endonucleases de restrição. Essas enzimas foram isoladas e purificadas de várias linhagens de bactérias e são capazes de digerir uma determinada sequência específica de DNA dupla fita, de quatro, cinco ou seis pares de bases (pb) de comprimento. Nas bactérias, as enzimas de restrição agem protegendo a célula contra a invasão de DNA exógeno. O DNA da própria bactéria é protegido da ação das endonucleases por um sistema específico de metilação. A denominação das enzimas de restrição tem relação com a linhagem bacteriana da qual foram isoladas, por exemplo, a enzima *Eco* RI, foi extraída da bactéria *Escherichia coli* e digere a sequência específica 5'-GAATTC-3'.

As enzimas de restrição começaram a ser de extrema utilidade nos laboratórios de genética molecular, encontrando diversas aplicações especialmente para a detecção de polimorfismos no DNA, este tipo de marcador foi denominado RFLP (do inglês, *Restriction Fragment Length Polymorphism*) e está baseado no padrão de fragmentos produzidos ao digerir um determinado DNA com uma endonuclease específica.

A variabilidade detectada através do RFLP reflete as variações nas sequências de DNA, as quais são a base da diversidade dentro de uma espécie. Em muitos casos, essa variação de sequência é fenotipicamente neutra, pois grande parte das mutações que alteram a sequência de nucleotídeos (e, portanto, a sequência de reconhecimento de uma determinada enzima de restrição) somente é mantida porque são mutações silenciosas, isto é, não causam efeito no indivíduo. O número dessas variações é elevado, podendo chegar a 1 de cada 100 nucleotídeos, e não está sujeito à influência do ambiente, representando, desta forma, uma rica fonte de polimorfismo a ser explorada.

O desenvolvimento de técnicas para o estudo do DNA através do RFLP propiciou um enorme e acelerado conhecimento da biossistemática, evolução e genética de animais e plantas. A grande sensibilidade destas técnicas, associada à descoberta de regiões com alto grau de variabilidade, tem permitido avanços significativos em estudos de genética de populações e biogeografia de diversos organismos, e em particular em populações humanas, onde foram desenvolvidas aplicações na medicina forense, na determinação de paternidade e no diagnóstico de doenças hereditárias.

Os marcadores do tipo RFLP apresentam algumas vantagens em relação a marcadores morfológicos, citológicos ou de isozimas. Entre eles, destacam-se: (a) são herdados como marcadores mendelianos livres de efeitos pleiotrópicos; (b) não são afetados pelo ambiente; (c) podem ser obtidos em número elevado; e (d) têm distribuição aleatória no genoma.

18.2. Princípios da Técnica de RFLP

A técnica de RFLP envolve a clivagem de moléculas de DNA por enzimas de restrição, separação dos fragmentos gerados por eletroforese em gel e visualização dos mesmos em forma de bandas. Cada banda corresponde a um grupo de moléculas de mesmo tamanho. Padrões diferentes de tamanho de fragmentos podem ser encontrados entre diferentes indivíduos, essa variabilidade é o reflexo de inserções, deleções, rearranjos (como inversões) que possam ter ocorrido em uma região ou, ainda, substituições de base na sequência de nucleotídeos, que é reconhecida por uma determinada enzima de restrição, alterando, dessa forma, o

número de sítios de clivagem e, como consequência, o tamanho dos fragmentos.

Para a realização dessa técnica, três fatores devem ser levados em conta: (1) o tipo de DNA que será analisado; (2) o substrato eletroforético a ser usado; e (3) os métodos de visualização dos fragmentos. Esses três fatores encontram-se intimamente relacionados. O tipo de DNA a ser estudado é um item que merece maior destaque e, portanto, será discutido em seção específica.

18.2.1. Substrato para eletroforese

Normalmente, o substrato mais utilizado para a separação dos fragmentos por eletroforese é a agarose, mas a poliacrilamida também é utilizada. Esses dois substratos formam uma malha fina por onde as moléculas de DNA migram, de acordo com a carga e o tamanho dos fragmentos, quando uma corrente elétrica é aplicada ao substrato. Fragmentos menores migram mais facilmente por essa malha que os fragmentos maiores. A molécula de DNA possui carga negativa quando em pH neutro, portanto migra em direção ao ânodo durante a eletroforese. A taxa de migração é determinada pelo tamanho dos fragmentos gerados após digestão com endonucleases. O tipo de substrato eletroforético a ser usado é determinado em função dos tamanhos dos fragmentos que se deseja separar. Um gel de agarose é indicado para separar fragmentos grandes, variando de 300 a 20.000 pb. Ainda é possível variar a concentração da agarose para uma boa resolução de fragmentos grandes ou pequenos, de acordo com os objetivos desejados. Por exemplo, géis menos concentrados (0,6 - 0,8%) são indicados quando se deseja separar com boa resolução fragmentos grandes e géis mais concentrados (1,0 - 2,0 %) quando se deseja separar fragmentos pequenos. Atualmente pode-se contar com outros tipos de agarose que permitem separar com boa resolução fragmentos menores que 300 pb. A poliacrilamida é indicada para a separação também de fragmentos pequenos. Podemos, de acordo com a concentração usada (3,5 a 20%), separar fragmentos de 10 a 1000 pb.

18.2.2. Visualização dos fragmentos

Depois da corrida eletroforética o passo seguinte é a visualização dos fragmentos, o que normalmente ocorre por métodos químicos. Quando se tem uma quantidade de DNA de mais de 50ng por banda, o brometo de etídeo é o indicado. Esse agente intercala a dupla fita de DNA e se torna fluorescente quando exposto à luz ultravioleta. Pelo fato de intercalar em moléculas de DNA, fragmentos maiores apresentam bandas mais intensas, podendo-se então fazer uma correlação direta entre a intensidade da banda e o tamanho do fragmento com a quantidade de DNA. É importante lembrar que o brometo de etídeo é um agente mutagênico, de modo que se deve sempre trabalhar com luvas e evitar contaminação de bancadas e cubas. A coloração por prata também tem sido empregada para ácidos nucléicos e, apesar de envolver várias etapas, é um método mais sensível, podendo detectar pequenas quantidades de DNA (picogramas). Um terceiro método de detecção de bandas em gel é por meio de material radioativo. Essa metodologia implica a marcação das moléculas de DNA com fósforo ou enxofre radioativos (P^{32} e S^{35}, respectivamente) antes do fracionamento dos fragmentos em gel. Esse método é extremamente sensível e a intensidade das bandas não possui uma relação direta com o tamanho do fragmento. A detecção dos fragmentos dá-se, nesse método, pela exposição do gel a um filme de raio X. Esse tipo de marcação é extremamente eficaz na detecção de fragmentos pequenos. Outra técnica altamente empregada, quando o rendimento de ácidos nucléicos, especialmente DNA, é baixo devido à má preservação do material ou mesmo à pouca quantidade de tecido disponível para a extração do DNA, é a técnica de "Southern blot" (Southern, 1975). O nome dessa técnica poderia ser traduzido como "mata borrão de Southern", devido ao sobrenome do pesquisador que a desenvolveu. Como uma espécie de trocadilho, as transferências de gel de eletroforese para membranas de proteínas e de RNA foram denominadas de "Western blot" e de "Northern blot", respectivamente. A metodologia designada como "Southern blot" envolve a digestão do DNA por enzimas de restrição, separação em gel de agarose e posterior transferência dos fragmentos para uma membrana de náilon ou nitrocelulose. A membrana é então hibridizada com uma sonda de DNA fria ou radioativa (Figura 18.1). As condições de hibridação, como temperatura ou concentração de sais, são determinadas de acordo com o grau de similaridade entre o DNA-alvo a ser detectado e o DNA da sonda. Normalmente, utilizam-se condições bem restritivas, ou seja, temperaturas altas (58 - 60°C), quando se trata de sonda derivada do mesmo organismo (sonda específica), e condições mais brandas (50°C) quando se realiza a hibridação entre organismos pertencentes a espécies, gêneros e mesmo ordens diferentes (sondas interespecíficas). O uso de sonda interespecífica muitas vezes pode levar a falsos resultados pelo fato de apresentar menor similaridade com o DNA alvo e poder se ligar a outras regiões que não aquelas de interesse. No entanto, não devem ser descartadas, mas sim analisadas com mais critério. No caso de uso de sondas radioativas, a revelação dos fragmentos dá-se pela exposição da membrana a um filme de raio X e, no caso de sonda fria, a revelação dá-se por imunoensaio e precipitação de corante na própria membrana. Para a análise de RFLP, podemos usar sondas para todo um genoma, como no caso do mitocondrial, ou somente para algumas regiões do genoma que são hipervariáveis, como as regiões de mini e microssatélites de origem nuclear.

Em todo gel a ser analisado, independentemente da metodologia aplicada, deve-se sempre acrescentar um marcador de tamanho molecular, o qual permitirá uma estimativa, por interpolação através de uma curva de calibração em escala logarítmica, do tamanho dos fragmentos gerados em suas amostras. Normalmente utilizamos o DNA do fago lambda digerido com *Hind*III para fragmentos que variam de 23.000 a 500 pb e o DNA do fago ϕ X174 RF digerido com *Hae*III para fragmentos entre 1300 e 300 pb. Existe uma grande variedade desses marcadores e devem ser escolhidos de acordo com o tamanho dos fragmentos que se esperam.

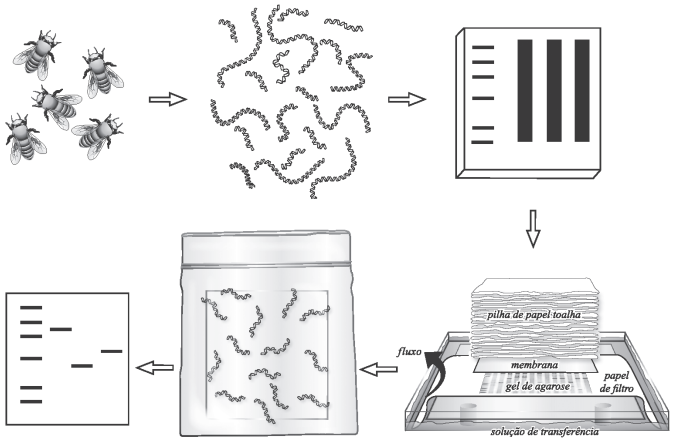


Figura 18.1. Esquema representando a técnica de "Southern Blot", incluindo extração de DNA total de um organismo (A e B), separação desse DNA em gel de agarose por eletroforese após digestão com enzimas de restrição (C), transferência dos fragmentos para uma membrana (D), hibridação da membrana com uma sonda específica (E) e revelação da membrana apresentando os fragmentos marcados (F).

Detalhes e protocolos sobre as técnicas para análise de RFLP poderão ser obtidos em Sambrook *et al.* (1989), Ausubel *et al.* (1993) e Hillis *et al.* (1996).

18.3. Natureza da Variação de Tamanho de Fragmentos

Enzimas de restrição (RFLP) têm sido amplamente utilizadas em estudos envolvendo o DNA (nuclear, mitocondrial e ou de cloroplasto) para análise da variabilidade genética intra e interespecífica. Como foi descrito no item anterior, a técnica de RFLP pode ser descrita sucintamente pela digestão de um determinado fragmento de DNA com uma ou várias enzimas de restrição, separação dos fragmentos gerados de acordo com seu peso molecular em um gel de eletroforese e finalmente a visualização e interpretação dos fragmentos gerados. Dessa forma, o DNA de cada indivíduo analisado apresentará seu próprio padrão de fragmentos específicos, denominado perfil de digestão, sendo que, entre várias amostras de DNA, podem ser detectados perfis invariáveis, ou seja, padrões de fragmentos idênticos para uma mesma endonuclease, ou perfis variáveis, isto é, padrões de bandas diferentes entre indivíduos.

As diferenças nos perfis de digestão para uma mesma endonuclease entre indivíduos pode ser resultado de vários eventos: (a) alteração no número e distribuição dos sítios de restrição devido à substituição de bases dentro dos sítios de reconhecimento (corte) da enzima; e (b) alterações no DNA devido a adições ou deleções de sequências (indels). Cada uma dessas forças de variação produz mudanças características no padrão de fragmentos gerados. Dessa forma, a variação pode ser detectada através do número ou do tamanho dos fragmentos gerados. Na Figura 18.2, estão representados alguns eventos que geram polimorfismo nos fragmentos de restrição e o padrão de bandas esperado quando analisadas em gel por eletroforese.

Cada enzima de restrição possui um sítio específico de reconhecimento onde é realizado o corte. Geralmente essas sequências são simétricas e possuem entre quatro e seis pares de bases, podendo produzir fragmentos com extremidades 3' e 5' protuberantes (coesivas) ou extremidades retas (Fig. 18.3). A substituição de bases ou pequenos eventos de inserção ou deleção podem criar ou eliminar sítios de restrição numa determinada sequência para uma enzima em particular, alterando desta forma o número de fragmentos gerados e o tamanho dos mesmos.

Na análise de RFLP, é possível identificar o polimorfismo como diferença de tamanho nos fragmentos gerados, mas não é possível identificar diretamente sua origem. Essa variação pode ser examinada por outras técnicas moleculares.

Rearranjos de sequência, como duplicações e inversões ou eventos de inserção ou deleção de segmentos maiores, tipicamente alteram o padrão de fragmentos não apenas para uma senão para várias enzimas simultaneamente, resultando numa alteração

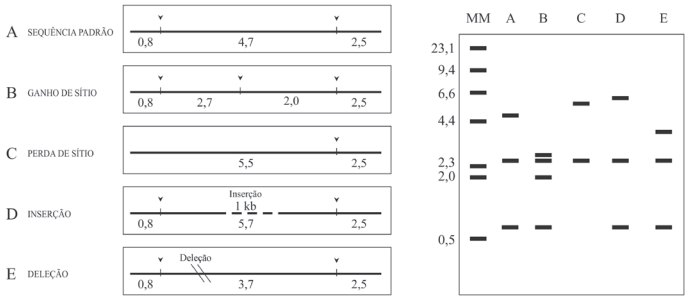


Figura 18.2. Mapa de restrição (A) indicando o padrão de corte de uma determinada enzima e a representação dos possíveis mecanismos que podem alterar esse padrão (B – E). O esquema à direita representa a separação dos fragmentos em gel de agarose e seus respectivos tamanhos em Kb. MM: marcador molecular λ/*Hind* III.

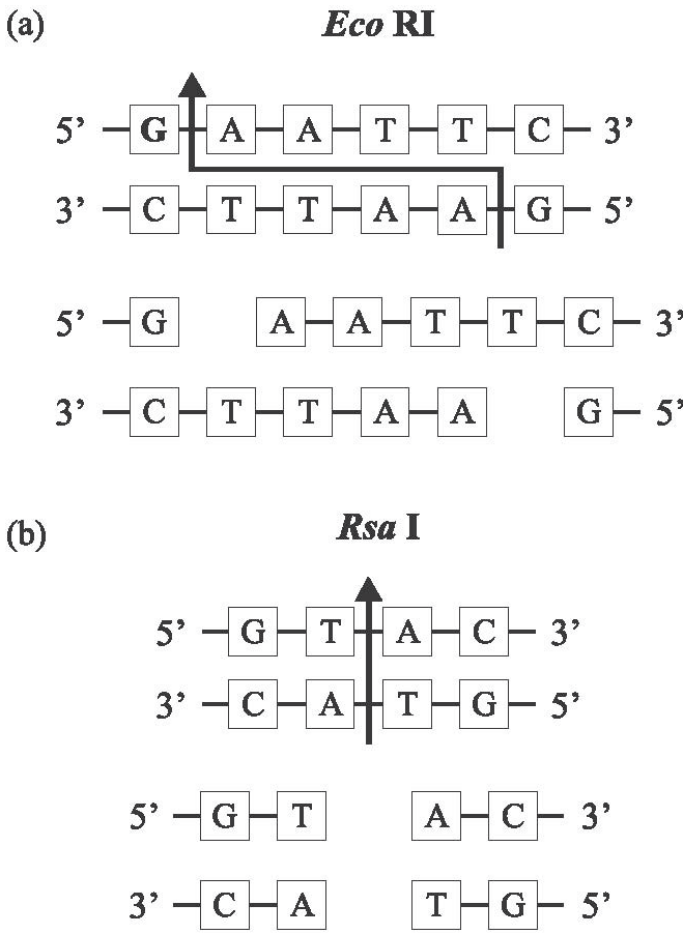


Figura 18.3. Sequências de nucleotídeos que constituem sítios de clivagem para duas enzimas de restrição. A enzima *Eco* RI reconhece uma sequência de seis pares de base e, ao cortá-la, gera extremidades coesivas (A); a enzima *Rsa* I reconhece uma sequência de quatro pares de base e a corta ao meio gerando pontas retas (B).

correlacionada com o padrão de restrição de diferentes enzimas, mostrando, dessa forma, a não independência entre caracteres, o que será discutido com maior detalhe na próxima seção.

18.4. Herança, Reprodutibilidade e Independência do Marcador RFLP

Quando se trabalha com fragmentos de DNA produzidos por uma enzima de restrição (RFLP), assume-se que os marcadores gerados são herdáveis, que podem ser reproduzidos, isto é, têm reprodutibilidade e que esses caracteres são independentes.

Assumindo-se que os fragmentos produzidos por RFLP são herdáveis, devem ser considerados dois elementos: fidelidade de transmissão e o modo de herança. A fidelidade de transmissão pode ser alterada no caso de sítios reconhecidos por enzimas sensíveis à metilação. A variação no estado de metilação da molécula de DNA pode imitar o estado de perda ou ganho de sítios de restrição, alterando a interpretação dos resultados. Esses artefatos gerados por diferentes estados de metilação podem ser evitados com a utilização de isoesquizômeros, que são enzimas de restrição isoladas de diferentes bactérias, que reconhecem a mesma sequência de DNA, mas que diferem em sua sensibilidade a sítios metilados. No entanto, a metilação aparentemente não é problema quando se trata de análise de RFLP de DNA mitocondrial (mtDNA) ou DNA de cloroplasto (cpDNA), mas pode resultar numa variação maior do que realmente existe e uma aparente homoplasia (veja Capítulo 12) de sítios específicos em sequências nucleares.

O modo de herança deve ser considerado com extremo cuidado quando se trabalha com RFLP, pois, dependendo da molécula (tipo de DNA) com que se trabalha, o modo de herança pode apresentar diferenças, levando a erros na interpretação do padrão de bandas obtido. Em geral, sabe-se que o DNA mitocondrial animal (com algumas exceções) é herdado por via materna. No caso do DNA de cloroplasto, há espécies que apresentam herança biparental (materna e paterna) e espécies que apresentam herança exclusivamente paterna. No caso do DNA nuclear, na maioria das vezes, o comportamento dos fragmentos gerados por RFLP é característico de herança mendeliana.

Outra característica importante do marcador RFLP é sua reprodutibilidade, ou seja, o padrão de fragmentos produzidos por uma determinada enzima de restrição para um dado DNA deve ser mantido e permanecer constante quando a digestão é refeita. Quando isto não ocorre, pode se pensar em artefatos de técnica produzidos por alteração em algum parâmetro específico, como tempo de corrida, concentração do gel e do tampão, concentração do DNA-alvo ou de enzima, temperatura e tempo de digestão, somente para citar alguns dos principais fatores que podem alterar a reprodutibilidade de um padrão de bandas no gel.

Finalmente, para a premissa de independência de caracteres no RFLP, é importante saber que, se as sequências de reconhecimento de duas enzimas de restrição se sobrepõem, esse marcador já não será independente. Para exemplificar, vejamos o caso da enzima *Mbo*I, cuja sequência de reconhecimento (GATC) se sobrepõe à da enzima *Bam*HI (GGATCC). Dessa maneira, a não independência de caracteres pode causar erros na estimativa da divergência de sequência ou da filogenia a partir de dados de RFLP.

18.5. Utilização de RFLP para a Construção de Mapas de Restrição

Uma das primeiras caracterizações normalmente feitas para moléculas de DNA ou mesmo para fragmentos isolados de DNA é a análise por enzimas de restrição. Conhecer quais enzimas cortam um dado fragmento e o número de sítios de restrição por elas gerados é extremamente útil em vários aspectos. Os sítios de restrição, quando mapeados, dão uma identidade para a molécula de DNA, ou melhor, caracterizam essa molécula, fornecendo pontos específicos, cujas distâncias entre eles são conhecidas. Esse mapeamento serve como base para a localização de genes, para a clonagem de regiões de interesse em vetores moleculares e também como marcador em estudos evolutivos.

Construir um mapa de restrição para uma dada molécula de DNA, seja apenas de um fragmento isolado ou de todo um genoma, como no caso do mitocondrial, significa posicionar os vários sítios de restrição gerados por várias enzimas uns em relação aos outros. Para obter essa relação, é necessário que digestões simples e duplas sejam realizadas, de modo que possamos comparar os padrões gerados por duas enzimas individualmente e quando combinadas, situação em que se espera que bandas novas surjam e outras desapareçam. É importante, para a construção de mapas de restrição, conhecermos o tamanho do fragmento ou genoma na sua forma íntegra, pois, após as digestões, a soma dos fragmentos gerados deve totalizar o tamanho do DNA em estudo. Bandas muito pequenas muitas vezes não são visualizadas em géis de agarose, porém a diferença entre o tamanho total conhecido e o obtido nas digestões pode nos dar uma idéia do tamanho do fragmento não visualizado. Muitas vezes, bandas duplas podem ser geradas. Isto é, fragmentos de regiões diferentes, porém com o mesmo tamanho, migram para a mesma posição no gel. Essas bandas normalmente aparecem mais espessas ou mais intensamente coradas com brometo de etídio, dando uma indicação desse tipo de situação. Outras técnicas de mapeamento, como digestões parciais ou hibridização em série, também podem ser aplicadas (Hillis *et al.*, 1996), mas a comparação do padrão de fragmentos originado por simples e duplas digestões é a mais frequentemente usada.

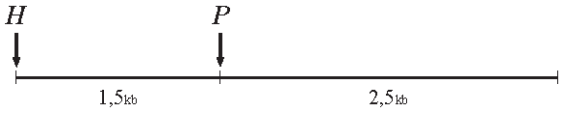
Como, então, proceder? Enzimas que reconhecem seis bases como seu sítio de clivagem geram um menor número de fragmentos, enquanto que aquelas que reconhecem quatro bases geram um número bem maior. Se o fragmento ou genoma que desejamos mapear for muito pequeno, provavelmente apresentará poucos sítios para enzimas que cortam em sequências de seis bases e um número razoável de sítios para enzimas que reconhecem quatro bases. Para fragmentos grandes ou mesmo para um genoma completo, como o mitocondrial, teremos um número razoável de fragmentos gerados pelas enzimas que reconhecem seis bases e um número muito grande de fragmentos para as enzimas que reconhecem quatro bases. Neste último caso, teremos muita dificuldade em analisar tanto o número de fragmentos gerados, como construir um mapa de restrição com essas enzimas. Um outro fator que deve ser levado em conta é o conteúdo de bases do genoma ou do fragmento de DNA a ser analisado. Genomas ricos em A+T, quando digeridos por enzimas que reconhecem uma sequência composta por C e G, sendo de seis ou mesmo quatro bases, resultarão em poucos fragmentos. Portanto, a escolha do tipo de enzima a ser utilizada dá-se de acordo com os objetivos, o tamanho e as características da sequência de DNA em análise.

O tamanho total do DNA em análise e a soma dos tamanhos dos fragmentos gerados devem ser consistentes, assim como o número de bandas observadas nas digestões simples deve ajudar a inferir o número de bandas esperado quando se realizam duplas digestões. Por exemplo, uma enzima que reconhece apenas um sítio resulta em apenas uma banda, quando se trata de genoma circular. Quando combinada com uma segunda enzima que reconhece dois sítios, e portanto produz dois fragmentos, espera-se como resultado dessa dupla digestão três bandas no gel, cuja soma de tamanhos deve totalizar o tamanho total do DNA em análise. Um número de bandas menor que o previsto pode ser devido a dois fatores: (1) fragmentos do mesmo tamanho migrando para a mesma posição (comigração); ou (2) sítios de restrição das duas enzimas muito próximos, gerando um fragmento tão pequeno que pode ter saído do gel durante a eletroforese. Neste último caso, a soma total dos fragmentos deverá ser apenas ligeiramente menor que a esperada para toda a molécula.

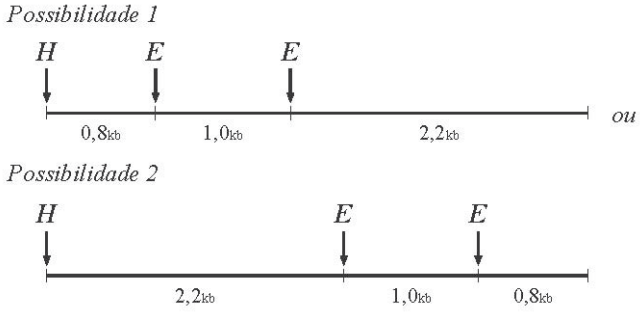
A realização de duplas digestões utilizando diferentes combinações de enzimas poderá esclarecer a proximidade dos sítios de restrição. A escolha das enzimas a serem combinadas duplamente deve dar-se de acordo com o número de vezes que cada uma corta individualmente a molécula em estudo. Normalmente começa-se por aquelas que cortam em número menor e, estando estas mapeadas, pode-se combinar aquelas que apresentam padrões mais complexos. Por hora, vamos tomar como exemplo, baseando-nos em Hillis *et al.* (1996), uma molécula de DNA circular de 4 Kb. Essa molécula foi digerida por quatro enzimas de restrição diferentes (*E*, *H*, *P* e *X*), individualmente e combinadas duas a duas, os fragmentos foram separados em gel de agarose e visualizados por coloração com brometo de etídeo. O tamanho dos fragmentos gerados foi calculado a partir do DNA padrão, cujos fragmentos são de tamanhos conhecidos, separados no mesmo gel; esses valores estão apresentados na Tabela 18.1.

Vamos começar com as enzimas *P* e *H*, que parecem cortar essa molécula de 4 Kb apenas uma vez cada uma; essas seriam as indicadas para o início do mapeamento. Vale ressaltar aqui a diferença no padrão de corte por enzimas entre genomas

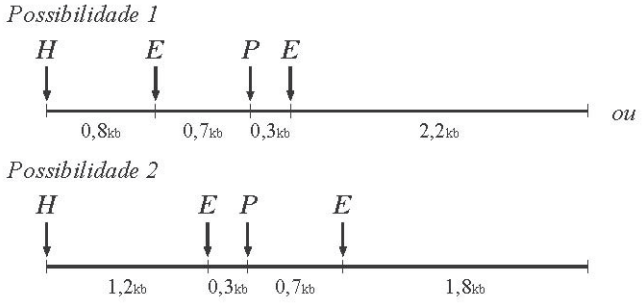
circulares e fragmentos ou genomas lineares. Normalmente, em genomas circulares, enzimas que cortam apenas uma vez geram fragmentos lineares cujo tamanho corresponde ao tamanho do próprio genoma intacto. Devido a essa característica, muitas vezes podemos incorrer no erro de pensar que a enzima não cortou o DNA em estudo. A certeza de que houve ou não clivagem só pode ser obtida mediante duplas digestões com enzimas que também supostamente cortem uma única vez ou poucas vezes. Esse problema não é visto quando se trabalha com genomas ou fragmentos lineares. Sempre que a enzima cortar, teremos como resultado duas bandas e, se não cortar, teremos o fragmento intacto. O único problema ocorre quando o sítio de corte é muito próximo a uma das extremidades do DNA linear em estudo. Voltando ao nosso exemplo, as enzimas *H* e *P* parecem cortar esse genoma circular apenas uma vez, tornando-o linear. A dupla digestão realizada apresentou dois fragmentos, um de 2,5 e outro de 1,5 kb. Assim, as duas enzimas cortam esse DNA e o mapa proposto é:



O ponto zero desse fragmento foi considerado no ponto de corte da enzima *H*. Uma terceira enzima (*E*) corta essa molécula em dois sítios, de maneira que, quando combinada com *H* ou *P*, apresenta três fragmentos (Tabela 18.1). As possibilidades para

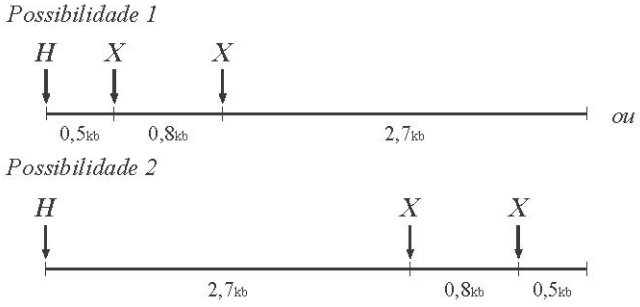


Os dados obtidos para essas duas enzimas não nos permitem concluir qual das duas possibilidades é a correta. Acrescentando dados de uma terceira enzima (*P*), poderemos chegar a uma conclusão. As possibilidades para a relação entre os sítios das enzimas *P* e *E* são:

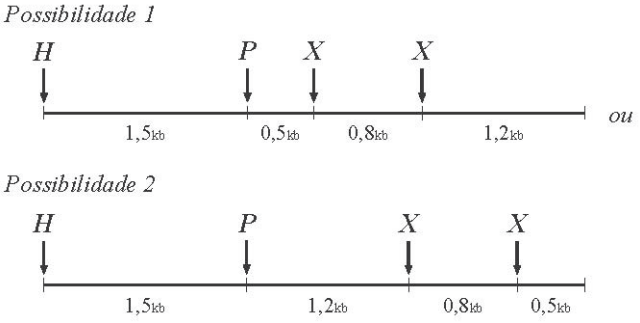


Como concluir qual das duas possibilidades é a correta? A resposta está no resultado obtido pela dupla digestão de *E* + *H*. Com base no tamanho dos fragmentos obtidos nessa dupla digestão (Tabela 18.1), concluimos que a possibilidade correta é a 1.

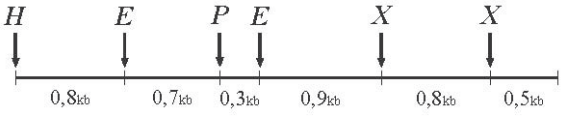
Incluindo em nossa análise a enzima *X*, que também corta essa molécula duas vezes e que, quando combinada separadamente com as enzimas *H* e *P,* gera três fragmentos, podemos ter diferentes possibilidades de mapa. Com relação à enzima *H:*



Com relação à enzima *P*, temos:



A resposta para as posições corretas no mapa vem da digestão *P* + *X* e do tamanho dos fragmentos gerados (Tabela 18.1), de modo que a possibilidade 2 é a correta. Para acompanhar todo esse raciocínio, é necessário ter o tamanho dos fragmentos gerados nas digestões simples e duplas, e verificar qual possibilidade de mapeamento dos sítios é condizente com o padrão e tamanho de bandas gerados. Nosso mapa final para essas quatro enzimas é o seguinte:



Uma vez tendo o mapa determinado, ele poderá servir de base para o mapeamento de novas enzimas e mesmo para o mapeamento de outros indivíduos. O grau de dificuldade neste último caso dependerá da divergência entre os organismos estudados, sendo que muitas vezes é necessário um mapeamento em separado para sequências cujos padrões são de difícil interpretação entre ganho e perda de sítios.

A Figura 18.4 ilustra um mapa de restrição para o genoma mitocondrial obtido para uma espécie de abelha-sem-ferrão do

Tabela 18.1. Tamanho dos fragmentos gerados pelas enzimas de restrição *P, H, E* e *X* após a digestão do nosso DNA circular hipotético de 4 kb. Os tamanhos estão apresentados em kb.

Enzimas	***P***	H	***PH***	***E***	***EP***	EH	***X***	***XP***	XH
Tamanho dos fragmentos	4	4	2,5	3	3	2,2	3,2	2	2,7
	-	-	1,5	1	0,7	1,0	0,8	1,2	0,8
	-	-	-	-	0,3	0,8	-	0,8	0,5
Total	4	4	4	4	4	4	4	4	4

Brasil, *Plebeia remota* (Francisco *et al.*, 2001). Para tanto, foram realizadas digestões simples e duplas, e o padrão de bandas foi visualizado por "Southern blot*"* utilizando sonda mitocondrial interespecífica, derivada da abelha *Apis mellifera*.

Em estudos de caráter evolutivo, o mapeamento de regiões homólogas entre diferentes organismos (populações, subespécies e espécies) fornece-nos dados sobre a homologia entre os sítios de restrição mapeados. Tais dados podem ser utilizados para gerar uma matriz com caracteres binários (presença e ausência de sítios), para análise filogenética posterior.

18.6. RFLP do DNA Mitocondrial Animal

O DNA mitocondrial (mtDNA) animal é uma molécula de fita dupla circular (genomas mitocondriais lineares são descritos na literatura, porém constituem exceções) e que codifica aproximadamente 5% de toda a maquinaria necessária para o funcionamento da mitocôndria. Foram descritos 37 genes, dos quais 13 codificam RNA mensageiros para proteínas envolvidas diretamente no processo do transporte de elétrons e fosforilação oxidativa, dois para subunidades ribossômicas e 22 para RNA transportadores. Existe uma região não codificadora, conhecida como A+T nos invertebrados, ou "alça D" ou "D-loop" nos vertebrados, que contém o controle da replicação e transcrição desse genoma. O mtDNA é tido como uma molécula econômica, pois toda sua sequência de nucleotídeos possui função codificadora. Regiões espaçadoras são raramente descritas e introns, pseudogenes e sequências repetitivas são estruturas ainda mais raras ou ausentes. A Figura 18.5 ilustra uma molécula de mtDNA animal. O conteúdo em termos de genes e o arranjo em sua ordem são extremamente conservados; algumas diferenças na ordem são relatadas normalmente em comparações entre organismos que divergiram há muito tempo, como entre táxons ao nível das categorias taxonômicas de ordem e classe. Na maioria das vezes, essas alterações de ordem, ou seja, translocações de genes, envolvem genes codificadores de RNA transportador.

Além dessa estrutura simples de organização, uma outra característica que torna essa molécula muito atrativa para estudos populacionais e evolutivos é o fato de apresentar herança citoplasmática, isto é, ser herdada via materna, de modo que ela não segue os padrões mendelianos de segregação e não sofre recombinação. Algumas exceções são descritas na literatura, mostrando contribuição paterna na herança do mtDNA (Zouros *et al.,* 1992) e evidências de recombinação entre moléculas de mtDNA (Ashkenas, 1997; Howell, 1997). A herança materna—e a não recombinação entre as moléculas—faz com que um genótipo ou haplótipo mitocondrial possa servir para traçar uma genealogia materna ou mesmo filogenia materna, o que muitas vezes pode ajudar a entender o modo de dispersão de muitos organismos, acasalamentos preferenciais etc.

Uma terceira característica fundamental dessa molécula é sua alta taxa de evolução. Acredita-se que seja dez vezes superior à de um gene de cópia nuclear única (veja também o Capítulo 7). Algumas explicações apontadas para esse fenômeno são uma baixa eficiência do sistema de reparo na mitocôndria, visto que todas as enzimas envolvidas nesse processo são codificadas pelo genoma nuclear e, portanto, importadas para a mitocôndria ou, ainda, sua alta exposição a radicais livres que possam ser gerados durante o processo da respiração. Apesar dessa alta taxa de evolução, os indivíduos normalmente são homoplásmicos, isto é, apresentam somente um tipo de molécula de mtDNA em todas as células de todos os tecidos. Parece existir uma tendência para a homogeneização dessas moléculas quando condições de heteroplasmia (isto é, existência de mais de um tipo de mtDNA em um mesmo indivíduo) ocorrem. Apesar da homogeneidade qualitativa das moléculas de mtDNA em um mesmo indivíduo, existe uma diferença entre a taxa de acúmulo de substituições de base entre os diferentes genes ou regiões do mtDNA. Alguns genes acumulam mais rapidamente essas substituições, dentre eles os genes codificadores das subuni-

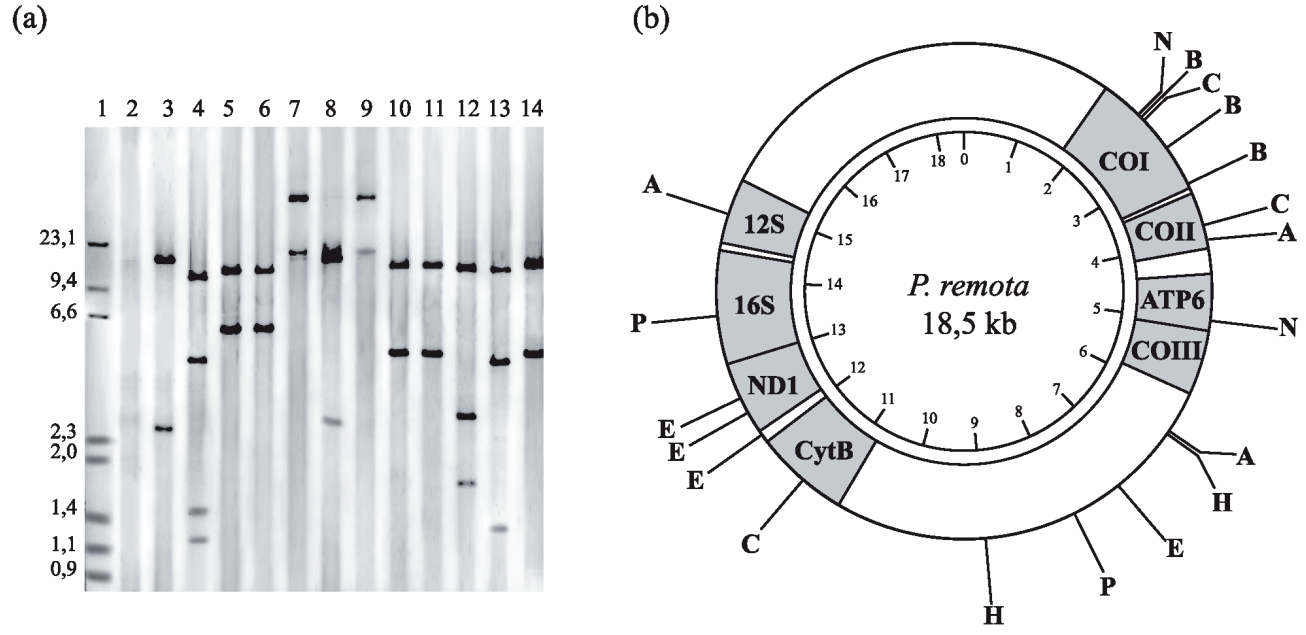


Figura 18.4. Exemplo de construção de um mapa de restrição para o genoma mitocondrial de uma espécie de abelha endêmica do Brasil (simplificado de Francisco *et al.*, 2001). (a) Esquema com o resultado de "Southern blot" utilizando sonda fria para detecção dos fragmentos de restrição resultantes de simples e duplas digestões (1. Marcadores de peso molecular λ/*Hind*III e ϕ X174/ *Hae*III; 2. DNA de *Apis mellifera*/*Hind*III; 3-14 DNA de *Plebeia remota* digerido com as respectivas enzimas: 3. *Hind*III; 4. *Hind*III/*Pst*I; 5. *Pst*I; 6. *Pst* I/*Xba*I ; 7. *Xba*I; 8. *Xba*I/*Hind*III; 9. *Bgl*II; 10. *Bgl*II/*Eco* RI; 11. *Eco* RI; 12. *Eco* RI/*Hind*III; 13. *Eco* RI/*Pst* I; 14. *Eco* RI/ *Xba*I). (b) Mapa de restrição de *P. remota* cujo tamanho total do genoma é aproximadamente 18,5 Kb. As letras indicam os sítios de restrição: A (*Hae*III), B (*Bcl*I), C (*Cla*I), E (*Eco* RI), H (*Hind*III), N (*Nde*I), P (*Pst*I). Para um maior refinamento, utilizou-se também a técnica PCR+RFLP, onde fragmentos conhecidos foram amplificados e, posteriormente, clivados pelas enzimas, permitindo não só o melhor posicionamento para sítios muito próximos, mas também a localização de alguns genes mitocondriais.

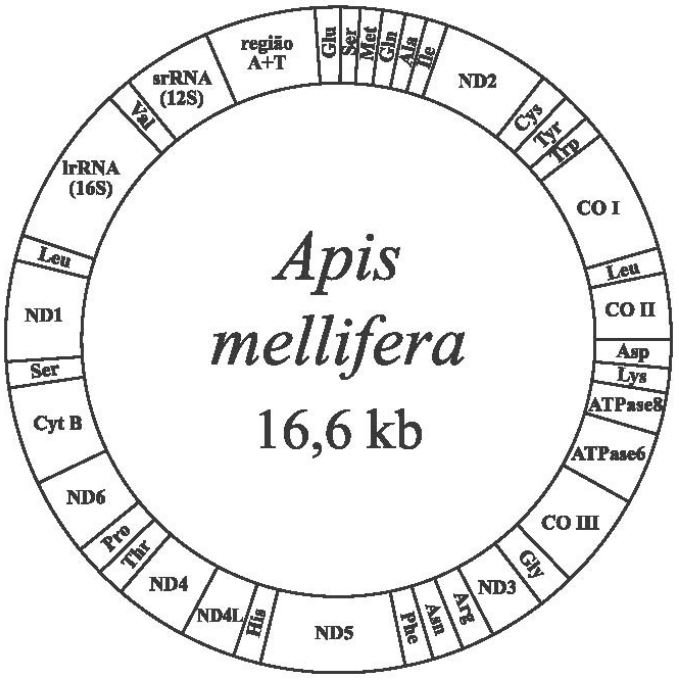


Figura 18.5. Genoma mitocondrial de *Apis mellifera* (adaptado de Crozier e Crozier, 1993).

dades da desidrogenase da NADH e a oxidase do citocromo e dos RNA transportadores. No entanto, os genes codificadores para as subunidades ribossômicas e para o citocromo *b* estão entre os mais conservados entre os organismos. As regiões A+T ou "D-loop" são as que mais acumulam mutações, podendo ser do tipo substituição de base ou inserção/deleção, esse último muito comum para essa região, levando a alterações no tamanho total dessa molécula.

Devido às características descritas acima, a partir da década de 1980, o mtDNA tem sido amplamente empregado em estudos genéticos e evolutivos, utilizando-se a técnica de RFLP, principalmente em comparações intra e interespecíficas. Os níves de variação populacional detectados com essa molécula são maiores do que aqueles obtidos com o uso de isozimas. Outra vantagem é que, com relação ao DNA nuclear, a análise do mtDNA é muito mais simples, pois não envolve recombinação gênica. A análise por RFLP do mtDNA é, portanto, muito sensível e tem fornecido informações relevantes em estudos de estrutura populacional de espécies, biogeografia, dispersão, fluxo gênico e colonização.

Por ser uma molécula circular, quando analisada por enzimas de restrição, o número de bandas geradas corresponde ao número de sítios de restrição reconhecidos pela enzima empregada. Em estudos populacionais ou de caracterização inicial dessa molécula, com a finalidade de estabelecer os sítios polimórficos, normalmente são utilizadas de dez a 20 enzimas de restrição, que reconhecem de cinco a seis bases como seus sítios de clivagem. Dessa forma, temos que um total de 250 a 600 pb são analisados por indivíduo. Em estudos populacionais, o número amostral geralmente é alto, chegando a ser por volta de várias centenas de indivíduos. A variação encontrada entre indivíduos, geralmente resulta de mutações pontuais que alteram a sequência reconhecida pela enzima de restrição, causando o aparecimento ou perda de sítios. Os padrões individuais gerados por uma dada enzima de restrição constituem-se nos haplótipos, geralmente designados por letras. Podemos ter, então, os haplótipos A, B, C etc. para a enzima *Eco* RI, com um, dois e três cortes, respectivamente, por exemplo. Quando se analisam indivíduos com diferentes enzimas de restrição e se determina o número de sítios de clivagem, podemos construir o haplótipo composto, isto é, a combinação dos haplótipos individuais por enzima, gerando combinações do tipo AACDB, a qual seria um haplótipo para um indivíduo resultante da análise com cinco enzimas de restrição, cada letra correspondendo ao haplótipo de cada enzima isoladamente.

Os haplótipos determinados por enzima podem ser analisados quanto ao padrão de interconversão entre eles (Fig. 18.6). Por probabilidade, é mais fácil se perderem sítios de restrição que ganhá-los, pois uma única mutação de ponto é suficiente para alterar a sequência de bases reconhecida por uma enzima. Portanto, os haplótipos que apresentam um maior número de sítios para uma dada enzima seriam considerados como os mais plesiomórficos e teriam dado origem aos outros haplótipos pela perda de sítios. No entanto, é difícil poder afirmar qual foi o curso da evolução sem análises adicionais.

Quando se trabalha com diferentes espécies e se pretende verificar as relações evolutivas entre elas, podemos estabelecer uma matriz de presença e ausência de sítios e depois submetê-la a análises filogenéticas. Alguns trabalhos de caráter filogenético apresentam matrizes de ausência ou presença de fragmentos, mas é preciso tomar certo cuidado com esse tipo de análise, pois fragmentos podem comigrar e fragmentos menores originar-se de um único maior. Desse modo, podemos subestimar as características compartilhadas entre os organismos em estudo. É, portanto recomendável que se construam matrizes baseadas em sítios que podem ou não ser compartilhados. Como obter tais resultados? Como já apresentado em detalhe anteriormente, quando se constrói um mapa de restrição para o genoma mitocondrial de um organismo ou indivíduo, estamos posicionando os sítios uns em relação aos outros. Na comparação entre indivíduos ou espécies, quando seus mapas estão determinados, fica fácil e claro detectar os sítios comuns e os únicos e, a partir dessa análise, pode-se construir uma matriz (presença ou ausência de sítios) e submetê-la às análises filogenéticas.

18.7. RFLP do DNA de Cloroplasto

O DNA do cloroplasto é dezenas de vezes maior que o das mitocôndrias, variando de 120 a 217 Kb, sendo que a maioria dos cpDNA possui tamanho ao redor de 140-160 Kb. As alterações de tamanho nesse genoma são devidas a dois fatores principais: quantidade de sequências repetidas e alterações na complexidade da sequência (expansão ou contração de duplicações invertidas que muitas vezes envolvem os genes para as subunidades ribossômicas). Apesar dessa variação em tamanho, a ordem dos genes tem se mostrado mais conservada que em relação ao genoma mitocondrial de plantas. A taxa

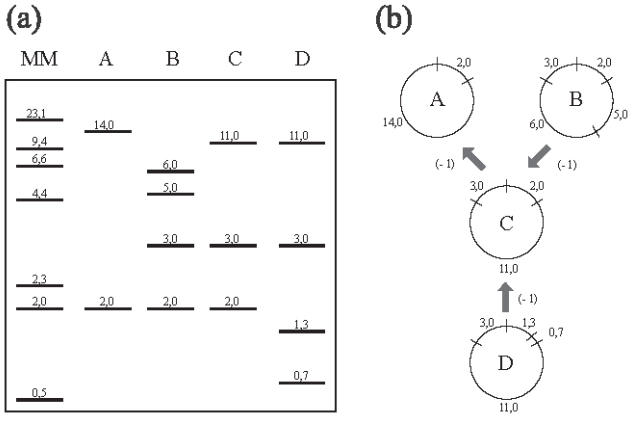


Figura 18.6. (a) Padrão de bandas para uma determinada enzima visualizado em gel. Cada padrão (A-D) constitui um haplótipo distinto. (b) Interconversão entre os haplótipos. As setas indicam o sentido de perda de sítio de restrição. MM: marcador molecular λ/*Hind* III. Os números indicam os tamanhos dos fragmentos em Kb.

de substituição de nucleotídeos é menor que a observada no genoma nuclear animal e de plantas e, portanto, menor ainda em comparação com aquela do genoma mitocondrial animal (veja o Capítulo 10). Uma das características mais intrigantes dessa organela é seu modo de herança. Enquanto o genoma mitocondrial, com raras exceções, é herdado através da via materna, o cloroplasto pode ser herdado também biparentalmente e paternalmente em várias angiospermas estudadas. Quando herdado biparentalmente, sua transmissão parece ser do tipo clonal, já que nenhuma evidência foi encontrada de recombinação entre as moléculas de diferentes cloroplastos. No entanto, alguns estudos têm mostrado que ocasionalmente ocorre transmissão de sequências de cpDNA para o núcleo; dados da literatura também apóiam que no passado tenham ocorrido trocas entre os genomas nuclear, mitocondrial e do cloroplasto.

Apesar de ser uma molécula com características muito peculiares, o fato de ser muito grande e apresentar baixa taxa de substituição de bases parece ter restringido o seu uso pelos pesquisadores a estudos filogenéticos e evolutivos. Estudos do tipo RFLP têm sido conduzidos apenas para algumas regiões do genoma e os resultados indicaram alguma variação intraespecífica.

18.8. RFLP do DNA Nuclear

Embora o DNA mitocondrial tenha sido a molécula mais extensivamente utilizada com a técnica de RFLP, sequências no genoma nuclear fornecem uma fonte rica de marcadores genéticos para análises de variação intraespecífica e interespecífica. Vários genes nucleares têm sido utilizados para inferências filogenéticas e mesmo para diagnóstico e prevenção de algumas doenças.

Entre as sequências nucleares estudadas através de RFLP, destacam-se as sequências de cópia simples scnDNA, ("*single copy nuclear DNA*", em inglês), conhecidas ou anônimas, isto é, sem um produto gênico identificado e o estudo de sequências repetidas. A análise de sequências de cópia simples do DNA nuclear tem sido realizada tradicionalmente utilizando-se o procedimento de "Southern blot*"*. Entre as sequências analisadas, têm sido utilizados genes de função conhecida e, nos diversos organismos, esse tipo de estudo tem revelado certo grau de polimorfismo e fornecido importantes informações sobre a história evolutiva e a diversidade nas populações, informações sobre os efeitos da seleção natural e da deriva genética, além de aspectos importantes da evolução molecular. Entretanto, regiões não codificantes dos genes—os íntrons—também estão sendo analisadas por RFLP e PCR devido a serem mais variáveis que as regiões codificantes.

Regiões anônimas do DNA nuclear de poucas cópias no genoma têm sido examinadas por RFLP, revelando padrões de polimorfismo intraespecífico. Frequentemente, a análise de DNA de cópia simples tem sido utilizada para se estabelecerem padrões de relações entre populações. Normalmente, populações diferentes são examinadas para verificar a presença ou ausência de determinado fragmento ou para verificar a frequência de determinado "alelo", para posteriormente serem realizadas análises filogenéticas.

Para a análise de sequências repetidas tradicionalmente, têm sido utilizados genes ribossômicos, ou famílias gênicas, que no núcleo da célula eucariótica existem como uma série de elementos repetidos. No DNA ribossômico, cada unidade de repetição possui uma sequência altamente conservada. Essas unidades de repetição podem existir em um ou vários sítios dos cromossomos. As regiões espaçadoras não transcritas dessas unidades de repetição têm revelado variação intraespecífica quando analisadas via RFLP, frequentemente devido à variação de tamanho (veja o Capítulo 8).

18.9. Análise por PCR - RFLP

A técnica de PCR, apresentada no Capítulo 19, tem sido extremamente utilizada para a obtenção de fragmentos de genoma, em especial do mitocondrial e de cloroplasto, para posterior digestão com as enzimas de restrição. Essa metodologia é frequentemente conhecida com PCR-RFLP. O sucesso na obtenção desses resultados depende inicialmente da amplificação da região pretendida e, posteriormente, da existência de sítios de restrição. A técnica de PCR tem uma limitação com relação à amplificação de fragmentos grandes e, portanto, quanto menor o fragmento menor a chance de existência de um número expressivo de sítios de corte para as enzimas de restrição.

Os fragmentos de 600 a 2.000 pb geralmente, são amplificados com facilidade. Nesse caso, recomenda-se o uso de enzimas que reconheçam sequências de quatro a cinco bases como seus pontos de clivagem, pois são mais prováveis de existirem nesses pequenos fragmentos. Desse modo, torna-se mais plausível a detecção de polimorfimos. A Figura 18.7 ilustra padrões de RFLP de um fragmento mitocondrial amplificado de *Apis mellifera*. Esse fragmento corresponde a uma região intergênica presente entre os genes COI e COII. Essa região, quando digerida com a enzima *Dra*I, gera padrões de RFLP de forma distinta. Esses padrões estão intimamente relacionados com a distribuição geográfica e com as linhagens evolutivas que deram origem ao grande número

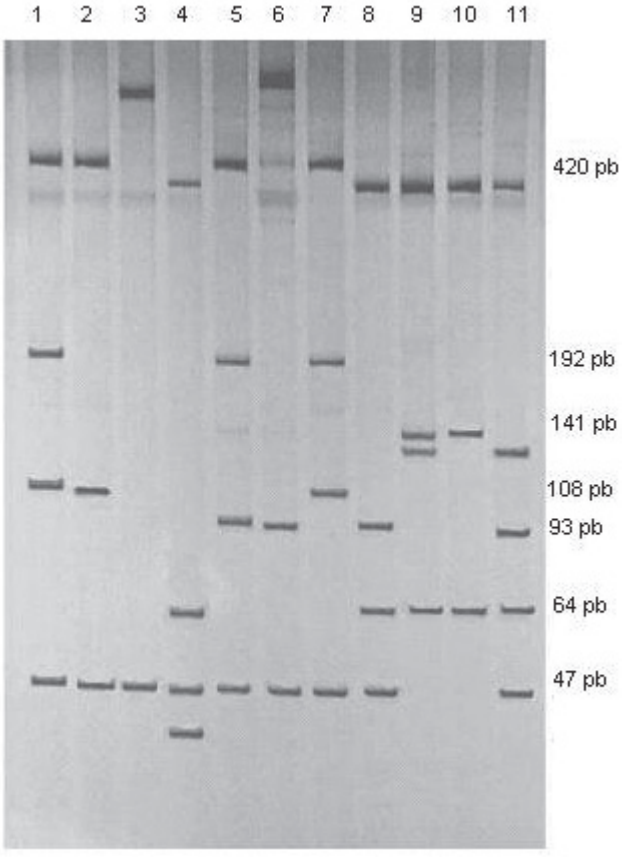


Figura 18.7. Padrões de RFLP verificados para a região intergênica COI-COII de *Apis mellifera*. O fragmento foi amplificado por PCR e digerido pela enzima de restrição *Dra*I. Cada raia corresponde a um padrão de RFLP, mostrando a alto grau de polimorfismo dessa região. Estão indicados os tamanhos de apenas alguns fragmentos (adaptada de Collet *et al.*, 2006).

de subespécies de *Apis mellifera* presentes na Europa e África (Garnery *et al.*, 1993; Collet *et al.*, 2006).

Utilizando essa metodologia, é possível amplificar várias regiões do genoma ou, em alguns casos, todo o genoma de organelas em uma série de fragmentos que se sobrepõem. Um mapa de restrição pode ser obtido e, como já descrito (seção 18.6), esses dados podem ser utilizados em análises intra e interespecíficas.

Uma extensa e interessante revisão dos tópicos comentados brevemente neste capítulo podem ser obtidas nos livros de Hillis *et al.* (1996) e de Avise (1994).

Agradecimentos

As autoras agradecem a valiosa colaboração de Daniela Silvestre Alves e Flávio de Oliveira Francisco pelo auxílio na confecção de figuras da primeira versão desse capítulo, Maria Regina de Siqueira Bueno Bruno e Daniel Gouw, pelo trabalho gráfico final.

Referências Bibliográficas

Ashkenas, J. (1997). Homologous recombination in human mitochondria?. **Am. J. Hum. Genet., 61**: 18.

Ausubel, F.M., Brent, R., Kinstone, R.E., Moore, D.D., Seidman, J.G., Smith, J.A. e Struhl, K.(Ed.). (1993). **Current Protocols in Molecular Biology**. Greene Publishing Associates, Inc. e John Willey e Sons, Inc. Harvard, Massachusetts, USA.

Avise, J.C. (1994). **Molecular Markers, Natural History and Evolution**. Chapman e Hall, New York, NY.

Collet, T., Ferreira, K.M., Arias, M.C. Soares, A.E.E. e Del Lama, M.A. (2006). Genetic structure of Africanized honeybee populations (*Apis mellifera* L.) from Brazil and Uruguay viewed through mitochondrial DNA COI-COII patterns. **Heredity** 97:329-335.

Crozier, R.H. e Crozier, Y.C. (1993). The mitochondrial genome of the honeybee *Apis mellifera*: complete sequence and the genome organization. **Genetics 133**: 97-117.

Francisco, F.O., Silvestre, D. e Arias, M.C. (2001). Mitochondrial DNA characterization of the five species of *Plebeia* (Apidae: Meliponini) : RFLP and restriction maps. **Apidologie** 32:323-332.

Garnery, L., Solignac, M., Celebrano, G. e Cornuet, J.M. (1993). A simple test using restricted PCR-amplified mitochondrial DNA to study the genetic structure of *Apis mellifera* L. **Experientia** 49:1016-1020.

Hillis, D.M., Moritz, C. e Mable, B.K. (Ed.). (1996). **Molecular Systematics**. Sinauer Associates, Inc. Sunderland, Massachusetts, USA.

Howell, N. (1997). mtDNA recombination: What do in vitro data mean? **Am. J. Hum. Genet., 61**: 19-22.

Hubby, J.L. e Lewontin, R.C. (1966). A molecular approach to the study of genic heterozygosity in natural populations. I. The number of alleles at different loci in *Drosophila pseudoobscura*. **Genetics 54**: 577-594.

Hunter, R.L. e Market, C.L. (1957). Histochemical demonstration of enzymes separated by zone electrophoresis in starch gels. **Science 125**: 1294-1295.

Lewontin, R.C. e Hubby, J.L. (1966). A molecular approach to the study of genic heterozygosity in natural populations. II Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. **Genetics 54**: 595-609.

Linn, S. e Arber, W. (1968). Host specificity of DNA produced by *Escherichia coli*. X. In vitro restriction of phage fd replicative form. **Proc. Natl. Acad. Sci. USA. 59**:1300-1306.

Meselson, M. e Yuan, R. (1968). DNA restriction enzyme from *E. coli*. **Nature 217**: 1110.

Sambrook, J; Fritsch, E.F. e Maniatis, T. (1989). **Molecular Cloning**. Cold Spring Harbor. Lab. Press, Cold Spring Harbor, New York.

Smithies, O. (1955). Zone electrophoresis in starch gels, group variations in the serum proteins of normal individuals. **Biochem. J. 61**: 629-641.

Southern, M. (1975). Detection of specific squences among DNA fragments separated by gel electrophoresis. **J. Mol. Biol. 98**: 503.

Zouros, E; Freeman, K.R; Ball A.O. e Pogson, G.H. (1992). Direct evidence for extensive paternal mitochondrial DNA inheritance in the marine mussel *Mytilus*. **Nature 359**: 412-414.

Página deixada em branco

Capítulo 19

Métodos baseados em PCR para análise de polimorfismos de ácidos nucléicos

Sergio Russo Matioli (srmatiol@ib.usp.br)
Departamento de Genética e Biologia Evolutiva
Instituto de Biociências
Universidade de São Paulo

Maria Rita dos Santos e Passos-Bueno (passos@ib.usp.br)
Departamento de Genética e Biologia Evolutiva
Instituto de Biociências
Universidade de São Paulo

"*E agora enleada na tênue cadeia*
Debalde minh'alma se embate, se irrita...
O braço, que rompe cadeias de ferro,
Não quebra teus elos,
Ó laço de fita!"
(Castro Alves, *Espumas Flutuantes*)

19.1. O Impacto do Método de PCR em Pesquisas Genéticas

O desenvolvimento da técnica da reação em cadeia da polimerase (PCR, do inglês, *polymerase chain reaction*) aumentou muito a eficiência de detecção de polimorfismos no nível do DNA ou RNA, traduzida em redução do tempo de execução dos experimentos, de seu custo e de sua complexidade. A concepção inicial da técnica foi de Kary Mullis, um bioquímico que trabalhava com síntese química de oligonucleotídeos em uma firma biotecnológica da Califórnia, a Cetus Corporation. O mais impressionante do desenvolvimento dessa técnica, em 1985, é que todos os elementos necessários para a sua execução já estavam disponíveis havia muito tempo dentro da área de Biologia Molecular. Segundo o historiador da Ciência Paul Rabinow (1996), foi possível para Kary Mullis desenvolver essa técnica, pois, além de possuir os conhecimentos necessários na área de Biologia Molecular, também programava computadores e, portanto, a idéia de ciclos que se repetem muitas vezes (tais como nos programas de verificação de números primos ou de cálculo de fatoriais) era, para ele, bem familiar.

19.2. Apresentação da Técnica de PCR

Um gene de cópia única em um genoma complexo, como o humano, representa um segmento de DNA muito raro. Sua detecção envolve o emprego de técnicas com sensibilidade muito grande, da ordem de partes por milhão. Antes do advento da PCR, era possível a detecção de polimorfismos por sondas marcadas radioativa ou quimicamente, as quais eram hibridadas em membranas contendo o DNA-alvo (Capítulo 18). Apenas dessa maneira obtinha-se sinal suficiente para detectarem-se essas sequências polimórficas. Essa metodologia necessita que, além do material radioativo, a quantidade de DNA-alvo utilizada seja muito grande. Essas dificuldades foram resolvidas com a introdução da técnica de PCR, que possibilita a obtenção de quantidades muito grandes de fragmentos específicos de DNA através de amplificação em ciclos. Teoricamente, cada ciclo de síntese duplica a quantidade de DNA-alvo. Assim, em 10 ciclos haveria 1024 vezes mais DNA alvo do que na amostra original ($1024 = 2^{10}$); em 20 ciclos, cerca de 1 milhão de vezes mais e, em 30 ciclos, mais de um bilhão de vezes mais! Essa é uma estimativa teórica, pois, na prática, não há eficiência de 100% na reação e cada ciclo é estequiometricamente diferente do anterior, pois há alteração na quantidade de substratos e produtos (Velikanov e Kapral, 1999). Na Figura 19.1, um dos ciclos está esquematizado. Na Figura 19.2, os três primeiros ciclos de uma PCR mostram a natureza exponencial do aumento do número de moléculas específicas.

Fatores que influenciam a PCR

Atualmente, a técnica de PCR é largamente difundida, sendo utilizada tanto em laboratórios acadêmicos como em privados, e em áreas bastante distintas. Isso é possível pois existem conjuntos-padrão (os chamados *kits*) vendidos por fabricantes de produtos acompanhados de protocolos muito fáceis de serem utilizados. Entretanto, para que essa técnica não seja utilizada apenas como uma "caixa-preta" e também para auxiliar na resolução de possíveis problemas, descreveremos cada um dos componentes da reação e sua função, bem como a justificativa das quantidades, volumes e concentrações dos reagentes empregados normalmente.

Uma PCR tipicamente realiza-se em volumes pequenos, de 10 a 100 microlitros. Atualmente esse volume pode ser muito menor, por exemplo em microcâmaras com volumes da ordem de dezenas de picolitros ou nas amplificações feitas em emulsões, como aquelas empregadas no processo de pirossequenciamento, descrito mais adiante. A quantidade de DNA que deverá ser adicionada depende fundamentalmente da quantidade estimada de DNA-alvo que exista na amostra. Teoricamente, basta uma única molécula de DNA-alvo na reação para que essa possa ser amplificada, mas, na prática, resultados melhores são obtidos

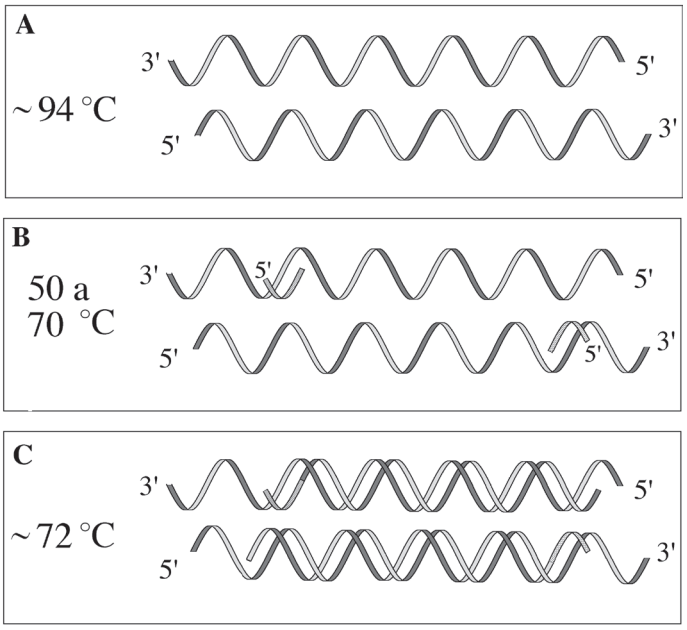


Figura 19.1. O primeiro ciclo de uma PCR. A. Na temperatura alta, as cadeias da fita dupla de DNA desnaturam. B. Quando a temperatura fica inferior à **Tm** dos *primers*, eles hibridizam nas regiões complementares a eles. C. A fita nova de DNA é elongada a partir da extremidade 3' dos *primers*.

com 10^3 a 10^4 moléculas originais. O fator mais importante é a complexidade do genoma a ser estudado. Por exemplo, em 1 ng de DNA de *Escherichia coli*, existem aproximadamente 240.000 genomas, enquanto que na mesma quantidade de DNA humano existem cerca de 300 genomas apenas. A qualidade do DNA também é importante, pois, se esse estiver degradado, apenas uma parcela das moléculas-alvo será amplificada (caso de DNA antigo, extraído, por exemplo, de organismos preservados em condições não ideais por muito tempo). O DNA para PCR pode ser extraído de maneira simplificada, pois os contaminantes (que podem influenciar a reação) acabam sendo bastante diluídos. Existem protocolos de extração que não passam de simples fervura, que serve apenas para inativar, por aquecimento, as enzimas DNAlíticas, como aquele descrito por Valsecchi (1998). O tamanho do fragmento amplificado também é importante para o sucesso da reação. Fragmentos maiores que 3 a 4 mil pares de bases podem apresentar problemas. A solução, nesses casos, é uma PCR com o emprego de mistura de polimerases (normalmente conhecida como *long PCR*; Cheng *et al.*, 1994), sendo que uma fração pequena delas tem a propriedade de "revisão" do processo de extensão (ou elongação), ou seja, essas polimerases têm a capacidade de remover nucleotídeos mal incorporados, os quais interrompem a síntese pela polimerase *Taq* do DNA.

