Parte 5

Era da Genômica

11 Manipulando o Gene | Técnicas de Biologia Molecular

Rodrigo da Silva Galhardo e Cristina Elisa Alvarez Martinez

Neste capítulo serão descritos os fundamentos das principais técnicas de biologia molecular. Essas metodologias são rotineiras em laboratórios de pesquisa, bem como em investigações clínicas e forenses. Frequentemente, os métodos de manipulação de DNA e RNA *in vitro* envolvem o uso de enzimas purificadas, oriundas de diversos organismos. Primeiramente, serão apresentadas as metodologias básicas de manipulação do DNA, que tornam possível a separação e a clonagem de fragmentos de DNA ou cDNA. Posteriormente, serão explicadas as metodologias que permitem a amplificação e o sequenciamento de DNA em pequena e em larga escala, assim como o estudo da expressão gênica. Algumas aplicações práticas, como a identificação humana por meio da análise de DNA, também serão abordadas.

Introdução

Até a metade do século 20, a manipulação de genes individuais ou de pequenos trechos regulatórios do DNA parecia uma tarefa intransponível para os geneticistas. Tal fato se deve principalmente à natureza da molécula de DNA e da informação nela contida. Apesar de os diferentes genes atuarem como entidades funcionalmente individuais, estes não são isolados em unidades fisicamente discretas nas células. Isso porque eles se encontram como componentes de uma grande molécula de DNA, que contém inúmeros outros genes e regiões não codificantes. No entanto, uma verdadeira revolução nos permitiu compreender melhor a estrutura e a função dos ácidos nucleicos, a partir das décadas de 1970 e 1980, ao ponto de, atualmente, ser comum isolar e estudar um trecho de DNA em particular. Esses avanços deram origem à disciplina à qual hoje nos referimos como biologia molecular, que nada mais é que a incorporação de metodologias de bioquímica e biofísica ao estudo da genética.

Curiosamente, as ferramentas que proporcionaram tal revolução não foram inventadas pelo ser humano, e sim originários da própria natureza. São enzimas provenientes majoritariamente de microrganismos, envolvidas nos processos fisiológicos básicos ligados ao metabolismo de DNA, tais como DNA polimerases, DNA ligases e nucleases. Essas enzimas podem ser purificadas em grandes quantidades e usadas para manipular DNA em um tubo de ensaio.

Neste capítulo, será abordado o conjunto básico de metodologias mais usadas em estudos de biologia molecular que proporcionaram (e continuam proporcionando) uma verdadeira revolução na biologia moderna. Contudo, é preciso considerar que é impossível descrever em detalhe todas as metodologias atualmente existentes e toda a gama de aplicações possíveis para elas. Pretende-se aqui lançar uma base sólida para o entendimento das ferramentas básicas de biologia molecular e algumas de suas aplicações mais comumente empregadas. Portanto, parte-se das técnicas mais fundamentais para separação, clonagem, detecção e amplificação de ácidos nucleicos. Em outros capítulos, serão apresentados alguns exemplos de experimentos nos quais essas técnicas fundamentais são empregadas para desvendar a função de genes em sistemas biológicos.

Enzimas de restrição | Ferramentas básicas da engenharia genética

O primeiro grande passo em direção à biologia molecular moderna foi, sem dúvidas, a descoberta das enzimas de restrição. A importância dessas enzimas como ferramentas de biologia molecular é tão grande que seus principais descobridores – Daniel Nathans, Werner Arber e Hamilton Smith – foram agraciados com o Prêmio Nobel de Medicina e Fisiologia em 1978.

Tais enzimas são endonucleases, ou seja, capazes de promover a clivagem de uma ligação fosfodiéster do DNA; elas estão presentes em bactérias e fazem parte do sistema chamado "restrição-modificação". A função biológica desse sistema é proteger as bactérias contra o ataque de bacteriófagos, pois as endonucleases de restrição clivam o DNA viral quando este entra na célula, impedindo que o bacteriófago inicie um ciclo infeccioso. A lógica desse sistema consiste em dois fatores. Em primeiro lugar, essas enzimas são altamente específicas: clivam o DNA apenas quando ele apresenta uma determinada sequência de nucleotídeos em particular, comumente chamada de *sítio de restrição*. Em segundo lugar, o reconhecimento dessa sequência é afetado pela metilação do DNA. Assim, a mesma bactéria que produz uma determinada enzima de restrição que reconhece a sequência "X" no DNA viral também protege o seu próprio DNA, por produzir uma enzima chamada metilase, impossibilitando o reconhecimento desse mesmo sítio "X" no seu genoma pela endonuclease. Desse modo, essas enzimas funcionam como um mecanismo de defesa para as bactérias que as contêm. A nomenclatura dessas enzimas se dá de acordo com o microrganismo do qual cada uma provém. Por exemplo, a enzima *Eco*RI foi isolada de *Escherichia coli*, e a enzima *Pst*I é produzida pela bactéria *Providencia stuartii*.

Como mencionado anteriormente, tais enzimas reconhecem sequências específicas no DNA. Existem três tipos de enzimas de restrição (tipos I, II e III). As enzimas de restrição do tipo II cortam o DNA em uma posição definida, dentro do próprio sítio de reconhecimento; até o momento, milhares de enzimas dessa classe já foram identificadas. Já as enzimas dos tipos I e III reconhecem o seu sítio específico, mas cortam a região em seu entorno de maneira aleatória (ou até mesmo a uma grande distância do sítio de reconhecimento) e, por essa razão, apresentam menor utilidade como ferramenta de biologia molecular.

Existe uma grande variedade de sítios de restrição e tipos de corte

Os sítios de enzimas de restrição apresentam características específicas. A maioria das enzimas reconhece sequências de 6 nucleotídeos, mas muitas outras reconhecem sequências de 4 ou 8 nucleotídeos. Algumas enzimas de restrição e seus sítios de reconhecimento estão ilustrados na Figura 11.1.

Embora haja algumas variações, a maioria das enzimas de restrição funciona de maneira semelhante. Primeiramente, os sítios de restrição costumam ser sequências palindrômicas diretas (p. ex., o sítio da enzima *Eco*RI, mostrado na Figura 11.1). A sequência 5'-GAATTC-3' é uma palíndrome, pois a sequência complementar que lida no mesmo sentido (5'-3') é idêntica. Outra característica das enzimas de restrição é a especificidade do corte. A ligação fosfodiéster que é clivada em uma fita é também a região clivada na outra fita da palíndrome. Consideremos as enzimas mostradas na Figura 11.1. A enzima *Eco*RI corta o seu sítio 5'-GAATTC-3' sempre da mesma maneira: entre o primeiro e o segundo nucleotídeo no sentido 5'-3' em ambas as fitas. O DNA resultante desse corte contém extremidades ditas *coesivas*. Tal denominação ocorre justamente pelo fato de que essas pontas podem se reassociar pela formação de pontes de hidrogênio entre as bases nitrogenadas das extremidades protuberantes. No caso de DNA cortado com *Eco*RI, ambas as pontas têm uma extremidade protuberante 5' de 4 nucleotídeos. Outras enzimas geram pontas coesivas diferentes, com protuberância 3' de 2 nucleotídeos, ou protuberâncias 5' de 2 ou 4 nucleotídeos. A enzima *Pst*I, mostrada na Figura 11.1, deixa extremidades coesivas com protuberância 3'. Existem outras enzimas que cortam exatamente no centro do palíndromo, tal como *Sma*I (Figura 11.1). Neste caso, as pontas de DNA geradas não apresentam protuberâncias, e são chamadas de pontas *cegas*.

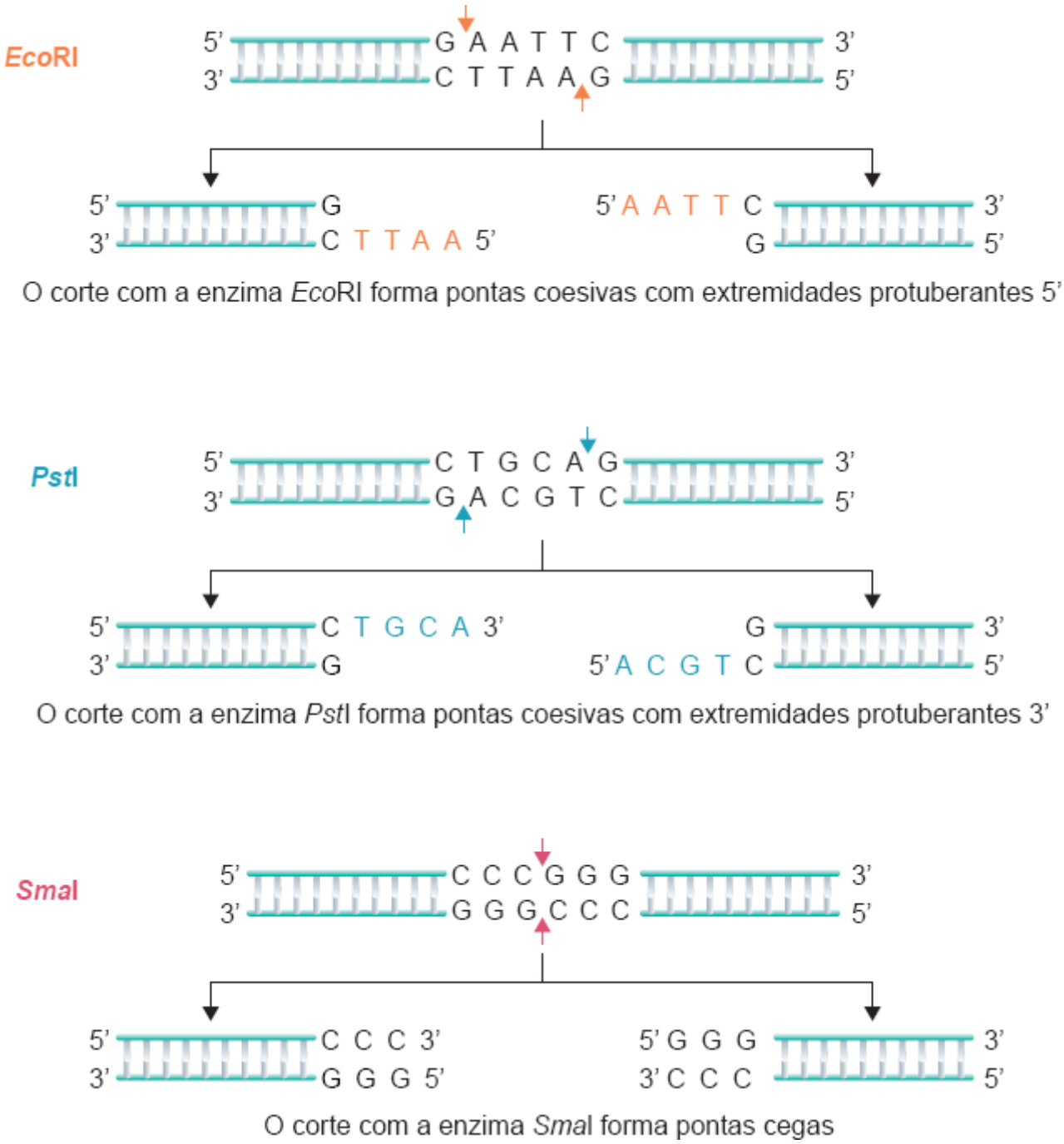


Figura 11.1 Mecanismo de ação de enzimas de restrição. A figura mostra os sítios de reconhecimento e o corte promovido no DNA pelas enzimas *Eco*RI, *Pst*I e *Sma*I.

A frequência de sítios de restrição é variável

Imagine uma sequência de DNA aleatória composta de proporções iguais de A, C, G e T. A probabilidade de um sítio qualquer de 6 nucleotídeos ocorrer nessa sequência é ¼ × ¼ × ¼ × ¼ × ¼ × ¼; ou seja, ¼.096. Isso quer dizer que há 1 sítio a cada 4 mil pares de base (pb), em média. Seguindo o mesmo raciocínio, enzimas que reconhecem sequências de apenas 4 nucleotídeos irão cortar, em média, 1 sítio a cada 256 pb, e enzimas que reconhecem sítios de 8 nucleotídeos irão cortar uma vez a cada 65.000 pb. No entanto, a sequência do DNA nos cromossomos dos seres vivos obviamente não é aleatória, e nem sempre moléculas de DNA apresentam igual proporção dos 4 nucleotídeos. Assim, até mesmo entre as enzimas com sítios de reconhecimento de mesmo tamanho existe uma grande variedade na frequência desses sítios em diferentes genomas. Logo, há flexibilidade no uso dessas ferramentas e, dependendo da aplicação, pode-se desejar usar uma enzima que corte com maior ou menor frequência o DNA em estudo.

Enzimas de restrição e os primeiros mapas genéticos moleculares

As enzimas de restrição são ferramentas básicas no processo mais fundamental da biologia molecular: a clonagem (discutida mais adiante neste capítulo). Contudo, cabe ressaltar que essas enzimas proporcionaram também outro avanço na genética: elas tornaram possível a construção dos primeiros *mapas físicos* de moléculas de DNA. Até a década de 1970, os mapas de cromossomos de organismos bem estudados eram *mapas genéticos*. No Capítulo 19, é mostrado em detalhes o modo como mapas genéticos humanos são obtidos. Basicamente, os mapas genéticos têm como base a posição relativa de genes que estão ligados (*i. e.*, no mesmo cromossomo), e dependem de dois fatores para sua construção: genes que produzem um fenótipo detectável quando alterados e análise de cruzamentos (ou análise de história familiar, em caso de humanos). Já os mapas físicos baseiam-se em características da molécula de DNA em si, ou seja, na sua composição de nucleotídeos.

Ácidos nucleicos de tamanhos diferentes podem ser separados por eletroforese

Antes de continuar a discussão sobre as ferramentas utilizadas em biologia molecular, vale apresentar um método fundamental para todas as demais técnicas: a eletroforese em gel. Essa técnica é utilizada para a análise de amostras de DNA de diversas origens, tais como plasmídeos e vírus, e para separar fragmentos de DNA obtidos após digestão com enzimas de restrição ou em vários outros procedimentos de biologia molecular que serão discutidos nas próximas seções. A eletroforese baseia-se na separação de moléculas de acordo com sua carga total. Assim, uma amostra qualquer a ser analisada é submetida a um campo elétrico, e as moléculas com carga total positiva ou negativa migram para o ânodo (polo negativo) ou cátodo (polo positivo), respectivamente. A eletroforese para a análise de amostras de DNA é realizada em fase sólida, utilizando-se uma matriz porosa chamada de *gel*. Uma vez que o DNA tem carga total negativa (pela presença do grupo fosfato na cadeia fosfodiéster externa), sua migração durante a eletroforese em gel depende apenas de conformação (linear ou circular, relaxada ou superenovelada) e de seu tamanho. As amostras de DNA a serem analisadas são colocadas em poços formados no gel, e a corrente elétrica é aplicada na presença de uma solução tampão, o que torna possível que o DNA migre através dos poros do gel na direção do cátodo (Figura 11.2). Quando o DNA está na conformação linear, sua migração passa a depender apenas de seu tamanho, possibilitando separar e identificar o tamanho de diferentes fragmentos de DNA em uma amostra. Para isso, também é colocada no gel uma amostra de DNA com fragmentos de tamanho conhecido, chamada de *marcador de peso molecular* (Figura 11.2). Após migração em gel, o DNA pode ser visualizado pelo uso de um corante fluorescente capaz de ligar-se ao DNA, como o *brometo de etídeo*, uma molécula que se intercala entre as bases nitrogenadas da fita de DNA, gerando fluorescência facilmente visualizada sob luz ultravioleta (Figura 11.2).

A separação de ácidos nucleicos por eletroforese pode ser feita em géis de agarose ou de poliacrilamida

De modo geral, a eletroforese torna possível separar fragmentos de DNA na faixa de 10 pb até 50 Kpb. No entanto, a matriz a ser utilizada varia de acordo com o grau de separação e a faixa de tamanho que se deseja analisar, podendo ser *agarose* ou *poliacrilamida*.

A agarose é um polissacarídeo que forma poros de 100 a 300 nm de diâmetro e é bastante utilizada para analisar amostras de DNA até 10 Kpb, sendo a mais comumente usada no dia a dia de um laboratório de biologia molecular. Além disso, é possível variar a concentração de agarose no gel de acordo com a faixa de tamanho do DNA que será analisado: géis contendo maiores concentrações de agarose formam poros menores, sendo usados para a análise de fragmentos pequenos; géis com menores concentrações de agarose formam poros maiores, que facilitam a migração de fragmentos grandes.

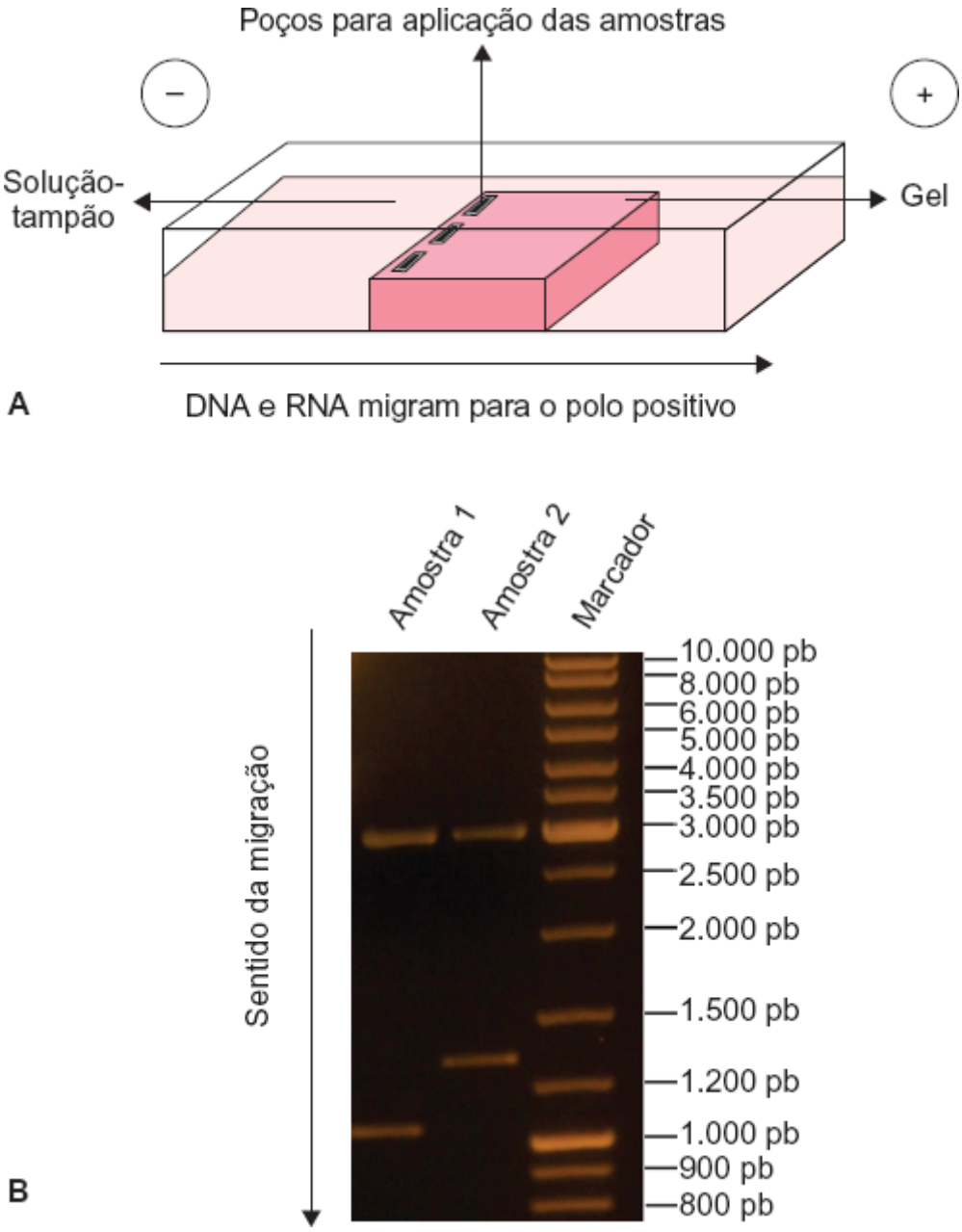


Figura 11.2 Separação de DNA por eletroforese em gel de agarose. **A.** Esquema de apa rato utilizado para a corrida eletroforética horizontal de DNA e RNA. **B.** Visualização sob luz ultravioleta de fragmentos de DNA depois da separação em gel e coloração com brometo de etídeo. Pela comparação com o marcador de peso molecular, é possível determinar os tamanhos dos fragmentos obtidos após digestão com enzimas de restrição.

No entanto, quando a amostra de DNA a ser analisada é composta por fragmentos pequenos, menores que 100 pb, ou quando se deseja obter um grau de separação maior de fragmentos de até 1.500 pb, utiliza-se a *poliacrilamida*. Nesse caso, o gel é composto por uma mistura de dois monômeros: acrilamida e bisacrilamida, e a proporção de cada uma delas, juntamente com a concentração final dos monômeros, determina o tamanho dos poros e, consequentemente, o grau de separação que será obtido. Esse tipo de gel torna possível identificar diferenças de apenas um nucleotídeo entre diferentes amostras, o que é bastante útil, por exemplo, para o sequenciamento de DNA.

Grandes trechos de DNA podem ser separados por eletroforese em campo pulsado

Fragmentos de DNA na faixa de dezenas de Kpb não são capazes de penetrar pelos poros formados pela agarose na eletroforese convencional e, por isso, migram a partir de uma das extremidades da molécula, independentemente do seu tamanho. Desse modo, não é possível obter um grau de separação satisfatório das moléculas nessa faixa de tamanho por meio da eletroforese convencional. Assim, a *eletroforese em campo pulsado* foi desenvolvida para possibilitar a análise precisa de grandes fragmentos de DNA. Em vez de se utilizar de um campo elétrico unidirecional, na eletroforese em campo pulsado a corrente elétrica é aplicada em pulsos provenientes de duas

direções, de maneira a causar um movimento em zigue-zague das moléculas de DNA. Assim, cada vez que a direção do campo elétrico é alterada, as moléculas são forçadas para a nova direção, o que ocorre de modo mais rápido com moléculas menores, facilitando a separação por tamanho dos diferentes fragmentos de DNA presentes na amostra (Figura 11.3). Essa variação na técnica de eletroforese tornou possível a análise de cromossomos bacterianos inteiros e até de cromossomos de eucariotos superiores, desde que previamente cortados em fragmentos menores, pelo uso de enzimas de restrição que reconhecem poucos sítios na molécula.

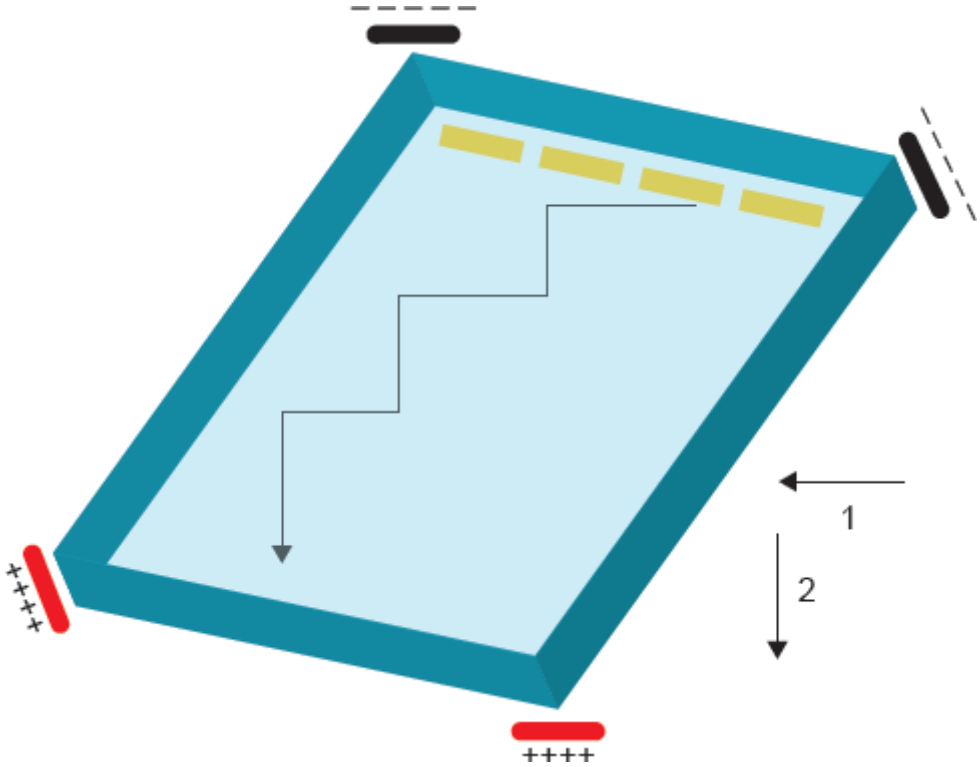


Figura 11.3 Esquema de funcionamento da eletroforese em campo pulsado. O campo elétrico é continuamente alternado entre as orientações 1 e 2, causando a reorientação do DNA e o seu movimento em zigue-zague, conforme indicado pela seta.

A eletroforese em gel também é comumente utilizada para a análise de amostras de RNA. A molécula de RNA forma inúmeras estruturas secundárias que interferem na análise de tamanho por eletroforese em gel. Assim, o RNA deve ser tratado com um agente desnaturante, geralmente o *formaldeído*, antes de ser aplicado no gel.

Fragmentos de DNA separados por eletroforese podem ser recuperados de géis

É interessante notar que a eletroforese é uma técnica analítica que possibilita separar vários fragmentos de DNA de uma mistura complexa e determinar seus tamanhos; no entanto, também pode ser usada para obter purificações de um DNA de interesse. A purificação de fragmentos de DNA a partir de amostras separadas por eletroforese é uma tarefa rotineira em laboratórios de biologia molecular. Imagine que você deseja obter uma preparação de apenas um único fragmento obtido por digestão de uma molécula de DNA com enzimas de restrição, sendo que essa digestão gerou quatro outros fragmentos de tamanhos diferentes. Após a eletroforese, o fragmento de interesse pode ser recuperado do gel, pois é possível recortar a banda do gel e purificá-la por meio de metodologias bioquímicas simples. Esse procedimento é fundamental em vários experimentos de biologia molecular, isto que, além de separar um fragmento específico dos demais, também purifica o DNA, livrando-o da presença de enzimas e sais usados em experimentos anteriores.

Tecnologia do DNA recombinante e clonagem de genes

Para estudos de biologia molecular, é desejável que se possa obter quantidades ilimitadas do fragmento de DNA de interesse, de modo que não seja necessário obter nova preparação de DNA genômico, digestão e purificação da região em estudo. Isso se tornou possível com o advento da *tecnologia do DNA recombinante*, mais popularmente conhecida como *engenharia genética*, que permitiu que fragmentos de DNA fossem clonados – ou seja,

reproduzidos em milhões de cópias idênticas. Um desdobramento da clonagem é a construção de *bibliotecas genômicas*, nas quais não se clona apenas um fragmento de interesse e, na verdade, todo o DNA de uma célula é fracionado em pequenos fragmentos e vários clones diferentes são produzidos.

Agora que já se conhece o funcionamento das enzimas de restrição, é possível imaginar que um gene (ou qualquer outra parte de uma sequência de DNA que seja de interesse para estudo) pode ser separado do restante da molécula de DNA por meio da clivagem com uma ou mais enzimas de restrição. No jargão da biologia molecular, essa clivagem é chamada de *digestão* com enzimas de restrição; tomemos como exemplo uma molécula de DNA relativamente simples, tal como o DNA do fago lambda, que tem 48,5 Kpb. É fácil imaginar que a digestão desse DNA e a posterior separação dos fragmentos resultantes por eletroforese isolem um ou mais genes de interesse do restante do DNA. Para genomas maiores, essa tarefa é mais difícil, pois muitos fragmentos de DNA do mesmo tamanho são gerados pela digestão com cada enzima de restrição. Nesse caso, a construção de bibliotecas genômicas ou a reação em cadeia da polimerase (PCR) são frequentemente empregadas, como será visto adiante. Contudo, é necessário introduzir duas ferramentas fundamentais no processo de clonagem: as DNA ligases e os vetores genéticos.

Uso de DNA ligases em biologia molecular | Novas combinações de sequências de DNA

Assim como as enzimas de restrição, as DNA ligases são enzimas que desempenham importante função na natureza, por realizarem uma função crucial nos processos de replicação e reparo de DNA. A reação que essas enzimas catalisam pode ser colocada em termos simples: a junção de uma extremidade 5'-P livre com outra extremidade 3'-OH livre (Figura 11.4).

Imagine dois DNA de origens diferentes, como ilustrados na Figura 11.5. Ambos foram cortados com a enzima *Eco*RI e, portanto, apresentam as extremidades coesivas, com protuberâncias 3' características. Como a enzima *Eco*RI deixa sempre as mesmas pontas coesivas complementares no DNA, estas podem se associar transientemente pelo emparelhamento de bases entre as protuberâncias em fita simples. Misturando-se os dois fragmentos de DNA de origens diferentes em um tubo de ensaio e adicionando a enzima DNA ligase, é possível produzir uma molécula de DNA *quimérica*, ou *recombinante* (Figura 11.5). Esse tipo de experimento é base fundamental da *tecnologia do DNA recombinante* e possibilita a construção de novas combinações de sequências de DNA. O emparelhamento de poucas bases proporcionado pelas pontas coesivas deixadas por enzimas de restrição aumenta bastante a eficiência das reações com a DNA ligase *in vitro*, por possibilitar uma associação transiente, mas específica, entre as moléculas. A DNA ligase também é capaz de ligar duas pontas cegas de DNA, tais como aquelas produzidas pela ação da enzima *Sma*I (Figura 11.5). Contudo, a eficiência desse processo é menor que a obtida na ligação de duas pontas coesivas compatíveis.

Vetores permitem a clonagem de genes

Conforme descrito anteriormente, é desejável que um fragmento de DNA em estudo possa ser clonado; para isso, utilizam-se os chamados vetores. De maneira geral, é possível descrever os vetores mais usados para clonagens como moléculas de DNA capazes de se replicar de maneira independente do restante do genoma, dentro de organismos-modelo de fácil manuseio em laboratório (tais como bactérias e leveduras). Por meio do uso da tecnologia do DNA recombinante, fragmentos de DNA de interesse podem ser inseridos em vetores, em uma reação catalisada por uma DNA ligase. Os vetores modificados pela presença do inserto de interesse podem ser reintroduzidos na célula hospedeira, na qual se mantêm de maneira estável, multiplicando-se ao longo das gerações deste organismo.

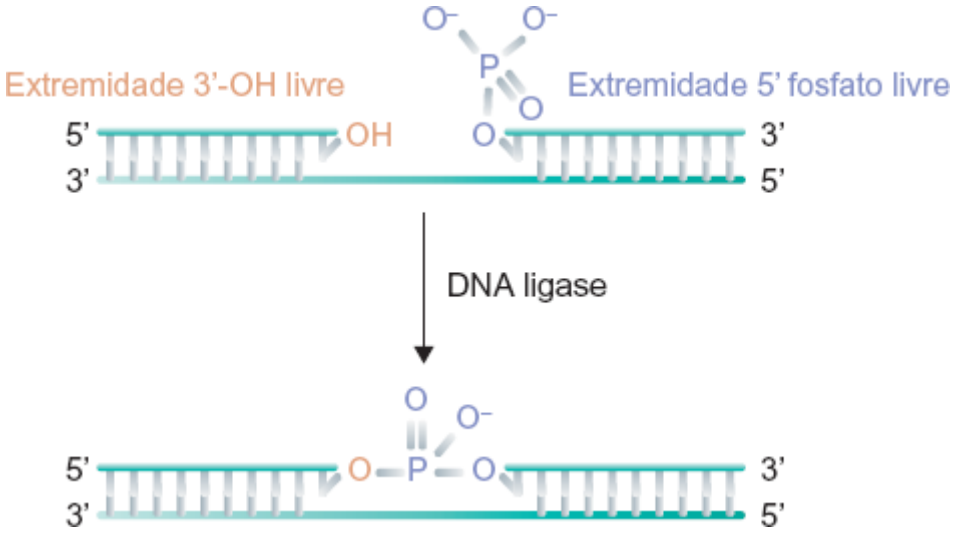


Figura 11.4 Mecanismo simplificado de ação da DNA ligase, promovendo a formação de uma ligação fosfodiéster entre extremidades 3'-OH e 5'P livres.

Os plasmídeos bacterianos são os principais vetores para clonagem

Os vetores de escolha para a maioria dos experimentos de clonagem de trechos pequenos de DNA (de até 10 Kpb) são os plasmídeos da bactéria *E. coli*. Fragmentos de DNA maiores são geralmente instáveis em plasmídeos, podendo ser clonados em outros vetores.