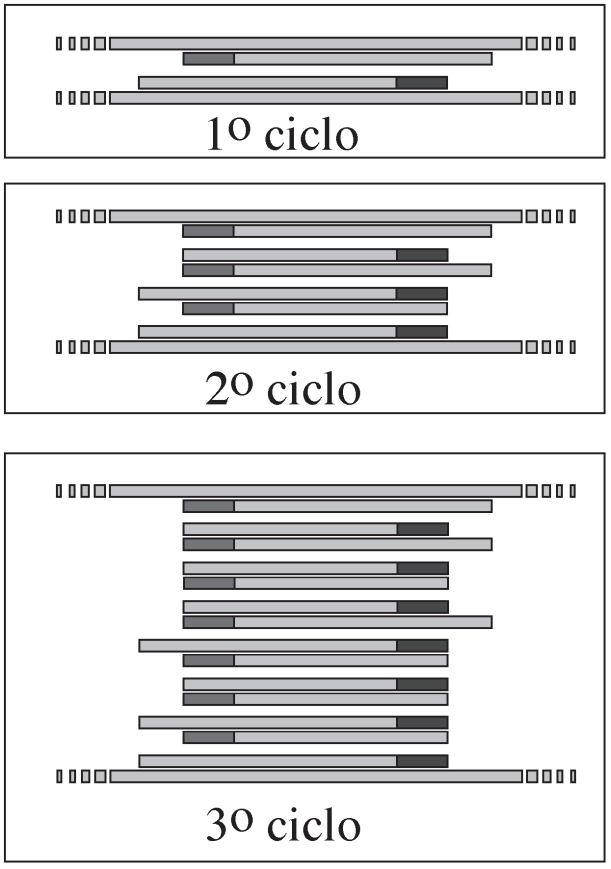


Figura 19.2. Os três primeiros ciclos de uma PCR mostrando a natureza exponencial da amplificação. Embora haja extensão das cadeias de DNA que "ultrapassam" as regiões complementares aos *primers* no primeiro ciclo, a amplificação restringe-se ao tamanho relativo à distância das extremidades 5' correspondentes às regiões de hibridização.

Outros componentes de uma PCR são os trifosfatos de desoxirribonucleotídeos, que, idealmente, devem existir na mesma proporção que aquela presente na molécula que será sintetizada. Na prática, utilizam-se proporções de dATP, dCTP, dGTP e dTTP equimolares; outras proporções diferentes podem ser empregadas quando se sabe de antemão que há desvios muito grandes dessas proporções. Tipicamente, a quantidade de cada um dos nucleotídeos varia entre 50 μM a 200 μM (sendo esse o máximo possível, na prática, por motivos de solubilidade). Uma quantidade correspondente a uma concentração de 50 μM de cada um dos nucleotídeos em um volume de 25 microlitros é suficiente para a síntese de cerca de 3 microgramas de DNA. Essa é uma quantidade apreciável. Como referência, uma banda de produto de PCR totalizando 50 nanogramas em um gel corado por brometo de etídeo pode ser visualizada. A concentração de sais (normalmente sob a forma de cloreto de potássio ou de sódio a 50 mM) é um fator crítico que deve ser levado em conta quando existem sais de cátions monovalentes na amostra de DNA (como o cloreto de sódio, por exemplo). Outro componente importante é a quantidade de cloreto de magnésio necessária para a síntese (Innis e Gelfand, 1990). Essa quantidade fica em torno de 2,5 mM, mas em amostras com muito EDTA ou outro agente que diminui a quantidade efetiva de cátions divalentes, essa concentração pode ser aumentada (uma única molécula de EDTA sequestra 4 cátions magnésio). O pH ideal de uma PCR é de 7,5, pois a enzima Polimerase *Taq* do DNA tem seu ótimo nessa faixa. Entretanto, os tampões utilizados têm valor mais alto de pH em temperatura ambiente (8,5 a 9,0), pois na temperatura ótima para a enzima, 72 ^{o}C, o pH diminui e chega naquele que é o ótimo para a enzima. A taxa de erros de incorporação também é dependente do pH, conforme a enzima empregada (Cline *et al.*, 1996). Certos protocolos empregam uma proteína na solução (albumina bovina, isenta de nucleases, ou gelatina) que adere aos possíveis sítios compatíveis que existam nas paredes do tubo, de tal forma que a polimerase não o faça e deixe de atuar na reação. Existem atualmente polimerases termo-resistentes que somente se tornam ativas depois que a temperatura atinge um valor alto (primeiro ciclo de desnaturação). Essas enzimas são conhecidas como "hot start" (início quente) e servem para evitar que a polimerização de produtos inespecíficos comece enquanto se prepara a solução de amplificação ou se aguarda para se colocar as placas ou tubos no termociclador.

A concentração de oligonucleotídeos (*primers*) gira em torno de 1 micromolar para cada um deles. Para um volume de 25 microlitros, isso significa a existência de um potencial de síntese de 1,5 x 10^{13} novas cadeias, ou seja, estão sempre presentes em excesso.

As temperaturas de cada ciclo correspondem à fase de desnaturação (onde as fitas são separadas), de hibridação com os oligonucleotídeos e da extensão da cadeia que será sintetizada. A temperatura de desnaturação (cerca de 94 ^{O}C) é suficientemente alta para permitir a separação das fitas de DNA e suficientemente baixa para não provocar degradação da enzima. Raramente se sugere alteração nessa temperatura, normalmente recomendada pelo fabricante. A duração dessa fase depende muito do termociclador e dos volumes e tubos empregados, pois há diferenças da condutividade térmica entre os sistemas. Não há necessidade de que essa fase tenha mais que 1 minuto, a não ser em situações excepcionais, como a amplificação em volumes elevados (maiores que 50 microlitros) ou de regiões ricas em GC.

A temperatura de hibridação do DNA-alvo com o oligonucleotídeo é a mais crítica da PCR. Essa depende da temperatura de fusão da molécula, que é a temperatura em que 50% das moléculas estão hibridadas, em um equilíbrio dinâmico. Essa temperatura é conhecida por **Tm** e é calculada, pelo fabricante ou por programas de computadores, a partir da sequência de bases e do tamanho do oligonucleotídeo. Oligonucleotídeos em geral têm 18-28 nucleotídeos com 50-60% de G+C. Uma regra geral para calcular a **Tm** é considerar 2 ^{O}C para A ou T e 4 ^{O}C para G ou C. A **Tm** deve ser semelhante para os dois oligonucleotídeos a serem utilizados em uma reação de PCR e a temperatura de hibridação deve ser um pouco inferior (de 0 a 10 ^{O}C) do que a **Tm** mais baixa dos oligonucleotídeos empregados. Pode haver situações, no entanto, em que se empregam temperaturas mais altas que a própria **Tm**, quando há necessidade de muita especificidade, em detrimento da eficiência. A duração dessa fase também depende do conjunto equipamento-tubo-volume e raramente deve ser maior que 1 minuto.

A temperatura de extensão também depende do ótimo da enzima empregada (no caso da polimerase *Taq* do DNA, é de 72 ^{O}C). A duração, nesse caso, depende também do tamanho do fragmento que será amplificado. É bom observar que a taxa de síntese da polimerase *Taq* do DNA a 72 ^{O}C chega a 150 nucleotídeos por segundo por molécula!

PCRs que produzem arrastos quando visualizados em gel após eletroforese (fragmentos com vários tamanhos) resultam de falta de especificidade da reação. Nesse caso, a solução pode ser (em ordem de fatores mais prováveis de resultar em sucesso):

1. Aumento da temperatura de hibridação, que pode ser mais alta nos primeiros ciclos e menor nos subsequentes. A essa prática, dá-se o nome de *touchdown* (Don *et al.*, 1991), aterrissagem em inglês, em uma analogia com a diminuição paulatina de altura. A razão disso é que o aumento da temperatura em todos os ciclos compromete a eficiência da reação quando há escassez relativa de substratos, o que acontece nos últimos ciclos.
2. Diminuição de DNA-alvo na amostra.
3. Diminuição na concentração de cloreto de magnésio.
4. Diminuição na concentração dos oligonucleotídeos.
5. Diminuição da concentração de dNTPs.

PCRs que "falham", ou seja, não apresentam amplificação alguma, também podem ser resultado de falta de especificidade (na verdade, há um arrasto, que corresponde a uma enormidade de fragmentos diferentes que são amplificados, mas que não são visualizados). Nesse caso, as soluções são as mesmas que aquelas apresentadas acima. Se houver excesso de especificidade, não há a hibridação dos *primers* com as cadeias alvo e não há, portanto, amplificação alguma. Nesse caso, as soluções são inversas com relação àquelas acima, na mesma ordem. A melhor estratégia para a otimização de uma PCR é a verificação da melhor temperatura de hibridação, mantendo-se constante os demais fatores, em geral empregando-se aqueles recomendados pelo fabricante dos reagentes e equipamentos. Se nenhuma temperatura de hibridação apresentar resultados satisfatórios, tentam-se outras mudanças. É importante que se façam sempre controles positivos (uma amostra de DNA que se sabe conter o fragmento alvo) e negativos (sem DNA) em cada experimento de otimização. Uma amplificação que ocorre na ausência de DNA adicionado pode ser resultado de contaminação de um dos reagentes, o que inclui a própria polimerase, normalmente purificada de uma bactéria recombinante. Descobrir a fonte de contaminação é um processo trabalhoso, caro e demorado, de modo que todos os cuidados devem ser tomados na prevenção da contaminação.

19.3. A Combinação de PCR com RFLP

A primeira aplicação de PCR na detecção de polimorfismos foi a combinação com a RFLP (Capítulo 18), apresentada inicialmente no mesmo trabalho em que a técnica de PCR foi descrita (Saiki *et al.*, 1985). Nesse caso, a etapa de amplificação precede a digestão. Um cuidado especial que deve ser tomado quando os fragmentos amplificados são digeridos é a verificação da compatibilidade dos tampões de amplificação e de digestão. As enzimas de digestão têm exigências variáveis no que se refere principalmente à concentração de sais. Quando as exigências são semelhantes, a digestão pode ser feita diretamente a partir da mesma solução onde ocorreu a amplificação. Quando as concentrações são muito diferentes, um ajuste no tampão ou uma etapa de purificação podem ser necessárias previamente à digestão. Com relação à análise dos fragmentos digeridos, toda a estratégia pode ser a mesma que aquela descrita no Capítulo 18, sendo que a obtenção dos mapas de sítios para enzima de restrição é fundamental para estudos filogenéticos ou de polimorfismos. A grande diferença é que, por causa da limitação prática de amplificação de fragmentos muito grandes, é possível apenas a análise de sítios em fragmentos pequenos, da ordem de alguns milhares de pares de bases. Nesse caso, escolhem-se enzimas de restrição com sítios de reconhecimento de 4 ou 5 pb, que são mais prováveis de existir em tais fragmentos.

19.4. Uso de *Primers* Flanqueadores de Regiões Repetitivas (Microssatélites)

Essa técnica, até recentemente, era a mais difundida na análise de polimorfismos. O termo "microssatélite", em uma primeira análise, parece referir-se a um objeto natural ou artificial que orbita um astro. A origem do termo no que se refere ao DNA realmente tem algo a ver com o termo astronômico. A técnica de ultracentrifugação em equilíbrio com cloreto de césio permite que se possa separar macromoléculas através de diferenças em densidade, pois uma solução desse sal, quando submetida a forças centrífugas muito grandes, tende a formar um gradiente de densidade e as macromoléculas concentram-se na região que tem a mesma densidade ao longo do gradiente formado. Quando amostras de DNA total são analisadas através dessa técnica, é normal que se observe, acima e abaixo da maior parte do DNA, frações que consistem em bandas definidas. O termo "DNA satélite" surgiu pois essas frações visualmente pareciam com órbitas em torno de um planeta (com um pouco de imaginação, evidentemente). Esse DNA revelou constituir-se, segundo estudos posteriores, em sequências repetitivas. Desde a década de 1960, com o emprego de experimentos de dissociação-reassociação de DNA, sabia-se

da existência de DNA altamente repetitivo, moderadamente repetitivo e de cópias únicas. Com a obtenção maciça de sequências de DNA, convencionou-se chamar de DNA satélite aquele que tem trechos longos repetidos, de DNA minissatélite aquele com sequências médias repetidas e de DNA microssatélite aquele com sequências bem curtas repetidas. Não existe, na literatura, um consenso exato sobre quais seriam os limites entre as classes, mas, de qualquer forma, são limites arbitrários.

No caso dos microssatélites, a natureza da variação é o número de repetições que existem das unidades repetidas (Tautz, 1989 e Weber e May, 1989, inicialmente demonstraram a utilidade desses marcadores na caracterização de polimorfismos). A estratégia para a análise é a amplificação do fragmento que contém as repetições, utilizando-se, como *primers*, oligonucleotídeos que hibridam com as regiões que flanqueiam as repetições. Essa técnica também é conhecida como STR (*short tandem repeats*, em inglês) ou SSR (*simple sequence repeats*).

O fator limitante para o emprego da técnica de microssatélites é a disponibilidade de informação (em termos de sequências de nucleotídeos) a respeito de regiões que possuam repetições para o organismo que se deseja estudar. No caso em que não esteja disponível o sequenciamento da região e/ou do genoma do organismo de interesse, para a obtenção de informações que permitam o emprego de *primers* correspondentes às regiões que flanqueiam os microssatélites, é necessária a construção de uma biblioteca genômica do organismo em questão. Essa biblioteca pode ser restrita a um determinado tamanho de fragmento clonado, por exemplo, de 200 a 600 pares de bases. Essa faixa de tamanho normalmente é escolhida porque é possível, através de eletroforese, a detecção de variação que consiste de diferença de dois a cinco nucleotídeos. Nessa faixa de tamanho de fragmentos, o processo de sequenciamento não apresenta dificuldades.

Após a construção da biblioteca genômica, é necessário que se encontrem, dentre os clones, aqueles que possuam sequências repetidas. Isso pode ser conseguido através de duas estratégias: na mais tradicional, hibrida-se o DNA dos clones imobilizados em uma membrana com sondas marcadas que consistem de oliogonucleotídeos sintetizados que se constituem em repetições. Por exemplo, se se quer detectar repetições ACACACAC…, os oligonucleotídeos 5´-$(ACAC)_4$ ou 5´- $(GTGT)_4$ podem ser utilizados, pois, no clone que está sendo testado, há ambas as fitas complementares do DNA. A mesma biblioteca pode ser testada para vários tipos de repetições. Uma vez que os clones que contenham as sequências repetitivas tenham sido detectados, é necessário que se obtenha a sequência das regiões que flanqueiam as repetições e que também se confirme a existência da repetição. Isso confirmado, a etapa seguinte é o desenho dos *primers* que deverão ser empregados na análise de genótipos.

Com a publicação de grande parte da sequência do cromossomo 22 humano, com 14 milhões de pares de bases, verificou-se a existência de cerca de 130.000 repetições de dinucleotídeos, 18.000 de trinucleotídeos, 47.000 de tetranucleotídeos e apenas 1.600 de pentanucleotídeos (Dunham *et al.*, 1999).

Embora os dinucleotídeos sejam os mais abundantes em repetições, a análise de genótipos com polimorfismos dessas regiões é um pouco mais difícil, pois a técnica a ser empregada deve ser sensível o suficiente para a detecção de diferenças de apenas dois nucleotídeos. Isso normalmente é possível apenas com o emprego de géis denaturantes. Diferenças como aquelas observadas nos polimorfismos de regiões de tetranucleotídeos repetidos podem ser detectadas com gel de agarose, muito mais simples de se lidar.

Atualmente existem outras estratégias que permitem a obtenção de dados de sequências de forma mais eficiente e rápida do que a descrita acima. Uma dessas estratégias é aquela em que há um enriquecimento prévio do DNA digerido, que será clonado com fragmentos que contenham as sequências repetitivas através do emprego de oligonucleotídeos sintéticos associados a partículas paramagnéticas (Refseth *et al.*, 1997). Utilizam-se oligonucleotídeos com regiões repetitivas (por exemplo, 5'-AGCAGCAGCAGCAGC-3') que sejam ligados quimicamente à biotina. Como a hibridação com esses oligonucleotídeos necessita que o DNA esteja em fita simples, é necessário um truque para que esse possa retornar ao estado de fita dupla. Esse truque é o emprego de "adaptadores" com sequências conhecidas que serão utilizadas para a amplificação por PCR. Os adaptadores são ligados às extremidades coesivas resultantes da digestão por enzimas de restrição com a ligase do DNA. Em seguida, o DNA é denaturado e hibridado com os oligonucleotídeos repetitivos. Partículas paramagnéticas recobertas com estreptavidina (uma proteína com alta afinidade por biotina) são então adicionadas e, em seguida, atraídas por ímãs. As partículas são lavadas e, quando aquecidas, liberam os fragmentos que contenham sequências 5'GCTGCTGCTGCT...3' (no caso do emprego do oligonucleotídeo exemplificado acima). A PCR possibilita tanto o enriquecimento da fração do genoma que tem as repetições desejadas, como a reconstituição da fita dupla, o que possibilita a clonagem posterior. Os clones individuais têm então seus insertos sequenciados e aqueles que possuem tanto as sequências com repetições e as regiões flanqueadoras são utilizados para o desenho dos *primers*.

19.5. *Primers* Arbitrários ou Aleatórios

Muitas vezes, não se tem a disponibilidade, em um laboratório, das técnicas e equipamentos necessários para obter o conhecimento de sequências de nucleotídeos que apresentem polimorfismos que possam ser analisados tanto pela técnica de PCR-RFLP ou de STR (microssatélites). Existem, nesse caso, alternativas que dispensam esse conhecimento para que se possa obter informação a partir de locos polimórficos em pesquisas genéticas. A primeira técnica, conhecida mais popularmente como RAPD, foi desenvolvida simultânea e independentemente por dois grupos de pesquisa (Welsh e McClelland 1990; Williams *et al.*, 1990). O termo RAPD significa "DNA polimórfico amplificado ao acaso" (*random amplified polymorphic DNA*, em inglês). O termo AP-PCR, menos usado, significa "PCR com *primers* arbitrários" (*arbitrarily primed PCR*, em inglês). O princípio da técnica é o seguinte: em primeiro lugar, constroem-se *primers* com sequências arbitrárias, que satisfaçam algum critério com relação a suas propriedades de hibridação. O procedimento mais comum é que tenham dez nucleotídeos e que pelo menos seis deles sejam C ou G. Isso garante que, para uma dada temperatura e concentração salina, haja hibridação caso as sequências dos *primers* sejam perfeitamente complementares às sequências existentes no genoma e que a hibridação seja bem menor, caso as sequências não sejam perfeitamente complementares. Utiliza-se, em geral, um único *primer*, de tal forma que, se houver duas sequências que hibridem com esse *primer*, dispostas de modo palindrômico a uma distância igual ou menor que 2 Kb, haverá amplificação do fragmento interno, tal como é mostrado na Figura 19.3. Caetano-Anolles *et al.* (1991) propuseram ainda a utilização de *primers* ainda menores, na técnica que foi por eles denominada de DAF (*DNA Amplification Fingerprinting*).

O polimorfismo é evidenciado quando há presença ou ausência de bandas após amplificação e eletroforese em gel. Normalmente, o padrão de herança de bandas produzidas por essa técnica é dominante, pois basta a presença de um único genoma que possibilite a amplificação do fragmento, que a banda será visualizada. No caso de indivíduos heterozigotos, a banda

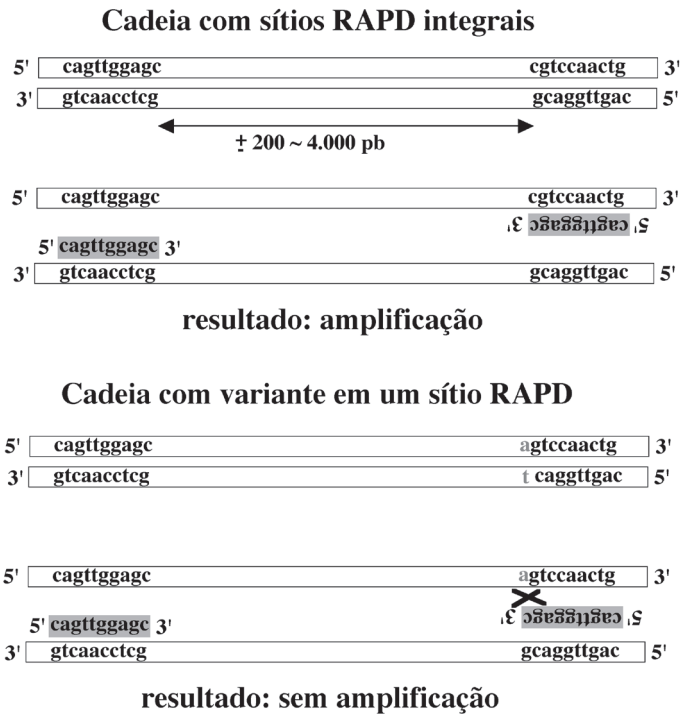


Figura 19.3. Esquema de detecção de polimorfismos através da técnica RAPD. O polimorfismo é detectado na região de complementaridade da cadeia de DNA com o *primer* empregado.

também será observada, impossibilitando a distinção de um dos homozigotos do heterozigoto. Se houver polimorfismos de indels (inserções/deleções) entre os sítios de ligação dos *primers* RAPD, as bandas poderão apresentar um comportamento de herança codominante, pois fragmentos de dois tamanhos diferentes serão evidenciados nos heterozigotos.

As limitações dessa técnica, além do mecanismo de herança dominante, dizem respeito à falta de reprodutibilidade dos resultados, principalmente entre laboratórios diferentes, pois as condições de hibridação para a detecção dos polimorfismos têm que ser discriminantes dentro de um espectro muito curto das condições experimentais. Algumas publicações destacam a produção de artefatos de técnica com o emprego de RAPD, como a existência de somente uma fração de bandas que se comporta mendelianamente (Bucci e Menozzi, 1993; Ellsworth *et al.*, 1993, por exemplo). Para empreender-se um estudo com RAPD, é importante que se disponha de meios para testar a herança dos padrões de bandas.

Na técnica de RAPD, utilizam-se *primers* curtos por uma razão probabilística. Existem 4^{10} sequências diferentes com 10 nucleotídeos, cerca de um milhão delas. Consideremos um genoma de 100.000.000 nucleotídeos. Supondo que a sequência seja completamente aleatória, é esperado que, em média, existam para cada uma das sequências de RAPD, cerca de 100 sequências que sejam idênticas, no genoma, à sequência de RAPD. Caso elas estejam dispostas palindromicamente, há a amplificação por PCR. Se uma sequência aleatória apresentasse 15 nucleotídeos, haveria uma probabilidade de apenas 1/10 de haver, por acaso, uma sequência idêntica em um genoma com 100 milhões de nucleotídeos, pois 4^{15} é aproximadamente 1 bilhão.

19.6. *Primers* Arbitrários Repetitivos

Conforme discutido no item anterior, o grande problema da técnica RAPD é a dificuldade de padronização, pois tem que haver uma discriminação exata entre a existência de um sítio RAPD que hibrida com o *primer* para que os polimorfismos possam ser caracterizados. O aumento do tamanho dos *primers*, que asseguraria uma melhor reprodutibilidade na amplificação, acarretaria uma diminuição na probabilidade de produção de bandas, conforme já mencionado anteriormente. Uma solução engenhosa foi proposta por Gupta *et al.* (1994), com o emprego de também um único *primer*, de tamanho maior, mas que consiste em repetições. Como se sabe da existência de repetições em vários genomas, o problema da diminuição de probabilidade com o aumento de tamanho não se aplica. A essa técnica foi atribuído o nome de SPAR (*Single Primer Amplification Reaction*). Nesse caso, ao contrário do que acontece na análise de locos de microssatélites, o que é amplificado é um fragmento que existe entre duas regiões repetitivas. Essa técnica também é conhecida como ISSR (*Inter Simple Sequence Repeats*).

A vantagem da técnica SPAR ou ISSR sobre a de RAPD é que há grande reprodutibilidade na primeira, havendo resultados semelhantes com variações grandes nas condições de amplificação.

19.7. AFLP

A sigla AFLP significa *amplified fragment length polymorphism*, ou seja, polimorfismo de tamanho de fragmentos amplificados. Essa técnica permite que fragmentos anônimos do genoma sejam amplificados por PCR após terem sido originados pela digestão do genoma por enzimas de restrição. Essa técnica foi descrita inicialmente por Vos *et al.* (1995) e envolve as seguintes etapas (mostradas na Figura 19.4): 1. digestão do genoma por uma enzima de restrição; 2. ligação de "adaptadores" aos fragmentos gerados; e 3. amplificação dos fragmentos a partir de *primers* que hibridam com os adaptadores.

Como as sequências que existem nos adaptadores são escolhidas arbitrariamente, a grande vantagem desse método é que as condições de amplificação são ótimas de acordo com parâmetros determinados livremente pelo pesquisador. A fase crítica diz respeito à quantificação cuidadosa do DNA digerido e dos adaptadores para a reação de ligação.

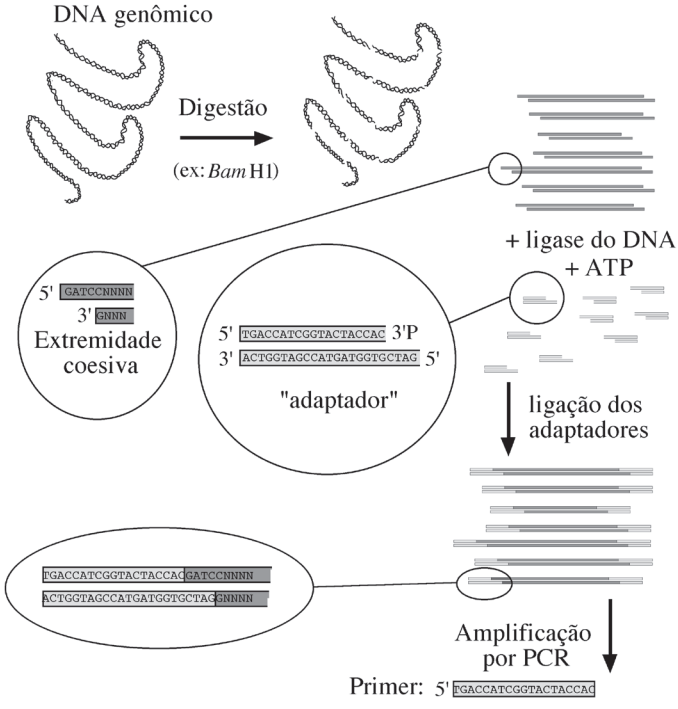


Figura 19.4. Esquema mostrando a técnica AFLP. Os polimorfismos detectados podem resultar tanto da existência de sítios de restrição como da existência de variação de tamanho de fragmentos entre os sítios.

19.8. DS-PCR

Na técnica de SPAR, ao contrário do que acontece com os RAPDs, o repertório de *primers* é reduzido, principalmente porque certas sequências estariam sujeitas ao artefato conhecido como dímero de *primer*. Por exemplo, se tentarmos utilizar um oligonucleotídeo com a sequência 5´GATCGATCGATCGATC, haveria a hibridação dele com ele mesmo (um *primer* no sentido 5´- 3´ e o outro no sentido 3´- 5´), o que reduziria sobremaneira a eficiência da PCR. Para combinar a versatilidade, em termos de sequências diferentes da técnica RAPD, com a estabilidade proporcionadas com os *primers* empregados em SPAR, Matioli e Brito (1995) desenvolveram a técnica denominada de DS-PCR (*double stringency PCR*, em inglês, "PCR de dupla restritividade"). Conforme a Figura 19.5, a PCR é dividida em duas fases distintas. Na primeira, a reação procede como na técnica de SPAR, mas durante um número de ciclos que não permite a visualização dos produtos. Na fase seguinte, com menor restritividade (proporcionada pela temperatura de hibridação mais baixa), há amplificação diferencial de fragmentos que contenham ou não sítios RAPD íntegros, o que gera, na maioria dos casos, bandas com segregação codominante. Essa técnica mostrou-se mais robusta que a de RAPD, pois a hibridação com os *primers* curtos ocorre a partir de uma fração previamente enriquecida por ciclos de alta restritividade.

19.9. SNPs

Os SNPs (*single nucleotide polymorphisms*) são resultantes de alterações pontuais, associadas a uma baixa taxa de mutação nova. Os SNPs são bialélicos, correspondendo a substituição de uma única base que ocupa uma localização em particular do genoma, sendo que a frequência do alelo menos comum é maior ou igual a 1%. Estima-se que, no genoma humano, há cerca de

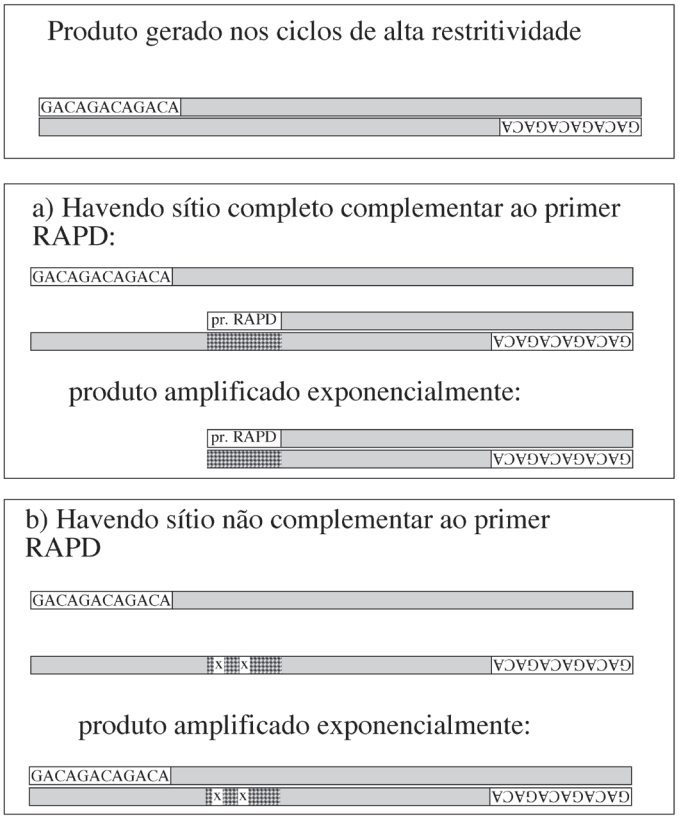


Figura 19.5. Esquema mostrando a técnica DS-PCR. No caso de haver heterozigotos para o trecho que hibrida com o *primer* RAPD, ambas as cadeias são amplificadas, gerando um padrão de bandas duplas para os heterozigotos.

um SNP a cada 1.000-2.000 bases e que há cerca de seis SNPs por região codificadora por gene (Collins *et al.*, 1998; The International SNP map working group, 2001). Uma vez que são abundantes (por exemplo há mais de 10 milhões no genoma humano já catalogados) e apresentam uma baixa taxa de mutação, os SNPs constituem bons indicadores para estudos evolutivos. O conhecimento da variabilidade entre os indivíduos poderá ter um papel importante na elucidação da história e organização do genoma humano, bem como na identificação de genes associados à susceptibilidade a determinadas doenças relativamente comuns, como diabetes, esquizofrenia, defeitos de tubo neural, fissuras lábio-palatinas, obesidade as quais têm um componente genético associado a fatores ambientais na sua etiologia. Atualmente, os SNPs representam o sistema polimórfico mais utilizado para o estudo dessas doenças comuns. Nesse contexto, foi estabelecido um mapa de alta densidade do genoma humano a partir da caracterização de milhares de SNPs ao longo do genoma, que constitui uma referencia mundial na área de genética humana e medica, o HapMap. O efeito funcional da grande maioria dos SNPs é desconhecida, mas há vários exemplos em que estão associadas a alterações transcricionais de um dado gene (http://hapmap.ncbi.nlm.nih.gov/).

Parte dos métodos atuais de análise de SNPs depende de uma reação inicial de PCR. Tais produtos podem ser analisados por diversas técnicas para a diferenciação dos dois alelos, que não utilizam necessariamente a eletroforese em gel para a observação dos resultados. Inicialmente os métodos possibilitavam a análise de SNPs individuais, como, por exemplo, o uso do sistema TaqMan® SNP Genotyping Assays (Applied Biosystems, EUA). Nessa abordagem, dois tipos de sequências são fornecidos: um contendo o alelo de menor frequência do polimorfismo e outro contendo o de maior frequência. Uma sonda e um corante fluorescente (específico para o alelo presente) são ligados a cada sequência e, durante a reação de polimerização do DNA, a sequência pareada emitirá fluorescência, revelando se o alelo do polimorfismo é o selvagem ou o polimórfico. A discriminação alélica, então, classificará as amostras como homozigotas ou heterozigotas. A maioria das técnicas de genotipagem de SNPs dependem de equipamentos específicos, como um PCR em tempo real no exemplo acima.

Atualmente se faz uso de técnicas de alta resolução, como os *microarrays* (*chips* de SNPs) de DNA, os quais se tratam de arranjos de dezenas a centenas de milhares de sondas diferentes de oligonucleotídeos ligados a um substrato inerte, como uma lâmina. Cada SNP é representado por um conjunto de sondas que correspondem às sequências de cada alelo possível e, com essa metodologia, é possível analisar os SNPs de um genoma inteiro, caso haja interesse. A maneira de como a amostra de DNA é processada dependerá de qual metodologia ou plataforma será utilizada, mas em qualquer das situações irá envolver reações de PCR. Os genótipos são gerados a partir da utilização de programas de informática com algoritmos específicos, os quais permitem classificar os SNPs em homozigotos para um dado alelo ou para o outro alelo, ou ainda em heterozigotos. Para essa classificação, o algoritmo toma por base a intensidade de sinal fluorescente emitido pelo conjunto de sondas relacionadas a cada alelo de cada SNP.

19.10 Sequenciamento direto

Até meados da década de 1970, a obtenção de sequências de ácidos nucléicos era um processo muito trabalhoso, análogo à obtenção de sequências de aminoácidos em polipeptídeos que

envolvia a purificação de RNA, sua fragmentação enzimática ou química e degradação, nucleotídeo a nucleotídeo, por métodos químicos ou enzimáticos e análise cromatográfica de cada um dos nucleotídeos (para uma revisão, veja Gilham, 1970). A partir desse ano, duas publicações sugeriram métodos bem mais simples. O método de Maxam e Gilbert (1977) envolve a marcação radioativa de uma das extremidades da fita dupla de DNA, a degradação química por quatro maneiras diferentes, o que gera fragmentos interrompidos em posições específicas (G, A+G, C, C+T). Uma separação eletroforética seguida de autorradiografia permite a inferência da sequência nucleotídea da fita marcada. Sanger *et al.*, (1977) propuseram um método no qual há síntese de cadeias complementares a uma fita molde que é interrompida em um nucleotídeo específico. São feitas quatro reações onde, em cada uma delas, adiciona-se um nucleotídeo modificado quimicamente em pequena proporção que não é reconhecido pela polimerase do DNA, o que promove interrupções na síntese da cadeia sempre no mesmo nucleotídeo. Como a fita sintetizada é marcada na extremidade 5' (por marcação do primer), os diferentes fragmentos são analisados por eletroforese seguida também de autorradiografia. Como o método de Sanger *et al.* (1977) é mais simples e pode ser facilmente comercializado em "kits", tornou-se muito difundido. Os grandes projetos de sequenciamentos genômicos inicialmente utilizaram o método de Sanger e de suas modificações, especialmente aquelas que utilizam-se de reagentes quimioluminescentes para distinguir cada uma das interrupções em bases.

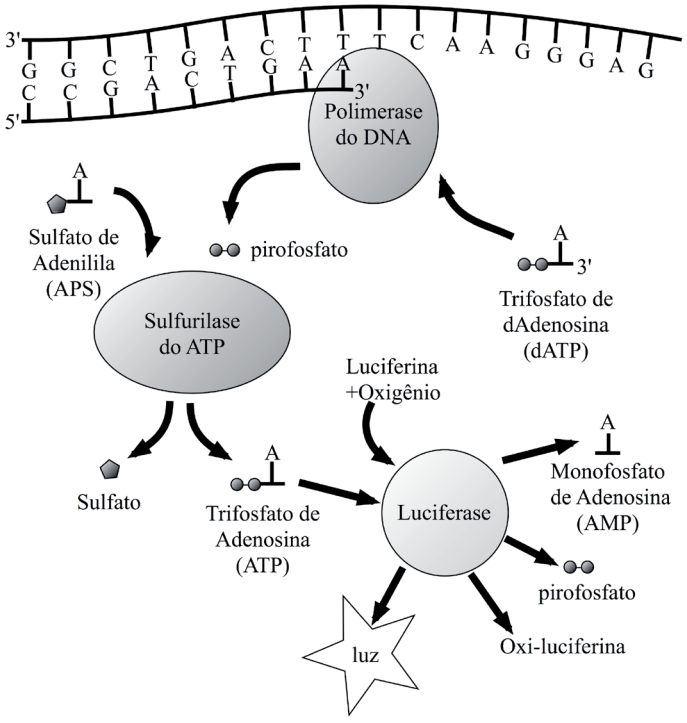


Figura 19.6. Esquema mostrando o princípio de detecção do nucleotídeo que está sendo incorporado com o emprego do método de pirosequenciamento. No caso, uma molécula de trifosfato de desoxiadenosina (dATP), ao ser incorporada, com a ação da polimerase do DNA, libera uma molécula de pirofosfato que reage com uma molécula de sulfato de adenila originando uma molécula de ATP. A detecção luminosa dá-se pela reação da luciferina com o ATP na presença de luciferase. Não confunda dATP (desoxinucleotídeo) com ATP (cofator energético).

19.11. Pirosequenciamento

O pirosequenciamento adota uma estratégia bastante diferente em relação aos procedimentos descritos acima. Esses procedimentos envolvem a interrupção da síntese complementar de uma cadeia de DNA e posterior eletroforese dos fragmentos obtidos. O pirosequenciamento foi desenvolvido em 1996 por Ronaghi *et al.* e não envolve a eletroforese de fragmentos. O nome "pirossequenciamento" deve-se à detecção do pirofosfato liberado cada vez que um nucleotídeo é incorporado à sequência complementar que está sendo sintetizada (Figura 19.6). O método de pirosequenciamento descrito a seguir é aquele que tem sido mais empregado recentemente e é realizado em um equipamento conhecido popularmente como 454 (454 Life Science Corp., Margulies *et al.*, 2005) Em primeiro lugar, um fragmento de DNA de fita simples é imobilizado pela ligação covalente de sua extremidade 3' a um substrato (uma esfera de resina), em condições tais em que a maioria das esferas tenha uma única molécula de DNA. As esferas então são colocadas em óleo mineral e emulsificadas com uma solução contendo os reagentes para PCR. Essa mistura é emulsificada de tal forma que as esferas ficam em "câmaras" imersas em óleo, que é submetida aos ciclos térmicos para amplificação. As fitas complementares, não ligadas às esferas, são removidas e as esferas são passadas para um aparato onde em cada célula cabe uma única esfera que, por sua vez, contém cópias de um único trecho de DNA. Em seguida, um primer 5'-3' é hibridado nessa extremidade e serve de ponto de iniciação da síntese. Soluções contendo apenas um dos nucleotídeos A, C, G ou T são adicionados e removidos sequencialmente (Figura 19.7). Quando há coincidência do nucleotídeo adicionado, a polimerase presente na solução catalisa a extensão da cadeia promovendo a liberação do íon pirofosfato detectado por uma reação de quimioluminescência. A intensidade da luz produzida é registrada, que é importante para a detecção de subsequências homopoliméricas, CCC, por exemplo. Nesse caso, a intensidade luminosa é teoricamente o triplo daquela produzida por um único nucleotídeo C. Os resultados são plotados em um gráfico, como mostrado na Figura 19.8. Outros métodos de sequenciamento direto de DNA vem sendo desenvolvidos de forma bastante célere, com tecnologias que vão desde o emprego de microarranjos de sondas até o desenvolvimento de um circuito eletrônico que "lê" as bases de moléculas de DNA de fita simples à medida que essa passa por um nanoporo. Para uma revisão recente sobre o assunto, a revisão de Shendure e Ji (2008) pode ser consultada.

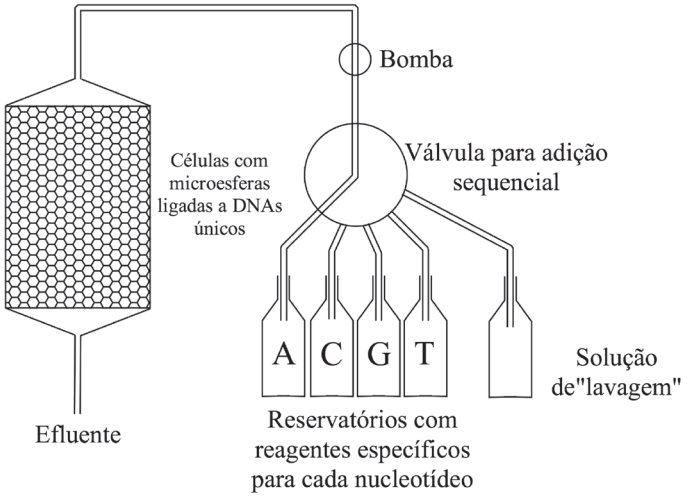


Figura 19.7. Esquema mostrando o mecanismo utilizado para o pirossequenciamento onde cada uma das células (à esquerda) contém uma microesfera ligada a DNAs de um único tipo. A adição sequencial da mistura de reação que contém um dos quatro tipos de nucleotídeos de cada vez permite o sequenciamento em paralelo através da detecção de luz em cada uma das células.

19.12. Exemplos de Utilização

19.12.1. Utilização de microssatélites no diagnóstico molecular e no mapeamento de genes humanos

Na Figura 19.8, está exemplificada a análise de um microssatélite polimórfico, D2S2109, em uma família brasileira com distrofia muscular progressiva do tipo cinturas e de herança autossômica recessiva (DMC2). As DMC2s são um grupo bastante heterogêneo de doenças, sendo que existem pelo menos oito locos distintos que podem causar essa doença (ver revisões em Passos-Bueno *et al.*, 1999; Zatz *et al.*, 2000). Como pode ser observado na Figura 19.9a, todos os afetados receberam dois alelos iguais do marcador microssatélite D2S2109, enquanto os indivíduos normais receberam apenas um ou nenhum desses alelos. O estudo dessa família exemplifica um caso de ligação entre um marcador de microssatélite mapeado no cromossomo 2, que é uma região candidata para DMC, e a doença presente nessa família. Esse resultado significa que a doença está sendo causada por uma mutação em um gene localizado próximo desse marcador. Essa sugestão de ligação foi também confirmada por testes estatísticos (o chamado teste de *Lod score;* Lathrop *et al.*, 1984). Já na Figura 19.9b, essa mesma família foi analisada com o marcador de microssatélite do cromossomo 17 (D17S250). Observa-se, contudo, que não há segregação entre qualquer alelo desse microssatélite e a doença, reforçando que o loco que causa a doença situa-se, de fato, no cromossomo 2.

A análise entre um marcador polimórfico, como o exemplificado acima, e uma doença é conhecida como "estudo de ligação". Ou seja, o estudo de ligação consiste em verificar-se a segregação de dois marcadores através da meiose: se dois marcadores segregarem sempre juntos, sugere-se que eles estão ligados, ou seja, estão fisicamente próximos; por outro lado, se a segregação entre esses marcadores for independente, conclui-se que esses dois locos não estão ligados e, portanto, estão genética e possivelmente distantes entre si fisicamente. No exemplo da DMC2, citado acima, foi realizada análise de segregação utilizando-se a informação das possíveis localizações dos genes que podem causar esse fenótipo. Contudo, em casos em que não se conhece a região cromossômica onde está o gene associado com a doença, pode-se utilizar o método de análise de ligação para mapear o loco em questão. Na Figura 19.10, está exemplificada uma família com a síndrome de Knobloch, que é uma condição genética de herança autossômica recessiva. O gene que causa essa doença foi mapeado no braço longo do cromossomo 21 (Sertié *et al.*, 1996). Para obter-se esse resultado, foram utilizados mais de 400 marcadores de microssatélites dispersos ao longo do

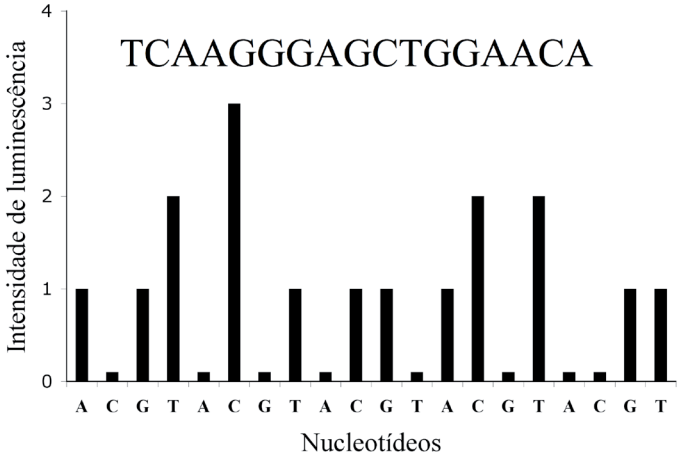


Figura 19.8. Esquema mostrando a intensidade luminosa obtida em uma célula do equipamento de pirossequenciamento ao longo da sequência de adição de nucleotídeos A, C, G e T que se repete, abaixo da interpretação da sequência complementar reversa.

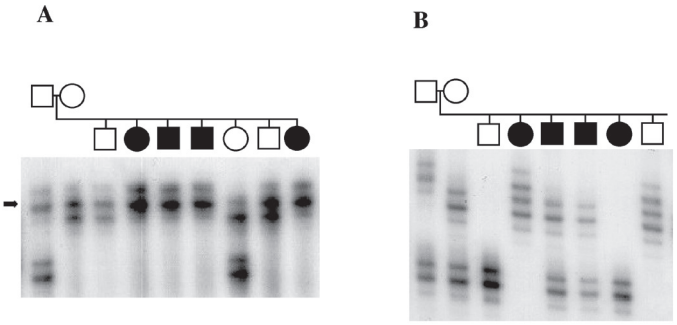


Figura 19.9. Análise de segregação entre um loco da distrofia muscular tipo Cinturas e um marcador de microssatélite (D2S2109) do cromossomo 2 (A) e um marcador de microssatélite (D17S250) do cromossomo 17 (B) humanos.

genoma e verificou-se que apenas os marcadores do cromossomo 21 segregavam junto com a doença (Figura 19.10). O mapeamento de genes é realizado quase que exclusivamente através do uso da PCR, uma vez que há marcadores polimórficos do tipo microssatélites identificados ao longo de todos os cromossomos humanos (www.ncbi.nlm.nih.gov).

Atualmente, em vez de se utilizarem os 400 marcadores de microssatelites para os estudos de ligação, utiliza-se *microarrays* de SNPs, o que significa um aumento na eficiência e rapidez quanto à obtenção dos resultados, pois, em um único experimento, é possível analisar o genoma humano inteiro com uma densidade de marcadores dispersos a cada 20cM.

19.12.2 Análises de locos para caracteres quantitativos

O emprego de marcadores moleculares para a detecção de variação genética com efeitos fenotípicos não se restringe apenas aos caracteres qualitativos, tais como aqueles relacionados a doenças genéticas de segregação mendeliana, conforme visto no item anterior. Existe também a possibilidade de empregar marcadores associados a caracteres de variação contínua, tais como peso, altura, velocidade etc. A essa área de estudo da genética contemporânea dá-se o nome de estudo de QTLs (do inglês, *Quantitative Trait Loci*). Para uma revisão a respeito desse assunto, recomendamos a consulta do arttigo de MacKay, 2001)

Há que se fazer uma ressalva quanto às características ditas quantitativas e as características que também são influenciadas por muitos locos mas que são qualitativas, como no caso de doenças multifatoriais, conforme exemplificadas na seção 19.9. Essas doenças são causadas por uma combinação de fatores genéticos e ambientais, quando esses atingem um determinado limiar. Nesse caso, não há acesso à variável que tem uma distribuição contínua de valores. De qualquer maneira, a abordagem é semelhante àquela utilizada nos estudos de QTLs, como descrito abaixo.

Para o estudo dos QTLs, existem duas estratégias básicas. A primeira delas refere-se ao estudo de "locos candidatos". A partir do estudo da variação genética em genes que podem estar envolvidos fisiologicamente no caráter que se pretende estudar e também a partir de análises estatísticas de grupos que se diferen-

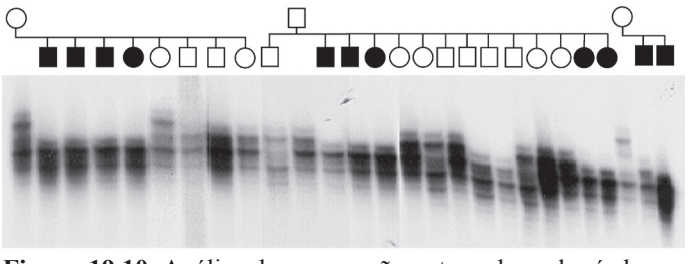


Figura 19.10. Análise de segregação entre o loco da síndrome de Knobloch e um marcador de microssatélite (D21S171) do cromossomo 21 humano.

ciam por classes genotípicas, pode-se testar a hipótese de que a variação genética existente no loco candidato esteja relacionada com a variação fenotípica. Esse tipo de abordagem tem sido empregado com sucesso em organismos dos quais se dispõe de conhecimentos razoáveis em termos do papel fisiológico desempenhado pelos genes. Mas, mesmo assim, não há garantias de que a variação genética importante seja localizada por essa estratégia, pois o desconhecimento do papel fisiológico da maioria dos genes é muito grande, mesmo em organismos simples.

A alternativa para o estudo de locos candidatos é o emprego de marcadores genéticos que apresentam variação genética, espalhados de maneira idealmente uniforme pelo genoma. Nesse caso, o que se observa nas análises é a segregação conjunta de genes que, de fato, são importantes para a característica fenotípica, com os marcadores polimórficos pela simples ligação física dessas entidades. Um exemplo do emprego dos marcadores para a caracterização de variação genética relativa à morfologia de populações de *Drosophila* está ilustrado por Matioli e Templeton (1999). Nesse estudo, ficou evidenciado que pode haver regiões do genoma que, individualmente, não contribuem com a variação fenotípica, mas que, em conjunto, apresentam efeitos significativos. Com o emprego de um grande número de marcadores moleculares e de métodos estatísticos sofisticados, a resolução de regiões do genoma que contêm variação genética importante para a variação fenotípica pode permitir a localização de genes e de interações entre eles que nem sequer poderiam ser considerados como suspeitos anteriormente, o que aumenta muito o potencial da caracterização funcional dos genomas. A metodologia de QTLs também permitiu, por exemplo, detectar-se interações de diferentes locos dsistribuídos pelo genoma de camundongos que interferem no comportamento materno de construção de ninhos (Peripato *et al.*, 2002), como do próprio tamanho da ninhada (Peripato *et al.*, 2004).

19.12.3. Análise de variação genética em populações naturais

Conforme já visto nos Capítulos 17 e 18, há grande interesse na caracterização da variação genética de populações naturais. Nos Capítulos 21 e 22, esse tema será retomado em profundidade, incluindo uma comparação do potencial de utilização dos marcadores genéticos polimórficos, descritos nesse capítulo, para programas de conservação da biodiversidade.

Referências Bibliográficas

Bray, M.S., Boerwinkle, E. e Doris, P.A. (2001). High-throughput multiplex SNP genotyping with MALDI-TOF mass spectrometry: Practice, problems and promise. **Hum. Mutat. 17**: 296-304

Bucci, G. e Menozzi, P. (1993). Segregation analysis of random amplified polymorphic DNA (RAPD) markers in *Picea abies* Karst. **Mol. Ecol. 2**: 227-232.

Caetano-Anolles, G., Bassam, B.J. e Gresshoff, P.M. (1991). DNA amplification fingerpriting using very short arbitrary oligonucleotide *primers*. **Biotechnology 9**: 553-557.

Collins F.S., Brooks L.D. e Chakravati A. (1998). A DNA polymorphism discovery resource for research on human genetic variation. **Genome Research 8**: 1229-1231.

Cheng, S., Fockler, C., Barnes, W.M. e Higuchi, R. (1994). Effective amplification of long targets from cloned inserts and human genomic DNA. **Proc. Natl. Acad. Sci. USA 91**: 5695-5699.

Cline, J., Braman, J.C., Hogrefe, H.H. (1996). PCR fidelity of Pfu DNA polymerase and other thermostable DNA polymerases. **Nucl. Acids Res. 24**: 3546-3551.

Don, R.H., Cox, P.T., Wainwright, B.J., Baker, K. e Matticks, J.S. (1991). Touchdown PCR to circumvent spurious priming during gene amplification. **Nucleic Acids Res. 19**: 4008-4008.

Dunham, I., Hunt,A.R., Collins, J.E., Bruskiewich, R., Beare, D.M., Clamp, M., Smink, L.J., Ainscough, R., Almeida, J.P., Babbage, A., Bagguley, C., Bailey, J., *et al.* (1999). The DNA sequence of human chromosome 22. **Nature 402**: 489 - 495.

Ellsworth, D.L., Rittenhouse, K.D. e Honeycutt, R.L. (1993). Artifactual variation in randomly amplified poplymorphic DNA banding patterns. **BioTechniques 14**: 214-217.

Franco, R.F., Araújo, A.G., Guerreiro, J.F., Elion, J. e Zago, M.A. (1998). Analysis of the C677T mutation of the methylenetetrahydrofolate reductase gene in different ethnic groups. **Thromb. Haemost. 79**: 119-121.

Gaspar, D.A., Pavanello, R.C., André, M., Zatz, M., Steman, S., Wyszynski, D., Matioli, S.R. e Passos-Bueno M.R. (1999). The role of the C677T polymorphism at the MTHFR gene on risk to nonsyndromic cleft lip with/without cleft palate: result from a case-control study in Brazil. **Am. J. Med. Genet. 87**: 197-199.

Gupta, M., Chyi, Y.-S., Romero-Severson, J., Owen, J.L. (1994). Amplification of DNA markers from evolutionarily diverse genomes using single *primers* of simple-sequence repeats. **Theor. Appl Genet 89**: 998-1006.

Hoogendoorn, B., Owen, M.J. Oefner, P.J., Williams, N, · Austin, J. e O'Donovan, M.C. (1999). Genotyping single nucleotide polymorphisms by primer extension and high performance liquid chromatography. **Hum. Genet. 104**:89–93.

Innis, M.A. e Gelfand, D.H. (1990). Optimization of PCRs. **In** Innis, M.A., Gelfand, D.H., Sninsky, J.J. e White, T.H. PCR protocols, a guide to methods and applications. Academic Press, San Diego, pp. 3-20.

Kuppuswamy, M.N., Hoffmann, J.,W, Kasper, C.K., Spitzer, S.G., Groce, S.L. e Bajaj, S.P.(1991) Single nucleotide primer extension to detect genetic diseases: experimental application to hemophilia B (factor IX) and cystic fibrosis genes. **Proc. Natl. Acad. Sci. USA 88**:1143-1147.

Lathrop, G.M., Lalouel, J.M., Julier, C. e Ott, J. (1984). Strategies for multilocus linkage analysis in humans. **Proc. Natl. Acad. Sci. USA 81**: 3443-3446.

MacKay, T.F.C. (2001). The genetic architecture of quantitative traits. **Ann. Rev. of Genet. 35**: 303-339.

Matioli, S.R. e Brito, R.A. (1995). Obtaining genetic markers by using double stringency PCR with microsatellite and arbitrary primer*s*. **BioTechniques 19**: 752-755.

Matioli, S.R. e Templeton, A.R. (1999). Coadapted gene complexes for morphological traits in *Drosophila mercatorum*. Two loci interactions. **Heredity 83**: 54-61.

Margulies *et al.* (2005). Genome sequencing in microfabricated high-density picolitre reactors. **Nature 437**:376-380.

Maxam, A.M. e Gilbert, W, (1977). A new method for sequencing DNA. Proc. Natl. Acad. Sci. USA 74:560–564.

Passos-Bueno, M.R., Wilcox, W.R., Jabs, E.W., Serti, A.L., Alonso, L.G. e Kitoh, H. (1999). Clinical spectrum of fibroblast growth factor receptor mutations. **Hum. Mutat. 14**: 115-125.

Peripato, A.C., de Brito, R.A., Vaughn, T.T., Pletscher, L.S., Matioli, S.R. e Cheverud, J.M. (2002). Quantitative trait loci for maternal performance for offspring survival in mice. **Genetics 162**(3):1341-1353.

Peripato A.C., de Brito, R.A, Matioli S.R., Pletscher, L.S., Vaughn, T., Cheverud, J.M. (2004). Epistasis affecting litter size in mice. **J. Evol. Biol. 17**(3):593-602.

Rabinow, P. (1996). **Making PCR: A story of biotechnology**. The University of Chicago Press, Chicago.

Refseth, U.H., Fangan, B.M. e Jakobsen, K.S. (1997). Hybridization capture of microsatellites directly from genomic DNA. **Eletrophoresis 18**: 1519-1523.

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M. e Nirén, P. (1996). Real-time DNA sequencing using detection of pyrophosphate release. Anal. Biochem. 242:84-89.

Saiki, R.K., Scharf, S., Faloona, F., Mullis, K.B., Horn, G.T., Erlich, H.A. e Arnheim, M. (1985). Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle-cell anemia. **Science 230**: 1350-1354.

Sanger, F., Nicklen, S. e Coulson A.R. (1977). DNA sequencing with chain-terminating inhibitors. **Proc. Natl. Acad. Sci. USA** 74: 5463–4467.

Sertié, A.L., Quimby M., Moreira E.S., Murray J,. Zatz M., Antonarakis S.E. e Passos-Bueno, M.R. (1996). A gene which causes severe ocular alterations andoccipital encephalocele (Knobloch syndrome) is mapped to 21q22.3. **Hum. Mol. Genet. 5**: 843-847.

Shendure, J. e Li, H. (2008). Next-generation DNA sequencing. **Nature Biotechnol. 26**:1135-1145.

Southern, E.M. (1996). DNA chips: Analysing sequence by hybridization to oligonucleotides on a large scale. **Trends in Genetics 12**: 110-115.

The International SNP map working group (2001). A map of human

genome sequence variation containing 1.42 million single nucleotide polymorphisms. **Nature 409**: 928-933.

Tautz, D. (1989). Hypervariability of simple sequences of a general source for polymorphic DNA markers. **Nucleic Acids Res. 17**: 6463-6471.

Valsecchi, E. (1998). Tissue boiling: a short-cut in DNA extraction for large-scale population screenings. **Mol. Ecol. 7**: 1243-1245.

Van der Put, N.M.J., Gabreels, F., Stevens, E.M.B., Smeitink, J.A.M., Trijbels, F.J.M. Eskes, T.K.B., van den Heuvel, L.P. e Blom, H.J. (1998). A second common mutation in the methylenetetrahydrofolate reductase gene: An additional risk factor for neutral-tube defects? **Am. J. Hum. Genet. 62**: 1044-1051.

Velikanov, M.V. e Kapral, R. (1999). Polymerase chain reaction: A Markov process approach. **J. Theoret. Biol. 201**: 239-249.

Vos, P., Hogers, R., Bleeker, M., Reijans, M., van de Lee, T., Hornes, M., Frijters A., Pot J., Peleman, J., Kuiper, M. e Zabeau, M. (1995). AFLP: a new technique for DNA fingerprinting. **Nucleic Acids Res. 23**: 4407-4414.

Wang, Z., Moult, J. (2001). SNPs, protein structure, and disease. **Hum Mutat 17**: 263-270.

Weber, R.D. e May, P.E. (1989). Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. **Am. J. Hum. Genet. 44**: 388-396.

Welsh, J. e McClelland, M. (1990). Fingerprinting genomes using PCR with arbitrary *primers*. **Nucleic Acids Res. 18**: 7213-7218.

Williams, J.G.K., Kubelik, A.R., Livak, K.J., Rafalski, J.A. e Tingey, S.V. (1990). DNA polymorphisms amplified by arbitrary *primers* are useful as genetic markers. **Nucleic Acids Res. 18**: 6531-6535.

Zatz, M., Vainzof, M. e Passos-Bueno, M.R. (2000). Limb-girdle muscular dystrophy: one gene with different phenotypes, one phenotype with different genes. **Current Opinion in Neurology 13**: 511-517.

Capítulo 20

Introdução às Árvores Genealógicas e à Teoria da Coalescência

Flora Maria de Campos Fernandes (flozinha123@yahoo.com.br)
Laboratório de Biologia Computacional
Instituto de Biologia - Universidade Federal da Bahia, UFBA

Tudo o que já foi é o começo do que vai vir.
Guimarães Rosa em *Grande Sertão Veredas*

Coalescent theory represents the most significant progress in theoretical population genetics in the past two decades of this century. It is now widely recognized as a cornerstone for rigorous statistical analyses of molecular data from populations. In the future, challenges from the rapidly expanding body of molecular data will continue to inject fresh blood into the development of coalescent theory.
Norm Bourg

20.1. Para Começar

Vamos começar do começo, esclarecendo uma dúvida muito comum: "Qual é a diferença entre genealogias e filogenias?" Bem, a diferença é enorme! Quando tratamos de filogenias, procuramos entender as relações evolutivas entre espécies (ou mesmo entre níveis taxonômicos superiores) diferentes, ou seja, relações evolutivas **interespecíficas**—o que quer dizer que nos baseamos em uma ou mais regiões genômicas homólogas para entendermos a história evolutiva de diferentes táxons (e aqui falamos sobre árvores de espécies). Diferentemente, quando tratamos de genealogias, o interesse está em saber como foi a história evolutiva de determinados genes e seus alelos e/ou haplótipos dentro de uma espécie particular (isto é, em suas populações), ou seja, será uma análise **intraespecífica** (neste caso, falamos sobre árvores de genes, Figura 20.1). Não importa que sejam utilizados os mesmos genes para as análises—os objetivos são completamente diferentes

Em algumas situações, árvores de genes e árvores de espécies podem ser coincidentes, mas nem sempre isso acontece. Uma olhada no clássico esquema de Nei (1987) é suficiente para entender (Figura 20.2). Como veremos em detalhe mais adiante, a deriva genética associada ao tempo de geração (T) e ao tamanho efetivo da população (*Ne*, parcela da população que efetivamente deixa descendentes) é que influencia nas diferenças ou semelhanças entre topologias de árvores de genes e de espécies. Por exemplo, no caso de populações com tamanho efetivo pequeno e tempo de geração grande e somente sob o efeito da deriva genética, a probabilidade de topologias diferentes é grande. Inversamente, populações com tamanho efetivo grande e tempo de geração pequeno, apenas a deriva genética faz com que a probabilidade de topologias coincidentes seja grande (Figura 20.3).

20.2. Genes, Alelos e Populações

Imagine uma população diplóide com N=2 indivíduos, que possua um gene com dois alelos (*D* e *d*). Quantos estados alélicos essa população pode apresentar com relação, por exemplo, ao alelo *D*? Simples, cinco! 0, 1, 2, 3, 4, no total, ou seja, ou o alelo *D* não está presente na população (estado 0), ou a população é composta por um heterozigoto (*Dd*) e um homozigoto recessivo (*dd*) (estado 1), ou por um homozigoto dominante (*DD*) e um recessivo (*dd*) (estado 2), ou por um heterozigoto (*Dd*) e um homozigoto dominante (*DD*) (estado 3) ou então por dois homozigotos dominantes (*DD* e *DD*) (estado 4). Isto quer dizer que as frequências alélicas são o que determinam os estados gênicos (ou alélicos) populacionais. Generalizando, podemos dizer que em uma população com *2N* genes com dois alelos, os estados alélicos populacionais variam de 0 (zero) a *2N*. No exemplo anterior, se olharmos para os estados extremos, teremos duas situações: 0, ausência total do alelo *D* (evidenciando a perda de um dos alelos), ou 4, presença total do alelo *D* (fixação alélica). Esses são os chamados estados de absorção, nos quais, na ausência de mutações diretas ou reversas (um alelo dominante tornar-se um recessivo e vice-versa), a população continuará do jeito que está em termos de frequência alélica. O processo evolutivo estocástico subjacente a essas situações é a deriva genética.

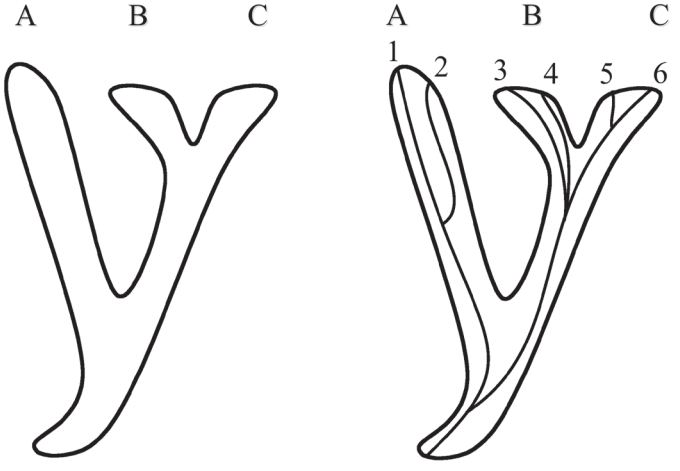


Figura 20.1. Árvore de espécies (esquerda) e árvore de genes (direita, representada pelas linhas internas). Letras indicam diferentes espécies, números indicam os alelos presentes em cada população.

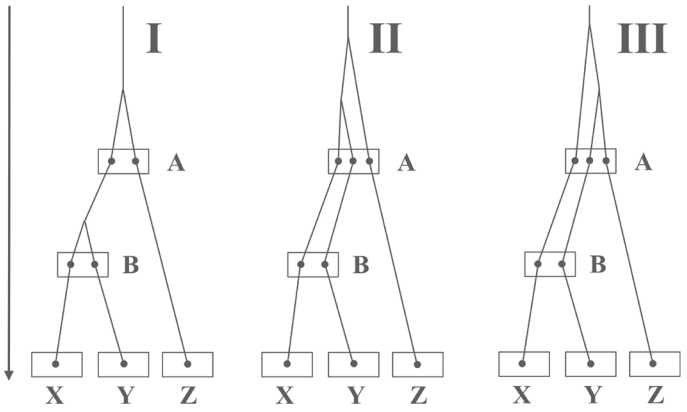


Figura 20.2. Três possíveis relações entre árvores de genes e árvores de espécies. A e B representam momentos de divergência gênica. Notar que apenas nas topologias I e II as histórias gênicas e as das espécies são coincidentes (modificado de Nei, 1987).

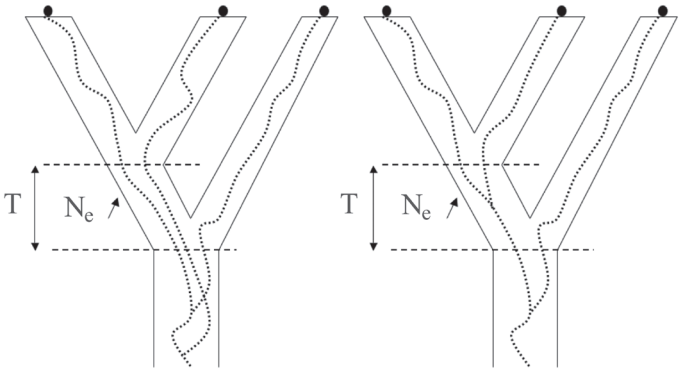


Figura 20.3. Tempo de geração (T) grande e tamanho efetivo (Ne) pequeno, probabilidade de topologias gênica e de espécie diferentes (esquerda). T pequeno e Ne grande, probabilidade de topologias gênica e de espécie iguais.

Para entendermos melhor as flutuações das frequências alélicas em uma população natural ao longo das gerações, podemos lançar mão de modelos matemáticos. O modelo matemático mais simples e elegante para fenômenos estocásticos que envolvem tempo (que é o nosso caso, pois trabalhamos com gerações) é aquele baseado em cadeias de Markov. Nesse caso, temos uma cadeia de variáveis aleatórias em que o estado futuro depende apenas do estado presente (lembrando que *estado*, aqui, representa a frequência alélica em uma população em uma dada geração). Por isso, dizemos que cadeias de Markov não têm memória ou qualquer tipo de "inércia", ou seja, estados passados não exercem qualquer influência na transição do estado presente para o estado da próxima geração.

Suponha que o estado atual de uma população seja *i* e o da próxima geração seja *j*; desse modo, a probabilidade de transição de *i* para *j* será:

$$P(X_{n+1} = j \mid X_n = i) = P_{ij}.$$

Assim, é possível construir uma matriz de probabilidades de transição de um estado qualquer para outro estado qualquer, a qual chamaremos de matriz *P*:

$$matriz\ P = \begin{pmatrix} P_{11} & P_{12} & P_{13} \ldots P_{1n} \\ P_{21} & P_{22} & P_{23} \ldots P_{2n} \\ . & . & . & . \\ P_{n1} & P_{n2} & P_{n3} \ldots P_{nn} \end{pmatrix}$$

$\sum_j P_{ij}$ e a deverá ser sempre 1 (um), pois as populações apresentam um número limitado de estados, *2N+1* para populações diplóides e *N+1* para populações haplóides.

Há classes de cadeias de Markov em que se chega a um limite de probabilidades, ou seja,

$$\pi_j = \lim_{n \to \infty} P_{ij}^n,$$

o que é independente de *i*. Em outras palavras, π_i pode ser interpretado como a proporção de tempo (gerações) em que a população permanece no estado *i*. Esse é o chamado *estado estacionário* de uma população, no qual a probabilidade de fixação de um dado alelo será igual à frequência desse alelo na população ancestral. Além disso, para populações em estado estacionário, a cadeia de Markov representará probabilidades *fixas* de transição de um estado para outro. Esses são conceitos fundamentais quando realizamos análises de coalescência.

Mas o que é coalescência? O que é coalescer quando estudamos a história de alelos ao longo das gerações de uma população? Coalescer significa unir, juntar. Mas como alelos se unem? Basta olharmos o tempo pregresso e veremos que alelos se unem em um ancestral comum (Figura 20.4). Como muito bem definido por Alan Templeton (com. pess.), a coalescência pode ser vista como o inverso temporal da replicação do DNA (Figura 20.5).

20.3. A Teoria da Coalescência

Há poucos anos, Joseph Felsenstein (2003) propôs uma analogia muito interessante para o processo da coalescência, a qual denominou "*bugs in a box*" (besouros numa caixa). Nessa analogia, *k* besouros extremamente vorazes, insaciáveis, hiperativos e nada discriminantes são colocados em uma caixa fechada. Dentro dela, permanecem voando incansavelmente. Ocasionalmente, dois insetos colidem e um deles devora o outro.

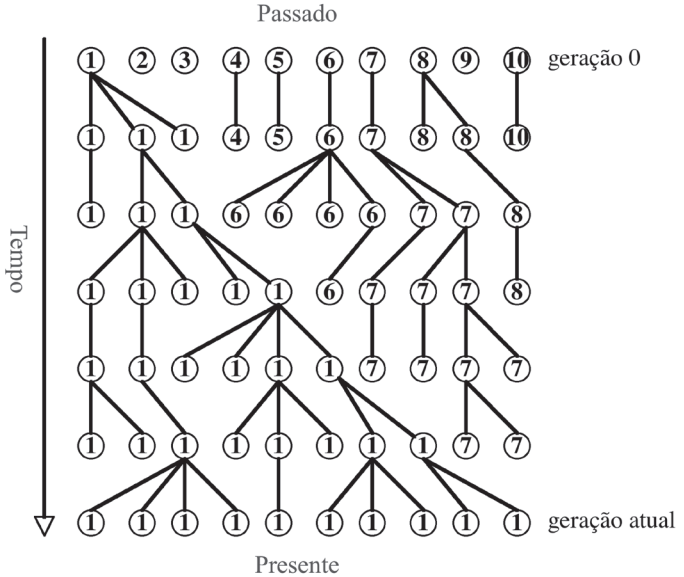


Figura 20.4. Esquema representativo do comportamento dos genes na população ao longo de seis gerações. Observar que as cópias dos alelos "1" da geração atual compartilham um único e mesmo alelo ancestral na geração 1. υ = tempo de coalescência.

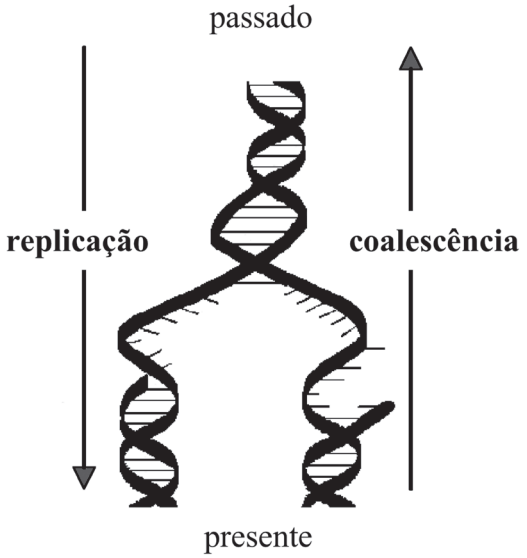


Figura 20.5. A coalescência é o inverso temporal da replicação.

Como são insaciáveis, o devorador retoma seu vôo e novas colisões acontecem (aqui, a coalescência, "dois besouros colidem e resta um"). Nesse processo, o número de besouros dentro da caixa vai se reduzindo de *k* para *k-1*, *k-2*, *k-3*… até que restará apenas um besouro. Assim, o número de pares de besouros que podem colidir será

$$\frac{k(k-1)}{2}.$$

Se na caixa há *2N* lugares que possam ser ocupados, então a probabilidade de colisões será de

$$\frac{k(k-1)}{4N},$$

e, uma vez que o volume da caixa representa o tamanho populacional, se dobrarmos este volume e não aumentarmos o número inicial de besouros, o processo de colisões será duas vezes mais lento

$$\frac{k(k-1)}{8N}.$$

Esse processo pode ser transferido para o comportamento dos alelos ao longo das gerações de uma população, apenas visto no sentido temporal contrário.

Formalmente, se examinarmos bem à frente no tempo em qualquer geração, chegaremos a um momento em que todos os alelos de um loco descenderão de uma das *2N* cópias de um gene da população atual. Igualmente, se examinarmos bem remotamente em qualquer geração, chegaremos a um momento em que todos os alelos atuais de um dado loco se encontram em uma única cópia de um gene no passado (Fu & Li, 1999; Kingman, 2000) (Figura 20.4). A maneira com que todas as cópias alélicas se remetem a um único gene ancestral é denominada *coalescência* e aquele gene ancestral único é chamado *coalescente* (Ridley, 2006) ou ancestral comum mais recente (*MRCA – most recent common ancestor*) (Hudson, 1991). Na Figura 20.4, também é possível observar que a existência de um único gene ancestral para todos os alelos atuais em um loco não significa que apenas um gene existiu naquele período (a geração zero possuía tantos genes quanto qualquer outra geração). No entanto, alguns genes foram se perdendo ao acaso ao longo das gerações. No caso de esquema apresentado na Figura 20.4, podemos dizer que o *tempo de coalescência* (representado pela letra grega upsilon - υ) do alelo "1" ocorreu há cinco gerações passadas. A Figura 20.6 traz a forma como a topologia da árvore gênica pode ser representada neste caso.

20.4. Trabalhando com Probabilidades

Como demonstrado por Wright nos anos 1960, a probabilidade de um descendente possuir um ancestral, na geração anterior, é igual a 1 e a probabilidade de um segundo descendente possuir o mesmo ancestral que o primeiro é igual a

$$\frac{1}{2N}.$$

Logo, a probabilidade de dois alelos compartilharem o mesmo ancestral na geração anterior será igual a

$$\frac{1}{2N},$$

sendo *N* o tamanho da população. Assim, a cada geração, há

$$\frac{1}{2N}$$

de chance de dois alelos coalescerem, de modo que a esperança do tempo de coalescência de dois alelos será dada por $E(\upsilon)=2N$. Mais tarde, Sir John Kingman (2000) generalizou o raciocínio para *k* cópias de um alelo (lembre-se dos besourinhos na caixa!):

$$probabilidade\,(\text{k cópias serem reduzidas para k - 1 cópias}) = \frac{k(k-1)}{4N}$$

assim,

$$E(\upsilon_k) = \frac{4N}{k(k-1)},$$

o que ficou conhecido como a "coalescência *n* de Kingman" (*Kingman's n-coalescent*).

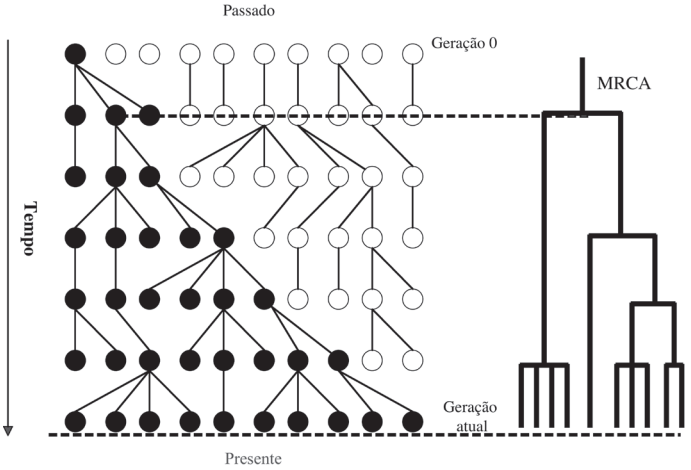


Figura 20.6. Esquema comparativo entre comportamento alélico e topologia que o representa.

20.5. O Parâmetro Θ (Theta)

Como foi dito no começo, se quisermos entender a história evolutiva dos genes, teremos que trabalhar com modelos matemáticos, probabilidades e, quando possível (o que é raro), com informações adicionais. Para isso, vamos introduzir mais alguns conceitos para entender o que é o parâmetro Θ. Primeiramente, como já vimos, a coalescência envolvendo *k* cópias alélicas é definida em termos de tamanho populacional (*N*) e tempo (gerações). No entanto, nesses estudos, não podemos medir tempo, mas podemos inferir divergências gênicas. De um modo geral, as equações matemáticas envolvidas podem ser formuladas em termos de tamanho populacional (N), tempo (gerações) e taxa de mutação (μ). Com isso, podemos descrever o parâmetro Θ como:

$$\Theta = 4N\mu \text{ para populações diplóides}$$
$$\Theta = 2N\mu \text{ para populações haplóides}.$$

Naqueles casos raros em que temos informações adicionais sobre a história evolutiva da população, é possível separar Θ de μ.

Na prática, esbarramos com algumas limitações em análises de coalescência, mas comecemos pelos casos em que essas limitações são reduzidas (em outras palavras, por casos hipotéticos, mas que modelos matemáticos não se valem de casos hipotéticos?). Outro ponto importante é que metodologias de análises filogenéticas podem ser valiosas em análises de coalescência, em especial os algoritmos de máxima verossimilhança (veja Capítulo 21).

Os dados que podemos obter de imediato são os dados atuais, ou seja, podemos detectar os alelos/haplótipos de um determinado gene na população vivente; segundo, também é possível estabelecer um modelo evolutivo de substituição de nucleotídeos (veja Capítulo 14) para aquela amostra em particular. E, isto feito, basta inverter temporalmente o modelo! Se for assim, os passos a serem seguidos são: {abordagem A, mais simples} 1 – coletar as sequências gênicas; 2 – estimar a árvore alélica (ou gênica) desprezando eventos de recombinação; 3 – pronto! A história estaria pronta para ser contada. Mas vamos ser mais realistas: {abordagem B}, 1 – sequências coletadas; 2 – considerar todas as possibilidades genealógicas, incluindo recombinação (lembrando que esta é bastante comum em alguns genomas), além de modelos evolutivos de substituição de nucleotídeos testados e não rejeitados (Capítulo 14); 3 – computar o valor de verossimilhança para cada genealogia obtida; 4 – estimar parâmetros que maximizem a verossimilhança dos dados e testar modelos evolutivos comparando a verossimilhança sob diferentes hipóteses.

Considerando os passos descritos acima, é importante lembrar que: (a) os estados alélicos de diferentes locos podem ser estatisticamente dependentes devido à ligação gênica; (b) os estados alélicos de diferentes haplótipos podem ser estatisticamente dependentes devido à ancestralidade comum; (c) tais dependências são resultados de uma única história de mutações, recombinações e coalescência. Assim, se desejamos uma análise estatisticamente coerente dos dados, mutação, recombinação e coalescência devem ser incorporadas. Além disso, vale a pena ressaltar que métodos filogenéticos por si sós não são suficientes para as análises, podendo levar a um viés na interpretação dos resultados. Por esse motivo, também lançamos mão de modelos estocásticos para modelar o passado. Isso é possível porque a coalescência é um processo estocástico, além de ser uma extensão natural dos modelos clássicos da genética de populações.

20.6. Construindo Genealogias

Imagine um conjuto de dados composto por sequências gênicas com 10kb cada, coletadas ao acaso de 20 indivíduos de uma população qualquer. Suponha, agora, que nesse conjunto de dados haja 100 sítios polimórficos. Numa situação como essa, após a reconstrução genealógica, a topologia de árvore revelará tanto a história da coalescência das linhagens como a história das mutações envolvidas nos sítios polimórficos! Suponha, agora, que não haja sítios polimórficos no conjunto de dados (mas mesmo assim, é claro que há história genealógica!). Nesse caso, pode-se pensar em restrições seletivas ou então que os indivíduos amostrados sejam extremamente relacionados em termos de parentesco (mas, se são escolhidos randomicamente, isso é difícil acontecer). A interpretação vai depender tão-somente da genealogia, da qual não se tem qualquer certeza de que seja a verdadeira. Por isso, temos que enxergar genealogias como randômicas, da mesma forma que fazemos com as mutações (veja Capítulo 5). Na Figura 20.7, está apresentado um esquema mostrando quatro topologias geradas com o mesmo modelo—"coalescência padrão para amostras com N=10". As variações observadas refletem apenas o acaso.

20.7. Modelo de Mutações e Coalescência Neutra

Nesse modelo, assume-se que as mutações se acumulam nas linhagens como uma variação de Poisson com uma taxa μ por unidade de tempo *t* (gerações), bem como o modelo dos sítios infinitos (em que cada sítio pode sofrer mutação apenas uma vez). Considerando *t* gerações a partir de um ancestral comum entre duas linhagens, teremos S_2 representando o número de mutações que ocorreram nas duas linhagens e que apresenta distribuição de Poisson com esperança

$$E[S_2] = 2\mu t,$$

em que o número 2 está indicando que cada uma das linhagens acumula μt mutações. Se *t* é o número de gerações, vamos considerar *Tc* como o tempo total da genealogia, ou Tempo de Coalescência. Além disso, no modelo dos sítios infinitos, *S* representará o número de sítios que variam. Conjugando o que foi visto, podemos escrever,

$$E[S] = E[\mu Tc] = \mu E[Tc],$$

e se considerarmos *n* alelos e o tempo (momento) na genealogia em que há *i* linhagens – *T(i)* -, teremos

$$E[Tc] = \sum_{i=2}^{n} iE[T(i)] = 2N\sum_{i=2}^{n}\frac{1}{i-1},$$

e se conhecemos *E[S]* e *E[Tc],* é possível associá-los como

Figura 20.7. Árvores geradas empregando-se um mesmo modelo para o mesmo tamanho populacional, evidenciando a aleatoriedade subjacente (figura adaptada de Rosemberg e Nordborg, 2002).

$$E[S_n] = \mu E[Tc] = 2N\mu \sum_{i=2}^{n} \frac{1}{i-1},$$

e, resgatando o parâmetro Θ (=*2Nμ*, populações haplóides e *4Nμ*, para populações diplóides), teremos

$$2N\mu = \Theta = \frac{S_n}{\sum_{1=2}^{n} \frac{1}{i-1}},$$

ou seja, Θ é um ótimo meio de se estimar *2Nμ*!

Fica fácil perceber na última fórmula que, quanto maior a taxa de mutação (μ), maior a variação das sequências (S_n), o que também vale para o tamanho da população (*N*).

20.8. Coalescência e Recombinação

A recombinação é a responsável pelo fato de sítios ligados poderem apresentar árvores genealógicas distintas. Isso ocorre porque recombinações dividem uma linhagem em duas ou mais. Árvores genealógicas envolvendo recombinação nem sequer podem ser chamadas de árvores, mas preferencialmente de gráficos (ARG – *Ancestral Recombination Graph*).

Dois fatores estão envolvidos nos efeitos da recombinação: o *crossing over* por gene por geração e o tamanho da população. Para esses casos, é necessário um meio de estimar a taxa de recombinação, mas até o momento não existe um meio ideal. Nas Figuras 20.8 e 20.9, está representada a história evolutiva de três linhagens envolvendo apenas um evento de recombinação e suas consequências para as inferências genealógicas, respectivamente.

Ao lado da recombinação, a estrutura populacional e as flutuações no tamanho da população também podem alterar topologias genealógicas. A taxa de coalescência pode ser alterada por fatores como sucesso reprodutivo, estrutura etária da população e razão sexual. As dificuldades não param por aqui. A seleção natural é outro fator complicante, pois com ela alguns genótipos deixam mais descendentes que outros (por possuírem

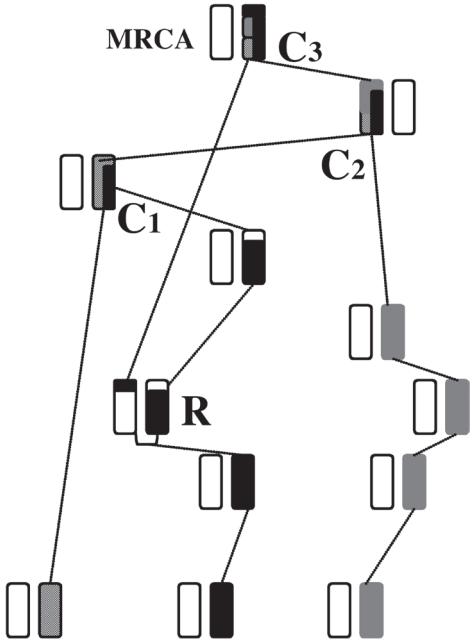


Figura 20.8. Efeito da recombinação na genealogia. MRCA: ancestral comum mais recente de todos os alelos destacados; C1, C2, C3: eventos de coalescência; R: evento de recombinação.

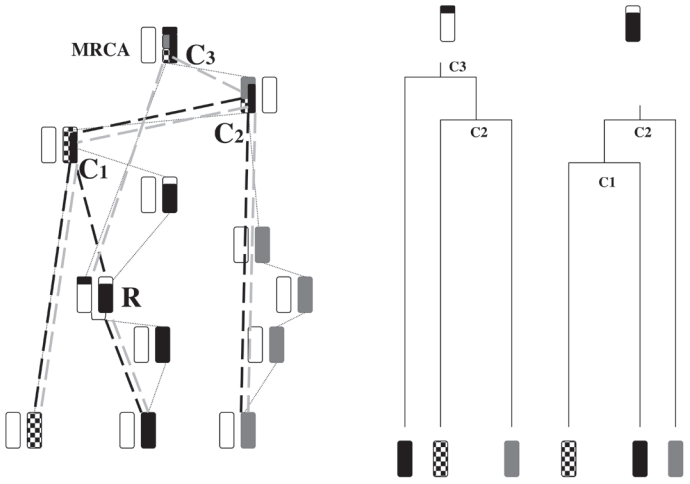


Figura 20.9. Topologias induzidas por cada uma das sequências originadas por recombinação. Siglas seguem as da figura anterior. Observar as diferenças flagrantes entre ambas.

maior valor adaptativo naquele momento evolutivo e ambiente), ou seja, as linhagens atuais não escolheram seus parentais e, nesse caso, muita informação é perdida. Outro processo evolutivo bastante complicador é a migração seguida de fluxo gênico…

20.9. Coalescência e Migração

Migrações podem ser uni-, bi- e multidirecionais, e sabemos que este último caso é bastante comum. Havendo fluxo gênico histórico entre populações migrantes e receptoras, a teoria analítica da coalescência encontra um grande problema. Porém, é possível simular algumas situações.

Tomemos dois casos: (a) duas populações de tamanho N em que haja alta taxa de migração bidirecional ao longo do tempo, com proporção constante de migrantes; e (b) duas populações de tamanho N em que haja baixa taxa de migração ao longo do tempo, com proporção constante de migrantes (Figura 20.10). Fica claro que, no primeiro caso, as frequências gênicas atuais serão similares interpopulacionalmente, ao passo que, no segundo caso, estas serão bastante diferentes. Ainda vale lembrar que excesso de migração pode interferir nas inferências genealógicas por mascarar divergências.

Em termos práticos, alguns cuidados devem ser tomados em análises genealógicas para que se evitem vieses ou erros de interpretação. Para isso, é recomendável: trabalhar com vários locos simultaneamente (pelo menos três); para inferência de migração, as sequências utilizadas devem ter um mínimo de 300 pb; para estimativas de recombinação, as sequências devem ter um mínimo de 1000 pb; e, de um modo geral, 20 indivíduos

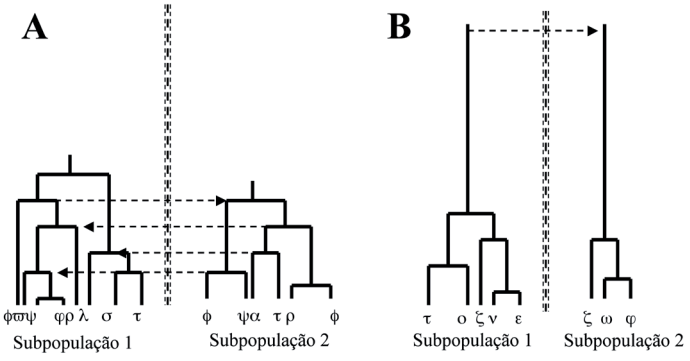


Figura 20.10. Genealogias representando possíveis consequências da troca alélica, devido à migração, entre duas subpopulações. A. Eventos recentes. B. Evento antigo. Note que eventos recorrentes e recentes podem levar a uma diminuição da diversidade alélica interpopulacional, enquanto que, em B, observa-se o contrário.

amostrados por população são suficientes para inferências genealógicas.

Em termos computacionais, existem vários programas com algoritmos robustos para análises de coalescência. Dentre eles, destacamos: LAMARC - *Likelihood Analysis with Metropolis Algorithm using Random Coalescence*; MIGRATE - *Estimation of migration rate and effective population size*; e RECOMBINE - *Metropolis-Hastings Markov Chain Monte Carlo Genealogy Sampler* (http://evolution.gs.washington.edu/lamarc/index.html). Todos eles estão disponíveis para Windows, Linux e Macintosh, e são comentados em maiores detalhes no Capítulo 21. Como todo e qualquer *software* empregado em análises genealógicas, alguns cuidados devem ser tomados para que as análises não fiquem "rodando" até o fim dos tempos. Entre eles, não efetuar muitas questões ao mesmo tempo, não estipular um modelo de análise restrito, pouca memória computacional (em alguns casos recomenda-se o uso de vários computadores em paralelo), muitos indivíduos sendo analisados ao mesmo tempo (como já mencionado, mais que 20 é desnecessário). Muitas vezes podemos dispor de mais de 20 amostras por população. Em situações como estas, recomenda-se a exclusão aleatória de sequências e não a exclusão sistemática de sequências similares.

20.10. Para lembrar

Alguns pontos fundamentais discutidos neste capítulo foram:

- A coalescência é definida em termos de tamanho populacional (N) e tempo (gerações);
- Não se pode medir o tempo "olhando" apenas para os genes, porém se pode medir divergências;
- Cálculos empregam equações reescalonadas em termos de N, tempo e taxa de mutação (μ);
- Não é possível estimar N, porém é possível estimar o parâmetro composto Θ;
- $\Theta = 4N\mu$ em populações diplóides e $2N\mu$ em populações haplóides;
- Informações adicionais podem dar maior consistência às inferências genealógicas.

20.11. Para Terminar

Dentre as muitas aplicações das análises genealógicas, queremos ressaltar que, nos dias atuais, com as condições ambientais deterioradas de nosso planeta, ecossistemas alterados, a poluição engolindo rios e mares, o efeito antrópico encolhendo áreas de fauna e flora e a desatenção com a saúde ambiental, informações genealógicas podem ser de grande valia para a detecção de estresses genéticos e ambientais em populações naturais, contribuindo para subsidiar programas de conservação de populações ameaçadas, buscando restituir uma vida melhor e mais equilibrada para a Terra.

Agradecimentos

É com grande carinho que agradeço a Sergio Russo Matioli por mais esta oportunidade científica, a Daniel Fernandes Bacellar, pelo constante estímulo à busca, ao meu grupo de pesquisa PANGEA, em especial a Antonio Marcio Martins Junior, André Andrade, Flávia Roberta Abbude, Ayling Martins Ng, Carlos Vaccari Gama, Fernanda Orpinelli, Leandro Benevides, Marcela Costa, Helber Crisnan Rodrigues, por me instigarem com questões preciosas sobre a vida, a Artur Trancoso Lopo de Queiroz pelos desafios, a José Geraldo de Aquino Assis por revisar carinhosamente este capítulo, a todos os meus bichos, que me dão alegria diária e vontade de viver.

Referências Bibliográficas

Felsenstein, J. (2003). **Inferring Phylogenies**. Sinauer, Sunderland, MA.

Fu, Y-X. e Li, W-H. (1999). Coalescing into the 21st century: an overview and prospects of coalescent theory. **Theor. Popul. Biol. 56**: 1-10.

Hudson, R.R. (1991). Gene genealogies and the coalescent process. **In** D Futuyma e J Antonovics (eds) **Oxford surveys in evolutionary biology**. Oxford University Press, Oxford.

Kingman, J.F.C. (2000). Origins of the coalescent. **Genetics 156**: 1461-1463.

Nei, M. (1987). **Molecular Evolutionary Genetics.** Columbia University Press, New York.

Ridley, M (2006). **Evolução**, 3 ed. Porto Alegre: Artmed.

Rosemberg N.A. e Nordborg, M. (2002). Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. **Genetics 6**: 380-390.

Wright, S. (1968). **Evolution and genetics of populations**. vol.1, University of Chicago Press, IL.

Capítulo 21

Análise filogeográfica

Haydée A. Cunha (haydeecunha@biologia.ufrj.br)
Laboratório de Biodiversidade Molecular, Departamento de Genética
Instituto de Biologia, Universidade Federal do Rio de Janeiro
Laboratório de Mamíferos Aquáticos, Faculdade de Oceanografia
Universidade do Estado do Rio de Janeiro

Antonio M. Solé-Cava (sole@biologia.ufrj.br)
Laboratório de Biodiversidade Molecular, Departamento de Genética
Instituto de Biologia, Universidade Federal do Rio de Janeiro
Rio de Janeiro, RJ

"*Much in evolution makes even more sense in the light of historical genealogy.*"
(Avise, 2000, em complemento a Dobzhansky, 1973)

21.1. Introdução

Filogeografia é o estudo dos princípios e processos, históricos e contemporâneos, que influenciaram a distribuição geográfica das linhagens genealógicas dentro de uma espécie ou entre espécies próximas (Avise *et al.*, 1987). Assim como as forças evolutivas (mutação, deriva genética, fluxo gênico e seleção natural), eventos demográficos (como colonização ou extinção de populações locais, expansão ou declínio populacional, e migração) deixam assinaturas na história genealógica das populações. A filogeografia visa decifrar essas assinaturas e entender como esses fatores moldaram a distribuição atual da variabilidade genética em uma espécie. Em síntese, a filogeografia preocupa-se com a história evolutiva de uma espécie, no espaço e no tempo.

A abordagem filogeográfica surgiu quando o biólogo John Avise, nos anos 1970, teve a idéia brilhante de analisar dados de polimorfismos de restrição (RFLP, veja Capítulo 18) do DNA mitocondrial de maneira qualitativa. John Avise estudava a genética de populações de roedores do gênero *Geomys* com alozimas e resolveu analisar os mesmos indivíduos usando RFLP do DNA mitocondrial. Também foi muito antes do advento da PCR, de modo que o estudo necessitava do isolamento das mitocôndrias e da purificação do DNA mitocondrial por ultracentrifugação em gradiente de cloreto de césio. A primeira observação importante dessa abordagem trabalhosa foi que o corte com enzimas de restrição do DNA das mitocôndrias de cada indivíduo não resultava em uma infinidade de bandas quando analisado por eletroforese. Isso significava que as mitocôndrias de cada indivíduo se comportavam de maneira clonal, ou seja, todas as mitocôndrias tinham basicamente o mesmo DNA. Por outro lado, existia variabilidade intrapopulacional nos padrões de restrição, condição necessária para o uso de uma molécula para estudos de genética populacional. O hábito de analisar genes com comportamento mendeliano inicialmente dificultava a abordagem do DNA mitocondrial, que é haplóide. John Avise teve, então, a idéia de usar para a análise uma abordagem semelhante à usada pelos antropólogos no estudo da herança dos sobrenomes humanos. A diferença é que os sobrenomes, na maior parte das sociedades, têm herança paterna, enquanto o DNA mitocondrial tem herança materna. Dessa forma, cada indivíduo foi tratado como uma *unidade taxonômica operacional* (UTO) independente e os dados não foram reduzidos a medidas de distância gênica entre populações, como feito com alozimas. Essa mudança de abordagem revolucionou os estudos populacionais desde então. Tratar indivíduos como UTOs permitiu analisá-los sem criar agrupamentos *a priori*. Além disso, como cada haplótipo passou a ser analisado individualmente, tornou-se possível o mapeamento das mutações (detectadas por polimorfimos de restrição) entre os indivíduos de forma muito semelhante à usada pelos cladistas nas análises filogenéticas com as quais John Avise já se tornara familiar na época. Houve, portanto, uma quebra de paradigma entre a análise populacional mendeliana (que era a única existente na época) e a análise de padrões mutacionais mapeados entre haplótipos de herança materna tratados de maneira independente. John Avise inovou também na maneira de apresentar os resultados: até então, os resultados de trabalhos de genética populacional eram visualizados na forma de árvores em que cada população ocupava uma posição terminal. Ao trabalhar com haplótipos, Avise decidiu visualisar as correlações entre indivíduos por meio de uma rede de haplótipos interconectados, em que as mutações eram indicadas nas linhas que ligavam os haplótipos mais semelhantes. O próximo passo, que também foi fundamental na criação dessa nova abordagem, foi a sobreposição da rede de haplótipos em um mapa com os locais onde os indivíduos haviam sido coletados (Avise, 1991). Anos mais tarde, esse novo campo de estudo foi denominado "filogeografia", por incorporar tanto dados das relações genéticas entre os indivíduos como seus posicionamentos geográficos (Avise *et al.*, 1987). Para um relato interessante sobre o amadurecimento da idéia da filogeografia, veja Avise (2006).

Duas características tornam os marcadores mitocondriais extremamente úteis para análises filogeográficas. A mais importante é que as sequências de genes ou regiões mitocondriais contêm informação genealógica, ou seja, elas possibilitam a reconstrução das linhagens genéticas das populações. A segunda vantagem é que, devido à herança exclusivamente materna e à ausência de recombinação, reconstruir a história genealógica de uma espécie a partir de marcadores mitocondriais é muito mais simples do que fazê-lo usando marcadores nucleares. Evidentemente, o marcador mitocondrial escolhido deve ter um nível de variabilidade adequado para a questão estudada. Inicialmente, o sequenciamento

de números grandes de indivíduos era muito custoso, o que limitava os estudos a espécies carismáticas ou economicamente importantes. Entretanto, durante os vinte anos passados desde o surgimento da filogeografia, o sequenciamento de DNA tornou-se cada vez mais acessível para estudos populacionais, de modo que a maioria dos estudos filogeográficos atuais se baseia na análise de sequências mitocondriais.

Recentemente, métodos de análise sofisticados foram desenvolvidos para aproveitar a quantidade de informação obtida com as sequências e os locos nucleares hiper-variáveis, como os microssatélites. Esses métodos utilizam as expectativas da Teoria da Coalescência, que foi formalizada por Kingman (1982) e Hudson (1991), e estabelece a base teórica que permite a reconstrução das genealogias a partir de alelos e sequências gênicas (veja Capítulo 20). Eles também usam abordagens de máxima verossimilhança ou Bayesianas para analisar simultaneamente vários locos gênicos e inferir parâmetros populacionais, como taxas de migração, tamanho populacional efetivo, sinal de crescimento ou declínio populacional, e tempo desde a separação das populações. Algumas dessas análises conseguem detectar o número de populações mais provável a partir dos dados sem qualquer informação *a priori* sobre a origem geográfica das amostras e também identificam o limite entre essas populações. Outras análises incorporam a informação geo-referenciada de coleta das amostras para definir populações. Os métodos desenvolvidos nos últimos anos prometem substituir as análises tradicionais de genética de populações, fornecendo parâmetros muito mais precisos que os que podiam ser estimados anteriormente.

As análises filogeográficas possibilitam uma compreensão mais ampla da história evolutiva das espécies, que é valiosa em qualquer estudo de genética de populações. Neste capítulo, são abordados os conceitos básicos usados na filogeografia. Também são apresentados os principais métodos de análise filogeográfica e discutidas suas vantagens em relação aos métodos tradicionalmente utilizados em genética de populações.

21.1.1.Abordagem filogeográfica x genética de populações clássica

Antes de tratar dos métodos filogeográficos, é importante considerar os métodos que eram utilizados anteriormente em Genética de populações. Esses métodos serão chamados, ao longo do texto, de métodos clássicos ou tradicionais, e são derivados dos trabalhos de Wright (1931, 1951, 1965) sobre a estruturação espacial da variação genética nas populações. Wright (1951) propôs três estatísticas sintéticas, chamadas estatísticas F, que descrevem a organização da variação genética em níveis hierárquicos em uma espécie. O índice de fixação F_{ST}, que mede a diferenciação genética entre populações, passou a ser amplamente utilizado em estudos de estruturação populacional (Weir e Cockerham 1984). Várias estatísticas análogas ao F_{ST}, adequadas às particularidades dos marcadores moleculares que surgiam, foram desenvolvidas, assim como testes da hipótese de panmixia a partir dessas estatísticas (*e.g.*, G_{ST}, Nei, 1978; θ, Weir e Cockerham, 1984; Φ_{ST}, Excoffier *et al.*, 1992; R_{ST}, Slatkin, 1995).

De acordo com a relação entre a diferenciação genética e o fluxo gênico prevista no equilíbrio migração-deriva ($F_{ST} = 1 / (4N_em + 1)$), é teoricamente possível estimar indiretamente o fluxo gênico entre populações a partir do F_{ST} (Slatkin 1985, 1987, Neigel 1997, 2002). Devido à dificuldade de estimar diretamente o fluxo gênico a partir de métodos ecológicos (tais como o método de marcação e recaptura), essa abordagem tornou-se popular. Entretanto, a estimativa do fluxo gênico a partir do F_{ST} baseia-se em uma série de pressupostos que são irreais em muitas situações, e pode ser tão imprecisa a ponto de tornar-se inútil (Bossart e Prowell, 1998a, 1998b; Whitlock e McCauley, 1999; Pearse e Crandall, 2004). A utilidade do próprio F_{ST} como descritor da estruturação genética também foi questionada, principalmente no caso de espécies ameaçadas, que frequentemente não estão em equilíbrio (Pearse e Crandall, 2004).

Nos últimos quinze anos, a redução do custo do sequenciamento e das genotipagens permitiu análises de números grandes de indivíduos em estudos populacionais. Como resposta ao acúmulo desses dados de alta resolução e também do aumento da capacidade de processamento dos computadores, foram desenvolvidos métodos analíticos que conseguem utilizar melhor a informação contida nos dados, além de relaxarem vários dos pressupostos assumidos pelos métodos convencionais (Emerson *et al.*, 2001; Pearse e Crandall, 2004). Além disso, os novos métodos fornecem estimativas muito mais precisas de parâmetros, como grau de diferenciação e fluxo gênico entre populações, e tamanhos populacionais efetivos. Por essa razão, eles são mais adequados aos estudos de conservação, onde parâmetros precisos são mais necessários, pois são usados na definição de táxons prioritários e das estratégias de manejo (Pearse e Crandall, 2004).

Ao mesmo tempo em que esses métodos surgiam, as limitações da abordagem tradicional, baseada no F_{ST}, ficaram mais evidentes. Uma importante desvantagem do F_{ST} (e de todos os seus análogos) é a redução da informação que acontece quando apenas as frequências são usadas, cujo resultado final é uma estimativa isolada ("estatística sintética"). Uma estimativa como o F_{ST} é incapaz de discriminar, por exemplo, entre populações isoladas que divergiram recentemente e populações que mantém alto fluxo gênico, ou entre populações que divergiram recentemente e apresentam baixa migração, e populações separadas há muito tempo, mas com muita migração (Templeton, 1998; Nielsen e Wakeley, 2001; Pearse e Crandall, 2004). Outros problemas com o F_{ST} resultam das premissas do modelo de ilhas no qual ele se baseia (Waples, 1998; Whitlock e McCauley, 1999; Taylor *et al., 20*00). O modelo de ilhas assume infinitas populações, todas com mesmo tamanho e mesma taxa de migração, sendo a migração simétrica e totalmente aleatória, além de ausência de seleção ou mutação. Ele também assume que as populações persistem indefinidamente e atingiram o equilíbrio deriva-migração. Todas essas premissas são irreais, principalmente nos estudos de conservação, pois as espécies ameaçadas frequentemente sofreram declínio e/ou fragmentação populacional (Pearse e Crandall, 2004).

A previsível violação de algumas das premissas pode ter efeitos graves sobre a estimativa do fluxo gênico (N_em) a partir do F_{ST} (Bossart e Prowell, 1998a; Whitlock e McCauley, 1999; Pearse e Crandall, 2004). Entretanto, apesar do reconhecimento das limitações do F_{ST}, ele ainda é considerado um bom estimador da estruturação por ser em geral bastante robusto frente às inevitáveis violações de suas premissas (Neigel, 2002). Ele também é mais adequado para análises com um número pequeno de marcadores nucleares, onde outras abordagens frequentemente são estatisticamente menos poderosas. Assim, o F_{ST} continua sendo usado com sucesso para análises populacionais. Além disso, seu uso permite que sejam feitas comparações com métodos mais recentes (Neigel, 2002).

21.1.2. Primeiros métodos filogeográficos

Por atuar na fronteira intra/inter-específica, inicialmente a filogeografia combinava métodos de genética de populações e de sistemática molecular. No primeiro estudo filogeográfico (Avise *et al.*, 1979), a sobreposição da rede de haplótipos mitocondriais com o mapa das localidades de coleta revelou uma grande divergência entre os ratos-toupeira do leste e os do oeste da região sudeste dos

Estados Unidos. Estudos subsequentes, com várias outras espécies terrestres e marinhas da mesma região, revelaram uma grande concordância nos padrões filogeográficos, correlacionados com o padrão de cobertura por gelo nas glaciações do Pleistoceno. Um dos primeiros estudos filogeográficos de espécies que ocorrem no Brasil foi realizado por Eizirik *et al.* (1998). Os autores usaram sequências da região de controle do genoma mitocondrial para investigar a filogeografia de dois felinos na América Central e do Sul: a jaguatirica (*Leopardus pardalis*) e o gato-maracajá (*L. wiedii*). As árvores filogenéticas construídas mostraram um padrão semelhante entre as duas espécies, refletindo subdivisões populacionais antigas. Foram observados quatro clados na filogenia da jaguatirica, e três clados na do gato-maracajá, com distribuições geográficas praticamente iguais. A partir do posicionamento dos clados nas árvores filogenéticas e de suas distribuições, os autores propuseram que as duas espécies teriam surgido no norte da América do Sul, colonizado a América Central e, posteriormente, o sul da América do Sul (Eizirik *et al., 199*8).

Em outro estudo, da Silva e Patton (1998) analisaram filogeneticamente sequências do gene mitocondrial citocromo b de 15 gêneros de roedores e marsupiais para testar hipóteses sobre a diversificação de espécies na Amazônia. Uma das hipóteses testadas foi a de "barreiras fluviais", proposta por Wallace em 1852, que explica a diversidade de espécies por eventos de vicariância causados pela formação dos rios da Bacia Amazônica, que teriam produzido as especiações. As filogenias de quatro pequenos roedores, construídas usando amostras de localidades ao longo e nas duas margens do Rio Juruá, são incompatíveis com essa hipótese, pois a maior diferenciação observada foi entre localidades de um mesmo lado do rio. Curiosamente, as filogenias das quatro espécies mostram uma divisão no mesmo local geográfico, que coincide com uma formação geológica conhecida como Arco Iquitos (ou Jutaí). Assim, é possível que a atividade tectônica tenha sido, direta ou indiretamente, responsável pela divergência das linhagens nas quatro espécies. O estudo também apoiou um cenário de diversificação mais antiga (Plioceno ou até Mioceno tardio) que tradicionalmente evocado (Pleistoceno tardio ou mais recente, da Silva e Patton, 1998).

Apesar dos exemplos apresentados acima, nem sempre os métodos filogenéticos conseguem detectar padrões filogeográficos. Isso acontece porque esses métodos foram desenvolvidos para o estudo das relações evolutivas entre espécies ou grupos supra-específicos e, por isso, têm dificuldades no nível intra-específico. Por exemplo, eles necessitam de um grande número de caracteres variáveis para conseguir determinar as relações filogenéticas e, em análises populacionais, os indivíduos apresentam muita similaridade em suas sequências. Outro problema é que os métodos filogenéticos representam a história evolutiva por meio de bifurcações, mas no nível populacional, um mesmo haplótipo pode originar vários outros (multifurcações). Nesse nível, também são mais frequentes as homoplasias e introgressões, que não são bem resolvidas em árvores filogenéticas. Por fim, os métodos filogenéticos não foram criados para lidar com situações em que ancestrais estão incluídos na amostra, mas em análises populacionais haplótipos ancestrais normalmente estão presentes (na verdade, espera-se que o haplótipo mais comum seja o ancestral de todos da população). As duas últimas limitações das árvores filogenéticas podem ser contornadas usando redes de haplótipos, que são uma representação gráfica mais conveniente e adequada da evolução no nível intra-específico (Posada e Crandall, 2001). Mas a observação direta de uma rede de haplótipos também pode ser insuficiente para revelar um padrão filogeográfico, porque populações em regiões distintas podem diferir quantitativamente, mas não qualitativamente (veja explicação na próxima seção).

21.1.3. Polimorfismo ancestral e separação das linhagens

Se acompanharmos a história das linhagens de uma população inicialmente panmítica a partir de um instante no qual ocorre subdivisão populacional, veremos que as linhagens presentes na população original se dividirão nas populações filhas (Figura 21.1). Em um primeiro momento, as populações filhas serão parafiléticas, porque alguns dos indivíduos presentes nelas serão mais próximos geneticamente aos da outra população do que dos da própria população. Com a extinção de algumas das linhagens ancestrais, as populações filhas passarão por estágios de parafiletismo até se tornarem, eventualmente, monofiléticas (quando os indivíduos de cada população compartilharem um ancestral mais recente do que o ancestral que compartilham com a outra população) (Tajima, 1983). Esse processo é chamado de separação das linhagens ("*lineage sorting*", em inglês) e termina quando as populações passam a ser monofiléticas entre si (separação completa das linhagens) (Avise, 2000).

O tempo requerido para que populações recém isoladas atinjam o monofiletismo recíproco é, em média, $4N_e$ gerações ($2N_e$ gerações para marcadores mitocondriais; Neigel e Avise, 1986, Avise, 2000). Durante esse tempo, as populações não serão monofiléticas devido ao compartilhamento de polimorfismos ancestrais. Por isso, métodos filogenéticos não são capazes de detectar padrões filogeográficos, nem estruturação genética. Enquanto a

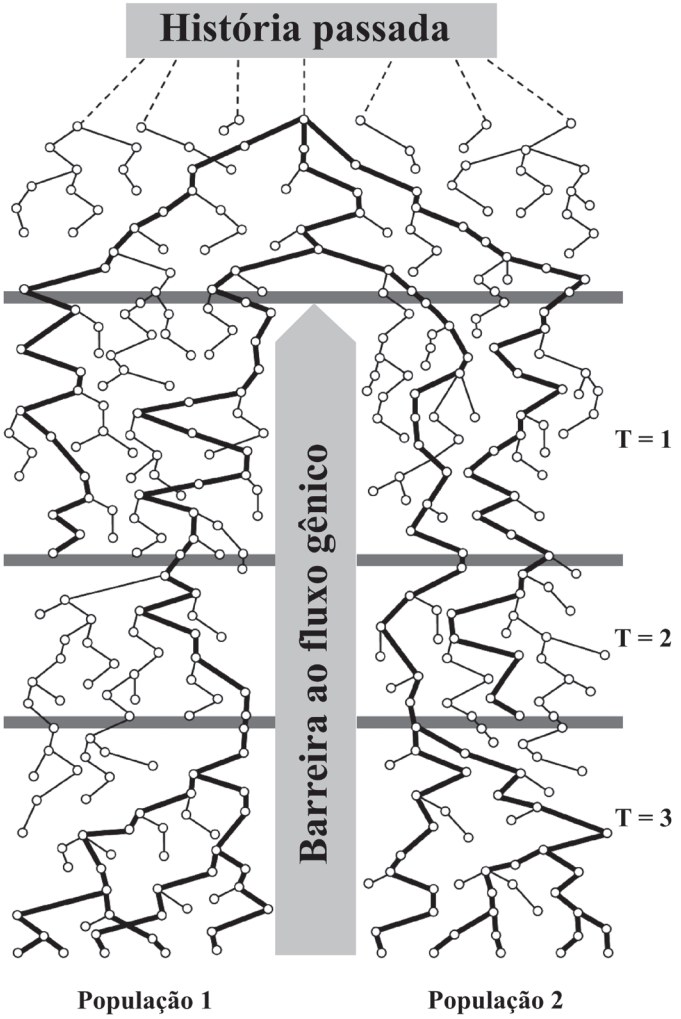


Figura 21.1. Processo de separação das linhagens ("lineage sorting") devido ao surgimento de uma barreira ao fluxo gênico em uma população originalmente panmítica. À medida que o tempo passa, as populações filhas (1 e 2) passam por estágios de parafiletismo (em T = 1 e 2), até se tornarem monofiléticas (T = 3), por causa da extinção de algumas linhagens. A separação das linhagens está completa quando as populações atingem o monofiletismo recíproco (modificado de Avise 2000).

separação das linhagens for incompleta, não haverá diferenças qualitativas entre as populações, embora diferenças quantitativas possam ser percebidas. Populações atualmente isoladas podem compartilhar haplótipos, mas diferir em suas frequências.

De acordo com o grau de separação genético e geográfico das linhagens, Avise *et al.* (1987) definiram cinco categorias filogeográficas, que integram um contínuo desde uma separação antiga, em que o monofiletismo recíproco claramente demonstra fragmentação alopátrica entre populações (categoria 1), até um sinal fraco de estruturação causado pela separação recente ou pela existência de fluxo gênico atual entre as populações (categorias 4 e 5, Figura 21.2).

Nas duas primeiras categorias filogeográficas, métodos filogenéticos tradicionais podem ser empregados para a construção de árvores de genes nas quais os indivíduos são posicionados nas UTOs. Nesses casos de separação mais antiga, aconteceu separação completa das linhagens no período entre as ramificações e, portanto, as árvores de genes (que mostram a separação das linhagens) e de populações (que refletem a separação das populações) são frequentemente iguais. Nas outras categorias, a probabilidade de que as linhagens não tenham se separado entre as ramificações é grande, porque o intervalo entre esses eventos é curto, o que leva a uma diferença entre a árvore de genes e de populações (Wakeley e Hey, 1997, Edwards e Beerli, 2000—veja próxima seção). Nesses casos, os métodos filogenéticos não conseguem recuperar a separação das populações, porque não há sinal filogenético, devido à ausência de diferenças fixadas ou ao pequeno número de caracteres variáveis. Ao longo deste capítulo, serão apresentados métodos filogeográficos adequados para o estudo de populações com todos os graus de divergência possíveis.

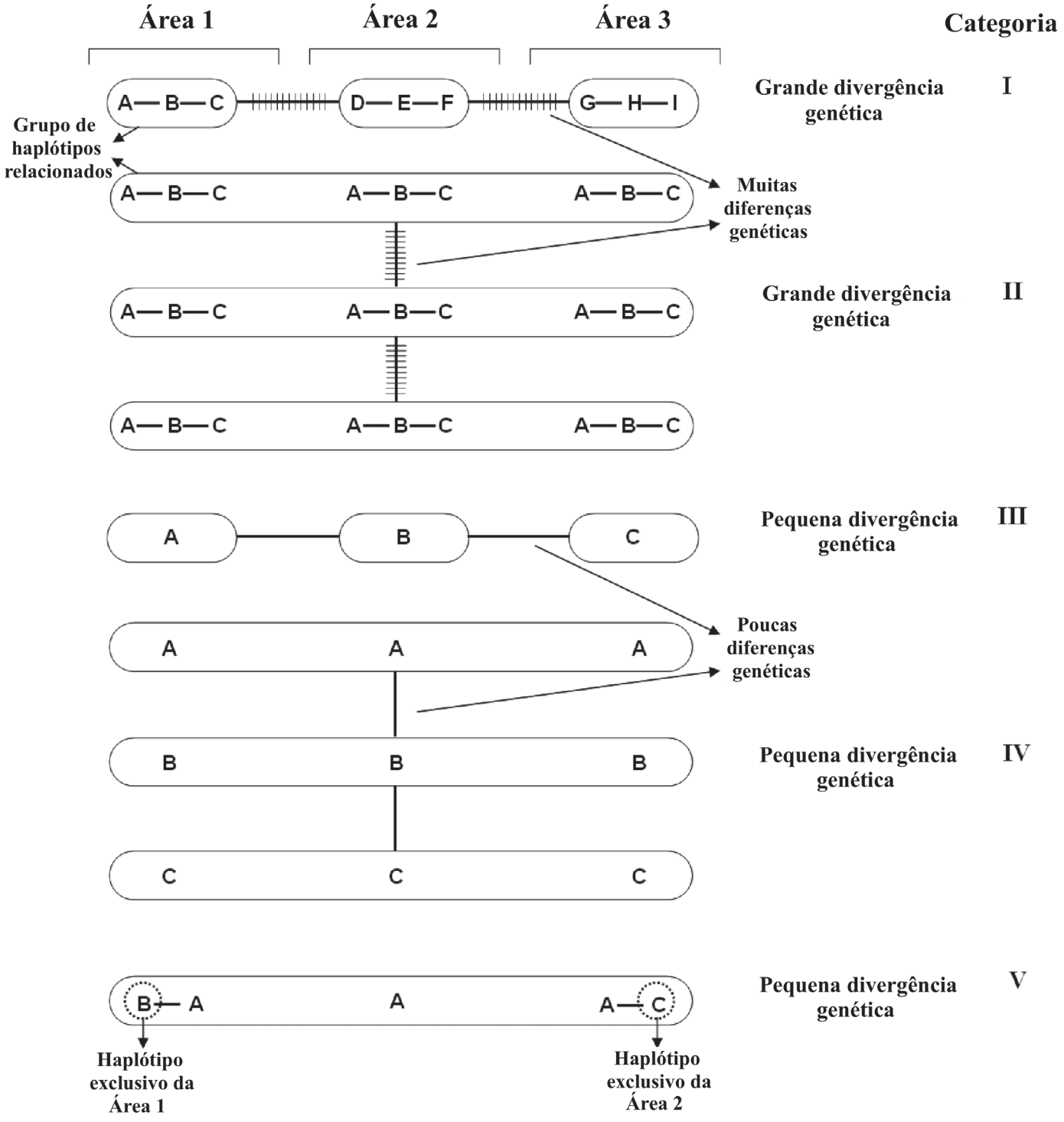


Figura 21.2. Categorias filogeográficas propostas por Avise *et al.* (1987), de acordo com o grau de divergência genética dos haplótipos e seus padrões de distribuição geográfica.

21.1.4. Abordagem genealógica e árvores de genes

Como grande parte dos estudos filogeográficos se encontra na interface da evolução dentro/entre espécies, é preciso utilizar uma abordagem alternativa aos métodos filogenéticos que seja adequada ao nível de divergência encontrado. O estudo da história evolutiva das populações deve adotar uma abordagem genealógica, na qual as linhagens da espécie estudada são reconstruídas. Mas a abordagem filogenética e a genealógica não diferem apenas por uma questão de escala de tempo: elas também retratam fenômenos diferentes. Ramificações em uma árvore de populações marcam eventos de fragmentação, enquanto as bifurcações nas árvores de genes correspondem a eventos de replicação (Avise, 1989, Maddison, 1997).

Como consequência dessa distinção, a separação das linhagens nas populações pode ou não corresponder temporalmente à divergência entre populações. No nível intra-específico, algumas separações de linhagens genealógicas acontecem antes e outras após a separação das populações (Figura 21.3, veja também o Capítulo 20). Assim, as árvores de genes não correspondem perfeitamente às árvores de populações. A discordância entre as duas pode ser provocada por separação incompleta das linhagens, por extinção, por hibridação ou outros fatores (Avise, 1989, Maddison, 1997). Em relação à separação incompleta das linhagens, quanto menor o intervalo entre os eventos de separação populacional, maior é a probabilidade de retenção de polimorfismos ancestrais entre as populações (Wakeley e Hey, 1997, Maddison, 1997). Além disso, como comentado antes, o processo de separação das linhagens é afetado pela demografia: quanto maior a população, maior o número de linhagens e maior a probabilidade de as populações compartilharem polimorfismos ancestrais, causando discordância entre a árvore de genes e a de populações (Maddison, 1997). Nesse caso, uma amostragem maior pode ser necessária para obter uma representação precisa da genealogia da espécie (Avise, 2000).

A perda das linhagens durante o processo de separação resulta da diferença nos tamanhos de prole entre as fêmeas, porque algumas não se reproduzem u geram apenas filhos machos, que não passam o DNA mitocondrial para as gerações seguintes. Ou seja, a separação das linhagens ocasionada pela extinção de algumas delas é um fenômeno probabilístico. A genealogia obtida depende de quais linhagens persistiram e, por isso, diferentes amostras e diferentes marcadores podem resultar em árvores genealógicas diferentes (Figura 21.4; Avise, 1989; Maddison, 1997). Ou seja, uma genealogia construída a partir de um loco gênico é um retrato isolado do passado, sujeito à estocasticidade do processo de separação das linhagens. A única forma de reduzir essa variância na reconstrução do passado da espécie é ampliar a amostragem de locos, porque cada um funciona como uma réplica independente do processo de coalescência (Edwards e Beerli, 2000; Arbogast *et al.*, 2002). Naturalmente, isso implica na análise de locos nucleares, pois o DNA mitocondrial não apresenta recombinação e, portanto, evolui conjuntamente como um único loco, mesmo quando são analisados vários genes mitocondriais diferentes.

Quando muito tempo se passou desde a divisão da população original, como em estudos entre espécies diferentes, a separação das linhagens para a maioria dos marcadores será completa e as árvores de genes serão equivalentes às árvores de espécies (Wakeley e Hey, 1997; Avise, 2004). Nesses casos, métodos filogenéticos são adequados. Devido à disparidade entre árvores de genes e de populações, análises filogeográficas baseadas em vários locos geram resultados mais confiáveis. Nos últimos anos, o campo da filogeografia tem mostrado uma tendência de uso de mais de um loco, acompanhada do desenvolvimento de métodos de análise multilocos. Também tem havido um esforço de desenvolvimento de análises que consideram as diversas genealogias possíveis e usam critérios para escolher as mais prováveis.

21.2. Filogeografia Estatística

Quando espécies que estão nas categorias filogeográficas 3, 4 e 5 são estudadas, deve-se optar por métodos de filogeografia estatística, que se dividem em dois grupos. O primeiro corresponde à Análise de Clados Hierarquizados (Nested Clade Analysis, NCA; ou Nested Clade Phylogeographic Analysis, NCPA—Templeton *et al.*, 1995; Templeton, 1998). Embora alguns autores não considerem a NCA como um método de filogeografia estatística (Knowles e Maddison, 2002, por exemplo), incluí-la nesse grupo de métodos é conveniente, porque a NCA foi o primeiro método filogeográfico a empregar testes estatísticos, ao contrário do método filogeográfico proposto inicial por Avise *et al.* (1987). Convém lembrar que até o surgimento da NCA, a análise filogeográfica era feita a partir da simples sobreposição de árvores ou redes de haplótipos sobre mapas e pela procura de padrões entre a filogenia e a geografia, que eram explicados por hipóteses *ad hoc*. Ou seja, era um método indutivista.

A NCA testa estatisticamente a associação entre a distribuição dos haplótipos em uma rede e suas distribuições geográ-

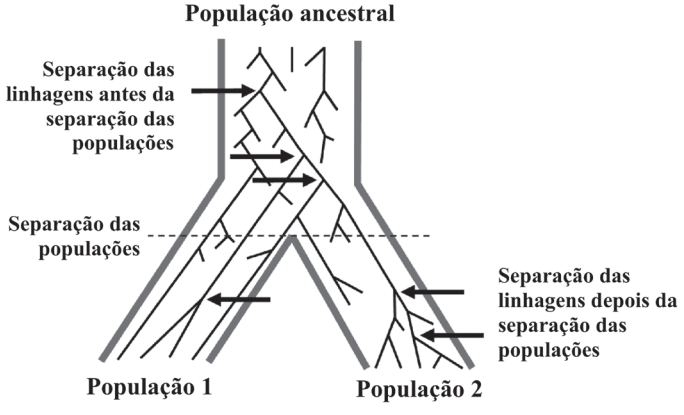


Figura 21.3. Distinção entre árvore de populações (desenhada com linha grossa) e árvore de genes (linha fina). A árvore de genes retrata eventos de replicação, enquanto a árvore de populações corresponde à interrupção do fluxo gênico (surgimento de populações diferenciadas). Por isso, pode haver separações de linhagens da árvore de genes anteriores e posteriores à separação das populações (linha tracejada) (modificado de Avise 2000).

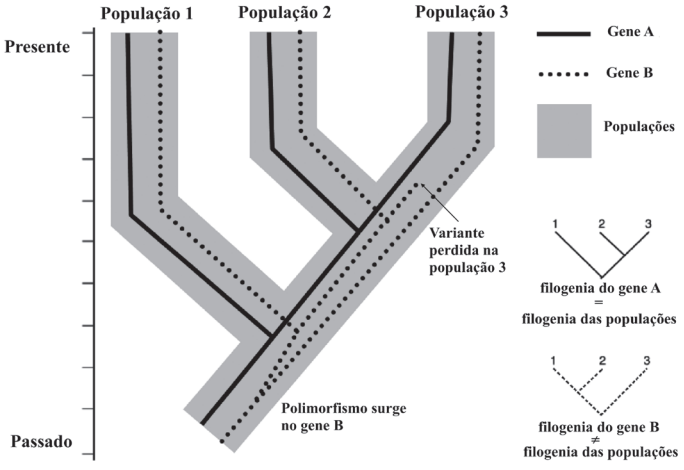


Figura 21.4. Discordância entre dois marcadores. Diferenças no processo de separação das linhagens de dois genes levam à reconstrução de diferentes árvores a partir dos mesmos indivíduos amostrados. A árvore do gene B não corresponde à árvore de populações (modificado de Bradley 2008).

ficas. Em primeiro lugar, é construída uma rede de haplótipos com o emprego do critério de máxima parcimônia (veja Capítulo 12), e aqueles que diferem por apenas um passo mutacional são agrupados em clados, que por sua vez são agrupados, um nível acima, com outros clados aos quais se ligam por apenas um passo mutacional etc., até que todos os clados sejam reunidos. Esse arranjo hierarquizado introduz um eixo temporal à análise, no qual clados de nível mais alto são mais antigos e os haplótipos (clados de nível zero) são a camada mais recente. Nos clados em que é encontrada associação significativa entre a distribuição genética e a geográfica, usam-se previsões baseadas na teoria da genética de populações e na teoria da coalescência para inferir as possíveis causas dos padrões encontrados. As inferências, feitas com auxílio de uma chave dicotômica, dividem-se em fragmentação alopátrica, aumento da distribuição e isolamento por distância (Templeton, 1998, 2001).

A grande vantagem dessa análise é que ela consegue separar o componente histórico (fragmentação alopátrica e aumento da distribuição) do atual (restrições ao fluxo gênico no presente). Outro ponto positivo é que a NCA é capaz de detectar inadequações do tamanho amostral, quando a análise é incapaz de produzir resultados conclusivos. Essas imperfeições da amostragem podem então ser corrigidas. Por fim, a NCA não requer hipóteses *a priori* (Templeton, 1998, 2001). Por isso mesmo, ela não pode ser considerada como um método completamente hipotético-dedutivista, já que as inferências não são testadas especificamente. O que a NCA testa é a hipótese nula de não associação entre a distribuição geográfica dos haplótipos e suas posições na rede. As inferências feitas não podem ser aceitas como hipóteses testadas. Assim, a NCA é uma análise exploratória e suas inferências devem ser testadas especificamente *a posteriori*.

Alternativamente à NCA, existem métodos baseados em modelos, que dependem de hipóteses *a priori* sobre o passado demográfico e evolutivo de uma espécie, assim como de eventos biogeográficos que podem tê-los influenciado. Esses métodos usam estratégias Bayesianas ou de máxima-verossimilhança para verificar a probabilidade de ocorrência de cada modelo alternativo de acordo com os dados observados—que serão apresentados mais adiante.

21.2.1. A Análise de Clados Hierarquizados (NCA)

O primeiro passo da NCA é a construção de uma rede de haplótipos a partir do método de 95% de parcimônia de Templeton *et al.* (1992), usando o programa *TCS* (Clement *et al.*, 2000). Nesse método, a rede é construída respeitando a probabilidade cumulativa (≥95%) de que todas as diferenças entre cada duas sequências sejam produto de substituições únicas (estado parcimonioso). Essa precaução é necessária porque, quando dois haplótipos são muito divergentes, aumenta a probabilidade de que eles compartilhem nucleotídeos não por ancestralidade comum, mas por convergências e paralalelismos. Assim, quando haplótipos muito diferentes são analisados, a probabilidade de que cada substituição nucleotídica entre dois haplótipos seja o resultado de um evento mutacional único diminui. Na abordagem de Templeton *et al.* (1992), quando existe uma probabilidade maior que 5% de que cada evento mutacional não seja único, a rede de haplótipos é rompida e os haplótipos muito divergentes são colocados fora dela.