Plasmídeos são moléculas de DNA de replicação autônoma que ocorrem naturalmente em bactérias, apresentando, em sua grande maioria, conformação circular. Os vetores mais utilizados em biologia molecular derivam desses plasmídeos naturais, mas foram modificados extensamente pelos pesquisadores para apresentarem uma série de características desejáveis. As moléculas de DNA plasmidial utilizadas em biologia molecular são relativamente pequenas, e podem ser separadas do DNA cromossômico e isoladas das células por meio de técnicas bioquímicas relativamente simples. Algumas das características comumente encontradas nesses vetores estão listadas a seguir e representadas pelo exemplo da Figura 11.6:

- Presença de um ou vários sítios de restrição: muitos dos vetores mais modernos têm uma região inserida artificialmente que contém um grande número de sítios únicos (ou seja, que só ocorrem uma vez na molécula do plasmídeo), chamada de *sítio múltiplo de clonagem*
- Presença de um *marcador selecionável*, que possibilita uma seleção positiva (ou seja, um crescimento seletivo) das células que apresentam o plasmídeo. Os marcadores selecionáveis mais utilizados são genes de resistência a antibióticos. Assim, apenas bactérias portadoras do plasmídeo podem crescer em meio de cultura contendo o antibiótico em questão
- *Genes repórteres* que possibilitam a distinção entre bactérias que carregam plasmídeos recombinantes (isto é, com inserto) daquelas que têm o plasmídeo "vazio". Nem todos os plasmídeos usados em clonagem apresentam essa característica, mas ela está presente nos mais populares. A maioria dos vetores mais usados tem parte do gene que codifica a enzima betagalactosidase, *lacZ*. A atividade desta enzima pode ser facilmente detectada em colônias de bactérias, se o meio de cultura for adicionado de um substrato cromogênico. Quando a enzima está presente, este substrato é metabolizado, dando cor azul às colônias. O sítio múltiplo de clonagem desses vetores se localiza precisamente dentro do gene *lacZ* (Figura 11.6). Quando um inserto é introduzido nesse vetor, o gene é interrompido, não produzindo mais a enzima funcional. Como consequência, colônias de bactérias que têm o plasmídeo vazio apresentam cor azul, enquanto aquelas com plasmídeos com inserto exibem cor branca.

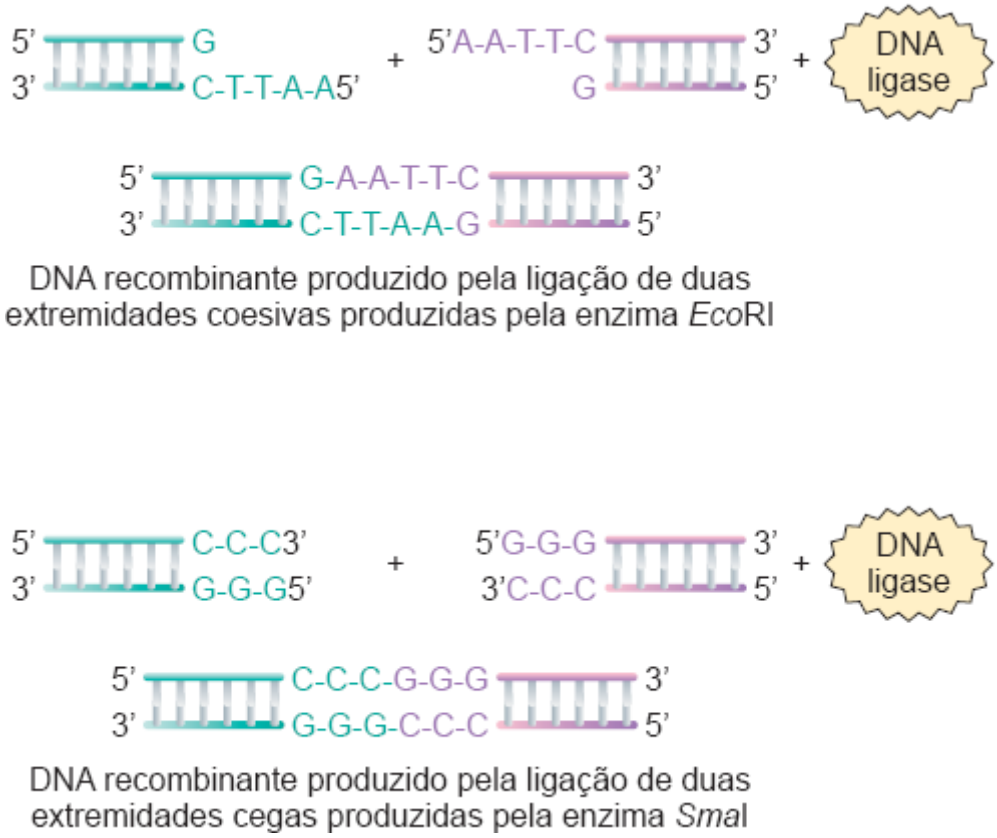


Figura 11.5 Produção de DNA recombinante. DNA de origens diferentes (representados por cores diferentes) que foram clivados com enzimas de restrição são covalentemente associados em um novo arranjo pela atividade da enzima DNA ligase.

Introduzindo genes em vetores plasmidiais

A Figura 11.7 mostra os principais passos em um experimento de clonagem usando um vetor plasmidial que reúne as características citadas anteriormente. Um trecho de DNA de interesse digerido com uma enzima de restrição foi isolado e ligado no plasmídeo previamente digerido com a mesma enzima, utilizando-se a enzima DNA ligase. Após a reação de ligação, essas moléculas são introduzidas nas células bacterianas por meio de *transformação genética*. Se as bactérias forem semeadas em meios de cultura contendo o antibiótico apropriado, apenas aquelas contendo plasmídeos irão crescer e formar colônias. Dentre as colônias formadas, é possível identificar aquelas que contêm plasmídeos recombinantes (ou seja, contendo o inserto de DNA desejado) pela coloração diferencial no meio de cultura, que foi acrescido do substrato cromogênico para a enzima betagalactosidase (o composto X-Gal).

O poder dessa clonagem pode ser ilustrado com uma conta simples. Ao crescer uma colônia bacteriana contendo um plasmídeo recombinante em cultura, obtém-se densidades de até 10^9 células/mℓ. Cada célula bacteriana carrega uma centena de cópias do plasmídeo. Portanto, em apenas 1 mℓ de cultura bacteriana, é possível obter 10^{11} cópias do plasmídeo portando o seu gene de interesse.

É necessário destacar que a gama de plasmídeos existentes para clonagem de genes é muito grande. Muitos desses plasmídeos apresentam características adicionais relevantes. Alguns são chamados de *vetores ponte*, pois replicam tanto em bactérias quanto em células eucarióticas; outros possibilitam que o gene clonado seja expresso, constitutivamente ou de maneira regulada, porque posicionam o gene inserido sob o controle de promotores conhecidos (para mais detalhes sobre promotores e controle da expressão gênica, ver Capítulo 6). Estes são chamados de *vetores de expressão*. Outra característica importante presente em alguns vetores é a possibilidade de expressar proteínas recombinantes em grandes quantidades e gerando fusões a pequenos peptídeos que permitem a sua purificação. Esses vetores plasmidiais são muito valiosos no estudo funcional de genes.

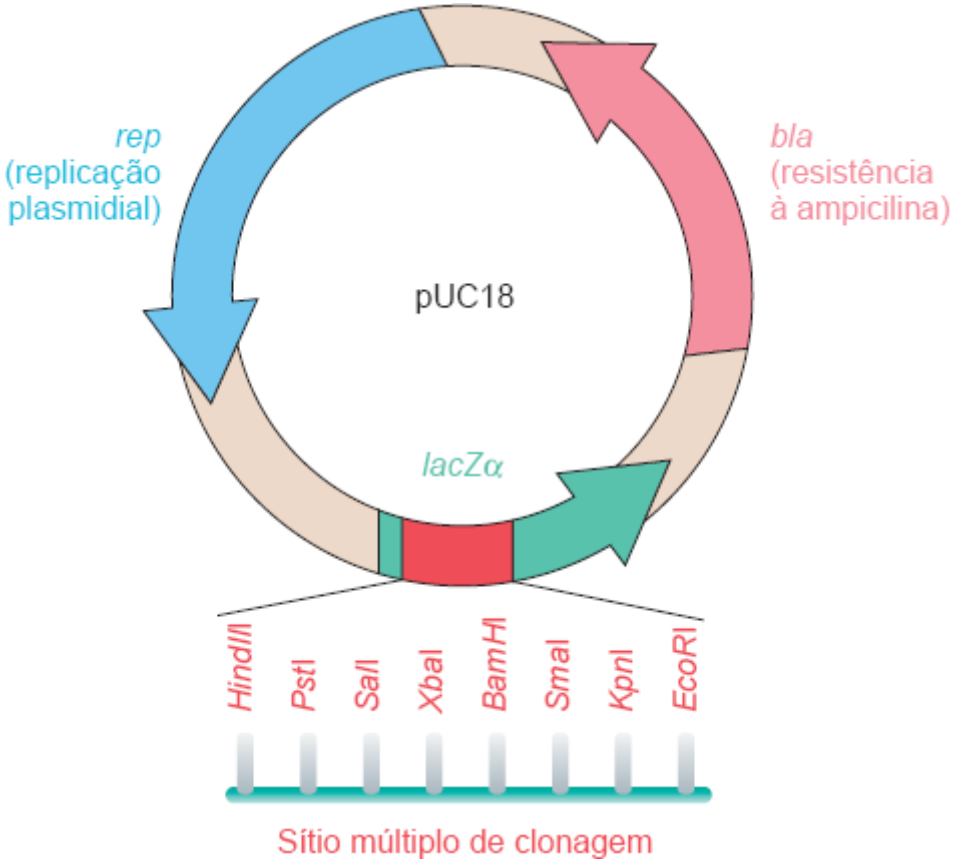


Figura 11.6 Mapa do vetor pUC18, mostrando suas principais características: gene de resistência a antibiótico (*bla*), sequências *rep* para replicação estável em bactérias e sítio múltiplo de clonagem inserido no gene *lacZ*.

Fragmentos grandes de DNA são clonados em outros tipos de vetores

Além dos plasmídeos, a técnica de clonagem pode ser feita com vários vetores diferentes. Em geral, esses vetores alternativos podem carregar fragmentos de DNA maiores que os suportados pelos plasmídeos.

Bacteriófagos

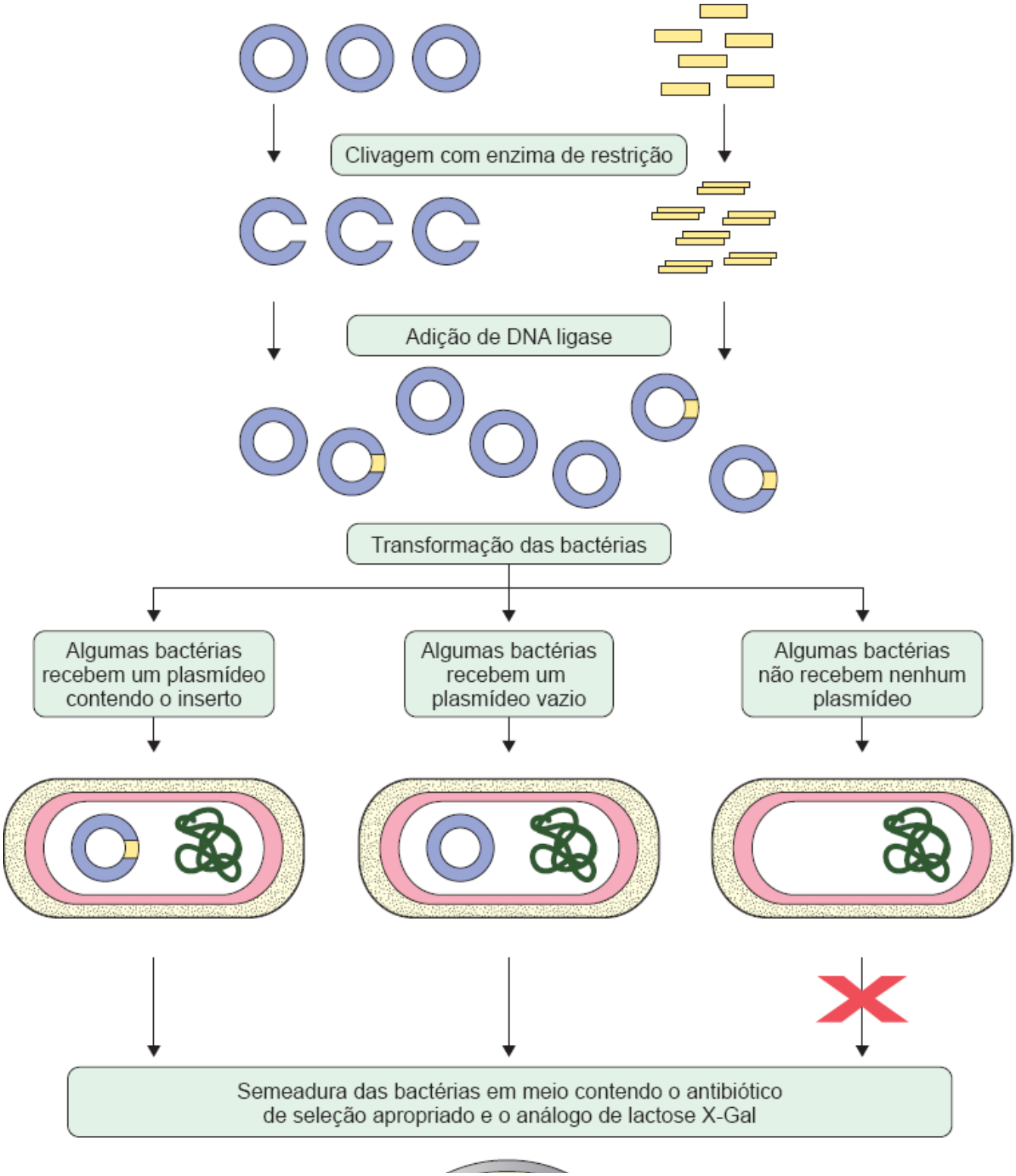
Os bacteriófagos proporcionaram uma série de avanços na genética. Por serem modelos de estudo bem conhecidos, não tardaram a ser utilizados como vetores genéticos com o advento da tecnologia do DNA recombinante. O bacteriófago mais usado para este fim é o fago lambda, que tem um genoma linear de 48,5 Kpb, que é empacotado no capsídeo viral. Destes, aproximadamente 15 Kpb codificam genes dispensáveis para o ciclo lítico do fago, e podem ser substituídos por DNA exógeno por meio da metodologia do DNA recombinante. Assim, o fago lambda pode ser usado para clonar fragmentos de tamanho superior aos suportados por plasmídeos, mas o tamanho dos insertos também tem um limite. Os capsídeos virais não incorporam moléculas de DNA menores que 37 Kpb, nem maiores que 52 Kpb. Uma vez que aproximadamente 33 Kpb de DNA endógeno do fago são essenciais para seu ciclo e multiplicação, o tamanho máximo de DNA exógeno comportado por esses vetores fica em torno de 18 Kpb.

Cosmídeos

Cosmídeos são vetores artificias, com características tanto de plasmídeos quanto do bacteriófago lambda. Essas moléculas podem ser mantidas como plasmídeos circulares em bactérias e geralmente apresentam tamanhos menores que 10 Kpb. Contudo, os cosmídeos têm as sequências conhecidas como *sítios cos* do fago lambda, necessárias para o empacotamento do DNA em partículas virais. Dessa maneira, podem ser empacotados por um sistema *in vitro* (uma mistura de enzimas purificadas do fago lambda), que reconhece os *sítios cos* e incorpora os cosmídeos em partículas virais, como se estes fossem o próprio genoma do vírus. Conforme visto anteriormente, as partículas virais comportam até 52 Kpb de DNA e, assim, um cosmídeo de 10 Kpb pode carregar fragmentos de até 42 Kpb.

Produzindo cromossomos artificiais | BAC e YAC

BAC e YAC (do inglês, *bacterial artificial chromosomes* e *yeast artificial chromosomes*) são cromossomos artificiais de bactérias e leveduras, respectivamente. Para a replicação estável como um cromossomo nas células da levedura *Saccharomyces cerevisiae*, é necessário que os YAC tenham todos os elementos funcionais de um cromossomo natural (ver Capítulo 10): *origens de replicação*, que garantem a duplicação do cromossomo na fase S; *centrômeros*, que garantem a partição correta dos cromossomos durante a divisão celular; e *telômeros*, necessários para a manutenção e replicação das extremidades do cromossomo. A capacidade desses vetores é impressionante: fragmentos de até 1 Mpb (megapares de base, ou seja, 1 milhão de pares de bases) podem ser clonados. No entanto, o DNA clonado em YAC tende a sofrer uma série de rearranjos, inserções e deleções. Assim, o uso desses vetores é limitado e, quando possível, outros vetores mais estáveis, como BAC devem ser escolhidos. Os BAC representam os vetores mais populares atualmente para clonagem de longos trechos de DNA, e foram extensamente usados nos primeiros projetos de sequenciamento completo de genomas (ver Capítulo 12). São derivados do plasmídeo F de *E.coli*, o qual se mantém em poucas cópias na célula e é capaz de carregar grandes trechos de DNA estavelmente. Os BAC comportam insertos de até 300 Kpb.



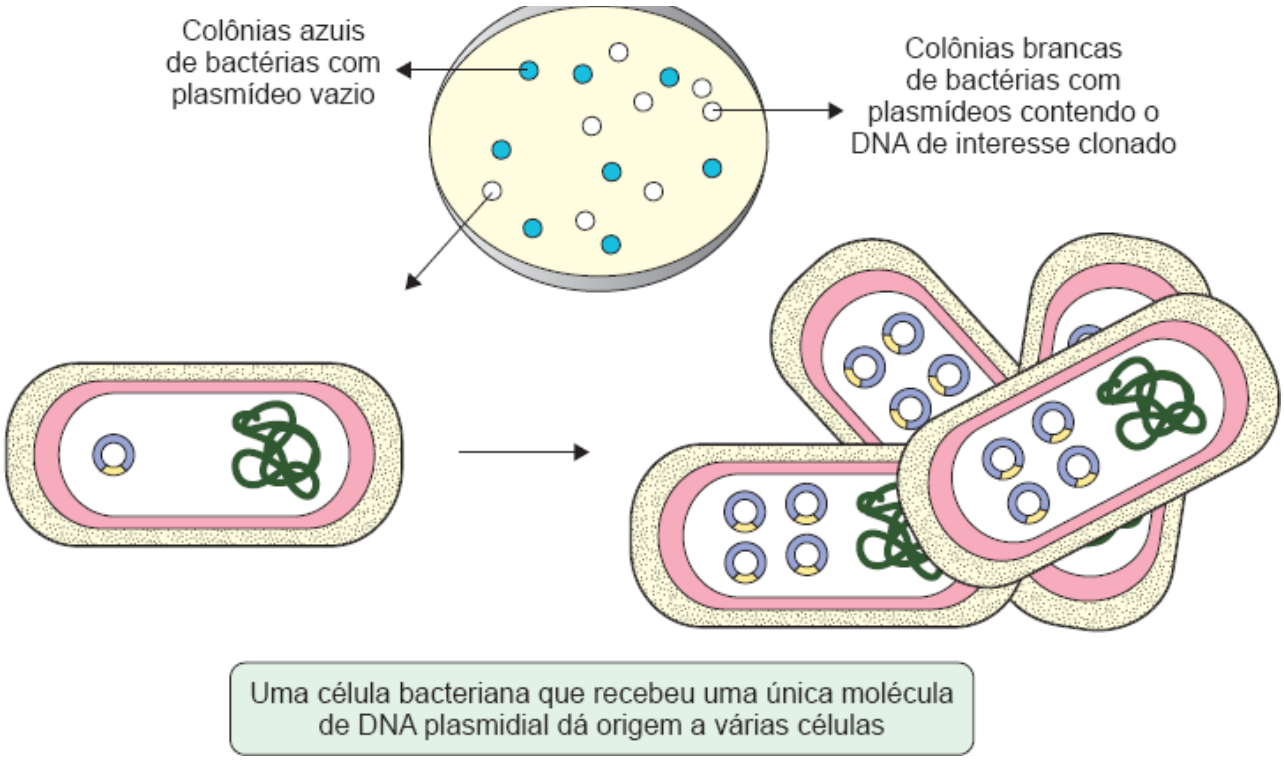


Figura 11.7 Processo de clonagem de um fragmento de DNA em um vetor plasmidial típico. Notar que os vários fragmentos ilustrados representam uma preparação pura de fragmentos idênticos, obtida por digestão de grande quantidade de DNA com uma enzima de restrição e purificação.

Conjuntos de genes podem ser clonados em bibliotecas genômicas e de cDNA

Além de clonar genes individuais, é possível obter um conjunto de clones que represente todo o genoma: são as chamadas *bibliotecas genômicas*. A construção de uma biblioteca genômica está ilustrada na Figura 11.8. Todo o DNA genômico é digerido com uma enzima de restrição, e clonado em um vetor apropriado. Assim, em vez de clonar um fragmento específico, todos os fragmentos obtidos por digestão de um longo trecho de DNA são clonados. Obviamente, na prática, é virtualmente impossível conseguir que uma biblioteca represente de fato todo o genoma, por vários fatores. Primeiramente, porque alguns fragmentos de DNA são intrinsicamente difíceis de serem clonados, pois são instáveis ou tóxicos para as células bacterianas. A distribuição de sítios de restrição também impossibilita a clonagem de alguns trechos de DNA. Por último, uma vez que a clonagem ocorre ao acaso, alguns genes serão representados várias vezes (ou seja, em vários clones), enquanto outros são pouco representados ou até mesmo ausentes, por simples variação estatística, que obedece a uma distribuição de Poisson. Apesar dessas limitações, bibliotecas genômicas são ferramentas poderosas em estudos de biologia molecular.

Em se tratando de organismos eucarióticos, é sabido que grande parte do genoma não é codificante. Portanto, uma imensa parcela dos clones de uma biblioteca de DNA genômico irá conter apenas DNA não codificante (Figura 11.8) ou genes incompletos. Além disso, muito frequentemente, o objetivo da construção de uma biblioteca é dispor de toda a coleção de genes de um determinado organismo ou os genes que estão sendo expressos em uma condição fisiológica em particular. Nesse caso, uma biblioteca genômica tem pouca utilidade, e pode-se construir uma *biblioteca de cDNA*. O esquema da construção de uma biblioteca de cDNA está representado na Figura 11.9. Para isso, é necessário produzir DNA a partir do RNA total presente na célula ou tecido de interesse, o que é possível graças à enzima *transcriptase reversa*. O Capítulo 6 mostrou que em eucariotos os RNA mensageiros são processados para a remoção dos introns e, posteriormente, são poli-adenilados. A presença da cauda poli-A é, portanto, uma característica dos RNA mensageiros maduros. Uma das maneiras de se produzir uma biblioteca de cDNA é obter uma preparação de RNA total de um organismo, que contém todos os RNA mensageiros não processados, RNA transportadores e RNA ribossômicos, além dos mRNA maduros. A síntese do cDNA pode ser feita utilizando-se um oligonucleotídeo poli-T como iniciador, de modo que apenas a fração poli-adenilada do RNA total será usada como molde. Este conjunto de cDNA pode ser então clonado para a produção de uma biblioteca.

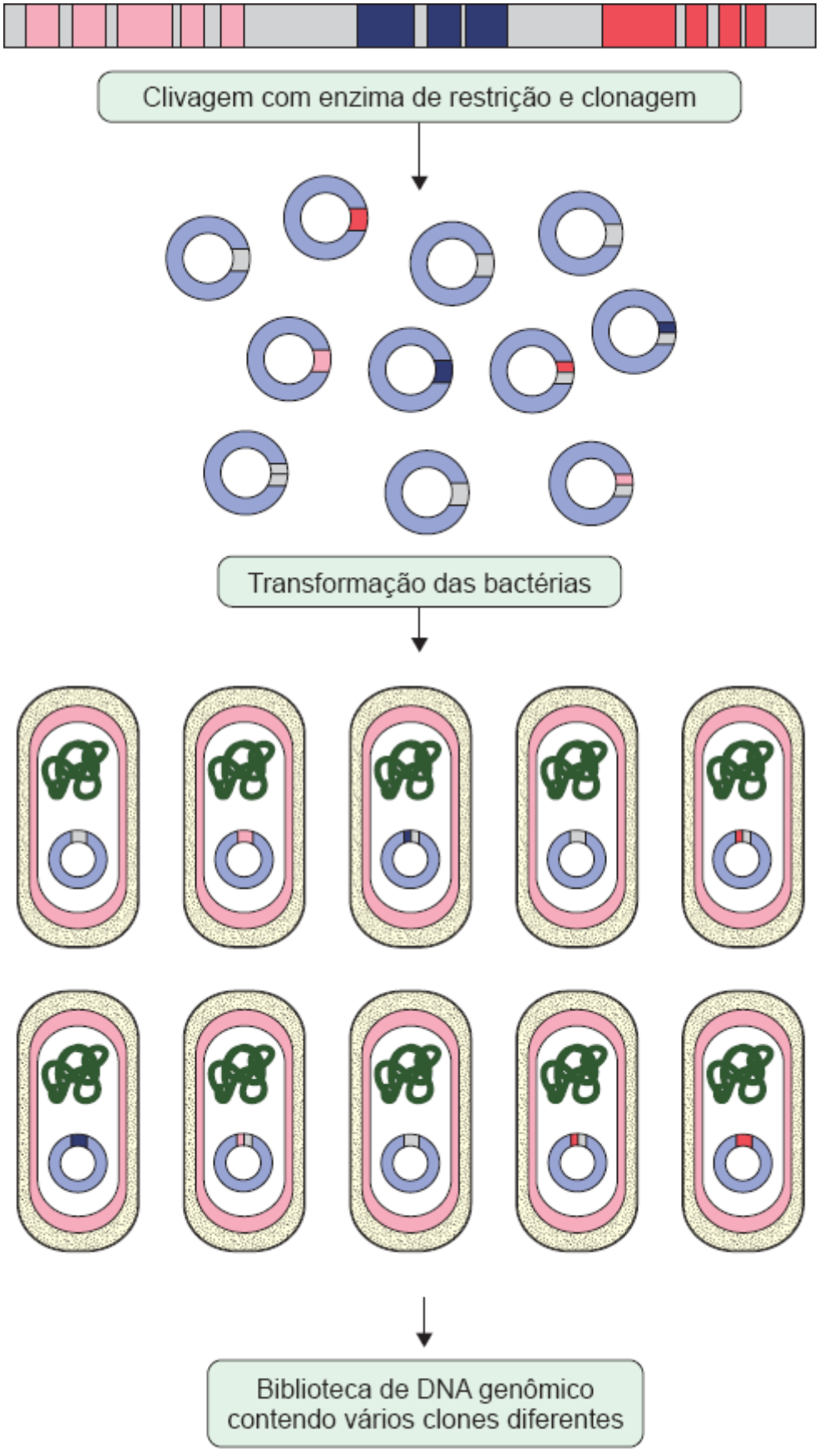
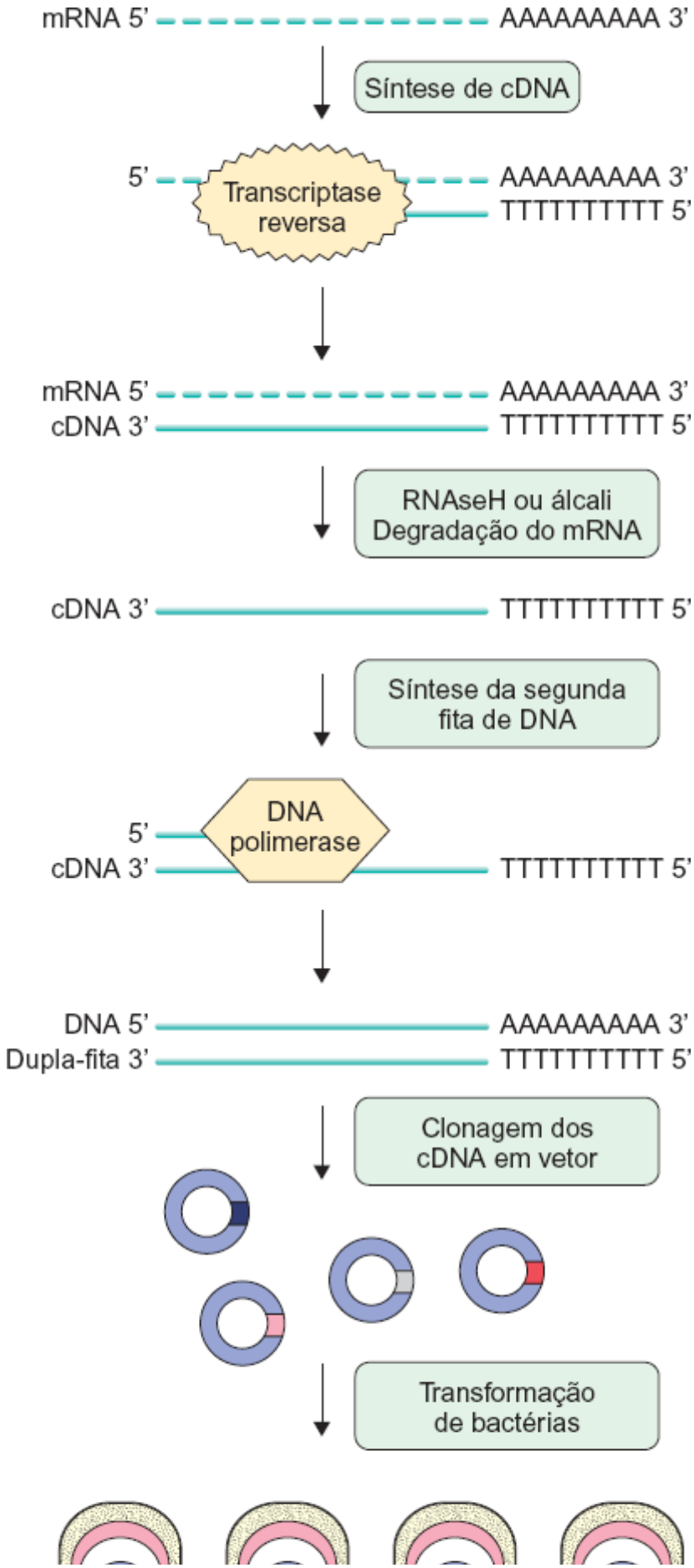


Figura 11.8 Esquema de construção de uma biblioteca genômica. O fragmento de DNA representado contém exons destacados em cores e introns e demais regiões não codificantes em cinza.

Bibliotecas são ferramentas úteis em biologia molecular

Bibliotecas genômicas e de cDNA são ferramentas muito úteis em biologia molecular, e muitas são as suas aplicações. Sua construção é o primeiro passo no sequenciamento de genomas completos utilizando a metodologia de Sanger. Uma biblioteca também pode ser usada para a obtenção de um clone de um gene de interesse. Nesse caso, é necessário identificar esse clone em meio aos demais. Uma maneira de realizar tal busca é pelo uso de

técnicas de hibridização, conforme explicado posteriormente neste capítulo e detalhado na Figura 11.14. Bibliotecas podem ser ainda construídas em uma série de vetores que possibilitem análises funcionais de genes, tais como vetores de expressão e vetores-ponte. Tais bibliotecas podem ser usadas em varreduras que busquem identificar genes que desempenham funções específicas na célula. Por último, é possível obter a sequência de vários clones de uma biblioteca de cDNA, de modo a se alcançar um perfil de expressão gênica em condições de interesse.



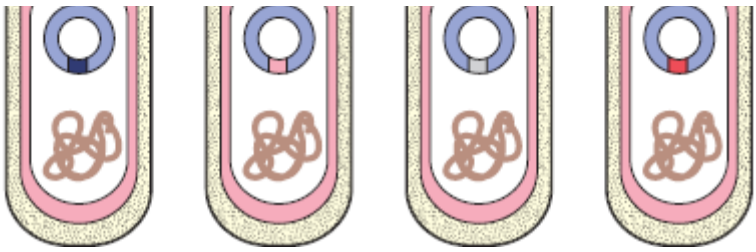


Figura 11.9 Esquema de construção de uma biblioteca de cDNA utilizando a enzima transcriptase reversa.

Técnicas de hibridação de ácidos nucleicos

Técnicas que se baseiam na hibridação de ácidos nucleicos surgiram a partir de meados dos anos 1970, viabilizando a realização de uma série de estudos de mapeamento e expressão de genes de interesse. O DNA forma uma fita dupla devido a interações entre bases nitrogenadas *complementares* de fitas independentes, ou dentro da mesma fita. Esse pareamento entre bases complementares pode ser facilmente desfeito por agentes como o calor e o pH alcalino, em um processo chamado *desnaturação*. Além disso, a desnaturação do DNA é reversível e, uma vez reestabelecidas as condições ideais, ocorre novamente o anelamento entre regiões complementares e a *renaturação* do DNA; ou seja, o pareamento entre essas bases e formação de um duplex. A *hibridação* de ácidos nucleicos consiste no anelamento entre fitas de DNA ou RNA de diferentes origens.

Sondas genéticas são usadas em experimentos de hibridação

Por meio da hibridação, pode-se localizar uma região de interesse em uma amostra de DNA ou RNA. Para isso, utiliza-se uma *sonda*, que consiste em um fragmento de DNA correspondente a uma parte da região a ser identificada, e que contém uma molécula ou átomo com marcação facilmente detectável, o que torna possível identificar o local em que ocorreu a hibridação ao DNA alvo.

O método de marcação de sondas mais utilizado tem como base o uso de um átomo radioativo, geralmente o ^{32}P ou ^{33}P, que produz sondas com alta sensibilidade de detecção. A marcação radioativa de sondas com ^{32}P pode ser feita pelo acoplamento desse átomo na extremidade do fragmento de DNA previamente isolado, ou pela incorporação do átomo radioativo durante a síntese *in vitro* da sonda (Figura 11.10). A marcação na extremidade da molécula de DNA é realizada pelo uso da enzima *polinucleotídeo quinase* e de ^{32}P acoplado à molécula do ATP na posição γ ([γ-^{32}P]ATP). A enzima fosforila o grupamento 5'-OH do DNA desfosforilado usando o átomo marcado. Na marcação por incorporação do átomo radioativo, a sonda é sintetizada a partir de um molde, utilizando uma das várias DNA polimerases disponíveis (a mais usada é a porção Klenow da DNA polimerase I de *E.coli*) ou a técnica de PCR. Um dos quatro desoxirribonucleotídeos presentes na reação de síntese contém ^{32}P na posição α ([α-^{32}P]dNTP), que é retido na molécula de DNA durante a síntese (Figura 11.10). A marcação por síntese é mais eficiente, uma vez que o átomo radioativo é incorporado ao longo de toda a molécula de DNA, geralmente sendo o método de escolha para estudos de hibridação de ácidos nucleicos. No entanto, as sondas marcadas na extremidade têm grande utilidade para outras técnicas de biologia molecular – que ainda serão discutidas neste capítulo. A presença de sondas marcadas radioativamente é facilmente detectada por *autorradiografia*, que consiste na detecção da radiação emitida após exposição em filmes de raios X.