Nem sempre as redes de haplótipos são únicas: algumas vezes existem redes alternativas, com o mesmo número de passos. Dizemos, então, que existe uma ambiguidade, que é representada em uma rede de consenso em que aparecem todas as conexões igualmente parcimoniosas. Essas ambiguidades na rede podem ser resolvidas de acordo com as regras descritas em Crandall e Templeton (1993). Basicamente, são usados três critérios: frequência, topologia e geografia. O critério de frequência corresponde à expectativa de que haplótipos mais frequentes sejam mais antigos e, portanto, teriam tido mais tempo para evoluir e dar origem a outros haplótipos, do que haplótipos pouco frequentes (e recentes). Assim, quando há incerteza sobre a conexão de um haplótipo com outros dois, pode-se resolver essa ambiguidade assumindo que o haplótipo em questão deve ter surgido a partir do mais frequente (ou seja, mais antigo), dentre as opções mostradas na rede. O critério da topologia é parecido, pois assume que haplótipos mais antigos tendem a ser mais internos na rede, pois teriam tido mais tempo para sofrer mutações que teriam gerado outros haplótipos. De fato, frequentemente se atribui a conformação em forma de estrela a um evento de expansão no qual vários haplótipos teriam surgido a partir do haplótipo central, ancestral (Slatkin e Hudson, 1991). Essa observação também auxilia na resolução das ambiguidades, pois se espera que o haplótipo sobre o qual há dúvida se ligue ao haplótipo mais interno na rede, dentre os possíveis. Por fim, o critério da geografia prevê que haplótipos geneticamente próximos provavelmente se localizam em uma mesma área geográfica, em vez de estarem em áreas distantes.

Se não for simples resolver as ambiguidades da rede seguindo as regras explicadas acima, aconselha-se fazer a análise de cada rede alternativa independentemente (Templeton *et al.*, 1995; Pfenninger e Posada, 2002; Brisson *et al.*, 2005). Algumas vezes, as ambiguidades não afetam o esquema de hierarquização dos clados e, portanto, não interferem na análise.

É possível usar outros algoritmos para a construção da rede de haplótipos. Por exemplo, Cassens *et al.* (2005) testaram quatro métodos com dados de sequência (parcimônia estatística – programa *TCS*; distâncias mínimas – programa *Arlequin*; e agrupamento médio – programa *Network*) e concluíram que o método de união de redes de parcimônia, por eles desenvolvido, pode funcionar melhor em alguns casos. Dessa forma, pode ser interessante explorar outros programas de construção de redes a fim de verificar diferenças entre as redes obtidas com outros métodos, e escolher a mais adequada.

Para prosseguir com a NCA, o desenho dos clados é traçado (manualmente) sobre a rede obtida. Os clados são sequencialmente hierarquizados desde o nível dos haplótipos (nível 0) até o nível mais alto, cada clado separado por uma substituição (Templeton, 1998, 2001). Os clados são nomeados usando a notação x-y, na qual x denota o nível do clado e y, sua identidade.

A associação geográfica dos clados hierarquizados é testada estatisticamente com o programa *GeoDis* (Posada *et al.*, 2000). Esse programa calcula a distância entre as localidades de coleta a partir de suas coordenadas geográficas. Alternativamente, uma matriz par-a-par das distâncias geográficas entre os centros das localidades de coleta pode ser usada, em vez das coordenadas geográficas, no caso de espécies distribuídas ao longo de um ambiente unidimensional (costeiras ou fluviais, por exemplo), ou quando há barreiras conhecidas para a dispersão dos indivíduos.

O teste da existência de associação entre a distribuição geográfica e genética dos clados é feito usando duas estatísticas: a distância do clado (D_c) e a distância do clado hierarquizado (D_n). D_c mede o alcance geográfico do clado (ou haplótipo) e D_n mede como um clado está geograficamente distribuído em relação a outros clados no mesmo nível hierárquico. Essas estatísticas são calculadas apenas para os clados que contêm variação genética e geográfica. O cálculo do D_c e do D_n emprega as distâncias geográficas (fornecidas em uma matriz ou calculadas pelo programa a partir das coordenadas geográficas, usando a fórmula para distâncias sobre um grande círculo) e as frequências dos haplótipos em cada localidade. As fórmulas estão explicadas em Templeton *et al.* (1995) e descritas detalhadamente em Posada *et al.* (2006).

Na NCA, também é usado o contraste entre clados internos e externos, que fornece informação sobre processos históricos. Assim, também são calculados D_c e D_n do contraste interior/ponta.

A significância estatística de cada valor de D_c e D_n é testada pelo método de Monte Carlo, por aleatorizações. A tabela de dados para cada clado e nível hierárquico é aleatorizada, mantendo-se as frequências dos haplótipos e tamanhos amostrais, de maneira que D_c e D_n são recalculados após cada aleatorização, a fim de construir suas distribuições nulas. Para um nível de significância de 0,05, devem ser usadas 1000 aleatorizações (Posada *et al.*, 2000, 2006).

Uma das críticas à NCA baseia-se na observação de um excesso de resultados significativos (falsos positivos) de até 70% em avaliações do método usando dados simulados de populações panmíticas (Knowles e Maddison, 2002; Panchal e Beaumont, 2007; Knowles, 2008). Esses estudos relatam um número excessivo de inferências de restrição ao fluxo gênico por isolamento por distância e de expansão da distribuição. Templeton (2008) defende que os falsos positivos observados podem ser minimizados com uma correção para testes múltiplos, como a de Bonferroni, Bonferroni sequencial (Rice, 1989) ou correção por permutação de Westfall e Young (1993). Assim, usando a correção de Bonferroni simples, os valores de P dos clados significativos devem ser multiplicados pelo número de testes realizados, que é igual ao número de clados analisados pelo *GeoDis*. Os novos valores de P são comparados com o nível de significância de 0,05 e aceitos se forem inferiores a ele. A correção sequencial de Bonferroni é um pouco menos rigorosa na detecção de falsos positivos, mas tem menor efeito no poder do teste (Rice, 1989; Verhoeven *et al.*, 2004). Nela, os valores de P são ordenados por ordem de maior significância e cada P é multiplicado pelo número de testes realizado até o momento no qual aquele valor de P foi calculado, em vez do número de testes total da análise (Rice, 1989). A correção de Westfall e Young (1993) é similar à sequencial, mas uma distribuição dos valores de P, obtida por reamostragem, é usada para correção dos valores de P observados. Uma outra abordagem promissora, que talvez represente o caminho a ser seguido para a NCA, é o uso simultâneo de locos gênicos diferentes e independentes. Espera-se que as associações espúrias que resultam de erros estatísticos do tipo I (rejeitar a hipótese nula da panmixia quando ela é verdadeira) em uma análise não devem se repetir em árvores com genes diferentes. Assim, a análise concomitante dos resultados de NCA de locos diferentes permite detectar as associações falsas, para reter apenas aquelas concordantes nas análises.

Os resultados significativos da análise de clados hierarquizados são interpretados usando a chave de inferência mais recente, que está disponível no sítio do programa *GeoDis*. A interpretação é feita a partir dos valores de D_c e D_n dos clados e do contraste interior/ponta que foram significativamente maiores ou menores que o esperado e da localização geográfica dos haplótipos de cada clado analisado. Informações sobre a biologia da espécie podem ser requeridas em alguns passos da chave, assim como testes adicionais para confirmar algumas inferências.

O programa *ANeCA* (Panchal, 2007) foi desenvolvido para automatizar a NCA e possibilitar testes do seu desempenho. Apesar de o *ANeCA* realizar todas as etapas da NCA de forma automática, sua utilização não é recomendada, já que a análise envolve diversos passos nos quais o pesquisador deve fazer escolhas e melhores decisões dependem da familiaridade com os dados e etapas da análise. Entretanto, pode ser interessante usar o programa posteriormente para comparar a análise "semi-manual" com a automática, detectando possíveis diferenças.

Um bom exemplo do uso da NCA foi o estudo filogeográfico da subespécie de mandioca selvagem *Manihot esculenta flabellifolia*, progenitora da mandioca domesticada, que ocorre no ecótono floresta-cerrado (Olsen, 2002). A NCA a partir de um gene nuclear (*G3pdh*) inferiu processos históricos que explicariam a configuração atual da variabilidade genética dessa subespécie no Brasil. Nos três níveis da análise, foram obtidas inferências de fluxo gênico restrito devido a isolamento por distância, mostrando que esse foi um processo recorrente no passado evolutivo da mandioca selvagem. Além disso, clados dos níveis mais recentes (1 e 2) indicaram uma fragmentação recente entre as áreas do extremo nordeste (norte do Tocantins) e noroeste (Acre e Rondônia) da distribuição da espécie, hoje separadas por uma barreira vegetacional (Floresta Amazônica). Essa conexão recentemente rompida entre as duas áreas explicaria tanto a similaridade genética entre elas, que é incongruente com a distribuição atual, quanto a maior diversidade que elas apresentam, compatível com uma origem mais antiga em relação às demais populações. A partir dos resultados da NCA, o autor propôs um cenário evolutivo no qual a distribuição original da mandioca selvagem se estendia entre essas duas áreas extremas, até ser fragmentada pela expansão da Floresta Amazônica sobrejacente, que teria causado o deslocamento de populações do meio da distribuição para áreas mais ao sul (Goiás e sul do Mato Grosso), possivelmente durante o último máximo glacial (há 18.000 anos; Olsen, 2002).

Conforme discutido anteriormente, as inferências da NCA devem ser tratadas como hipóteses que precisam ser testadas. Essas inferências dividem-se em três grupos: fragmentação alopátrica, expansão da distribuição e isolamento por distância. Cada inferência da análise deve ser testada usando métodos independentes da NCA. A seguir, são discutidos possíveis testes dessas hipóteses.

A fragmentação alopátrica inferida pela NCA pode ser antiga e corresponder a um evento de especiação. Templeton (2001) detalha um procedimento para a detecção de especiações usando a NCA. Um exemplo do uso da NCA com esse objetivo foi a investigação da situação taxonômica dos golfinhos do gênero *Sotalia*, até então considerados uma única espécie (*Sotalia fluviatilis*), com um ecótipo fluvial e outro marinho. A NCA de sequências da região controle mitocondrial detectou uma fragmentação alopátrica antiga entre a linhagem fluvial e a marinha, compatível com a existência de duas espécies (Figura 21.5A). Essa hipótese foi testada também com métodos filogenéticos (Máxima-Verossimilhança, Neighbour-Joining e Parcimônia), que demonstraram que os animais marinhos e fluviais estão separados em grupos reciprocamente monofiléticos (Figura 21.5B). A partir desses resultados e de diferenças morfológicas e ecológicas previamente conhecidas, a conclusão foi que os golfinhos marinhos e fluviais constituem espécies diferentes (*S. guianensis* e *S. fluviatilis*, respectivamente; Cunha *et al.*, 2005).

Quando inferida em um nível mais recente de diferenciação, a fragmentação alopátrica corresponde à diferenciação das populações com uma descontinuidade geograficamente abrupta entre elas (Figura 21.6) e pode ser verificada com testes sequenciais de subdivisão populacional.

A análise preferencialmente utilizada para o estudo da estruturação populacional com dados de sequência é a Análise de Variância Molecular (AMOVA, Excoffier *et al.*, 1992), realizada no programa *Arlequin* (Schneider *et al.*, 2000). Esse programa computa as estatísticas Φ, análogas às estatísticas F (Wright, 1978), que incorporam a informação sobre a distância molecular para separar a variância molecular em níveis hierárquicos e testar diferentes hipóteses de estruturação. Dessa maneira, a fragmentação indicada pela NCA pode ser testada como cenário de estruturação. Embora a AMOVA tenha maior sensibilidade para detectar diferenciação populacional recente que os métodos tradicionais, ela também não é capaz de identificar eventos do

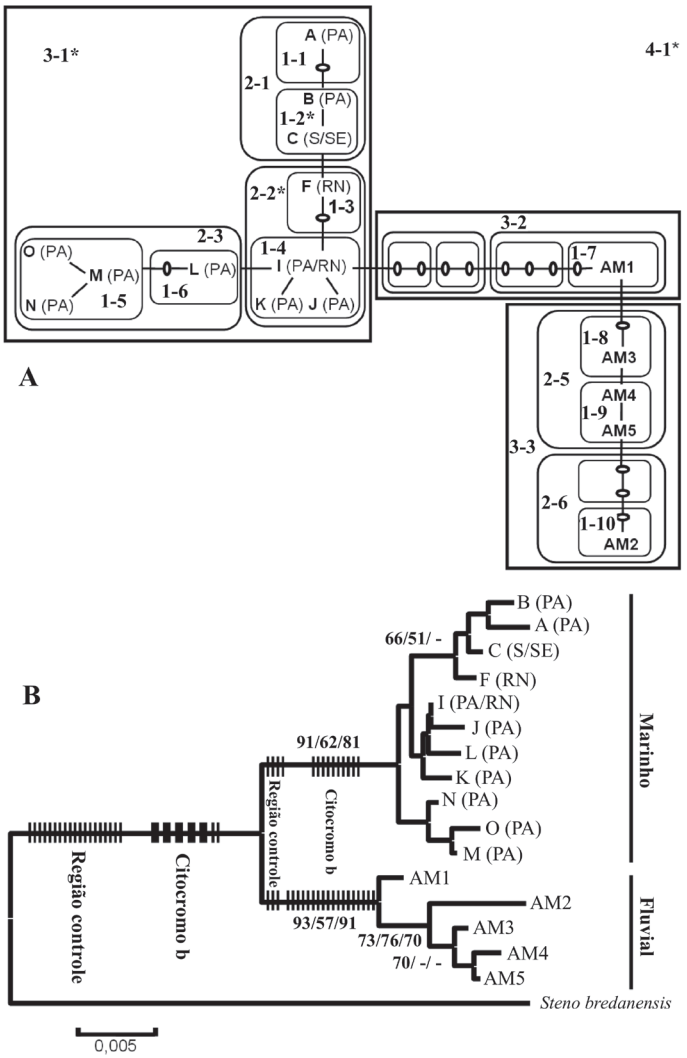


Figura 21.5. A. Rede de haplótipos da região controle de *Sotalia guianensis* (haplótipos A a O) e *S. fluviatilis* (haplótipos AM), com o desenho de clados hierarquizados. Ovais representam intermediários faltantes. Clados significativos ($P < 0,05$) estão marcados com um asterisco. Os códigos das localidades de coleta estão entre parênteses. O nível hierárquico está denotado como 1-x para o primeiro nível, 2-x para o segundo etc., onde x identifica o clado. B. Árvore de Neighbour-Joining (NJ) (distância p) entre os haplótipos da região controle de *S. guianensis* e *S. fluviatilis*. Máxima-Verossimilhança (ML) e Parcimônia (P) recuperaram a mesma topologia. Os valores de *bootstrap* (NJ/ML/P) maiores que 50% são mostrados. Sinapomorfias hipotéticas dos haplótipos da região controle e do citocromo b dos grupos marinho e fluvial são indicadas pelos traços verticais. O traço mais grosso representa dez sinapomorfias.

passado evolutivo da espécie e requer a atribuição *a priori* dos indivíduos às populações para o teste de estruturação genética (Templeton, 1998, Pearse e Crandall, 2004). Essa atribuição normalmente é feita por localidade de coleta e frequentemente é baseada em critérios logísticos, mais do que hipóteses *a priori* de diferenciação, de modo que os agrupamentos acabam sendo definidos *a posteriori* a partir das suas significâncias. Se é feita uma análise de NCA inicial, os agrupamentos detectados podem ser testados como hipótese *a priori* na análise de variância molecular.

Uma análise relacionada é a SAMOVA, ou Análise Espacial de Variância Molecular, realizada pelo programa de mesmo nome (Dupanloup *et al*., 2002). O programa encontra o melhor cenário de estruturação dentre os possíveis considerando agrupamentos das localidades de coleta adjacentes. Isso é feito por uma AMOVA de todas as combinações de k populações (grupos de localidades), sendo que o usuário estipula o número de populações

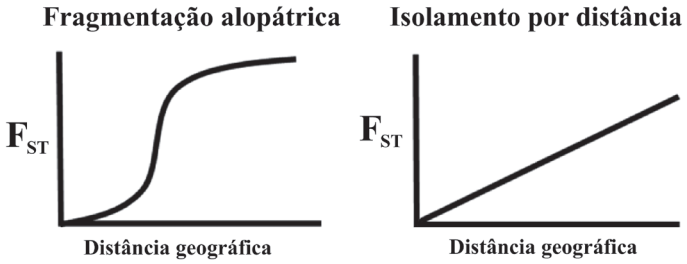


Figura 21.6. Diferenciação entre estruturação causada por isolamento por distância e por uma descontinuidade abrupta, como a causada por fragmentação alopátrica. As distâncias geográficas são calculadas a partir de um ponto em um dos extremos da distribuição.

a ser testado em cada análise e o programa informa a combinação com maior índice de fixação significativo. O SAMOVA usa as coordenadas geográficas das localidades para determinar quais são adjacentes e estabelecer os grupos possíveis. Como na AMOVA, a significância dos índices de fixação é testada pela permutação dos haplótipos, indivíduos ou populações entre indivíduos, populações ou grupos de populações. Após cada permutação, os índices são recalculados para obtenção de suas distribuições nulas e pelo menos 1000 permutações devem ser usadas para um nível de significância de 0,05. Subsequentemente à AMOVA ou SAMOVA a partir de sequências mitocondriais, o cenário de estruturação pode ser reavaliado usando outros marcadores e métodos (como, por exemplo, microssatélites e métodos de agrupamento). Apesar da abordagem SAMOVA ter a vantagem de comparar todos os agrupamentos possíveis de populações, seu resultado continua sendo uma hipótese *ad hoc* de estruturação. A associação com os resultados da NCA permite uma inferência mais robusta da estruturação populacional, por representarem abordagens radicalmente diferentes, apesar de usarem os mesmos dados.

Outra inferência possível da NCA é a expansão da distribuição, que pode corresponder a um aumento da distribuição ou a um aumento populacional. Essas duas hipóteses podem ser testadas usando uma Análise da Distribuição das Diferenças ("Mismatch Distribution Analysis", Rogers e Harpending, 1992), que pode ser feita no programa *Arlequin*. O princípio dessa análise é o fato de que populações que experimentam crescimento populacional súbito apresentam uma distribuição unimodal do número de diferenças entre pares de sequências, enquanto populações estáveis mostram distribuições multimodais. O *Arlequin* compara a distribuição observada com a esperada sob a hipótese nula de crescimento populacional e verifica se as duas são estatisticamente diferentes a partir da comparação da soma dos quadrados dos desvios. Se a diferença não for significativa, os dados são compatíveis com um cenário de expansão demográfica. Mais recentemente, Ray *et al.* (2003) demonstraram que a expansão geográfica também pode causar uma distribuição unimodal, de modo que o programa *Arlequin* passou a testar as duas hipóteses (de aumento demográfico ou da distribuição geográfica). As duas hipóteses são interessantes para testar a inferência de expansão feita inicialmente pela NCA.

A expansão populacional também pode ser verificada por testes de neutralidade que, como o nome indica, foram criados para testar a neutralidade dos marcadores, mas funcionam bem na detecção de sinais de expansão populacional. Os mais usados são os que usam os índices F_S (Fu, 1997) e o D (Tajima, 1989), ambos executáveis pelo *Arlequin*. Além desses, outros índices usados em testes de crescimento populacional são o F* e o D* (Fu e Li, 1993), calculados no programa *DnaSP* (Rozas *et al., 20*03). Os programas também obtêm a significância dos índices por aleatorizações, sendo que valores significativos mostram

evidência de expansão. Para o F_S, o nível de significância deve ser de 0,02 e valores negativos significativos são indício de crescimento populacional (Fu, 1997).

A última inferência que pode resultar da análise pela NCA é a de isolamento por distância, que sugere que as populações estão estruturadas de acordo com esse modelo, no qual o fluxo gênico entre as populações diminui com o aumento da distância geográfica. Essa hipótese pode ser verificada com um teste de Mantel. Como a divergência genética entre as populações pode aumentar com a distância geográfica, mesmo quando elas não estão estruturadas, é interessante poder testar a hipótese de isolamento por distância apenas quando há um indício independente de sua ocorrência, como quando a NCA faz essa inferência. No teste de Mantel, é verificada a correlação entre duas matrizes: de distância genética e geográfica entre as populações. As matrizes são aleatorizadas e é calculada a correlação entre elas a cada aleatorização. Se menos de 5% das correlações aleatórias forem maiores que a observada, existe evidência de isolamento por distância. O teste de Mantel pode ser feito em vários programas, dentre eles o *Arlequin*, o *FSTAT* (Goudet *et al.*, 1995) e o *SPAGeDi* (Hardy e Vekemans, 2002).

A NCA tem sofrido críticas desde que surgiu, no fim dos anos 90. Algumas deficiências da análise foram apontadas durante esse intervalo e ela passou por pequenas modificações, dentre as quais a inclusão de alguns testes suplementares (Templeton, 2001, 2004, 2008). As principais críticas à NCA são: a ocorrência de falsos positivos quando conjuntos de dados simulados são testados (Knowles e Maddison, 2002; Panchal e Beaumont, 2007), o fato de se basear em uma determinada rede de haplótipos (em vez de considerar todas as genealogias possíveis) de um único marcador (Wakeley, 2003; Knowles e Maddison, 2002) e sua relativa subjetividade (Knowles e Maddison, 2002). Templeton (2004, 2008) tem discutido as críticas, usando-as para refinar a análise ou rebatendo-as. Apesar de haver opositores da NCA, ela é o método filogeográfico mais popular atualmente: 700 artigos citaram o artigo de Posada *et al.* (2000) em que é feita a descrição do programa *GeoDis*, o que sugere que grande parte desses usaram a análise (dados obtidos no Web of Knowledge, em novembro de 2010).

O problema dos falsos positivos e uma possível solução para ele já foram comentados acima. Knowles e Maddison (2002) e Panchal e Beaumont (2007) relataram um excesso de resultados significativos levando a falsas inferências de isolamento por distância e fragmentação alopátrica. Templeton (2008) frisa que esses estudos se basearam em dados simulados, que não retrataram situações reais ou violaram limites previamente reconhecidos da análise ou seus pressupostos. Em uma avaliação feita a partir de conjuntos de dados reais com 150 expectativas *a priori* de fragmentação alopátrica e de expansão geográfica, a NCA teve um bom desempenho (Templeton, 2004). Porém, ainda não está descartada a possibilidade de que a NCA gere algum excesso de inferências de expansão da distribuição falsas: algumas das inferências de expansão obtidas nesse estudo podem ser verdadeiras, mas não esperadas, o que teria inflado artificialmente a taxa de falsas inferências (Templeton, 2004, 2008). Além de sugerir que uma correção para testes múltiplos seja aplicada, Templeton (2008) propõe duas formas de testar se inferências de isolamento por distância e expansão da distribuição são falsas, pela verificação de padrões de agrupamento temporal dessas inferências a partir de um conjunto de dados. Esses testes baseiam-se na previsão de que as expansões, sendo eventos localizados no tempo, devem agrupar-se temporalmente em um ou poucos momentos, enquanto o isolamento por distância não deve se restringir a alguns momentos, pois não é um evento discreto. Se as inferências forem falsas, ambas devem apresentar o mesmo padrão de distribuição uniforme ao longo do tempo.

O fato de ser uma análise baseada em um único loco é uma deficiência da NCA—compartilhada por todas as análises feitas a partir de um único marcador— que deixa de ser um problema importante se a NCA for reconhecida como exploratória, porque suas inferências podem ser posteriormente avaliadas usando outros marcadores. Quando é feita a NCA de diferentes marcadores, devemos aceitar apenas as inferências em comum entre eles, como uma forma de "validação cruzada" (Templeton, 2002, 2008).

Por fim, devido às várias etapas manuais, alguma subjetividade é introduzida na análise de NCA, o que não é algo tão grave, se considerarmos que os métodos alternativos a ela requerem hipóteses *a priori*, o que certamente introduz subjetividade na análise. Afinal, definir populações *a priori* com base em suas localidades de coleta, o que é feito rotineiramente nessas análises, também é subjetivo. Além disso, Panchal e Beaumont (2007) compararam as respostas de diversos pesquisadores a um questionário sobre como eles realizam a NCA e encontraram pouca variação nas respostas, o que sugere que a aparente subjetividade, na verdade, tem pouco impacto na análise.

Os oponentes da análise de NCA têm recrudescido seus ataques a partir das simulações usando o programa *ANeCA*, chegando a propor que a análise seja abandonada (Petit, 2008; Beaumont e Panchal, 2008; Knowles, 2008), enquanto outros a defendem (Pearse e Crandall, 2004; Templeton, 2004, 2008). Esses ataques podem ter sido responsáveis por uma diminuição no número de artigos publicados usando NCA a partir de 2008, após uma grande aceleração desde a sua criação (Figura 21.7).

Garrick *et al.* (2008) ponderam que, apesar de suas limitações, a NCA oferece uma série de vantagens, que mais que justificam seu uso como análise exploratória. Segundo eles, descartar a NCA seria como "jogar fora o bebê junto com a água da banheira". Afinal, trata-se da única análise filogeográfica que considera simultaneamente vários processos evolutivos históricos e contemporâneos, que podem ter acontecido em locais diferentes da distribuição de uma espécie ao mesmo tempo ou em um mesmo local em tempos diferentes. Além disso, há um esforço contínuo de aperfeiçoamento da análise que merece ser acompanhado de mais avaliações de seu desempenho (Templeton, 2004, 2008; Garrick *et al.*, 2008). Por fim, vale lembrar que, se a NCA for aplicada como uma análise exploratória e se suas inferências forem especificamente testadas, a maioria das críticas deixa de ser aplicável. Nós acreditamos que, de fato, apesar dos avisos iniciais do autor da abordagem sobre suas limitações (Templeton, 2002),

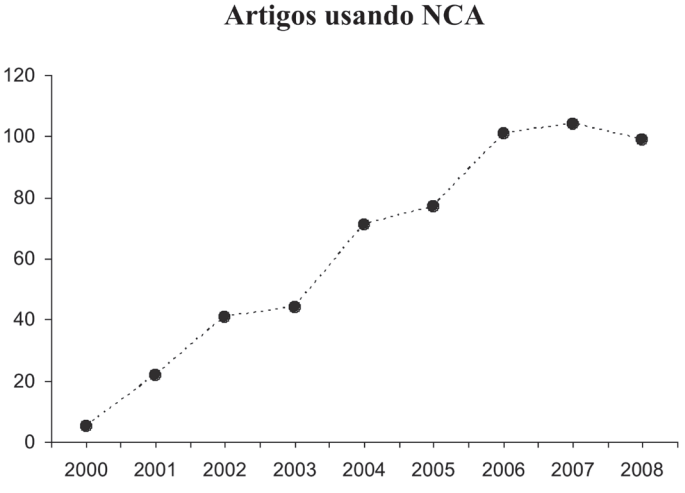


Figura 21.7. Evolução no uso da análise de clados hierarquizados (NCA) desde a publicação do programa *GeoDis* (pesquisa feita de artigos até novembro de 2010).

ela foi aplicada indiscriminadamente e muitos artigos usaram seus resultados de maneira pouco crítica. Possivelmente, muitas das análises publicadas, particularmente em que são relatados eventos de expansão de distribuição e isolamento por distância (Panchal e Beaumont, 2007), estão incorretas, o que é obviamente um resultado indesejável para o uso de um programa. Esse fato deve-se, em grande parte, à falta de atenção para os pressupostos de cada abordagem por parte dos pesquisadores, especialmente quando métodos novos se tornam disponíveis sem que eles tenham sido testados adequadamente. Apesar dos inúmeros avisos da limitação da abordagem quando é usado apenas um loco, mais de 95% de todos os trabalhos publicados com NCA fizeram exatamente isso. Curiosamente, esse contrassenso não parece ter ocorrido por impossibilidade de produzir dados de sequência de locos diferentes, mas principalmente porque não foram publicados programas de computador que permitam esse tipo de análise. É provável que esse quadro mude consideravelmente quando a metodologia de NCA multilocos automática se torne disponível. De qualquer forma, mesmo para análises de um loco e levando-se em conta as limitações de seus pressupostos, a abordagem de NCA pode ser extremamente útil como geradora de hipóteses a serem testadas com outros métodos independentes, e seu caráter subjetivo pode ser, de fato, um ponto positivo nessa análise exploratória dos resultados.

21.2.2. Outros métodos de filogeografia estatística

Um outro método filogeográfico estatístico baseia-se na modelagem de cenários populacionais alternativos a partir da teoria da coalescência e na comparação dos padrões esperados produzidos por esses modelos com os observados. Esse método precisa de hipóteses *a priori* e se apóia em outros métodos para obter parâmetros para a construção dos modelos alternativos (veja seção "Modelagem de cenários evolutivos alternativos", abaixo).

Vários métodos estimam parâmetros populacionais interessantes para análises filogeográficas. A maioria deles usa uma abordagem de Máxima-Verossimilhança ou Bayesiana para verificar a probabilidade de observação dos dados de acordo com o(s) modelo(s) escolhido(s). A partir das genealogias mais prováveis, são estimados parâmetros como tamanho populacional efetivo, taxas de migração assimétricas e tempo de divergência entre as populações, e também pode ser investigada a ocorrência de crescimento ou declínio demográfico. Esses métodos utilizam dados de sequência ou de frequência (principalmente de microssatélites), de um ou mais marcadores, sendo que alguns programas aceitam dados de mais de um tipo de marcador ao mesmo tempo. Os princípios nos quais esses métodos se baseiam e descrições de alguns métodos filogeográficos alternativos interessantes, são apresentados a seguir.

O principal avanço teórico que possibilitou o desenvolvimento da maioria desses métodos foi a teoria da coalescência, apresentada no Capítulo 20. O modelo de coalescência permite que a informação genealógica contida nas sequências ou nos alelos seja utilizada, ao contrário do que ocorre com os métodos tradicionais de genética de populações, nos quais apenas as frequências são usadas. Dessa forma, os novos métodos aproveitam melhor a informação disponível (Emerson *et al.*, 2001; Nielsen e Wakeley, 2001; Pearse e Crandall, 2004).

Uma importante previsão da teoria da coalescência é a relação entre o tempo de coalescência e o tamanho populacional. Populações pequenas coalescem em pouco tempo, porque a extinção por acaso de poucas linhagens ocorre mais rápido que a de muitas linhagens. Assim, em populações pequenas, o ancestral comum de todos os indivíduos está em um ponto mais recente do que em populações grandes, sendo que o tempo estimado para a coalescência de todos os indivíduos, a partir de sequências mitocondriais, é N_e gerações. Quanto menor o tamanho populacional, mais eventos de coalescência são esperados. Se a população estiver declinando, o número de eventos coalescentes recentes será grande. Já quando a população está em expansão, a maioria das coalescências ocorrerá no passado mais distante. Assim, reduções ou aumentos no tamanho populacional (número de linhagens) deixam sinais na genealogia, que são usados para inferir o passado demográfico da espécie (Emerson *et al.*, 2001; Rosenberg e Nordborg, 2002; Pearse e Crandall, 2004).

Com base na teoria da coalescência, o polimorfismo das sequências, o número de linhagens e o tempo estimado até seu ancestral comum são usados para inferir parâmetros populacionais. Para isso, além do modelo básico de coalescência, modelos mais sofisticados têm sido desenvolvidos. Programas como o *Migrate* (Beerli e Felsenstein, 2001), *Genetree* (Bahlo e Griffiths, 2000) e *Lamarc* (Kuhner, 2006) usam modelos elaborados baseados na coalescência para estimar fluxo gênico, tamanho populacional e tempo de divergência das populações, a partir de estratégias de máxima verossimilhança ou Bayesianas.

Muitos dos métodos aqui apresentados baseiam-se em abordagens Bayesianas ou de máxima verossimilhança para estimativa de parâmetros populacionais.

As abordagens de verossimilhança assumem que os dados observados derivam de um modelo probabilístico no qual os parâmetros são desconhecidos. A estimativa dos parâmetros é feita usando vários valores possíveis (função de verossimilhança) e verificando qual apresenta maior verossimilhança (Máxima-Verossimilhança, MV) de acordo com o modelo escolhido e os dados observados. Na análise de MV, é feita uma estimativa pontual correspondente ao valor do parâmetro que maximiza a probabilidade de ocorrência dos dados (Wade, 2000; Beaumont e Rannala, 2004).

Na inferência Bayesiana (IB), informações *a priori* podem ser adicionadas ao modelo para otimizar as estimativas. Em vez de uma estimativa pontual do parâmetro, na IB são obtidas probabilidades posteriores, que são definidas em função de probabilidades *a priori* ("prior probabilities"). A probabilidade *a priori* refere-se à distribuição dos valores do parâmetro antes da observação dos dados, e a probabilidade posterior retrata a distribuição desses valores após considerar os dados. As inferências são feitas a partir da probabilidade posterior, que é o resultado da integração do produto da função de verossimilhança (simulação com todos os valores possíveis dos parâmetros) com a probabilidade *a priori*. A própria distribuição posterior fornece uma medida da incerteza relacionada com a estimativa, e sua média pode ser usada como estimativa pontual (Wade, 2000; Beaumont e Rannala, 2004).

Nos métodos desenvolvidos recentemente, todas as genealogias possíveis são consideradas, enquanto vários parâmetros populacionais são estimados ao mesmo tempo a partir dos dados. Assim, os cálculos da função de verossimilhança e da distribuição posterior são somas ou integrais complexas, que não podem ser resolvidas analiticamente e cuja aproximação geralmente é feita por simulação Monte Carlo de cadeias de Markov (MCMC) (Pritchard *et al.*, 2000; Beerli e Felsenstein, 2001; Beaumont e Rannala, 2004).

As cadeias de Markov modelam processos estocásticos com um número finito de estados nos quais o estado futuro depende apenas do estado presente do sistema e nunca do estado em momentos passados (Emerson *et al.*, 2001; Beaumont e Rannala, 2004). Elas são usadas para obter a função de verossimilhança ou de probabilidade posterior de diferentes genealogias (e parâmetros) dentre as possíveis, já que é impossível calcular todas elas devido à demanda computacional. A cadeia move-se pelo espaço

de soluções possíveis, calculando a cada passo a verossimilhança de uma mudança na genealogia e comparando-a com a atual. Se a verossimilhança (no caso da MV) ou a distribuição posterior (IB) da nova genealogia for maior que a atual, a cadeia aceita a mudança. Se for menor, o valor da razão nova:atual é comparado com um valor entre 0 e 1 gerado aleatoriamente e a nova genealogia é descartada somente se a razão for também inferior a esse valor. Em seguida, a cadeia torna a avaliar outra mudança, e assim por diante. Dessa forma, são amostradas as genealogias que contribuem para estimativas melhores (mais prováveis) dos parâmetros.

Quando a estratégia MCMC é adotada, os primeiros estágios da cadeia de Markov são descartados ("burn-in"), a fim de gerar uma distribuição na qual os valores encontrados são menos influenciados pelos valores iniciais, privilegiando-se, assim, a busca na distribuição estacionária. Além disso, procura-se simular cadeias de Markov longas o suficiente, assim como réplicas (cadeias) independentes, para assegurar que o espaço de soluções foi eficientemente explorado (o que é avaliado pela convergência entre as iterações). As inferências são, então, feitas a partir dessa distribuição. Na análise Bayesiana, a distribuição obtida é a probabilidade posterior conjunta de um ou mais parâmetros (Beaumont e Rannala, 2004).

Uma estratégia alternativa à de MV ou IB com MCMC é o cálculo Bayesiano aproximado (ABC, "Approximate Bayesian Computation", em inglês; Beaumont *et al.*, 2002). Pelo ABC, são inferidos parâmetros demográficos a partir da aproximação de suas distribuições posteriores, sem cálculos de verossimilhança. Primeiro é gerado um grande número de dados usando parâmetros conhecidos em um determinado modelo, e para cada conjunto de dados são calculadas estatísticas sintéticas. Essas estatísticas sintéticas simuladas são comparadas com as observadas e as que forem mais diferentes do que o limite de tolerância escolhido são rejeitadas. É feita, então, uma regressão linear local dos valores dos parâmetros associados às estatísticas aceitas, que aproxima a distribuição posterior e pode ser usada para estimar os parâmetros de interesse. O ABC é vantajoso nos casos em que é impossível estimar as verossimilhanças devido à grande complexidade demográfica dos modelos (Beaumont *et al.*, 2002). Usando essa abordagem, podem ser estimados tamanhos populacionais efetivos (Talmon *et al.*, 2004) e taxas de migração recentes após eventos de expansão, além do tempo decorrido desde então (Hamilton *et al.*, 2005).

Nesses métodos, são estimados os parâmetros θ e M, proporcionais ao tamanho populacional efetivo e à taxa de migração de longo prazo, respectivamente. Como é impossível desvincular o tamanho populacional (N_e) e a taxa de migração (m) da taxa de mutação (μ), estima-se θ e M, sendo $\theta = xN_e\mu$, e $M = m/\mu$, onde x é 1 para marcadores mitocondriais e 4 para nucleares (Beerli, 2004).

O pacote *Lamarc* reúne vários programas (*Fluctuate*, *Migrate*, *Coalesce* e *Recombine*) e estima simultaneamente vários parâmetros populacionais a partir de genealogias usando um esquema de Máxima-Verossimilhança ou Bayesiano de inferência, ambos a partir de MCMC (Kuhner, 2006; Kuhner e Smith, 2007). A versão atual do *Lamarc* estima o tamanho populacional efetivo de longo-prazo e as taxas de migração assimétricas entre as populações, detecta crescimento ou declínio populacional e calcula a taxa de crescimento exponencial usando dados de sequência ou frequência, combinados ou não.

O programa *Migrate*, que integra o *Lamarc*, também existe em versão solo, na qual a inferência Bayesiana é implementada de forma diferente da do *Lamarc*. Esse programa estima as taxas de migração assimétricas entre n populações, assim como seus tamanhos populacionais efetivos de longo-prazo (Beerli e Felsenstein, 1999, 2001). Métodos para estimar fluxo gênico, como o implementado no *Migrate*, são muito mais úteis do que os disponíveis anteriormente baseados no F_{ST} (veja discussão no início do capítulo). Além de não apresentar as limitações dos métodos baseados no F_{ST}, a estimativa de fluxo gênico por MV ou IB possui outras vantagens, como, por exemplo, a possibilidade de obtenção de uma estimativa da migração assimétrica entre duas populações, que é uma expectativa bastante realista, e o fato de fornecer intervalos de confiança para os parâmetros (Beerli e Felsenstein, 2001; Beerli, 2004).

Usando o *Migrate*, Roman e Palumbi (2003) estimaram o tamanho populacional efetivo de longo prazo das baleias jubarte, fin e minke (*Megaptera novaeangliae*, *Balaenoptera physalus* e *B. acutorostrata*, respectivamente) no Atlântico Norte a partir de sequências da região controle mitocondrial. Essas três espécies de baleia tiveram suas populações reduzidas devido à caça, mas dados sobre seus tamanhos populacionais anteriores à exploração não são confiáveis, porque os registros são incompletos. Por outro lado, é importante obter dados sobre a abundância dessas espécies, já que há alegações de que elas estejam atualmente próximas da capacidade suporte, o que justificaria a retomada da caça. Os autores observaram tamanhos populacionais históricos de 240.000, 360.000 e 265.000 para as jubartes, fin e minkes, respectivamente. Esses valores são dez vezes superiores aos previamente supostos (para as jubartes e fin) e muito maiores que os atuais (10.000, 56.000 e 150.000). A análise mostrou também que, para que a atual diversidade genética fosse explicada pelos tamanhos históricos previamente propostos, o tempo de geração e a taxa de mutação precisariam ser muito maiores do que se acredita hoje (Roman e Palumbi, 2003).

Outro exemplo é o estudo da estruturação populacional de baleias-nariz-de-garrafa-do-Norte (*Hyperoodon ampullatus*), com amostras de três áreas no Atlântico Norte e utilizando marcadores mitocondriais e microssatélites (Dalebout *et al.*, 2006). Os autores compararam as estimativas de fluxo gênico obtidas pelo método tradicional e pelo *Migrate*, que fornece a taxa de migração entre cada par de populações, nos dois sentidos. Além de conseguir estimar o fluxo gênico mesmo em uma situação onde não seria possível estimá-la pelo F_{ST} (pois o F_{ST} entre as populações não foi significativamente diferente de zero), o *Migrate* permitiu detectar uma assimetria da migração entre as populações. Esse foi um resultado importante, porque mostrou que a população do Gully (um canyon submarino a 300 Km da Nova Escócia) recebia em média apenas um ou nenhum migrante das outras populações a cada geração. Dessa forma, a população do Gully, que foi extensivamente caçada, requer medidas de conservação, pois sua extinção dificilmente será seguida de recolonização a partir das outras populações.

O programa *Genetree* (Bahlo e Griffiths, 2000) é semelhante ao *Migrate*, mas utiliza somente dados de sequências. Ele também estima o fluxo gênico assimétrico entre populações e seus tamanhos populacionais efetivos a partir de MV e MCMC, produzindo resultados similares aos obtidos com o *Migrate*, aproximadamente com mesmo esforço e tempo computacional (Beerli e Felsenstein, 2001). Uma das principais diferenças entre os dois programas está na estratégia exclusivamente de MV do *Genetree*, o que pode ser limitante. Beerli (2006) concluiu que a IB funciona melhor em conjuntos de dados de um único loco ou de vários locos com baixa variabilidade e que a MV pode ser incapaz de atingir boas estimativas nesses casos. Além disso, o *Migrate* permite definir diferentes modelos de mutação, enquanto o *Genetree* assume o modelo de sítios infinitos. Por outro lado, o *Genetree* faz outras inferências sobre a história das populações, como o tempo decorrido desde o ancestral comum mais recente

de todos os indivíduos analisados e de cada população, em que local os ancestrais estavam originalmente e onde cada mutação surgiu (Bahlo e Griffiths, 2000).

Usando o programa IM, a partir de dados de um loco ou de vários locos, é possível testar se duas populações originadas de uma mesma população ancestral divergiram recentemente e trocam poucos migrantes, ou se estão separadas há muito tempo, mas apresentam alto fluxo gênico (Nielsen e Wakeley, 2001, Hey e Nielsen, 2004). Esse programa usa MCMC e MV ou IB para comparar os dois modelos e decidir qual é o mais provável e também estima o tempo desde a divergência das populações e seus tamanhos populacionais efetivos em relação à população ancestral. Uma versão mais recente (*IMa*) incorpora um novo método de inferência, mais rápido e preciso, mas não modela todos os parâmetros incluídos no *IM* (Hey e Nielsen, 2007). Esses dois programas não assumem que as populações estão trocando migrantes por um longo período de tempo (ou seja, não assumem que atingiram o equilíbrio migração-deriva) e, portanto, seriam mais adequados para populações recentemente separadas (Hey e Nielsen, 2004).

Além dos métodos citados acima, existem outras formas de inferir a demografia passada de uma espécie. Um exemplo é o método de "skyline plot", que se aproveita da relação entre a distribuição do número de nós em uma genealogia (eventos de coalescência) e o tamanho populacional efetivo. Usando sequências, é construída uma genealogia a partir da qual é calculado o parâmetro M_i, proporcional ao tamanho populacional efetivo de longo prazo e à taxa de mutação. Os valores de M_i são plotados contra os intervalos de tempo entre os nós, fornecendo um gráfico que mostra a demografia passada da espécie ("skyline plot"; Pybus *et al.*, 2000). Uma modificação foi posteriormente introduzida ao método, reduzindo a quantidade de ruído no gráfico ("generalized skyline plot"). Nos dois métodos, diferentes modelos demográficos produzem gráficos distintos, o que possibilita a comparação entre o gráfico observado a partir dos dados com os previstos pelos diferentes modelos (Emerson *et al.*, 2001; Pybus e Rambaut, 2002).

A desvantagem dos "skyline plots" é que eles se baseiam em uma genealogia estimada que, devido à estocasticidade do processo coalescente e da reconstrução filogenética, pode não corresponder à genealogia verdadeira. Para resolver esse problema, foi desenvolvido o "skyline plot" Bayesiano, que incorpora intervalos de confiança que fornecem uma medida dessas incertezas (Drummond *et al., 20*05). Esse método, implementado no programa *BEAST* (Drummond e Rambaut, 2007), usa MCMC para estimar as probabilidades posteriores dos tamanhos populacionais efetivos ao longo do tempo a partir das sequências, de acordo com um modelo de substituição especificado, que pode ser escolhido com ajuda do programa *Modeltest* (Posada e Crandall, 1998).

Usando o "skyline plot" Bayesiano, Drummond *et al.* (2005) re-analisaram dados de sequências mitocondriais obtidas de bisões (*Bison* cf. *priscus*) congelados na "permafrost" da Beringia. O bisão foi um dos grandes mamíferos mais abundantes durante o final do Pleistoceno. Shapiro *et al.* (2004) analisaram essas sequências e observaram um longo crescimento populacional, seguido de um declínio há cerca de 37.000 anos, que os autores atribuíram às mudanças climáticas do Pleistoceno. O modelo do "skyline plot" Bayesiano mostrou um cenário parecido com o anteriormente observado, porém mais detalhado. Nele, ao fim do período de declínio, percebe-se um gargalo populacional ainda mais acentuado, há cerca de 10.000 anos, que coincide com a extinção de vários representantes da megafauna da América do Norte e com o primeiro assentamento humano no Alasca. Embora o início do declínio anteceda a presença humana na América do Norte em ambos os modelos, o "skyline plot" Bayesiano demonstra que o papel dos humanos na redução populacional do bisão foi maior do que se imaginava (Drummond *et al.*, 2005).

A ocorrência de gargalos evolutivos também pode ser testada usando testes como o implementado no programa *Bottleneck* (Piry *et al.*, 1999) e o teste de "M-Ratio" (Garza e Williamson, 2001). O *Bottleneck* consegue detectar gargalos usando dados de frequência genotípica de pelo menos quatro locos a partir de 20 a 30 indivíduos em uma população. O programa baseia-se no fato de que, durante reduções drásticas do tamanho populacional, alelos raros são perdidos mais rapidamente do que a heterozigosidade. Assim, um excesso de heterozigosidade esperada (H_e), relativamente à heterozigosidade esperada no equilíbrio (H_{eq}), indica um gargalo, porque o cálculo de H_{eq} leva em consideração o número de alelos (Cornuet e Luikart, 1996; Luikart e Cornuet, 1998). A distribuição de H_{eq} é obtida por simulações do processo coalescente considerando uma amostra do mesmo tamanho da analisada e dois modelos evolutivos. Esses dois modelos (modelo de alelos infinitos e modelo de mutação passo-a-passo) representam tipos extremos de evolução e tanto o modelo de alelos infinitos quanto qualquer modelo intermediário levam a um excesso de heterozigosidade como consequência de gargalos. Dentre os três testes da hipótese nula de ausência de gargalo ($H_e = H_{eq}$), o de Wilcoxon é o que tem maior poder (Piry *et al.*, 1999).

Um outro teste que usa propriedades dos microssatélites e a rápida perda de alelos raros para inferir gargalos evolutivos a partir de frequências genotípicas é o da razão M ("M-Ratio", Garza e Williamson 2001). Nele, calcula-se a razão entre o número de alelos em um locus (k) e a distância, em pares de bases, entre o alelo de menor e o de maior tamanho (r): M = k / r. Como os alelos raros desaparecem rapidamente após um gargalo, enquanto a distância entre o menor e o maior alelo tende a permanecer igual, a razão M geralmente é menor em populações que passaram por estrangulamentos populacionais drásticos. Uma distribuição do valor de M no equilíbrio é simulada e M é considerado significativamente menor que o esperado se menos de 5% dos valores simulados forem menores que o observado. Por meio de simulações, também é calculado M_c, que é o valor crítico de M abaixo do qual uma população mostra um sinal de declínio abrupto recente. O teste pode ser feito usando dois programas disponíveis na página do Dr. J.C. Garza (*M_P_Val* e *Critical_M*). Como o *Bottleneck* e o teste da razão M baseiam-se em diferentes propriedades dos microssatélites e testam diferentes sinais de gargalo, quando ambos fornecem resultados concordantes, há uma maior confiança na inferência de estrangulamento populacional (Pearse e Crandall, 2004).

O tempo de divergência entre as populações pode ser estimado pelo *IM, IMa* (Hey e Nielsen, 2007) e *Genetree* (Bahlo e Griffiths, 2000) (descritos acima) ou pelo *BEAST* (Drummond *et al.*, 2006). Divergências mais antigas (entre espécies) também podem ser datadas usando qualquer programa que implemente o relógio molecular clássico ou relaxado, como o *BEAST* e o *Multidivtime* (Kishino *et al.*, 2001).

A datação da divergência entre linhagens foi impossibilitada por muito tempo porque dependia da veracidade dos relógios moleculares na evolução, até que surgiram modelos que relaxam o relógio molecular. Esses métodos não assumem que a taxa evolutiva é constante ao longo dos ramos e entre as linhagens—em vez disso, modelam as taxas de evolução ao longo da filogenia. Thorne *et al.* (1998) e Kishino *et al.* (2001) desenvolveram um método para estimar a idade de separação de ramos em árvores filogenéticas usando IB e MCMC, e um modelo de relógio molecular relaxado, com autocorrelação das taxas evolutivas entre as linhagens. Esse método utiliza os programas *PAML* (Yang,

1997, 2007) e *Multidivtime*. As análises são feitas a partir de filogenias prontas.

O programa *BEAST* usa IB e MCMC para a análise filogenética ou genealógica de sequências. A grande diferença desse programa é que as árvores que são construídas possuem informação sobre o tempo decorrido entre os nós, assumindo um relógio molecular estrito, ou modelos de relógio relaxado (Drummond *et al.*, 2006, Drummond e Rambaut, 2007). Diferentemente do *Multidivtime*, no *BEAST* são implementados modelos não correlacionados de relógio molecular. Embora os tempos de divergência estimados sejam altamente sensíveis à escolha do modelo de relógio, ainda não há avaliações definitivas sobre o desempenho dos diferentes modelos. Lepage *et al.* (2007) compararam estimativas feitas com vários dos modelos de relógio existentes e também com dois novos e concluíram que os modelos com taxas autocorrelacionadas fornecem estimativas mais confiáveis.

Diferentes modelos de divergência populacional podem ser testados com o programa *Mesquite* (Maddison e Maddison, 2008). Nessa abordagem, são usados modelos de cenários demográficos alternativos, cada um descrito por uma árvore de populações diferente. O programa é capaz de simular, a partir do modelo de coalescência, árvores de genes sob diferentes cenários de divergência populacional (ou seja, árvores de populações) com ou sem migração e também de mudança de tamanho das populações.

Com a finalidade de testar hipóteses filogeográficas alternativas, o *Mesquite* simula árvores de genes de acordo com um determinado modelo (digamos, modelo A) e mede o grau de discordância entre elas e a árvore de populações prevista sob um modelo alternativo (modelo B). A partir dessas simulações, é construída uma distribuição nula dos valores de discordância. A discordância é medida por uma estatística baseada no número de coalescências antigas (Maddison, 1997). O valor observado de discordância é comparado com o esperado (distribuição nula): se a discordância observada for menor que aquela observada em 95% dos valores da distribuição esperada, não é possível rejeitar o modelo B. Dessa forma, é possível avaliar o nível de confiança estatístico de que uma genealogia corresponda a um modelo específico (Knowles e Maddison, 2002).

Usando esse método, Knowles *et al.* (2007) testaram se populações atuais do gafanhoto *Melanoplus marshalli*, restritas a topos de montanhas dos Estados Unidos, se originaram de um mesmo refúgio durante a última glaciação do Pleistoceno, ou de múltiplos refúgios. Se o modelo de refúgio único fosse real, a discordância observada entre as árvores de genes e a de populações seria maior que a esperada a partir das simulações usando o modelo de múltiplos refúgios. Como a discordância observada foi menor, os autores concluíram que as populações atuais se formaram a partir de vários refúgios glaciais.

O método baseado em modelagem é vantajoso em situações de divergência recente, quando ainda não houve separação completa das linhagens, já que considera várias árvores de genes possíveis a partir dos dados e fornece uma estimativa da confiança estatística das inferências. Entretanto, aceitar um modelo pela rejeição estatística do alternativo não significa necessariamente que o modelo aceito reflita o passado verdadeiro, porque a evolução real do grupo pode simplesmente não ter sido representada em nenhum dos modelos testados. Além disso, aceitar um modelo porque ele não foi rejeitado é uma inferência estatística fraca, pois resulta da aceitação da hipótese nula. A forma como os modelos foram definidos também pode tendenciar os resultados (Pearse e Crandall, 2004; Excoffier e Heckel, 2006; Templeton, 2008). Outra desvantagem é que o número de cenários evolutivos que podem ser modelados é limitado. A fim de guiar a elaboração dos modelos, alguns autores recomendam o uso de dados paleoambientais e fósseis para a construção de paleodistribuições que, pela comparação com as distribuições atuais, ajudem a formular hipóteses (Knowles e Carstens, 2007, Richards *et al.*, 2007).

A estimativa das taxas de migração recentes pode ser obtida a partir de análises de IB e MCMC sobre genótipos multilocos. Essa abordagem é implementada no programa *BayesAss* (Wilson e Rannala, 2003), que também infere a ancestralidade dos indivíduos, as frequências alélicas das populações e seus coeficientes de endocruzamento. O *BayesAss* não estipula que as populações devam estar em equilíbrio de Hardy-Weinberg, mas assume que todas os grupos populacionais diferentes daquela espécie na natureza foram amostrados, o que é um pressuposto dificilmente verdadeiro. Para estimar o fluxo gênico recente, o programa faz uma análise de atribuição ("assignment") na qual o genótipo multiloco de cada indivíduo é usado para designar qual é sua população de origem e para identificar migrantes e descendentes de migrantes, a partir dos quais são obtidas as taxas de migração assimétricas entre populações. Por ser um tipo de teste de atribuição, ele não tem a limitação dos métodos baseados na coalescência, que precisam assumir que as populações mantiveram tamanho constante ou cresceram exponencialmente nas últimas $4N_e$ gerações (o que é irreal para espécies com N_e grandes ou que vivem em ambientes sujeitos a grandes perturbações) (Wilson e Rannala, 2003).

A possibilidade de estimar o fluxo gênico recente é interessante porque permite uma comparação entre estimativas de longo prazo e atuais. Por exemplo, Palstra *et al.* (2007) usaram dados de microssatélies para analisar o fluxo gênico entre salmões (*Salmo salar*) desovando em vinte rios do Canadá. As taxas de migração de longo prazo assimétricas foram estimadas usando o programa *Migrate* e as taxas de migração assimétricas recentes foram obtidas pelo *BayesAss*. Dessa forma, foi possível observar que o fluxo gênico atual entre as populações é mais limitado que ao longo do passado da espécie, o que tem consequências para seu manejo. A direção do fluxo gênico também foi um resultado importante, que contrariou as expectativas. O manejo dessas populações de desova do salmão considerava que sua dinâmica obedecia a um modelo de "populações fonte e escoadouro" ("source-sink", em inglês), ou seja, que populações grandes serviam de fonte de migrantes para as populações pequenas. As análises de Palstra *et al.* (2007) mostraram que, em muitos casos, a migração ocorria no sentido inverso. Isso significa que a estratégia de manejo para a espécie precisa ser revista, estendendo a proteção às populações pequenas.

O desempenho do *BayesAss* foi avaliado por Faubet *et al.* (2007). Os autores concluíram que o programa consegue estimar com acurácia as taxas de migração quando a história demográfica da espécie corresponde às premissas do modelo e quando a diferenciação entre as populações não é muito baixa. Nesses casos, mesmo taxas de migração elevadas podem ser precisamente estimadas. Em situações muito diferentes das assumidas pelo modelo, porém, o programa só produz estimativas confiáveis quando a taxa de migração é muito baixa e a diferenciação, alta. Assim, é necessário avaliar com cautela o quanto a espécie estudada se encaixa no modelo, que prevê baixa taxa de migração (a proporção de migrantes em cada geração deve ser igual ou menor a 1/3) e assume que a migração e a deriva não alteraram as frequências alélicas das populações ao longo das últimas gerações.

21.3. Abordagem Integrada

Embora as duas abordagens da filogeografia estatística sejam conceitualmente diferentes, elas não são mutuamente excludentes (Templeton, 2004, 2008). Na realidade, o ideal seria usar as duas

de forma complementar, permitindo que hipóteses geradas pela NCA sejam testadas usando os demais métodos. É conveniente lembrar que os resultados obtidos com os métodos filogeográficos alternativos, baseados na modelagem de cenários evolutivos, serão satisfatórios apenas se os modelos construídos para as análises forem realistas—e em grande parte dos qauis, não existe informação *a priori* para sua elaboração (Pearse e Crandall, 2004; Excoffier e Heckel, 2006; Templeton 2008). Assim, com uma abordagem integrada, é possível tirar vantagem do melhor de cada um dos dois grupos de métodos: considerando múltiplos fatores causais (como na NCA), em vez de um modelo demográfico simplificado, e também testando a confiança estatística de diferentes árvores de populações hipotéticas (como nos demais métodos; Knowles e Maddison, 2002; Wakeley, 2003; Templeton, 2004, 2008).

21.4. Filogeografia e Delimitação de Populações: Fluxo Gênico no Presente

Em contraposição aos métodos filogeográficos, que tentam reconstruir o passado evolutivo das espécies, análises modernas de estruturação populacional visam detectar padrões de fluxo gênico recentes ou atuais, que possibilitem a delimitação geográfica das populações. As duas abordagens, porém, são importantes para a compreensão dos padrões de distribuição da variabilidade genética atual e dos processos que devem ser mantidos para assegurar a preservação dessa variabilidade, assim como das populações e das espécies.

A delimitação de populações é uma questão fundamental no caso das espécies ameaçadas, assim como nas espécies selvagens exploradas comercialmente (Avise, 1989, Allendorf e Luikart, 2006). Nas duas situações, pretende-se obter informação, a partir de marcadores genéticos, que ajude a garantir a persistência de unidades demográficas independentes. Essas unidades são chamadas de "estoques", quando aplicadas às espécies exploradas, e "populações", "Unidades de Manejo" ("Management Units", MU) ou "Unidades Evolutivamente Significativas" ("Evolutionarily Significant Units", ESU), para as espécies ameaçadas.

Identificar o limite geográfico entre as populações é mais difícil no caso de espécies com distribuição contínua. Até recentemente, esse era um dos maiores desafios no estudo da estruturação genética, porque os métodos tradicionais requerem a definição *a priori* das populações. Essa definição geralmente era feita usando a localidade de origem das amostras e cada localidade era tratada como uma "população" potencial. A identificação dos agrupamentos de localidades que formam "populações" era feita, por exemplo, pelo teste da significância dos índices F_{ST} entre cada par de populações. Dessa forma, a hipótese de panmixia não pode ser rejeitada para localidades que não possuem valores de F_{ST} significativos entre si e o limite entre as populações é marcado por valores de F_{ST} significativos. A desvantagem desse método é que, conforme discutido no início do capítulo, o F_{ST} confunde eventos do passado com a estruturação atual e não permite uma estimativa minimamente precisa do fluxo gênico entre as populações (Bossart e Prowell, 1998a, b; Whitlock e McCauley, 1999; Pearse e Crandall, 2004).

Nos últimos anos, a redução do custo de genotipagem produziu um incremento do uso dos microssatélites, com aumento tanto do número amostral quanto de locos analisados. Como consequência, foram desenvolvidos métodos analíticos que partem da informação no nível do indivíduo para detectar padrões de estruturação e estimar parâmetros populacionais atuais e passados. Esses métodos, chamados "métodos de atribuição", usam estratégias de Máxima-Verossimilhança ou Inferência Bayesiana para calcular a probabilidade de observar os diferentes genótipos em cada população e atribuem os indivíduos às populações de acordo com as probabilidades de seus genótipos pertencerem a elas. Ou seja, eles conseguem descobrir a qual população um indivíduo pertence, independentemente de seu local de coleta, o que é um enorme avanço sobre os métodos disponíveis até então (Pearse e Crandall, 2004, Beaumont e Rannala, 2004).

Os métodos de atribuição usam a informação no nível dos indivíduos com diferentes finalidades em estudos populacionais (Paetkau *et al.*, 2004; Manel *et al.*, 2005; Excoffier e Heckel, 2006). A seguir, serão apresentados métodos de atribuição com duas finalidades (além da fundamental, que é a atribuição dos indivíduos às populações): delimitação de populações e estimativa da taxa de migração contemporânea.

21.4.1. Quantas e quais populações: métodos de agrupamento

Para serem capazes de encontrar os limites entre populações, alguns métodos de atribuição também são métodos de agrupamento. O método original e mais popular é o de Pritchard *et al.* (2000), implementado no programa *Structure*. Esse programa usa inferência Bayesiana para, a partir dos genótipos multilocos dos indivíduos, estimar o número de populações que melhor explica os dados, ao mesmo tempo em que os indivíduos são atribuídos às populações. O critério para agrupamento em populações é o de minimizar os desequilíbrios de Hardy-Weinberg e de ligação dentro delas. Assim, as populações são definidas pelos dados, sem qualquer inferência *a priori*. Posteriormente, a informação sobre a origem geográfica das amostras pode ser incorporada à análise para aumentar a resolução dos limites entre as populações (Pritchard *et al., 20*00, Pritchard e Wen, 2004).

Um exemplo do uso do *Structure* é um estudo recente sobre a estruturação populacional dos ameríndios, que usou uma ampla amostragem geográfica (83 povos indígenas, sendo 29 nativos das Américas) e genômica (678 locos de microssatélites) (Wang *et al.*, 2007). A análise de agrupamento foi feita em escala inter e intracontinental. No primeiro caso, o *Structure* mostrou que os ameríndios formam um grupo distinto dos povos indígenas de outros continentes, mas existe um gradiente de similaridade em relação aos siberianos, que decresce com o aumento da distância em relação ao Estreito de Bering. Esse resultado corrobora a teoria de que a colonização das Américas aconteceu por essa rota (Figura 21.8). Na análise intracontinental, o número de grupos mais provável foi sete, sendo quatro deles formados exclusivamente por povos isolados do Brasil (Karitiana e Surui), Colômbia (Ticuna) e Paraguai (Ache). Os povos nativos do norte da América do Norte e os Pima do México formam dois outros grupos. Apesar da distância geográfica, os povos da América Central e dos Andes formam um mesmo agrupamento, o único que corresponde a um grupo linguístico (Chibchan-Paezan). A partir desses resultados, pelo menos três hipóteses sobre a demografia histórica dos ameríndios foram corroboradas: a colonização a partir da Sibéria, uma dispersão costeira, em contraposição às rotas interiores, e um processo de divergência populacional que pode ter sido influenciado por diferenças linguísticas (Wang *et al., 20*07).

Em avaliações a partir de dados simulados, o *Structure* funcionou bem em condições de alta diferenciação e baixo fluxo gênico, mas, como na maior parte das outras abordagens, teve dificuldade para identificar populações nos cenários de alta migração (Rosenberg *et al.*, 2001; Evanno *et al.*, 2005; Waples e Gaggiotti, 2006). Outros fatores que afetaram o desempenho foram o número de locos e seu grau de polimorfismo, e o número de indivíduos analisados (Waples e Gaggiotti, 2006).

Programas similares ao *Structure* são o *Partition* (Dawson e Belkhir, 2001), o *BAPS* (Corander *et al.*, 2003, 2006), o *Gene-*

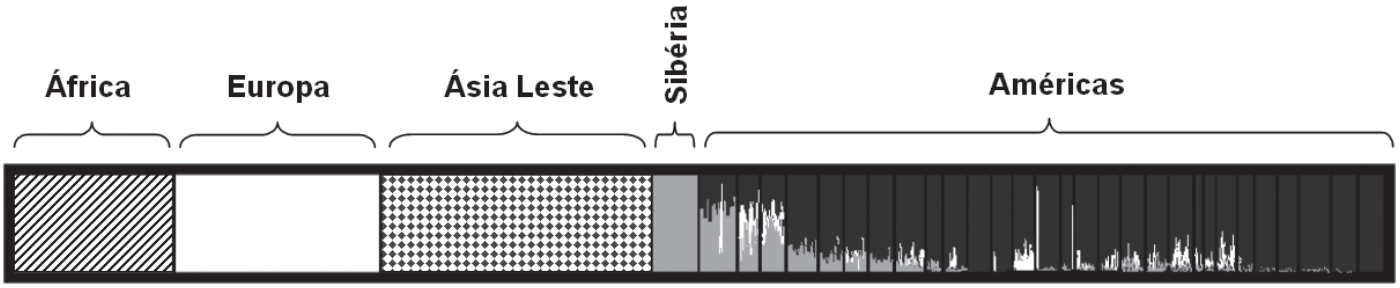


Figura 21.8. Gráfico da proporção do genótipo multilocos de cada indivíduo que corresponde a cada uma das cinco populações mundiais detectadas pelo programa *Structure*. Os ameríndios mostram um gradiente de identidade com os siberianos, que decresce com o aumento da distância em relação ao estreito de Bering (adaptado de Wang *et al.* 2007).

land (Guillot *et al.*, 2005), o *Geneclust* (François *et al.*, 2006) e o *TESS* (Chen *et al.*, 2007). Os quatro últimos diferem do *Structure* por incorporar explicitamente a informação geográfica, usando-a para determinar os agrupamentos. O *BAPS* também é diferente do *Structure*, porque, além de usar a informação de origem das amostras, trata as populações como unidades, ou seja, não faz a análise no nível do indivíduo. Esse programa estima as frequências alélicas de cada população e determina quais são diferentes, em vez de separar os indivíduos em populações com base em seus genótipos multilocos. Já o *Partition* tem a desvantagem de não permitir que os indivíduos tenham ancestralidade misturada, ao contrário do *Structure*.

Rowe e Beebee (2007) compararam os resultados obtidos com três desses métodos (*Structure*, *BAPS* e *Geneland*), usando dados de microssatélites de uma rã (*Bufo calamita*) ameaçada de extinção na Grã-Bretanha. Os três métodos mostraram resultados similares, mas houve algumas inconsistências, principalmente na região geográfica que os autores acreditam ter uma estruturação mais complexa (South Cumbria). Nesse caso, o programa *Geneland* foi o que apresentou resultados mais consistentes.

Latch *et al.* (2006) usaram dados simulados para comparar o desempenho dos programas *Structure*, *Partition* e *BAPS* em cenários de baixa diferenciação populacional ($F_{ST} \leq 0,1$). O *Partition* conseguiu detectar o número de populações corretamente apenas em situações acima de $F_{ST} = 0,09$, enquanto *Structure* e *BAPS* funcionaram muito bem mesmo em valores de F_{ST} entre 0,02 e 0,03. O número de atribuições corretas de indivíduos chegou a mais de 97% quando $F_{ST} \geq 0,05$. Os autores frisam que a substituição da MCMC por um algoritmo estocástico de otimização ("greedy optimization") tornou a versão nova do *BAPS* (Corander *et al.*, 2006) muito veloz, mas defendem que ele seja usado junto com o *Structure*, já que a concordância entre os dois gera uma maior confiança nos resultados. De uma maneira geral, essa conclusão aplica-se a todas as abordagens: é interessante observar como os diversos algoritmos, cada qual com seus pressupostos e modelos diferentes, particionam as populações. Pontos de concordância entre as várias análises podem ser considerados mais robustos, enquanto que pontos de discordância devem ser analisados com maior detalhe, levando em conta as vantagens e limitações de cada abordagem face ao tipo de dados e a biologia das espécies estudadas.

21.4.2. Métodos de estimativa do fluxo gênico atual

Os métodos filogeográficos que estimam fluxo gênico e tamanho populacional efetivo fornecem uma média desses parâmetros ao longo do tempo evolutivo, que pode não corresponder aos valores atuais. Entretanto, os valores contemporâneos são importantes para auxiliar nas decisões para conservação e manejo das espécies. Nos últimos anos, análises que partem do nível dos indivíduos foram desenvolvidas para possibilitar a estimativa de parâmetros populacionais na escala de tempo ecológico. Alguns desses métodos são análogos genéticos de técnicas de marcação-recaptura, que utilizam amostras dos animais para identificá-los individualmente usando vários marcadores genéticos hipervariáveis e acompanhá-los no tempo e no espaço. A análise pode ser planejada para detectar as relações entre pais e filhos, usando um número de locos ainda maior que o necessário para a identificação individual, o que maximiza a informação obtida a partir das amostras (Palsbøll, 1999; Pearse *et al.*, 2001). Essa abordagem foi batizada de "marcação genética" ("genetic tagging").

Os métodos de atribuição são capazes de detectar indivíduos migrantes na geração atual de forma direta a fim de estimar fluxo gênico no presente. Esses métodos usam reamostragens Monte Carlo para obter a distribuição dos genótipos de indivíduos de uma população e fornecer os valores de verossimilhança críticos para determinar se um indivíduo amostrado pertence àquela população. Assim, a verossimilhança de um indivíduo pertencer àquela população serve de critério para identificá-lo como migrante, sendo que a hipótese nula é a que ele tenha nascido na população onde foi amostrado. A partir da identificação dos migrantes e de suas populações de origem, é possível calcular as taxas de migração (Paetkau *et al., 20*04, Manel *et al.*, 2005).

Uma limitação importante dessa abordagem é que os métodos de atribuição acertam mais quando a diferenciação entre as populações é grande e, por conseguinte, a migração é rara. Nesses casos, mesmo uma amostragem muito grande poderia ser incapaz de detectar migrantes (Manel *et al., 20*05). Paetkau *et al.* (2004) testaram dois métodos, implementados nos programas *Geneclass* (Cornuet *et al.*, 1999) e *IMMANC* (Rannala e Mountain, 1997). Eles observaram, usando dados simulados, que as reamostragens precisavam preservar o desequilíbrio de ligação causado pela entrada recente de migrantes nas gerações e a variância da amostragem. Por isso, os autores desenvolveram outro esquema de reamostragem Monte Carlo, que teve alto poder de detecção de migrantes quando um grande número de amostras foi analisado e quando todas as populações de origem potenciais foram consideradas. Esse novo esquema de reamostragem foi incorporado à nova versão do *Geneclass* (*Geneclass2*).

21.5. Limitações e Perspectivas

Os pesquisadores que desejam realizar análises filogeográficas precisam explorar os novos métodos de análise, mas é fundamental que compreendam suas premissas e limitações. É preciso verificar se o método escolhido já foi avaliado por autores diferentes dos que o desenvolveram, usando dados empíricos ou simulados. Especialmente se o desempenho do método tiver sido insatisfatório sob condições potencialmente compatíveis com as da espécie estudada, os resultados devem ser interpretados com cautela.

De forma geral, como esses métodos somente puderam ser desenvolvidos após avanços teóricos e computacionais recentes, eles ainda não foram extensivamente avaliados. Falhas e limitações detectadas a partir de avaliações mais completas poderão ser corrigidas graças ao aumento da capacidade de processamento dos computadores e ao aperfeiçoamento das teorias que fundamentam os métodos. Além disso, o barateamento das análises genéticas também favorecerá o uso de vários marcadores em estudos no nível populacional. Assim, espera-se que os métodos filogeográficos sejam cada vez mais capazes de recriar o passado evolutivo das espécies de forma ainda mais precisa.

Naturalmente, como as análises são em geral muito complexas e impossíveis de serem refeitas manualmente, é fundamental que o pesquisador analise os resultados de maneira crítica, procurando avaliar o significado biológico do observado. Existe atualmente uma tendência lamentável da valorização excessiva dos programas mais novos em detrimento dos já estabelecidos e bem testados. Essa pressa tem provocado, muitas vezes, o uso dos programas como "caixas pretas", de modo que os pesquisadores limitam seu esforço a aprender como fazer o arquivo de entrada de dados, e os resultados são aceitos diretamente, sem considerar a filosofia e os limites de cada abordagem. Essas decisões não são inócuas, no sentido que o uso de métodos inadequados propõem reconstruções equivocadas—e às vezes inclusive decisões equivocadas de conservação. Não devemos perder de vista que nosso objetivo é a compreensão dos fenômenos biológicos e, para isso, a análise crítica e o bom senso podem ser instrumentos mais valiosos que o uso de programas mais "modernos" ou aparentemente sofisticados.

Agradecimentos

Gostaríamos de agradecer ao Dr. Sergio R. Matioli pela oportunidade de apresentar esse assunto tão importante para a Biologia Evolutiva atual. Agradecemos também ao Dr. Carlos G. Schrago pelas discussões sobre MV, IB e estimativa de tempos de divergência, e ao Dr. Cristiano Lazoski pelo auxílio na confecção das figuras. O Laboratório de Biodiversidade Molecular é apoiado por projetos do CNPq, CAPES, FAPERJ, SEAP-DF e FINEP.

Referências Bibliográficas

Allendorf, F.W., Luikart, G. (2006). **Conservation and the Genetics of Populations**. Blackwell Publishing, Oxford.

Arbogast, B.S., Edwards, S.V., Wakeley, J., Beerli, P. e Slowinski, J.B. (2002). Estimating divergence times from molecular data on phylogenetic and population genetic timescales. **Annu. Rev. Ecol. Syst. 33**:707-740.

Avise, J.C., Giblindavidson, C., Laerm, J., Patton, J.C. e Lansman, R.A. (1979). Use of restriction endonucleases to measure mitochondrial-DNA sequence relatedness in natural populations. 2. Mitochondrial-DNA clones and matriarchal phylogeny within and among geographic populations of the pocket gopher, *Geomys pinetis*. **Proc. Natl. Acad. Sci. USA 76** (12):6694-6698.

Avise, J.C., Arnold, J., Ball, R.M., Bermingham, E., Lamb, T., Neigel,

Tabela 21.1. Sítios da Internet onde os programas estão disponíveis.

Programa	*Sítio*
ANeCA	http://www.rubic.rdg.ac.uk/~mahesh/software.html
Arlequin	http://cmpg.unibe.ch/software/arlequin35/Arl35Downloads.html
BAPS	http://web.abo.fi/fak/mnf/mate/jc/software/baps.html
BayesAss	http://www.rannala.org/?page_id=245
BEAST	http://beast.bio.ed.ac.uk/Main_Page
Bottleneck	http://www.ensam.inra.fr/URLB/bottleneck/bottleneck.html
DnaSP	http://www.ub.edu/dnasp/
FSTAT	http://www2.unil.ch/popgen/softwares/fstat.htm
Geneclass	http://www1.montpellier.inra.fr/URLB/index.html
Geneclust	http://odin.mdacc.tmc.edu/~kim/geneclust/
Geneland	http://www2.imm.dtu.dk/~gigu/Geneland/
Genetree	http://www.stats.ox.ac.uk/~griff/software.html
GeoDis	http://darwin.uvigo.es/software/geodis.html
IM	http://genfaculty.rutgers.edu/hey/software
IMMANC	http://www.rannala.org/?page_id=245
Lamarc	http://evolution.gs.washington.edu/lamarc/index.html
Mesquite	http://mesquiteproject.org/mesquite/mesquite.html
Migrate	http://popgen.sc.fsu.edu/Migrate/Migrate-n.html
M-Ratio	http://swfsc.noaa.gov/textblock.aspx?Division=FED&id=3298
Multdivtime	http://statgen.ncsu.edu/thorne/multidivtime.html
Network	http://www.fluxus-engineering.com/sharenet.htm
PAML	http://abacus.gene.ucl.ac.uk/software/paml.html
Partition	http://www.genetix.univ-montp2.fr/partition/partition.html
SAMOVA	http://cmpg.unibe.ch/software/samova/
SpaGeDi	http://ebe.ulb.ac.be/ebe/Software.html
Structure	http://pritch.bsd.uchicago.edu/software.html
TCS	http://darwin.uvigo.es/software/tcs.html
TESS	http://membres-timc.imag.fr/Olivier.Francois/tess.html

J.E., Reeb, C.A. e Saunders, N.C. (1987). Intraspecific phylogeography - The mitochondrial-DNA bridge between population genetics and systematics. **Annu. Rev. Ecol. Syst. 18**:489-522

Avise, J.C. (1989). Gene trees and organismal histories - a phylogenetic approach to population biology. **Evolution 43** (6):1192-1208.

Avise, J.C. (1991). 10 Unorthodox perspectives on evolution prompted by comparative population genetic findings on mitochondrial DNA. Annu. Rev. Genet. 25:45-69.

Avise, J.C. (2000). **Phylogeography :the history and formation of species**. Harvard University Press.

Avise, J.C. (2004). **Molecular markers, natural history and evolution**. 2ª ed. Chapman & Hall, New York.

Avise, J.C. (2006). **Evolutionary pathways in nature:a phylogenetic approach**. Cambridge University Press.

Beaumont, M.A., Zhang, W.Y. e Balding, D.J. (2002). Approximate bayesian computation in population genetics. **Genetics 162**:2025-2035.

Beaumont, M.A. e Panchal, M. (2008). On the validity of nested clade phylogeographical analysis. **Mol. Ecol. 17**:2563-2565.

Beaumont, M.A. e Rannala, B. (2004). The Bayesian revolution in genetics. **Nature Rev. Genet. 5**(4):251-261.

Beerli, P. e Felsenstein, .J (1999). Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. **Genetics 152**:763-773.

Beerli, P., Felsenstein J (2001). Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. **Proc. Natl. Acad. Sci. USA 98**(8):4563-4568.

Beerli, P. (2004). Migrate Documentation. Disponível em: *http://popgen.csit.fsu.edu/migrate.download.html.*

Bossart, J.L. e Prowell, D.P. (1998a). Genetic estimates of population structure and gene flow:limitations, lessons and new directions. **Trends. Ecol. Evol. 13**:202-206.

Bossart, J.L. e Prowell, D.P. (1998b). Is population genetics mired in the past? Reply from J.L. Bossart and D. Pashley Prowell. **Trends. Ecol. Evol. 13**:360-360.

Bradley, B.J. (2008). Reconstructing phylogenies and phenotypes:a molecular view of human evolution. **J. Anat. 212**:337-53.

Brisson, J.A., De Toni, D.C., Duncan, I. e Templeton, A.R. (2005). Abdominal pigmentation variation in *Drosophila polymorpha*:Geographic variation in the trait, and underlying phylogeography. **Evolution 59**:1046-1059.

Cassens, I., Van Waerebeek, K., Best, P.B., Tzika, A., Van Helden, A.L., Crespo, E.A. e Milinkovitch, M.C. (2005). Evidence for male dispersal along the coasts but no migration in pelagic waters in dusky dolphins (*Lagenorhynchus obscurus*). **Mol. Ecol. 14** (1):107-121.

Chen, C., Durand, E., Forbes, F. e François, O. (2007). Bayesian clustering algorithms ascertaining spatial population structure:a new computer program and a comparison study. **Mol. Ecol. Notes 7**:747-756.

Clement, M., Posada, D. e Crandall, K.A. (2000). TCS:a computer program to estimate gene genealogies. **Mol. Ecol. 9**:1657-1659.

Cornuet, J.M., Piry, S., Luikart, G., Estoup, A. e Solignac, M. (1999). New methods employing multilocus genotypes to select or exclude populations as origins of individuals. **Genetics 153**:1989-2000.

Crandall, K.A. e Templeton, A.R. (1993). Empirical tests of some predictions from coalescent theory with applications to intraspecific phylogeny reconstruction. **Genetics 134**:959-969.

Corander, J., Waldmann, P., e Sillanpaa, M.J. (2003). Bayesian analysis of genetic differentiation between populations. **Genetics 163**:367-374.

Corander, J., Marttinen, P. e Mantyniemi, S. (2006). A Bayesian method for identification of stock mixtures from molecular marker data. Fish. Bull. 104:550-558.

Cornuet, J.M. e Luikart, G. (1996). Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. **Genetics 144**:2001-2014.

Cunha, H.A., da Silva, V.M.F., Lailson-Brito, J. Jr, Santos, M.C.O., Flores, P.A.C., Martin, A.R., Azevedo, A.F., Fragoso, A.B.L., Zanelatto, R.C. e Solé-Cava, A.M. (2005). Riverine and marine ecotypes of *Sotalia* dolphins are different species. **Mar .Biol. 148**:449-457.

da Silva, M.N.F. e Patton, J.L. (1998). Molecular phylogeography and the evolution and conservation of Amazonian mammals. **Mol. Ecol. 7**:475-486.

Dalebout, M.L., Ruzzante, D.E. , Whitehead, H. e Oien, N.I. (2006). Nuclear and mitochondrial markers reveal distinctiveness of a small population of bottlenose whales (*Hyperoodon ampullatus*). in the western North Atlantic. **Mol. Ecol. 15**:3115-3129.

Dawson, K.J. e Belkhir, K. (2001). A Bayesian approach to the identification of panmictic populations and the assignment of individuals. **Genet. Res. 78**:59-77.

Dupanloup, I., Schneider, S. e Excoffier, L. (2002). A simulated annealing approach to define the genetic structure of populations. **Mol. Ecol. 11**(12):2571-2581.

Drummond, A.J., Rambaut, A., Shapiro, B. e Pybus, O.G. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. **Mol. Biol. Evol. 22**:1185-1192.

Drummond, A.J., Ho, S.Y.W., Phillips, M.J. e Rambaut A (2006). Relaxed phylogenetics and dating with confidence. **PLOS Biology 4**:699-710

Drummond, A.J. e Rambaut, A. (2007). BEAST:Bayesian evolutionary analysis by sampling trees. **BMC Evol. Biol. 7**:214.

Edwards S.V. e Beerli P. (2000). Perspective: Gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. **Evolution 54**:1839-1859.

Eizirik, E., Bonatto, S.L., Johnson, W.E., Crawshaw, P.G., Vie, J.C., Brousset, D.M., O'Brien, S.J., Salzano, F.M. (1998). Phylogeographic patterns and evolution of the mitochondrial DNA control region in two neotropical cats (Mammalia, Felidae). **J. Mol. Evol. 47**:613-624.

Emerson, B.C., Paradis, E. e Thebaud, C. (2001). Revealing the demographic histories of species using DNA sequences. **Trends. Ecol. Evol. 16** (12):707-716.

Evanno, G., Regnaut, S. e Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE:a simulation study. **Mol. Ecol. 14** (8):2611-2620.

Excoffier, L., Smouse, P.E. e Quattro, J.M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes:application to human mitochondrial DNA restriction data. **Genetics 131**:479-491.

Excoffier, L. e Heckel, G. (2006). Computer programs for population genetics data analysis:a survival guide. **Nature Rev. Genet. 7**(10):745-758.

Falush, D., Stephens, M. e Pritchard, J.K. (2003). Inference of population structure using multilocus genotype data:Linked loci and correlated allele frequencies. **Genetics 164**:1567-1587.

Faubet, P., Waples, R.S. e Gaggiotti, O.E. (2007). Evaluating the performance of a multilocus Bayesian method for the estimation of migration rates. **Mol. Ecol. 16**:1149-1166.

François, O., Ancelet, S. e Guillot, G. (2006). Bayesian clustering using hidden Markov random fields in spatial population genetics. **Genetics 174**:805-816.

Fu, Y.-X. e Li, W.-H. (1993). Statistical tests of neutrality of mutations. **Genetics 133**:693-709.

Fu, Y.-X. (1997). Statistical tests of neutrality of mutations against population growth, hitchhiking and backgroud selection. **Genetics 147**:915-925.

Garrick, R.C., Dyer, R.J., Beheregaray, L.B. e Sunnucks, P. (2008). Babies and bathwater: a comment on the premature obituary for nested clade phylogeographical analysis. **Mol. Ecol. 17**:1401-1403.

Garza, J.C. e Williamson, E.G. (2001). Detection of reduction in population size using data from microsatellite loci. **Mol. Ecol. 10**:305-318.

Goudet, J. (1995). FSTAT (version 1.2): a computer program for calculating F-statistic. **J. Hered. 86**:485-486.

Guillot, G., Mortier, F., Estoup, A. (2005). Geneland:a computer package for landscape genetics. **Mol. Ecol. Notes 5**:712-715.

Hamilton, G., Currat, M., Ray, N., Heckel, G., Beaumont, M. e Excoffier, L. (2005). Bayesian estimation of recent migration rates after a spatial expansion. **Genetics 170**:409-417.

Hardy, O.J. e Vekemans, X. (2002). SPAGEDi:a versatile computer program to analyse spatial genetic structure at the individual or population levels. **Mol. Ecol. Notes 2**(4):618-620

Hey, J. e Nielsen, R. (2004). Multilocus methods for estimating population sizes, migration rates and divergence time, with Applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. **Genetics 167**:747-760.

Hey, J. e Nielsen, R. (2007). Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. **Proc. Natl. Acad. Sci. USA 104**(8):2785-2790.

Hudson, R.R. (1991). Gene genealogies and the coalescent process. In:Futuyma D, Antonovics J (eds). **Oxford Surveys in Evolutionary Biology 7**:1-44.

Kingman, J.F.C. (1982). The coalescent. **Stochast. Proc. Appl. 13**:235-48.

Kishino, H., Thorne, J.L. e Bruno, W.J. (2001). Performance of a divergence time estimation method under a probabilistic model of rate evolution. **Mol. Biol. Evol. 18**(3):352-361.

Knowles, L.L. e Maddison, W.P. (2002). Statistical phylogeography. **Mol. Ecol. 11**:2623-2635

Knowles, L.L. e Carstens, B.C. (2007). Estimating a geographically explicit model of population divergence. **Evolution 61**:477-493.

Knowles, L.L., Carstens, B.C. e Keat, M.L. (2007). Coupling genetic and ecological-niche models to examine how past population distributions contribute to divergence**. Curr. Biol. 17**:940-946

Knowles, L.L. (2008). Why does a method that fails continue to be used? **Evolution 62**(11):2713-2717

Kuhner, M.K. (2006). LAMARC 2.0:maximum likelihood and Bayesian estimation of population parameters. **Bioinformatics 22**:768-770.

Kuhner, M.K., Smith, L.P. (2007). Comparing Likelihood and Bayesian coalescent estimation of population parameters. **Genetics 175**:155-165.

Latch, E.K., Dharmarajan, G., Glaubitz, J.C., Rhodes, O.E. (2006). Relative performance of Bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation. **Conserv. Genet. 7**:295-302.

Lepage, T., Bryant, D., Philippe, H. e Lartillot, N. (2007). A general comparison of relaxed molecular clock models. **Mol. Biol. Evol. 24**(12):2669-2680.

Luikart, G. e Cornuet, J.M. (1998). Empirical evaluation of a test for identifying recently bottlenecked populations from allele frequency data. **Conserv. Biol. 12**:228-237.

Maddison, W.P. (1997). Gene trees in species trees. **Syst. Biol. 46**(3):523-536.

Maddison, W.P. e Maddison, D.R. (2008). **Mesquite:a modular system for evolutionary analysis**. Version 2.5, http://mesquiteproject.org.

Manel, S., Gaggiotti, O.E. e Waples, R.S. (2005). Assignment methods: matching biological questions techniques with appropriate techniques. **Trends. Ecol. Evol. 20**(3):136-142.

Nei, M. (1978). Estimation of average heterozygosity and genetic distance from a small number of individuals. **Genetics 89**:583-590.

Neigel, J.E. (1997). A comparison of alternative strategies for estimating gene flow from genetic markers. **Annu. Rev. Ecol. Syst. 28**:105-128.

Neigel, J.E. (2002). Is F_{ST} obsolete? **Cons. Genet. 3**(2):167-173.

Neigel, J.E. e Avise, J.C. (1986). Phylogenetic relationships of mitochondrial DNA under various demographic models of speciation. In:**Evolutionary processes and theory**. Nevo, E., Karlin, S. (eds). Academic Press, Nova Iorque, pp 515-534.

Nielsen, R. e Wakeley, J. (2001). Distinguishing migration from isolation:A Markov chain Monte Carlo approach. **Genetics 158**(2):885-896.

Olsen, K.M. (2002). Population history of *Manihot esculenta* (Euphorbiaceae). inferred from nuclear DNA sequences. **Mol. Ecol. 11**:901-911.

Paetkau, D., Slade, R., Burden, M. e Estoup, A. (2004). Genetic assignment methods for the direct, real-time estimation of migration rate: a simulation-based exploration of accuracy and power**. Mol. Ecol. 13**(1):55-65.

Palsboll, P.J. (1999). Genetic tagging:contemporary molecular ecology**. Biol. J. Linn. Soc. 68**:3-22.

Panchal, M. (2007). The automation of Nested Clade Phylogeographic Analysis. **Bioinformatics 23**:509-510.

Panchal, M. e Beaumont, M.A. (2007). The automation and evaluation of the Nested Clade Phylogeographic Analysis. **Evolution 61**:1466-1480.

Pearse, D.E., Eckerman, C.M., Janzen FJ e Avise, J.C. (2001). A genetic analogue of 'mark-recapture' methods for estimating population size:an approach based on molecular parentage assessments. **Mol. Ecol. 10**:2711-2718.

Pearse, D.E. e Crandall, K.A. (2004). Beyond F-ST:Analysis of population genetic data for conservation. **Cons Genet 5**(5):585-602.

Petit, R.J. (2008). The coup de grâce for the nested clade phylogeographic analysis? **Mol. Ecol. 17**:516-518.

Pfenninger, M. e Posada, D. (2002). Phylogeographic history of the land snail *Candidula unifasciata* (Helicellinae, Stylommatophora):Fragmentation, corridor migration, and secondary contact. **Evolution 56**:1776-1788.

Palstra, F.P., O'Connell, M.F. e Ruzzante, D.E. (2007). Population structure and gene flow reversals in Atlantic salmon (*Salmo salar*). over contemporary and long-term temporal scales:effects of population size and life history. **Mol. Ecol. 16**:4504–4522.

Posada, D. e Crandall, KA (1998). Modeltest:testing the model of DNA substitution. **Bioinformatics 14**:817-818.

Posada, D., Crandall, K.A. e Templeton, A.R. (2000). GeoDis:a program for the cladistic nested analysis of the geographical distribution of genetic haplotypes. **Mol. Ecol. 9**:487-488.

Posada, D., Crandall, K.A. e Templeton, A.R. (2006). Nested clade analysis statistics. **Mol. Ecol. Notes 6**:590-593.

Posada, D. e Crandall, K.A. (2001). Intraspecific gene-genealogies:trees grafting into networks. **Trends. Ecol. Evol. 16**:37-45.

Piry, S., Luikart, G. e Cornuet, J.M. (1999). Bottlenek: a computer program for detecting recent reduction in effective population size using allele frequency data. **J. Hered. 90**:502-503.

Pritchard, J.K., Stephens, M. e Donnelly, P. (2000). Inference of population structure using multilocus genotype data. **Genetics 155**(2):945-959.

Pritchard, J.K. e Wen, W. (2004). **Documentation for structure software**: Version 2.1. Disponível em: *http://pritch.bsd.uchicago.edu*.

Pybus, O.G., Rambaut, A. e Harvey, P.H. (2000). An Integrated Framework for the Inference of Viral Population History From Reconstructed Genealogies. **Genetics 155**:1429-1437.

Pybus, O.G. e Rambaut, A. (2002). GENIE:estimating demographic histories from molecular phylogenies. **Bioinformatics 18**:1404-1405.

Rannala, B. e Mountain, J.L. (1997). Detecting immigration using multilocus genotypes. **Proc. Natl. Acad. Sci. USA 94**:9197-9201.

Ray, N., Currat, M. e Excoffier, L. (2003). Intra-deme molecular diversity in Spatially expanding populations. **Mol. Biol. Evol. 20**(1):76-86.

Richards, C.L., Carstens, B.C., Knowles, L.L. (2007). Distribution modelling and statistical phylogeography:an integrative framework for generating and testing alternative biogeographical hypotheses. **J Biogeogr 34**:1833-1845.

Rice, W.R. (1989). Analyzing tables of statistical tests. **Evolution. 43**:223-225.

Rogers, A.R., Harpending, H. (1992). Population-growth makes waves in the distribution of pairwise genetic-differences. **Mol. Biol. Evol.9**(3):552-569.

Roman, J. e Palumbi, S.R. (2003). Whales before whaling in the North Atlantic. **Science 301**:508-510.

Rosenberg, N.A., Burke, T., Elo, K., Feldman, M.W., Freidlin, P.J., Groenen, M.A.M., Hillel, J., Maki-Tanila, A., Tixier-Boichard, M., Vignal, A., Wimmers, K. e Weigend, S. (2001). Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds. **Genetics 159**:699-713.

Rosenberg, N.A. e Nordborg, M. (2002). Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. **Nature Rev. Genet. 3**:380-390.

Rowe, G. e Beebee, T.J.C. (2007). Defining population boundaries:use of three Bayesian approaches with microsatellite data from British natterjack toads (*Bufo calamita*). **Mol. Ecol. 16**:785-796.

Rozas, J., Sánchez-DelBarrio, J.C., Messenguer, X, e Rozas, R. (2003). DNAsp, DNA polymorphism analyses by the coalescent and other methods. **Bioinformatics 19**:2496-2497.

Shapiro, B., Drummond, A.J., Rambaut, A., Wilson, M.C., Matheus, P.E., Sher, A.V., Pybus, O.G., Gilbert, M.T.P., Barnes, I., Binladen, J., Willerslev, E., Hansen, A.J., Baryshnikov, G.F., Burns, J.A., Davydov, S., Driver, J.C., Froese, D.G., Harington, C.R., Keddie, G., Kosintsev, P., Kunz, M.L., Martin, L.D., Stephenson, R.O., Storer, J., Tedford, R., Zimov, S. e Cooper, A. (2004). Rise and fall of the Beringian steppe bison. **Science 306** (5701):1561-1565.

Slatkin, M. (1985). Gene flow in natural populations. **Ann. Rev. Ecol. Syst. 16**:393-430.

Slatkin, M. (1987). Gene flow and the geographic structure of natural populations. **Science 236**:787-792.

Slatkin, M., Hudson, R.R. (1991). Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. **Genetics 129**:555-562.

Slatkin, M. (1995). A measure of population subdivision based on microsatellite allele frequencies. **Genetics 139**(3):1463-1463.

Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. **Genetics 105**:437-460.

Tajima, F. (1989). The effect of change in population size on DNA polymorphism. **Genetics 123**:597-601.

Tallmon, D.A., Luikart, G. e Beaumont, M.A. (2004). Comparative evaluation of a new effective population size estimator based on approximate Bayesian computation. **Genetics 167**:977-988

Taylor, B.L., Chivers, S.J., Sexton, S. e Dizon, A.E. (2000). Evaluating dispersal estimates using mtDNA data: Comparing analytical and simulation approaches. **Conserv. Biol. 14**(5):1287-1297.

Templeton, A.R., Crandall, K.A. e Sing, C.F. (1992). A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. **Genetics 132**:619-633.

Templeton, A.R., Routman, E. e Phillips, C.A. (1995). Separating population structure from population history: a cladistic analyses of the

geographical distribution of mitochondrial DNA haplotypes in the tiger salamander, *Ambystona tigrinum*. Genetics 140:767-782

Templeton, A.R. (1998). Nested clade analyses of phylogeographic data:testing hypotheses about gene flow and population history. **Mol. Ecol. 7**:381-397.

Templeton, A.R. (2001). Using phylogeographic analyses of gene trees to test species status and processes. **Mol. Ecol. 10**:779-791.

Templeton, A.R. (2002). Out of Africa again and again. **Nature 416**:45-51.

Templeton, A.R. (2004). Statistical phylogeography:methods of evaluating and minimizing inference errors. **Mol. Ecol. 13**:789-810.

Templeton, A.R. (2008). Nested clade analysis:an extensively validated method for strong phylogeographic inference. **Mol. Ecol. 17**:1877-1880.

Thorne, J.L., Kishino, H. e Painter, I.S. (1998). Estimating the rate of evolution of the rate of molecular evolution. **Mol. Biol. Evol. 15**(12):1647-1657.

Verhoeven, K.F.J., Simonsen, K.L. r McIntyre, L.M. (2004). Implementing false discovery rate control: increasing your power. **OIKOS 108**:643-647.

Wade, P.R. (2000). Bayesian methods in conservation biology. Conserv. Biol. 14 (5):1308-1316

Wakeley, J. e Hey J. (1997). Estimating ancestral population parameters. **Genetics 145**:847-855.

Wakeley, J. (2003). Inferences about the structure and history of populations: coalescents and intraspecific phylogeography. **In:** Singh, R.S., Uyenoyama, M.K. (eds). **The evolution of population biology**. Cambridge University Press. pp. 193-199.

Wallace, A.R. (1852). On the monkeys of the Amazon. **Proc. Zool. Soc. Lond. 20**:107-110

Wang, S., Lewis, C.M., Jakobsson, M., Ramachandran, S., Ray, N., Bedoya, G., Rojas, W., Parra, M.V., Molina, J.A., Gallo, C., Mazzotti, G., Poletti, G., Hill, K., Hurtado, A.M., Labuda, D., Klitz, W., Barrantes, R., Bortolini, M.C., Salzano, F.M., Petzl-Erler, M.L., Tsuneto, L.T., Llop, E., Rothhammer, F., Excoffier, L., Feldman, M.W., Rosenberg, N.A. e Ruiz-Linares, A. (2007). Genetic Variation and Population Structure in Native Americans. **PLOS Genetics 3** (11):185.

Waples, R.C. (1998). Separating the wheat from the chaff:patterns of genetic differentiation in high gene flow species. **J. Hered. 89** (5):438-450.

Waples, R.S. e Gaggiotti, O. (2006). What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. **Mol. Ecol. 15**:1419-1439.

Weir, B.S., Cockerham, C.C. (1984). Estimating f-statistics for the analysis of population-structure. **Evolution 38**(6):1358-1370.

Westfall, P.H. e Young, S.S. (1993). On adjusting P-values for multiplicity. **Biometrics 49**:941-944.

Whitlock, M.C. e McCauley, D.E. (1999). Indirect measures of gene flow and migration:F-ST not equal 1/(4Nm+1). **Heredity 82**:117-125.

Wilson, G.A. e Rannala, B. (2003). Bayesian inference of recent migration rates using multilocus genotypes. **Genetics 163**:1177–1191.

Wright, S. (1931). Evolution in Mendelian populations. **Genetics 16**:97-159.

Wright, S. (1951). The genetical structure of populations. **Ann. Eugen. 15**:323-354.

Wright, S. (1965). The interpretation of population structure by F-statistics with special regard to systems of mating. **Evolution 19**:395-420.

Wright, S. (1978). **Evolution and the genetics of populations**. The University of Chicago Press, Londres.

Yang, Z.H (1997). PAML:a program package for phylogenetic analysis by maximum likelihood. **Comp. Appl. Biosci. 13**:555-556.

Yang, Z.H. (2007). PAML:Phylogenetic analysis by maximum-likelihood. **Mol. Biol. Evol. 24**:1582-1591.

Página deixada em branco

Capítulo 22

A genética e a conservação da Natureza

Antonio M. Solé-Cava (sole@biologia.ufrj.br)
Laboratório de Biodiversidade Molecular, Departamento de Genética
Instituto de Biologia, Universidade Federal do Rio de Janeiro

Haydée A. Cunha (haydeecunha@biologia.ufrj.br)
Laboratório de Biodiversidade Molecular, Departamento de Genética
Instituto de Biologia, Universidade Federal do Rio de Janeiro
Laboratório de Mamíferos Aquáticos, Faculdade de Oceanografia
Universidade do Estado do Rio de Janeiro

"(...) *em um mundo cheio de feridas* (...) *um cientista pode endurecer-se e fingir que as consequências da ciência que faz não lhe dizem respeito, ou pode ser o doutor que vê os sinais da morte em uma população que acredita que está bem e que não quer ouvir dizer que não está."* (Aldo Leopold, 1943)

22.1. Introdução

O desenvolvimento de nossa espécie permitiu prodígios incríveis, como a comunicação entre locais distantes, o tratamento de muitas doenças, as altas tecnologias, o desenvolvimento da ciência, das artes e do lazer, e o transporte em grande escala de produtos agrícolas e industriais. Esses desenvolvimentos permitiram que a população humana crescesse de maneira sem precedentes na história de nosso planeta, de modo que somos hoje uma das espécies animais mais abundantes: estima-se que, em 2050, a população mundial chegará a quase nove bilhões de pessoas (Cohen, 2003). Uma consequência desse crescimento foi a destruição de grandes áreas ocupadas anteriormente por outras espécies. A biomassa de nossa espécie em 2008 era de aproximadamente 240 milhões de toneladas e consumimos cerca de 30% de toda a produção primária terrestre, incluídas a produção agrícola, pecuária e de florestas para produção de papel e madeira (Haberl *et al.*, 2007). Nunca uma única espécie consumiu uma proporção tão grande dos recursos naturais. Podemos dizer, portanto, que o crescimento da população humana foi feito, até hoje, à custa da destruição da biodiversidade, causando a extinção de um número enorme de espécies (http://www.iucnredlist.org/).

Apesar de ser muito popular hoje em dia, a expressão "diversidade biológica" somente começou a ser usada na literatura há pouco tempo (Norse e McManus, 1980). A expressão "biodiversidade" é mais recente ainda, tendo sido usada pela primeira vez em 1985, por W.G. Rosen, para uma reunião do Foro Nacional de Biodiversidade (norte-americano), em Washington. Desde sua origem, a expressão diversidade biológica já trazia a idéia do conjunto da variabilidade ecológica (número de espécies de uma comunidade e suas interações) e genética (diversidade de alelos nos vários locos de uma espécie). O componente genético da biodiversidade é fundamental, pois é a variação genética que fornece o material básico para a seleção natural e, portanto, para a evolução de todas as espécies (Allcock *et al.*, 1995). O objetivo central da genética aplicada à conservação é o uso de marcadores moleculares e sua interpretação ecológica e evolutiva para ajudar a minimizar os danos causados pelas atividades humanas nas populações das espécies.

A ciência da genética para a conservação (do inglês, "conservation genetics") foi criada há cerca de 30 anos e os primeiros livros a reverem o assunto foram publicados logo em seguida (Soulé e Wilcox, 1980; Frankel e Soulé, 1981; Schonewald-Cox *et al.*, 1983). Naquele momento, o campo resumia-se praticamente a estimativas de variabilidade genética (heterozigosidade) e sua extrapolação para a estimativa do tamanho efetivo de populações ameaçadas ou que haviam sofrido estrangulamentos populacionais recentes ("bottlenecks" em inglês, também chamados de gargalos populacionais; Soulé, 1980). Por causa dessa limitação, a aplicação da genética para a conservação foi criticada no final dos anos 80 como um desperdício de dinheiro e de esforços, que poderiam ser mais bem usados na manutenção de parques e reservas ambientais, pois as questões demográficas—como o número absoluto de indivíduos e variações estocásticas nesses números—seriam mais importantes do que as questões genéticas, pelo menos da forma que eram apresentadas na época (Lande, 1988). No entanto, com a maior compreensão pelos geneticistas dos problemas enfrentados pelos conservacionistas, que, por sua vez, compreenderam melhor o potencial que marcadores genéticos têm para a abordagem de seus problemas, a genética voltou a ser vista como uma ciência útil para a conservação. Dessa forma, ao contrário dos livros produzidos nos anos 80, uma grande diversidade de problemas foi abordada em publicações recentes sobre genética aplicada à conservação (boas revisões recentes podem ser vistas em Avise e Hamrick, 1996; Avise, 1998; Allendorf e Luikart, 2006—para trabalhos feitos no Brasil, veja Galetti Jr. *et al.*, 2008) e o campo está claramente em expansão, incluindo a publicação de uma revista inteiramente dedicada ao assunto ("Conservation Genetics"), assim como de livros-texto de boa qualidade (*e.g.*, Frankham *et al.*, 2002; Frankham *et al.*, 2004; Allendorf e Luikart, 2006; Mills, 2006).

A variabilidade genética, também chamada de Biodiversidade Molecular, além de importante para a evolução, pode ser usada como instrumento de investigação por ecólogos e sistematas em diversos ramos, como para verificar as afinidades e os limites entre as espécies, para identificar a origem de bioinvasões e auxiliar em seu controle, para detectar modos de reprodução

e estrutura familiar, para estimar níveis de migração e dispersão nas populações e até mesmo para ajudar na identificação de restos animais, como conteúdos estomacais e produtos industrializados (principalmente peles e carne) de espécies ameaçadas de extinção (Avise, 2004). Os dados básicos para esses estudos são os chamados marcadores moleculares, que são locos gênicos que apresentam alguma variabilidade no escopo do problema a ser estudado (Silva e Russo, 2000; Avise, 2004).

Lamentavelmente, os recursos disponíveis para a preservação da biodiversidade são, em geral, muito reduzidos por conta de decisões políticas dos governos que raramente priorizam a conservação ambiental. Portanto, é necessário evitar que recursos tão necessários na criação, manutenção e fiscalização de parques e reservas nacionais sejam desviados, apenas por questões de modismo, para pesquisas genéticas (que são geralmente bastante caras), a não ser nos casos em que os problemas estão claramente definidos e para os quais a abordagem genética realmente pode fornecer dados originais impossíveis de obter de outras maneiras. Esse é um dos problemas mais frequentes em trabalhos na área: apesar de os artigos terem objetivos bem definidos para pesquisa básica, muitas vezes eles têm implicações vagas para a Conservação em suas conclusões (Vernesi *et al.*, 2008). Genética para conservação não deveria apenas ser a genética de populações ou filogeografia aplicada a organismos ameaçados de extinção ou carismáticos. Para que um trabalho se pudesse intitular "genética para conservação", ele precisaria ter implicações claras para a definição de políticas e estratégias para a conservação das espécies estudadas. Por exemplo, em 2006, noventa e seis artigos indexados no Web of Science tinham a expressão "conservation genetics" em seus títulos, mas apenas 25% deles tinham indicações claras nas conclusões sobre o emprego dos resultados no gerenciamento de espécies (Vernesi *et al.*, 2008). Uma análise rápida de publicações autointituladas como de "genética para conservação" revela que muitas delas, em suas conclusões, apenas adicionam a frase "os resultados desta pesquisa terão consequências para a conservação da espécie" (e suas variantes), sem indicar quais são, exatamente, essas consequências. Esse fato é preocupante, porque os tomadores de decisão não necessariamente são treinados para a interpretação dos resultados, tornando necessário que os pesquisadores explicitem quais medidas são apoiadas pelos resultados obtidos (Latta, 2008). Se o objetivo desses trabalhos é guiar estatégias de manejo e conservação eficientes, não é aceitável que os autores deleguem a interpretação dos resultados aos tomadores de decisão.

Naturalmente devemos lutar, a partir da educação e da sensibilização da opinião pública (em última análise, a moeda corrente com maior valor nas decisões governamentais), para que a escolha não seja entre o estudo e a preservação da biodiversidade. Mas também devemos ser seletivos na aplicação de recursos na genética para conservação, a fim de evitar desperdícios. Muitos problemas que podem ser abordados por ecólogos ou sistematas também poderiam ser abordados molecularmente, com maior ou menor eficiência. Entretanto, seria lamentável se os recursos para projetos de conservação, já tão limitados, fossem concentrados em abordagens moleculares, que geralmente são bastante caras, apenas por estarem mais na moda ou por terem o "*sex appeal*" das tecnologias modernas de DNA. Portanto, devemos concentrar o uso da genética para conservação naqueles aspectos que sejam importantes e ao mesmo tempo difíceis—ou impossíveis—de serem abordados de outras formas. Assim, a genética aplicada à conservação revela-se uma ciência útil para o trabalho dos outros profissionais interessados na conservação da biodiversidade e a colaboração com eles passa a ter efeito multiplicativo nos resultados obtidos.

O objetivo deste texto é apresentar, brevemente, como os marcadores moleculares, interpretados sob a luz da genética de populações, podem ser usados para auxiliar no estudo e na conservação da biodiversidade.

22.2. Marcadores Moleculares

A matéria bruta dos estudos em biodiversidade molecular é a mesma envolvida na evolução das espécies: a variabilidade gênica. É essa variabilidade que nos permite comparar indivíduos, populações ou espécies diferentes. Em outros capítulos deste livro, foram apresentadas as principais técnicas de estimativa da variabilidade gênica (aloenzimas, no Capítulo 17; RFLP, no Capítulo 18; RAPDs, microssatélites, sequenciamento de DNA, no Capítulo 19) e, portanto, não vamos nos estender aqui na sua descrição (outras revisões sobre marcadores moleculares podem ser encontradas em Hoelzel, 1992; Parker *et al.*, 1998; Silva e Russo, 2000; Avise, 2004). O mais importante, do ponto de vista da genética para a conservação, é que marcadores moleculares diferentes podem ter taxas de substituição/evolução diferentes, de modo que, através de uma escolha judiciosa desses marcadores, podemos estudar desde problemas de identificação de indivíduos até a detecção de espécies crípticas ou formulação de hipóteses filogenéticas em grupos supra-específicos. Os marcadores disponíveis atualmente podem ser classificados de acordo com a existência de dominância: padrões com dominância, como os gerados com RAPD ou AFLP, são menos úteis em análises populacionais, pois exigem um pressuposto importante e frequentemente violado—o equilíbrio de Hardy-Weinberg—para que possam ser estimadas as frequências gênicas a partir dos dados brutos obtidos. Podem ser classificados ainda em relação à possibilidade de viés sexual (DNA mitocondrial em geral é transmitido apenas pelas fêmeas), às taxas evolutivas (microssatélites e RAPDs evoluem muito rapidamente, aloenzimas evoluem mais lentamente), entre outros critérios que estão relacionados na Tabela 22.1.

A escolha do método a ser usado na abordagem de cada problema depende de vários critérios científicos e não científicos. Entre os critérios científicos está, em primeiro lugar, a adequação do grau de variabilidade do marcador molecular escolhido ao nível de divergência que se deseja estudar (Capítulo 16 deste livro). Marcadores que evoluem rapidamente são úteis para o estudo de indivíduos, famílias e populações, enquanto que marcadores que evoluem mais lentamente são mais bem utilizados no estudo de espécies ou táxons supra-específicos. Essa escolha é crítica: se escolhermos um marcador que evolui de forma demasiadamente lenta para o nível estudado, teremos pouca variabilidade e uma saturação de plesiomorfias nos dados (ou seja, todas as semelhanças observadas serão devidas à ancestralidade dos alelos e não haverá eventos novos capazes de discriminar os grupos estudados). Se, por outro lado, escolhermos um marcador que evolui rápido demais para o nível escolhido, teremos um excesso de variabilidade e uma saturação de homoplasias nos dados (ou seja, as semelhanças observadas serão frequentemente devidas à convergência acidental dos alelos—devido ao caráter finito do espaço amostral—e cometeremos erros na discriminação dos grupos).

Outro critério importante na escolha é o tipo de material disponível para estudos. Em estudos com aloenzimas, por exemplo, necessitamos de quantidades maiores de material

Tabela 22.1. Comparação dos principais métodos de estimativa da biodiversidade molecular. P&D – custo de pesquisa e desenvolvimento para adequação da técnica a cada projeto; Dados – velocidade na produção de dados; Técnica – facilidade de interpretar geneticamente os dados. Abreviações: AFLP, "*amplified fragment length polymorphism*"; PCR/RFLP, "*Polymerase chain reaction/ Restriction fragment length polymorphism*"; SNP, "*single nucleotide polymorphism*"; RAPD, "*Random amplified polymorphic DNA*".

Abordagem	P&D	Custo	Dados	Herança	Técnica	Maior problema
AFLP	baixo	médio	rápido	dominante	média	Falta de homologia, dominância
Aloenzimas	baixo	médio	muito rápido	codominante	difícil	Precisa de material fresco
Microssatélites	alto	médio	rápido	codominante	média	Alelos nulos associados com o uso de primers heterólogos[1]. Alto custo de desenvolvimento de primers específicos.
Minissatélites	baixo	médio	rápido	codominante	difícil	Falta de homologia (exceto SL-VNTR)
PCR/RFLP	baixo	baixo	rápido	variável	fácil	Escolha do gene
SNP	médio	médio	muito rápido	codominante	fácil	Busca de *primers*, desenvolvimento dos sistemas. Difícil em espécies pouco conhecidas
RAPD	baixo	baixo	muito rápido	dominante	fácil	Falta de homologia, dominância, baixa consistência metodológica
Sequenciamento	baixo	médio	rápido	variável	fácil	Escolha do gene

[1] Primers desenvolvidos para uma espécie diferente da estudada.

biológico (pelo menos 200 mg para uma análise completa de 30 locos enzimáticos), que deve necessariamente estar fresco ou congelado (as enzimas se desnaturam ou são digeridas por proteases quando mantidas à temperatura ambiente, mesmo quando o material está preservado em solventes como o álcool). Essa limitação pode ser importante se pretendemos estudar espécies raras (a coleta de tecido geralmente é destrutiva) ou muito pequenas. Esse problema não existe para a maioria dos métodos com DNA graças ao advento da PCR, que permite o uso de quantidades muito pequenas de tecido, incluindo aquelas preservadas em álcool ou desidratadas e até mesmo de ossos. Embora técnicas como sequenciamento de DNA possam ser onerosas demais em estudos populacionais em que um grande número de indivíduos deve ser analisado, estudos com espécies ameaçadas evidentemente nunca empregam um tamanho amostral tão grande.

Entre os critérios não científicos que costumam determinar a escolha do marcador está a tradição do laboratório; frequentemente, os laboratórios se tornam proficientes para um número limitado de técnicas e as aplicam ao maior número possível de problemas (Solé-Cava e Thorpe, 1994). Esses autores lembram um velho ditado russo que diz que "para quem tem um martelo novo, tudo parece prego". Outro critério não científico refere-se à moda: técnicas mais "modernas" ou "sofisticadas" são frequentemente consideradas melhores—e mais fáceis de obter financiamentos—do que técnicas mais antigas, mesmo quando as duas oferecem os mesmos tipos de resultados (Allcock *et al.*, 1995). Um aspecto não científico, mas muito importante na escolha do método a ser usado é o custo: para qualquer problema biológico, podemos escolher um número grande de abordagens. Faz parte do bom senso sermos financeiramente parcimoniosos quando tivermos que escolher entre técnicas igualmente informativas mas com custos muito diferentes.

A escolha do marcador molecular a ser usado depende, portanto, de vários fatores. O fundamental é que o problema a ser estudado esteja bem definido, que exista uma adequação do grau de polimorfismo do marcador escolhido ao tipo de divergência evolutiva a ser estudada, que os pressupostos da análise dos dados sejam bem explicitados e, por fim, que o pesquisador evite o fetiche da técnica, não perdendo de vista o problema original em seu estudo.

22.3. Aplicações de Marcadores Moleculares na Conservação das Espécies

Como dito acima, marcadores moleculares podem ser usados como auxiliares para pesquisas em diferentes campos. Marcadores desse tipo foram usados, com sucesso, por exemplo, para: (a) estimar os níveis de heterozigosidade e relacioná-los com parâmetros importantes na sobrevivência das espécies, como eficiência reprodutiva e resistência a doenças; (b) analisar estruturas familiares, os efeitos da reprodução assexuada na população e determinar o sexo de animais com pouco dimorfismo sexual externo; (c) estimar o tipo de distribuição espacial e temporal das populações em relação ao fluxo gênico, permitindo desenhar políticas adequadas de parques e reservas; (d) verificar a biodiversidade nominal e os níveis de endemismo e cosmopolitismo das espécies, por meio de estudos de sistemática molecular; (e) identificar e acompanhar a dispersão de espécies bioinvasoras; e (f) identificar a origem de produtos industrializados para controlar a comercialização fraudulenta de espécies de venda proibida ou restrita.

Essas aplicações serão brevemente comentadas a seguir.

22.3.1. Heterozigosidade e fragilidade populacional

Populações naturais normalmente têm níveis altos de variação genética. Essa variação é introduzida continuamente nas populações por mutação ou migração de indivíduos de outras populações e é perdida por deriva genética, por endocruzamento e, no caso de genes não neutros, pela maior parte dos tipos de seleção natural (Nei, 1987). As medidas de variabilidade são tratadas no Capítulo 21, mas basicamente são estimadas ou a variabilidade nas frequências dos alelos de um gene, pela heterozigosidade ou diversidade haplotípica, ou as variabilidades qualitativas entre as sequências desses alelos, pela diversidade nucleotídica.

Um conceito importante para o estudo de populações ameaçadas é o seu **tamanho efetivo,** N_e (Wright, 1931). O tamanho efetivo das populações não é igual ao número total de indivíduos, como normalmente se considera em ecologia. O tamanho efetivo leva em conta o número de indivíduos que está **efetivamente** contribuindo para a evolução de uma população. Ou seja, o tamanho efetivo de uma população tem um componente histórico e, em um dado momento, considera apenas os indivíduos participando na produção da próxima geração, excluindo os indivíduos jovens ou

velhos demais para a reprodução. Por exemplo, quando a proporção sexual na população se afasta de 1:1 entre machos e fêmeas, os indivíduos do sexo mais raro acabam contribuindo mais para a próxima geração, o que diminui a variabilidade total. Assim, o tamanho efetivo de população é afetado pela proporção sexual, de acordo com a fórmula

$$N_e = 4N_mN_f / (N_m + N_f),$$

onde N_m e N_f são, respectivamente, os números de machos e fêmeas contribuindo para a reprodução da espécie. Por exemplo, as populações do lobo-marinho-do-sul (*Arctocephalus australis*) foram muito exploradas na costa do Peru, além de sofrerem reduções populacionais devido à escassez de alimento causada pelos eventos periódicos de aquecimento das águas (fenômeno conhecido como El Niño). Em um senso populacional feito em 1996, foram contados 24.481 indivíduos, dos quais 10.720 eram fêmeas em idade reprodutiva e 2.903 eram machos em idade reprodutiva. O lobo-marinho-do-sul apresenta um sistema poligínico (mais de uma fêmea para cada macho), com formação de haréns, que faz com que alguns machos fecundem um grande número de fêmeas, enquanto muitos não conseguem se reproduzir. Usando a fórmula acima, percebe-se que o N_e dos 13.623 indivíduos adultos observados era, de fato, de apenas 9.138 (de Oliveira *et al.*, 2006).

O efeito da proporção sexual no tamanho efetivo populacional tem implicações importantes para o manejo de populações ameaçadas: por exemplo, se temos espaço/recursos para manter 50 tigres em cativeiro em um programa de recolonização, poderia ser demograficamente mais interessante ter um macho e 49 fêmeas, pois em última análise é o número de fêmeas que vai determinar o número de filhotes produzidos. No entanto, aplicando-se a fórmula verificamos que, nesse caso, o tamanho efetivo de população seria de apenas (4 X 1 X 49) / (1+49), ou seja, o N_e seria apenas 3,92 indivíduos! Se, por outro lado, tivéssemos 25 machos e 25 fêmeas, todos adultos e em idade reprodutiva, teríamos um tamanho efetivo de 50, mas talvez nesse caso o número de filhotes produzidos fosse pequeno para fins de recolonização. A decisão na administração dessa população vai depender, portanto, da obtenção de uma proporção sexual que maximize o número de filhotes produzidos com um prejuízo menor na perda de variabilidade gênica.

O tamanho efetivo de população, como dissemos, também incorpora um fator histórico: uma população com um número grande de indivíduos, originados há poucas gerações de um número pequeno de ancestrais, terá um tamanho efetivo menor do que populações que permaneceram demograficamente estáveis há muito tempo. O tamanho efetivo de uma população em um dado momento pode ser estimado pela média harmônica dos tamanhos efetivos das *n* gerações passadas, de acordo com a fórmula

$$N_e = n / [\Sigma(1/N_i)],$$

onde N_i é o tamanho efetivo da população na geração i. Imaginemos, por exemplo, que, no caso anterior dos tigres, o macho e as 49 fêmeas da primeira geração produzissem em sua vida um total de 200 filhotes (100 machos e 100 fêmeas) e que esses filhotes, por sua vez, produzissem uma F_2 de 800 filhotes (400 machos e 400 fêmeas). Ao final dessas três gerações o tamanho efetivo da população dessa F_2 seria de apenas $N_e = 3/[(1/3,92)+(1/200)+(1/800)] = 11,48$ indivíduos. Isso indica que a recuperação do tamanho efetivo de população é muito lenta após um estrangulamento populacional. Voltando ao nosso exemplo dos lobos marinhos, vimos que o tamanho efetivo populacional, corrigido para a proporção sexual, em 1996 no Peru foi de 9.138. No período 1997-1998, houve um forte evento de El Niño, de modo que, em 1999, o tamanho efetivo populacional havia sido reduzido para 1.220 (de Oliveira *et al.*, 2006). Assim, o tamanho efetivo populacional, levando em conta as oscilações temporais, seria de $2/[(1/9.138)+(1/1.220)] = 2.153$ indivíduos.

A relação existente entre variação genética e tamanho efetivo da população pode ser expressa na seguinte fórmula, com base no modelo de alelos infinitos (Wright, 1978):

$$h_e = \frac{4N_e\mu}{(4N_e\mu+1)},$$

onde h_e é a heterozigosidade esperada (sob equilíbrio de Hardy-Weinberg) e μ é a taxa de mutação para o gene estudado. A diminuição do tamanho efetivo da população é, portanto, um dos principais responsáveis pela perda de variabilidade em populações ameaçadas de extinção.

Um problema adicional enfrentado por populações pequenas (<1000 indivíduos) é a fixação aleatória de alelos deletérios pré-existentes ou oriundos de novas mutações. As populações de tamanho normal costumam apresentar um grande número de alelos deletérios com frequências reduzidas (a chamada "carga genética" da população). Esses alelos, geralmente recessivos, são mantidos em frequências baixas pela ação leve, mas constante, da seleção natural sobre os homozigotos. No entanto, em populações pequenas, a não ser em regimes seletivos muito fortes, pode acontecer a fixação aleatória desses alelos deletérios. Uma vez fixados, esses alelos não são mais sujeitos à ação da seleção natural, pela ausência de alternativa genética que possa ser selecionada. Dessa forma, enquanto que, para os alelos letais ou altamente deletérios, o endocruzamento significa uma "purificação", para alelos levemente deletérios ele pode significar a imposição de uma carga genética crônica que pode levar a população ao colapso (Lynch, 1995).

Alguns autores estimam que um tamanho populacional efetivo mínimo para evitar depressão por endocruzamento seja de 50 indivíduos (Soulé e Wilcox, 1980; Franklin e Frankham, 1998). Esse número sobe para 500 indivíduos se se deseja também evitar a perda de variabilidade devido à deriva gênica (Franklin e Frankham, 1998). Esses valores, embora tenham sido empregados como padrão, originaram-se de estudos realizados com drosófilas na década de 1950 e dependem naturalmente da variabilidade genética e da arquitetura genética dessa variação, que é importante fenotipicamente. Por esse motivo, devemos ser cautelosos na determinação de valores mínimos de tamanhos populacionais para fins de conservação. Por exemplo, uma vez perdida a variabilidade gênica, ela só é recuperada muito lentamente (por mutação ou migração), de modo que, mesmo tomando providências para aumentar o tamanho populacional de uma espécie, ela pode continuar ameaçada de extinção (Avise, 2004). Um conceito interessante para a conservação é o de **tamanho mínimo viável de uma população** (MVPS, "minimum viable population size"), que expressa o tamanho (em número de indivíduos) que uma população deve ter para poder persistir na natureza (Shaffer, 1981). Estima-se que esse número seja bem maior que os números calculados para populações manejadas e uma meta-análise recente, de dados de 212 espécies, chegou a um valor mediano de MVPS de mais de 4.000 indivíduos (intervalo de 95% de confiança = 3577–5129 indivíduos; Traill *et al.*, 2007).

Um dos exemplos mais claros do tempo que leva para recuperar a variabilidade perdida é o caso dos elefantes-marinhos-do-norte, *Mirounga angustirostris*. Essa espécie foi caçada de 1820 a 1880 devido ao alto valor de seu óleo. A caça foi tão intensa

que, por volta de 1884, a espécie foi considerada extinta, até que se encontrou uma população sobrevivente na Ilha de Guadalupe, na Baixa Califórnia. Alguns censos indicaram que menos de 20 indivíduos haviam sobrevivido, quando os governos americano e mexicano tomaram medidas radicais, aprovando leis rígidas para impedir a extinção desses animais. A partir daí a população começou a se expandir e fundar novas colônias, aumentando o tamanho populacional, hoje estimado em 175.000 indivíduos. No entanto, esse estrangulamento do tamanho populacional não ocorreu sem consequências. Como o tamanho efetivo da população é influenciado fortemente pelos estrangulamentos passados, os níveis de heterozigosidade nessa espécie permaneceram extremamente baixos, a ponto de vários estudos não conseguirem detectar variação genética alguma em 55 locos de aloenzimas (Bonnell e Selander, 1974) e pouquíssima variação em marcadores com alto grau de polimorfismo, como a região controle do DNA mitocondrial (Slade, 1998). Poderia ser argumentado que os elefantes-marinhos, como muitos grandes mamíferos, teriam normalmente variabilidade genética baixa. No entanto, outra espécie do mesmo gênero, o elefante-marinho-do-sul (*M. leonina*), que apresenta tamanhos populacionais elevados e alta conectividade entre populações (Fabiani *et al.*, 2003), apresenta variabilidade gênica cerca de cinco vezes maior (Slade, 1998), o que indica que a baixa variabilidade não é uma característica do gênero e, mais provavelmente, está realmente relacionada ao estrangulamento populacional recente dos elefantes marinhos do norte. Isso foi confirmado por comparações diretas de genes extraídos de peles de elefante-marinho, que mostraram que, antes do grande estrangulamento populacional de 1890, a variabilidade gênica na região controle do DNA mitocondrial era dezenas de vezes maior do que a atual (Weber, 2000). A redução da diversidade genética também foi documentada no golfinho-de-Hector (*Cephalorhynchus hectori*), uma espécie ameaçada endêmica da Nova Zelândia. Na Ilha Norte, há uma população pequena (134 indivíduos), que não recebe migrantes de outras áreas, e que apresenta apenas um haplótipo na região controle mitocondrial. Porém, quando amostras de museu dessa população foram analisadas, foram encontrados outros haplótipos, mostrando que historicamente ela possuía diversidade maior. A curva da diversidade haplotípica ao longo do tempo mostrou uma queda acentuada nos anos 70, coincidindo com a introdução das redes de espera na Ilha Norte. Essas redes passaram a ser a maior causa de mortalidade desses golfinhos (Pichler e Baker, 2000). Dessa forma, foi possível identificar não só a perda da diversidade nessa população, mas sua causa mais provável. Outro exemplo de espécie ameaçada com baixíssimo nível de heterozigosidade é o lobo-cinza (*Canis lupus*), que tem 50% da heterozigosidade de outros canídeos e apenas um haplótipo mitocondrial nas populações naturais (Randi, 1993; Randi *et al.*, 1995; Wayne, 1996), apesar de estar atualmente mostrando sinais de recuperação populacional, com expansão de sua faixa de distribuição nos Alpes Italianos e Suíços (Fabbri *et al.*, 2007).

Os guepardos (*Acinonyx jubatus jubatus*) também têm baixa variação gênica tanto em aloenzimas (52 locos monomórficos) como no complexo principal de histocompatibilidade (cuja sigla, em inglês, é MHC, de *Major Histocompatibility Complex*, um grupo de proteínas ligadas ao sistema imunológico dos vertebrados). Esse complexo costuma apresentar alta variabilidade, associada a uma maior resistência imunológica. Guepardos são geneticamente tão homogêneos que seus indivíduos aceitam enxertos de pele uns dos outros sem nenhuma rejeição (O'Brien *et al.*, 1983; Menotti-Raymond e Stephen, 1993; O'Brien, 1994). O mesmo ocorre com o leão-asiático (*Panthera leo persica*), cuja população é muito reduzida (cerca de 250 animais, depois de um estrangulamento populacional que fez com que a população chegasse a 20 indivíduos, no início do século XX), e, consequentemente, não tem nenhuma variação aloenzimática (em 50 locos testados; Wildt *et al.,* 1987) ou no sistema MHC (avaliada com RFLP; O'Brien, 1994).

A perda de variabilidade gênica frequentemente indica altos graus de endocruzamento, que podem estar associados a problemas importantes na reprodução e biologia das espécies. Como foi visto, o guepardo apresenta baixíssimos níveis de heterozigosidade e nessa espécie é observada contagem de esperma dez vezes menor que em outros felinos, elevado índice de malformações em embriões (79%) e também de assimetria bilateral do esqueleto—todos fenômenos comuns em espécies com alto grau de endocruzamento (O'Brien, 1994). No lobo-cinza-mexicano (*Canis lupus baileyi*), as linhagens endocruzadas apresentam uma redução significativa de qualidade em seu esperma, tanto sob análise microscópica como na eficiência reprodutiva (Asa *et al.*, 2007). Da mesma forma, o leão-asiático possui contagem de esperma dez vezes menor que o leão-da-Tanzânia (*Panthera leo*), que tem tamanhos populacionais muito mais elevados. Anomalias nos espermatozóides, que ocorrem em 25% dos espermatozóides dos leões-da-Tanzânia, atingem 65% dos espermatozóides dos leões-asiáticos. Finalmente, os níveis de testosterona circulante nos leões-asiáticos são dez vezes menores (0,1 ng/ml) que nos leões-da-Tanzânia (1,3 ng/ml; Wildt *et al.*, 1987). O endocruzamento pode reduzir, portanto, a fecundidade das populações envolvidas. Baseado nesses e em outros resultados, foi sugerido que 10% na redução do tamanho da população acarretaria, devido à depressão de endocruzamento, uma redução de 10 a 25% em seu desempenho reprodutivo (Franklin e Frankham, 1998). Isso, por sua vez, pode causar uma nova redução no tamanho efetivo da população e, portanto, reduzir a habilidade de persistência da população ao longo do tempo (Gilpin e Soulé, 1986). Espécies presas nesse círculo vicioso eventualmente se extinguem, principalmente se levarmos em conta que, além dos problemas genéticos, existem outros elementos comportamentais (algumas espécies precisam de uma densidade populacional mínima para se reproduzirem) e demográficos que podem precipitar a extinção dessas espécies (Andrewartha e Birch, 1954).

Além dos efeitos negativos na reprodução, a baixa heterozigosidade também pode estar correlacionada à taxa de mortalidade de uma espécie. Essa correlação já foi demonstrada, por exemplo, nos leões-marinhos-da-Califórnia (*Zalophus californianus*), nos quais filhotes homozigotos para vários locos mostraram maior vulnerabilidade a infecções pelo verme-gancho (*Uncinaria* spp.), que é a maior causa isolada de mortalidade de filhotes da espécie (causando até mais de 40% das mortes) (Acevedo-Whitehouse *et al.*, 2006). A mesma associação foi observada na foca-do-porto (*Phoca vitulina*): indivíduos jovens homozigotos foram mais susceptíveis à infecção pelo verme-do-pulmão (*Otostrongylus circumlitus*) (Rijks *et al.*, 2008).

Parece claro, então, que um dos objetivos a serem buscados na conservação de espécies ameaçadas é a diminuição do endocruzamento, principalmente em espécies mantidas em cativeiro para recolonização (Borlase *et al.*, 1993; Avise e Hamrick, 1996).

Existem alguns dados, no entanto, que indicam que a baixa heterozigosidade talvez não seja tão problemática. Por exemplo, os elefantes-marinhos, discutidos anteriormente, apesar da baixa heterozigosidade, parecem estar vivendo muito bem, com tamanhos populacionais elevados e ainda em crescimento. O mesmo foi observado em castores da Escandinávia, que quase se extinguiram devido à caça no século passado e agora apresentam, previsivelmente, uma baixíssima heterozigosidade, mas têm populações estáveis de mais de 100.000 animais (Ellegren *et al.*, 1993). Outro exemplo é o tordo-preto ("black robin", *Petroica*

traversi) das Ilhas Chatham, na Nova Zelândia, uma ave cuja população permaneceu abaixo de 200 indivíduos pelos últimos 100 anos, tendo atingido um mínimo de cinco indivíduos em 1980, dos quais um casal conseguiu se reproduzir e a população, sob intensa proteção, voltou a aumentar novamente até 200 indivíduos nos anos 90. Esses indivíduos têm baixíssima heterozigosidade, mas são saudáveis e se reproduzem normalmente, indicando que essa baixa variação gênica não parece estar prejudicando suas chances de sobrevivência, pelo menos nas condições atuais (Ardern e Lambert, 1997).