Recentemente, vários métodos não radioativos para a marcação de sondas foram desenvolvidos, e têm se tornado cada vez mais comuns por evitar riscos associados à manipulação de radioatividade. Em um desses métodos, a sonda é marcada por um nucleotídeo acoplado à molécula digoxigenina (digoxinenina-11-dUTP). Para a detecção da sonda após a hibridação, utiliza-se um anticorpo antidigoxinenina, e o complexo antígeno-anticorpo é detectado por *fluorescência*, *quimioluminescência* ou *métodos colorimétricos* (Figura 11.11). A detecção por fluorescência é direta, uma vez que o anticorpo está acoplado a uma molécula que emite luz quando excitada em comprimento de onda específico. Nos métodos indiretos, o anticorpo está acoplado a uma enzima, que catalisa uma reação na presença de substrato específico, formando um produto que emite luz (*método quimioluminescente*) ou um precipitado com coloração visível a olho nu (*método colorimétrico*). Além da digoxigenina, outras moléculas têm sido usadas para marcação de sondas não radioativas. No entanto, em geral, são utilizados princípios similares aos descritos aqui.

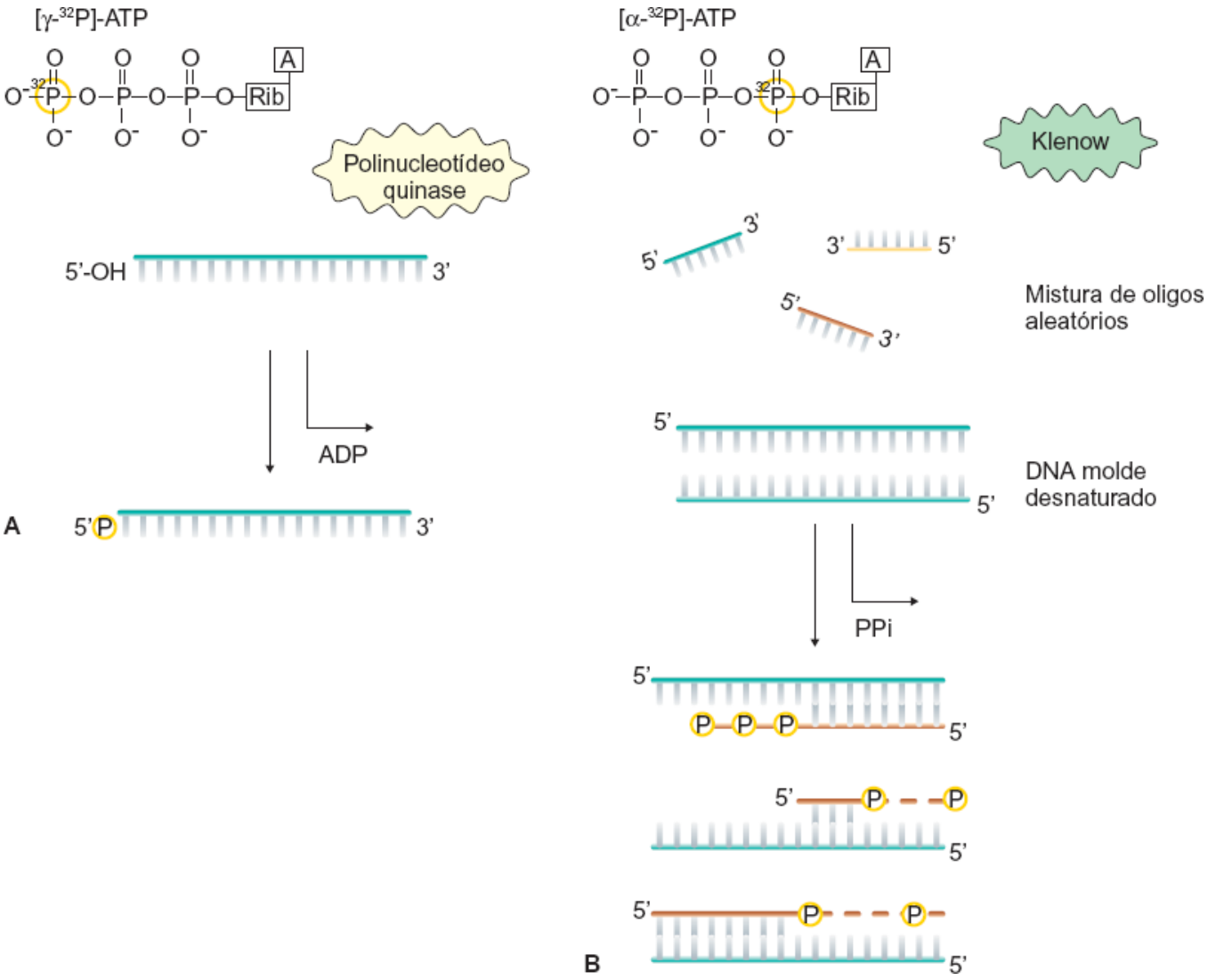


Figura 11.10 Representação de dois métodos de marcação de sondas radioativas. **A.** Marcação na extremidade 5': consiste em uma reação de fosforilação, e apenas o fosfato da posição γ do ATP é incorporado na sonda. Notar que o oligonucleotídeo sintetizado *in vitro* possui uma extremidade 5'-OH (não fosforilada). **B.** Marcação por síntese: a molécula de AMP é incorporada na sonda – liberando pirofosfato –, e o átomo radioativo deve estar na posição α do ATP.

A técnica de *Southern blot* detecta sequências de DNA em amostras separadas por eletroforese

A primeira das técnicas de hibridação usando sondas foi desenvolvida por Edwin Southern em 1975, sendo por isso denominada *Southern blot* (Figura 11.12). Esta técnica tornou possível a detecção e a identificação de um gene de interesse em uma amostra de DNA, ou até mesmo o seu mapeamento em relação a sítios de restrição no cromossomo. Inicialmente, é necessário cortar o DNA pelo tratamento com uma ou mais enzimas de restrição, de modo a obter fragmentos de DNA dentro da faixa de separação de um gel de agarose convencional. Essa mistura de fragmentos é submetida à eletroforese em gel de agarose produzindo, normalmente, um rastro de migração, em função da presença de vários fragmentos de todos os tamanhos possíveis.

Após a eletroforese, o gel é submerso em solução alcalina (pH > 13) para total desnaturação do DNA. Em seguida, é realizada a *transferência do DNA*, ou seja, é obtida uma réplica do gel em uma folha de papel de náilon, chamada de *membrana de hibridação*. A transferência pode ocorrer simplesmente por capilaridade, colocando-se a membrana sobre o gel em ambiente umedecido; ou então o processo é acelerado pelo uso de vácuo, com o auxílio de equipamentos específicos. A membrana contendo a réplica do gel é incubada com a sonda contendo o DNA correspondente à região de interesse, possibilitando a hibridação e a retenção da sonda na região da membrana onde está o DNA de sequência idêntica. Em geral, utiliza-se um excesso de sonda, o que torna mais provável o anelamento e a hibridação do DNA alvo com a sonda, em vez da renaturação do DNA. Utilizando-se um DNA marcador de peso molecular no gel de agarose, é possível determinar o tamanho do fragmento que contém o gene ou a região de interesse. É importante destacar que a hibridação ocorre até mesmo entre fragmentos que não

apresentam 100% de complementariedade, basta apenas que uma parte significativa ou várias partes distintas da sonda tenham complementariedade com o alvo. Assim, é possível encontrar um gene ainda não identificado em um organismo usando como sonda um gene já conhecido de outro organismo, desde que eles apresentem similaridade de sequência significativa.

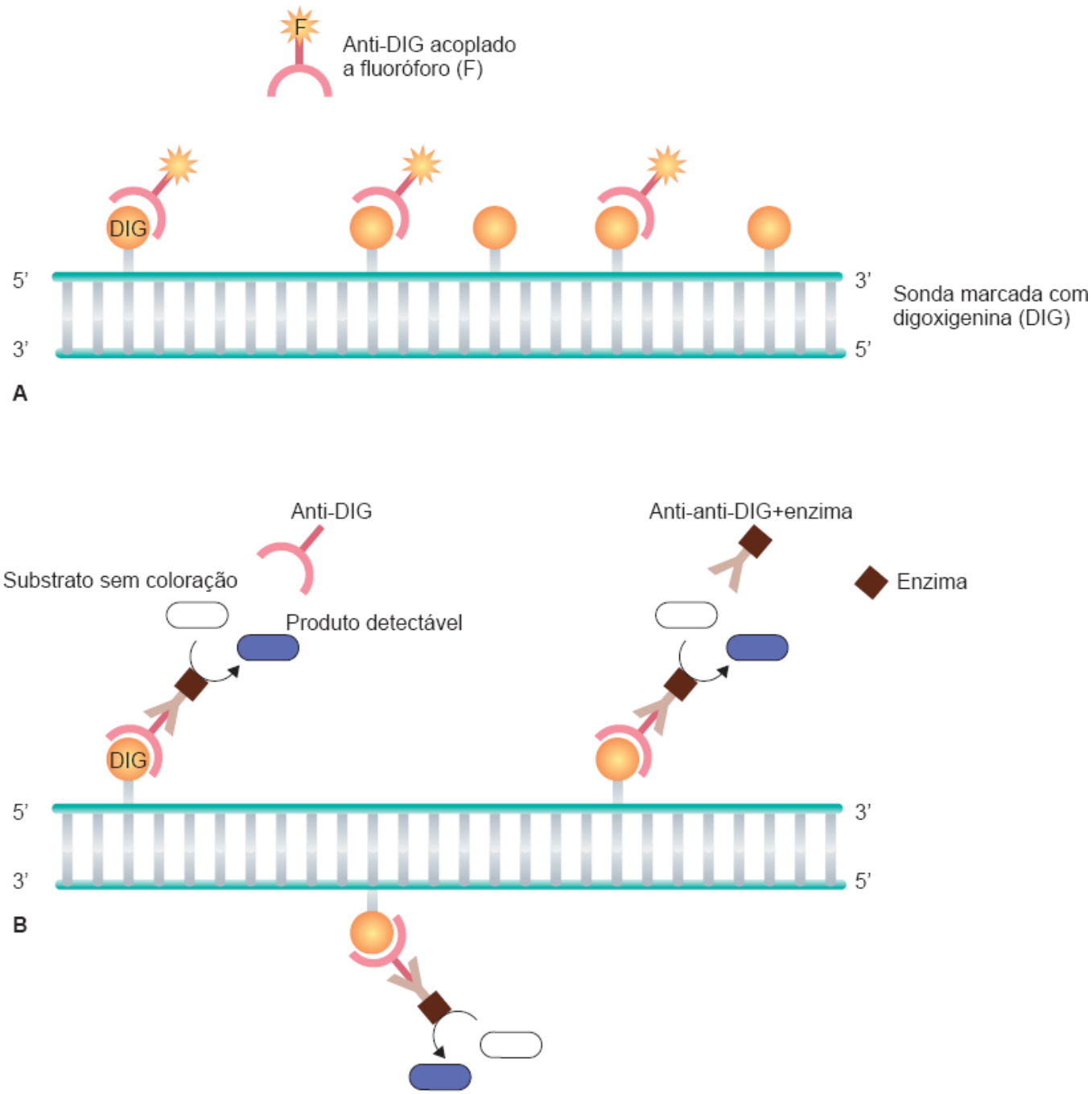


Figura 11.11 Representação de métodos de detecção de sondas marcadas com dUTP-digoxigenina (DIG) por fluorescência. **A.** Detecção direta: após a hibridação, o complexo sonda-alvo é detectado pelo uso de anticorpo antidigoxigenina (anti-DIG). **B.** Detecção indireta (método colorimétrico): utiliza-se também um anticorpo secundário (anti-anti-DIG), que reconhece o anticorpo anti-DIG e está acoplado a uma enzima que catalisa reação detectada por método colorimétrico ou quimioluminescente.

A técnica de *Northern blot* detecta a presença de RNA de um gene de interesse

Pouco tempo após o surgimento da técnica de *Southern blot*, em 1977, foram desenvolvidas modificações que possibilitaram a realização de estudos similares em amostras de mRNA extraído de células, eucarióticas ou procarióticas. A técnica, denominada *Northern blot*, permite *detectar*, *quantificar* e estimar o *tamanho* dos transcritos de um gene de interesse em uma mistura contendo todas as moléculas de RNA de uma célula, sendo muito usada em estudos de expressão gênica. Ela segue os mesmos princípios do *Southern blot*: a amostra de RNA também é fracionada por tamanho por meio de eletroforese em gel de agarose; no entanto, não há necessidade de

tratamento prévio, uma vez que as moléculas de RNA são bem menores e de tamanho conveniente para separação em gel. Além disso, neste caso o gel de agarose deve conter um agente desnaturante, a fim de desfazer estruturas secundárias presentes nas moléculas de RNA. Após a hibridação com sonda específica para o gene de interesse, a quantidade de sonda retida em cada amostra é uma medida da quantidade de transcritos do gene ali presentes e, portanto, dos níveis de expressão do gene. Para isso, deve-se utilizar um excesso de sonda, o que garante que a sua quantidade não será o fator limitante na hibridação, mas sim a quantidade de RNA complementar presente na amostra.

Microarranjos de DNA monitoram a expressão de milhares de genes ao mesmo tempo

A técnica de microarranjos de DNA também se utiliza do princípio de hibridação de ácidos nucleicos, e evoluiu a partir das técnicas de *Southern* e *Northern blot* para tornar possível a análise da expressão gênica global de uma célula. Assim, fragmentos representando todos ou a maioria dos genes conhecidos de um organismo são imobilizados em uma lâmina em alta densidade (Figura 11.13). Para isso, utilizavam-se originalmente fragmentos de DNA obtidos a partir de uma biblioteca de cDNA do organismo de interesse; contudo, com o advento do sequenciamento de genomas em larga escala, passaram a ser utilizados fragmentos pequenos sintetizados *in vitro* com base na sequência conhecida dos genes do organismo. Amostras de cDNA total de células ou tecidos são marcadas com moléculas fluorescentes, os *fluoróforos*, e, assim, todas as moléculas de cDNA presentes na amostra funcionam como sondas, hibridando por complementariedade na posição correspondente na lâmina (Figura 11.13). A intensidade de fluorescência emitida em cada posição da lâmina representa o número total de moléculas de cDNA presentes e, consequentemente, o número de transcritos de cada um dos genes. Conforme será possível ver com mais detalhes no Capítulo 13, essa técnica é muito usada para comparar os níveis de expressão de todos os genes representados em uma lâmina entre duas condições diferentes, em um único experimento. Para isso, basta que as amostras de cDNA de cada uma das condições a ser comparada estejam marcadas com um fluoróforo diferente (Figura 11.13). Logo, a fluorescência emitida em cada posição da lâmina será uma mistura dos dois fluoróforos e representará, portanto, a quantidade relativa de cDNA marcados com os fluoróforos 1 e 2.

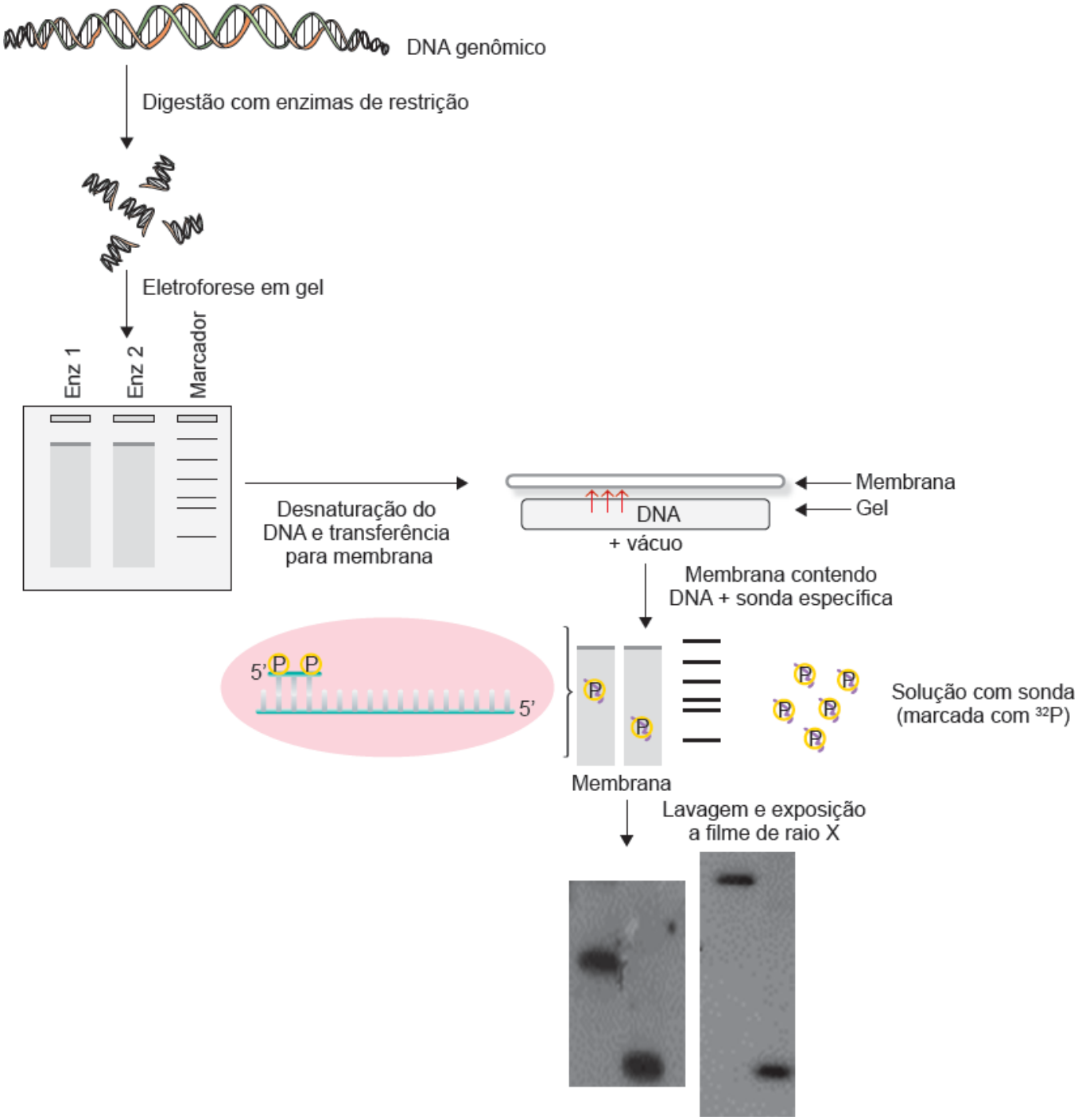


Figura 11.12 Esquema da técnica de *Southern blot*. A técnica possibilita identificar um gene de interesse em uma mistura de fragmentos obtidos por digestão de um DNA genômico com uma ou mais enzimas de restrição. Quando diferentes enzimas são utilizadas, o gene ou parte dele é encontrado em fragmentos de tamanho distintos, como mostrado na figura.

Técnicas de hibridação são úteis para buscar genes em bibliotecas

Neste ponto, você já sabe o que são e como são construídas bibliotecas de DNA e cDNA. Muitas vezes, desejamos isolar e clonar um gene de interesse de um determinado organismo. Isso pode ser feito se tivermos à disposição uma biblioteca de clones desse organismo e uma sonda correspondente a uma região do gene de interesse, usando-se a técnica de hibridação chamada de *colony blot* (Figura 11.14). Por ela, colônias bacterianas representando vários clones independentes da biblioteca são dispostas em uma placa de petri contendo meio de cultura apropriado. Após o crescimento, elas são transferidas para uma membrana de hibridação, produzindo uma réplica idêntica à placa. Essa transferência é feita por contato direto da membrana com a placa. Em seguida, a membrana é submetida a uma série de tratamentos que lisam as células, desnaturam e fixam o DNA. A membrana é então incubada com a sonda para o gene de estudo. O local onde ocorre hibridação entre a sonda e o alvo corresponde a um clone da biblioteca

que contém um inserto com parte ou todo o gene de interesse. Uma vez que a membrana é uma réplica perfeita da placa original, é fácil identificar a colônia de bactéria que contém o clone desejado (Figura 11.14).

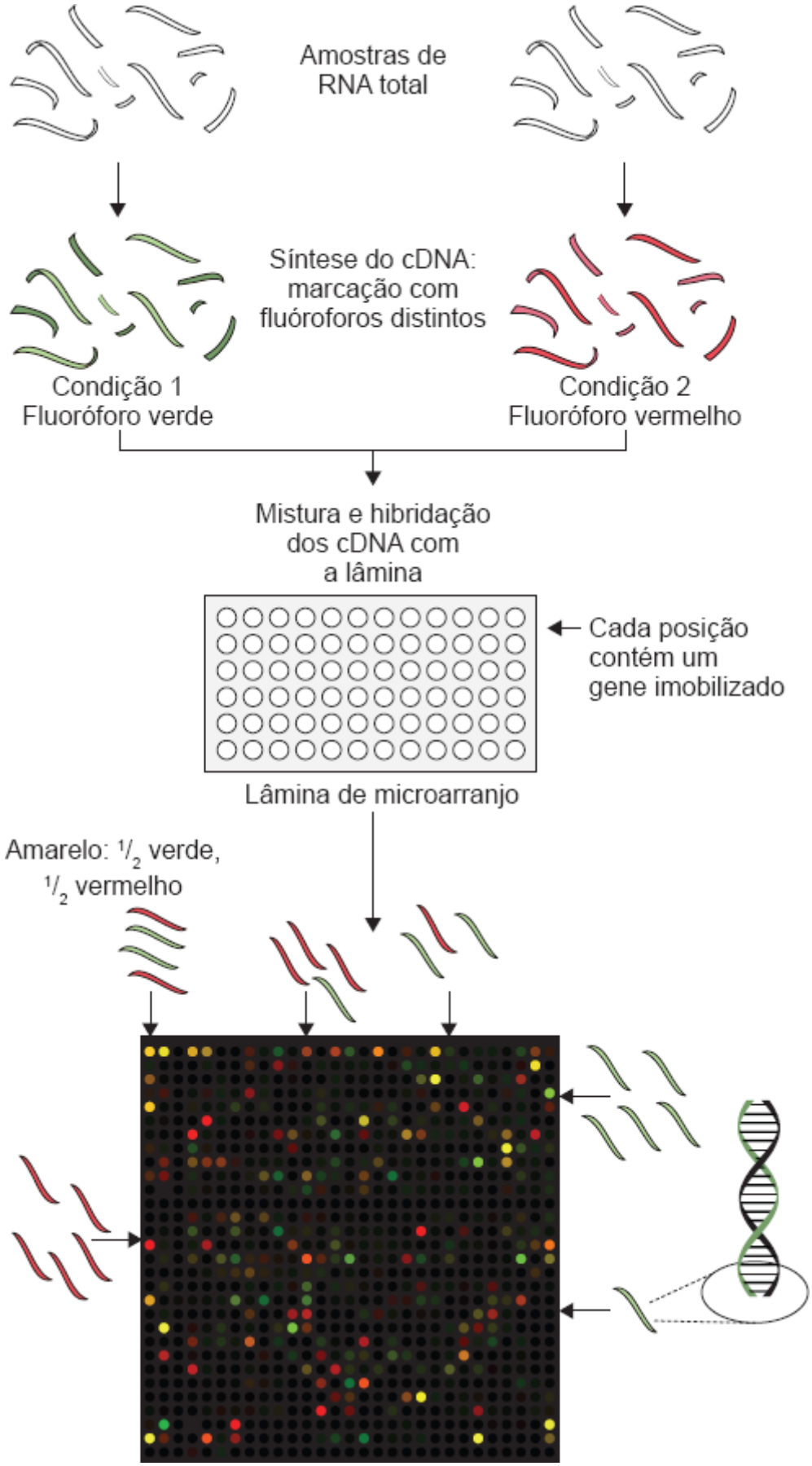


Figura 11.13 Representação da técnica de microarranjos de DNA para a comparação do padrão global de expressão gênica entre duas condições diferentes. A coloração amarela indica que uma mesma quantidade de moléculas marcadas com fluoróforo vermelho e verde deu origem a híbridos na posição.

Experimentos de hibridação podem ser usados para localizar genes em cromossomos

A técnica de *Southern blot* depende da extração do DNA total de células ou tecidos e sua digestão com enzimas de restrição, não sendo capaz de fornecer informações sobre a localização de genes no cromossomo, muito importante para estudos citogenéticos, por exemplo. Em 1969, antes até do desenvolvimento da técnica de *Southern blot*, Joseph Gall e Mary Lou Pardue descreveram uma técnica que usava a hibridação para determinar a posição de uma sequência de DNA de interesse no cromossomo de células *in situ*; ou seja, no seu contexto celular. Na época, a técnica se utilizava de sondas radioativas, o que dificultava bastante a sua utilização. Posteriormente, em 1977, as sondas radioativas foram substituídas por marcadores fluorescentes, cuja detecção tem como base os mesmos princípios já descritos na Figura 11.11. A técnica foi então chamada de *FISH*, do termo em inglês *fluorescence in situ hybridization*. Na hibridação por FISH, as células ou tecidos são fixados a uma lâmina e tratadas com agentes que desnaturam o DNA sem alterar a estrutura dos cromossomos. Após a hibridação com a sonda e a lavagem do material, a localização da fluorescência emitida pela sonda é detectada em microscópio apropriado. Usando-se sondas marcadas com fluoróforos diferentes, é possível localizar vários genes de interesse simultaneamente. A técnica de FISH foi fundamental para os primeiros estudos de mapeamento de genes em cromossomos humanos e, atualmente, é muito usada no diagnóstico clínico de doenças genéticas associadas a rearranjos cromossômicos. Também tem sido muito utilizada para a localização de transcritos de RNA no interior da célula. Nesse caso, a hibridação ocorre sem uma etapa de desnaturação, o que garante que o DNA celular não será hibridado com a sonda, somente o RNA.

Reação em cadeia da polimerase (PCR)

No final da década de 1980, Mullis e Faloona desenvolveram uma técnica que revolucionou e impulsionou a biologia molecular: a PCR (do inglês, *polymerase chain reaction*), ou *reação em cadeia da polimerase*. Tal técnica possibilita a obtenção de milhares de cópias de um gene ou sequência de interesse a partir de poucas ou apenas uma molécula de DNA molde (o que chamamos de *amplificação do DNA*), em uma reação simples realizada *in vitro*. O boxe "Como a PCR surgiu e dominou a biologia molecular" mostra a história da concepção dessa técnica.

A reação de PCR consiste, portanto, na síntese de novas moléculas de DNA, mediada por uma enzima do tipo DNA polimerase. A síntese de DNA pelas DNA polimerases ocorre a partir de uma extremidade 3'-OH de uma pequena molécula iniciadora. Do mesmo modo, a reação de PCR também necessita de iniciadores, pequenos fragmentos de DNA complementares às extremidades da região a ser amplificada, chamados de *oligonucleotídeos*. Os oligonucleotídeos de qualquer sequência desejada podem ser facilmente sintetizados quimicamente, sendo também usados como sondas em experimentos de hibridação ou nas técnicas de sequenciamento de DNA – mostradas na próxima seção deste capítulo. Portanto, para amplificação de um gene de interesse por PCR, é preciso ter informações sobre a sequência da região a ser amplificada ou pelo menos das extremidades dessa região. Com o sequenciamento de genomas, é cada vez mais fácil obter informações sobre a sequência dos genes de interesse, e a técnica de PCR vem cada vez mais substituindo o uso de bibliotecas de DNA para o isolamento de genes.

O uso dos oligonucleotídeos iniciadores confere outra característica importante à PCR, a *especificidade*. O uso de dois iniciadores de tamanho relativamente pequeno (em torno de 20 pares de base cada um) costuma ser suficiente para garantir que a região a ser amplificada seja única, mesmo em genomas complexos, como o humano. Isso torna comum a tarefa de isolamento de uma região do genoma. Fragmentos amplificados por PCR podem ser facilmente clonados em diversos vetores genéticos, ou usados como sondas em experimentos de hibridação.

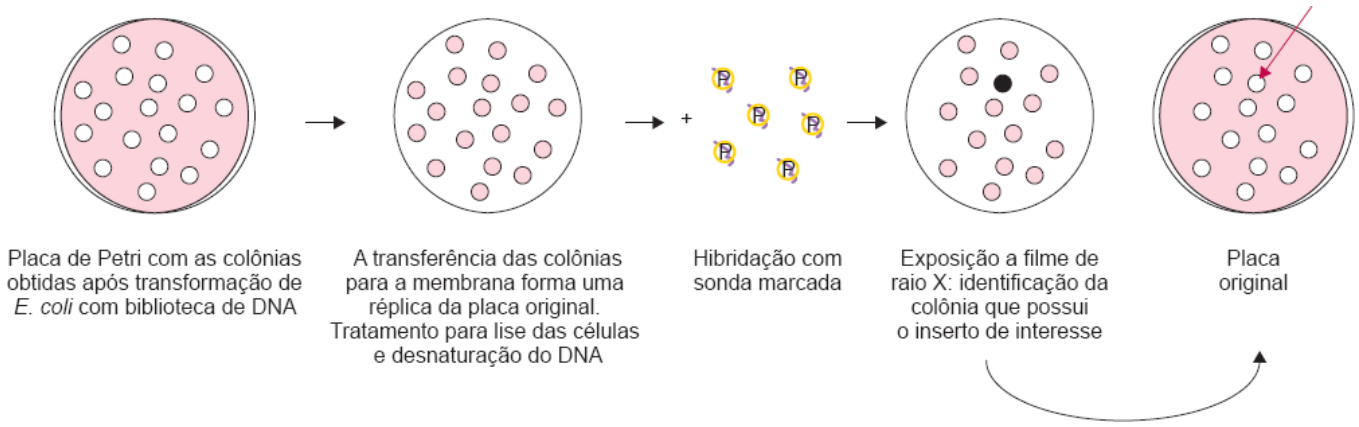


Figura 11.14 Esquema da identificação de uma colônia contendo um inserto de interesse em uma biblioteca de DNA usando a técnica de *colony blot*.

Como a PCR surgiu e dominou a biologia molecular

O primeiro trabalho que descreveu a técnica de PCR foi publicado na revista *Science*, em dezembro de 1985. O estudo original descreve a amplificação *in vitro* de um gene de cópia única de mamíferos, com o objetivo de desenvolver um método diagnóstico rápido e mais sensível para uma doença genética humana. O pesquisador responsável pela descoberta, o químico Kary Mullis, trabalhava na empresa de biotecnologia Cetus Corporation (Califórnia, EUA) e concebeu a técnica enquanto dirigia seu carro, em uma das longas viagens que fazia todos os fins de semana.

No entanto, quase 15 anos antes, o pesquisador indiano Har Gobind Khorana (Prêmio Nobel de Medicina em 1968 por seus estudos sobre o código genético) e seus colaboradores já haviam proposto uma ideia similar para síntese de genes de RNA transportador, sem, contudo, colocá-la em prática. A ideia de Khorana foi esquecida rapidamente, pois, na época, parecia inviável: não havia informações disponíveis sobre as sequências de genes, e a síntese química de oligonucleotídeos para serem usados como iniciadores era um processo extremamente complexo.

Da mesma maneira, a técnica de PCR (como foi proposta no trabalho original de 1985 e descrita em detalhes em um artigo posterior, publicado em 1987) era bastante trabalhosa e jamais teria sido tão impactante se não tivesse sido modificada, 3 anos depois, pelo uso da enzima termoestável Taq DNA polimerase. No início, utilizava-se como polimerase para a reação a enzima Klenow DNA polimerase I purificada de *E. coli*, que perdia a atividade a cada ciclo de desnaturação do DNA. Assim, era necessário adicionar mais enzima nos tubos de reação a cada novo ciclo, o que impedia a automatização do processo. Em 1988, Kary Mullis e seus colegas da Cetus aprimoraram a tecnologia com o uso da enzima purificada de *Thermus aquaticus*, Taq DNA polimerase, a mais usada até o momento. A bactéria *T. aquaticus* foi isolada e descrita por Thomas Brock em 1969, e é encontrada em nascentes termais a temperaturas de até 90°C. A enzima Taq DNA polimerase era bem mais eficiente que a Klenow e facilitou a automatização da técnica de PCR. Além disso, o uso da enzima termoestável possibilitou que os passos de anelamento e extensão fossem realizados a temperaturas mais elevadas, o que reduzia a amplificação de produtos inespecíficos. A partir daí, a técnica de PCR se popularizou definitivamente e, no fim dos anos 1980, tomou conta dos laboratórios de biologia molecular. O pesquisador Kary Mullis ganhou o Prêmio Nobel de Química em 1993 pela invenção da técnica de PCR.