Uma possível explicação para esse sucesso reprodutivo, apesar da baixa variabilidade, é que os exemplos dados são de animais que teriam passado suas crises de endocruzamento há muito tempo e teriam sobrevivido, saindo, no processo, "purificados" das mutações deletérias. Como comentado anteriormente, um dos problemas do endocruzamento é a ocorrência de homozigose de alelos deletérios recessivos, existentes normalmente nas populações naturais. Portanto, poderia ser argumentado que inúmeras espécies que no século XIX sofreram reduções populacionais já se extinguiram por colapso de endocruzamento, e que os elefantes-marinhos-do-norte, os castores dinamarqueses e os tordos-pretos são alguns dos poucos exemplos de espécies que escaparam desse cataclismo genético por terem populações iniciais muito grandes, poucos alelos levemente deletérios (o que diminuiria a chance de fixação desses alelos sub-ótimos por deriva genética) ou apenas por acaso. Como as taxas de mutação para caracteres morfológicos de herança quantitativa são de até três ordens de grandeza maiores do que aquelas observadas para a maioria dos caracteres moleculares (com exceção dos microssatélites; Slatkin, 1995), é possível que elas já tenham inclusive recuperado parte de seu potencial evolutivo nesses caracteres, apesar da baixa variabilidade molecular. Outras espécies, como o guepardo e os lobos-cinzas. estariam ainda lutando em seu purgatório evolutivo, seu futuro dependendo, em última análise, da proporção de genes deletérios fixados ou dos acidentes demográficos que suas populações possam sofrer. Claramente, o que parece ser o problema nas populações ameaçadas não é tanto a baixa variabilidade genética, mas sim o alto endocruzamento (entretanto, veja abaixo discussão sobre alguns locos onde a variabilidade é importante, como aqueles do sistema imunológico). Baixa variabilidade genética, em última análise, representa um problema em termos de potencial de resposta a desafios ambientais, pois a variabilidade é, como Darwin bem demonstrou, a condição necessária para a seleção natural. O alto endocruzamento representa, contudo, um problema bem maior, pela questão da expressão da carga genética da população, causada pela homozigose de genes deletérios.

Existem caracteres moleculares com grande importância adaptativa, entretanto, para os quais o grau de polimorfismo é fundamental. Para esses, a "purificação do endocruzamento" é, na verdade, muito prejudicial, por reduzir a variabilidade gênica e comprometer, assim, as chances de sobrevivência da espécie a curto e médio prazo. Um sistema gênico típico desse caso é o de proteínas do sistema MHC. Essas proteínas são responsáveis pelo reconhecimento de peptídeos virais ou de outros patógenos e pela transmissão dessa mensagem aos linfócitos T no processo de resposta imune às infecções (Klein *et al.*, 1993). Existe uma co-evolução constante entre os organismos patogênicos e o sistema de reconhecimento. Nesse processo, cada mutação que altera os peptídeos virais, diminuindo sua chance de reconhecimento pelo sistema imunológico, é favorecida e, conseqüentemente, aumenta em freqüência na população viral. Uma vez que tal peptídio mutante passe a ser comum nesses vírus, cada mutação no sistema de reconhecimento que permita a detecção do novo peptídeo também é favorecida (Hughes e Nei, 1992a, b). Esse processo de ação e reação constante também é conhecido como a hipótese evolutiva da "Rainha Vermelha" (*Red Queen*, em inglês), por referência à frase em "Alice atrás do espelho", de Lewis Carroll, "Você tem de estar sempre correndo para poder continuar no mesmo lugar" (Van Vallen, 1973; para uma crítica à hipótese da Rainha Vermelha, veja Rios, 2008). Organismos com um número maior de locos heterozigotos no sistema MHC são favorecidos nessa co-evolução (seleção dependente de frequência) e, consequentemente, o grau de polimorfismo desses sistemas em populações naturais está entre os mais elevados para genes não neutros (Hughes, 1988, Hughes e Nei, 1990).

Um exemplo de espécie com baixa variabilidade no MHC é a vaquita (*Phocoena sinus*), um pequeno golfinho com distribuição restrita à parte norte do Golfo da Califórnia, cuja única população vivente está estimada em 567 indivíduos. As vaquitas não apresentam variabilidade alguma na região controle mitocondrial (Rosel *et al.*, 1999) e baixíssima variabilidade no MHC (apenas um alelo no locus DQB e dois no DRB; Munguia-Vega *et al.*, 2008). Simulações demográficas mostraram que a homogeneidade genética da vaquita provavelmente se deve a um evento fundador na origem da espécie, que teria mantido um baixo tamanho populacional desde então (Taylor e Rojas-Bracho, 1999). Assim, a vaquita teria expurgado seus alelos de efeito deletério médio e elevado. Por outro lado, alelos levemente deletérios teriam escapado da seleção e aumentado em frequência por causa da forte deriva. Isso explicaria a alta incidência de malformações esqueléticas, como a polidactilia e fusões vertebrais, que aparentemente não afetam a sobrevivência ou a fertilidade desses animais. Embora nenhuma doença infecciosa tenha sido até hoje observada na vaquita e sua carga parasitária seja normal, a falta de variabilidade no MHC pode significar alta susceptibilidade da espécie a novos agentes infecciosos (Munguia-Vega *et al.*, 2008).

Outras espécies que apresentam graus anormalmente baixos de polimorfismo no sistema MHC devido a estrangulamentos populacionais são: os castores-escandinavos (Ellegren *et al.*, 1993), nos quais o MHC é monomórfico; as populações remanescentes do bisão-europeu (*Bison bonasus*; Udina e Shaikhaev, 1998); uma espécie ameaçada de peixe de água-doce dos EUA (*Poeciliopsis o. occidentalis*; Hedrick e Parker, 1998); os leões-asiáticos (Packer *et al.*, 1991); e o guepardo (Yuhki e O'Brien, 1990). Com essa baixa variabilidade no sistema MHC, essas espécies estão potencialmente mais vulneráveis a doenças virais. De fato, uma epidemia do vírus da peritonite felina, que em gatos domésticos (que têm variabilidade normal no sistema MHC) costuma matar no máximo 5% dos gatos afetados, praticamente dizimou uma população experimental de guepardos nos EUA nos anos 80, matando mais de 60% de toda a população e severamente debilitando os sobreviventes (O'Brien *et al.*, 1985).

As espécies com baixa variabilidade do sistema MHC devido ao endocruzamento podem estar, portanto, vivendo uma situação de bomba-relógio, na qual o aparecimento de uma cepa viral que consiga driblar o sistema imunológico de um indivíduo rapidamente se espalhe por toda a população. Elas teriam vencido a batalha contra os próprios genes deletérios durante a crise do endocruzamento, se tornando, no processo, mais puras e homogêneas, para serem, em seguida, derrotadas por um agente externo justamente por sua homogeneidade e consequente falta de potencial evolutivo. Assim, a perda de heterozigosidade tem um efeito inicial forte, pela depressão de endocruzamento e expressão de alelos deletérios raros, que leva a um decréscimo populacional importante, que depois pode ter dois destinos distintos: a entrada em um processo de *genetic meltdown* (colapso genético), com a consequente extinção da espécie; ou a diminuição

em frequência, por seleção natural, dos alelos deletérios, de forma que a espécie fica "purificada" e razoavelmente imune a efeitos posteriores do endocruzamento. Isso é o que ocorre, em parte, com animais domésticos, onde o cruzamento entre irmãos se tornou, ao longo do tempo, biologicamente viável pela baixa frequência de alelos deletérios recessivos em suas populações atuais. Por outro lado, mesmo passada a crise inicial da depressão de endocruzamento, as espécies que perdem variabilidade gênica continuam evolutivamente frágeis por causa da consequente baixa imunidade a doenças, devido à homogeneidade genética nos sistemas de defesa e reconhecimento imunológico. Apesar da relação demonstrada em muitas espécies entre variabilidade no sistema MHC e resistência a doenças, nem sempre essa relação é observada. Por exemplo, no elefante-marinho-do-norte existem doenças, como em outros mamíferos aquáticos, sem que elas atinjam proporções epidêmicas (Colegrove *et al.*, 2005). Da mesma forma, populações de bois selvagens no norte da Inglaterra (*Bos taurus*) apresentam baixíssima variabilidade no sistema MHC, mas possuem resistência a doenças parasitárias (Visscher *et al*., 2001). Portanto, apesar da relação entre variabilidade no sistema MHC e resistência a doenças, encontrada em várias espécies ameaçadas, essa relação é complexa e seguramente outros sistemas gênicos estão envolvidos, o que demanda cuidado na interpretação dos resultados dessa variação (Acevedo-Whitehouse e Cunningham, 2006).

De qualquer forma, nosso papel não deve ser de observadores passivos dessa roleta russa do endocruzamento, deixando que o acaso decida sobre a sobrevivência das espécies. Como vimos, em geral, a variabilidade genética é importante para a persistência evolutiva das espécies, mesmo em curto prazo, e programas de tentativa de recuperação de populações reduzidas devem se preocupar com a manutenção da pouca variabilidade restante. A crise do endocruzamento pode ser evitada se os tamanhos populacionais forem mantidos acima de um nível crítico, com o monitoramento regular dos níveis de heterozigosidade e com cruzamentos que maximizem a variabilidade gênica e minimizem o endocruzamento (Borlase *et al.*, 1993). Em casos extremos, em que a espécie está claramente debilitada e em vias de extinção, pode ser até considerada a hibridação com subespécies próximas para tentar diminuir a depressão de endocruzamento. Esse procedimento normalmente não é recomendado, pois a hibridação é frequentemente vista como uma outra maneira, mais sutil, de extinguir uma espécie, por meio da sua "diluição" em genomas de outra linhagem evolutiva (Rhymer e Simberloff, 1996) ou mesmo pela entrada de agentes patogênicos novos na população (Nishi *et al.*, 2002), trazidos pelos indivíduos que serviriam para rejuvenecê-la. Entretanto, se o destino mais provável de alguns poucos indivíduos restantes de uma espécie é a extinção devido a problemas reprodutivos e outros decorrentes do endocruzamento, essa medida pode ser considerada. No caso dos leões-asiáticos, por exemplo, observou-se que os cruzamentos em cativeiro eram raramente bem-sucedidos (muitos machos tinham dificuldades de cruzar e, daqueles que cruzavam, apenas 15% das gestações chegavam a termo nos zoológicos; O'Brien, 1994). No entanto, em um parque nos EUA, observou-se, inesperadamente, um grande sucesso reprodutivo nesses leões em uma colônia começada com apenas cinco indivíduos. Um estudo posterior feito com aloenzimas mostrou que, desses cinco indivíduos, dois eram, na verdade, leões-africanos, que pertencem a uma outra subespécie. O sucesso da reprodução em cativeiro tinha sido, portanto, provavelmente devido ao vigor híbrido (Wildt *et al.*, 1987). Introgressões contínuas e controladas desses leões revigorados com leões asiáticos puros podem servir para injetar alguma variabilidade gênica nessa subespécie, sem que sejam perdidos em demasia seus genes originais. Em populações altamente endocruzadas de lobos-cinza na Escandinávia e com baixa capacidade de recuperação, a chegada de apenas um lobo migrante foi suficiente para fazer com que a população voltasse a crescer (Vila *et al.*, 2003). A hibridação pode ser a última esperança para uma das espécies de tartaruga-gigante-de-Galápagos (*Geochelone abingdoni*), cujo único representante vivo é Lonesome George (George Solitário). Após três décadas de infrutíferas tentativas de reprodução de Lonesome George com fêmeas de outras espécies do gênero *Geochelone*, marcadores moleculares revelaram, em uma ilha vizinha, um indivíduo de origem híbrida, cujo pai era *G. abingdoni* e a mãe, híbrida entre *G. hoodensis* e *G. becki* (Russello *et al.*, 2007). Infelizmente, o indivíduo em questão era macho. Entretanto, como somente oito tartarugas foram analisadas nessa ilha, é possível que ainda existam também fêmeas híbridas descendentes de *G. abingdoni*, que seriam excelentes candidatas a companheiras de Lonesome George e poderiam salvar a espécie da extinção.

A estratégia de hibridização pode falhar, no entanto, se houver baixo sucesso reprodutivo dos híbridos. Na Ilha Macquarie, no Pacífico Sul, os lobos-marinhos foram extintos no século XIX devido à caça. Nos anos 1940, três espécies (*Arctocephalus gazella*, *A. forsteri* e *A. tropicalis*) recolonizaram a ilha e atualmente um alto nível de hibridização entre elas é observado (17-30%). Marcadores moleculares mostraram que os machos híbridos, apesar de terem a mesma aptidão para conquistar e manter territórios em relação aos machos puros, tiveram um menor número de filhotes (Lancaster *et al.*, 2007). Além disso, alguns autores alertam contra o uso disseminado de tentativas de recuperação de variabilidade genética pelo intercruzamento de indivíduos que provenham de áreas geográficas distintas (justificado mais adiante).

Os estudos apresentados acima são, em sua maioria, com espécies ditas "carismáticas", que atraem a atenção do público e dos órgãos financiadores. No entanto, processos semelhantes devem estar acontecendo com todas as espécies ameaçadas de extinção, de modo que devemos procurar sempre a preservação de seus tamanhos populacionais acima de um nível crítico, para evitar tanto os acidentes demográficos, como a depressão por endocruzamento.

22.3.2. Indivíduos e estruturas familiares

Marcadores moleculares com altas taxas de mutação, como SNPs, mini e microssatélites e, de maneira mais limitada, aloenzimas, RAPD ou AFLP, podem ser usados com vantagem no estudo das estruturas clonais e das composições familiares nas espécies (Hughes, 1998; Avise, 2004). Usando marcadores de RAPD, por exemplo, se pôde estabelecer que a última população remanescente de uma espécie de eucalipto australiana (*Eucalyptus phylacis*) provavelmente era constituída de um único clone, ou seja, geneticamente apenas um indivíduo desse eucalipto sobreviveu à expansão da agricultura naquele país (Rossetto *et al.*, 1999).

Além disso, marcadores moleculares são úteis para revelar o sistema de acasalamento das espécies, uma informação fundamental para a definição das estratégias de manejo a serem adotadas para as que estão em risco. Com o uso de marcadores moleculares, foi possível, por exemplo, verificar que a suposta monogamia de muitas espécies de aves estava errada, ao se verificar que, para marcadores nucleares polimórficos, muitas ninhadas apresentavam mais de quatro alelos, o que seria impossível se todos fossem filhos de um único casal (Haig *et al.*, 1996; Petrie, 1998; Alderson *et al.*, 1999; revisão em Griffith *et al.*, 2002).

Outro exemplo é o dragão-de-Komodo (*Varanus komodoensis*), uma espécie de lagarto ameaçada pela fragmentação de hábitat. Uma estratégia para evitar a extinção da espécie tem sido a manutenção de indivíduos em cativeiro, para posterior recolonização de áreas recuperadas. Entretanto, muitos zoológicos mantêm apenas fêmeas e os machos são transportados entre as instituições. Essa estratégia foi recentemente confrontada por um estudo que demonstrou que os dragões-de-Komodo realizam partenogênese, alternadamente com a reprodução cruzada. O "DNA fingerprinting" revelou que embriões de duas fêmeas mantidas em cativeiro, isoladas de machos, eram todos homozigotos para um ou outro dos alelos das mães. Esse resultado sugere que, a fim de manter a diversidade genética da espécie, os zoológicos devem evitar manter fêmeas sem a presença de machos (Watts *et al.*, 2006).

Outra aplicação importante da genética para a conservação é na determinação não invasiva do sexo em espécies sem dimorfismo sexual, como ocorre em muitas aves, onde cerca de 50% das espécies não apresentam dimorfismo sexual visível (Griffiths *et al.*, 1998). Frequentemente, pessoas lidando com conservação e criação em cativeiro de espécies de aves ameaçadas de extinção se deparam com o problema de não saber quais aves devem ser pareadas para formar casais reprodutores. O perna-longa *Himantopus novaezelandiae*, por exemplo, é uma das aves mais ameaçadas de extinção do mundo, com apenas 24 adultos existentes. Muitas tentativas de reprodução dessas aves em cativeiro foram mal sucedidas por envolverem pares do mesmo sexo, já que a formação dos pares nas condições de criação era aleatória. Com o uso de marcadores moleculares ligados aos cromossomos sexuais (nas aves o sexo heterogamético é o feminino, que apresenta em seu cariótipo os cromossomos sexuais W e Z, enquanto que os machos são ZZ), esses pássaros puderam ser agrupados em casais, o que aumentou o sucesso na reprodução da espécie em programas de reintrodução na natureza (Millar, 1997). Marcadores de sexo em aves, portanto, são muito importantes e várias estratégias têm sido seguidas, como a busca empírica de bandas sexo-específicas em padrões de RAPD (Lessells e Mateman, 1998) ou de minissatélites (Miyaki *et al.*, 1997; Miyaki *et al.*, 1998; Wink *et al.*, 1998). Esses últimos trabalhos são particularmente interessantes e tem envolvido cientistas brasileiros, que usam marcadores moleculares no estudo de populações de várias espécies de papagaios (Psittacidae).

Uma abordagem que parece ser muito promissora, semelhante à usada por Miyaki *et al.* (1998), é o uso de genes de uma enzima ligada à replicação do DNA—os genes da helicase de ligação cromossômica (CHD – "chromo-helicase-DNA binding", em inglês), que são evolutivamente muito conservados e que estão localizados nos cromossomos sexuais. O gene CHD-W existe apenas nas fêmeas, enquanto que o CHD-Z existe nos dois sexos. Os dois genes têm introns de tamanhos diferentes, mas têm homologia e similaridade em regiões conservadas, para as quais podem ser desenhados iniciadores para PCR. Assim, uma única reação de PCR produz para cada gene um fragmento de DNA de tamanho diferente. Esses produtos de PCR podem ser facilmente visualizados por eletroforese, de forma que o macho apresenta apenas uma banda e a fêmea, duas bandas. Esses genes são evolutivamente muito conservados e aparentemente ocorrem em todas as aves exceto nos avestruzes e em outros Struthioniformes, onde estão em cromossomos autossômicos (Griffiths *et al.*, 1998). Dessa forma, eles poderão ser usados facilmente como marcadores moleculares na determinação do sexo da maioria das aves onde não exista dimorfismo sexual externo, facilitando enormemente todos os estudos com reprodução em cativeiro e reintrodução na natureza de espécies ameaçadas.

Marcadores moleculares podem ser usados também para a sexagem de mamíferos, quando a sexagem morfológica é dificultada ou porque o material disponível não inclui genitálias em bom estado (como em amostras degradadas, que incluem apenas alguns tecidos ou provenientes da análise direta de pele, fezes ou de pelos). Nesse caso, a abordagem é semelhante à feita com aves, usando-se, entretanto, marcadores do cromossomo Y, como o fator de diferenciação testicular SRY, ou regiões pseudohomólogas com sequências diferentes no X e no Y, como os genes das proteínas "zinc finger" (ZFX e ZFY). Quando o SRY é usado, machos produzem uma banda (referente ao fragmento presente no Y) e fêmeas não apresentam nenhuma banda (ausência do Y). Por isso, é essencial que controles positivos sejam co-amplificados, a fim de distinguir entre falhas na amplificação (que levariam à designação equivocada do sexo) e ausência do marcador sexo-específico na amostra. O sistema ZFX/ZFY é um pouco diferente, porque se baseia na amplificação simultânea de fragmentos presentes nos cromossomos X e Y: nos machos, são produzidas duas bandas, enquanto, nas fêmeas, uma banda é amplificada (devido à presença do cromossomo X apenas). Assim, o próprio ZFX funciona como controle positivo. Os sistemas ZFX/ZFY e SRY têm sido empregados, por exemplo, no boto-cinza e no tucuxi (*Sotalia guianensis* e *S. fluviatilis*) (Cunha e Solé-Cava, 2007; Figura 22.1). Como a maioria dos cetáceos, o boto-cinza e o tucuxi não apresentam dimorfismo sexual externo facilmente observável, o que atrapalha estudos de estrutura social baseados na foto-identificação, uma técnica que possibilita o acompanhamento temporal e espacial dos indivíduos usando marcas naturais. Determinar visualmente o sexo dos indivíduos foto-identificados é dificílimo, porque depende da observação de sua região ventral, um evento raro no campo. A aplicação dos sistemas ZFX/ZFY e SRY nessas espécies tem permitido a determinação molecular do sexo de indivíduos foto-identificados, amostrados por biópsia remota. A técnica também é capaz de sexar carcaças cuja putrefação impede a sexagem convencional, feita a partir do exame morfológico. Como carcaças são a principal fonte de amostras para os pesquisadores que trabalham com essas espécies, a possibilidade

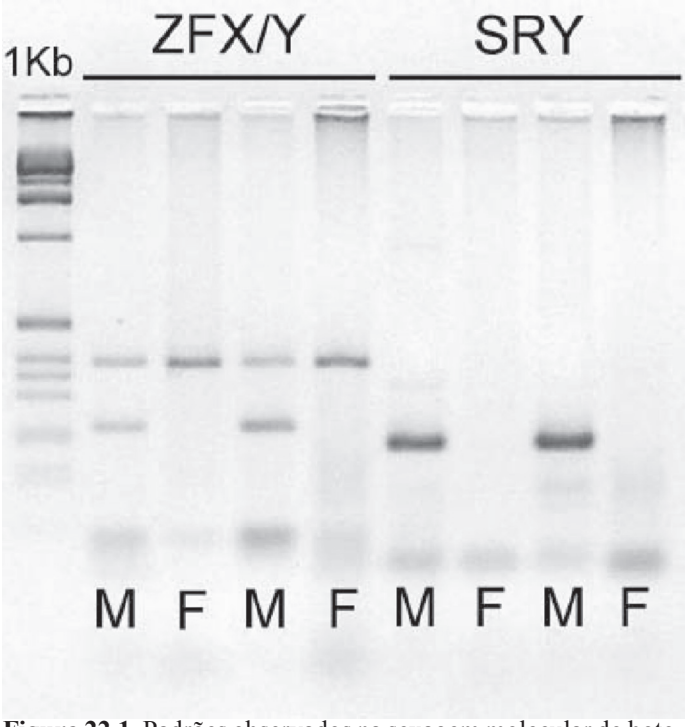


Figura 22.1. Padrões observados na sexagem molecular do boto-cinza (*S. guianensis*) usando os sistemas ZFX/ZFY e SRY. M: machos; F: fêmeas; 1Kb: marcador de tamanho. Fonte: Cunha e Solé-Cava (2007).

de sexá-las é valiosa, por aumentar o conteúdo informativo para as análises, assim como o próprio tamanho amostral.

Embora as técnicas moleculares de sexagem venham sendo cada vez mais aplicadas em espécies ameaçadas, muitos estudos não realizam uma avaliação da performance dos sistemas na identificação correta do sexo dos indivíduos (Robertson e Gemmell, 2006). Por essa razão, é aconselhável que os estudos que usam a sexagem molecular incluam uma etapa de teste com um número razoável de indivíduos de sexo conhecido, como foi feito no caso dos botos-cinza e tucuxis.

As estruturas familiares também têm importância quando consideramos espécies ameaçadas. Em populações naturais, grupos geneticamente mais relacionados ("famílias") frequentemente se distribuem espacialmente de maneira agregada. As consequências desse comportamento são várias, inclusive o aumento de endocruzamento e, algumas vezes, a necessidade da manutenção de tais estruturas familiares para a reprodução da espécie (Avise, 2004). Marcadores moleculares permitiram verificar que várias espécies de baleias (Amos *et al.*, 1993), tartarugas (Bowen *et al.*, 1996; Bowen e Karl, 2007) e inúmeras aves (Avise, 1996; Haig *et al.*, 1996) e peixes (Barluenga e Meyer, 2004) têm comportamentos migratórios com fidelidade à região natal (filopatria) e frequentemente se mantêm agregadas em estruturas familiares muito fechadas, com pouco intercâmbio entre grupos (Avise, 2004). Essa característica tem influência direta para a conservação, porque afeta os padrões de migração das espécies e leva a situações de não-panmixia (veja próxima seção).

22.3.3. Estruturação genética e estratégias de conservação

O estudo das estruturas populacionais por meio de técnicas moleculares é talvez a parte mais importante da genética para conservação e tem sido útil tanto no estudo de populações exploradas comercialmente (ou seja, abundantes, mas com riscos populacionais devido à superexploração), como nas espécies já ameaçadas de extinção. Naturalmente, a própria definição do que seja uma população é assunto complexo (Waples e Gaggiotti, 2006), mas para as pessoas envolvidas em conservação os conceitos mais usados são os de "unidades evolutivamente significativas" ("Evolutionarily Significant Unit", ESU; Ryder, 1986) e de "unidades de manejo" ("Management Unit", MU; Moritz, 1994). As ESU geralmente são identificadas com base no monofiletismo recíproco das linhagens, enquanto que, para a delimitação de MU, podem ser usadas diferenças quantitativas (por exemplo, nas frequências alélicas) (Moritz, 1994). Assim, as MU diferenciam-se das ESU por serem menos restritivas e mais próximas do presente demográfico da espécie (Moritz, 1994; Palsboll *et al.*, 2007).

Peixes e outros organismos marinhos são os únicos animais selvagens consumidos por nossa espécie em grande escala pela exploração direta de populações naturais. Assim, os problemas (e as soluções) encontrados em sua exploração são completamente diferentes daqueles oriundos da exploração das plantas e animais domésticos. O risco de extinção, por exemplo, não é um problema para a exploração de animais terrestres como vacas e ovelhas, mas se torna um problema constante na exploração de populações marinhas. As baleias, por exemplo, que eram abundantes até o século XIX, constituindo recurso alimentar e industrial notável, quase se extinguiram em meados do século XX (Baker *et al.*, 1996), e a pesca ilegal continuada de algumas espécies poderá ainda levá-las à extinção (Baker *et al.*, 2000b). De fato, a caça descontrolada foi o fim de outros mamíferos aquáticos, como a vaca-marinha-de-Steller (*Hydrodamalis gigas*) e a foca-monge-do-Caribe (*Monachus tropicalis*). A vaca-marinha era uma espécie de sirênio gigante (que alcançava 8m) que foi extinta pela caça apenas 27 anos após sua descoberta, que aconteceu em 1741, antes mesmo de sua descrição (publicada em 1780). No primeiro relato sobre o animal, seu descobridor citava o delicioso sabor de seu óleo ("que lembra o de amêndoas doces"; Steller, 1751). A foca-monge-do-Caribe também foi superexplorada por causa de sua gordura e o último indivíduo da espécie foi observado em 1952.

O colapso da pesca de várias espécies de vertebrados e invertebrados é um exemplo de como populações, que pareciam ser (e, frequentemente, eram) enormes, se mostram frágeis quando exploradas intensivamente (Grant e Bowen, 1998). A genética das espécies marinhas exploradas comercialmente, também conhecida como "genética pesqueira", é um campo de pesquisas enorme e interessante. O leitor poderá encontrar mais informações sobre esse assunto em várias revisões (Ryman e Utter, 1987; Carvalho e Hauser, 1995; Thorpe *et al.*, 2000; Ward, 2000).

Antes de falar especificamente sobre a questão das implicações da estruturação genética sobre as abordagens de conservação em populações ameaçadas, é necessário que definir brevemente estruturação gênica e mostrar como podemos estimar os níveis de fluxo gênico (dispersão e migração) entre populações naturais.

Populações naturais frequentemente não mantêm panmixia, ou seja, a probabilidade de reprodução entre dois indivíduos quaisquer da população não é sempre a mesma, e depende de fatores biológicos e geográficos. Os motivos para essa limitação em relação à panmixia são muitos e, frequentemente, complexos (revisões desses conceitos e processos podem ser encontradas em Wright, 1978; Lewontin, 1985). Alguns fatores que limitam a panmixia são:

- Pouco deslocamento dos adultos ou das larvas/sementes (para uma panmixia real, o deslocamento potencial de uma espécie deveria ser igual a sua faixa de distribuição);
- Em organismos fixos, como plantas e vários organismos marinhos, pouca viabilidade gamética (o que leva a uma probabilidade maior de cruzamentos entre vizinhos);
- Seleção de hábitat com fidelidade natal (filopatria);
- Cruzamentos com escolha de parceiro ("assortative mating", em inglês);
- Recrutamento (entrada de novos indivíduos na população) caótico (alta dispersão, mas dispersantes endocruzados migrando em conjunto e de maneira caótica, com ciclos dispersivos muito longos e de seleção pós-fixação).

Essa ausência de panmixia provoca uma estruturação (subdivisão) das populações em subpopulações. Os motivos são: (1) novidades evolutivas (mutações) surgidas em algumas partes vão levar muito tempo para se espalharem para toda a população; (2) endocruzamento local diminuirá a variabilidade genética de cada subpopulação; e (3) deriva genética e seleção poderão levar a uma divergência entre indivíduos agrupados em diferentes partes da distribuição. Dessa forma, populações estruturadas vão apresentar um equilíbrio dinâmico entre fatores de diferenciação (mutação, deriva genética e seleção natural direcional ou disruptiva diferente em cada área) e fatores de homogeneização (migração, seleção natural purificadora e seleção natural balanceada ou direcional igual em cada área). A importância relativa de cada um dos fatores vai variar de acordo com o tamanho populacional e as diferenças ambientais ao longo da faixa de distribuição da espécie.

De acordo com a bionomia da espécie—em particular o tipo de dispersão—,podemos observar três tipos principais de estruturação de populações (Wright, 1978):

- Isolamento por distância, em que o fator principal para a diferenciação é a limitação da dispersão em função da dis-

tância geográfica, dentro de uma população sem subdivisões aparentes. Nesse modelo, não existe panmixia, mas também não existe uma diferenciação abrupta ou descontinuidade que permita a delimitação de subpopulações;

- Modelo de ilhas, em que a diferenciação entre as populações não depende da distância entre elas, os recrutas sendo oriundos de uma única população de tamanho infinito (modelo infinito; Wright, 1978) ou de outras ilhas, sem relação com sua posição espacial (modelo finito; Slatkin, 1985);
- Modelo passo-a-passo ("stepping stones"; Wright, 1978), em que cada subpopulação somente pode trocar migrantes com as populações vizinhas.

Uma maneira gráfica de visualizar os diferentes modos de estruturação gênica em populações naturais é considerar que, em uma dada região geográfica, foram amostradas três localidades ao longo de uma linha (por exemplo, três praias diferentes ao longo de uma costa). Se denominarmos os gametas produzidos pelos indivíduos dos três pontos contíguos, respectivamente, de A, B e C (Figura 22.2), teremos panmixia quando a probabilidade de cruzamentos entre gametas de uma mesma localidade ($P_{A\text{-}A}$, $P_{B\text{-}B}$ e $P_{C\text{-}C}$) for igual à probabilidade dos cruzamentos entre gametas de localidades diferentes ($P_{A\text{-}B}$, $P_{B\text{-}C}$ e $P_{A\text{-}C}$). Por outro lado, teremos uma população estruturada quando as probabilidades de cruzamentos dentro de cada localidade ($P_{A\text{-}A}$, $P_{B\text{-}B}$ e $P_{C\text{-}C}$) forem maiores do que as probabilidades de cruzamentos entre gametas de localidade diferentes ($P_{A\text{-}B}$, etc).

Usando essa notação, podemos dizer que, tomando como referência a população A, na Figura 22.2:

(a) Panmixia: $P_{A\text{-}A} = P_{B\text{-}B} = P_{C\text{-}C} = P_{A\text{-}B} = P_{A\text{-}C}$
(b) Isolamento por distância: $P_{A\text{-}A} = P_{B\text{-}B} = P_{C\text{-}C} > P_{A\text{-}B} > P_{A\text{-}C}$
(c) Ilhas: $P_{A\text{-}A} = P_{B\text{-}B} = P_{C\text{-}C} \gg P_{A\text{-}B} = P_{A\text{-}C}$
(d) Passo-a-passo: $P_{A\text{-}A} = P_{B\text{-}B} = P_{C\text{-}C} \gg P_{A\text{-}B} \gg P_{A\text{-}C} \cong 0$

De uma maneira geral, é esperado que espécies com alta capacidade de dispersão, mas exibindo algum tipo de filopatria, como várias espécies de aves e peixes, se apresentem estruturadas conforme o modelo de ilhas. Por outro lado, espécies em que os dispersores têm vida curta ou por algum outro motivo não podem cobrir toda a distância da distribuição da espécie, se apresentarão estruturadas conforme o modelo de isolamento por distância ou passo-a-passo (Avise, 2004). Naturalmente, esses modelos, como todos os modelos, são simplificações da realidade e, portanto, podemos observar na natureza casos de transição entre os modelos (Neigel, 1997).

A verificação da existência de estruturação gênica nas populações naturais e a estimativa do fluxo gênico entre essas subpopulações pode ser feita a partir da variância das frequências gênicas entre localidades diferentes (F_{ST}, Wright, 1978), como visto no Capítulo 21 (estruturação populacional). O uso do F_{ST} como medida de estruturação populacional pode ser problemático em espécies nas quais o equilíbrio entre deriva e migração (o chamado equilíbrio de Wright-Fisher) ainda não foi atingido (Whitlock e McCauley, 1999). Entretanto, apesar de não ser recomendado para a estimativa direta do fluxo gênico entre populações, o F_{ST} ainda pode ser útil mesmo em populações ameaçadas (e, portanto, fora do equilíbrio) como uma medida razoável de sua estruturação histórica (Neigel, 2002).

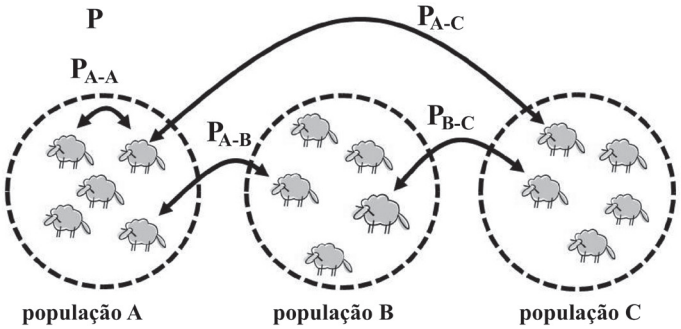


Figura 22. 2. Probabilidades de cruzamentos entre indivíduos em três populações e dentre essas populações.

Um outro tipo de análise da estruturação de populações é a Análise de Variância Molecular (AMOVA; Excoffier, 1994). Nessa abordagem, de forma semelhante à usada na análise estatística de covariância (Sokal e Rohlf, 1995), a variância nas frequências gênicas é subdividida em vários componentes hierarquicamente inclusivos, como a variância entre grupos de subpopulações (Φ_{CT}) e a variância entre subpopulações (Φ_{SC}). Uma das vantagens desse tipo de abordagem é que ele pode ser usado com um número grande de tipos de marcadores moleculares (ao contrário do F_{ST} e do G_{ST}, que se aplicam mais a dados de marcadores codominantes). A interpretação dos valores de F_{ST}, G_{ST} ou Φ_{CT} é semelhante: valores não significativos nos dizem que a hipótese nula de panmixia não pode ser rejeitada e indicam a necessidade de outros estudos para verificação da homogeneidade. A AMOVA também pode ser complementada com o uso das informações da distribuição espacial dos indivíduos, considerando todos os agrupamentos *ad hoc* possíveis, em uma análise chamada de AMOVA espacial (SAMOVA; Dupanloup *et al.*, 2002). Uma das limitações em análises *ad hoc* como F_{ST} e AMOVA é a necessidade de se definirem os grupos *a priori*. Na prática, os grupos em geral são definidos como grupos de amostragem, ou seja, os organismos coletados em um determinado local. Isso significa que locais de mistura populacional serão tratados como se fossem homogêneos e muitas localidades amostradas nas quais ocorre a mesma população poderão viciar a análise da estruturação global. Uma maneira de contornar esse problema é tratar cada individuo separadamente, em procedimentos chamados de Análises de Atribuição. Esses tipos de análise, geralmente feitos com abordagens Bayesianas empregando algoritmos de MCMC (Markov Chain Monte Carlo), são explicados no Capítulo 21.

As implicações das análises populacionais para estudos de conservação são muito importantes: se uma espécie ameaçada que ocupa uma determinada área se apresenta estruturada, então a estratégia de conservação deve procurar preservar a diversidade da espécie em toda a sua distribuição, pois já podem existir adaptações locais que se perderiam no caso de as populações serem misturadas ou serem extintas. Por outro lado, se a espécie é homogênea ao longo de toda a área de ocorrência, então é viável concentrar sua proteção em apenas uma área, usando indivíduos dessa área para recolonização das outras quando necessário (Haig, 1998). Em outras palavras, para estudos de recolonização ou conservação de espécies ameaçadas, deve ser levado em conta não a espécie, mas as unidades de manejo, que são as subpopulações diferenciadas geneticamente (Berg *et al.*, 1996).

Por exemplo, um estudo com sequenciamento da região controle mitocondrial do veado-dos-pampas (*Ozotoceros bezoarticus*) revelou que suas populações estavam estruturadas (F_{ST}=0,20), indicando também uma relação entre diferenciação populacional e distância geográfica ao longo de um transecto entre as latitudes 5 e 41º sul, desde o cerrado brasileiro até a Argentina (Gonzalez *et al.*, 1998b). A consequência para a conservação dessa espécie é que a preservação dos hábitats deve ocorrer ao longo de toda sua distribuição. Se políticas de reintrodução forem implementadas, elas devem ser feitas a partir de estoques reprodutores específicos para cada região (estoques autóctones).

As populações da tartaruga-de-pente (*Eretmochelys imbricata*), apesar de sua grande capacidade de dispersão,

também estão muito estruturadas, desta vez seguindo um modelo de ilhas (Bass *et al.*, 1996). Graças ao conhecimento das diferenças genéticas entre as várias populações dessa tartaruga, foi possível verificar que os indivíduos existentes em uma região de alimentação, em Porto Rico, vinham de toda a região caribenha. Esse resultado indicou que projetos de conservação das regiões de desova dessas tartarugas, apesar de importantes, seriam inúteis se não fossem acompanhados de sua preservação também nas áreas de alimentação, que podiam estar muito distantes (Bowen *et al.*, 1996). Uma consequência interessante desse resultado é que, para espécies com alto grau de dispersão, as fronteiras políticas são irrelevantes e a conservação do estoque nativo de uma espécie precisa ser feita de maneira coordenada entre vários países (Bowen *et al.*, 1996).

Outro exemplo de espécie que requer uma ação coordenada entre diferentes países é a toninha (*Pontoporia blainvillei*), única espécie de golfinho considerada ameaçada de extinção no Brasil. Em um estudo usando sequenciamento da região controle mitocondrial de amostras brasileiras da espécie, foi observada uma alta diferenciação (Φ_{CT} = 0,401; P<0,0001) entre o Rio de Janeiro e o Rio Grande do Sul (Secchi *et al.*, 1998). Com base nesses dados e em parâmetros biológicos e populacionais, foram propostas quatro unidades de manejo para a espécie, se estendendo entre Brasil, Uruguai e Argentina (Secchi *et al.*, 2003). Posteriormente, a estruturação genética das toninhas no Uruguai e na Argentina foi estudada e, de maneira geral, apoiou a proposta feita por Secchi *et al.* (2003) (Lazaro *et al.*, 2004; Mendez *et al.*, 2008). A delimitação das unidades de manejo é fundamental para a avaliação do impacto das capturas acidentais em redes de pesca, a maior causa de mortalidade não natural nessa espécie. Como a taxa de mortalidade por captura excede a taxa de natalidade das populações, a conservação da espécie depende da redução das capturas acidentais.

Na Tabela 22.2, apresentamos uma compilação de resultados de vários estudos moleculares feitos com populações de espécies ameaçadas. De uma maneira geral, observamos que populações de espécies em declínio se encontram frequentemente estruturadas, o que é esperado, pois em geral a degradação ambiental leva à formação de refúgios ou "oásis", onde pequenas populações das espécies persistem, sem contudo poder trocar genes através das áreas impactadas, que são o "deserto" em volta delas.

A idéia geral, portanto, é que, ao se manejarem populações de espécies ameaçadas, devemos procurar preservar também sua diversidade geográfica, a não ser nos casos de pouca diferenciação genética, ou seja, nos casos em que as populações estão pouco estruturadas. No entanto, se a variabilidade local já tiver sido perdida devido ao endocruzamento, mas ainda houver uma alta diversidade entre populações isoladas, pode-se considerar, em casos extremos, o cruzamento de indivíduos de populações diferentes para diminuir a depressão por endocruzamento. Por exemplo, a foca-monge-havaiana (*Monachus schauinslandi*), apesar de décadas de proteção e esforços de aumento das cinco populações conhecidas, se manteve até hoje com um número muito reduzido de indivíduos. Um estudo com sequenciamento da região controle do mtDNA e com minissatélites encontrou um baixíssimo nível de variabilidade dentro de cada população, mas uma alta diferenciação entre elas (F_{ST}=0,17; Kretzmann *et al.*, 1997). A baixa eficiência reprodutiva dessas focas foi interpretada como resultado do endocruzamento e a nova política de proteção e tentativas de proliferação de suas populações serão feitas a partir do cruzamento de indivíduos de colônias diferentes, para tentar reverter o quadro de depressão por endocruzamento.

Dentro desse espírito, também foram reintroduzidas, com sucesso, indivíduos de duas espécies de jacu, *Penelope obscura bronzina* e *P. superciliaris jacupemba*, como parte de um esquema de reflorestamento posterior à construção de uma represa hidrelétrica em São Paulo. Marcadores moleculares (minissatélites) foram usados para acompanhar o processo de liberação e introgressão das aves reintroduzidas nas populações nativas (Pereira e Wajntal, 1999), que parecem ter se adaptado bem à reintrodução.

Como vimos, a estimativa da estruturação populacional é um passo importante em estudos de genética para a conservação, pois permite direcionar os esforços de conservação para um uso mais eficiente dos recursos disponíveis. Naturalmente, os programas de conservação sugeridos por esses estudos são cenários ideais, que nem sempre podem ser seguidos por questões de logística ou vontade política. De qualquer forma, os instrumentos existem e podem ser usados como guias para a formulação de políticas de conservação, principalmente quando são realizados estudos comparativos de várias espécies de um mesmo ambiente, permitindo a corroboração mútua das melhores estratégias (o que foi chamado por John Avise de "perspectiva regional da conservação"; Avise e Hamrick, 1996; Avise, 1998).

22.3.4. Revelando a biodiversidade escondida

Muitas vezes estudos moleculares revelam a existência de espécies biológicas diferentes dentro do que os taxônomos haviam considerado ser apenas uma espécie. Quando essas espécies têm pequenas diferenças, que haviam sido consideradas pelos taxônomos como simples variabilidade intraespecífica, elas são chamadas de espécies semicrípticas. Quando são, de fato, morfologicamente indistinguíveis, são denominadas espécies crípticas. Ambos os casos têm consequências importantes para a conservação (Bickford *et al.*, 2007).

A biodiversidade do planeta foi estimada em 13 milhões de espécies, das quais cerca de 1,75 milhões já foram descritas (Heywood e Watson, 1996), distribuídas em 64 filos, 146 classes, 869 ordens e cerca de 7.000 famílias. Os animais, com 1,5 milhões de espécies (das quais cerca de 900.000 são insetos), excedem em muito o número de espécies de plantas (60.000 conhecidas, 200.000 em total estimado).

Além da subestimativa da biodiversidade, causada pela deficiência de financiamentos para estudos de zoologia e de botânica de regiões pouco conhecidas (Wheeler, 1995), existe também uma outra fonte de subestimativa: a existência de espécies crípticas dentro de espécies consideradas como bem conhecidas (Knowlton, 1993; Thorpe e Solé-Cava, 1994; Knowlton, 2000). Neste campo, a genética tem contribuído enormemente, descobrindo inúmeras espécies novas, com a consequente redução da distribuição geográfica suposta para espécies anteriormente consideradas cosmopolitas. Pelo menos para o ambiente marinho, parece que, com exceção das espécies bioinvasoras (veja item 3.5), devem ser muito raras as espécies verdadeiramente cosmopolitas.

As consequências da descoberta de espécies crípticas dentro de espécies anteriormente consideradas cosmopolitas é particularmente importante, porque, justamente por sua suposta abundância e ampla distribuição, essas espécies são frequentemente usadas como modelos para estudos biológicos e para pesquisas em produtos naturais (Thorpe e Solé-Cava, 1994).

O mexilhão-europeu, *Mytilus edulis*, por exemplo, foi escolhido por sua ubiquidade no hemisfério norte como modelo de estudos de controle de poluição ("The mussel watch programme"; Daskalakis, 1997). No entanto, estudos moleculares indicaram que essa espécie tem regiões de ampla hibridação com outras duas espécies do gênero (*M. trossulus* e *M. galloprovincialis*), o que complica seu uso para controle ambiental (Skibinski *et al.*, 1978; Quesada *et al.*, 1995). Da mesma forma, descobriu-se que o anfípode *Hyalella azteca*, que há anos vinha

Tabela 22.2. Exemplos de populações naturais ameaçadas analisadas quanto aos seus níveis de estruturação gênica.

Grupo	**Nome vulgar**	Espécie	**Local**	**Gene**	**Método**	**Estruturação**	**REF**
Ave	"*Shrike*"	*Lanius ludovicianus*	Canadá	mtDNA	sequenciamento	baixa	1
Ave	Pica-pau	*Picoides borealis*	EUA	nuclear	minissatélites	alta	2
Ave	"*Clapper rail*"	*Rallus longirostris*	EUA	ambos	RAPD	baixa	3
Ave	"*Grouse*"	*Lagopus lagopu*	Escócia	nuclear	microssatélites	alta	4
Ave	Tangará dançarino	*Chiroxiphia caudate*	Brasil	nuclear	microssatélites	baixa	5
Coral	Coral chifre-de-veado	*Acropora cervicornis*	Caribe	ambos	sequenciamento	Moderada (nuclear) e alta (mtDNA)	6
Coral	Coral vermelho	*Corallium rubrum*	Mediterrâneo	nuclear	sequenciamento, microssatélites	alta	7
Aranha	Aranha	*Macrothele calpeiana*	Europa e Norte da África	mtDNA	sequenciamento	alta	8
Crustáceo	Pitu	*Austropotamobius pallipes*	Suíça	nuclear	aloenzimas	alta	9
Crustáceo	Pitu	*Austropotamobius pallipes*	Inglaterra	mtDNA	PCR/RFLP	baixa	10
Inseto	Mariposa	*Papilio machaon*	Inglaterra	ambos	RAPD	baixa	11
Inseto	Libélula	*Coenagrion mercuriale*	Inglaterra	nuclear	microssatélies	moderada	12
Inseto	Cigarra	*Cycas fairylakea*	China	ambos	AFLP	alta	13
Inseto	Besouro	*Hydrophilus piceus*	Inglaterra	nuclear	microssatélites	moderada	14
Mamífero	Foca cinza	*Halichoerus grypus*	Europa/Canadá	mtDNA	PCR/RFLP	alta	15
Mamífero	Boto do porto	*Phocoena phocoena*	Inglaterra	mtDNA	PCR/RFLP	baixa	16
Mamífero	Boto de Dall	*Phocoenoides dalli*	Europa	mtDNA	PCR/RFLP	baixa	17
Mamífero	Toninha	*Pontoporia blainvillei*	Brasil	mtDNA	sequenciamento	alta	18
Mamífero	Toninha	*Pontoporia blainvillei*	Argentina, Uruguai	mtDNA	sequenciamento	moderada	19, 20
Mamífero	Morcego-raposa	*Pteropus scapulatus*	Europa	nuclear	RAPD / aloenzimas	baixa	21
Mamífero	Koala	*Phascolarctos cinereus*	Austrália	ambos	RAPD	alta	22
Mamífero	Veado-dos-pampas	*Ozotoceros bezoarticus*	Brasil/Argentina	mtDNA	sequenciamento	alta	23
Mamífero	Urso	*Ursus arctos*	Japão	mtDNA	sequenciamento	alta	24
Mamífero	Bisão	*Bison bison*	EUA	nuclear	microssatélites	baixa	25
Mamífero	Morcego	*Nyctalus azoreum*	Açores	nuclear	microssatélites	alta	26
Mamífero	Preguiça	*Bradypus torquatus*	Brasil	mtDNA	sequenciamento	alta	27
Molusco	Mexilhão-de-água-doce	*Pyganodon grandis*	EUA	ambos	aloenzimas / RFLP	alta	28
Molusco	Haliote da Califórnia	*Haliotis cracherodii*	Califórnia, EUA	ambos	sequenciamento, microssatélites, AFLP	moderada	29
Réptil	Tartaruga radiada	*Geochelone radiata*	Madagascar	nuclear	microssatélites	moderada	30
Réptil	Tartaruga de pente	*Eretmochelys imbricata*	Caribe	mtDNA	sequenciamento	alta	31
Réptil	Lagarto	*Podarcis atrata*	Espanha	mtDNA	sequenciamento	alta	32
Réptil	Cascavel	*Sistrurus catenatus*	Canadá	ambos	RAPD, microssatélites	alta	33, 34
Réptil	Serpente-marinha	*Aipysurus laevis*	Austrália	mtDNA	sequenciamento	alta	35
Anfíbio	Sapo	*Litoria áurea*	Austrália	mtDNA	sequenciamento	alta	36
Anfíbio	Sapos "*corroboree*"	*Pseudophryne corroboree* *Pseudophryne pengilleyi*	Austrália	ambos	sequenciamento, microssatélites	alta	37
Anfíbio	Sapos	*Proceratophrys boiei* e *Ischnocnema gr. ramagii*	Brasil	ambos	sequenciamento	alta	38
Peixe	"*Razzorback sucker*"	*Xyrauchen texanus*	EUA	mtDNA	PCR/RFLP	alta	39
Peixe	Cyrprinodontideo	*Valencia*	Espanha	nuclear	aloenzimas	alta	40
Peixe	Espátula	*Polyodon spathula*	EUA	mtDNA	sequenciamento	alta	41
Peixe	Tubarão-baleia	*Rhincodon typus*	Todos os oceanos	mtDNA	sequenciamento	moderada	42
Peixe	Bagre gigante	*Pangasianodon gigas*	Tailândia e Camboja	ambos	sequenciamento, microssatélites	baixa	43
Peixe	Espadarte	*Xiphias gladius*	Atlântico	mtDNA	sequenciamento	moderada	44
Planta	Pau-brasil	*Caesalpinia echinata*	Brasil	ambos	RAPD	alta	45
Planta		*Armeria maritima*	Dinamarca	nuclear	aloenzimas	alta	46

Tabela 22.2. Exemplos de populações naturais ameaçadas analisadas quanto aos seus níveis de estruturação gênica. (cont.)

Planta	Orquídea	*Cypripedium kentuckiense*	EUA	nuclear	aloenzimas	alta	47
Planta	Araucária	*Araucaria angustifolia*	Brasil	nuclear	AFLP, microssatélites	moderada e alta	48
Planta	Bromélia	*Alcantarea glaziouana*	Brasil	nuclear	microssatélites	alta	49
Planta	Bromélia	*Tillandsia achyrostachys*	México	nuclear	aloenzimas	alta	50
Planta	Cagaiteira	*Eugenia dysenterica*	Brasil	nuclear	aloenzimas	alta	51

Referências: 1 - Mundy et al., 1997; 2 - Haig et al., 1993; 3 - Nusser et al., 1996; 4 - Piertney et al., 1998; 5 – Francisco et al., 2007; 6 - Vollmer e Palumbi, 2007; 7 - Costantini et al., 2007; 8 - Arnedo e Ferrández, 2007; 9 - Lortscher et al., 1998; 10 - Grandjean et al., 1997; 11 – Hoole et al., 1999; 12 - Watts et al., 2004; 13 - Jian et al., 2006; 14 – Beebee, 2007; 15 - Boskovic et al., 1996; 16 - Walton, 1997; 17 - McMillan e Bermingham, 1996; 18 – Secchi et al., 1998; 19 – Lazaro et al., 2004; 20 – Mendez et al., 2008; 21 - Sinclair, 1996; 22 – Fowler et al., 1998; 23 – Gonzalez et al., 1998a; 24 – Matsuhashi et al., 1999; 25 - Wilson e Strobeck, 1999; 26 - Salgueiro et al., 2008; 27 - Lara-Ruiz et al., 2008; 28 - Liu et al., 1996; 29 – Gruenthal e Burton, 2008; 30 - Paquette et al., 2007; 31- Bass et al., 1996; 32 – Castilla et al., 1998; 33 – Gibbs et al., 1994; 34 – Gibbs et al., 1997; 35 - Lukoschek et al., 2007; 36 - Burns et al., 2007; 37 - Morgan et al., 2008; 38 - Carnaval e Bates, 2007; 39 - Dowling et al., 1996; 40 - Perdices et al., 1996; 41 - Epifanio et al., 1996; 42 - Castro et al., 2007; 43 - Ngamsiri et al., 2007; 44 - Bremer et al., 2005; 45 - Cardoso et al., 1998; 46 - Weidema et al., 1996; 47 - Case et al., 1998; 48 - Stefenon et al., 2007; 49 – Barbará et al., 2008; 50 - Gonzalez-Astorga e Castillo-Campos, 2004; 51 – Telles et al., 2003.

sendo usado como indicador em testes de ecotoxicologia para estudo de qualidade da água em vários laboratórios do mundo, é, na verdade, um aglomerado de pelo menos quatro espécies diferentes (Hogg *et al.*, 1998).

Outra consequência importante da subestimativa da biodiversidade para a conservação e a exploração racionais das populações naturais acontece quando populações de um organismo explorado comercialmente são constituídas, na verdade, de mais de uma espécie (ver Thorpe *et al.*, 2000, para uma revisão desse problema para pesca de invertebrados marinhos). Ao se considerarem espécies diferentes exploradas comercialmente como se fossem apenas uma espécie, corre-se um grande risco de extinguir a espécie ecologicamente mais frágil sem que sequer se tenha percebido. Se o controle de estoques pesqueiros depende muito da delimitação dos estoques, conforme visto acima, ele depende ainda mais da detecção de espécies crípticas, que podem ser vistas como um caso mais extremo de diferenciação populacional. No Brasil, por exemplo, se considerava a existência de apenas uma espécie de cação-anjo (*Squatina argentina*), uma espécie pescada comercialmente principalmente na região sul do país. No entanto, um estudo molecular provocado por diferenças observadas por biólogos pesqueiros na dinâmica das populações deste cação revelou que, na verdade, existiam três espécies diferentes do gênero no sul do Brasil, que eram confundidas com "variação ontogenética" ou "polimorfismos naturais" de *S. argentina* (Solé-Cava *et al.*, 1983; Solé-Cava e Levy, 1987). Também foi descoberto por análises moleculares que a espécie de camarão comercialmente mais importante da costa brasileira (*Farfantepenaeus subtilis*) é, na verdade, a soma de duas espécies distintas (Gusmão *et al.*, 2000). O mesmo foi observado com o camarão-sete-barbas (*Xiphopenaeus kroyeri*), que, de fato, inclui duas espécies diferentes no Brasil (Gusmão *et al.*, 2006). Da mesma forma, dificuldades na aquacultura da ostra-do-mangue, *Crassostrea rhizophorae*, podem ser explicadas pelo fato dela corresponder, de fato, a duas espécies (Ignacio *et al.*, 2000). A dificuldade em muitos cruzamentos dessas ostras em laboratório se devia, portanto, à tentativa involuntária de efetuar cruzamentos entre espécies diferentes.

Outros exemplos são a tartaruga-lora (*Lepidochelys kempii*)—uma espécie rara, confundida com a tartaruga-oliva (*L. olivacea*) — e a orquídea do Arkansas (*Cypripedium kentuckiense*)—interpretada inicialmente como variedade de uma orquídea comum nos EUA, *C. parviflorum.* Nos dois casos, o táxon mais raro teria sido extinto se não fossem consideradas suas diferenças evolutivas e, consequentemente, tomadas medidas específicas de proteção. Isso só foi possível com trabalhos de genética, usando mtDNA, no caso da tartaruga (Bowen *et al.*, 1991), e aloenzimas, no caso da orquídea (Case *et al.*, 1998). Assim como nesses exemplos, também é provável que existam espécies crípticas de peixe-boi-marinho (*Trichechus manatus*), que corresponderiam às ESU detectadas usando a região controle mitocondrial (Garcia Rodriguez *et al.*, 1998; Vianna *et al.*, 2006). Esse resultado tem consequências importantes para a preservação, pois mostra que o peixe-boi-marinho, ameaçado em toda sua distribuição, é ainda mais frágil do que se pensava.

Espécies crípticas são encontradas em todos os grupos de organismos (Pfenninger e Schwenk, 2007), não se restringindo, portanto, apenas a grupos obscuros de invertebrados. Por exemplo, análises moleculares feitas com sequências de quatro genes nucleares (Roca *et al.*, 2001) revelaram que o elefante-africano era constituído, na verdade, de duas espécies diferentes, uma nas savanas (*Loxodonta africana*) e outra nas florestas (*L. cyclotis*). No caso das girafas, os resultados foram ainda mais surpreendentes: o que era considerada uma única espécie (*Giraffa camelopardalis*) revelou ser, após uma análise de sequências mitocondriais e microssatélites, um conjunto de seis espécies crípticas (Brown *et al.*, 2007).

Um caso recente de identificação de espécies crípticas em um mamífero aquático no Brasil foi o dos golfinhos do gênero *Sotalia*. Após mais de um século de incerteza taxonômica, marcadores moleculares mostraram que os "ecótipos" marinho e fluvial de *Sotalia fluviatilis* na realidade são espécies diferentes (Cunha *et al.*, 2005). A existência de duas espécies havia sido sugerida por um estudo morfológico (Monteiro *et al.*, 2002) e também era indicada por diferenças em parâmetros biológicos e ecológicos dos dois ecótipos. Com a verificação das diferenças genéticas, a espécie *Sotalia guianensis* foi revalidada para os animais marinhos, enquantos os fluviais retiveram o binômio *Sotalia fluviatilis*. Esse resultado teve importantes implicações para a conservação dessas espécies, pois *S. fluviatilis* passou a ser o único delfinídeo exclusivamente fluvial do mundo e também uma espécie endêmica da Bacia Amazônica (Cunha *et al.*, 2005).

Estudos de biodiversidade molecular aplicada à estimativa da biodiversidade nominativa, portanto, serviram para aumentar, na maioria das vezes, o número de espécies conhecidas. No entanto, em alguns casos, os estudos genéticos indicaram o contrário, ou seja, que espécies consideradas distintas eram, na verdade, variedades ou híbridos de outras espécies. Esses resultados têm muita relevância para a conservação, pois permitem que recursos não sejam desperdiçados na proteção ou tentativa de manutenção em cativeiro de indivíduos que não têm independência evolutiva. É como se, ao observarmos um eventual declínio na população de burros e mulas, resolvêssemos investir grandes somas em sua preservação em zoológicos etc., sem saber que, de fato, eles são híbridos de jumentos com éguas... Como a biologia dos burros

e mulas é bem conhecida, isso não aconteceria. Mas e se isso acontecesse com espécies menos conhecidas? Por exemplo, uma subespécie de um pardal-caiçara-americano (o "dusky sea-side sparrow", *Ammodramus maritimus nigrescens*) foi descrita no século XIX e era relativamente abundante na Flórida até a metade do século XX. Quando a destruição de seus hábitats pela especulação imobiliária levou a uma grande diminuição da população dessa ave, foram feitos programas na tentativa de preservação desses pardais. Esses programas foram mal-sucedidos até que o último exemplar morreu em cativeiro em 1987. No entanto, um estudo posterior, feito com marcadores de DNA mitocondrial, revelou que o *A. maritimus nigrescens* era geneticamente indistinguível de outras supostas raças dessa espécie na costa leste Americana (mas diferente das raças da parte próxima ao Golfo do México), sendo, provavelmente, apenas uma forma melânica de alguma delas (Avise e Nelson, 1989). Apesar da perda daquela população ser de qualquer forma lamentável, ela foi, do ponto de vista genético, uma perda limitada de alguns genes diferenciados naquela população.

A genética, portanto, tem tido um papel fundamental na reavaliação da biodiversidade nominal. Já se passaram quase 40 anos desde a publicação do artigo visionário de John Avise sobre o uso de métodos moleculares na sistemática (Avise, 1974). Desde então, um número enorme de trabalhos científicos demonstrou que os sistematas frequentemente haviam sido conservadores demais na atribuição de espécies, aceitando como variações intraespecíficas diferenças sutis que, na verdade, marcavam espécies diferentes. Os pesquisadores trabalhando atualmente com biodiversidade estão mais conscientes desse problema e mais preocupados em analisar melhor a biologia das espécies (e, quando necessário, sua genética) antes de construírem amplas listas de sinonímia ou de expandirem a distribuição geográfica de espécies pela citação de novas ocorrências. A sistemática molecular certamente nunca substituirá os métodos descritivos criteriosos da sistemática baseada em morfologia (para uma crítica sobre os limites na identificação molecular rápida e unidimensional, por uma abordagem chamada de Barcodes de DNA, veja Solé-Cava, 2008a) ou seu trabalho de construção do sistema completo conectando as espécies em grupos mais abrangentes. No entanto, a sistemática molecular tem auxiliado enormemente na detecção de espécies crípticas (Thorpe e Solé-Cava, 1994; Knowlton, 2000) e poderá ser muito útil, também, na definição de taxa supra-específicos, talvez funcionando como base de um padrão universal que permita atribuir a esses níveis algum sentido evolutivo (Avise e Johns, 1999).

22.3.5. Bioinvasões

Se, por um lado, um dos resultados dos estudos genéticos tem sido a descoberta de que as distribuições geográficas amplas atribuídas a muitas espécies estão exageradas (ver acima), por outro, esses estudos têm servido também para detectar e acompanhar um processo inverso—o da invasão, com auxílio antropogênico direto ou indireto, de várias regiões do planeta por espécies exóticas (Holland, 2000).

As espécies mais intimamente associadas a nossa espécie, como nossos parasitas e comensais, têm distribuição tão cosmopolita quanto nós, de modo que podemos encontrar, por exemplo, piolhos, camundongos e ácaros das mesmas espécies distribuídos desde as regiões equatoriais até os círculos polares. O mesmo se observa com os animais e plantas domesticados, que atualmente têm distribuições completamente cosmopolitas.

A questão das bioinvasões, no entanto, refere-se a espécies não tão diretamente associadas a nossa e, consequentemente, menos conhecidas, sem por isso serem menos importantes. O impacto causado por bioinvasões é muito maior do que geralmente se reconhece. Somente nos Estados Unidos, por exemplo, foram registradas, até 1993, mais de 4.500 espécies invasoras (Mills *et al.*, 1994). Estima-se que bioinvasores sejam o segundo maior problema para a conservação no planeta, perdendo apenas para a destruição de hábitats (Clavero e Garcia-Berthou, 2005).

Um número enorme de espécies marinhas teve sua distribuição geográfica aumentada pelo transporte involuntário nos cascos de navios (Carlton e Hodder, 1995; Southward *et al.*, 1998) e, mais recentemente, nos tanques de lastro (Carlton e Geller, 1993). Os navios são vetores importantes na introdução de espécies exóticas: existem atualmente mais de 35.000 navios cargueiros circulando no globo (Carlton, 1985). Considerando-se que um único grande cargueiro pode carregar mais de 150.000 toneladas de água de lastro (Schormann *et al.*, 1990), podemos verificar o potencial desse tipo de transporte para as bioinvasões (Pierce *et al.*, 1997). Várias das espécies bioinvasoras marinhas representam verdadeiras pragas. Elas podem provocar: destruição de hábitats naturais (Lafferty *et al.*, 2005); competição (Russell *et al.*, 1994; Fleming, 2000) ou predação (Kuris, 1991) de espécies nativas; produção de toxinas (Hamer, 1991); e encrustamento de estruturas construídas pelos seres humanos, como pilares em cais de portos, tubos de refrigeração de usinas nucleares, plataformas de petróleo etc. (Lafferty e Kuris, 1996; Cohen e Carlton, 1998). Marcadores moleculares podem ser usados para auxiliar tanto na identificação das espécies invasoras (May e Marsden, 1992; Geller *et al.*, 1994; Southward *et al.*, 1998) e de sua origem geográfica (Holland, 2000), como no acompanhamento da evolução do processo invasivo (Cohen *et al.*, 1995; Holland, 2000) e no controle das populações invasoras (Hampton *et al.*, 2004).

Populações invasoras em geral têm baixa variabilidade gênica, principalmente se as invasões ocorrem a partir de um número pequeno de organismos fundadores (Holland, 2000). No entanto, também já foram observadas variabilidades gênicas altas em espécies invasoras e chegou a ser argumentado que altas heterozigosidades facilitariam o sucesso da bioinvasão (Pappert *et al.*, 2000), embora isso não tenha sido ainda testado. Essa relação entre evento de invasão e baixa heterozigosidade foi observada na mosca invasora *Zaprionus indianus*, que invadiu o Brasil provavelmente a partir da África no final do século XX (Mattos-Machado *et al.*, 2005). Por outro lado, altos níveis de variabilidade (Lizarralde *et al.*, 2008) foram observados nos castores invasores da Terra do Fogo (*Castor canadensis*), cuja população foi fundada por 25 casais em 1946 (para criação e comércio de peles), mas que se tornou espécie invasora e atualmente conta com mais de 100.000 indivíduos.

Um fator importante no sucesso de espécies invasoras é a "saúde" do ambiente invadido: da mesma forma que um corpo fragilizado por doenças é mais facilmente invadido por bactérias e vírus, os ambientes abalados por poluição ou outros desequilibrios também são mais facilmente colonizados. Em um lago no estado de Nova Iorque, por exemplo, um estudo molecular com formas dormentes do cladócero asiático invasor *Daphnia curvirostris* encontrados nas diversas camadas dos sedimentos revelou que a invasão ocorreu nos anos 50, atingindo um pico populacional nos anos 60 e 70, durante a época em que o lago foi mais poluído. No entanto, a recuperação do lago, após programas de controle da poluição nos anos 80, levou a espécie invasora a desaparecer, sendo substituída por espécies locais (Duffy *et al.*, 2000).

No ambiente terrestre, as bioinvasões também são um grande problema (Manchester e Bullock, 2000). Um exemplo bastante próximo de nós foi a bioinvasão causada pela liberação acidental de abelhas africanizadas da espécie *Apis mellifera* no Estado de São Paulo nos anos 1950. Essa variedade é bastante

agressiva e, desde então, se espalhou por todo o país, chegando a atingir, através da América Central, o sul dos EUA (uma boa revisão deste incidente pode ser encontrada em Sheppard e Smith, 2000). Existia um debate se a alta dispersão dessa variedade de abelha se devia aos zangões, que regularmente abandonariam a colônia e migrariam, cruzando com a variedade européia aonde chegassem, ou se seriam enxames completos que de vez em quando fariam uma migração. Essas hipóteses foram testadas pela comparação da estruturação genética das populações dessas abelhas nas regiões invadidas, estimada usando marcadores nucleares (transmitidos pelos zangões e pelas rainhas) e usando marcadores mitocondriais (transmitidos apenas pelas rainhas). O resultado observado foi que, apesar da ocorrência de alguma introgressão entre as duas variedades (Lobo *et al.*, 1989), ao longo da distribuição atual das abelhas africanas, o padrão de mtDNA não mudava muito e era diferente do padrão da abelha européia, apoiando, assim, a hipótese dos enxames (Hall, 1992).

Bioinvasores podem interferir na evolução das espécies nativas, como foi recentemente documentado na bacia do Rio Colorado, EUA. Marcadores mitocondriais e nucleares mostraram que uma espécie de peixe (o "white sucker", *Catostomus commersoni*), introduzida nessa bacia no início do século XX, está hibridizando com duas espécies nativas: o "flannelmouth sucker" (*C. latipinnis*) e o "bluehead sucker" (*C. discobolus*). A análise de atribuição revelou ainda a existência de híbridos de segunda ou mais gerações, que possuem genótipos com contribuição das três espécies (os "muttsuckers"). Ou seja, a introdução dos "white suckers" levou à ruptura do isolamento reprodutivo entre as espécies nativas, que antes não hibridizavam entre si. A situação é mais grave para o "flannelmouth", que parece hibridizar mais facilmente com a espécie introduzida (McDonald *et al.*, 2008). Apesar de esses peixes não serem ameaçados, processos semelhantes podem estar acontecendo com espécies em risco submetidas ao contato com bioinvasores.

Um exemplo de aplicação de marcadores genéticos no controle de populações invasoras é o dos porcos ferais, que são uma praga na Austrália. Esses bioinvasores são um risco para as espécies nativas ameaçadas, por causa da predação, competição, por modificarem os hábitats, e também pela disseminação de doenças. Como a erradicação de vertebrados bioinvasores é considerada extremamente difícil ou impossível (Bomford e O'Brien, 1995), o controle em muitos casos restringe-se a minimizar o impacto das populações invasoras. A fim de controlar o tamanho das populações de porcos ferais, esses animais são periodicamente abatidos. Entretanto, a abundância de algumas populações não diminuía após os abates. Para entender esse fenômeno, dados de microssatélites foram usados em uma análise de atribuição com amostras coletadas ao longo de uma grande área no leste da Austrália. O resultado mostrou que as populações que não reduziam de tamanho eram geneticamente similares a outras populações porque recebiam migrantes dessas populações logo após os abates. Como consequência, a estratégia de controle somente funcionaria se as populações que trocam migrantes forem agrupadas em Unidades de Erradicação (*sensu* Robertson e Gemmel, 2004) e se os abates forem realizados concomitantemente em toda a extensão de cada Unidade de Erradicação. Naquele estudo, foram detectadas cinco Unidades de Erradicação para os porcos ferais na região (Cowled *et al.*, 2008).

Um outro caso interessante de bioinvasão foi detectado recentemente, mostrando que os guaxinins das Índias Ocidentais (Bahamas, Guadalupe e Barbados) não constituem espécies insulares endêmicas e ameaçadas (*Procyon maynardi*, de Bahamas, e *P. minor*, de Guadalupe) e tampouco uma espécie extinta (*P. gloveralleni*, de Barbados). Em vez disso, as populações dessas ilhas pertencem à espécie invasora *P. lotor* (dos Estados Unidos), conforme demonstrado tanto pela morfologia quanto pelo DNA mitocondrial (Helgen *et al.*, 2008). Observando a rede de haplótipos, verifica-se que os haplótipos insulares se posicionam no meio dos continentais, o que também indica que houve múltiplas introduções no Caribe. Esse resultado tem consequências importantíssimas, já que essas "espécies" de guaxinins eram consideradas ameaçadas e recebiam grandes esforços de conservação, incluindo a introdução de indivíduos em outras ilhas. No entanto, com a conclusão de que os guaxinins são invasores, passa a ser necessário inverter a estratégia, buscando erradicá-los, já que esses animais predam espécies verdadeiramente nativas e ameaçadas, como iguanas, tartarugas marinhas e aves que nidificam no solo, em especial o papagaio-das-Bahamas (*Amazona leucocephala bahamensis*) (Helgen *et al.*, 2008).

Um problema na aplicação de marcadores moleculares no estudo de populações bioinvasoras é que os modelos de estruturação gênica geralmente usados foram desenvolvidos para um estado de equilíbrio de endocruzamento, que leva muitas gerações para ser atingido depois de eventos de fundação de novas áreas. Por exemplo, se estudarmos uma população de uma ascídia que ocorre em Hong Kong e a compararmos com uma outra, introduzida por acidente há 100 anos no Rio de Janeiro, poderemos encontrar, a partir de um estudo com marcadores moleculares, um F_{ST} baixo, indicando um suposto alto fluxo gênico entre as duas áreas, que de fato não tem ocorrido desde então! Isso ocorre porque as frequências gênicas não terão tido tempo para divergir, deflacionando, assim, nossa estimativa de endocruzamento local (Crochet, 1996). Uma solução parcial para esse problema é o uso de marcadores com taxas de mutação muito elevadas (como os microssatélites), que podem ser interpretados usando-se modelos genéticos de desequilíbrio (Davies *et al.*, 1999). Portanto, os métodos capazes de detectar estruturação em situações fora do equilíbrio migração-deriva são extremamente úteis tanto no caso de espécies ameaçadas quanto nas bioinvasões (exemplos desses métodos são apresentados no Capítulo 21).

22.3.6. Identificação forense para conservação

Um aspecto importante da conservação da biodiversidade é a formulação de leis de controle do uso, comércio e exportação de produtos de animais e plantas ameaçadas de extinção. O problema é que, pelo menos até a cultura do consumo daquele produto se extinguir, o aumento da demanda causado pelo controle—e consequente escassez—dele faz com que seu preço aumente, estimulando os comerciantes a burlar a lei (Baker *et al.*, 1996; De Salle e Birstein, 1996; Rehbein *et al.*, 1997; Palumbi e Cipriano, 1998). Os valores envolvidos podem ser muito altos. Por exemplo, os atuns maiores podem render mais de U$ 6.000 por peixe (Ward, 1995); as ovas de esturjão custam U$ 2.400 por quilo (De Salle e Birstein, 1996); a pele de crocodilo pode custar de U$ 200 a US$ 500 por pele (Brazaitis *et al.*, 1998); e a carne de várias espécies de baleia pode custar mais de U$ 50 por quilo (Sweijd *et al.*, 2000b).

O problema é ainda mais complicado quando existem espécies próximas cujo comércio é legalizado ou quando a espécie protegida pode ser comercializada se proveniente de cultivo. Nesses casos, as fazendas de cultivo podem ser usadas para a legalização fraudulenta de produtos das espécies protegidas e sua inserção no mercado. Já se observou, por exemplo, que a comercialização mundial de peles de jacarés (gênero *Caiman*) é superior a um milhão de peles por ano, das quais apenas a metade vem realmente de fontes legalizadas (Brazaitis *et al.,* 1998). A espécie brasileira mais vulnerável do gênero é o jacaré-do-pantanal (*Caiman yacare*) e sua importação pelos EUA está proibida. No

entanto, numa tentativa de driblar esse bloqueio de importação, foi argumentado recentemente que essa espécie teria três subespécies, uma das quais poderia ser explorada (ver Brazaitis *et al.*, 1996, para um histórico do caso). Felizmente, estudos genéticos com sequenciamento de DNA demonstraram que essa subdivisão não era justificada (Brazaitis *et al.*, 1998) e a importação de pele de *C. yacare* continua proibida nos EUA.

No caso do jacaré, a identificação das peles pode ser feita pelo padrão de manchas e pelo seu relevo. No entanto, em outras situações, como na comercialização da carne de animais protegidos, essa identificação não é tão simples. Nesses casos, marcadores moleculares podem ser de extrema valia ao permitirem a identificação não ambígua mesmo de produtos industrializados, como carnes salgadas, cozidas, defumadas ou enlatadas (Ram *et al.*, 1996; Quintero *et al.*, 1998). Métodos moleculares foram usados, por exemplo, para demonstrar que um carregamento de carne de caranguejo (*Paralithodes camtschatica)* havia sido obtido em uma área de sua distribuição onde era ilegal a pesca (Seeb *et al.*, 1990). Em outro caso, esses métodos foram usados para provar na justiça que a carne do haliote (os gastrópodes de maior valor comercial do mundo, do gênero *Haliotis*) vendida na África do Sul como "haliote australiano" na verdade pertencia a *Haliotis midae*, uma espécie local cuja pesca estava proibida (Sweijd *et al.*, 2000).

A demonstração mais dramática e elegante do uso forense de marcadores moleculares foi o trabalho feito pelo grupo de Stephen Palumbi sobre o comércio ilegal de carne de baleia no Japão. As baleias eram animais abundantes nos oceanos até o advento, no início do século XX, dos navios a vapor e seus canhões de arpão, seguidos, após a Segunda Guerra Mundial, dos navios-fábrica. Estima-se, por exemplo, que de 1920 até 1986, quando uma moratória na caça indiscriminada das baleias foi imposta, mais de um milhão de baleias tenham sido mortas pelos baleeiros (Baker *et al.*, 1996). A moratória teve um efeito importante na proteção das baleias. No entanto, um buraco na lei permite que baleias caçadas "para fins científicos" sejam consumidas nos países que as caçaram (ou seja, a lei proíbe a exportação, mas permite o consumo, Baker *et al.*, 2000a). Os resultados disso foram que: (a) o Japão passou a ser o país com o maior número de pesquisas sobre baleias (o Japão é também o maior consumidor de carne de baleia do mundo...); e (b) o comércio de carne de baleia "legalizada" nos mercados japoneses continuou.

O trabalho feito por S. Palumbi e seu grupo consistiu na construção inicial de uma "biblioteca de sequências"— um banco de dados com as sequências de DNA da região controle mitocondrial de todas as baleias conhecidas, incluindo amostras intraespecíficas de várias partes do globo, coletadas por dardos de biópsia, que retiram uma pequena quantidade de tecido sem gerar qualquer ferimento grave na baleia. A região controle mitocondrial é altamente variável em mamíferos, permitindo a distinção fácil de todas as espécies e, inclusive, em muitos casos, possibilitando até a determinação da origem geográfica de cada baleia (Baker e Palumbi, 1994). Tendo construído essa biblioteca, amostras rotuladas como "carne de baleia" foram compradas em feiras e mercados do Japão por agentes da ONG ambientalista Earthtrust e enviadas clandestinamente para os laboratórios de Palumbi nos EUA para um estudo preliminar. Nesse estudo, foi descoberto que, das 17 amostras compradas, apenas nove pertenciam a baleias-minke-do-sul (*B. bonaerensis*), as quais o governo japonês tem permissão de caçar (para fins científicos). As outras oito amostras eram provenientes de baleias de captura proibida, como a jubarte, cuja caça é proibida desde 1966, ou de botos e golfinhos (ou seja, enganando também o consumidor japonês).

Esse estudo preliminar estimulou um novo estudo, dessa vez com uma preocupação forense, a fim de processar as companhias responsáveis pela comercialização ilegal da carne da baleia. Como a exportação de produtos de baleia é proibida e o DNA é um produto da baleia, os dados do estudo, se envolvessem DNA de baleia exportado do Japão para os EUA, não poderiam ser usados como prova de acusação por terem sido obtidas ilegalmente (Palumbi e Cipriano, 1998; Palumbi e Cipriano, 1999). A solução encontrada foi montar um "laboratório portátil" de extração de DNA, PCR e purificação dos produtos do PCR e usá-lo *in loco* no Japão. O procedimento realizado foi comprar a carne com documentação fotográfica, extrair o DNA e fazer o PCR usando *primers* biotinados. Depois do PCR, a biotina ligada aos *primers* foi usada para auxiliar a remoção do DNA original da baleia através do uso de minúsculas bolas magnéticas cobertas de streptavidina, que se liga fortemente à biotina que estava nos *primers* e, portanto, aos fragmentos de DNA amplificados por PCR. Esse DNA "livre de DNA original de baleia" foi então exportado para os EUA, onde foi analisado em sequenciadores automáticos. Nesse segundo estudo, também foram usados *primers* de genes nucleares para evitar o possível argumento pela defesa de que as espécies estavam identificadas incorretamente porque teriam vindo de hibridações na natureza (em cujo caso seriam identificadas erroneamente como se pertencessem à espécie da mãe). O resultado desse estudo foi que, num total de 237 amostras analisadas, 149 pertenciam a baleias minke (sendo 107 minkes do hemisfério sul, *B. bonaerensis*), 49 eram golfinhos e botos, e as demais amostras eram de outras espécies de baleias. Ou seja, mais de um terço das amostras não pertenciam a baleias caçadas legalmente (Palumbi e Cipriano, 1998). Devido ao sucesso dessas pesquisas, a Comissão Baleeira Internacional (IWC) decidiu, em 1999, empregar a técnica desenvolvida por Palumbi e seu grupo como metodologia oficial de controle da comercialização da carne de cetáceos.

Outra espécie ameaçada de extinção, por causa da intensa exploração comercial, é o tubarão-elefante (*Cetorhinus maximus*). A comercialização de sua barbatana pode alcançar valores elevados no mercado (essas barbatanas podem custar mais de U$ 50.000 cada uma!). A identificação específica das barbatanas de tubarão é muito difícil, o que atrapalha o controle da comercialização das espécies ameaçadas. Entretanto, a identificação genética das barbatanas foi possível a partir de PCR multiplex de espaçadores ribossomais transcritos (ITS), que permitiram detectar a qual espécie pertencia cada barbatana, revelando que barbatanas de tubarão-elefante eram vendidas em várias localidades do mundo, inclusive nos EUA, onde a venda de espécies protegidas é fortemente regulada (Magnussen *et al.*, 2007).

Uma aplicação direta da genética forense para conservação que teve resultados concretos imediatos foi a análise feita com carne de tartaruga em Porto Rico, onde mais de 1500 tartarugas marinhas eram mortas por ano para o preparo de guisados em restaurantes. Apesar de os donos de restaurantes dizerem que as carnes usadas eram de espécies de caça permitida, a análise por PCR/RFLP do gene mitocondrial do citocromo b revelou que a maior parte dos restaurantes testados comercializava carne de captura ilegal, o que resultou em 11 condenações (Moore *et al.*, 2003).

No Brasil, também há exemplos de estudos com genética forense aplicada à conservação. A identificação forense já foi aplicada em amostras de olhos e genitálias de botos comercializadas como talismãs nos mercados populares da região norte do País. De acordo com superstições locais, os olhos e genitálias do boto-vermelho (*Inia geoffrensis*) possuem poderes mágicos e são capazes de atrair parceiros sexuais para as pessoas

que carregam tais amuletos. A análise genética de olhos vendidos em três mercados (em Belém, Manaus e Porto Velho) revelou que todos os olhos obtidos de golfinhos pertenciam a outra espécie, o boto-cinza (*Sotalia guianensis*). Das amostras de Porto-Velho, 90% na verdade eram olhos de porco ou ovelha (Gravena *et al.*, 2008). Em outro estudo, amostras de genitálias do mercado Ver-O-Peso, de Belém, foram analisadas e também foram identificadas como *S. guianensis* (Sholl *et al.*, 2008).

As abordagens Bayesianas de atribuição permitiram que fossem desenvolvidas análises de identificação forense mais sensíveis. Com essas abordagens, é possível identificar não apenas a espécie à qual pertence uma amostra, mas também sua origem populacional. Por exemplo, uma análise de atribuição baseada em microssatélites permitiu verificar que alguns indivíduos de cervos-vermelhos (*Cervus elaphus*) em Luxemburgo haviam sido ilegalmente translocados de outras regiões para agregar valor às terras de um fazendeiro (Frantz *et al.*, 2006).

22.4. Perspectivas

Como pudemos ver ao longo deste texto, a genética pode ter aplicações importantes na conservação da natureza (Amos e Balmford, 2001). Esse é um ramo fascinante da ciência, com um número grande de abordagens e desafios diferentes. Marcadores moleculares de vários tipos podem ser usados para a identificação de produtos de uma espécie protegida, servindo de subsídio para o cumprimento da lei. Eles podem servir para determinar o sexo de animais antes de serem tentados cruzamentos em cativeiro, servem para a descoberta de espécies crípticas—permitindo a reavaliação da biodiversidade—e para a verificação da heterogeneidade espacial na biodiversidade molecular de cada espécie.

A maior parte dos estudos em genética para a conservação feitos até hoje envolveu espécies carismáticas, que atraem grande interesse público e, portanto, bons financiamentos para pesquisa. No entanto, é certo que esses processos tenham acontecido e estejam acontecendo em muitas outras espécies menos felizes em sua atração do interesse público. Muitos desses organismos já podem ter se extinguido sem que sequer tenhamos notado que eles existiam.

É fundamental que as políticas de preservação das espécies contemplem a questão da variabilidade gênica como um todo, para que não preservemos apenas espécies "aleijadas" geneticamente, que permanecerão indefinidamente sob custódia humana. O custo da proteção de uma espécie é, em geral, exponencial e inversamente proporcional a sua abundância: proteger uma área, pela demarcação e fiscalização de parques nacionais, tem um custo infinitamente menor do que tentar manter/reproduzir em cativeiro cada uma das espécies daquela área uma vez que estejam ameaçadas (Adams e Carwardine, 1990; Solé-Cava, 1993). Estudos genéticos podem ser úteis também na sugestão de políticas de preservação em seus estágios iniciais e mais simples, como na determinação de estruturas populacionais e na análise filogeográfica. No entanto, devido à miopia evolutiva de nossa espécie, e particularmente daqueles em posições com poder para tomada de decisões, a ciência tem sido usada mais frequentemente no estudo de espécies com populações já muito reduzidas, muitas vezes em fase terminal, como pôde ser visto ao longo deste artigo. A genética certamente é instrumento poderoso nesses estudos, juntamente com a ecologia. Mas a prevenção do processo de extinção por meio da proteção ambiental seria um uso mais inteligente dos recursos. Lamentavelmente, a preocupação com a cura, não só no Brasil, mas na maioria dos países, vem somente após a catástrofe.

Os trabalhos em conservação têm uma urgência particular quando comparados com outros estudos em ecologia molecular, pois o próprio objeto das pesquisas desaparece rapidamente. Assim, faz parte da motivação das pessoas que querem trabalhar na área a compaixão pelas espécies e a admiração da beleza da natureza (Miller, 2005; veja também Orr, 1992 e outros do mesmo autor). De certa forma, a ciência muitas vezes acaba funcionando de maneira contrária à conservação. Isso não é devido apenas ao fato de a ciência ter sido a base do grande desenvolvimento industrial que tem levado à destruição da natureza, mas também por uma questão psicológica, ao se apresentar como salvadora eventual de todas as crises (Orr, 2002). Há alguns exemplos claros. Um trabalho recente com os resultados da criação de camundongos transgênicos com uma pequena sequência de DNA obtida de material de museu de uma espécie extinta (o lobo da Tasmânia) foi recebido com entusiasmo exagerado pela mídia. Muitas pessoas viram nesse experimento a esperança da redenção de nossos pecados ambientais. Entretanto, mais do que oferecer uma solução, esse tipo de publicidade pode ter um efeito muito negativo, ao diminuir nossa sensação de responsabilidade para com as espécies que extinguimos (Solé-Cava, 2008b).

É fundamental que as pessoas que trabalham com genética para conservação não fiquem apenas em seus laboratórios cercados de micropipetas e termocicladores: apenas com o contato direto com o problema, realizando trabalhos de campo e colaboração estreita com as pessoas diretamente envolvidas com os problemas de conservação, poderemos realmente fazer alguma diferença. Como disse uma vez o grande divulgador da biologia evolutiva, Stephen J. Gould, "para ganhar a luta de salvar as espécies e os ambientes, é necessário estabelecer uma ligação emocional com a Natureza... Devemos encontrar espaço para a Natureza em nossos corações".

Agradecimentos

Agradecemos a Sergio R. Matioli pela oportunidade de apresentar esse assunto fascinante, aos orientados e amigos do LBDM pelo apoio constante, e aos alunos dos cursos de Genética para Conservação e de Biodiversidade Molecular, por tantas perguntas instigantes. O Laboratório de Biodiversidade Molecular é apoiado por projetos do CNPq, CAPES, FAPERJ, SEAP-DF e FINEP.

Referências Bibliográficas

Acevedo-Whitehouse, K. e Cunningham, A.A. (2006). Is MHC. enough for understanding wildlife immunogenetics? **Trends Ecol. Evol. 21**:433-438.

Acevedo-Whitehouse, K., Spraker, T.R., Lyon,s E., Melin, S.R., Gulland, F., Delong, R.L. e Amos W (2006). Contrasting effects of heterozygosity on survival and hookworm resistance in California sea lion pups. **Mol. Ecol. 15**:1973-1982

Adams, D. e Carwardine, M. (1990) **Last chance to see**. Pan Books, London.

Alderson, G.W., Gibbs, H.L. e Sealy, S.G. (1999) Parentage and kinship studies in an obligate brood parasitic bird, the brown-headed cowbird (Molothrus ater), using microsatellite DNA markers., J. **Hered. 90**:182-190.

Allcock, A.L., Chauvet, M., Crandall, K.A., Given, D.R., Hall, S.J.G., Iriondo, J.M., Lewinsohn, T.M., Lynch, S.M., Mace, G.M., Solé-Cava, A.M., Stackebrandt, E., Templeton, A.R. e Watts, P.C. (1995) Genetic diversity as a component of biodiversity. In: Heywood, V.H. e Watson, R.T. (eds) **Global Biodiversity Assessment**. Cambridge University Press, Cambridge, pp. 57-88.

Allendorf, F.W., Luikart, G. (2006). **Conservation and the Genetics of Populations**. Blackwell Publishing, Oxford.

Amos,, B., Schlotterer, C. e Tautz, D. (1993) Social structure of pilot whales revealed by analytical DNA profiling. **Science 260**:670-672.

Amos, W. e Balmford, A. (2001). When does conservation genetics mat-

ter? **Heredity 87**:257-265.

Andrewartha, H.G. e Birch, L.C. (1954). **The distribution and abundance of animals**. University of Chicago Press, Chicago, EUA.

Ardern, S.L. e Lambert, D.M. (1997). Is the black robin in genetic peril? **Molec. Ecol. 6**:21-28.

Arnedo, M.A., e Ferrández, M.A. (2007) Mitochondrial markers reveal deep population subdivision in the European protected spider Macrothele calpeiana (Walckenaer, 1805) (Araneae, Hexathelidae). **Conserv. Genet. 8**:1147–1162.

Asa, C., Miller, P., Agnew, M., Rebolledo, J.A.R., Lindsey, S.L., Callahan, M. e Bauman, K. (2007) Relationship of inbreeding with sperm quality and reproductive success in Mexican gray wolves. **Anim. Conserv. 10**:326-331.

Avise, J.C. (1974). Systematic value of electrophoretic data. **Syst. Zool. 23**:465-481.

Avise, J.C. (1996). Three fundamental contributions of molecular genetics to avian ecology and evolution. **Ibis** 138: 16-25.

Avise, J.C. (1998). Conservation genetics in the marine realm., J. **Hered.89**:377-382.

Avise, J.C. (2004). **Molecular markers, natural history and evolution**. Chapman & Hall., London.

Avise, J.C. e Hamrick, J.L. (1996). **Conservation genetics : case histories from nature**. Chapman & Hall, New York.

Avise, J.C. e Johns, G.C. (1999). Proposal for a standardized temporal scheme of biological classification for extant species. **Proc. Natl. Acad. Sci. USA 96**:7358-7363.

Avise, J.C. e Nelson W.S. (1989). Molecular genetic relationships of the extinct Dusky Seaside Sparrow. **Science 243**:646-648.

Baker, C.S., Cipriano F. e Palumbi, S.R. (1996). Molecular genetic identification of whale and dolphin products from commercial markets in Korea and Japan. **Mol. Ecol. 5**:671-685.

Baker, C.S., Lento, G.M., Cipriano F., Dalebout, M.L. e Palumbi, S.R. (2000a) Scientific whaling: Source of illegal products for market? **Science 290**:1695-1695.

Baker, C.S., Lento, G.M., Cipriano F. e Palumbi, S.R. (2000b) Predicted decline of protected whales based on molecular genetic monitoring of Japanese and Korean markets. **Proc. Royal Soc. of Lond. B: 267**:1191-1199.

Baker, C.S. e Palumbi, S.R. (1994). Which whales are hunted - a molecular genetic approach to monitoring whaling. **Science 265**:1538-1539.

Barbará, T., Lexer, C., Martinelli, G., Mayo, S., Fay M,F. e Heuertz M (2008). Within-population spatial genetic structure in four naturally fragmented species of a neotropical inselberg radiation, Alcantarea imperialis, A. geniculata, A. glaziouana and, A. regina (Bromeliaceae). **Heredity 101**:285-296.

Barluenga, M. e Meyer, A. (2004). The Midas cichlid species complex: incipient sympatric speciation in Nicaraguan cichlid fishes? **Mol. Ecol. 13**: 2061-2076.

Bass, A.L., Good, D.A., Bjorndal, K.A., Richardson, J.I., Hillis Z.M., Horrocks, J.A. e Bowen, B.W. (1996). Testing models of female reproductive migratory behaviour and population structure in the Caribbean hawksbill turtle, Eretmochelys imbricata, with mtDNA sequences. **Mol. Ecol. 5**:321-328.

Beebee, T.J.C. (2007). Population structure and its implications for conservation of the great silver beetle Hydrophilus piceus in Britain. **Freshwater Biol. 52**:2101-2111

Berg, D.J, Guttman, S.I. e Cantonwine E.G. (1996). Geographic variation in unionid genetic structure: Do management units exist?, J. **Shellfish Res 15**:484.

Bickford, D., Lohman, D.J, Sodhi N.S., Ng, P.K.L., Meier, R. e Winker, K., Ingram, K.K. e Das I. (2007). Cryptic species as a window on diversity and conservation. **Trends Ecol. Evol. 22**:148-155.

Bomford, M. e O'Brien P. (1995). Erradication or control for vertebrate pests. **Wildl. Soc. Bull. 23**:249-255.

Bonnell, M.L. e Selander, R.K. (1974). Elephant seals: genetic variation and near extinction. **Science 184**:908-909.

Borlase, S.C., Loebel, D.A., Frankham, R., Nurthen, R.K., Briscoe, D.A. e Daggard, G.E. (1993). Modeling Problems in Conservation Genetics Using Captive Drosophila Populations - Consequences of Equalization of Family Sizes. **Conservat. Biol. 7**:122-131.

Boskovic, R., Kovacs, K.M., Hammill, M.O. e White, B.N. (1996). Geographic distribution of mitochondrial-Dna haplotypes in grey seals (Halichoerus grypus). **Can., J. Zool. 74**:1787-1796.

Bowen, B.W., Bass, A.L., Garcia-Rodriguez, A., Diez, C.E., Vandam, R., Bolten, A., Bjorndal, K.A., Miyamoto, M.M. e Ferl, R.J. (1996). Origin of hawksbill turtles in a caribbean feeding area as indicated by genetic-markers. **Ecol. Appl. 6**:566-572.

Bowen, B.W. e Karl, S.A. (2007). Population genetics and phylogeography of sea turtles. **Mol. Ecol. 16**:4886–4907.

Bowen, B.W., Meylan, A.B. e Avise, J.C. (1991). Evolutionary distinctiveness of the endangered Kemps Ridley sea turtle. **Nature 352**:709-711.

Brazaitis P., Rebêlo, J.H., Yamashita, C., Odierna E.A. e Watanabe, M.E. (1996). Threats to Brazilian crocodilian populations. **Oryx 30**:275-284.

Brazaitis P., Watanabe, M.E. e Amato, G. (1998). The caiman trade. **Sci. Am. 278**:52-58.

Bremer, J.R.A., Viñas, J., Mejuto, J., Ely, B. e Pla, C. (2005). Comparative phylogeography of Atlantic bluefin tuna and swordfish: the combined effects of vicariance, secondary contact, introgression, and population expansion on the regional phylogenies of two highly migratory pelagic fishes. **Mol. Phylogenet. Evol. 36**:169–187.

Brown, D.M., Brenneman, R.A., Koepfli, K.-P., Pollinger, J.P., Milá, B., Georgiadis N.J., Louis Jr E.E., Grether, G.F., Jacobs, D.K. e Wayne, R.K. (2007). Extensive population genetic structure in the giraffe. **BMC. Biology 5**: 57.

Burns E.L., Eldridge, M.D.B., Crayn, D.M. e Houlden, B.A. (2007). Low phylogeographic structure in a wide spread endangered Australian frog *Litoria aurea* (Anura: Hylidae). **Conserv. Genet. 8**:17–32.

Cardoso, M.A., Provan, J., Powell, W., Ferreira P.C.G. e De Oliveira, D.E. (1998). High genetic differentiation among remnant populations of the endangered *Caesalpinia echinata* Lam. (Leguminosae-Caesalpinioideae). **Mol. Ecol. 7**:601-608.

Carlton, J.T. (1985). Transoceanic and interoceanic dispersal of coastal marine organisms: the biology of ballast water. **Oceanogr. Mar. Biol. Annu. Rev. 23**:313-371.

Carlton, J.T. e Geller, J.B. (1993). Ecological roulette: the global transport of indigenous organisms. **Science 261**:78-82.

Carlton, J.T. e Hodder, J. (1995). Biogeography and dispersal of coastal marine organisms: experimental studies on a replica of a 16th-century sailing vessel. **Mar. Biol.** 121:721-730.

Carnaval, A.C. e Bates, J.M. (2007). Amphibian DNA shows marked genetic structure and tracks Pleistocene climate change in Northeastern Brazil. **Evolution 61**:2942–2957.

Carvalho, G.R. e Hauser L. (1995). Molecular genetics and the stock concept in fisheries. **Molecular Genetics in Fisheries**. Chapman & Hall, London. p. 55–79.

Case, M.A., Mlodozeniec H.T., Wallace L.E. e Weldy T.W. (1998). Conservation genetics and taxonomic status of the rare Kentucky lady's slipper: Cypripedium kentuckiense (Orchidaceae). **Am. J Bot. 85**:1779-1786.

Castilla, A.M., Fernandez-Pedrosa V., Backeljau T., Gonzalez, A., Latorre, A. e Moya, A. (1998). Conservation genetics of insular Podarcis lizards using partial cytochrome b sequences. **Mol. Ecol. 7**:1407-1411.

Castro, A.L.F., Stewart, B.S., Wilson, S.G., Hueter, R.E., Meekan, M.G., Motta P.J., Bowen, B.W. e Karl, S.A. (2007). Population genetic structure of Earth's largest fish, the whale shark (Rhincodon typus). **Mol. Ecol. 16**:5183–5192.

Clavero, M. e Garcia-Berthou E. (2005). Invasive species are a leading cause of animal extinctions. **Trends Ecol. Evol. 20**:110.

Cohen, A.N. e Carlton, J.T. (1998). Accelerating invasion rate in a highly invaded estuary. **Science 279**:555-558.

Cohen, A.N., Carlton, J.T. e Fountain, M.C. (1995). Introduction, dispersal and potential impacts of the green crab *Carcinus maenas* in San-Francisco Bay, California. **Mar. Biol. 122**:225-237.

Cohen, J.E. (2003). Human Population: The Next Half Century. **Science 302**:1172-1175.

Colegrove, K.M., Lowenstine L.J. e Gulland F.M.D. (2005). Leptospirosis in northern elephant seals (*Mirounga angustirostris*) stranded along the California coast., J. **Wildl. Dis. 41**:426-430.

Costantini F., Fauvelot, C. e Abbiati, M. (2007). Genetic structuring of the temperate gorgonian coral (*Corallium rubrum*) across the western Mediterranean Sea revealed by microsatellites and nuclear sequences. **Mol. Ecol. 16**:5168–5182.

Cowled, B.D., Aldenhoven, J., Odeh I.O.A., Garrett T., Moran, C. e Lapidge, S.J. (2008). Feral pig population structuring in the rangelands of eastern Australia: applications for designing adaptive management units. **Conserv Genet 9**:211–224.

Crochet P.A. (1996). Can measures of gene flow help to evaluate bird dispersal? Acta Oecologica-International, J. **Ecol. 17**:459.

Cunha H.A., da Silva V.M.F., Lailson-Brito, J., Santos, M.C.O., Flores P.A.C., Martin, A.R., Azevedo, A.F., Fragoso, A.B.L., Zanelatto, R.C. e Solé-Cava, A.M. (2005). Riverine and marine ecotypes of *Sotalia*

dolphins are different species. **Mar. Biol. 148**:449-457.
Cunha H.A. e Solé-Cava, A.M. (2007). Molecular sexing of tucuxi dolphins (*Sotalia guianensis* and *Sotalia fluviatilis*) using samples from biopsy darting and decomposed carcasses. **Genet. Mol. Biol. 30**:1186-1188.
Daskalakis, K.D., O'Connor T.P. e Crecelius E.A. (1997). Evaluation of digestion procedures for determining silver in mussels and oysters. **Environ. Sci. Technol. 31**:2303.
Davies N., Villablanca F.X. e Roderick, G.K. (1999). Bioinvasions of the medfly *Ceratitis capitata*: Source estimation using DNA sequences at multiple intron loci. **Genetics 153**:351-360.
de Oliveira L.R., Arias-Schreiber, M., Meyer, D. e Morgante, J.S. (2006). Effective population size in a bottlenecked fur seal population. **Biol. Conserv. 131**:505-509.
De Salle, R. e Birstein V.J. (1996). PCR identification of black caviar. **Nature 381**:187-188.
Dowling T.E., Minckley, W.L., Marsh P.C. e Goldstein E.S. (1996). Mitochondrial DNA variability in the endangered razorback sucker (*Xyrauchen texanus*) - analysis of hatchery stocks and implications for captive propagation. **Conservat. Biol. 10**:120-127.
Duffy, M.A., Perry L.J., Kearns, C.M. e Weider L.J. (2000). Paleogenetic evidence for a past invasion of Onondaga Lake, New York, by exotic Daphnia curvirostris using mtDNA from dormant eggs. **Limnol. Oceanogr. 45**:1409-1414.
Dupanloup I., Schneider, S. e Excoffier L. (2002)., A. simulated annealing approach to define the genetic structure of populations. **Mol. Ecol. 11**:2571-2581.
Ellegren H., Hartman, G., Johansson, M. e Andersson L. (1993). Major Histocompatibility Complex Monomorphism and Low-Levels of DNA-Fingerprinting Variability in a Reintroduced and Rapidly Expanding Population of Beavers. **Proc. Natl. Acad. Sci. USA 90**:8150-8153.
Epifanio, J.M., Koppelman, J.B., Nedbal, M.A. e Philipp, D.P. (1996). Geographic variation of paddlefish allozymes and mitochondrial DNA. **Trans. Am. Fish. Soc. 125**:546-561.
Excoffier L. e Smouse P.E. (1994). Using allele frequencies and geographic subdivision to reconstruct gene tree within a species: molecular variance parsimony. **Genetics 136**:343-359.
Fabbri E., Miquel, C., Lucchini V., Santini, A., Caniglia, R., Duchamp, C., Weber, J.M., Lequette, B., Marucco F., Boitani L., Fumagalli L., Taberlet P. e Randi E. (2007). From the Apennines to the Alps: colonization genetics of the naturally expanding Italian wolf (Canis lupus) population. **Mol. Ecol. 16**:1661-1671.
Fabiani, A., Hoelzel, A.R., Galimberti F. e Muelbert, M.M,C., (2003). Long-range paternal gene flow in the southern elephant seal. **Science 299**:676-676.
Fleming I.A., Hindar, K., Mjolnerod I.B., Jonsson, B., Balstad T. e Lamberg, A. (2000). Lifetime success and interactions of farm salmon invading a native population. **Proc., R. Soc. Lond. 267B**:1517-1523.
Fowler E.V., Hoeben P. e Timms P. (1998). Randomly amplified polymorphic DNA variation in populations of eastern Australian koalas, *Phascolarctos cinereus*. **Biochem. Genet. 36**:381-393.
Francisco, M.R., Gibbs, H.L., Galetti, M., Lunardi, V.O. e Galetti, P.M. (2007). Genetic structure in a tropical lek-breeding bird, the blue manakin (*Chiroxiphia caudata*) in the Brazilian Atlantic Forest. **Mol. Ecol. 16**:4908–4918.
Frankel, O.H. e Soulé, M. (1981). **Conservation and Evolution**. Cambridge University Press, Cambridge.
Frankham, R., Ballou J.D. e Briscoe, D.A. (2002). **Introduction to Conservation Genetics**. Cambridge University Press, Cambridge.
Frankham, R., Ballou, J.D., Briscoe, D.A. e McInnes, K.H. (2004). **A Primer of Conservation Genetics**. Cambridge University Press, Cambridge.
Franklin, I.R. e Frankham, R. (1998). How large must populations be to retain evolutionary potential? **Anim. Conserv. 1**:69-70.
Frantz, A,C., Pourtois, J,T., Heuertz, M., Schley, L., Flamand, M,C., Krier, A., Bertouille, S., Chaumont, F. e Burke T. (2006). Genetic structure and assignment tests demonstrate illegal translocation of red deer (*Cervus elaphus*) into a continuous population. **Mol. Ecol. 15**:3191-3203.
Galetti Jr, P.M., Rodrigues, F.P., Solé-Cava, A.M., Miyaki, CY, Carvalho, D., Eizirik, E., Veasey, E.A., Santos, F.R., Farias, I.P., Vianna, J.A., Oliveira, L.R., Weber, L.I., Almeida-Toledo, L.F., Francisco, M.R., Redondo, R.A.F., Siciliano, S., Del Lama, S.N., Freitas, T.R.O., Hrbek, T. e Molina, W.F. (2008). Genética da conservação brasileira. In:Frankham, R., Ballou, J.D. e Briscoe, D.A. (eds) **Fundamentos de Genética da Conservação**. Editora da Sociedade Brasileira de Genética, Ribeirão Preto, SP., pp 244-274.
Garcia-Rodriguez, A.I., Bowen, B.W., Domning, D., Mignucci-Giannoni, A.A., Marmontel, M., Montoya-Ospina, R.A., Morales-Vela, B., Rudin, M., Bonde R.K. e McGuire P.M. (1998). Phylogeography of the West Indian manatee (*Trichechus manatus*): how many populations and how many taxa? **Mol. Ecol. 7**:1137-1149.
Geller, J.B., Carlton, J.T. e Powers, D.A. (1994). PCR based detection of mtDNA haplotypes of native and invading mussels on the Northeastern Pacific Coast - Latitudinal pattern of invasion. **Mar. Biol. 119**:243-249.
Gibbs, H.L., Prior, K.A. e Weatherhead, P.J. (1994). Genetic analysis of populations of threatened snake species using RAPD. markers. **Mol. Ecol. 3**:329-337.
Gibbs, H.L., Prior, K.A., Weatherhead, P.J. e Johnson, G. (1997). Genetic structure of populations of the threatened eastern massasauga rattlesnake, *Sistrurus c catenatus*: evidence from microsatellite DNA markers. **Mol. Ecol. 6**:1123-1132.
Gilpin M.E. e Soulé, M.E. (1986). Minimum viable populations: proceses of species extinction. In: Soulé, M.E. (ed) **Conservation Biology: the science of scarcity and diversity**. Sinauer Associates, Sunderland, Massachusetts, pp 19-34.
Gonzalez-Astorga, J. e Castillo-Campos, G. (2004). Genetic variability of the narrow endemic tree *Antirhea aromatica* Castillo-Campos & Lorence, (Rubiaceae, Guettardeae) in a tropical forest of Mexico. **Ann. Bot. 93**:521-528.
Gonzalez, S., Maldonado, J.E., Leonard, J.A., Vila, C., Duarte, J.M.B., Merino, M., Brum-Zorrilla, N. e Wayne, R.K. (1998a) Conservation genetics of the endangered Pampas deer (*Ozotoceros bezoarticus*). **Mol. Ecol. 7**:47-56.
Gonzalez, S., Maldonado, J.E., Leonard, J.A., Vila, C., Duarte, J.M.B., Merino, M., Brum-Zorrilla, N. e Wayne, R.K. (1998b) Conservation genetics of the endangered Pampas deer (*Ozotoceros bezoarticus*). **Mol. Ecol. 7**:47.
Grandjean, F., SoutyGrosset, C., Raimond, R. e Holdich, D.M. (1997). Geographical variation of mitochondrial DNA between populations of the white-clawed crayfish *Austropotamobius pallipes*. **Freshwater Biol. 37**:493-501
Grant, W.S. e Bowen, B.W (1998). Shallow population histories in deep evolutionary lineages of marine fishes:insights from sardines and anchovies and lessons for conservation., J. **Hered. 89**:415-426.
Gravena, W., Hrbek, T., da Silva, V.M.F. e Farias, I.P. (2008). Amazon River dolphin love fetishes: from folklore to molecular forensics. **Mar. Mam. Sci. 24**:969–978.
Griffith, S.C., Owens, I.P.F. e Thuman, K.A. (2002). Extra pair paternity in birds: a review of interspecific variation and adaptive function. **Mol. Ecol. 11**:2195-2212.
Griffiths, R., Double, M.C., Orr, K. e Dawson, R.J.G. (1998)., A. DNA test to sex most birds. **Mol. Ecol. 7**:1071.
Gruenthal, K.M. e Burton, R.S. (2008). Genetic structure of natural populations of the California black abalone (*Haliotis cracherodii* Leach, 1814), a candidate for endangered species status., J. **Exp. Mar. Biol. Ecol. 355**:47–58.
Gusmão, J., Lazoski, C., Monteiro, F.A. e Solé-Cava, A.M. (2006). Cryptic species and population structuring of the Atlantic and Pacific seabob shrimp species, *Xiphopenaeus kroyeri* and *Xiphopenaeus riveti*. **Mar. Biol.149**:491-502.
Gusmão J., Lazoski, C. e Solé-Cava, A.M. (2000)., A new species of *Penaeus* (Crustacea : Penaeidae) revealed by allozyme and cytochrome oxidase I analyses. **Mar. Biol. 137**:435-446.
Haberl, H., Erb, K.-H., Krausmann, F., Gaube, V., Bondeau, A., Plutzar, C., Gingrich, S., Lucht, W. e Fischer-Kowalski, M. (2007). Quantifying and mapping the human appropriation of net primary production in earth's terrestrial ecosystems. **Proc. Natl. Acad. Sci. USA 104**:12942-12947.
Haig, S.M. (1998). Molecular contributions to conservation. **Ecology 79**:413-425.
Haig, S.M., Belthoff, J.R. e Allen, D.H. (1993). Examination of population structure in red-cockaded woodpeckers using DNA profiles. **Evolution 47**(1):185-194.
Haig, S.M., Bowman, R. e Mullins, T.D. (1996). Population structure of red-cockaded woodpeckers in south Florida: RAPDs revisited. **Mol. Ecol. 5**:725-734.
Hall, H.G. (1992). Suspected african honeybee colonies in Florida tested for identifying DNA markers. **Fla. Entomol. 75**:257-266.
Hamer, M. (1991). Toxic hithhikers conquer the world's oceans. **New Sci 129**:23.
Hampton, J.O., Spencer, P.B.S., Alpers, D.L., Twigg, L.E., Woolnough,

A.P., Doust, J., Higgs, T. e Pluske, J. (2004). Molecular techniques, wildlife management and the importance of genetic population structure and dispersal: a case study with feral pigs., J. **App. Ecol. 41**:735–743.

Hedrick, P.W. e Parker, K.M. (1998). MHC variation in the endangered Gila topminnow. **Evolution 52**:194-199.

Helgen, K.M., Maldonado, J.E., Wilson, D.E. e Buckner, SD. (2008). Molecular confirmation of the origin and invasive status of west indian raccoons., J. **Mammal. 89**:282–291.

Heywood, V.H. e Watson, R.T. (1996). **Global biodiversity assessment**. Cambridge University Press, New York

Hoelzel, A.R. (1992). **Molecular genetic analysis of populations**. IRL Press, Oxford

Hogg, I.D., Larose, C., de Lafontaine, Y. e Doe K.G. (1998). Genetic evidence for a *Hyalella* species complex within the Great Lakes St Lawrence River drainage basin: implications for ecotoxicology and conservation biology. **Can., J. Zool. 76**:1134-1140.

Holland, B.S. (2000). Genetics of marine bioinvasions. **Hydrobiologia 420**:63-71.

Hoole, J.C., Joyce, D.A. e Pullin, A.S. (1999). Estimates of gene flow between populations of the swallowtail butterfly, *Papilio machaon* in Broadland, UK and implications for conservation. **Biol. Conserv. 89**:293-299.

Hughes, A. e Nei, M. (1988). Pattern of nucleotide substitution ar major histocompatibility complex class I loci reveals overdominant selection. **Nature 335**:167-170.

Hughes, A. e Nei, M. (1990). Evolutionary relationships of class II major histocompatibility-complex genes in mammals. **Mol. Biol. Evol. 7**:491-514.

Hughes, A. e Nei, M. (1992a) Models of host-parasite interaction and MHC. polymorphism. **Genetics 132**:863-864.

Hughes, A. e Nei, M. (1992b) Maintenance of MHC Polymorphism. **Nature 355**:402-403.

Hughes, C. (1998). Integrating molecular techniques with field methods in studies of social behavior: A revolution results. **Ecology 79**:383.

Ignacio, B.L., Absher, T.M., Lazoski, C. e Solé-Cava, A.M. (2000). Genetic evidence of the presence of two species of Crassostrea (Bivalvia : Ostreidae) on the coast of Brazil. **Mar. Biol.136**:987-991.

Jian, S., Zhong, Y., Liu, N., Gao, Z., Wei, Q., Xie, Z,. e Ren, H. (2006). Genetic variation in the endangered endemic species *Cycas fairylakea* (Cycadaceae) in China and implications for conservation. **Biodiversity Conserv. 15**:1681–1694.

Klein, J., O'hUigin, C., Figueroa, F., Mayer, W.E. e Klein, D. (1993). Different modes of *Mhc* evolution in primates. **Mol. Biol. Evol. 10**:48-59.

Knowlton, N. (1993). Sibling species in the sea. **Ann. Rev. Ecol. Syst. 24**:189-216.

Knowlton, N. (2000). Molecular genetic analyses of species boundaries in the sea. **Hydrobiologia 420**:73-90.

Kretzmann, M.B., Gilmartin, W.G., Meyer, A., Zegers, G.P., Fain, S.R., Taylor, B.F. e Costa, D.P. (1997). Low genetic variability in the Hawaiian monk seal. **Conservat. Biol. 11**:482-490.

Kuris, A.M. (1991). A review of patterns and causes of crustacean brood mortality. **Crust. Issues 7**:117-141.

Lafferty, K.D., Smith, K.F., Torchin, M.E., Dobson, A.P. e Kuris, A.M. (2005). The role of infectious disease in natural communities: what introduced species tell us. In: Sax, D.F., Stachowicz, J.J. e Gaines, S.D. (eds) **Species Invasions: Insights into Ecology, Evolution, and Biogeography**. Sinauer, Sunderland, Mass.

Lafferty, K.D. e Kuris, A.M. (1996). Biological control of marine pests. **Ecology 77**:1989-2000,

Lancaster, M.L., Bradshaw, C.J.A., Goldsworthy, S.D. e Sunnucks, P. (2007). Lower reproductive success in hybrid fur seal males indicates fitness costs to hybridization. **Mol. Ecol. 16**:3187–3197.

Lande, R. (1988). Genetics and demography in biological conservation. **Science 241**:1455-1460

Lara-Ruiz, P., Chiarello, A.G. e Santos, F.R. (2008). Extreme population divergence and conservation implications for the rare endangered Atlantic Forest sloth, *Bradypus torquatus* (Pilosa: Bradypodidae). **Biol. Conserv. 141**:1332 –1342.

Latta, R.G. (2008). Conservation genetics as applied evolution: from genetic pattern to evolutionary process. **Evol. Appl. 1**:84–94.

Lazaro, M., Lessa, E.P. e Hamilton, H. (2004). Geographic genetic structure in the franciscana dolphin (*Pontoporia blainvillei*). **Mar. Mamm. Sci. 20**:201-214

Lessells, C.M. e Mateman, A.C. (1998). Sexing birds using random amplified polymorphic DNA (RAPD) markers. **Mol. Ecol. 7**:187-195.

Lewontin, R.C. (1985). Population genetics. **Annu. Rev. Genet. 19**:81-102.

Liu, H.P., Mitton, J.B. e Herrmann, S.J. (1996). Genetic differentiation in and management recommendations for the freshwater mussel, *Pyganodon grandis* (Say, 1829). **Am. Malacol. Bull. 13**:1-2.

Lizarralde, M.S., Bailliet, G., Poljak, S., Fasanella, M. e Giulivi, C. (2008). Assessing genetic variation and population structure of invasive North American beaver (*Castor canadensis* Kuhl, 1820) in Tierra Del Fuego (Argentina). **Biol. Invasions 10**:673-683.

Lobo, J.A., Del Lama, M.A. e Mestriner, M.A. (1989). Population differentiation and racial admixture in the africanized honeybee (*Apis mellifera* L.). **Evolution 43**:794-802.

Lortscher, M., Claluna, M. e Scholl, A. (1998). Genetic population structure of *Austropotamobius pallipes* (Lereboullet 1858) (Decapoda : Astacidae) in Switzerland, based on allozyme data. **Aquat. Sci. 60**:118-129.

Lukoschek, V., Waycott, M. e Marsh, H. (2007). Phylogeography of the olive sea snake, *Aipysurus laevis* (Hydrophiinae) indicates Pleistocene range expansion around northern Australia but low contemporary gene flow. **Mol. Ecol. 16**:3406–3422.

Lynch, M., Conery, J. e Bürger, R. (1995). Mutation accumulation and the extinction of small populations. **Am. Nat. 146**:489-518.

McDonald, D.B., Parchman ,T.L., Bower M.R., Hubert W.A. e Rahel F.J. (2008). An introduced and a native vertebrate hybridize to form a genetic bridge to a second native species. **Proc. Natl. Acad. Sci. USA 105**:10837–10842.

McMillan W.O. e Bermingham E. (1996). The phylogeographic pattern of mitochondrial DNA variation in the Dall's porpoise *Phocoenoides dalli*. **Mol. Ecol. 5**:47-61.

Magnussen, J.E., Pikitch, E.K., Clarke, SC., Nicholson, C., Hoelzel, A.R. e Shivji, M.S. (2007). Genetic tracking of basking shark products in international trade. **Anim. Conserv. 10**:199-207.

Manchester, S.J. e Bullock, J.M (2000). The impacts of non-native species on UK biodiversity and the effectiveness of control., J. **Appl. Ecol. 37**:845-864.

Matsuhashi, T., Masuda, R., Mano, T. e Yoshida, M.C. (1999). Microevolution of the mitochondrial DNA control region in the Japanese brown bear (*Ursus arctos*) population. **Mol. Biol. Evol. 16**:676-684.

Mattos-Machado, T.M., Solé-Cava, A.M., David, J.R. e Bitner-Mathé, B.C. (2005). Allozyme variability in an invasive drosophilid, *Zaprionus indianus* (Diptera: Drosophilidae): comparison of a recently introduced Brazilian population with Old World populations. **Ann. Soc. Entomol. France 41**:7-13.

May, B. e Marsden, J.E. (1992). Genetic identification and implications of another invasive species of dreissenid mussel in the Great Lakes. **Can., J. Fish. Aquat. Sci. 49**:1501-1506.

Mendez, M., Rosenbaum, H.C. e Bordino, P (2008). Conservation genetics of the franciscana dolphin in Northern Argentina: population structure, by-catch impacts, and management implications. **Conserv. Genet. 9**:419–435.

Menotti-Raymond, M.O.B. e Stephen, J. (1993). Dating the genetic bottleneck of the African cheetah. **Proc. Natl. Acad. Sci. USA 90**:3172-3176.

Millar, C.D., Reed, C.E.M., Halverson, J.L. e Lambert, D.M. (1997). Captive management and molecular sexing of endangered avian species: An application to the black stilt *Himantopus novaezelandiae* and hybrids. Biol. **Conserv. 82**:81-86.

Miller, J.R. (2005). Biodiversity conservation and the extinction of experience. **Trends Ecol. Evol. 20**:430-434.

Mills, E.L., Leach, J.H., Carlton, J.T. e Secor, C.L. (1994). Exotic species and the integrity of the Great Lakes. **BioScience 44**:666-676.

Mills, L.S. (2006). **Conservation of Wildlife Populations: Demography, Genetics and Management**. Blackwell Publishing, Oxford

Miyaki, C.Y, Duarte, J.M.B., Caparroz, R., Nunes, A.L.V. e Wajntal, A. (1997). Sex identification of South American Parrots (Psittacidae, Aves) using the human minisatellite probe 33.15. **Auk 114**:516-520.

Miyaki, C.Y., Griffiths, R., Orr, K., Nahum, L.A., Pereira, S.L. e Wajntal, A. (1998). Sex identification of parrots, toucans, and curassows by PCR: Perspectives for wild and captive population studies. **Zoo Biology 17**:415.

Monteiro, E.L.D., Monteiro, L.R. e dos Reis, S.F. (2002). Skull shape and size divergence in dolphins of the genus *Sotalia*:, A. tridimensional morphometric analysis., J. **Mammal 83**:125-134.

Moore, M.K., Bemiss, J.A., Rice, S.M., Quattro, J.M. e Woodley, C.M. (2003). Use of restriction fragment length polymorphisms to identify sea turtle eggs and cooked meats to species. **Conserv. Genet. 4**:95-103.

Morgan, M.J, Hunter, D., Pietsch, R., Osborne, W. e Keogh J.S. (2008).

Assessment of genetic diversity in the critically endangered Australian corroboree frogs, *Pseudophryne corroboree* and *Pseudophryne pengilleyi*, identifies four evolutionarily significant units for conservation. **Mol. Ecol. 17**:3448–3463.

Moritz, C. (1994). Applications of mitochondrial-DNA analysis in conservation - a critical review. **Mol. Ecol. 3**:401-411.

Mundy, N.I., Winchell, C.S. e Woodruff, D.S. (1997). Genetic differences between the endangered San Clemente Island loggerhead shrike *Lanius ludovicianus mearnsi* and two neighbouring subspecies demonstrated by mtDNA control region and cytochrome b sequence variation. **Mol. Ecol. 6**:29-37.

Munguia-Vega, A., Esquer-Garrigos, Y., Rojas-Bracho L., Vazquez-Juarez, R., Castro-Prieto, A. e Flores-Ramirez, S. (2008). Genetic drift vs. natural selection in a long-term small isolated population: major histocompatibility complex class II variation in the Gulf of California endemic porpoise (*Phocoena sinus*). **Mol. Ecol. 16**:4051–4065.

Nei, M. (1987). **Molecular evolutionary genetics**. Columbia University Press, New York.

Neigel, J.E. (1997)., A. comparison of alternative strategies for estimating gene flow from genetic markers. **Annu. Rev. Ecol. Syst. 28**:105-128.

Neigel, J.E. (2002). Is FST obsolete? **Conserv. Genet. 3**:167-173.

Ngamsiri, T., Nakajima, M., Sukmanomon, S., Sukumasavin, N., Kamonrat, W., Na-Nakorn, U. e Niguchi, N. (2007). Genetic diversity of wild Mekong giant catfish *Pangasianodon gigas* collected from Thailand and Cambodia. **Fish. Sci. 73**:792–799.

Nishi, J.S., Stephen, C. e Elkin, B.T. (2002). Implications of agricultural and wildlife policy on management and eradication of bovine tuberculosis and brucellosis in free-ranging wood bison of northern Canada. **Ann. N. Y. Acad. Sci. 969**:236–244.

Norse, E.A. e McManus, R.E. (1980). Ecology and living resources biological diversity. In: Council of Environmental Quality (eds) **Environmental Quality: The eleventh annual report of the Council on Environmental Quality**. Council on Environmental Quality, Washington, DC.

Nusser, J.A., Goto, R.M., Ledig, D.B., Fleischer, R.C. e Miller, M.M. (1996). RAPD. analysis reveals low genetic variability in the endangered light-footed clapper rail. **Mol. Ecol. 5**:463-472.

O'Brien, S.J. (1994)., A. role for molecular genetics in biological conservation. **Proc. Natl. Acad. Sci. 91**:5748-5755.

O'Brien, S.J., Roelke, M.E., Marker, L., Newman, A., Winkler, C.A., Meltzer, D., Colly, L., Evermann, J.F., Bush, M. e Wildt, D.E. (1985). Genetic basis for species vulnerability in the cheetah. **Science 227**:1428-1434.

O'Brien, S.J., Wildt, D.E., Goldman, D., Merril, C.R. e Bush, M. (1983). The cheetah is depauperate in genetic variation. **Science 221**:459-462.

Orr, D.W. (1992). For the Love of Life. **Conservat. Biol. 6**:486-487.

Orr, D.W. (2002). Four challenges of Sustainability. **Conservat. Biol. 16**:1457-1460.

Packer, C., Gilbert, D.A., Pusey, A.E. e O'Brien ,S.J. (1991). Kinship, co-operation and inbreeding in African lions: a molecular genetic analysis. **Nature 351**:562-565.

Palsboll, P.J., Berube, M. e Allendorf, F.W (2007). Identification of management units using population genetic data. **Tends Ecol. Evol. 22**:11-16.

Palumbi, S. e Cipriano, F. (1999). Harpoons fly in whale wars - reply. **Nature 398**:366-366.

Palumbi, S.R. e Cipriano, F. (1998). Species identification using genetic tools: The value of nuclear and mitochondrial gene sequences in whale conservation., J. **Hered. 89**:459-464.

Pappert, R.A., Hamrick, J.L. e Donovan, L.A. (2000). Genetic variation in *Pueraria lobata* (Fabaceae), an introduced, clonal, invasive plant of the southeastern United States. **Am., J. Bot 87**:1240-1245.

Paquette, S.R., Behncke, S.M., O'Brien, S.H., Brenneman, R.A., Louis Jr., E.E. e Lapointe, F.J. (2007). Riverbeds demarcate distinct conservation units of the radiated tortoise (*Geochelone radiata*) in southern Madagascar. **Conserv. Genet. 8**:797–807.

Parker, P.G., Snow, A.A., Schug, M.D., Booton, G.C. e Fuerst, P.A. (1998). What molecules can tell us about populations: Choosing and using a molecular marker. **Ecology 79**:361-382.

Perdices, A., Machordom, A. e Doadrio, I. (1996). Allozymic variation and relationships of the endangered cyprinodontid genus *Valencia* and its implications for conservation., J. **Fish Biol. 49**:1112-1127.

Pereira, S.L. e Wajntal, A. (1999). Reintroduction of guans of the genus *Penelope* (Cracidae, Aves) in reforested areas in Brazil: assessment by DNA fingerprinting. **Biol. Conserv. 87**:31-38.

Petrie, M.K.B. (1998). Extra-pair paternity in birds: explaining variation between species and populations. **Trends Ecol. Evol. 13**:52-58.

Pfenninger, M. e Schwenk, K. (2007). Cryptic animal species are homogeneously distributed among taxa and biogeographical regions. **BMC. Evol. Biol. 7**:121.

Pichler, F.B. e Baker, C.S. (2000). Loss of genetic diversity in the endemic Hector's dolphin due to fisheries-related mortality. **Proc. Royal Soc. London B267**:97-102.

Pierce, R.W., Carlton J.T., Carlton, D.A. e Geller J.B. (1997). Ballast water as a vector for tintinnid transport. **Mar. Ecol. Prog. Ser. 149**:295-297.

Piertney, S.B., MacColl, A.D.C., Bacon, P.J. e Dallas, J.F. (1998). Local genetic structure in red grouse (*Lagopus lagopus scoticus*): evidence from microsatellite DNA markers. **Mol. Ecol. 7**:1645-1654.

Quesada, H., Beynon, C.M. e Skibinski, D.O.F. (1995). Mitochondrial DNA discontinuity in the mussel *Mytilus galloprovincialis* Lmk: pleistocene vicariance biogeography and secondary intergradation. **Mol. Biol. Evol. 12**:521-524.

Quintero, J., Sotelo C.G., Rehbein, H., Pryde, S.E., Medina, I., Perés-Martin, R.I., Rey-Méndez, M. e Mackie, I.M. (1998). Use of mtDNA direct polymerase chain reaction (PCR) sequencing and PCR-restriction fragment length polymorphism methodologies in species identification of canned tuna., J. **Agric. Food. Chem. 46**:1662-1669.

Ram, J.L., Ram, M.L. e Baidoun, F.F. (1996). Authentication of canned tuna and bonito by sequence and restriction site analysis of polymerase chain reaction products of mitochondrial DNA., J. **Agric. Food Chem. 44**:2460-2467.

Randi, E. (1993). Effects of fragmentation and isolation on genetic variability of the italian populations of wolf *Canis lupus* and brown bear *Ursus arctos*. **Acta Theriol. 38**:113-120.

Randi, E., Francisci F. e Lucchini V. (1995) Mitochondrial DNA restriction fragment length monomorphism in the italian wolf (*Canis lupus*) population., J. **Zool. Syst. Evol. Res. 33**:97-100.

Rehbein, H., Kress, G. e Schmidt, T. (1997). Application of PCR-SSCP to species identification of fishery products., J. **Sci. Food and Agric. 74**:35-41.

Rhymer, J.M. e Simberloff, D. (1996). Extinction by hybridization and introgression. **Annu. Rev. Ecol. Syst. 27**:83-109.

Rijks, J.M., Hoffman, J.I., Kuiken, T., Osterhaus, A.D.M.A. e Amos W. (2008). Heterozygosity and lungworm burden in harbour seals (*Phoca vitulina*). **Heredity 100**:587-593.

Rios, R.I. (2008). **O funcionamento dos ecossistemas: a natureza é macunaímica** Educação Pública. CECIERJ, Rio de Janeiro.

Robertson, B.C. e Gemmell, N.J. (2004). Defining eradication units to control invasive pests., J. **Appl. Ecol. 41**:1042–1048.

Robertson, B.C. e Gemmell, N.J. (2006). PCR-based sexing in conservation biology: Wrong answers from an accurate methodology? **Conserv. Genet. 7**:267-271.

Roca, A.L., Georgiadis, N.J, Pecon-Slattery, J. e O'Brien, S.J. (2001). Genetic evidence for two species of elephant in Africa. **Science 293**:1473-1477.

Rosel, P.E., France, S.C., Wang, J.Y. e Kocher, T.D. (1999). Genetic structure of harbour porpoise *Phocoena phocoena* populations in the northwest Atlantic based on mitochondrial and nuclear markers. **Mol. Ecol. 8**:S41-S54.

Rossetto, M., Jezierski, G., Hopper S.D. e Dixon K.W. (1999). Conservation genetics and clonality in two critically endangered eucalypts from the highly endemic south-western Australian flora. Biological Conservation 88:321-331.

Russell, W.C., Thorne, E.T., Oakleaf, R. e Ballou, J.D. (1994). The genetic basis of Black-Footed Ferret reintroduction. **Conservat. Biol. 8**:263-266.

Russello, M.A., Beheregaray, L.B., Gibbs, J.P., Fritts, T., Havill, N., Powell, J.R. e Caccone, A. (2007). Lonesome George is not alone among Galápagos tortoises. **Current Biol.17**:317-318.

Ryder, O.A. (1986). Species conservation and systematics - the dilemma of subspecies. **Trends Ecol. Syst. 1**:9-19.

Ryman, N. e Utter, F. (1987). **Population genetics and fishery management**. Washington University Press, London.

Salgueiro, P., Palmeirim, J.M., Ruedi, M. e Coelho, M.M. (2008). Gene flow and population structure of the endemic Azorean bat (*Nyctalus azoreum*) based on microsatellites: implications for conservation. **Conserv. Genet. 9**:1163–1171.

Schonewald-Cox, C.M., Chambers, S.M., MacBryde, B. e Thomas, L. (1983). **Genetics and conservation**. Benjamin/Cummings, Menlo Park, California.

Schormann, J., Carlton, J.T. e Dochoda, M.R. (1990). The ship as a vector in biotic invasions. **Mar. Engin. Digests Oct 1990**:18-22.

Secchi, E.R., Danilewicz, D. e Ott, P.H. (2003). Applying the phylogeo-

graphic concept to identify franciscana dolphins stocks: implications to meet management objectives., J. **Cetac. Res. Manag. 5**:61-68.

Secchi, E.R., Wang, J.Y., Murray, B.W., Rocha-Campos, C.C. e White, B.N. (1998). Population differentiation in the franciscana (*Pontoporia blainvillei*) from two geographic locations in Brazil as determined from mitochondrial DNA control region sequences. **Can., J. Zool. 76**:1622-1627.

Seeb, J.E., Kruse, G.H., Seeb, L.W. e Weck, R.G. (1990). Genetic structure of red king crab populations in Alaska facilitates enforcement of fishing regulations. **Alsk. Sea Grant Coll. Prog. Rep. 90**:91-102.

Shaffer, M.L. (1981). Minimum Population Sizes for Species Conservation. **Bioscience 31**:131–134.

Sheppard, W.S. e Smith, D.R (2000). Identification of African-derived bees in the Americas:, A. survey of methods. **Ann. Entomol. Soc. Am. 93**:159-176.