A reação de PCR usa uma DNA polimerase termoestável e amplifica o DNA exponencialmente

A Figura 11.15 representa um esquema do funcionamento da técnica de PCR. Utiliza-se uma amostra de DNA *molde*, que contém a região a ser amplificada, um par de oligonucleotídeos, sendo cada um deles complementar a uma das extremidades do alvo, a enzima DNA polimerase e uma mistura equimolar dos quatro desoxirribonucleotídeos, que serão incorporados durante a síntese das novas moléculas. A técnica consiste em ciclos sucessivos de três etapas principais: *desnaturação*, *anelamento* e *extensão*. O primeiro passo é a *desnaturação* do DNA molde, que ocorre pelo tratamento com alta temperatura (94°C) por um tempo curto. Em seguida, a temperatura é reduzida para aproximadamente 60°C, o que possibilita o *anelamento* de cada um dos oligonucleotídeos a uma das fitas-molde do DNA desnaturado. Eventualmente, pode ocorrer a renaturação do DNA molde, mas a utilização de um excesso de oligonucleotídeos na reação favorece o pareamento molde-oligonucleotídeo. A terceira etapa é a *extensão*, ou seja, a síntese das novas moléculas de DNA a partir da extremidade 3'-OH de cada um dos pares molde-oligonucleotídeo que foram formados. Essa etapa ocorre a 72°C, temperatura ideal para funcionamento da DNA polimerase usada na reação. Como a reação de PCR envolve várias incubações em temperatura de 94°C, é necessário utilizar uma DNA polimerase termoestável, ou seja, que não sofra perda de atividade em altas temperaturas. A enzima Taq DNA polimerase isolada da espécie de bactéria termofílica *Thermus aquaticus* passou a ser a mais usada para amplificação de DNA por PCR, por sua alta eficiência nas condições necessárias para a reação.

Como mostra a Figura 11.15, o primeiro ciclo da PCR dá origem a duas moléculas novas e longas, com a extremidade 5'P definida por cada um dos oligonucleotídeos e término 3'-OH aleatório, que geralmente vai bem além do limite da região de interesse. No entanto, a cada novo ciclo de PCR, as novas moléculas sintetizadas se tornam moldes para a síntese de mais moléculas, de modo que, a partir do segundo ciclo da PCR, são obtidos apenas fragmentos com extremidades delimitadas pelos dois oligonucleotídeos usados na reação (Figura 11.15). Um total de aproximadamente 30 ciclos é realizado em uma reação de PCR padrão. No início da reação, quando os reagentes estão presentes em excesso, cada ciclo de amplificação produz o dobro de moléculas presentes no ciclo anterior, e a amplificação ocorre em escala exponencial. Em seguida, os reagentes disponíveis se tornam limitantes e a taxa de amplificação começa a reduzir, até que o processo chegue à saturação.

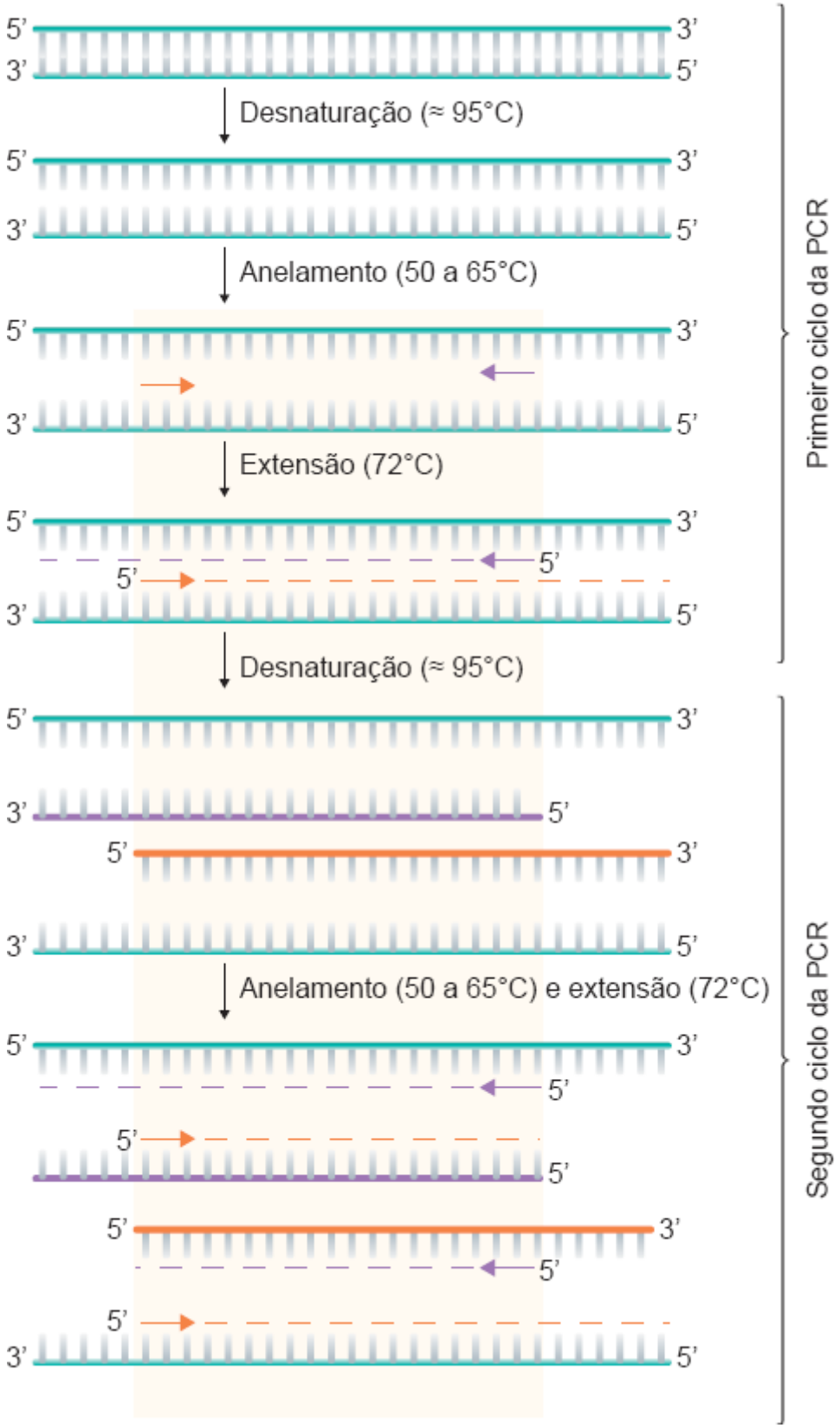


Figura 11.15 Representação dos dois primeiros ciclos de amplificação de DNA pela técnica de PCR. O primeiro ciclo dá origem a dois fragmentos, com extremidade 5' definida pelos oligonucleotídeos e região 3' indefinida. A partir do segundo ciclo, já começam a ser obtidos os fragmentos de tamanho delimitado pelos oligonucleotídeos nas duas extremidades.

Técnica de PCR

Agora que você sabe o funcionamento da técnica de PCR, sua simplicidade e eficiência, fica mais fácil compreender como ela rapidamente tomou conta dos laboratórios de biologia molecular, substituindo, em muitos casos, o uso de técnicas mais complexas e trabalhosas, como construção de bibliotecas de DNA e *Southern blot*. Com ela, ficou muito mais fácil isolar um gene de sequência conhecida em diversos organismos. Além disso, como mostra a

próxima seção, a técnica possibilitou o desenvolvimento das estratégias de sequenciamento em larga escala. A PCR facilita também a introdução de sítios de enzimas de restrição nas extremidades dos fragmentos amplificados, bastando, para isso, que tais sítios sejam incluídos nas extremidades dos oligonucleotídeos usados para a amplificação. Assim, esses fragmentos podem ser facilmente clonados em vetores de interesse diversos, usando-se os sítios de restrição introduzidos por PCR. A PCR permitiu também a obtenção de genes contendo mutações de interesse, o que chamamos de *mutagênese sítio-dirigida*. A mutação é inserida na sequência de oligonucleotídeos iniciadores, e os produtos da primeira amplificação passam a conter a mutação, servindo de molde para novos ciclos de amplificação.

Em razão de sua especificidade e da alta sensibilidade da técnica, que é capaz de amplificar amostras que contenham apenas uma molécula do alvo, a PCR é muito usada também no diagnóstico de doenças genéticas e no de doenças infecciosas (pela detecção de sequências de DNA de patógenos em amostras clínicas). Além disso, a PCR veio a contribuir bastante com estudos arqueológicos, possibilitando a análise de DNA obtido a partir de fragmentos de ossos e outros fósseis. A fim de elucidar como essa técnica impactou a identificação humana e a genética forense, ver o boxe "A identificação de indivíduos pela análise de polimorfismos".

cDNA de interesse também podem ser obtidos por PCR

A técnica de PCR também foi adaptada para a amplificação de DNA a partir de amostras de RNA, o que chamamos de RT-PCR (do inglês, *reverse transcriptase* PCR). A RT-PCR tem sido muito utilizada para a detecção de transcritos pouco abundantes (de difícil detecção por *Northern blot*), e também possibilita a construção de bibliotecas de cDNA quando apenas pequenas quantidades de mRNA estão disponíveis. No RT-PCR, é necessário adicionar uma etapa de síntese de cDNA a partir do RNA inicial, uma vez que a Taq DNA polimerase não é capaz de sintetizar DNA a partir de RNA. A síntese de cDNA é realizada pelo uso de uma enzima do tipo *transcriptase reversa*, uma classe de proteínas provenientes de retrovírus. Estas enzimas são capazes de gerar uma fita simples de DNA complementar ao RNA molde (Figura 11.18). Em seguida, o cDNA produzido é utilizado como molde na reação de PCR padrão. É importante notar que a amostra de RNA não pode conter nenhuma contaminação com DNA celular; caso contrário, o DNA contaminante pode causar amplificação indesejável na reação de PCR. Possíveis contaminações com DNA são eliminadas pelo tratamento com enzimas do tipo desoxirribonucleases (DNAse), que reconhecem e degradam exclusivamente moléculas de DNA.

A identificação de indivíduos pela análise de polimorfismos

A técnica de PCR também possibilitou a identificação de indivíduos a partir de amostras de baixa qualidade e/ou contendo quantidades mínimas de DNA, como traços de sêmen, saliva e até um fio de cabelo. Esse tipo de teste é especialmente importante na área forense, para identificação de criminosos a partir de amostras obtidas em cenas de crime.

Mesmo antes do uso da PCR, a identificação de indivíduos já se baseava na análise de *polimorfismos* da sequência de DNA. Polimorfismos são variações na sequência de bases em um mesmo local do cromossomo entre os indivíduos de uma população; podem ser variações de uma ou poucas bases e que ocorrem a aproximadamente cada 1.000 pb no cromossomo, ou podem ser variações no tamanho de uma região contendo repetições de uma sequência pequena. Tais regiões são chamadas de microssatélites, e têm tamanho bastante variável de pessoa para pessoa. Assim, cada indivíduo tem o seu padrão de polimorfismo.

Muitas vezes, esses polimorfismos modificam o padrão de fragmentos de DNA obtido após digestão com uma ou mais enzimas. Essas variações no padrão de clivagem por enzimas de restrição são chamadas RFLP (do inglês, *restriction fragment lenght polymorphisms*). Assim, as técnicas tradicionais de identificação de indivíduos consistiam na análise por *Southern blot* do padrão de digestão obtido em amostras de suspeitos, com o padrão gerado pela amostra coletada na cena do crime. Pela análise dos perfis de digestão obtidos com um grupo de sondas diferentes, é possível identificar com alta precisão o indivíduo entre um grupo de suspeitos. No entanto, havia muita dificuldade em obter amostras de DNA de qualidade e quantidade suficiente para a análise por *Southern blot*. Por possibilitar a amplificação do DNA da amostra inicial para posterior análise por *Southern blot*, a PCR eliminou esse problema.

Alternativamente, os polimorfismos podem ser caracterizados diretamente por PCR, detectando-se diferenças no tamanho de fragmentos produzidos pela amplificação de regiões polimórficas. A Figura 11.16 mostra um exemplo de genotipagem de indivíduos usando essa técnica. Para uma identificação sem ambiguidade, vários *loci* polimórficos diferentes devem ser analisados.

Uma técnica similar é utilizada para testes de paternidade, conforme mostrado na Figura 11.17. Para isso, é obtida uma amostra de DNA do filho (geralmente de sangue, mas pode ser feito a partir de saliva ou fio de cabelo) e uma ou mais amostras dos possíveis pais e da mãe. Para cada *locus* analisado, um dos alelos do filho é proveniente da mãe e outro do pai, o que torna possível identificar os padrões paterno e materno.

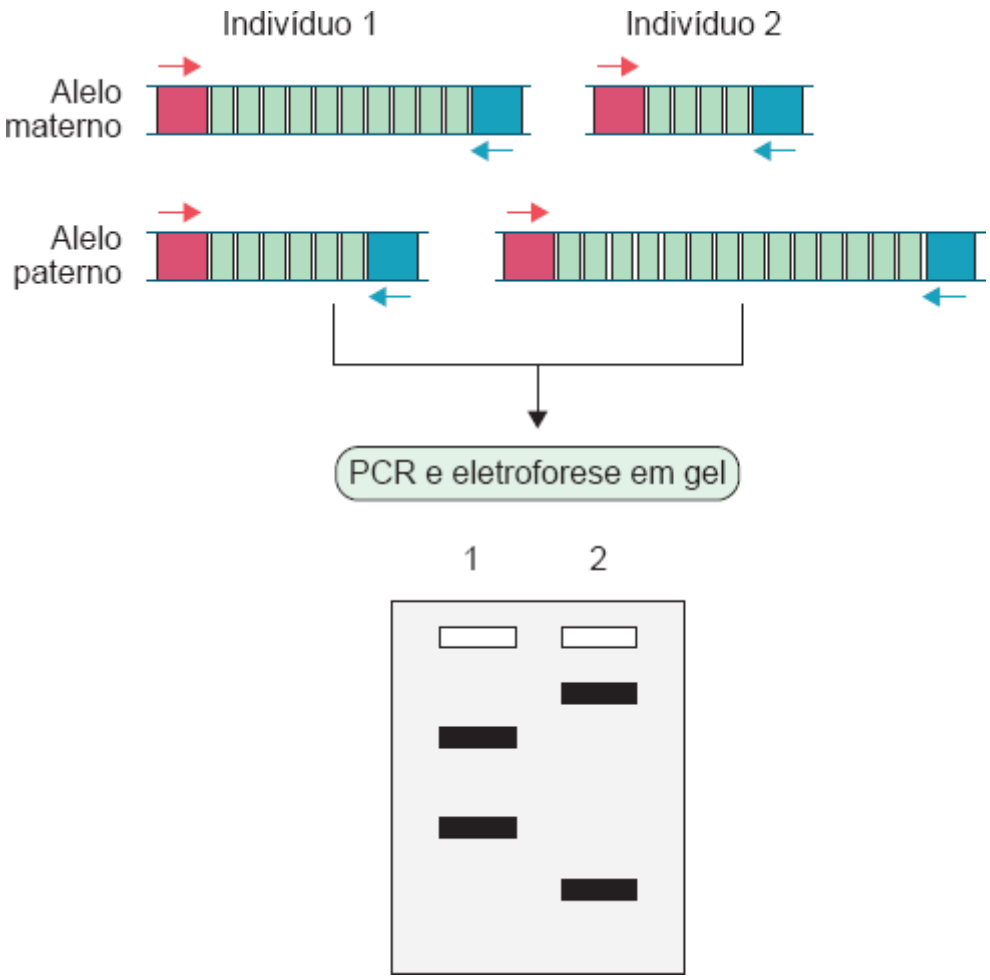


Figura 11.16 Genotipagem humana utilizando PCR. Regiões polimórficas correspondendo a microssatélites (caixas verdes) são flanqueadas por regiões não variáveis (representadas em rosa e azul). Iniciadores de PCR (setas rosa e azul) podem ser utilizados para amplificar a região de microssatélites. O produto de amplificação de um indivíduo pode conter até duas bandas de tamanhos diferentes, uma correspondente ao alelo materno e outra ao alelo paterno.

A síntese do cDNA pode ser realizada utilizando-se um oligonucleotídeo específico para um transcrito de interesse, ou usando um oligonucleotídeo inespecífico, capaz de anelar em todas as moléculas de RNA disponíveis na amostra, o que vai gerar uma amostra de cDNA total celular (Figura 11.18). O oligonucleotídeo inespecífico mais utilizado é o *oligo(dT)*, que anela na cauda poliA de mRNA eucarióticos. Após a síntese do cDNA, a reação de PCR é feita com um par de oligonucleotídeos específico para o gene de interesse. Alternativamente ao uso de oligo(dT), pode-se usar os *random hexamer*, que consistem em uma mistura de fragmentos de 6 nucleotídeos distribuídos de maneira aleatória, capazes de anelar em várias regiões distintas ao longo da molécula de RNA.

Quantificando DNA e RNA | PCR em tempo real

A técnica de PCR também pode ser usada para *quantificar* amostras de DNA e RNA. A quantificação de DNA por PCR é muito usada para a quantificação do número de cópias de genes em estudos de amplificação gênica e no diagnóstico clínico, como, por exemplo, na determinação de carga viral em amostras de pacientes. No caso do RNA, esse tipo de análise é cada vez mais usado em estudos de expressão gênica nas quais se deseja verificar um ou poucos transcritos de interesse. No entanto, a alta sensibilidade da técnica de PCR dificulta bastante a realização de análises quantitativas, pois variações mínimas na eficiência entre amostras são maximizadas pela amplificação em escala exponencial dos produtos, levando a resultados incorretos. Assim, os primeiros métodos de quantificação por PCR com base na análise dos produtos finais da reação eram bastante imprecisos.

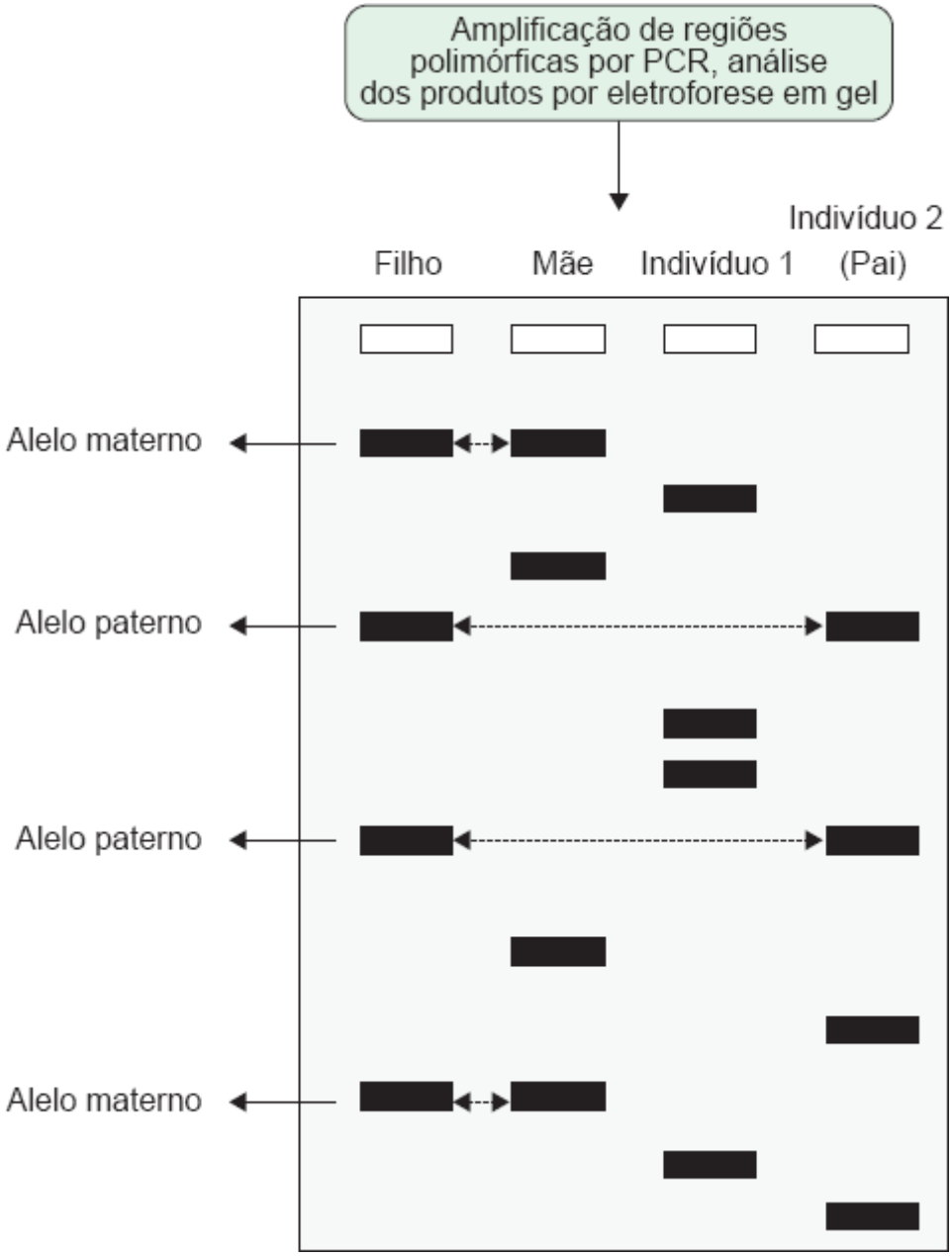


Figura 11.17 Exemplo de análise para determinação de paternidade. Produtos de amplificação de regiões de microssatélites são comparados em corrida eletroforética. Metade das bandas do filho deve ser igual às bandas obtidas com o DNA da mãe. A outra metade deve ter correspondência no padrão paterno, para determinar a paternidade. O indivíduo 2 apresenta o padrão de bandas compatível com a paternidade da criança. Neste exemplo simplificado, dois *loci* diferentes foram analisados, mas um número bem maior de *loci* devem ser analisados na prática, para determinar com altos níveis de confiança estatística.

Em 1992, Higuchi *et al*. desenvolveram um equipamento de PCR que possibilitava detectar a fluorescência emitida pelo brometo de etídeo conforme essa molécula fluorescente ia sendo incorporada nas novas moléculas de DNA fita-dupla que iam se formando durante a amplificação: surgia o PCR em tempo real. Essa técnica é muito mais precisa e eficaz para quantificação por PCR, já que, em vez de detectar os produtos finais da reação, identifica os produtos de amplificação conforme estão sendo formados a cada novo ciclo. A tecnologia para detecção dos produtos de reação foi bastante aprimorada e, atualmente, utilizam-se marcadores fluorescentes bem mais sensíveis e menos tóxicos que o brometo de etídeo, e que também se ligam aos produtos durante sua síntese.

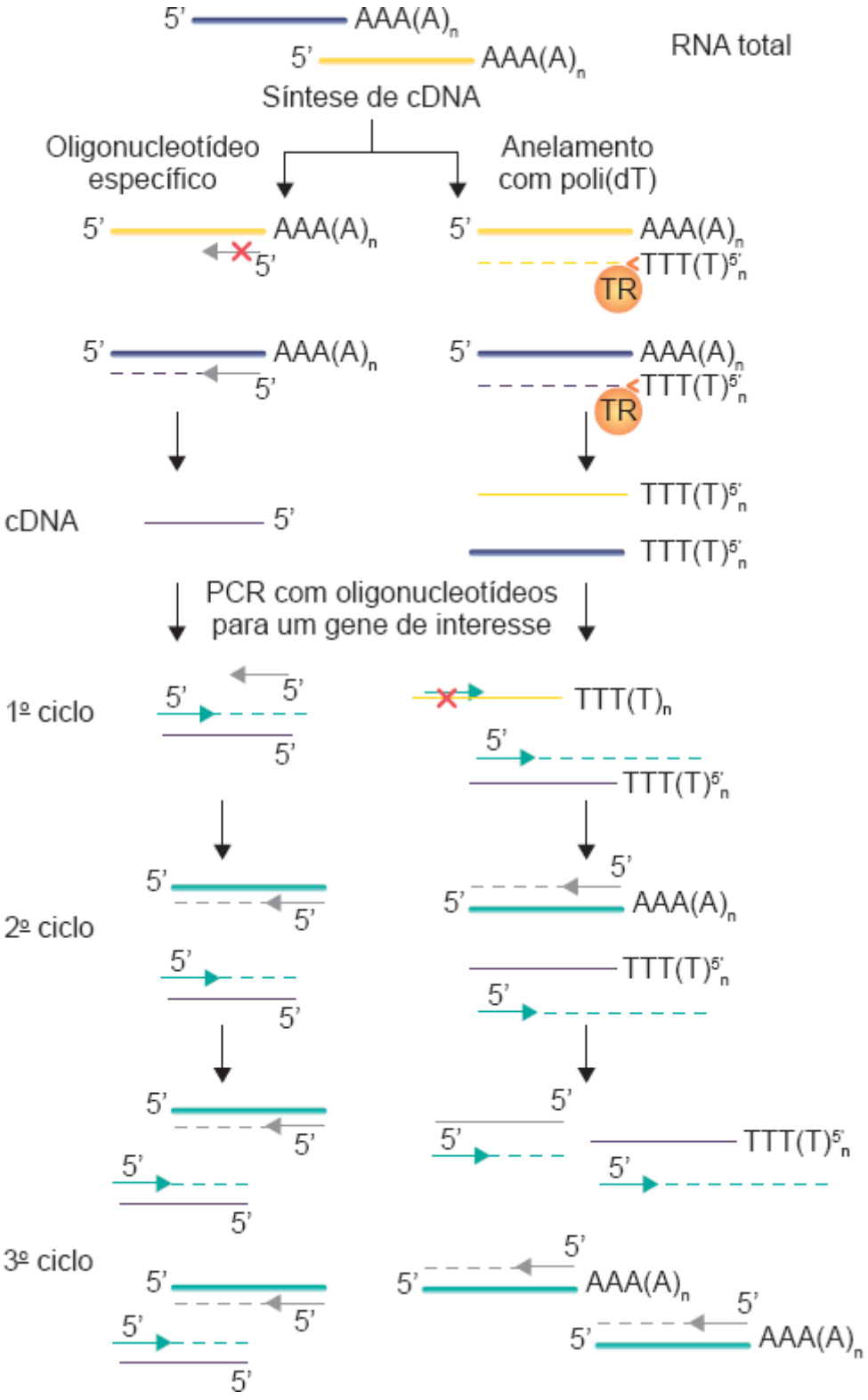


Figura 11.18 Esquema da amplificação de um transcrito de interesse por RT-PCR. Após a síntese do cDNA específico (lado esquerdo da figura) ou do cDNA total (lado direito da figura) pela transcriptase reversa (TR), os cDNA servem de molde em uma reação de PCR padrão com oligonucleotídeos para o gene de interesse (indicado pelas setas verdes).

PCR em tempo real usa marcadores fluorescentes para detectar a quantidade de DNA em cada ciclo de amplificação

Dois tipos de marcadores fluorescentes são os mais utilizados: os que intercalam na dupla-fita de DNA recém-formada após a extensão; e sondas fluorescentes, que hibridam em uma pequena região específica do DNA fita simples durante as fases de anelamento e extensão (Figura 11.19). Nenhum deles interfere na eficiência da reação de amplificação, característica essencial para sua funcionalidade.

A molécula SYBR Green® é o marcador intercalante mais utilizado atualmente e se liga a qualquer DNA fita dupla, o que causa aumento na sua fluorescência. Este tipo de marcação é inespecífico, ou seja, detecta qualquer DNA que esteja sendo amplificado na reação, podendo ser utilizado para qualquer gene de interesse. Uma desvantagem do uso desses marcadores é que eventuais produtos inespecíficos da reação também são detectados, interferindo na análise.

A marcação por sonda fluorescente detecta apenas o produto de interesse, não sofrendo a influência de produtos inespecíficos formados na reação. As sondas do tipo TaqMan® são muito utilizadas nesses casos, e contêm uma molécula fluorescente na extremidade 5' e outra que absorve a fluorescência na extremidade 3' (no termo em inglês, *quencher*) (Figura 11.19). Assim, quando a sonda está intacta, o *quencher* está bem próximo da fonte de fluorescência e não permite que a fluorescência emitida seja detectada. Conforme a reação progride e surgem novas fitas simples, a sonda hibrida nessas fitas, mantendo-se associada a elas. Durante a extensão das fitas pela Taq DNA polimerase, sua atividade exonucleásica 5'→3' cliva a sonda, liberando a molécula fluorescente, que se distancia de seu *quencher*. A molécula passa a emitir a fluorescência, cuja intensidade é proporcional à quantidade de DNA presente na reação. Outras sondas fluorescentes utilizadas em PCR em tempo real são as do tipo *Scorpions* e *Molecular Beacons*, que utilizam princípios similares de controle da emissão de fluorescência pela proximidade a uma molécula *quencher* descritos para a sonda TaqMan®.

Diferenças nas quantidades de DNA são detectadas por PCR em tempo real

A PCR em tempo real utiliza equipamentos que amplificam e quantificam o produto de amplificação ao mesmo tempo. A alta sensibilidade do método possibilita que os produtos sejam detectados na fase exponencial de amplificação, quando a reação é altamente eficiente e depende apenas da quantidade inicial de molde presente na amostra. Nessa fase, quanto maior a quantidade de molde, menos ciclos de amplificação serão necessários para desenvolver uma quantidade detectável de produto. A quantificação se baseia na análise de um gráfico de amplificação que compara número de ciclos *versus* intensidade de fluorescência (Figura 11.20). O número de ciclos no qual a fluorescência detectada ultrapassa o limite mínimo de detecção é chamado de C_T (Figura 11.20). Quanto mais DNA molde está presente na reação, menos ciclos serão necessários para detecção dos produtos, e menor será o C_T. Para a quantificação absoluta do número de cópias de um gene, o valor de C_T obtido na amostra é comparado a uma curva-padrão de uma amostra de concentração conhecida. No entanto, muitas vezes, desejamos fazer uma quantificação relativa, que compara a quantidade de um gene entre duas condições diferentes. Nesse caso, não é necessária uma curva-padrão, e a comparação entre os C_T de cada condição demonstra a variação nas quantidades relativas entre as condições. Esse tipo de análise é muito usado para confirmar dados de variação de expressão de genes, como os obtidos em ensaios de microarranjos de DNA de estudos de transcriptoma – conforme mostra o Capítulo 13.

Sequenciamento de DNA

A habilidade de determinar a sequência de nucleotídeos de moléculas de DNA é parte fundamental da biologia molecular moderna. Frederick Sanger e Walter Gilbert, idealizadores das primeiras técnicas de sequenciamento, foram laureados com o Prêmio Nobel de Química, em 1980, juntamente com Paul Berg, um dos pioneiros na tecnologia do DNA recombinante. Nesta seção, é possível observar que as técnicas de sequenciamento vêm sendo alvo de intensas pesquisas, e estão em constante desenvolvimento. Enquanto as primeiras inventadas podiam determinar a sequência de apenas dezenas ou, no máximo, centenas de bases de cada vez, as mais modernas podem determinar sequências de genomas inteiros em questão de dias.

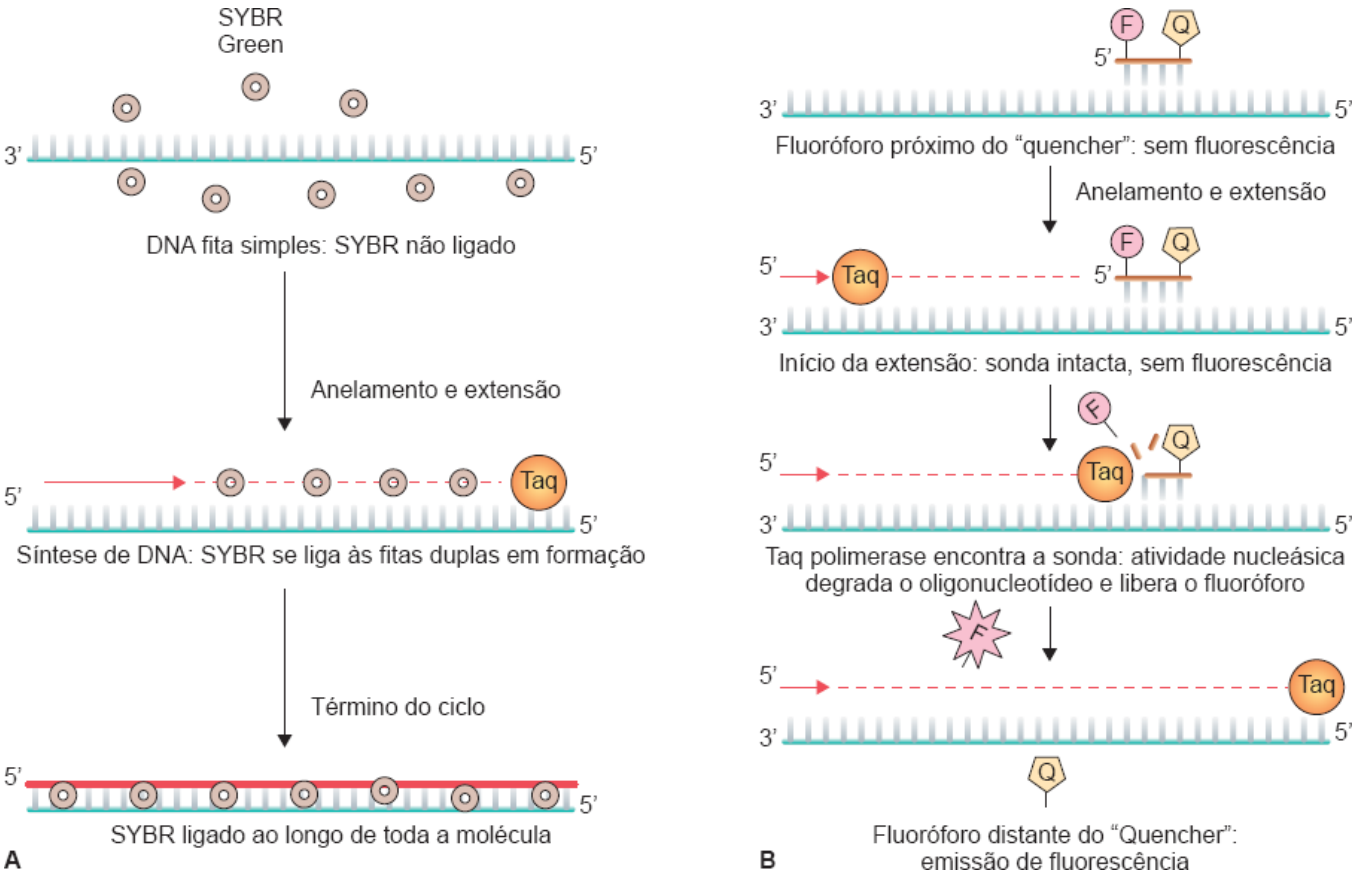


Figura 11.19 Representação dos dois métodos mais usados para detecção dos produtos de amplificação durante a PCR em tempo real. **A.** Marcador inespecífico: SYBR Green®. A fluorescência é detectada logo após a fase de extensão de cada ciclo, quando o DNA está em fita dupla. **B.** Marcador específico: TaqMan®. A fluorescência é detectada após a extensão pela Taq DNA polimerase.