Sholl, T.G.C., Nascimento, F.F., Leoncini, O., Bonvicino, C.R. e Siciliano, S. (2008). Taxonomic identification of dolphin love charms commercialized in the Amazonian region through the analysis of cytochrome b DNA., J. **Mar. Biol. Assoc. U.K. 88**:1207–1210.

Silva, E.P. e Russo, C.A.M. (2000). Techniques and statistical data analysis in molecular population genetics. **Hydrobiologia 420**:119-135.

Sinclair, EA., Webb, N.J, Marchant, A.D. e Tidemann, C.R. (1996). Genetic variation in the little red flying-fox *Pteropus scapulatus* (chiroptera: pteropodidae): implications for management. **Biol. Conserv. 76**:45-50.

Skibinski, D.O.F., Ahmad, M. e Beardmore, J.A. (1978). Genetic evidence for naturally occurring hybrids between *Mytilus edulis* and *Mytilus galloprovincialis.* **Evolution 32**:354-364.

Slade, R.W., Moritz, C., Hoelzel, A.R. e Burton, H.R. (1998). Molecular population genetics of the southern elephant seal *Mirounga leonina.* **Genetics 149**:1945-1957.

Slatkin, M. (1985). Gene flow in natural populations. **Ann. Rev. Ecol. Syst. 16**:393-430.

Slatkin, M. (1995). Hitchhiking and associative overdominance at a microsatellite locus. **Mol. Biol. Evol. 12** (3):473-480.

Sokal, R.R. e Rohlf, F.J (1995). **Biometry: the principles and practice of statistics in biological research**., W.H. Freeman and Co., San Francisco.

Solé-Cava, A.M. (1993). Vasculhando as Cinzas. **Bioletim 1** (1):26-27.

Solé-Cava, A.M. (2008a) Códigos de barras de DNA: o Rabo que abana o cachorro. **Ciencia Hoje 41**:65-67.

Solé-Cava, A.M. (2008b) Ressuscitando espécies? **Ciencia Hoje 42**:16-17.

Solé-Cava, A.M. e Levy, J.A. (1987). Biochemical evidence for a third species of angel shark (*Squatina*) off the East coast of South America. **Biochem. Syst. Ecol. 15**:135-144.

Solé-Cava, A.M. e Thorpe, J.P. (1994). Evolutionary genetics of marine sponges. In: van Soest, R.W.M., van Kempen, T.M.G. e Braekman, J.C. (eds) **Sponges in time and space**., A., A. Balkema, Rotterdam, pp 55-63.

Solé-Cava, A.M., Vooren, C.M. e Levy, J.A. (1983). Isozymic differentiation of two sibling species of *Squatina* (Chondrichthyes) in South Brazil. **Comp. Biochem. and Physiol. 75B**:354-358.

Soulé, M.E. (1980). Thresholds for survival: maintaining fitness and evolutionary potential. In: Soulé, M.E. e Wilcox, B.A.(eds) **Conservation Biology: An evolutionary-ecological perspective**. Sinauer Associates, Sunderland, Massachusetts, pp 111-124.

Soulé, M.E. e Wilcox, B.A. (1980). **Conservation Biology: An evolutionary-ecological perspective**. Sinauer Associates, Sunderland, Massachusetts.

Southward, A.J, Burton, R.S., Coles, S.L., Dando, P.R., DeFelice, R., Hoover, J., Parnell, P.E., Yamaguchi, T., Newman, W.A. (1998). Invasion of Hawaiian shores by an Atlantic barnacle. **Mar. Ecol. Prog. Ser. 165**:119-126.

Stefenon, V.M., Gailing, O. e Finkeldey, R. (2007). Genetic structure of *Araucaria angustifolia* (Araucariaceae) populations in Brazil: Implications for the in situ conservation of genetic resources. **Plant Biol. 9**:516-525.

Steller, G.W. (1751) De bestiis marini. **Novi Commentarii. Acad. Sci. Imp. Petropoli 2**:289-398.

Sweijd, N.A., Bowie, R.C., Evans, B.S. e Lopata, A.L. (2000). Molecular genetics and the management and conservation of marine organisms. **Hydrobiologia 420**:153-164.

Taylor, B.L. e Rojas-Brach,o L. (1999). Examining the risk of inbreeding depression in a naturally rare cetacean, the vaquita (*Phocoena sinus*). **Mar. Mamm. Sci. 15**:1004-1028.

Telles, M.P.D., Coelho, A.S.G., Chaves, L.J., Diniz, J.A.F. e Valva, F.D. (2003). Genetic diversity and population structure of *Eugenia dysenterica* DC. ("cagaiteira"-Myrtaceae) in Central Brazil: Spatial analysis and implications for conservation and management. **Conserv. Genet. 4**:685-695.

Thorpe, J.P. e Solé-Cava, A.M. (1994). The use of allozyme electrophoresis in invertebrate systematics. **Zool. Scripta 23**:3-18.

Thorpe, J.P., Sole-Cava A.M. e Watts, P.C. (2000). Exploited marine invertebrates: genetics and fisheries. **Hydrobiologia 420**:165-184.

Traill, L.W., Bradshaw, C.J.A. e Brook, B.W. (2007). Minimum viable population size:, A. meta-analysis of 30 years of published estimates. **Biol. Conserv. 139**:159-166

Udina, I.G. e Shaikhaev, G.O (1998). Restriction fragment length polymorphism (RFLP) of exon 2 of the MhcBibo-DRB3 gene in European bison *Bison bonasus*. **Acta Theriol. Vol.Sup**.:75-82.

Van Vallen, L. (1973)., A. new evolutionary law. **Evol. Theory 1**:1-30

Vernesi, C., Bruford, M.W., Bertorelle, G., Pecchioli, E., Rizzoli, A. e Hauffe, HC. (2008). Where's the Conservation in Conservation Genetics? **Conservat. Biol. 22**:802-804.

Vianna, J.A., Bonde, R.K., Caballero, S., Giraldo, J.P., Lima, R.P., Clark, A., Marmontel, M., Morales-Vela, B., De Souza, M.J, Parr. L., Rodriguez-Lopez, M.A., Mignucci-Giannoni, A.A., Powell, J.A. e Santos, F.R. (2006). Phylogeography, phylogeny and hybridization in trichechid sirenians: implications for manatee conservation. **Mol. Ecol. 15**:433-447.

Vila, C., Sundqvist. A.K., Flagstad. O., Seddon. J., Bjornerfeldt, S., Kojola. I., Casulli, A., Sand. H., Wabakken, P. e Ellegren, H. (2003). Rescue of a severely bottlenecked wolf (*Canis lupus*) population by a single immigrant. **Proc. Royal Soc. London 270B**:91-97.

Visscher, P.M., Smith, D., Hall. S.J.G. e Williams, J.A. (2001). A viable herd of genetically uniform cattle. **Nature 409**:203.

Vollmer. S.V. e Palumbi, S.R. (2007). Restricted gene flow in the Caribbean staghorn coral *Acropora cervicomis*: Implications for the recovery of endangered reefs., J. **Hered.98**:40-50.

Walton. M.J. (1997). Population structure of harbour porpoises *Phocoena phocoena* in the seas around the UK and adjacent waters. **Proc. Royal Soc. London 264B**: 89-94.

Waples, R.S. e Gaggiotti. O. (2006). What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. **Mol. Ecol. 15**:1419-1439.

Ward, R.D. (2000). Genetics in fisheries management. Hydrobiologia 420:191–201

Ward, R.D. (1995). Population genetics of Tunas. J. **Fish Biol. 47**:259-280.

Watts, P.C., Rouquette, J.R., Saccheri, I.J., Kemp, S.J. e Thompson, D.J. (2004). Molecular and ecological evidence for small-scale isolation by distance in an endangered damselfly, *Coenagrion mercuriale*. **Mol. Ecol. 13**:2931–2945

Watts, P.C., Buley, K.R., Sanderson, S., Boardman, W., Ciofi, C. e Gibson, R. (2006). Parthenogenesis in Komodo dragons. **Nature 444**:1021-1022

Wayne, R.K. (1996). Conservation genetics in the Canidae In: Avise, J.C. e Hamrick, J.L.(eds) **Conservation genetics**: **Case histories from nature**. Chapman and Hall, New York, pp 75-118

Weber, D.S., Stewart B.S., Garza, J.C. e Lehman, N. (2000). An empirical genetic assessment of the severity of the northern elephant seal population bottleneck. **Current Biol. 10**:1287-1290.

Weidema, I.R., Siegismund, H.R. e Philipp, M. (1996). Distribution of genetic variation within and among Danish populations of *Armeria maritima*, with special reference to the effects of population size. **Hereditas 124**:121-129.

Wheeler, Q.D. (1995). Systematics, the scientific basis for inventories of biodiversity. **Biodivers. Conserv. 4**:476-489.

Whitlock, M.C. e McCauley, D.E. (1999). Indirect measures of gene flow and migration: FST is different from 1/(4Nm+1). **Heredity 82**:117-125

Wildt, D.E., Bush, M., Goodrowe, K.L., Packer, C., Pusey, A.E., Brown, J.L., Joslin, P. e O'Brien, S.J. (1987). Reproductive and genetic consequences of founding isolated lion populations. **Nature 329**: 1751-1755

Wilson, G.A. e Strobeck, C. (1999). Genetic variation within and relatedness among wood and plains bison populations. **Genome 42**:483-496

Wink, M., Sauer-Gurth, H., Martinez, F., Doval, G., Blanco, G. e Hatzofe O. (1998). The use of (GACA)(4) PCR to sex Old World vultures (Aves : Accipitridae). **Mol. Ecol.7**: 779

Wright, S. (1931). Evolution in mendelian populations. Genetics 16:97-159

Wright, S. (1978). **Evolution and the genetics of populations**. The University of Chicago Press, London

Yuhki, N. e O'Brien, S.J. (1990). DNA variation of the mammalian major histocompatibility complex reflects genomic diversity and population history. **Proc. Nat. Acad. Sci. USA. 87**: 836-840.

Glossário

Sergio Russo Matioli (srmatiol@ib.usp.br)
Departamento de Genética e Biologia Evolutiva,
Instituto de Biociências, USP

As palavras estrangeiras estão em itálico (L. latim e I. inglês). As palavras definidas no próprio glossário estão sublinhadas.

A

Abiótico. Sem a presença de vida.
Ácido graxo. Ácido orgânico com uma longa cadeia alifática (de carbono e hidrogênio).
Actina. Proteína que participa da contração muscular e de outro tipos de movimentos celulares.
Aeróbico. Que depende de oxigênio para viver.
Aflatoxina. Toxina produzida pelo fungo *Aspergillus flavus*.
Agarose. Polissacarídeo purificado do agár-agár, material gelificante extraído de algas.
Alelos. Formas alternativas de um gene para um determinado loco.
Aleurona. Uma das camadas externas do endosperma da semente de monocotiledôneas.
Algoritmo. Série de procedimentos logicamente encadeados para a resolução de um problema.
Alozimas. Isozimas decorrentes de formas alélicas de seus genes.
Alquilação. Reação química onde um radical alquila (acetila, metila, propila, etc.) é adicionado a uma molécula orgânica.
Aminoácido. Composto orgânico que se caracteriza por possuir um grupo carboxila e um gupo amina ligados a um átomo de carbono. São os componentes das proteínas
Anabolismo. Reações químicas do metabolismo envolvidas na síntese de substâncias.
Anaeróbico. Que não depende de oxigênio para viver.
Aneuploidia. Condição em que uma célula ou indivíduo contem um número cromossômico que não é múltiplo do número básico característico da espécie.
Ânion. Íon carregado negativamente (ex.: Cl^-, $SO4^{--}$).
Ânodo. Eletrodo com carga positiva que atrai ânions.
Anticorpos. Proteínas sintetizadas por linfócitos que têm a capacidade de se agregar com substâncias estranhas ao corpo.
Antígeno. Substância que desencadeia a síntese de anticorpos específicos para essa substância.
Antissenso. Macromolécula que, por ser complementar a uma determinada macromolécula sintetizada na célula, bloqueia a atuação desta.
Apomórfico. Condição de um caráter que apresenta o estado derivado.
Apoptose. Morte celular efetuada por mecanismos endógenos.
Arginina. Aminoácido de característica básica por possuir um grupo amina em sua cadeia lateral.
Arquéia. Sinônimo de Arqueobactéria.
Arqueobactéria. Organismo procarioto que possui diéter de isoprenila glicerol e lipídios tetraéter como constituintes de sua parede celular. Um dos três domínios dos seres vivos. Os outros dois são as eubactérias e os eucariotos.
ATP. Sigla em inglês do trifosfato de adenosina, ribonucleotídeo rico em energia.
ATPase. Enzima que catalisa a reação de hidrólise de um grupo fosfato do ATP.
Autocatalítico. Propriedade de uma molécula que permite catalisar uma reação química em si própria.
Autossomo. Cromossomo presente sob a forma diploide nos dois sexos que não possuem genes para a determinação primaria do sexo.
Autótrofo. Organismo que obtêm energia a partir de reações químicas que independem de produtos de outros organismos.
Auxotrófico. Estado em que um organismo depende de um determinado nutriente para sobreviver.
Avidina. Proteína que se liga à biotina com alta afinidade.

B

Bacteriófago. Vírus que infecta bactérias provocando sua lise ou que pode incorporar-se ao genoma do hospedeiro.
Bipedalismo. Propriedade de animais que se locomovem por ação de seus membros posteriores.
Biogeografia. Disciplina relativa ao estudo da distribuição geográfica dos organismos.
Bioinformática. Disciplina que emprega técnicas computacionais no estudo de biopolímeros, genomas e demais constituintes das células e tecidos.
Bioinvasora. Qualidade de uma espécie que pode colonizar hábitats em regiões nas quais não era encontrada anteriormente.
Biopolímero. Polímero de origem biológica.
Biotina. Composto orgânico heterocíclico com dois anéis. Vitamina H ou B7.
Bipedalismo. Propriedade de organismos que andam sobre duas patas.
Bootstrap (I.) Procedimento em que se verifica a proporção de vezes em que uma determinada estimativa ocorre além de um limite prestabelecido através de sorteios repetidos, com reposição, dos dados originais.

C

Cadeia de Markov. Processo estocástico descontínuo que depende das probabilidades de mudança de estado e dos estados anteriores.
Cadeia respiratória. Série de reações químicas que resulta na oxidação de moléculas orgânicas pelo oxigênio.
Capsídeo. Envólucro proteico de vírus.
Carcinogênese. Processos envolvidos na origem de neoplasias.
Caspase. Família de proteínas com função de proteases que apresenta um resíduo de cisteína em seu sítio ativo.
Catabolismo. Reações químicas do metabolismo envolvidas na degradação de substâncias.
Catalisador. Agente que participa de uma reação química aumentando a velocidade da mesma, sem no entanto se modificar durante a reação.
Cátion. Íon carregado positivamente (ex:. Na^+, Mg^{++}).
Cátodo. Eletrodo com carga negativa que atrai cátions.
cDNA. DNA obtido a partir de RNA pela ação da transcriptase reversa.
Células germinativas. Células de um organismo diploide que originarão gametas.
Células somáticas. Células de um organismo que não passam material genético aos descendentes.
Centrômero. Porção do cromossomo onde o mesmo se liga às fibras do fuso durante a divisão celular.
Chaperona. Proteína que é sintetizada em condições de estresse (altas temperaturas por exemplo) e que participa do dobramento correto de outras proteínas.
Chip (I.) Arranjo bidimensional de circuitos eletrônicos.
***Chip* de DNA** (I.) Arranjo bidimensional de oligonucleotídeos ou polinucleotídeos.
Cianobactérias. Também chamadas de algas azuis, são microorganismos procarióticos que têm a capacidade de realizar fotossíntese.
Ciliados. Grupo de protistas que se locomovem pela ação de cílios.
Cistron. Região do genoma bacteriano ou virótico que contém a

sequência codificante para um único polipeptídeo.

Citocromo. Molécula constituída por uma cadeia polipeptídica e por um grupo prostético heme que participa da cadeia respiratória.

Citoesqueleto. Conjunto de elementos estruturais do citoplasma, composto principalmente por proteínas de natureza filamentosa e microtúbulos.

Citogenética. Disciplina relativa ao estudo dos cromossomos.

Citoquímica. Disciplina relativa ao estudo de componentes químicos e bioquímicos das células através de visualização *in situ* com corantes.

Clado. Representação gráfica de um grupo monofilético em uma filogenia.

Cladogênese. Evento de especiação.

Clone. 1. Conjunto de células que são cópias originadas a partir de reprodução assexuada originários de uma única célula. 2. Conjunto de réplicas de macromoléculas originadas pela inserção, em um organismo, de uma única macromolécula. 3. Indivíduo produzido por reprodução assexuada.

Clorofila. Pigmento que atua na fotossíntese.

Cloroplasto. Organela membranosa onde ocorre a fotossíntese.

Código genético. Correspondência de cada um dos 64 códons possíveis com a ação que ocorre durante a síntese protéica (incorporação de um aminoácido ou término da síntese).

Codominância. Modo de herança onde o indivíduo heterozigoto possui um fenótipo intermediário relativo aos indivíduos homozigotos para os alelos que possui.

Códon. Sequência de três bases nitrogenadas que corresponde a um aminoácido durante a síntese protéica, de acordo com o código genético.

Coenzima. Cofator que não permanece ligado com a enzima.

Coevolução. Processo evolutivo em que as interações ecológicas entre duas ou mais espécies resulta em adaptações recíprocas (ex.: aumento concomitante de velocidades em presas e predadores, desenvolvimento de mecanismos químicos de defesa em hospedeiros e metabolização dos produtos tóxicos pelos parasitas, etc.).

Cofator. Molécula ou íon necessário para que uma reação seja catalizada por uma enzima.

Complexo de Golgi. Organela membranosa relacionada com o retículo endoplasmático e com o processo de secreção.

Complexo de Histocompatibilidade. Conjunto de genes relacionados com a histocompatibilidade.

Condritos. Minerais rochosos constituintes de meteoros formador por poeira e grânulos que estiveram presentes durante a formação do Sistema Solar.

Cotilédone. Folha embrionária presente em sementes de espermatófitas (Gimnospermas e Angiospermas).

Covariância. Medida de codispersão de dois tipos de dados de uma mesma amostra.

cpDNA. DNA de cloroplastos.

Criofratura. Técnica em que um tecido é rompido em temperaturas muito baixas para a visualização de estruturas tridimensionais por microscopia de varredura.

Cromátide. Uma das réplicas de cromossomos produzidas durante a divisão destes.

Cromatina. Fração do núcleo de células eucarióticas que têm afinidade por corantes ácidos.

Cromatografia. Migração de substâncias pelo deslocamento de um solvente através de um substrato sólido poroso.

Cruciforme. Que tem a forma de uma cruz.

D

Datação. Estimativa da época de ocorrência de um determinado evento.

Deleção. Tipo de mutação onde um ou mais nucleotídeos são removidos no meio de uma cadeia. Quando se refere a grandes segmentos perdidos que podem ser detectados em análises cromossômicas é dita deleção cromossômica.

Dendrograma. Representação gráfica em forma de ramos e nós onde as entidades são agrupadas de acordo com suas similaridades.

Desaminação. Reação de clivagem de um gupo amina ($-NH_2$).

Desidrogenase. Classe de enzimas que catalizam a reação de transferência de dois átomos de hidrogênio para um cofator.

Desoxirribonucleotídeo. Nucleotídeo cujo açúcar é a desoxirribose.

Desoxirribose. Açúcar com estrutura semelhante à da ribose diferindo desta por possuir um hidrogênio no carbono 2 ao invés de uma hidroxila.

Dicotiledôneas. Plantas cujas sementes têm dois cotilédones.

DNA. Sigla em inglês do ácido desoxirribonucléico.

Dominância. Modo de herança onde o indivíduo heterozigoto possui um fenótipo igual a um dos indivíduos homozigotos.

dsRNA. RNA de fita dupla ("double stranded RNA" em inglês)

Duplicação gênica. Processso pelo qual um gene origina duas cópias ou mais que são transmitidas para um mesmo descendente.

E

Ecotoxicologia. Disciplina relativa ao estudo de toxidez de produtos químicos que são lançados no ambiente.

Efetora. Substância qua altera a atividade de uma enzima diminuíndo-a ou aumentando-a.

Elementos móveis. Sequências do genoma que têm a capacidade de mudar de posição, deixando ou não uma cópia no local original.

Eletrodo. Terminal elétrico metálico em contato com uma solução.

Eletroforese. Migração de substâncias eletricamente carregadas movidas pela ação de um campo elétrico.

Enantiômeros. Cada um dos isômeros ópticos de moléculas orgânicas com carbono assimétrico, ou seja, que possui quatro ligantes diferentes.

Endemismo. Distribuição geográfica restrita.

Endonuclease. Classe de enzimas que catalisam a hidrólise da cadeia polinucleotídica a partir de uma posição interior da mesma.

Endonuclease de restrição. Endonuclease que cliva uma molécula de fita dupla de DNA em um trecho que contém uma seqência específica de nucleotídeos.

Endosperma. Tecido de sementes rico em nutrientes.

Endossimbiose. Relação ecológica onde um dos organismos vive no citoplasma das células de um outro organismo.

Epigênese. 1. Propriedade relativa à formação de caracteres que resulta da interação entre componentes individuais. 2. Teoria segundo a qual o desenvolvimento de um organismo se dá pelo aumento gradativo de complexidade, em contraposição à teoria preformacionista, segundo a qual há somente aumento de tamanho.

Epigenético. Relativo à epigênese. Diz-se dos mecanismos de transmissão de informações não relacionados diretamente ao DNA.

Enzima. Catalisador de origem biológica que possui polipeptídios em sua estrutura

Eócito. Grupo hipotético de arqueobactérias ancestral dos eucariotos.

Especiação. Processo pelo qual uma espécie origina uma ou mais espécies diferente(s) da espécie ancestral.

Especialista. Organismo com pequena amplitude ecológica.

Espécie. 1. Grupo de indivíduos que compartilham o mesmo conjunto de características biologicamente relevantes e que não podem ser divididos em outros grupos mais inclusivos pelo mesmo critério. 2. Grupo de indivíduos que reproduzem-se entre si e não o fazem com outros grupos semelhantes. 3. Grupo de indivíduos que compartilham o mesmo destino evolutivo. 4. Grupo de indivíduos semelhantes. (N. do E. há dezenas de conceitos de espécies na literatura, não havendo consenso devido a circunstâncias nas quais um dos conceitos

é mais aplicável que outro).
Espectrometria de massa. Método químico analítico que permite a identificação de moléculas orgânicas.
Espermatogênese. Processo de formação de espermatozoides.
Esteroide. Classe de lipídios derivados do ciclopentano peridrofenantreno (ex. colesterol, testosterona, estrógeno).
Estocástico. Processo que é influenciado por efeitos do acaso.
Estrógeno. Hormônio esteroide produzido pelos ovários das fêmeas de mamíferos.
Estromatólito. Rocha formada por camadas oriundas de atividade bacteriana.
Eubactéria. Organismo procarioto que possui derivados de diacila glicerol como constituinte de sua parede celular. Um dos três domínios dos seres vivos. Os outros dois são as arqueobactérias e os eucariotos.
Eucarioto. Organismo cujas células possuem um ou mais núcleos delimitados por membrana lipoprotéica (carioteca).
Excisão. Corte interno seguido de emenda.
Exon. Sequência que efetivamente codifica para a síntese do polipeptídeo correspondente.
Exonuclease. Classe de enzimas que catalisam a hidrólise da cadeia polinucleotídica a partir de uma das extremidades da mesma.

F

Fago. O mesmo que bacteriófago.
Fagocitose. Processo pelo qual uma partícula ou um microorganismo é introduzido em uma célula através do seu encapsulamento pela membrana plasmática da célula.
Fagossomo. Cápsula intracelular resultante do processo de fagocitose.
Família gênica. Conjunto de genes de sequências similares que se originaram por duplicação gênica.
Fenotípico. Referente ao fenótipo.
Fenótipo. 1. A aparência de uma característica de um organismo. 2. O resultado da interação entre o genótipo de um organismo e o ambiente onde ele se desenvolveu.
Filogenia. Esquema que representa a relação evolutiva de parentesco entre as entidades consideradas.
Filogeografia. Estudo histórico da distribuição geográfica de organismos.
Filopatria. Tendência à reprodução na região geográfica natal.
Fingerprinting (I.) Técnica de detecção de locos polimórficos correspondentes a minissatélites.
Fitness (I.). Aptidão darwiniana, estado de estar adaptado ao ambiente.
Fitopatógeno. Organismo que causa doença em vegetais.
Fluorescência. Propriedade de uma molécula que consiste na emissão de luz quando atingida por luz de um determinado comprimeto de onda. A luz emitida tem um comprimento de onda maior que a luz atingida.
Fosfatase. Enzima que catalisa a hidrólise de um grupo fosfato.
Fosfodiéster. Ligação que existe entre uma molécula de açúcar e um fosfato.
Fosfodiesterase. Classe de enzimas que catalisam a hidrólise de ligações fosfodiéster.
Fóssil. Resto mineralizado ou marca deixada por um organismo morto há muito tempo.
Fotoliase. Classe de enzimas que utilizam a energia da luz absorvida na realização de reparo direto do DNA.
Fotorreativação. Fenômeno no qual organismos expostos à luz apresentam menor quantidade de mutantes após irradiados com luz ultravioleta do que aqueles que permaneceram na escuridão.
Fotossíntese. Conjunto de reações químicas onde a energia proveniente da luz é utilizada na síntese de moléculas orgânicas a partir de CO_2 e H_2O.

G

Gama (distribuição). Distribuição de frequências que é um caso geral de outras distribuições com formas mais definidas.
Genoma. O conjunto de genes de um lote haploide de um organismo.
Genótipo. O potencial de um organismo referente ao seu material genético.
Glicerol. Trihidroxi propano. O mesmo que glicerina.
Glicólise. Série de reações anabólicas a partir da glicose.
Glicoproteína. Proteína associada a polissacarídeos.
Glicosilase. Classe de enzimas que catalisam a hidrólise de ligações glicosídicas.
Glicossomo. Organela membranosa que contém enzimas da glicólise de tripanosomatídeos.
Globina. Polipeptídeo encontrado na hemoglobina.
Gluteína. Uma das proteínas encontradas no glúten, que é a fração protéica da farinha de trigo, cevada ou centeio.
Grupo prostético. Cofator que se liga de forma duradoura com a enzima.

H

Hábitat. Conjunto dos espaços físicos ocupado por um organismo.
Haploide. Célula, indivíduo ou ornanismo que contém uma cópia do conjunto dos cromossomos característicos da espécie .
Haplótipo. Combinação de alelos de um segmento colinear de *DNA*.
Hematopoiese. Porcesso de produção de células sanguíneas.
Heme. Grupo prostético dos citocromos, mioglobinas e hemoglobinas, constituído por anéis de porfirina e um íon de Ferro.
Hemoglobina. Molécula constituída por globinas e um grupo prostético heme.
Herança materna. Modo de herança uniparental onde o material genético origina-se do ancestral fêmea.
Herança uniparental. Modo de herança onde o material genético é herdado de um dos pais.
Herdabilidade. Fração da variância fenotípica que é devida à variância genética aditiva.
Heterocromatina. Região cromossômica que é intensamente corada após tratamento alcalino, que se condensa durante a intérfase.
Heteroduplex. Fita dupla de ácido nucléico onde cada uma das fitas tem origem diferente.
Heterogamético. Sexo no qual os cromossomos sexuais são diferentes, por exemplo, o sexo masculino XY em mamíferos e o sexo feminino ZW em aves.
Heterose. Condição onde o descendente híbrido possui características aumentadas em relação a ambos os pais.
Heterótrofo. Organismo que obtêm energia a partir de reações químicas que dependem de produtos de outros organismos.
Heterozigosidade. Proporção de heterozigotos em uma amostra populacional.
Heterozigoto. Que possui um loco com dois alelos diferentes.
Heurístico. Procedimento algorítmico simplificado e eficiente mas que não garante a solução exata de um determinado problema.
Hidrofobicidade. Medida que reflete o quanto uma molécula é insolúvel em água.
Hidrogenossomo. Organela membranosa de protistas anaeróbicos relacionada com fermentação.
Hidrolase. Classe de enzimas que catalizam a reação de hidrólise.
Hidrólise. Reação química em que uma ligação covalente é rompida com a produção de uma molécula de água.
Hipercolesterolemia. Taxa de colesterol do sangue mais elevada que aquela encontrada em um intervalo considerado como normal.
Histocompatibilidade. Propriedade que um tecido tem de não sofrer o processo de rejeição por parte do sistema imune de um receptor após um transplante.
Histona. Proteína de caráter básico encontrada nos cromossomos e que participa da estrutura destes associando-se ao DNA.
Homeobox. Sequência consenso de cerca de 180 nucleotídeos que codificam para o homeodomínio.

Homeodomínio. Sequência consenso de cerca de 60 aminoácidos que está presente em fatores de transcrição que atuam durante o desenvolvimento, tais como os genes homeóticos.
Homeotérmicos. Organismos com alta capacidade endógena de regulação térmica.
Homeótico. Tipo de gene que atua durante o desenvolvimento relacionado com a presença de estruturas específicas para os diferentes segmentos do corpo.
Homogamético. Diz-se do sexo que é determinado pela presença de dois cromossomos sexuais do mesmo tipo, por exemplo o sexo feminino XX em mamíferos ou o sexo masculino ZZ em aves.
Homologia. Propriedade de estruturas biológicas com uma origem evolutiva comum.
Homoquiralidade. Propriedade de uma substância que possui somente um tipo de isômero óptico.
Homoplasia. Semelhança entre condições apomórficas surgidas de forma paralela ou convergente. Semelhança não-homóloga.
Homozigoto. Que possui um loco com dois alelos iguais.

I

Imino. Grupo químico C=NH.
Imunoensaio. Método de detecção de um antígeno pela reação com seu anticorpo.
Imunoglobulina. Proteína com a propriedade de se ligar a antígenos. Anticorpo.
In situ (L.). Situação experimental que é feita no próprio material biológico.
In vitro (L.). Situação experimental independente de tecidos vivos.
In vivo (L.). Situação experimental dependente de tecidos vivos.
Indel. Evento que envolveu uma deleção ou inserção nucleotídica, pressuposto através da comparação de cadeias nucleotídicas homólogas com comprimentos diferentes.
Indutivismo. Premissa epistemológica segundo a qual o conhecimento pode ser obtido através do acúmulo de observações.
Inserção. Tipo de mutação onde um ou mais nucleotídeos são inseridos no interior de uma cadeia nucleotídica.
Interferência do RNA. Processo no qual moléculas transcritas de RNA interferem na expressão de outras moléculas de RNA por possuírem sequências complementares a estas.
Interferon. Proteína de cadeia curta liberada por células infectadas por vírus.
Intergênica. Refere-se à região dos genomas situada entre genes.
Introgressão. Fluxo gênico unidirecional como consequência de cruzamentos entre indivíduos de espécies, raças, subespécies, variedades ou populações distintas.
Intron. Sequência de DNA que intercala sequências codificantes mas cuja informação não é traduzida no polipeptídeo correspondente.
Invaginação. Movimento de uma membrana em direção ao conteúdo que a mesma recobre.
Íon. Átomo carregado eletricamente.
Isócoros. Sequências de DNA com relação dos nucleotídeos AG/CT semelhantes.
Isoenzimas. O mesmo que isozimas.
Isoesquizômeros. Endonucleases de restrição de organismos diferentes que catalizam a hidrólise da mesma sequência nucleotídica.
Isômeros. Moléculas com a mesma fórmula química mas com estrutura diferente.
Isótopos. Átomos de um mesmo elemento que diferem em massa pela quantidade do número de nêutrons em seu núcleo.
Isozimas. Formas de enzima*s* diferentes que catalizam a mesma reação.

K

Kb. Kilobase, ou 1.000 pb.
Knock-down. [gênico] (I.). Supressão experimental da expressão de um gene.

L

Ligase. Classe de enzimas que catalizam a formação de uma ligação fosfodiéster entre duas cadeias de polinucleotídeos.
Linfócito. Célula da série branca do sangue relacionada com a resposta imune.
Lipídios. Compostos químicos insolúveis em água existentes nas células nas membranas ou sob a forma de gotículas. O mesmo que gordura.
Lipossomo. Organela membranosa rica em lipídios.
Loco. Posição de um gene no cromossomo.

M

Meia-vida. Tempo que leva para que a metade dos elementos de um conjunto seja extinta ou transformada.
Meiose. Processo de divisão celular onde uma célula diploide origina quatro células haploides.
Merofilético. Grupo não monofilético (parafilético ou polifilético)
Metabolismo. Conjunto de reações químicas de um organismo.
Metilação. Adição de um radical metila a uma molécula orgânica.
Metilase. Classe de enzimas que catalizam a alquilação de uma molécula com uma metila.
Microcromossomo. Cromossomo excepcionalmente muito pequeno.
Microhábitats. Subdivisões do hábitat de um organismo.
Mícron. Um milésimo de milímetro. O mesmo que micrômetro, 10^{-6} m.
Microssatélite. Sequência de DNA que consiste em repetições de subsequências muito curtas (2 a 10 nucleotídeos).
Microssomal. Referente aos microssomos.
Microssomo. Vesícula intracelular que se origina pela fragmentação do retículo endoplasmático.
Microtúbulos. Estruturas proteicas em forma de tubos que existem no citoplasma das células.
Minicromossomo. Cromossomo excepcionalmente pequeno.
Minissatélite. Sequência de DNA que consiste em repetições de subsequências curtas (11 a 100 nucleotídeos).
Mitocôndria. Organela membranosa dos eucariotos onde ocorrem as reações da cadeia respiratória.
Monocotiledôneas. Plantas cujas sementes têm apenas um cotilédone.
Monofilético. Diz-se de um grupo taxonômico de organismos constituído por todos os descendentes de uma única espécie ancestral.
Monômeros. Cada uma das estruturas químicas básicas da qual um polímero é formado.
Mosaicismo. Presença em um organismo, órgão ou tecido de células diferentes em seu conteúdo genético.
mRNA. *RNA* mensageiro.
miRNA. Micro-RNA, cadeia curta de RNA (20 a 26 ribonucleotídeos) com efeito na expressão de genes.
mtDNA. DNA mitocondrial.
multienzimático. Diz-se de complexos proteicos que possuem diversas propriedades catalíticas diferentes.
Mutação. Alteração do material genético que é herdada.
Mutagênese. Processos envolvidos na origem de mutantes.
Mutagênico. Condição de natureza física ou química que aumenta a chance de produção de uma lesão no material genético e, consequentemente, de uma mutação genética.
Mutante. Indivíduo ou macromolécula que apresenta uma mutação.

N

ncRNA. RNA não codificante, ou seja, que é transcrito mas não codifica para polipeptídeos.
nDNA. DNA do núcleo da célula.
Nemátodo. Verme da classe Nematoda, do filo Aschelminthes.
Neoplasia. Tumor maligno.
Neutralismo. Pressuposição de que a maioria dos polimorfismos e substituições moleculares não têm influência na adaptação

dos organismos.

Northern blot (I.) Processo de transferência por capilaridade de fragmentos de RNA de um gel de eletroforese para uma membrana de nitrocelulose ou de nylon para que posteriormente sejam detectados com sondas específicas de DNA ou RNA marcadas. O nome "Northern" (do norte) trata-se de um trocadilho feito a partir da analogia com o Southern blot, pois Southern significa, além do nome do autor da técnica, "do Sul", em inglês.

Nucleosídeo. Molécula composta por uma base nitrogenada e um açúcar.

Nucleotídeo. Molécula composta por uma base nitrogenada, um açúcar e um grupo fosfato.

O

Oligonucleotídeo. Polímero de nucleotídeos de cadeia curta (até cerca de 35 monômeros).

Oligopeptídeo. Polímero de aminoácidos com poucos resíduos.

Oncogene. Gene que tem o potencial de causar neoplasias.

Oogênese. Processo de formação de óvulos.

ORF. Sigla em inglês de Quadro Aberto de Leitura (de "*Open Reading Frame*"), ou seja, uma sequência de nucleotídeos que se inicia com o códon de início de leitura distante de um número arbitrário de nucleotídeos correspondentes a um códon sem sentido ou de parada.

Organela. Corpúsculo intracelular mebranoso presente no citoplasma de organismos procarióticos.

Ortologia. Homologia decorrente de especiação.

Oxidase. Classe de enzimas que catalizam uma reação de oxi-redução.

Ovíparo. Modo de reprodução através de ovos onde o desenvolvimento do descendente ocorre após a postura.

Ovovivíparo. Modo de reprodução através de ovos, onde o desenvolvimento do descendente ocorre antes da postura.

Ozônio. Molécula em anel costituída por três átomos de oxigênio.

P

Paleobiologista. Especialista no estudo de formas de vida antigas.

Paleoclima. Clima de eras passadas.

Palíndromo. Forma de anagrama onde uma sequência de letras é a inversa da original (Ex.: Amor, Roma). Quando se fala em palíndromos em uma molécula de DNA, a sequência palindrômica está na cadeia complementar (ex.: 5' ACGGCCGT 3', que é complementar a 3' TGCCGGCA 5').

Panmítica. Propriedade de uma população onde há panmixia.

Panmixia. Cruzamentos ao acaso.

Parafilético. Agrupamento, em uma filogenia, que exclui um ou mais descendentes de um mesmo ancestral.

Paralogia. Homologia decorrente de duplicação gênica.

Paralelismo. Mudanças evolutivas independentes de caracteres para o mesmo estado.

Paramagnético. Que apresenta a propriedade de ser atraído por um campo magnético.

Parcimônia. Adoção da hipótese mais simples.

Partenogênese. Forma de reprodução onde fêmeas não fecundadas produzem descendência com indivíduos de um mesmo sexo, sendo arrenótoca aquela que produz somente machos e telítoca aquela que produz somente fêmeas.

Patógeno. Organismo que causa alguma doença.

pb. Pares de bases. Medida de comprimento de ácidos nucléicos de fita dupla que equivale a 1 par de nucleotídeos.

PCR. Sigla em inglês de Reação em Cadeia da Polimerase.

Pecilotérmicos. Organismos com reduzida capacidade endógena de regulação térmica.

Penetrância. Probabilidade empírica de manifestação fenotípica de uma característica genética, utilizada principalmente em características com herança dominante.

Peritonite. Inflamação do peritônio, membrana que reveste a cavidade abdominal.

Permeabilização. Ato de tornar permeável, de possibilitar a passagem de moléculas.

Peroxissomo. Organela membranosa rica em peroxidases.

pH. Cologaritmo da concentração hidrogeniôncia de uma solução. Reflete a acidez (pH < 7,0) ou a basicidade (pH > 7,0) da mesma.

Pirimidina. Base nitrogenada que consiste em uma cadeia fechada com carbono, nitrogênio e hidrogênio.

Plastos. Organelas encontradas em plantas e algas que podem ser pigmentadas (cromoplastos) ou não (leucoplastos).

Pleitotropia. Influência de um gene em mais de um caráter.

Plesiomorfia. Estado primitivo de carateres.

Poisson (distribuição de). Tipo de distribuição de frequências, onde o desvio-padrão é maior que a média, originando-se de um caso particular da distribuição binomial $(\mathbf{p}+\mathbf{q})^n$, onde **p** é muito maior que **q**.

Policistrônico Diz-se de uma região de DNA cujo transcrito codifica para dois ou mais cistrons.

Poliacrilamida. Polímero hidrossolúvel utilizado como matriz porosa para eletroforese.

Poliadenilação. Polimerização de uma cadeia polinucleotídica com o nucleotídeo Adenina como monômero na ausência de molde.

Poliamina. Moléculas orgânicas com múltiplos radicais amina (NH2) que possuem natureza policatiônica. Ex.: Putrescina (diamina), Espermidina (triamina).

Polifilético. Diz-se de um grupo, em uma filogenia, constituído por descendentes de mais de um ancestral.

Polimerase. *Enzima* que catalisa uma reação de polimerização.

Polimerização. Reação química onde monômeros são unidos resultando em cadeias mais longas.

Polimorfismo. Propriedade de um loco gênico em uma população onde há dois ou mais alelos segregando, cujas frequências são superiores a 1% ou 5%.

Polipeptídeo. Polímero resultante da polimerização de aminoácidos através da ligação de um grupo amina de um aminoácido com um grupo carboxila de outro (ligação peptídica).

Poliploidia. Estado da célula, tecido ou organismo que apresenta três ou mais cópias do conjunto cromossômico característico para a espécie.

Polipurina. Sequência de nucleotídeos cujas bases nitrogenadas são purinas.

Polissacarídeo. Polímero cujo monômero é uma molécula de açúcar.

Polissomia. Estado da célula que apresenta um cromossomo duplicado.

Politênico. Cromossomo gigante que resulta da replicação das cromátides sem que estas se separem.

Politômico. Que se separa em mais de dois ramos.

Ponto isoelétrico. pH no qual a molécula possui carga elétrica líquida neutra.

Porfirina. Molécula com 4 anéis alifáticos que se associa a um íon de metal pesado, normalmente o Ferro.

Predação. Relação ecológica onde um organismo (predador) alimenta-se de outro (presa), provocando a morte deste.

***Primer* (I.).** Oligonucleotídeo que inicia uma reação de polimerização a partir de sua hibridação com a cadeia molde.

Procarioto. Organismo cujas células não possuem núcleo delimitado por membrana lipoprotéica

Progênie. O conjunto dos indivíduos que são resultantes da reprodução de um indivíduo ou de um casal.

Progenota. Ancestral hipotético de todos os organismos.

Protease. Enzima que catalisa a reação de hidrólise de polipeptídeos.

Proteína. Macromolécula cuja cadeia principal é constituída por um ou mais polipeptídeos.

Protistas. Reino dos eucariotos que compreende protozoários e algas unicelulares.

Protocélula. Estrutura ancestral da célula.

Proto-oncogene. Genes que tem o potencial de tornarem-se oncogenes por meio de certos tipos de mutação.

Prototrófico. Estado em que um organismo não depende de nutrientes adicionais.
Pseudoaleatório. Número produzido através de um algoritmo computacional, cuja série tem propriedades que, idealmente, não a permitem distinguir de uma série de números aleatórios.
Pseudogenes. Sequências polinucleotídicas do genoma que não são funcionais, mas que possuem similaridades com sequências funcionais.
Purina. Base nitrogenada que consiste em duas cadeias fechadas com carbono, nitrogênio e hidrogênio.

Q

Quiasma. Figura em forma de cruz resultante de recombinação entre cromátides de cromossomos homólogos que pode ser visualizada durante a meiose.
Quiralidade. Propriedade de moléculas com relação aos seus isômeros ópticos.
Quinase. Enzima que catalisa a transferência de um fosfato de um *ATP* para outra molécula.

R

Radioativo. Material que, ao se desintegrar, emite radiação ionizante.
rbcL. O mesmo que Rubisco.
rDNA. Sequências de DNA do genoma que codificam para rRNA.
Recombinação. Processo em que o material genético de um indivíduo forma uma nova combinação com o material genético proveniente de outro indivíduo.
Reducionismo. Pressuposto filosófico em que os fenômenos que são estudados em uma das áreas da Ciência podem ser explicados pelos fenômenos que são estudados nas áreas com níveis menores de complexidade.
Redutase. Classe de enzimas que catalisam uma reação de oxirredução preferencialmente no sentido da redução do substrato que não é o cofator.
Replicação. Processo pelo qual uma entidade origina duas entidades iguais à original.
Replicase. Enzima hipotética que catalisa a reação que produz réplicas dela própria.
Retículo endoplasmático. Conjunto de membranas duplas existente nos citoplasmas de eucariotos.
Retroelementos. Transposons que se inserem no genoma a partir de *RNA*
Retrotransposição. Transposição onde há a produção de uma molécula intermediária de RNA.
Retrotransposons. Elementos móveis de DNA que, por transcrição reversa, originam cópias sob a forma de RNA.
Retrovírus. Vírus que tem o RNA como seu material genético.
Ribonuclease. Classe de enzimas que catalizam a hidrólise do RNA.
Ribonucleoproteína. Molécula formada por RNA e polipeptídeos.
Ribonucleotídeo. Nucleotídeo cujo acúcar é a ribose.
Ribose. Açúcar (polihidroxi aldeído ou cetona) com 5 carbonos em uma cadeia fechada.
Ribossomo. Corpúsculo citoplasmático composto por RNA e proteínas onde se dá a síntese protéica.
Ribozima. RNA com propriedades catalíticas.
RNA. Sigla em inglês do ácido ribonucléico, polímero de ribonucleotídeos.
RNAi. Interferência de RNA, fenômeno em que pequenos trechos de RNA de fita dupla interferem na expressão de genes com sequências complementares a uma de suas fitas.
RNAse. Ribonuclease.
rRNA. **RNA** ribossômico.
Rubisco. Proteína encontrada em abundância nos cloroplastos. (carboxilase do difosfato 1,5 da ribose).

S

Selecionismo. Pressuposição de que a maioria dos polimorfismos e substituições moleculares têm influência na adaptação dos organismos e estão, portanto, sujeitos à ação da seleção natural.
Sequenciamento. Obtenção da sequência de monômeros de um biopolímero.
Silenciamento. Supressão da expressão de um gene.
Simbiogênese. Origem a partir de simbiose.
Simbiose. Associação ecológica entre dois organismos de origens evolutivas diferentes.
Similaridade. 1. Semelhança. 2. Proporção de sítios idênticos (para sequências de macromoléculas).
Sinergia. Ação de dois ou mais fatores que resulta em um efeito maior que a simples soma de cada um dos fatores considerados isoladamente.
SINES. Sequências de microssatélites dispersas pelo genoma.
Sintetase. Enzima que catalisa a síntese de uma molécula a partir de outras duas ou mais moléculas menores.
siRNA. Moléculas curtas de RNA que têm a propriedade de participar no processo de interferência do RNA. ("small interfering RNA", em inglês).
Sítio ativo. Região de uma macromolécula onde ocore a catálise.
snoRNA. RNA pequeno nucleolar.
snuRNA. RNA pequeno nuclear.
Sociobiologia. Ciência que estuda os determinantes biológicos dos comportamentos sociais sob uma perspectiva evolutiva.
Southern blot (I.) Literalmente significa "mata borrão de Southern", devido ao pesquisador (Southern) que desenvolveu a técnica de transferência por capilaridade de fragmentos de DNA de um gel de eletroforese para uma membrana de nitrocelulose ou de nylon para que posteriormente seja hibridizada com sondas específicas de DNAcomplementar marcado.
Splicing (I.) Processamento de mRNA que consiste na junção dos segmentos correspondentes aos exons.
Superexpressão. Expressão de genes muito aumentada em relação ao seu nível normal.

T

Tautômero. Tipo de isômero que ocorre durante frações do tempo da existência de uma molécula.
Táxon. Conjunto dos organismos pertencentes a um dos níveis taxonômicos (ex. gênero, tribo, família).
Telomerase. Enzima da classe das transcriptases reversas que participa da síntese de sequências repetitivas de DNA encontradas nos telômeros.
Telômero. Estrutura presente nas extremidades dos cromossomos formada por DNA repetitivo.
Terapia gênica. Tecnologia que é usada no tratamento de doenças cujas causas são genéticas em que se procura alterar as causas das manifestações clínicas dos pacientes.
Termociclador. Aparelho que altera ciclicamente a temperatura de soluções utilizado para executar reações de amplificação de DNA.
Termófilo. Que vive em temperaturas altas.
Terpenoide. Ou terpenos. Classe diversificada de moléculas orgânicas derivadas do ácido mevalônico, presentes em plantas como metabólitos secundários.
Testosterona. Hormônio esteroide produzido pelos testículos dos machos de mamíferos.
Topologia. Estudo das propriedades das formas e de suas modificações.
Transcriptase reversa. Enzima que catalísa a polimerização de uma molécula de *DNA* a partir de um molde de *RNA*.
Transecto. Linha imaginária que liga dois pontos geográficos diferentes na superfície terrestre.
Transesterificação. Reação química onde os substratos são um álcool e um éster e os produtos são também um álcool e um éster nas moléculas de forma recíproca.
Transfecção. Processo pelo qual material genético é introduzido

passivamente para o interior de uma célula.

Transferase. Enzima que cataliza a reação de transferência de um grupo químico de uma molécula para outra.

Transição. Mutação onde uma purina dá lugar a outra purina ou uma pirimidina dá lugar a outra pirimidina.

Transmembrânico. Parte da estrutura de uma macromolécula que atravessa uma membrana lipoprotéica.

Transposase. Enzima que cataliza a reação de transposição do *DNA*.

Transposição. Reação de transferência de uma sequência nucleotídica para o interior de uma região do genoma.

Transposon. Sequência de nucleotídeos que tem a capacidade de mudar sua posição no genoma deixando ou não uma cópia no local original.

Transversão. Mutação onde uma purina dá lugar a uma pirimidina ou vice-versa.

tRNA. RNA transportador.

Trófico. Que diz respeito à alimentação de um organismo.

Tubulina. Proteína constituinte de microtúbulos.

U

Ultracentrifugação. Centrifugação a velocidades muito elevadas (acima de 50.000 rotações por minuto)

Ultraestrutura. Estrutura intracelular observada a partir de ampliações obtidas com microscópio eletrônico (5.000 X a 100.000 X).

UTO. Sigla de "unidade taxonômica operacional".

UV. Sigla de ultravioleta, emissão eletromagnética com comprimento de onda inferior a 340 nm.

V

Variância. Medida de dispersão de dados.

Variável. Em matemática, diz-se de um quantidade que pode ser qualquer valor que depende do contexto. Usa-se em contraposição a uma constante, cujo valor é sempre o mesmo.

Verossimilhança. 1. Adoção da hipótese mais provável. 2. Estimativa numérica da probabilidade.

Vivíparo. Modo de reprodução através de parto.

Voltagem. Diferença de potencial elétrico.

W

Western blot (I.) Processo de transferência por capilaridade de polipeptídeos de um gel de eletroforese para uma membrana de nitrocelulose ou de nylon para que posteriormente sejam detectados com sondas específicas que consistem em anticorpos marcados. O nome "Western" (do oeste) trata-se de um trocadilho feito a partir da analogia com o Southern blot, pois Southern significa, além do nome do autor da técnica, "do Sul", em inglês.

X

Xenologia. Homologia decorrente de transferência horizontal de material genético.

Xeroderma pigmentosum. Doença de pele onde há pigmentação acentuada e outras malformações com grande sensibilidade à luz solar.

Z

Zigoto. Célula diploide que se origina pela fecundação do óvulo pelo espermatozoide.

Página deixada em branco

Índice remissivo

A

A priori, probabilidade 147-150
AFLP 185, 218-219, 223, 228-229
Alelos, genealogia de 191, 194-195, 198, 201, 206-208
Algoritmos
  branch-swapping 117
  busca exaustiva 116-118
  busca min-mini 117
  decomposição de politomia 117
  evolução mínima 118, 128-129
  exatos 116-117
  Fitch-Margoliash 129
  heurísticos 116-117, 128, 136
  MCMC 147, 150-151, 155, 206-209, 211, 226
  MCMCMC 152
  stepwise addition 117-118
  troca de vizinhos próximos 117
Alinhamento de sequências 76, 78, 124, 126, 130, 133, 141-142, 145, 158-161, 163-164
Ames, teste de 52
Análise de clados hierarquizados 202-206
Ancestral comum mais recente 105, 114-116, 140, 193, 195, 207
Ancestral universal 21
Anisogamia 55-56
Argonauta (proteína) 41
Arqueobactérias 42, 80, 84
Árvores filogenéticas
  enraizadas e não enraizadas 114-116, 119, 137, 155
Atração de ramos longos 121
Auto-replicação 18, 23

B

Bactérias
  domínios 25-27, 50
  primitivas 13
Bayesiana, inferência filogenética 71, 123, 141, 147-155, 202, 206-207, 210, 226, 233
Biodiversidade
  estimativas de 227
  molecular 93, 97-98, 217-219, 229
  subestimativa 227, 229
Bioinformática 50, 82, 91, 101
Bioinvasões 217, 219, 227, 230-231
Bootstrap 127, 129-130, 145, 149-151, 158

C

Cadeia de Markov
  e coalescência 192
  e Monte Carlo 151-153, 207
Carcinogênese
  e mutagênese 51-52
Chips de DNA 101, 186
Citoesqueleto
  origem 27-28
Clados hierarquizados 201-205
Classificação, sistemas de 113
Clorofila **a** 27, 29-30
Cloroplastos 25-26, 28-29, 65-66, 68, 99-100, 173-174, 177-178
Coacervados 15-16, 17
Coalescência 191-196, 198, 201-202, 206, 209
Código genético
  origem 23-24
Códons
  origem dos 22
  utilização de 47, 62, 64-65, 93-95
Comprimento de ramos 67, 115, 121, 123-124, 126-129, 133-134, 136-138, 140-141, 147, 154-155
Conteúdo G+C 93-95

D

Daniel
  nós te amamos
Deriva genética 45, 61, 64, 70, 96-97, 118, 168, 178, 191, 197, 219, 222, 225
DICER (proteína) 40-41
Dimorfismo sexual
  determinação de 219
Distância
  genética 168, 205
  p 123-124, 127, 163, 204
  Gama-Poisson 125-126
  Jukes e Cantor 124, 134
  Kimura 2 parâmetros 67, 124, 126, 129-130
  PAM 126
  Tajima e Nei 125
  Tamura 3 parâmetros 125-126
  Tamura e Nei 125-127, 135 Distribuição gama 125, 138-139, 162
DNA
  microarrays 101, 186, 188
  microssatélite 98-99, 183-185, 188, 198, 204, 206-208, 210-211, 218-219, 222-223, 228-229, 231, 233
  minissatélite 94, 98-99, 184, 219, 224, 227-228
  satélite 97-99, 183-184
  sondas de 172, 176, 181, 184, 186-187 DNA de cloroplasto
  organização 100
  RFLP 173, 177-178
DNA-lixo 33, 37, 39, 99
DNA mitocondrial
  heteroplasmia 176
  haplótipos 177, 198
  organização do 176
  RFLP do 176
  taxas de evolução 66
DNA nuclear 33-34, 39,
  RFLP 171, 173, 178
Dogma central da biologia molecular 43
Ds, elemento 80
DS-PCR 186
Duplicação gênica
  e evolução de genomas 24, 26

E

Elementos transponíveis
  classes 84-89
  evolução dos 37, 42, 86
Eletroforese
  de fragmentos de DNA 166, 171-174, 183-184, 186-187, 197, 224
  de isozimas 165, 167-169, 171
ENCODE, projeto 34, 92
Endocruzamento 168, 209, 219-223, 225, 227, 231
Endonuclease AP 48
Endonucleases de restrição 171
Endossimbiose 25, 27-30
Enzimas de restrição
  e metilação 41, 46, 49, 53, 171, 173
  extremidades coesivas 173, 184
  extremidades retas 173
  isoesquizômeros 173
  sítios de reconhecimento 23, 173, 183
Escolha de genes
  para problemas filogenéticos 158, 161-162, 169, 218-219
Espécies
  raras 219
  ameaçadas e estrutura gênica 228-229
Estados ancestrais 134, 155
Estela
  eu te amo
Estrangulamento populacional
  e coalescência 208, 217
  efeitos na variabilidade 220-222
Estromatólitos 13, 25
Estrutura genética 167
  e coalescência 195
  e conservação 225-226
  estimativas de 168
  e filogeografia 198-199
  isolamento por distância 167, 202-206
  modelo de ilhas 198, 226-227
  modelo passo a passo 226
  panmixia 198, 203, 210
Eubactérias 22, 24-26
  e elementos tranponíveis 80
Eucariotos
  origem 25-29
Evolução
  das mitocôndrias 26-28
  de alelos novos 45, 53, 61
  de elementos transponíveis 86-89
  de introns 23
  do código genético 23-24
  dos cloroplastos 26-29
  do RNA não codificante 34-36
  dos genes nucleares 64-66
  dos genomas 91-101
  dos sistemas de reparo 53-53
  e mutagênese 53
  em concerto 96-98,159
  in vitro 24
  taxas de 61-71, 120
Experimento de Miller 14

F

Família *En/Spm* de elementos 83-84
Família *hAT* de elementos 83
Família *mariner-Tc1* de elementos 83
Família *Mu* de elementos 81, 83
Família *p* de elementos 81, 83
Famílias gênicas 26, 65, 93, 95-97
  parálogas 97, 159, 162
Filogeografia 197-212
  estatística 201-210
  limitações 211-212
  programas 212
Fósseis 13, 16
  moleculares 21, 23, 25
  calibração de relógio molecular 71, 154
Fotossíntese 27-28
Função de genes
  efeito em filogenias 162
Funções novas
  origem das 98, 109

G

Gama (distribuição) 125-126, 138-139, 158, 162
Gargalo evolutivo 168
  efeitos na variabilidade 168
Genes de RNA ribossômico 75-77
  cópias múltiplas 159
  evolução 75-79
  organização no genoma 75, 95
Genética da conservação 198, 207, 211-212, 217-233
Genoma
  não codificante 98
Genomas de organelas 65-66
  evolução 88
Genomas
  bancos de dados 91-92
  evolução 91-102
  mecanismos de evolução 96-99
  tamanho dos 92-93
Genômica
  e RNAi 41
  funcional 100-101
  perspectivas 101
Grupo externo
  e enraizamento de árvores 119
  para teste de taxas relativas 67

Grupos merofiléticos 115
Grupos parafiléticos 115
Grupos polifiléticos 115

H
Haplótipos 177, 191, 194, 197, 198-199
Heterozigosidade
e conservação 219-223, 230
estimativas de 167-168
perda de 208
HIV
evolução 63
máxima verossimilhança e 142
Homeóticos, genes 30, 95, 107
Homologia 98, 105, 109, 158-159
primária 158
profunda 110
Hox, genes 107-108

I
Identificação forense 99, 157
e conservação 231-233
Indels 61-62, 161, 173
Inferência Bayesiana
de filogenias moleculares 147-155
estados ancestrais 155
relógio molecular 153-154
Introns
autocatalíticos 21-25
precoces 24
Invariantes de Lake 121
Inversões cromossômicas 118
Isócoros 66
hipóteses neutralista e adaptacionista 94
Isogamia 55

L
Lara
eu te amo
Lesões do DNA 45-51, 53
reparo 48-51

M
Mapas
de restrição 173-176-179, 183
genéticos 95, 99
Marcadores moleculares 169, 188
comparação entre os 218-221
obtidos com PCR 181-189
Matriz de distâncias 127
Matriz de pesos 120
Matriz de substituição 134-135
Matriz de transição 126
Máxima parcimônia 116-122
pesos e 120
princípio 118-119
vantagens e desvantagens 121-122
variações da 119-121
Máxima verossimilhança 133-143
atribuição de migrantes 211
coalescência 194
e filogeografia estatística 206-207
e inferência bayesiana 147-148, 152-154
estados ancestrais 155
histórico 133-134
modelos de evolução e 134-136
na calibração de relógio molecular 71, 139-141
princípio 134
programas 145
seleção positiva 154-155
vantagens e desvantagens 126, 142-143
Meiose
e reparo do DNA 50
origem da 58-59
Meteoros
e origem de matéria orgânica 15
MHC
polimorfismos do 155-156, 221-223
Microfósseis 13-14
Migração 198
de genes para organelas 100
e coalescência 195
estimativa 168, 207-211
Miller, experimento de 14
Mitocôndrias
genoma das 65
origem das 26-29
Mobilização de elementos
controle 84-86
Modelos de substituição
F81 135
F84 135
GTR 135-136
HKY85 135
Jukes e Cantor 124, 134
Kimura 2 parâmetros 134
proporcional 135
TN93 135
Modularidade 108-109
Monofilia 115
Mundo de DNA 24-25
Mundo de RNA 21-24
Mundo de RNP 23-24
Mutações 45-49
e carcinogênese 52
pontuais 47-48, 61, 177
sinônimas e não sinônimas 63, 66, 68
sem sentido 47, 61
taxas 49, 51, 53, 57, 58, 63, 186, 194-195, 220
Mutagênese 45-48
teste de 52
Mutases 53

N
Navalha de Ockham 117
ncRNA 34-38
Nível de confiança
em filogenias 130
Northern blot 101, 172
Núcleo da célula, origem 26-27

O
Ockham 117
Organelas
genomas 65
origem 25-30
taxas de evolução 99
Origem da vida 13-19
extraterrestre 16-17
Ortólogos,
genes 98, 107, 159
OTUs 114-115

P
Panspermia 17
Parálogos,
genes 97, 159
Parcimônia
princípio da máxima 118-119
Partenogênese 57, 224
PCR 91, 181-187, 232
e RFLP 178, 219
Plesiomorfia 113, 115-116, 119
Pontos quentes
mutacionais 53, 121
recombinacionais 98
Posterior
distribuição 207
probabilidade 207
Probabilidade de uma árvore 136, 139
Progenota 18, 21, 25
fotossíntese do 27
genes do 26
Projeto ENCODE 34, 92
Projetos Genoma 26
Proteômica 100-101
bancos de dados 92
Protocélula artificial 18
Pseudogenes 65
conteúdo G+C 64
e famílias gênicas 95-96, 98
taxas de evolução dos 69, 139-140, 159, 176

Q
QTLs 188-189
Quebra cabeça de quartetos 137
Quebras cromossômicas
causadas por introns 79-80
Quiralidade, origem 15

R
Rainha vermelha, hipótese da 57, 222
Razão de verossimilhança
testes de 140-141, 153, 155
Recombinação 65
desigual 95
e coalescência 195
e elementos transponíveis 85
e reparo do DNA 50-51
e sexo 55-57
pontos quentes 98
Reconstrução filogenética
e classificação 116
histórico 114
algoritmos para 116-118, 123, 127-129, 136-138, 141, 145, 151-152
Regiões codificadoras
constância de tamanho 34
interrompidas 22
proporção 42
Regiões não codificadoras 33-38, 39
Relógio molecular 66-71, 139-141
bases 61
calibração 129
local 67
teste de taxas relativas 140
bayesiano 141, 153-154
programas 145
relaxado 208-209
Reparo do DNA 45-54
de emparelhamento errado 46
direto 49-50
e doenças genéticas 5.4
e isócoros 65
e sexo 58-59
e taxas de evolução 68-70, 162, 176
evolução dos sistemas de 24, 52-53
por excisão de bases 48-49
por excisão de nucleotídeos 50
recombinacional 50-51
sujeito a erro 51
Replicação
com fidelidade reduzida 51
do DNA 33, 45-46, 53, 65
do RNA 18, 23
e coalescência 193
e taxas de evolução 68-70
Replicadores precoces 18
Reprodução assexuada 55-57, 219
Reprodução sexuada 55-58
consequências da 56
Retrotransposons 81-87
diversidade 82
vegetais 82-83
e cérebro 37
e embaralhamento de exons 26
e seleção natural 88
e tamanho de genomas 93
e variação 88-89
estrutura 81
Ribossomo
e síntese protéica 23
origem 24
RFLP 171-179
análise de biodiversidade 219
de DNA de cloroplastos177-178
de DNA mitocondrial 176-177
de DNA nuclear 178
de produtos de PCR 178-179, 183
emprego em conservação
mapas de restrição 174-176
princípios 172-173
reprodutibilidade 173-174
RISC (proteína) 40-42
Ritmos da evolução molecular 71
RNA catalítico 21-25, 27, 33, 39-40, 78
RNA não codificante 33-38, 39
miRNA 39
RNAi 39
siRNA 39
RNA nucleolar 24, 75
RNA ribossômico 25, 28, 39
estruturas secundárias 76-78

evolução 75-78
RNP 23-25

S

Segmentos de expansão 76-77
papel funcional 77
Seleção balanceada 165, 225
Seleção estabilizadora 169
Seleção natural 17-18, 37, 45, 52-53, 59, 61, 64-65, 88, 95, 168, 169, 178, 195, 217, 219, 220, 222-223, 225
Seleção positiva 45, 52, 53, 63, 78, 154-155, 169
Sexo
determinação 56
e reprodução 55
Simbiogênese 27-28
Síntese pré-biótica 14-15
Sistemática filogenética 113-114
Sítios informativos 119, 121
SNPs 186
Sondas de DNA 172, 184, 186
heteroespecíficas 172,176
homólogas 172
Sopa primitiva 14-16, 23
Southern blot 172
SPAR 185
SSR 184-186
STR 184
Sucesso reprodutivo 110, 195, 222-224

T

Tamanho efetivo
de populações 191-192, 217, 219-221
Tamanhos de ramos 67, 115, 121, 123-124, 126-129, 133-134, 136-138, 140-141, 147, 154-155
Taxas de erro 46, 53
Taxas de expressão 64-65
Taxas de evolução 53, 61-71, 120-121, 154, 157, 159
Taxas de migração, estimativa 168, 207-211
Taxas de mutação 49, 51, 53, 57, 58, 63, 186, 194-195, 220
Taxas metabólicas 69-70
Taxas relativas 67-68
Taxas de substituição 61, 62-69, 87, 94-95, 120-121, 123-124, 127-128-129, 133-136, 147, 149, 152, 154, 161-162
em organelas 65-66, 176-178
heterogeneidade 138-141
Tempo de geração 191-192
hipótese de 68-70, 162, 207
Teste de Ames 52
Testes de hipóteses
comparação entre modelos evolutivos 141
comparação entre topologias 141
critério de informação de Akaike 141
Kishino e Hasegawa 141-142
razão de verossimilhança 140-141, 153, 155
Transcriptoma 101
Transferência lateral de genes 26, 28-29, 86-87, 92, 97, 99, 159
Transições e transversões
definição 47, 61
distâncias 124-125, 129
inferência bayesiana 147, 149
máxima parcimônia 120-121
máxima verossimilhança 134-135
mutações 47
taxas 62-63, 68, 139, 158

U

Unidades taxonômicas operacionais 114, 134, 138

V

Variegação
causada por íntrons 79
Vida artificial 17-18
Vida, definição de 17

W

Western blot 72

X

Xenologia 159

Página deixada em branco