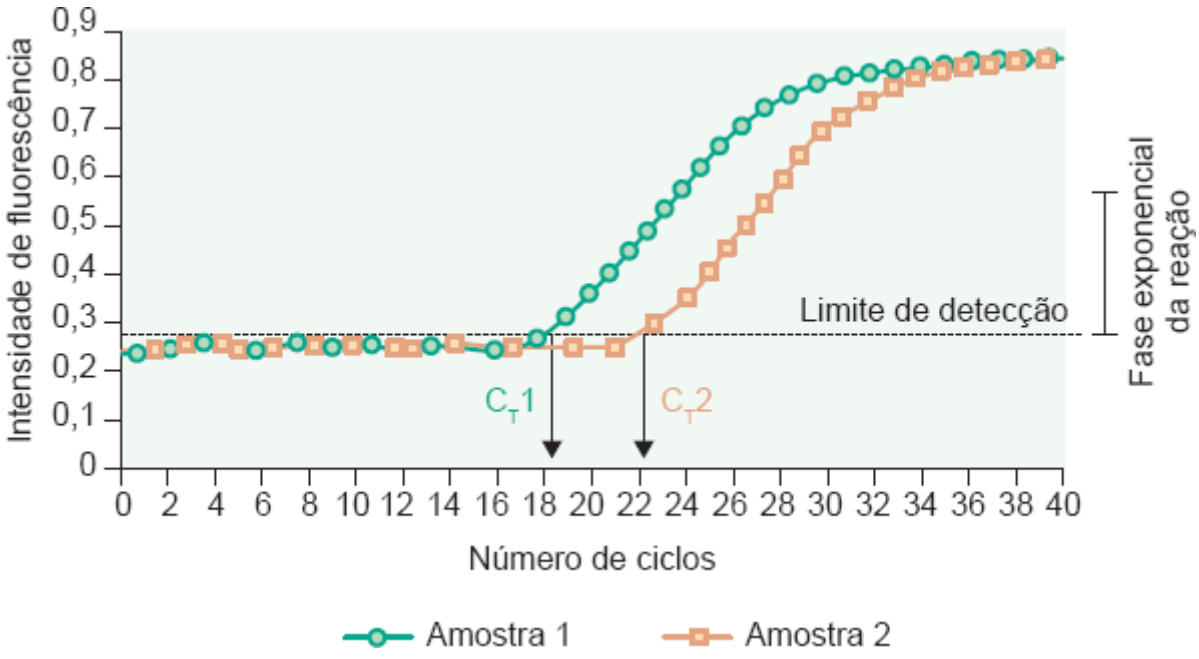


Figura 11.20 Gráfico representando a variação da fluorescência com relação ao número de ciclos de amplificação durante PCR em tempo real. A amostra 2 contém menos quantidade de molde inicial, pois o valor C_T é maior que o obtido na amostra 1.

Sequenciamento químico de Maxam e Gilbert

Um dos primeiros métodos para sequenciamento foi a técnica desenvolvida por Maxam e Gilbert. Bastante laboriosa, requer o uso de reagentes extremamente tóxicos e, por isso, caiu em desuso, sendo largamente suplantada pelos métodos de terminação de cadeia. Contudo, seu valor histórico é enorme, e segue sendo usada em experimentos específicos. O sequenciamento de Maxam e Gilbert é conhecido como *sequenciamento químico*, e se

baseia em quatro tratamentos químicos diferentes, cujo resultado final é a formação de quebras no DNA em localizações específicas. Um dos tratamentos promove uma clivagem a 5’ de G, outro a 5’ de G ou A, outro a 5’ de T, e outro a 5’ de C ou T. Um exemplo de sequenciamento de um pequeno trecho de DNA por essa técnica é mostrado na Figura 11.21. O fragmento de DNA em fita simples a ser sequenciado é marcado na ponta 5’ com fosfato radioativo, e então submetido aos quatro tratamentos separadamente, e os produtos de cada reação são analisados por eletroforese desnaturante em géis de poliacrilamida, o que possibilita a separação de fragmentos diferindo de apenas um nucleotídeo. A exposição autorradiográfica desses filmes torna possível a detecção do padrão de bandas e, a partir daí, a sequência do DNA pode ser facilmente deduzida, com a “leitura do gel”.

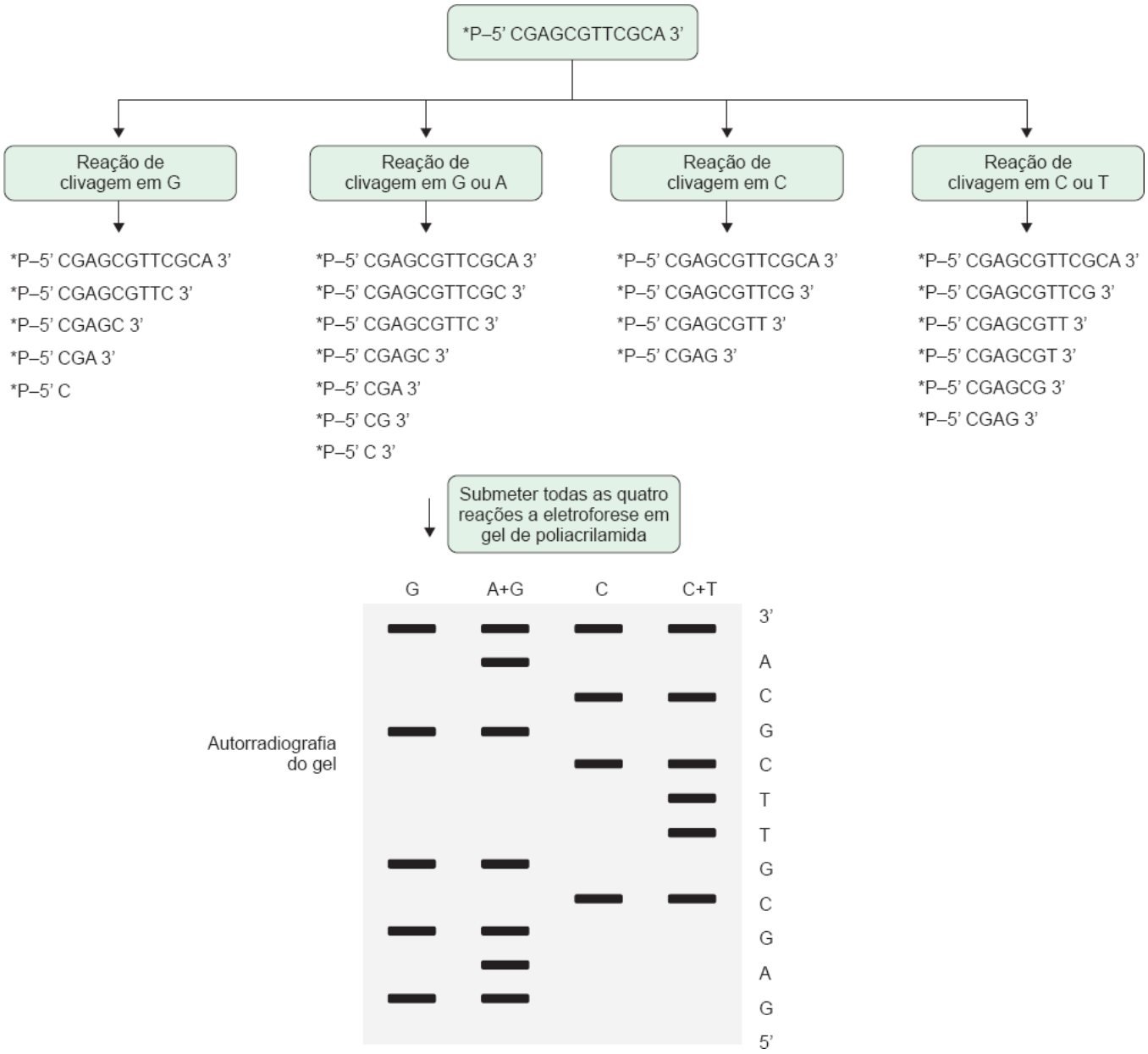


Figura 11.21 Sequenciamento de um pequeno fragmento de DNA pelo método de Maxam e Gilbert.

Sequenciamento enzimático de Sanger envolve reações de polimerização do DNA e terminação de cadeia por didesoxirribonucleotídeos

O método de sequenciamento mais usado até o momento foi desenvolvido por Frederick Sanger. Assim como a tecnologia do DNA recombinante e a PCR, tal técnica também faz uso de enzimas purificadas – neste caso, DNA polimerases. Para compreender melhor como funcionam os métodos de terminação de cadeia, é preciso recordar a estrutura dos nucleotídeos e o modo de ação das DNA polimerases. Desoxirribonucleotídeos são substratos para DNA polimerase, que requer extremidades 3’-OH livres para a síntese de DNA. A estrutura de um desoxirribonucleotídeo está representada na Figura 11.22. Durante o sequenciamento de DNA de Sanger, são adicionados didesoxirribonucleotídeos (ddNTP) às reações de síntese de DNA *in vitro*. Estes são diferentes dos

desoxirribonucleotídeos por apresentarem uma substituição do grupo 3'-OH por um hidrogênio (Figura 11.22). Logo, esses nucleotídeos podem ser utilizados normalmente pela DNA polimerase durante a síntese de DNA; contudo, uma vez incorporados, impedem que a síntese de DNA continue, por não terem grupo 3'-OH, essencial para a formação de uma ligação fosfodiéster subsequente.

O sequenciamento de Sanger original utiliza quatro reações diferentes de polimerização de DNA *in vitro*. Cada uma dessas reações contém um excesso de desoxirribonucleotídeos normais (dNTP) e um dos quatro didesoxirribonucleotídeos. Tome como exemplo a reação feita com dNTP + ddATP mostrada na Figura 11.23. Eles são misturados ao fragmento de DNA que se deseja sequenciar, juntamente com um iniciador marcado radioativamente na sua ponta 5', e uma DNA polimerase. A DNA polimerase começará a incorporar os nucleotídeos na síntese de uma fita complementar, mas eventualmente incorporará o ddATP, causando a interrupção da síntese. Imagine que essa reação está sendo feita com milhares de cópias idênticas desse fragmento de DNA ao mesmo tempo no tubo. Assim, o produto final é uma coleção de fragmentos de tamanhos diferentes, todos eles terminados em A por incorporação do ddATP em diferentes posições. Agora considere que foram feitas reações também com ddCTP, ddGTP e ddTTP. As reações são então submetidas à eletroforese em gel de poliacrilamida, e as bandas do gel reveladas por autorradiografia. É possível então determinar a sequência de nucleotídeos do DNA em questão (Figura 11.23).

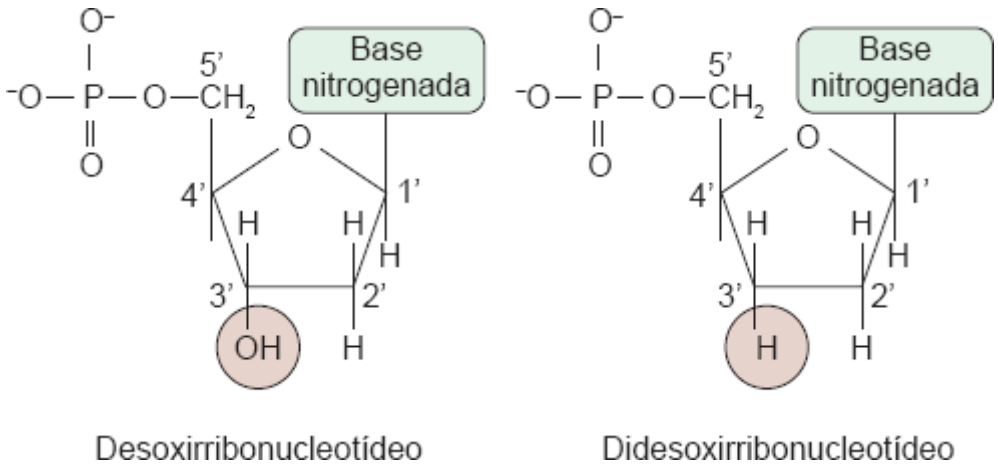


Figura 11.22 Comparação entre as estruturas de um desoxirribonucleotídeo e um didesoxirribonucleotídeo. Os grupamentos presentes na porção 3' de cada um estão destacados pelos círculos marrons.

É importante notar que o sequenciamento de Sanger requer o uso de uma DNA polimerase. Consequentemente, para que ocorra a reação de sequenciamento, é necessário um iniciador que se anele na região imediatamente a montante (ou seja, a 5') da região de interesse. Já foi mostrado neste capítulo que pequenos oligonucleotídeos em fita simples podem ser sintetizados *in vitro*, um processo rotineiro também no uso da tecnologia da reação em cadeia da polimerase. No entanto, a síntese de um iniciador para sequenciamento requer um conhecimento prévio da sequência que está imediatamente a 5' da região de interesse. Como em muitas situações tal informação não está disponível, frequentemente um fragmento de DNA de sequência desconhecida é primeiramente clonado em vetores plasmidiais. Uma vez obtido o clone, ele pode ser sequenciado usando-se iniciadores que anelam na região do plasmídeo a 5' do sítio onde o fragmento foi clonado, tornando possível, assim, o sequenciamento pelo método de Sanger.

O método de Sanger foi aprimorado com o uso de didesoxirribonucleotídios fluorescentes

O método de Sanger foi adaptado para o uso de marcadores fluorescentes. Essa metodologia se tornou muito popular e é até hoje o método mais usado para sequenciamento de DNA. É interessante notar que o salto qualitativo dado com essa nova metodologia foi muito grande, sendo possível somente graças aos avanços em outros campos do conhecimento humano, como a engenharia e a informática, o que possibilitou a construção de máquinas e computadores poderosos nos quais este sequenciamento é feito.

O método fluorescente também se baseia em didesoxirribonucleotídeos, mas, neste caso, uma única reação é feita, em vez das quatro reações necessárias no método original de Sanger. Isso é possível porque cada um dos quatro didesoxirribonucleotídeos é conjugado a uma molécula fluorescente diferente, sem prejudicar seu uso pela

DNA polimerase. Pode parecer uma mudança sutil, mas essa simplificação representou um avanço enorme na tecnologia, pois a redução no número de reações por um fator de 4 vezes facilitou a automação do sequenciamento, característica fundamental para processos de larga escala, tais como o sequenciamento de genomas completos. Centenas de genomas de vírus, bactérias, fungos e até mesmo o genoma humano foram sequenciados usando esta tecnologia.

Na Figura 11.24 é mostrado o esquema do sequenciamento usando fluorescência. Nessa reação, o DNA a ser sequenciado é misturado com o iniciador, DNA polimerase, dNTP, e ddNTP. O diferencial é que cada um dos ddNTP é marcado com um corante fluorescente que emite luz em um comprimento de onda diferente. Assim, a terminação de cadeia causada pela incorporação de ddATP desenvolve fragmentos com uma cor de fluorescência; aquelas causadas pela incorporação de ddCTP dão origem fragmentos de outra cor, e assim por diante.

Nos aparelhos modernos, os produtos dessa reação são separados por eletroforese em géis de poliacrilamida contidos em pequenos capilares, e a detecção da fluorescência de cada banda é feita pelo próprio aparelho, que transmite os dados para um computador acoplado. A fluorescência emitida por cada banda é lida em picos de intensidade, e o resultado desse processo é a produção de um *cromatograma de sequência* (Figura 11.24). Cada reação individual pode sequenciar até aproximadamente 800 pb, e a maioria dos equipamentos atuais possibilita a análise dos produtos de muitas reações em paralelo, por apresentarem vários capilares independentes.

Sequenciamento de nova geração | Obtenção de dados em larga escala

Imagine o potencial de uma tecnologia capaz de sequenciar o genoma humano inteiro em questão de dias. A perspectiva de uma medicina individualizada, a partir do conhecimento das predisposições genéticas de cada paciente, pode parecer ficção científica, mas talvez aconteça em um futuro mais próximo do que se pensa. Neste ponto, é desnecessário dizer que muitos dilemas éticos também deverão ser encarados quando essas metodologias se encontrarem largamente difundidas.

Não é difícil imaginar que existe um interesse comercial enorme nesse tipo de tecnologia. Assim, várias companhias de biotecnologia desenvolveram plataformas diferentes, todas capazes de realizar sequenciamento de alta escala; ou seja, de vários gigabases de DNA por corrida do aparelho. Neste capítulo serão discutidas apenas algumas plataformas mais populares, embora existam outras em uso e/ou desenvolvimento.

É importante salientar que a complexidade dessas metodologias é enorme, mas serão explicadas aqui porque seu impacto no futuro da biologia e da medicina será, sem dúvida, avassalador. Antes de discutir cada uma delas, pode-se apontar o que elas têm em comum. O sequenciamento de nova geração (*next-generation sequencing*) é também chamado frequentemente de *sequenciamento de segunda geração* ou *sequenciamento maciço em paralelo*. Este último ilustra bem o princípio básico por trás dessas plataformas de próxima geração, que é precisamente sequenciar várias moléculas diferentes (ou vários fragmentos diferentes provenientes do mesmo genoma) em paralelo. Logo, elas diferem fundamentalmente das metodologias de Sanger e de Maxam e Gilbert, em que uma única sequência de DNA, oriunda de um trecho específico, é obtida em cada reação. Isso é possível graças à integração de ferramentas de nanotecnologia, robótica e informática em aparelhos de alto desempenho.

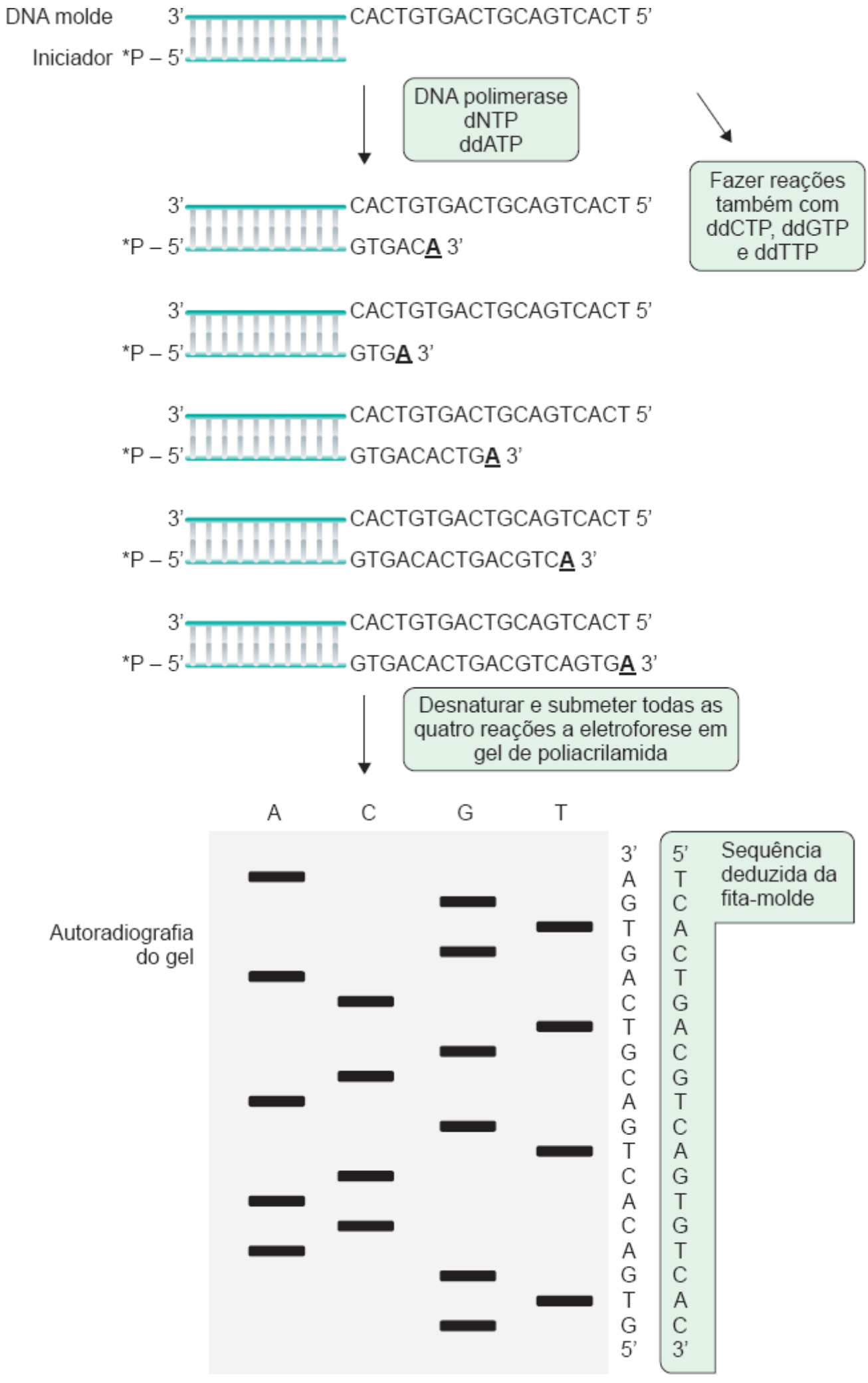
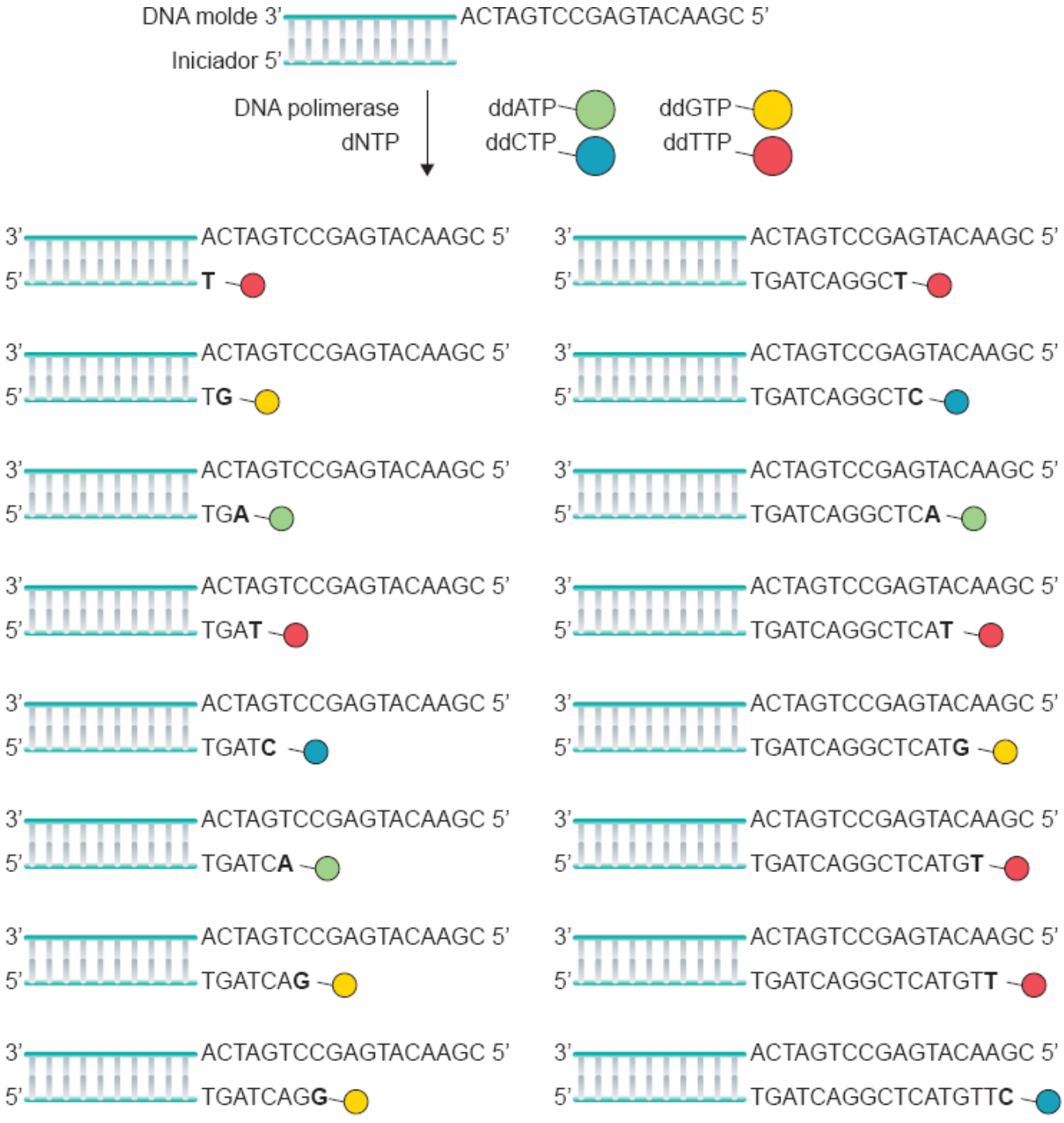

DNA molde
3'
CACTGTGACTGCAGTCACT 5'
Iniciador
*P – 5'
DNA polimerase
dNTP
ddATP
Fazer reações também com ddCTP, ddGTP e ddTTP
3'
CACTGTGACTGCAGTCACT 5'
*P – 5'
GTGACA 3'
3'
CACTGTGACTGCAGTCACT 5'
*P – 5'
GTGA 3'
3'
CACTGTGACTGCAGTCACT 5'
*P – 5'
GTGACACTGA 3'
3'
CACTGTGACTGCAGTCACT 5'
*P – 5'
GTGACACTGACGTCA 3'
3'
CACTGTGACTGCAGTCACT 5'
*P – 5'
GTGACACTGACGTCAGTGA 3'
Desnaturar e submeter todas as quatro reações a eletroforese em gel de poliacrilamida
A
C
G
T
Autoradiografia do gel
3' A G T G A C T G C A G T C A C A G T G 5'
5' T C A C T G A C G T C A G T G T C A C 3'
Sequência deduzida da fita-molde

Figura 11.23 Sequenciamento de DNA pelo método de Sanger. Notar que, para simplificação, apenas a reação com ddATP foi mostrada, mas reações similares são feitas com ddCTP, ddGTP e ddTTP.

As diferentes plataformas apresentam limitações diferentes e, em geral, obtém-se apenas a sequência de pequenos fragmentos de DNA, por limitações das técnicas que vão além do escopo deste texto. No entanto, é importante considerar a capacidade assombrosa dessas tecnologias: em seus modelos mais avançados, cada corrida desses aparelhos pode sequenciar dezenas de gigabases de DNA. Para ter uma ideia do que isso representa, vale relembrar os tamanhos dos diferentes genomas. Por exemplo: o genoma da bactéria *Escherichia coli* tem 4,6 Mpb de DNA, e o genoma humano tem aproximadamente 3 Gpb de DNA.

Em primeiro lugar, para compreender o modo como esses métodos de segunda geração mais usados funcionam, serão examinados os seus princípios fundamentais em comum:

1. Construção de uma *biblioteca* que represente o genoma em análise. Ao contrário das bibliotecas genômicas e de cDNA vistas anteriormente, ela não é produzida por clonagem em vetores. Essas bibliotecas são, na verdade, compostas de milhões de fragmentos de DNA diferentes, isolados fisicamente uns dos outros, e cada um deles multiplicado em várias cópias de pequenos agrupamentos (chamados *clusters*) de moléculas idênticas.



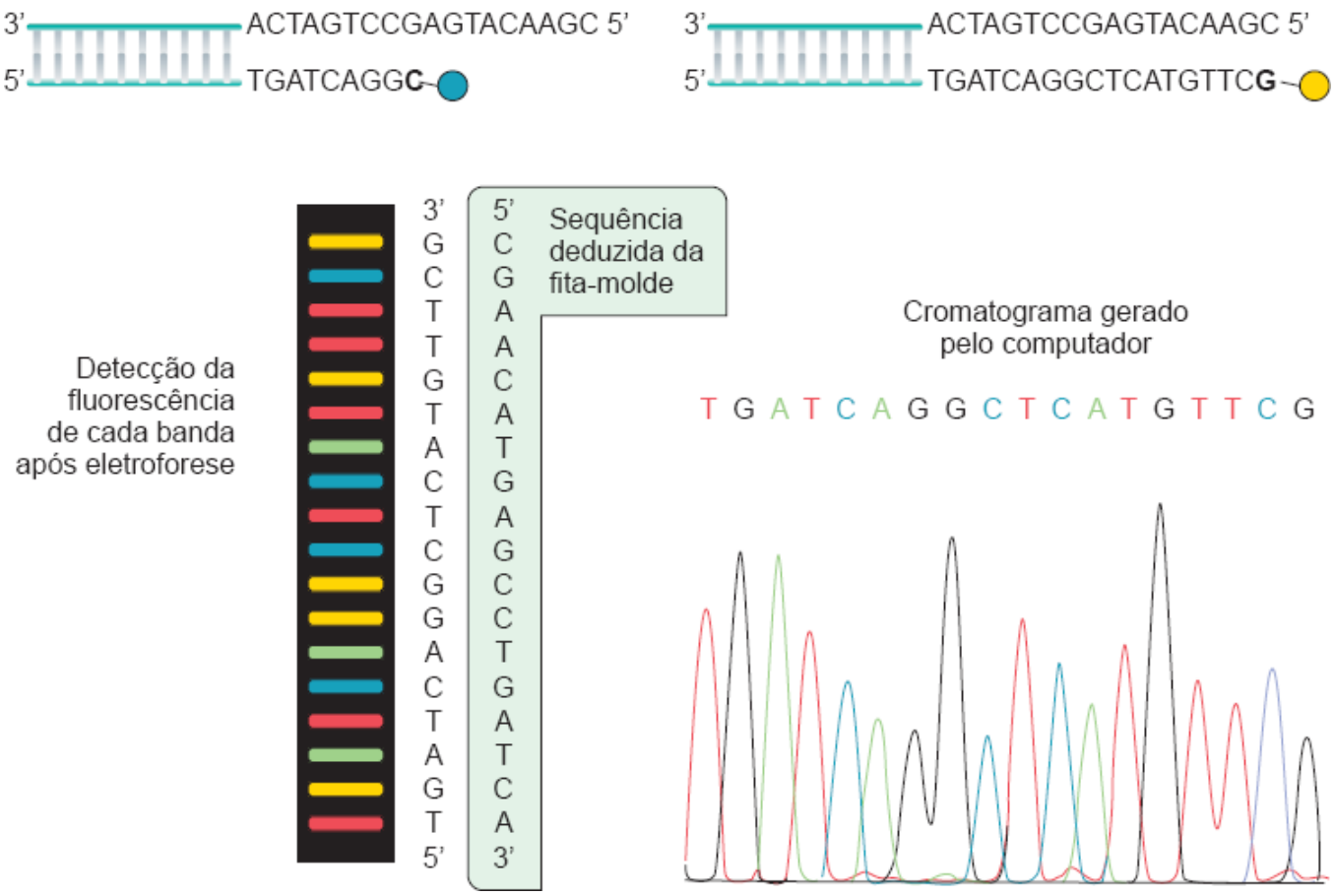


Figura 11.24 Sequenciamento de DNA pelo método de Sanger usando didesoxirribonucleotídeos fluorescentes.

2. Cada um desses pequenos *clusters* é construído por fragmentação do DNA original em pequenos fragmentos (geralmente por força mecânica) e ligação de adaptadores (pequenos trechos de DNA de sequência conhecida) nas duas extremidades dessas moléculas. A coleção de fragmentos ligados aos adaptadores é chamada de "biblioteca" no jargão do sequenciamento de nova geração, embora não represente uma biblioteca de fragmentos de DNA clonados em vetor. As sequências conhecidas desses adaptadores são usadas para a produção de várias cópias de cada molécula por *PCR*. A produção de várias cópias em *cluster* é fundamental, pois amplifica o sinal obtido pelas reações de sequenciamento subsequente.
3. Detecção do processo de sequenciamento em tempo real. Ou seja, a cada ciclo de reação nesses aparelhos, se obtém informação sobre uma ou mais bases de cada um dos milhões de fragmentos de DNA que estão sendo sequenciados em paralelo. Essa detecção ocorre pela emissão de luz ou fluorescência por cada um dos *clusters* da biblioteca. As emissões de cada um dos *clusters* são registradas pelo computador a cada ciclo, obtendo a sequência de cada um deles ao final do processo.
4. Em todas as metodologias, o resultado final é a sequência de milhões de pequenos trechos de DNA. A tarefa de remontar um genoma inteiro a partir de vários fragmentos pequenos é um desafio enorme na bioinformática, e vários avanços vêm sendo realizados nesse sentido. Contudo, tais metodologias já então sendo largamente empregadas para várias aplicações. É possível usar os dados desse sequenciamento para detectar polimorfismos no genoma humano, por comparação das sequências pequenas obtidas com o genoma humano de referência. Além disso, os sequenciamentos de nova geração também têm enorme valia nos estudos de expressão gênica.

Plataforma Illumina

A construção de bibliotecas para sequenciamento com a plataforma Illumina requer a fragmentação do DNA e a ligação de adaptadores a suas extremidades. Os fragmentos de DNA são ligados à superfície de uma lâmina de vidro. Nesta, já estão ligadas milhões de cópias de oligonucleotídeos complementares aos adaptadores, que servem como iniciadores para a amplificação local dos fragmentos por PCR, conforme ilustrado na Figura 11.25. Essa "amplificação-ponte", que ocorre com DNA se dobrando sobre si mesmo, produz dezenas de milhões de pequenos *clusters* de fragmentos espalhados pela placa.

O processo de sequenciamento usa uma DNA polimerase e nucleotídeos modificados com duas características: primeiro, cada um dos quatro diferentes nucleotídeos é marcado com corantes fluorescentes diferentes. Além disso, esses nucleotídeos têm o terminal 3' bloqueado quimicamente, e tanto a fluorescência quanto o bloqueio 3' podem ser revertidos por tratamentos químicos. O sequenciamento nessa plataforma envolve a adição dos nucleotídeos marcados e bloqueados, que são então incorporados pela DNA polimerase. Como seu terminal 3' está bloqueado, apenas uma incorporação é possível de cada vez para cada *cluster*. A fluorescência emitida por cada cluster é lida pelo sistema óptico acoplado e registrada no computador. Na sequência, uma etapa de tratamento químico remove tanto o bloqueio 3' dos nucleotídeos, quanto a fluorescência conjugada. Assim, um novo ciclo pode ser iniciado, adicionando-se novamente nucleotídeos fluorescentes bloqueados na porção 3'. Esta metodologia foi batizada de *sequenciamento por síntese*. A Illumina pode produzir uma sequência de milhões de pequenos fragmentos de até 150 pb em seus aparelhos mais modernos.

Plataforma Roche 454

A plataforma Roche 454 foi a pioneira em termos comerciais, portanto, tem importância histórica. Foi utilizada até 2016, quando cessou a comercialização da plataforma.

A construção de bibliotecas para o sequenciamento com a plataforma 454 tem como base o princípio de PCR em emulsão. Basicamente, o DNA fragmentado é ligado a adaptadores e misturado a um conjunto de minúsculas esferas de agarose contendo milhares de cópias de oligonucleotídeos ligados covalentemente a sua superfície (Figura 11.26). Estes são complementares às sequências dos adaptadores. Um óleo é misturado a esses componentes em solução aquosa contendo os reagentes necessários para uma reação de PCR, resultando na formação de pequenas micelas desse óleo. A estequiometria dos reagentes é ajustada de modo que a maioria das micelas irá conter apenas uma esfera com um único fragmento de DNA ligado por complementariedade de bases aos iniciadores. Essa mistura de micelas é submetida a ciclos de amplificação por PCR e, dessa maneira, cada uma produzirá milhares de cópias de cada fragmento individual, todas ligadas à esfera de agarose. A mistura de esferas de agarose é colocada em microplacas que contêm poros de tamanho suficiente para acomodar uma única esfera. O material está pronto para ser submetido à reação de pirosequenciamento.

Conforme explicado anteriormente, os nucleotídeos precursores da síntese de DNA são trifosfatados. Quando usados pela DNA polimerase, um *pirofosfato* é liberado, e a energia proveniente da hidrólise da ligação deste com o restante do nucleotídeo é usada para a formação da nova ligação fosfodiéster. O princípio do pirosequenciamento é a detecção da liberação de pirofosfato durante síntese de DNA. Logo, esse método de sequenciamento também envolve uma reação com iniciadores e DNA polimerase *in vitro*. Para detecção da liberação de pirofosfato, utiliza-se uma sequência de reações enzimáticas (Figura 11.26), que culmina com a emissão de luz pela enzima luciferase.

Para o sequenciamento, esferas ainda menores são adicionadas, contendo todas as enzimas necessárias ao pirossequenciamento ligadas. Soluções com os diferentes nucleotídeos são adicionadas sequencialmente. Imagine uma placa com milhares de microporos, cada um deles contendo um único *cluster* de moléculas de DNA idênticas. Ao adicionar dATP, por exemplo, naqueles fragmentos propriamente anelados com um iniciador e cuja próxima base é T, a DNA polimerase irá incorporar dATP, liberando pirofosfato e culminando com a emissão de luz. Assim, um leitor óptico acoplado a um computador registra todos os microporos nos quais houve emissão de luz. Ao realizar diversas vezes os ciclos com todos os nucleotídeos, é possível determinar sequências de bases dos fragmentos contidos em cada microporo. Assim, as sequências de milhões de diferentes fragmentos pequenos são geradas em paralelo. A plataforma 454 pode sequenciar fragmentos de até 700 pb.

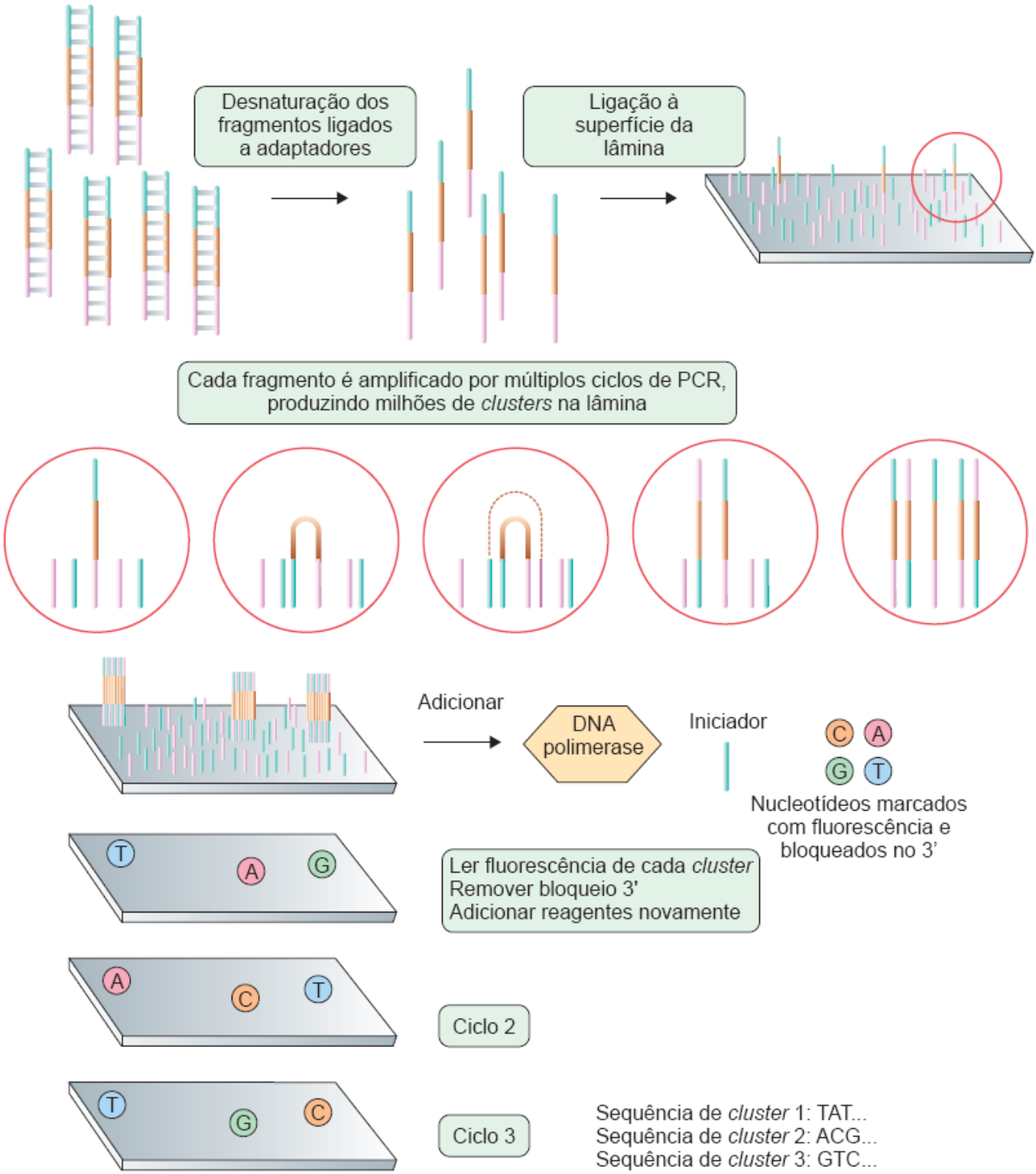


Figura 11.25 Sequenciamento de DNA na plataforma Illumina. Apenas três *clusters* são mostrados para simplificação.

Plataforma Ion Torrent (Thermo)

A plataforma Ion Torrent é, em muitos aspectos, similar à plataforma 454 da Roche. Assim como na metodologia 454, PCR em emulsão é usada para a amplificação das bibliotecas, e as esferas com milhares de cópias de cada fragmento são separadas uma das outras em pequenos poços. A metodologia de sequenciamento também envolve a detecção de produtos oriundos da atividade da DNA polimerase. A incorporação de um nucleotídeo à cadeia nascente a partir do dNTP libera não só o pirofosfato, mas também um íon H^+. A detecção da liberação de H^+ em cada rodada é a base desse sequenciamento. Assim, se um nucleotídeo é adicionado por vez, a incorporação do mesmo em determinado poço é detectada pela liberação de H^+, uma vez que o chip onde o sequenciamento ocorre funciona como um semicondutor, transformando o sinal detectado em voltagem, registrada pelo aparelho. O processo é repetido diversas vezes, utilizando-se cada um dos quatro dNTP sequencialmente. Assim como na

plataforma 454, o sequenciamento de regiões homopoliméricas (repetições de um nucleotídeo, como a sequência 5'-GGGGG-3', por exemplo) apresenta maior taxa de erro.

Considerações finais e perspectivas

As técnicas básicas de análise e manipulação de ácidos nucleicos descritas neste capítulo permitiram um enorme avanço na compreensão da genética e da biologia dos organismos vivos, sendo o arcabouço para o desenvolvimento de novas metodologias de alta resolução e larga escala que estão surgindo muito rapidamente. Lembre-se, por exemplo, de que a técnica de PCR foi usada pela primeira vez em 1985 e, em apenas 30 anos, passou a ser aplicada em metodologias como o sequenciamento rápido de gigabases de DNA e a quantificação de ácidos nucleicos em alta sensibilidade e tempo real em praticamente qualquer amostra biológica de interesse. Hoje é possível, portanto, determinar em poucos dias a sequência completa de um genoma humano, abrindo novas perspectivas para a medicina baseada nas características genéticas de cada indivíduo. Da mesma forma, é possível determinar todo o conjunto de microrganismos presente em um ambiente (microbioma) a partir do sequenciamento de todo o DNA ali encontrado.

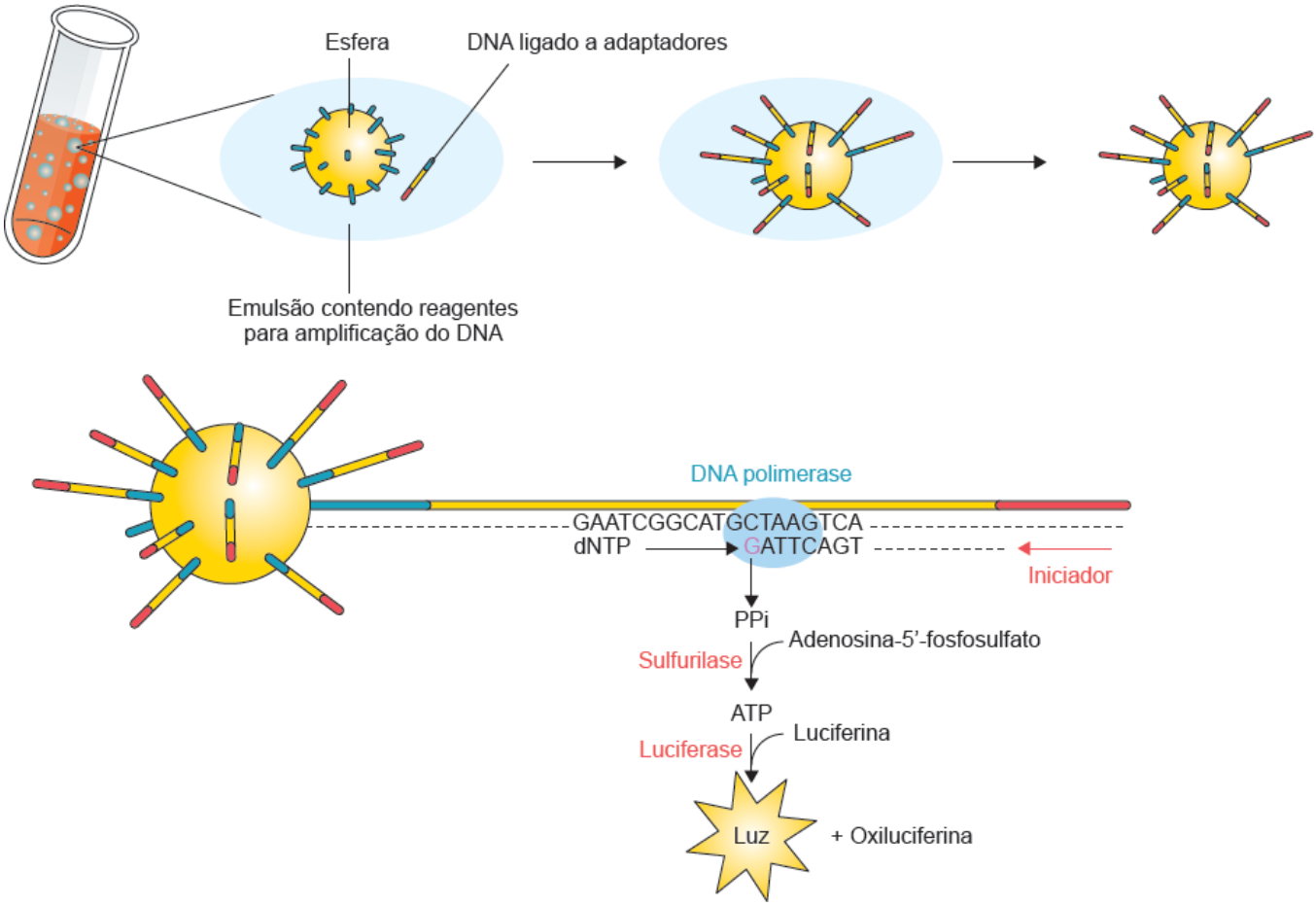


Figura 11.26 Sequenciamento de DNA na plataforma 454. Na parte superior, é mostrado um esquema de produção das esferas contendo um *cluster* de fragmentos idênticos. Na parte inferior, a detecção de cada ciclo de pirosequenciamento é detalhada.

Novas áreas de estudo também se desenvolveram em função dessas tecnologias nos últimos anos, como a Biologia Sintética, que busca sintetizar genes e até criar genomas para introduzir características de interesse. Agora é possível, por exemplo, sintetizar genes inteiros ligados a vetores de interesse, assim como obter versões com mutações alterando individualmente cada um dos aminoácidos da proteína por ele codificada.

Independentemente das novas metodologias que surgirão, as técnicas descritas neste capítulo são as mais usadas para análise e manipulação de ácidos nucleicos. A compreensão dos princípios básicos dessas técnicas é fundamental para estudos em diversas áreas da Biologia moderna.

Bibliografia

Brown TA. Gene cloning and DNA analysis: an introduction. 6. ed. Oxford: Blackwell Scientific Publishers; 2010.

Hutchison CA 3rd. DNA sequencing: bench to bedside and beyond. Nucleic Acids Res. 2007; 35(18):6227-37.

Maxam AM, Gilbert W. A new method for sequencing DNA. Proceedings of the National Academy of Sciences of the United States of America. 1977;74(2):560-4.

Mullis KB. The unusual origin of the polymerase chain reaction. Sci Am. 1990;262(4):56-61,64-5.

Mullis KB, Faloona FA. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. Meth Enzymol. 1987;155:335-50.

Roberts RJ. How restriction enzymes became the workhorses of molecular biology. Proc Natl Acad Sci U S A. 2005;102(17):5905-8.

Saiki RK, Scharf S, Faloona F, Mullis KB, Horn GT, Erlich HA, Arnheim N. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science. 1985;230(4732):1350-4.

Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America. 1977;74(12):5463-7.

12 Genômica | A Revolução na Análise dos Genes

Claudia Barros Monteiro Vitorello, Erika Cristina Jorge e Luis Eduardo Aranha Camargo

O genoma consiste das informações genéticas de um indivíduo e determina todas as características herdáveis. Neste capítulo, serão abordados os conceitos e as metodologias dos estudos que envolvem o sequenciamento completo de genomas de indivíduos de diversas espécies, as comparações que levam a um melhor entendimento das diferenças e as aplicações do conhecimento que vêm se acumulando há mais de uma década. A ciência genômica possibilitou mudanças nas definições de gene e de regiões regulatórias e na maneira de se estudar e entender a função dos genes. Essas mudanças tiveram como base a análise de sequências de DNA e de proteínas, utilizando métodos computacionais para predição de sequências gênicas e de suas funções. Os métodos computacionais associados à construção de bancos de dados de acesso público aplicados à ciência genômica viabilizaram avanços no conhecimento antes inimagináveis.

Introdução

O termo genômica (*genomics*) é atribuído ao geneticista Thomas H. Roderick (1986) e, inicialmente, tinha como objetivo "promover o entendimento da estrutura, função e evolução de genomas em todos os reinos da vida com aplicações a problemas em biologia e medicina". Depois, o termo apareceu impresso no editorial de Victor McKusick e Frank Ruddle (1987) como parte do primeiro volume da revista científica criada com o mesmo nome (*Genomics*).

A tecnologia de sequenciamento desenvolvida em 1977 concomitantemente por Frederick Sanger e dois outros cientistas, Allan Maxam e Walter Gilbert, possibilita a leitura dos nucleotídeos de pequenos fragmentos de DNA. Por conta do tempo e do custo elevados, somente genes e proteínas de grande interesse dos cientistas tinham suas sequências determinadas. A partir de 1986, a automatização da técnica tornou possível aos cientistas vislumbrar e planejar o sequenciamento completo de genomas, incluindo os cromossomos humanos. A revolução da genômica teve realmente início a partir da publicação das sequências completas dos genomas das bactérias *Haemophilus influenzae* e *Mycoplasma genitalium*, em 1995, marcada por modificações nos passos iniciais do mapeamento genético, que foram substituídos pela estratégia de *shotgun*, proposta por J. Craig Venter (ver boxe "Estratégia de sequenciamento aleatório de genomas (*shotgun* sequencing)" e Figura 12.1). Com essa nova estratégia – a redução significativa nos custos e o aumento na capacidade de sequenciamento dos últimos 15 anos –, o número de sequências de genomas completos disponíveis no início de 2015 passa de 3.000 e inclui aqueles de organismos eucariotos (178), de leveduras a humanos, e procariotos (3.121 bactérias e 179 arqueias). O primeiro genoma de um fitopatógeno totalmente sequenciado foi realizado por laboratórios de pesquisa distribuídos em universidades e institutos do estado de São Paulo, em uma iniciativa em rede virtual inédita no país (ver boxe "O genoma de *Xylella fastidiosa*" e Figura 12.1). Além disso, já foi ultrapassado o número de 1.000 genomas humanos sequenciados, o que torna possível fazer um estudo sem precedentes de variabilidade humana.

Quando o objetivo do projeto é o sequenciamento completo de um genoma, é preciso obter uma cobertura entre 10 e 12 vezes equivalente à sequência completa; ou seja, para a montagem completa de um genoma de 1 milhão de pares de bases, é necessário o sequenciamento das extremidades de 10.000 insertos clonados, considerando que cada sequência obtida tenha pelo menos 700 pb, levando a um total de 10 milhões de pares de bases sequenciadas. A

montagem da sequência original é obtida a partir da sobreposição de sequências comuns nos fragmentos por programas computacionais – é como resolver um quebra-cabeça em que as peças se complementam, não por formato, mas por sequências comuns (Figura 12.2). Um dos programas utilizado para a montagem de genomas completos, disponível para pesquisas acadêmicas livre de custos, foi desenvolvido por Phil Green e é conhecido como Phrap. Ao final da etapa de sequenciamento aleatório, muitas regiões permanecerão ainda não resolvidas. São os chamados *gaps,* regiões entre *contigs*, que, por diversos motivos, permanecem não sequenciadas (Figura 12.3). Nesses casos, estratégias de sequenciamento direto devem ser utilizadas.

A segunda fase da revolução na era genômica apresenta novas tecnologias que, em conjunto, são conhecidas como estratégias *next generation sequencing* (NGS) e contam com equipamentos sofisticados que chegam a determinar a sequência de um bilhão de bases em 1 h. O primeiro relato entre as novas tecnologias data de 2005 (Margulies *et al*., 2005) (Figura 12.4). Comparativamente à descrição feita inicialmente por Sanger, de até 200 bases sequenciadas por dia de trabalho, essas novas tecnologias representam um aumento significativo na capacidade de gerar informação genética. Desde então, o genoma de vários organismos foi completamente sequenciado, marcando essa última década e meia desde o primeiro genoma descrito (Tabela 12.1).

Além das sequências de genomas completos de espécies representantes de todos os reinos, fragmentos de DNA sequenciados de mais de 300 mil espécies também compõem o repertório de informação acessível aos pesquisadores (Figura 12.5). Da mesma maneira, desde 2004, estão sendo realizados experimentos para conhecer a diversidade microbiana e o sequenciamento de regiões genômicas, sem a necessidade de cultivo dos organismos. O número de espécies microbianas conhecidas para cada ambiente varia entre 0,1 e 10%, dependendo do ambiente estudado. Esses microrganismos representam a diversidade possível de cultivar em laboratório. As metodologias de estudo independentes do cultivo em laboratório, juntamente com as técnicas de sequenciamento, têm possibilitado a coleta de informações também sobre o repertório de genes e funções presentes na natureza, antes desconhecidas dos pesquisadores. Essa nova abordagem, além das tecnologias associadas a ela, vem sendo coletivamente chamada de *metagenômica*, termo utilizado pela primeira vez pela pesquisadora Jo Handelsman, em 1998.

Estratégia de sequenciamento aleatório de genomas (*shotgun sequencing*)

Para o sequenciamento de grandes fragmentos de DNA ou projetos de sequenciamento em grande escala, que inclui genomas completos, a estratégia dominante é a conhecida como *shotgun* ou sequenciamento aleatório (Figura 12.1). A estratégia descrita inicialmente por Anderson (1981) e com aplicações na década de 1980 – que incluíam o sequenciamento de insertos completos de clones genômicos, fragmentos gerados por enzimas de restrição ou grandes *amplicons* de PCR – foi modificada para ser utilizada no sequenciamento completo do primeiro genoma bacteriano em 1995 (Fleischmann *et al*., 1995). A estratégia de *shotgun* envolve os seguintes passos:

- Fragmentação ao acaso do DNA de interesse purificado, utilizando uma estratégia que promova quebras aleatórias no DNA como, por exemplo, a fragmentação física pela ação de um sonicador (as extremidades dos fragmentos gerados devem ser reparadas de modo a possibilitar a clonagem)
- Seleção dos fragmentos gerados por tamanho, sendo que a clonagem deve ser realizada em vetores apropriados: vetores plasmidiais (insertos entre 2 e 4 Kpb), vetores cosmidiais (insertos de 40 Kpb), vetores cromossômicos (BAC ou YAC), fragmentos maiores de DNA
- Sequenciamento aleatório das extremidades dos insertos clonados
- Montagem das sequências obtidas por sobreposição de leituras comuns. No processo de fragmentação do DNA, a ordem dos fragmentos é perdida – ou seja, não é conhecida a localização das sequências obtidas no DNA original. Assim, a montagem dos fragmentos em sequências contínuas ou *contigs* (Figura 12.1) deve ser alcançada com a utilização de programas computacionais.

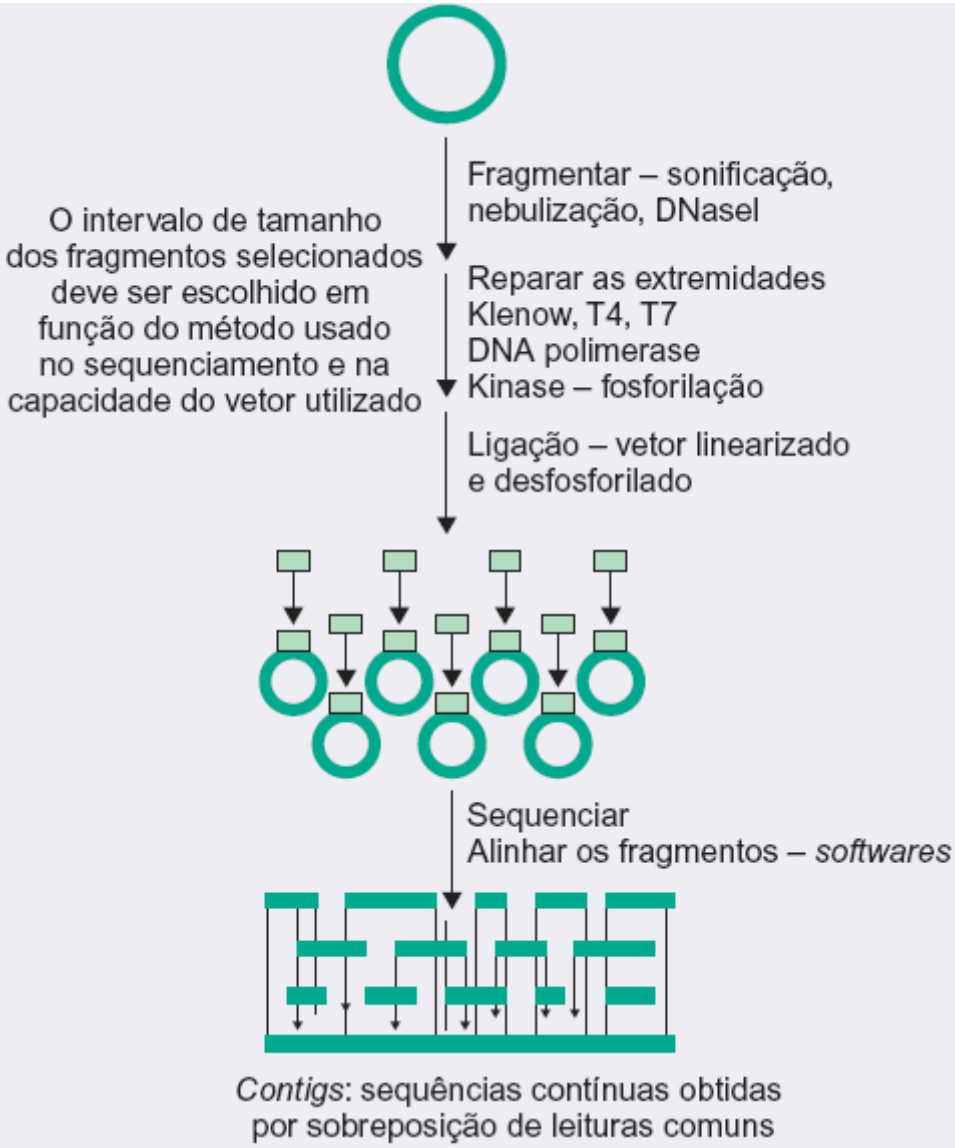


Figura 12.1 Etapas da estratégia de sequenciamento randômico.

O genoma de *Xylella fastidiosa* | O primeiro genoma de uma bactéria fitopatogênica completamente sequenciado

Xylella fastidiosa é o agente causal da doença clorose variegada dos citros (CVC), popularmente conhecida como amarelinho, um dos problemas fitossanitários mais importantes da citricultura brasileira. A linhagem 9a5c de *X. fastidiosa* teve o seu genoma completamente sequenciado e disponibilizado no ano 2000.

Foi o primeiro projeto de sequenciamento completo de genoma realizado no Brasil e o primeiro de fitopatógeno no mundo. Essa iniciativa foi financiada principalmente pela Fapesp, uma agência de fomento e pesquisa, e realizada por 32 grupos de pesquisadores do estado de São Paulo, alocados em laboratórios de universidades estaduais, federais e privadas, e institutos de pesquisa. O projeto foi o marco para o desenvolvimento de competências e infraestrutura adequadas a estudos desta natureza. Essa rede virtual de laboratórios distribuídos pelo estado (ONSA, *Organization for Nucleotide Sequencing and Analysis*) realizou o sequenciamento do genoma de outras cinco bactérias, além de vários projetos de transcriptoma de organismos eucarióticos.

O genoma dessa bactéria compreende um cromossomo circular de aproximadamente 2,7 milhões de bases e dois plasmídeos de 52 Kpb e 1,2 Kpb. Um total de 2.904 genes foi predito na sequência completa de nucleotídeos, e 47% desses genes puderam ter uma função atribuída por homologia a outras sequências depositadas em bancos de dados de acesso público. Dentre as funções associadas à interação e adaptação ao hospedeiro, foram encontrados genes que codificam toxinas, antibióticos, proteínas envolvidas na assimilação de ferro e síntese de exopolissacarídeos e adesinas. O genoma de *Xylella* apresentou também grande quantidade de sequências derivadas de bacteriófagos (7%), indicando que a transferência horizontal de genes tem papel importante na evolução desse patógeno (Simpson *et al*., 2000). Atualmente, há seis outras sequências completas de genomas de diferentes linhagens de *X. fastidiosa* que causam doenças em vários hospedeiros responsáveis por diferentes sintomas.

Os avanços na tecnologia de sequenciamento são acompanhados por avanços na capacidade de processamento e armazenamento das informações produzidas em grande escala. As sequências produzidas são depositadas em bancos de dados de acesso público e disponíveis *on-line* – um vasto reservatório de informações que está ainda a ser decifrado em sua totalidade. Neste capítulo, serão discutidos inicialmente a mudança de paradigma sobre a descoberta de genes e os avanços quanto à interpretação de suas funções com base na análise comparativa de sequências dos organismos diferentes; as aplicações da genômica e as descobertas de novas funções por meio da metagenômica.

Do sequenciamento de genes a genomas completos | A mudança de paradigma

Gene é a unidade funcional da hereditariedade. O conceito de gene passou por várias modificações ao longo dos anos, desde as descobertas dos "fatores hereditários" por Gregor Mendel (1866) e a criação do termo "gene" por Wilhelm Johannsen (1909), até a definição com uma abordagem mais molecular que incluiu as regiões regulatórias ao conceito de gene. Assim, gene pode ser definido como uma unidade de informação genética que controla a síntese de um polipeptídeo ou de uma molécula de RNA. Inclui as regiões 5' e 3' não codificantes que estão envolvidas na regulação da transcrição e tradução e todas as regiões intrônicas daquele gene – definição esta posterior aos estudos envolvendo o sequenciamento do DNA.

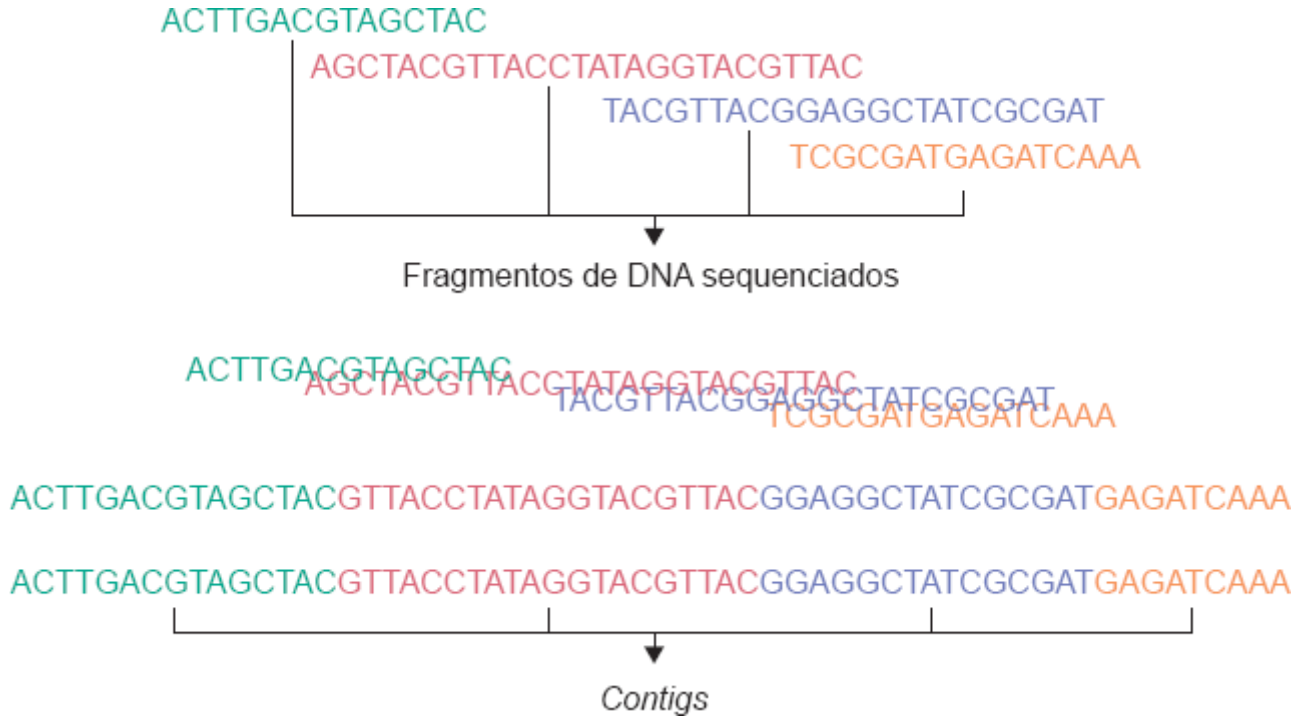


Figura 12.2 Representação da montagem dos *contigs* por meio da sobreposição de sequências.

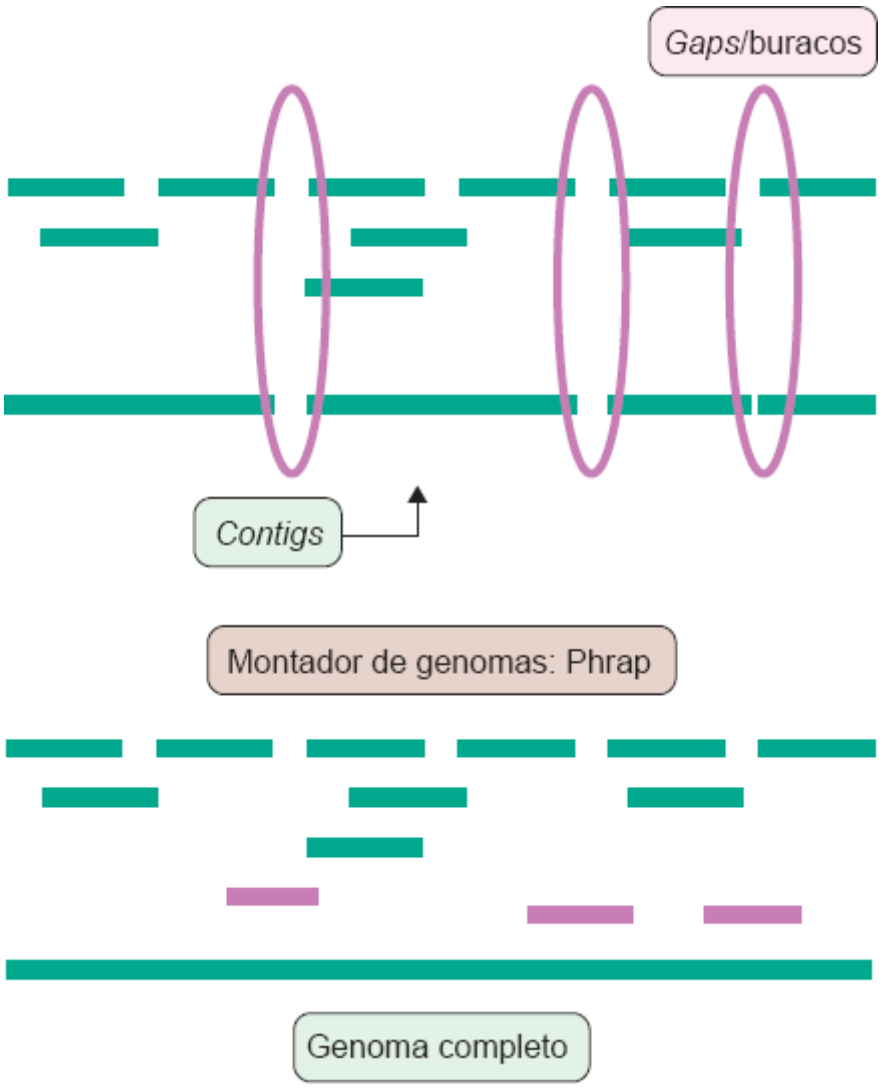


Figura 12.3 Representação da montagem de um genoma completo, passando pela utilização de programas computacionais que, pela sobreposição de sequências, recuperam a sequência contínua do DNA original. As linhas rosas representam estratégias de sequenciamento direto para revolver *gaps* entre os *contigs*.

Até o surgimento da genômica, a descoberta de um gene e sua função tinha como base a presença de variabilidade fenotípica em uma população para uma determinada característica, como estudado nos trabalhos de Mendel. Muitas vezes, uma vida científica inteira poderia ser dedicada à geração do conhecimento sobre a função de uma determinada proteína em um dado organismo, sendo o ápice do trabalho representado pela clonagem e sequenciamento do gene em questão. É interessante lembrar que, nessa época, a maioria das sequências disponíveis era de genes e/ou proteínas cujas funções eram conhecidas, ou seja, determinadas experimentalmente.

Atualmente, com o sequenciamento completo de genomas, pode-se fazer predições quanto ao repertório total de genes de um dado organismo sem conhecer previamente as funções por ele realizadas ou o seu papel ecológico. É possível, por exemplo, conhecer as vias presentes e ausentes para biossíntese de todos os aminoácidos, vitaminas e cofatores a partir da presença ou ausência de genes codificadores de enzimas associadas a essas vias. Ou a presença e ausência de genes para cada uma das categorias funcionais conhecidas responsáveis pela manutenção celular; ou, ainda, a identificação de novas características antes desconhecidas para um determinado grupo taxonômico. Um exemplo bastante interessante se refere a conhecer o repertório de genes contido no genoma da bactéria *Deinococcus radiodurans* que, como o próprio nome diz, é um organismo resistente à radiação, sendo 200 vezes mais tolerante à radiação ionizante e 20 vezes mais tolerante à radiação ultravioleta que a bactéria *Escherichia coli* (White *et al*., 1999).Além de ser interessante do ponto de vista biológico, essa característica faz com que *D. radiodurans* tenha grande potencial para ser utilizada em programas de biorremediação de sítios contaminados. Várias descobertas foram feitas analisando o genoma dessa bactéria com relação à sobrevivência desse microrganismo nesse ambiente inóspito. Aparentemente, *D. radiodurans* utiliza um mecanismo de reparo de DNA altamente redundante em número de genes, apresenta um sistema diferenciado para exportar nucleotídeos danificados para fora da célula e um funcionamento diferente do mecanismo de reparo induzido pela RecA – em

relação ao que geralmente se encontra em outras bactérias. Esses dados obtidos a partir do genoma estão ainda sob investigação para entender todos os mecanismos envolvidos nesse estilo de vida tão especial. Assim, a genômica revela os genes que devem ser estudados *a posteriori* para a comprovação fenotípica. Contrariamente ao que se fazia no início, nas descobertas de novos genes.

Bancos de dados | Organização da informação genômica

O avanço rápido e crescente do número de bases depositadas em bancos de dados reflete as inovações tecnológicas em pleno desenvolvimento, ficando evidente a necessidade de organizar essa informação para análises *a posteriori*.

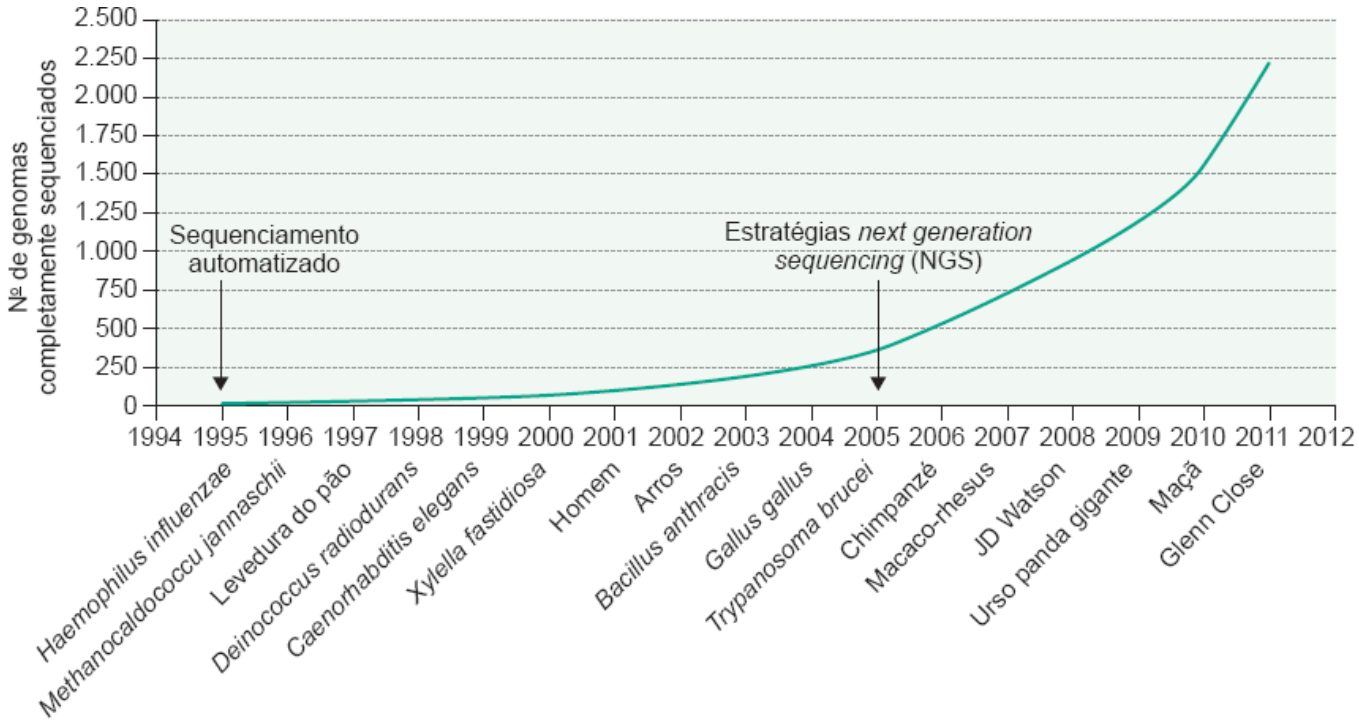


Figura 12.4 Linha do tempo das metodologias de sequenciamento e os diferentes organismos que tiveram o seu genoma completamente sequenciado.

O cenário atual, no entanto, é muito diferente daquele quando foi preparado o primeiro catálogo impresso de sequências, ou *Atlas of Protein Sequence and Structure*, em 1965, por Margaret Dayhoff, pesquisadora do National Biomedical Research Foudation (NBRF), Washington D.C., e que continha a sequência de aminoácidos de 70 proteínas. A Dra. Dayhoff estabeleceu o primeiro banco de dados *on-line* em 1980, contendo 1.660 sequências de proteínas, com acesso via telefone e que empregava um sistema sofisticado, para a época, para recuperação das sequências pelo usuário. Tal atlas constituiu a base do que, a partir de 1984, foi chamado de *Protein Information Resource* (PIR). Dayhoff foi também uma das pioneiras no desenvolvimento de ferramentas computacionais para a análise de sequências de proteínas e nucleotídeos.

Ainda, ao final da década de 1970, discussões lideradas pelas instituições de pesquisa National Science Foudation (NSF*)* e Rockefeller University, nos EUA, tiveram como objetivo estabelecer uma base de dados internacional para o armazenamento de sequências de DNA e proteínas. Em 1982, foi criado o GenBank, com financiamento do National Institute of Health (NIH). Após a criação do GenBank, discussões anteriormente iniciadas no *European Molecular Biology Laboratory* (EMBL) levaram ao estabelecimento de um segundo banco de dados, o *EMBL Nucleotide Sequence Database*. Rapidamente, os grupos (norte-americano e europeu) perceberam a necessidade de uma colaboração para a troca de informações entre si e também com um terceiro banco de dados, criado na Ásia com a mesma finalidade, o *DNA Data Bank of Japan* (DDBJ). Esses três bancos de dados armazenam, desde 1987, todas as sequências de nucleotídeos e proteínas geradas por diferentes grupos ao redor do mundo e dos mais diversos organismos. A comunicação sincronizada diariamente dos dados depositados em qualquer um dos três bancos é mantida por meio de uma colaboração conhecida como *International Nucleotide Sequence Database Collaboration* (INSDC). A visualização gráfica e as ferramentas de análise de sequências variam entre os sítios de acesso dos três bancos de dados; no entanto, as informações de sequenciamento são

exatamente as mesmas. Acompanham as sequências depositadas nesses bancos de dados: informações sobre a função biológica, quando existente; dados taxonômicos do organismo a que pertence; características da sequência inferidas pelos autores como, por exemplo, presença de uma sequência codante, intergênica ou intrônica; publicação científica mais relevante; e dados dos autores e suas instituições de origem.

A criação dos bancos de dados foi acompanhada pelo desenvolvimento de ferramentas para a análise destas, dando origem aos bancos de dados secundários, que capturam dados de bancos primários, como os citados anteriormente, e acrescentam análises de vários tipos. Uma das ferramentas mais utilizadas, conhecida como BLAST, consiste em um programa de busca por comparação de sequências em um determinado banco de dados. No grupo de cientistas pioneiros, responsáveis pela criação de banco de dados secundários, destaca-se o Dr. Amos Bairoch, fundador do SwissProt. Em 1984, Bairoch disponibilizou a primeira versão do pacote computacional NAPDB (*Nucleic-acid and Protein Data Bank*), contendo 15 programas para análise de sequências e 1.200 sequências de proteínas. O NAPDB foi o embrião do que viria a ser o banco de dados mais completo do mundo, que compreende informações manualmente curadas (ou seja, informações que foram avaliadas por especialistas em cada área do conhecimento) sobre proteínas, o *UniProt Knowledgebase* (Swiss-Prot KB).

A bioinformática tornou-se parte integrante da biologia molecular a partir do grande número de sequências de nucleotídeos e proteínas disponíveis nesses bancos de dados contendo informações biológicas que não poderiam ser compreendidas sem o auxílio de ferramentas computacionais. Bioinformática é um termo utilizado para descrever o uso de métodos computacionais, matemáticos e estatísticos na organização e análise de dados biológicos, em grande escala e de maneira integrada (Figura 12.6).

São inúmeros os bancos de dados e ferramentas de análises disponíveis com acesso *on-line*, e uma visão geral do que pode ser encontrado é publicada anualmente em dois volumes da revista científica *Nucleic Acids Research*. Alguns dos bancos de dados são apresentados na Tabela 12.2.

Tabela 12.1 Alguns dos genomas sequenciados que marcaram a era genômica.

Organismo	Domínio	Características	Ano
Haemophilus influenzae	Bacteria	Patógeno humano, primeiro organismo com genoma completamente sequenciado	1995
Methanocaldococcus jannaschii	Archeae	Hipertermófilo, isolado de fonte termal, de regiões profundas do oceano, EUA	1996
Saccharomyces cerevisiae	Eukarya	Levedura do pão, linhagem de laboratório, primeiro genoma de eucarioto completamente sequenciado	1997
Deinococcus radiodurans	Bacteria	Isolado de carne enlatada irradiada	1999
Caenorhabditis elegans	Eukarya	Nematoide	1998
Xylella fastidiosa	Bacteria	Isolado de laranja doce, variedade Valência, Brasil. Primeiro organismo fitopatogênico com o genoma completamente sequenciado	2000
Homo sapiens	Eukarya	Primeiro *draft* do genoma humano	2001
Oryza sativa ssp. Japônica	Eukarya	Arroz	2002
Takifugu rubripes	Eukarya	Peixe baiacu asiático	2002
Bacillus anthracis	Bacteria	Agente causador do antraz	2003
Gallus gallus	Eukarya	Galinha	2004
Trypanosoma brucei	Eukarya	Protozoário	2005
Psychrobacter arcticus	Bacteria	Isolado de pergelissolo congelado há mais de 20 milhões de anos, Sibéria,	2005

		Rússia	
Entamoeba histolytica	Eukarya	Protozoário ameba	2005
Pan troglodytes verus	Eukarya	Chimpanzé comum	2006
Canis lupus familiaris	Eukarya	Cão Boxer	2007
Macaca mulatta	Eukarya	Macaco-*rhesus*	2007
Vitis vinifera	Eukarya	Videira	2007
Homo sapiens JD Watson	Eukarya	Primeiro genoma de um homem conhecido sequenciado completamente	2008
Ornithorhynchus anatinus	Eukarya	Ornitorrinco	2008
Woolly-Mammoth	Eukarya	Mamute extinto	2008
Ailuropoda melanoleura	Eukarya	Urso panda gigante	2009
Equus caballus	Eukarya	Cavalo	2009
Zea mays	Eukarya	Milho	2009
Malus domestica	Eukarya	Maçã *Golden Delicius*	2010
Neanderthal	Eukarya	Ossos de Neandertal de Vindija Cave, Croácia	2010
Homo sapiens Glenn Close	Eukarya	Primeiro genoma de uma mulher conhecida sequenciado completamente	2010
Sarcophilus harrisii	Eukarya	Demônio da Tasmânia	2011
Solanum tuberosum L.	Eukarya	Batata	2011

Primeiros frutos da genômica de procariotos

Em 1969, em importante artigo publicado na revista *Science*, Robert Whittaker propôs um esquema de classificação dos seres vivos composto pelos Reinos Plantae (Metaphyta), Animalia (Metazoa), Fungi, Protista (Protoctista) e Monera. Tratava-se de um sistema inovador em relação aos anteriores, pois reconheceu o valor de características celulares até então pouco considerados na classificação dos seres vivos, como a presença/ausência de núcleo organizado, modo de nutrição da célula (ingestão, absorção ou nutrição autotrófica) e organização celular (pluri ou unicelular). Nesse sistema, popularmente conhecido como os cinco reinos de Whittaker, os seres vivos foram classificados em dois grandes grupos: de um lado, os eucariotos, agrupados em quatro reinos; do outro, os procariotos, agrupados no Reino Monera. O sistema de Whittaker, por isso, pode ser visto como um sistema dicotômico, pois enxerga duas vias evolutivas: uma de eucariotos e outra de procariotos.

No final da década de 1970, no entanto, teve início outra reclassificação dos seres vivos, com profundas consequências para o Reino Monera de Whittaker. Essa "revolução" teve como pioneiros Carl Woese e G. E. Fox (1977). Em artigo publicado na revista da Academia de Ciências dos EUA, esses dois pesquisadores sugeriram que "a evolução progride de maneira quantizada". Levando em conta que Woese era físico e não biólogo, com o termo "quantizar" emprestado da física, os autores expressaram a visão da evolução como um processo contínuo, cujo progresso se verifica de maneira discreta, em que um nível de organização dá origem a outro, mais complexo. Esses níveis foram denominados domínios. Assim, parte dos procariotos e todos os eucariotos representam, respectivamente, os domínios Bacteria e Eukarya. A grande contribuição foi a identificação de um terceiro domínio que, até então, havia passado insuspeito, que abriga a outra parte dos procariotos. Tal domínio foi inicialmente denominado Archaeabacteria, cujo nome posteriormente foi reduzido para Arqueia, para não induzir ao erro que esses seres se assemelham mais a bactérias que aos eucariotos do ponto de vista genético, visto que arqueias e bactérias são semelhantes do ponto de vista morfológico. A proposta desse novo sistema de classificação tem

alcance mais profundo no nosso modo de interpretar a evolução da vida na Terra, pois reconhece uma terceira "via" evolutiva (em vez das duas preconizadas por Whittaker), representada por seres que, embora procarióticos, em muitos outros aspectos se diferenciam das bactérias e se aproximam mais dos seres eucarióticos.

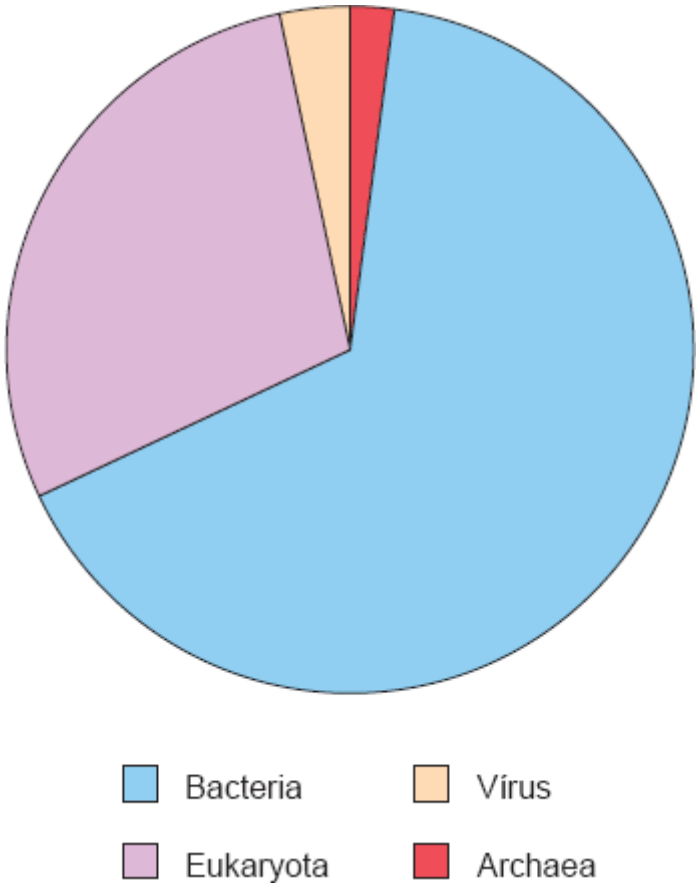


Figura 12.5 Distribuição taxonômica das sequências depositadas no banco de dados TrEMBL.

A base da revolução de Woese está na forma inédita pela qual os organismos foram analisados. Esta reside na análise comparativa das sequências dos genes que codificam para as moléculas de RNA (em especial, os genes que codificam as subunidades 16S ou 18S) presentes nos ribossomos (para relembrar, essas moléculas de RNA se combinam com um conjunto de proteínas específicas, resultando na montagem do ribossomo). A escolha desses genes, obviamente, não foi por acaso. Para comparar geneticamente seres que são evolutivamente muito distintos, a escolha de um gene deve recair sobre aquele que seja universal (ou seja, que desempenhe a mesma função e que esteja presente em todos os organismos). Além disso, a taxa de modificação da sequência desse gene deve ser lenta, de modo a possibilitar comparações entre organismos evolutivamente muito distantes. Se adotássemos um gene de evolução rápida, a comparação de sequências entre um eucarioto e procarioto, por exemplo, não seria possível, pois teriam divergido além do ponto onde fosse possível qualquer comparação.

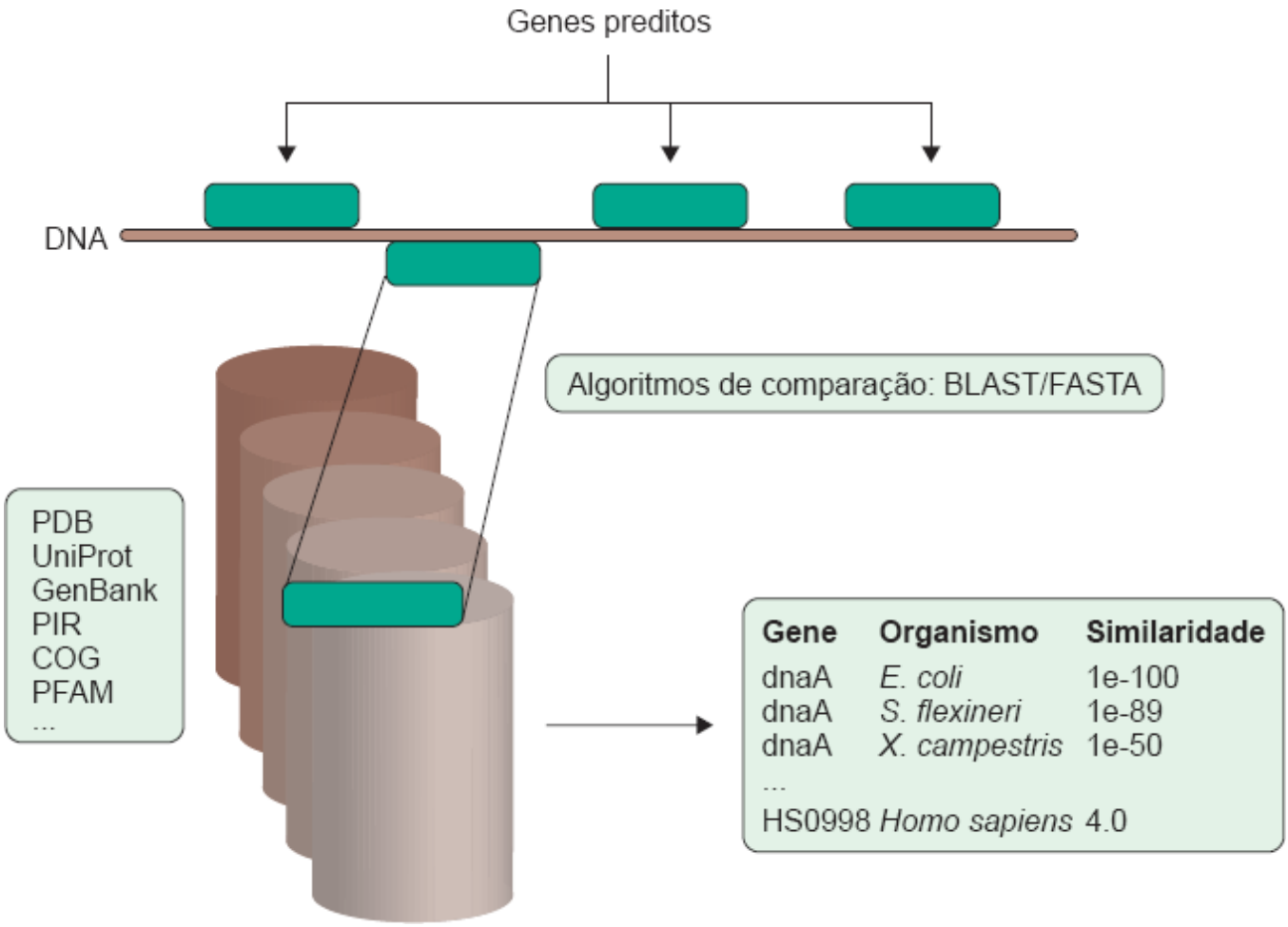


Figura 12.6 Representação esquemática da organização e tipos de análise dos dados biológicos acumulados nos bancos de dados de sequência.

Atualmente, parece lógico que a escolha do gene para reconstrução da filogenia dos seres vivos devesse recair sobre os genes do 16S (no caso de procariotos) ou 18S rRNA (no caso de eucariotos), mas não foi o caso na época, visto que a ideia de usá-los não foi aceita de imediato. Aliás, para muitos, a ideia de uma terceira via evolutiva era absurda e, além disso, muito incômoda, pois atestava que os biólogos haviam ignorado ou classificado de maneira errônea, até então, um terço das formas vivas do planeta. Woese seguiu a sina de vários cientistas que provaram estar muito adiante de sua época e, por muito tempo, foi criticado por importantes membros da comunidade científica, incluindo o professor de zoologia da Universidade de Harvard (EUA), Ernst Mayr. Esse ilustre evolucionista reconhecia que as arqueias eram, de fato, diferentes das bactérias, mas relutava em dar-lhes um *ranking* semelhante ao dado aos eucariotos. Sobre isso, o cientista afirmou:

> Ambas as subdivisões dos procariotos [Bacteria e Archaeabacteria] compartilham a ausência de um núcleo, além da mitose [...] Eu não tenho dúvida de que os eucariotos são derivados das archaeabactérias [...] No entanto, dar às archaeabactérias o mesmo ranking que os eucariotos seria produzir uma classificação totalmente distorcida.

Tabela 12.2 Alguns bancos de dados secundários mais utilizados em genômica.

Banco	Características	Endereço web
UniProt	Informações manualmente curadas sobre a função de proteínas	http://www.ebi.ac.uk/interpro
KEGG	Informações sobre o posicionamento de enzimas em vias metabólicas	http://www.genome.jp/kegg/
ENZYME	Informações sobre classificação enzimática	http://ca.expasy.org/enzyme/
Merops	Informações sobre peptidases	http://merops.sanger.ac.uk/
TCDB	Informações sobre proteínas transportadoras de membrana	http://www.tcdb.org/

OMA	Grupos de genes ortólogos	http://omabrowser.org
CGAP	Cancer Genome Anatomy Project	http://cgap.nci.nih.gov/
MIPS	Informações sobre análise comparativa de genomas de eucariotos	http://www.helmholtz-muenchen.de/en/ibis
Ensembl	Informações sobre análise comparativa de genomas de eucariotos	http://www.ensembl.org/index.html
MBGD	*Microbial Genome Database for Comparative Analysis.* Informações sobre análise comparativa de genomas de procariotos	http://mbgd.genome.ad.jp/
CMR	*Comprehensive Microbial Resource.* Informações sobre análise comparativa de genomas de procariotos	http://cmr.jcvi.org/tigr-scripts/CMR/CmrHomePage.cgi
IMG	*The Integrated Microbial Genomes System.* Informações sobre análise comparativa de genomas de procariotos	http://img.jgi.doe.gov/
SGD	*Saccharomyces Genome Database.* Informações específicas sobre *Saccharomyces*	http://www.yeastgenome.org/
FlyBase	*A Database of Drosophila Genes & Genomes.* Informações específicas sobre *Drosophila*	http://flybase.org/
XDB	*The Xylella fastidiosa comparative genome database.* Informações específicas sobre *Xylella*	http://www.xylella.lncc.br

Sequenciamento do genoma de *Methanococcus jannaschii*

Woese e Fox, em seu memorável artigo citado anteriormente, já reconheciam a importância da determinação da sequência completa do genoma de um organismo para ajudar a traçar seu histórico evolutivo de maneira mais precisa do que simplesmente analisar a sequência de um ou dois genes. "O genoma de um organismo parece ser o último registro de seu histórico evolutivo". Nesse contexto, o sequenciamento do genoma de *Methanococcus jannaschii*, em 1996, foi um fato marcante na consolidação de sua tese da existência de uma terceira via evolutiva, ilustrando, já naquele momento, o impacto que a genômica viria a ter no estudo dos seres vivos.

Como várias das arqueias, *M. jannaschii* é um organismo extremófilo, pois vive em condições extremas – nesse caso, junto às fontes hidrotermais oceânicas, a 2.600 metros abaixo do nível do mar. É anaeróbio e sobrevive com uma dieta um tanto peculiar, reduzindo dióxido de carbono na presença de hidrogênio, resultando em metano. O sequenciamento de seu genoma (o terceiro genoma de um procarioto a ser sequenciado) foi aguardado com muito interesse, pois, afinal, poderíamos saber se as arqueias eram de fato tão diferentes das bactérias quanto a análise dos genes ribossômicos dava a entender – justificando, assim, a criação de um domínio somente para elas. À parte desse fato, o Departamento de Energia dos EUA (que financiou o projeto) tinha interesse em conhecer um pouco mais sobre esse procarioto metanogênico, na esperança de poder usá-lo, de alguma maneira, como fonte geradora de energia renovável no futuro.

Para a redenção de Carl Woese, o genoma de *M. janaschii* mostrou-se muito diferente dos genomas das bactérias *Haemophilus influenza* e *Mycoplasma genitalium*. De acordo com um dos coautores do projeto de sequenciamento, J. Craig Venter observou:

> Nós ficamos pasmos ao verificar que dois terços dos genes [de M. janaschii] não se parecem com nada que tenha sido visto até agora em biologia. Isso encerra a questão se arqueias são formas distintas e separadas de vida.

Por meio da análise comparativa de sequências gênicas, chegou-se à conclusão de que *H. influenza* e *M. genitalium* contêm 83% dos genes em comum. No entanto, quando os genes de *M. janaschii* (1.738 genes arranjados em 1,7 milhão de pares de nucleotídeos) foram comparados com os dessas duas bactérias, essa porcentagem foi de apenas 11% no primeiro caso e 17% no segundo. Ou seja, ficou evidente que os genes dessa arqueia eram bem

distintos, o que deu amplo suporte para a tese do terceiro domínio de Woese – ele estava certo: até então, uma importante linha evolutiva da vida na Terra tinha passado completamente despercebida.

Com base no sequenciamento completo, foi possível verificar, dentre outras características, que as arqueias guardam semelhanças com eucariotos em nível molecular com relação à replicação do DNA, transcrição e tradução. O códon de iniciação desses organismos é uma metionina, ao passo que, em bactérias, é uma versão modificada desse aminoácido, denominado formilmetionina. Arqueias também contêm TATA- boxes, que orientam a RNA polimerase para o início da transcrição; e sua RNA polimerase é mais parecida com a RNA polimerase II de eucariotos do que com a de bactérias. Essas características explicam outra semelhança com eucariotos, que é a sensibilidade à toxina diftérica, produzida por *Corynebacterium diphtheriae*, capaz de inibir a síntese protética, mas que não tem efeito em bactérias. Por outro lado, a recíproca é verdadeira; isto é, alguns antibióticos que afetam células bacterianas não afetam células de arqueias e de eucariotos, como a estreptomicina.

Genômica comparativa em procariotos | 1.500 genomas de procariotos sequenciados

Considera-se como o início da era genômica a publicação das sequências completas dos genomas das bactérias *Haemophilus influenza* e *Mycoplasma genitalium*, em 1995. Claro que uma considerável quantidade de sequências genômicas já estava disponível antes desses eventos, como vimos anteriormente. No entanto, a novidade foi a disponibilização da sequência completa desses genomas. A grande diferença é que, com a sequência completa, podemos comparar genomas de organismos diferentes, o que deu origem ao importante ramo da genômica denominado genômica comparativa (GC). A comparação de genomas possibilita identificar genes ortólogos entre duas ou mais espécies; isto é, genes que compartilham um mesmo ancestral. Além disso, é possível identificar genes presentes ou ausentes em determinado genoma em comparação com o outro. Somados, esses dois produtos da GC possibilitam inferências sobre o metabolismo de um dado organismo pouco conhecido e também sobre os caminhos de sua evolução genômica por meio da comparação das sequências de seu genoma com a de outro, mais conhecido. No caso específico de procariotos patogênicos, por exemplo, a GC tem sido especialmente importante para identificar genes presentes ou ausentes em estirpes patogênicas comparadas com estirpes não patogênicas.

Outras sequências genômicas vieram se juntar às sequências dos genomas de *Haemophillus* e *Mycoplasma* de maneira rápida e estonteante. Após um período inicial de crescimento irregular, o acúmulo de sequências genômicas completas de procariotos acusa uma tendência de crescimento exponencial. Em janeiro de 2015, 3.300 genomas estavam completamente sequenciados, de acordo com o European Bioinformatics Institute (EBI), sendo 3.121 de bactérias representantes de vários grupos bacterianos e 179 de arqueia.

Genômica comparativa e predição da função dos genes

A análise das mais de 1.500 sequências de genomas de procariotos tornou possível visualizar um panorama global da organização genômica desses seres que, antes, jamais poderia ser visualizado. A primeira observação diz respeito ao tamanho do genoma de procariotos. Em uma escala de tamanho, genomas de vírus são, em média, menores que os de procariotos, que, por sua vez, são menores que os de eucariotos; mas isso considerando valores médios, pois é possível encontrar alguns genomas bacterianos menores que os de alguns vírus e maiores que de alguns eucariotos.

Os menores genomas de procarioto encontrados até o momento são genomas de proteobactérias gama intracelulares simbiontes de insetos, denominados psilídeos, que sugam a seiva elaborada de plantas. São eles: *Candidatus* Hodgkinia *cicadicola*, com apenas 143 mil pares de bases (143 Kpb) e *Carsonella rudii*, com 154 Kpb; e o maior é o de *Sorangium cellulosum*, também uma proteobactéria, mas de vida livre e habitante do solo, com aproximadamente 13 Mpb (Figura 12.7). A título de comparação, genomas virais variam entre 3.569 nucleotídeos (bacteriófago MS2) e 1,2 milhão de pares de bases, no caso de um mimivírus parasita de amebas (ou seja, aproximadamente 10 vezes maior que o de *Carsonella*). O menor genoma eucariótico (2,9 Mpb distribuídos em 11 cromossomos), por sua vez, é o do parasita *Encephalitozoon cuniculi*, pertencente ao grupo dos microsporídeos, que são parasitas unicelulares que produzem esporos (alguns classificam esses organismos como protozoários; outros, como fungos). O maior genoma é o de *Amoeba dubia* (um protozoário), com 670 bilhões de pares de bases. Esse tamanho equivale a nada menos que cerca de 200 vezes o genoma humano (Figura 12.7). O número de genes de eucariotos é, em geral, superior ao encontrado em procariotos, porém, com uma vairação estimada de 6.000 a 40.000 genes (Figura 12.8).

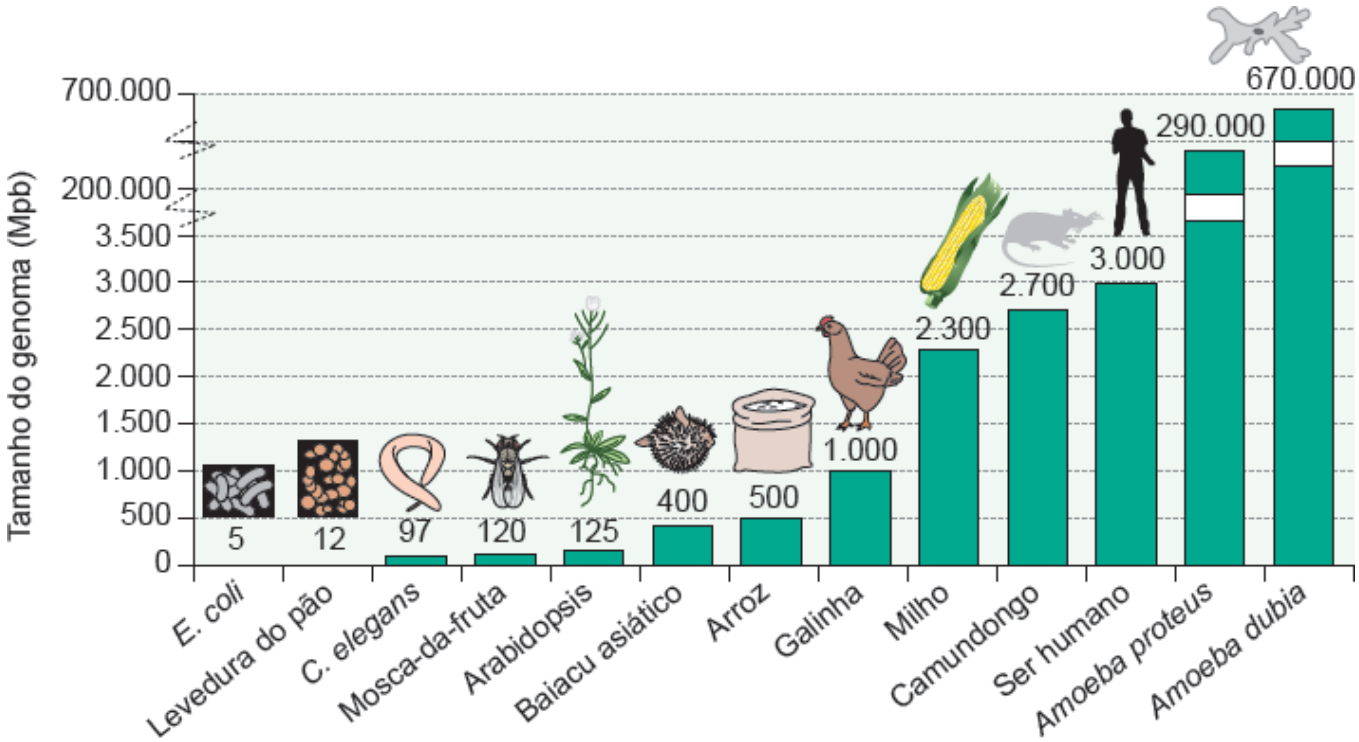


Figura 12.7 Comparação entre o tamanho dos genomas de algumas espécies.

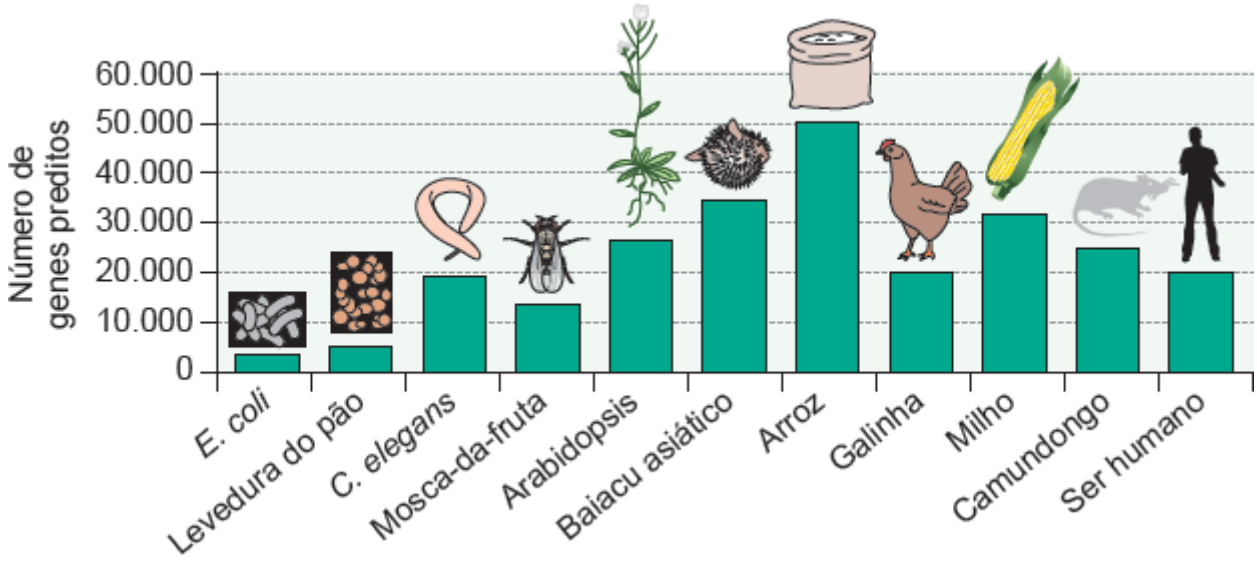


Figura 12.8 Comparação do número de genes em alguns organismos.

Genomas de procariotos | Pequenos e grandes

Ao analisar a distribuição dos tamanhos de genomas tanto em arqueias como em bactérias, nota-se que há uma distribuição contínua. Em arqueias, a maioria dos genomas tem tamanho médio, ao redor de 2 Mpb (de 0,5 Mpb em *Nanoarchaeum equitans* até 5,5 Mpb em *Methanosarcina barkeri*), ao passo que, em bactérias, os tamanhos giram em torno de dois valores: 2 Mpb e 5 Mpb (Figura 12.9). Curiosamente, em procariontes, o tamanho médio dos genes (ORF) é 1 Kpb.

Bactérias parasitas

Outra constatação muito interessante que surgiu por meio da genômica comparativa é que todas as bactérias de genomas muito pequenos (menor que 2 Mpb) são organismos que vivem dentro de outras células (Figura 12.10), seja de forma parasítica ou simbiótica. A explicação é que, vivendo desse modo, a bactéria se vale de muitos metabólitos produzidos pela célula hospedeira e, assim, passam a ser desnecessários seus próprios genes que codificam para esses mesmos metabólitos. Uma vez desnecessários, não estão sob pressão de seleção. Assim, se forem inativados em decorrência de uma mutação, essas formas mutantes não são eliminadas pela seleção e, em um

primeiro momento, o genoma começa a acumular esses genes defeituosos, que se tornam pseudogenes. Em uma etapa mais avançada, os genes podem ser eliminados e, em consequência, o tamanho do genoma fica menor. Esse processo de acúmulo de genes mutantes e posterior "encolhimento" do genoma é denominado "decaimento genômico", que será discutido adiante mais detalhadamente. Se, por um lado, usar os genes da célula hospedeira para suprir suas funções descomplica a situação para a bactéria (dando a ideia de uma bactéria *bon vivant*), por outro lado, a bactéria pode passar a depender muito mais dessa relação, ao ponto de não poder mais viver fora de uma célula, tornando-se um caso extremo de parasita: o de um parasita ou simbionte *obrigatório* intracelular.

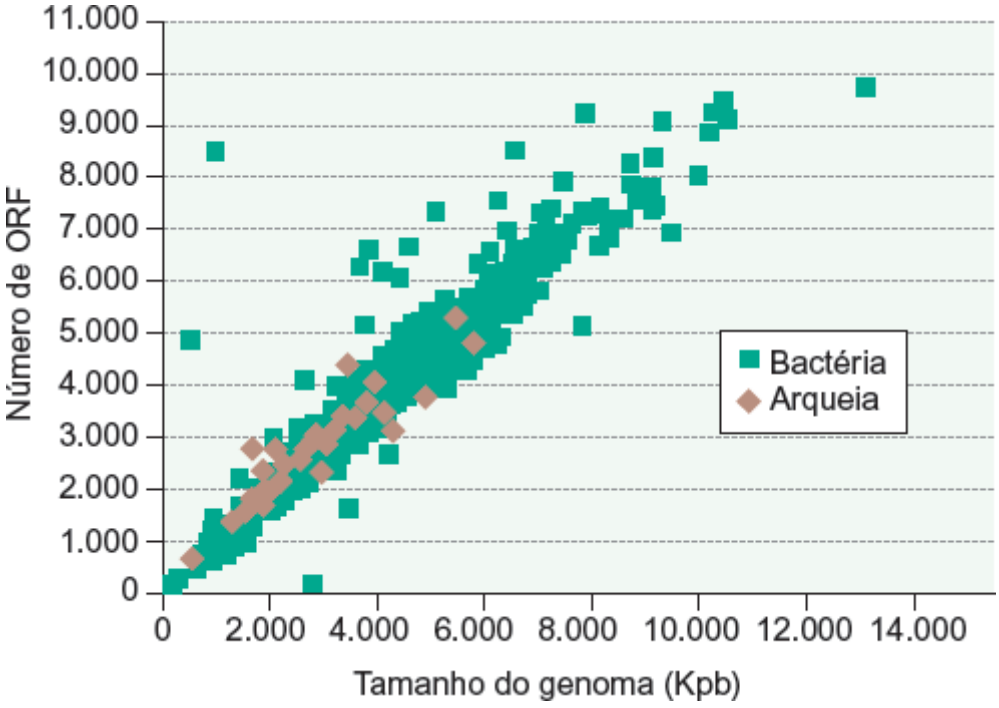


Figura 12.9 Distribuição do tamanho dos genomas de bactérias e arqueias em relação ao número de genes preditos (ORF). Foi considerado somente o genoma sequenciado de um único membro de cada espécie, sendo 795 genomas de bactérias e 73 de arqueias.

Com quantos genes se faz uma bactéria de vida livre?

Se, por um lado, o genoma de bactérias intracelulares é pequeno, o oposto vale para bactérias de vida livre: elas apresentam um genoma maior, com mais genes para que possam responder aos desafios de uma sobrevivência independente, sob diversas condições ambientais (em muitos casos, bem inóspitas). É importante manter em mente, no entanto, que essa regra não se aplica de maneira indiscriminada (como tantas outras em biologia). A bactéria *Mycoplasma genitalium* fica no meio termo, pois contém um genoma muito pequeno (580.070 pb) e é de vida livre (pode também parasitar células epiteliais do trato urinário e respiratório, mas, neste caso, fica fora das células). Pelo fato de ter um genoma muito pequeno, mas ainda assim ser de vida livre, os genes dessa bactéria (apenas 521, dos quais 482 codificam para proteínas) são reconhecidos como o conjunto mínimo necessário para um procarioto ter uma vida livre. *M. genitalium* é membro da classe Mollicutes, que evoluiu a partir de bactérias Gram-positivas com genomas de baixo conteúdo de bases G e C.

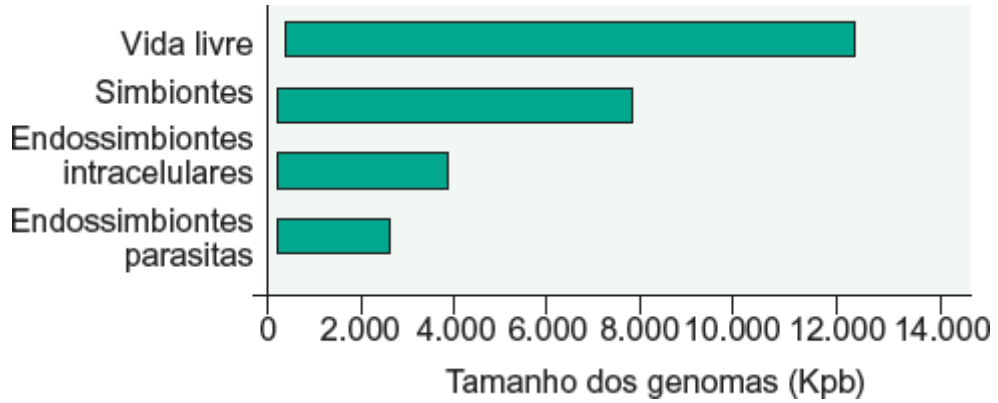


Figura 12.10 Distribuição dos genomas de acordo com as informações do banco de dados *Gold* quanto ao tamanho do genoma e estilo de vida.

Dentre os mais de 30 genomas completamente sequenciados de membros dessa classe, *M. mycoides* merece destaque. O genoma desse organismo, com cerca de 1,2 milhão de pares de bases contendo 985 genes, serviu de modelo para um projeto ambicioso de sintetizar o primeiro genoma de uma bactéria *in vitro* e produzir um organismo em laboratório. *M. mycoides* é o agente causal de uma doença altamente contagiosa em ruminantes: pleuropneumonia bovina contagiosa (CBPP). Um genoma de 1,08 milhão de pares de bases foi sintetizado artificialmente, tendo como base o genoma completamente sequenciado de *M. mycoides* e inserido em células de outra espécie, a *M. capricolum.* A linhagem de *M. mycoides,* JCVI-syn 1.0, apresenta uma versão modificada do genoma original da bactéria. Foram introduzidas marcas na sequência de DNA para que essa nova versão pudesse ser identificada. Dentre elas, está uma sequência de nucleotídeos que, quando traduzida acompanhando o código genético, forma a frase: "Viver, errar, triunfar, criar vida a partir da vida", do romancista e poeta irlandês James Joyce, um dos autores de maior relevância do século 20. Outras marcas formam o nome dos principais autores do trabalho. As novas células criadas apresentam as características fenotípicas esperadas com base no genoma, e são capazes de se duplicar continuamente.

Com esse projeto, os pesquisadores querem demonstrar a possibilidade da genômica sintética, pela qual, em um futuro não muito distante, pedaços de genomas de vários organismos poderiam ser sintetizados em um só, criando novas formas de vida – como, por exemplo, organismos que sejam extremamente eficientes na digestão de resíduos industriais ou, ainda, na digestão de celulose, visando à produção de energia.

Genomas de procariotos | "Compactos" em comparação aos genomas de eucariotos

Embora o tamanho de genomas procarióticos seja variável, a densidade gênica do genoma não varia. Isso significa que o número de genes que codificam para proteínas, dividido pelo número de pares de bases do genoma, é conservado. Para a maioria das bactérias, esse valor (densidade gênica média) gira em torno de 0,9 genes a cada 1.000 pares de bases (1 Kpb), ao passo que, para arqueias, esse número é ligeiramente maior, envolvendo aproximadamente de 1 a 1,1 gene/Kpb. Posto de outro modo, em geral, é possível encontrar um gene a cada 1.000 pares de bases (o que, portanto, corresponde ao tamanho médio de um gene). Se esses números forem divididos pelo tamanho dos respectivos genomas, cerca de 90% do genoma de um procarioto codifica proteínas, e indica também que esses genomas são bem compactos – no sentido de que toda informação ali contida é prioritariamente armazenada para codificar proteínas; os espaços entre genes (os restantes 10%), se existirem, são muito pequenos se comparados com genomas de eucariotos. A diferença é gritante, visto que, em eucariotos, os genes podem ser separados por milhares de bases. Tomando o genoma humano para ilustrar, ao dividir seu tamanho (que é de aproximadamente 3 bilhões de pares de bases) pelo total de 20.000 genes que codificam proteínas (esse número é uma estimativa e pode variar entre uma análise e outra, dependendo das ferramentas de bioinformática usadas para encontrar os genes nas sequências de genomas), conclui-se que apenas 0,8% do genoma codifica para proteínas, sendo o restante composto por sequências regulatórias, introns, DNA repetitivo, DNA espaçador etc. A densidade gênica, portanto, é uma grande diferença entre genomas eucarióticos e procarióticos, e as teorias usadas para explicar as causas disso são variadas. No entanto, todas concordam que existe uma pressão evolutiva que age favorecendo genomas compactos em procariotos, e essa pressão é, muito provavelmente, a seleção.

Sequências de genes procarióticos ORFãs

Após o sequenciamento de um genoma, o passo seguinte é comparar as sequências de prováveis genes com as de genes sequenciados em outros organismos. Ao fazermos isso, estamos conjecturando sobre a identidade de um gene com base em sua similaridade com outro. Isso é feito por meio de programas que usam como base de comparação sequências depositadas em bancos de dados. A predição de genes de procariotos realizada dessa maneira é uma ferramenta muito importante da genômica e funciona relativamente bem em procariotos, uma vez que as sequências que codificam proteínas são consideravelmente conservadas dentre esse grupo. Em termos práticos, isso significa que, atualmente, ao sequenciar pela primeira vez o genoma de uma espécie procariótica e comparar os genes que são encontrados com o de outras bactérias, para cerca de 85 a 90% dos genes, será encontrado um gene similar já sequenciado em algum outro genoma. Por outro lado, conclui-se que cerca de 10 a 15% dos genes são inéditos, no

sentido de que não há nada similar a eles nos bancos de dados. São as sequências ORFãs, um termo que aproveita a sigla em inglês ORF, usada para designar uma possível sequência gênica (de *open reading frame*) – para fazer uma alusão divertida ao fato de esses genes não terem "parentes". Para completar a brincadeira com nomes, foi proposto até um ORFanato, que é um banco de dados somente de sequências ORFãs.

Por motivos óbvios, essa porcentagem era muito maior no início da genômica (cerca de 50 a 40%), quando o número de genomas sequenciados era pequeno, uma vez que os bancos de dados não continham muita informação: há 10 anos, havia uma grande probabilidade de sequenciar um gene inédito. À medida que a disponibilidade de sequências de genoma aumentou, a porcentagem de genes inéditos diminuiu, como era de se esperar.

Retomando a questão das ORFãs, o que são tais sequências? Há algumas explicações, e a mais simples é a de que, na verdade, não representam genes. São erros cometidos pelos programas de computadores que tentam adivinhar onde começam e terminam os genes em uma sequência de nucleotídeos; são, portanto, falsas predições. Outra explicação é que esses genes existem em outros genomas, mas não os detectamos com as ferramentas da bioinformática, pois apresentam sequências muito divergentes em consequência de uma rápida taxa de evolução. Finalmente, existe a hipótese de que certa fração dessas ORFãs veio de seres ainda desconhecidos, sendo incorporados aos genomas de procariotos por meio de mecanismos de transferência horizontal. Os seres desconhecidos doadores dos genes, neste caso, poderiam ser vírus. Tal hipótese leva em conta que os vírus representam a entidade biológica mais numerosa e menos conhecida do planeta Terra. Um estudo de ORFãs em *Escherichia coli*, por exemplo, revelou a existência de ORFãs específicas para cada estirpe dessa espécie. A estirpe MG1655, por exemplo, contém 64 ORFãs que ocorrem somente em seu genoma. Análises mais refinadas mostraram que essas ORFãs codificam proteínas (e, portanto, são genes de verdade e não erros de bioinformática), mas diferem do restante dos genes de *E. coli*, pois suas sequências são mais ricas em adenina e timina que em citosina e guanina. Nesse sentido, portanto, as ORFãs de *E. coli* se parecem mais com genes de vírus do que da própria *E. coli*, visto que genes de vírus também são mais ricos em AT.

Elementos genéticos móveis e evolução de genomas de procariotos

O grande volume de informações produzidas a partir de análises comparativas mostra que tanto a aquisição de DNA exógeno quanto a redução via deleções têm papel importante na evolução de genomas bacterianos. Essa plasticidade genômica demonstrada para várias espécies de bactérias – dentre as quais, *E. coli*, *Helicobacter pylori*, *Mycobacterium leprae* e *tuberculosis* e *Leifsonia xyli* – tem um impacto importante sobre a biologia dos procariotos, especialmente no que refere a sua adaptação ecológica.

Análises comparativas entre genomas de variantes patogênicas e não patogênicas de uma mesma espécie revelam que as diferenças são, muitas vezes, resultantes da presença e/ou ausência de genes que ocorrem de maneira agrupada no genoma e que foram potencialmente adquiridos em eventos de transferência horizontal de genes (*horizontal gene transfer* – HGT). As regiões que agrupam genes específicos desse modo são conhecidas como *ilhas genômicas*. Estas foram primeiramente descritas como regiões contendo fatores de virulência no cromossomo da estirpe uropatogênica da bactéria *E. coli*. Essas regiões podem ser identificadas *in silico* por uma variação no conteúdo de bases GC diferente da média definida para um dado genoma. Ressalta-se que genomas de procariotos variam enormemente quanto à composição de bases GC: alguns grupos taxonômicos chegam a ter menos que 20% do total de bases do tipo AT, como, por exemplo, *Carsonella ruddii* (com 16,5%), e outros, como *Anaeromyxobacter dehalogens*, chegam a conter 75% de bases AT (Figura 12.11). Variações no conteúdo de bases podem ser detectadas utilizando ferramentas computacionais; assim, essas ilhas genômicas podem ser detectadas em um dado genoma quando o conteúdo de bases de uma determinada região é muito discrepante da média definida para todo o genoma.

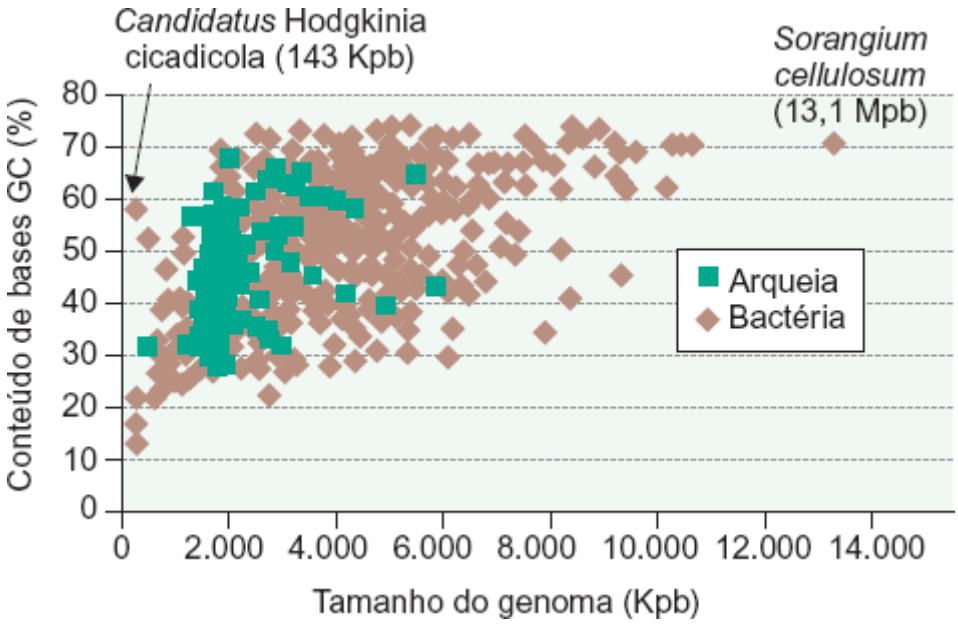


Figura 12.11 Distribuição do tamanho do genoma e conteúdo de bases GC em arqueias e bactérias. Foi considerado somente o genoma sequenciado de um único membro de cada espécie, sendo 795 genomas de bactérias e 73 de arqueias.

Muitas vezes, essas ilhas estão integradas dentro ou próximas a um tRNA, podendo conter elementos que favorecem a sua inserção (sequências de inserção, transposons, integrases, recombinases) e sequências repetidas e invertidas nas extremidades, oriundas do evento de inserção. Podem variar em tamanhos de 10 e 500 Kpb e, quando contêm genes associados à virulência, são denominadas *ilhas de patogenicidade*, sendo encontradas em uma grande variedade de bactérias patogênicas de plantas e animais. As ilhas genômicas podem conter outros genes que não estão associados à virulência. Em *Shigella flexinery,* por exemplo, genes de resistência a antibióticos estão agrupados em regiões de ilhas genômicas e em *Mesorhizobium loti*, uma bactéria fixadora de nitrogênio que vive em associação simbiótica a raízes de plantas leguminosas. Uma ilha genômica de simbiose contém os genes associados ao processo de fixação de nitrogênio.

Dentre os elementos genéticos móveis envolvidos em eventos de HGT estão os bacteriófagos, os plasmídeos, os transposons e as sequências de inserção (elementos IS – *insertion sequences*). Bacteriófagos são vírus que infectam células de bactérias, podendo inserir o seu material genético no cromossomo bacteriano. Um exemplo importante de HGT por essa via é a aquisição de genes que codificam toxinas associadas à patogenicidade de *Vibrio cholerae* e de *Pseudomonas aeruginosa*.

As sequências de inserção (IS) e transposons são pequenos segmentos de DNA transponíveis, podendo estar presentes em muitas cópias no genoma de bactérias. Esse grande número de cópias favorece eventos de rearranjos, atuando como sítios para recombinação homóloga. Por exemplo, em *Leptospira interrogans*, as diferenças encontradas entre os genomas de dois *serovares* (*lai* e *conpenhagni*) são decorrentes principalmente da variação encontrada nesses elementos, onde a localização de elementos IS coincide com as regiões do genoma que sofreram grandes rearranjos.

As regiões adquiridas com o auxílio de elementos genéticos móveis podem manter a sua capacidade de transferência. No entanto, muitas vezes, os genes associados à integração são perdidos e essas regiões tornam-se permanentemente integradas ao genoma. As características das ilhas genômicas são também determinadas pelo mecanismo associado à transferência entre organismos que variam entre os clássicos mecanismos de recombinação em bactérias: conjugação, transdução e transformação. A transferência de material genético, assim como em outros processos associados à evolução, cria uma variabilidade que, sob os efeitos da seleção, formata o conteúdo de um dado genoma em resposta às condições ambientais.

Perdas de segmentos do DNA | Decaimento genômico

Se, por um lado, procariotos podem ganhar genes por meio de eventos de duplicação ou por HGT, por outro lado também podem perder genes pela mutação (Figura 12.12). Há casos em que o genoma acumula genes mutantes inativos (pseudogenes) que, com o passar do tempo, são eliminados, resultando na redução do tamanho do genoma.

Tal processo de redução é denominado “decaimento genômico”. Exemplos desse fenômeno tornaram-se conhecidos por meio da genômica comparativa de espécies próximas, que revelou sua importância para a evolução de genomas de procariotos e também para a adaptação dos organismos a novos ambientes e estilos de vida, principalmente no que se refere à especialização de *habitat* dependente de hospedeiro (como é o caso de microrganismos parasitas ou simbiontes intracelulares).

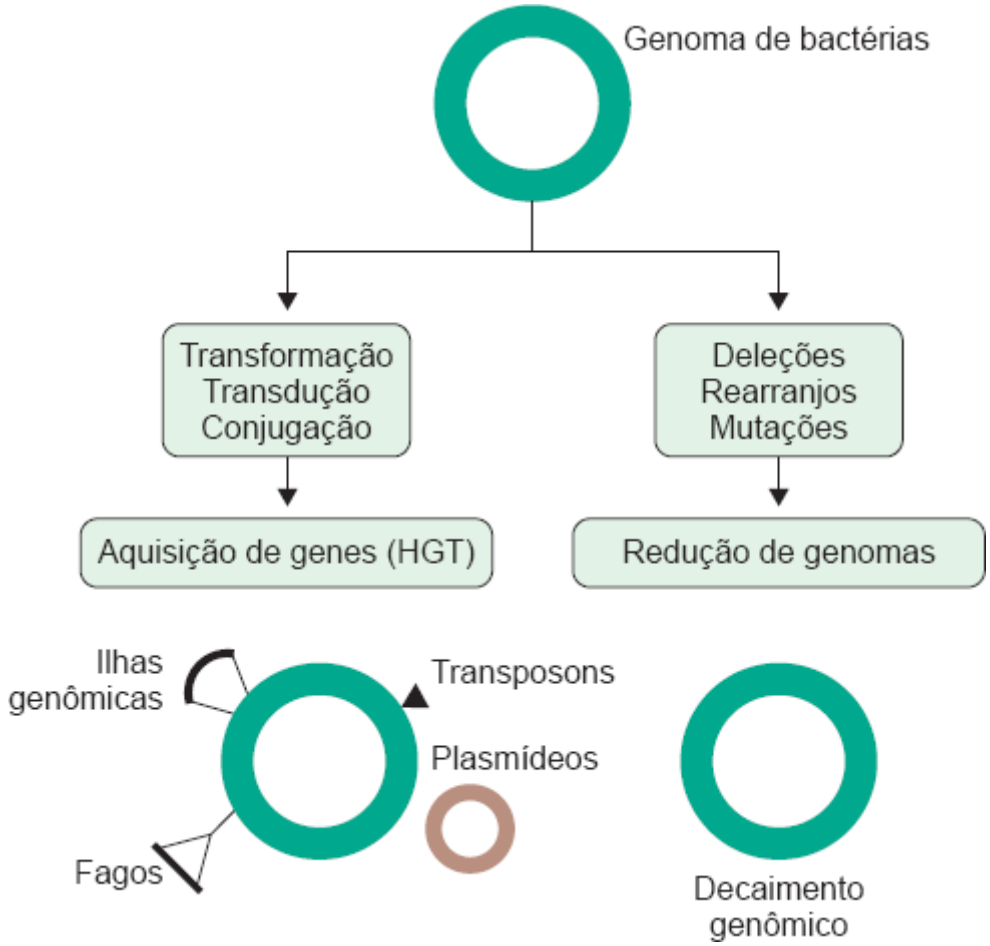


Figura 12.12 Representação esquemática da movimentação genômica em bactérias.

A presença de um grande número de pseudogenes em um determinado genoma é sintoma de decaimento. Durante o processo evolutivo, pseudogenes vão acumulando mutações até que não sejam mais reconhecidos quando comparados com a versão original não mutante do gene. Para avaliar o decaimento genômico, duas metodologias são preferencialmente utilizadas: uma é a simples contagem do número de pseudogenes e a outra é a análise da ausência de genes em uma determinada espécie comparada com outra geneticamente próxima. A perda de genes pode ocorrer por meio da deleção de grandes segmentos de DNA que carregam vários genes, ou alterações por pequenas mutações em genes envolvidos em uma via comum, que pode acontecer quando a biossíntese de um determinado metabólito não for mais necessária, relaxando a pressão seletiva para a manutenção daquela função. Quando um gene perde a função, aqueles que dependem da sua atuação também são perdidos durante o processo evolutivo, tal qual um “efeito dominó”.

A redução de genomas foi evidenciada primeiramente nos genomas das organelas mitocôndrias e cloroplastos, nas quais muitos dos genes presentes foram transferidos para o núcleo da célula hospedeira. Os genomas de mitocôndrias e de cloroplastos podem ser tão pequenos quanto 6 Kpb e 35 Kpb, respectivamente. A redução no tamanho do genoma está ainda bem documentada em parasitas e simbiontes intracelulares.

O exemplo mais extremo de decaimento genômico é o caso do patógeno intracelular obrigatório *Mycobacterium leprae*, agente causal da hanseníase, no qual aproximadamente 50% do genoma compreende pseudogenes. Comparando os genomas de *M. leprae* e da espécie próxima *M. tuberculosis*, a primeira apresenta menos da metade dos genes presentes em *M. tuberculosis* (3.959 genes). O genoma de *M. leprae* compreende 1.116 pseudogenes e 1.604 genes funcionais, sendo 1.439 comuns a *M. tuberculosis*. A redução do genoma eliminou atividades metabólicas importantes, como vários sistemas catabólicos e grande parte do metabolismo energético.

Genoma essencial *versus* genoma assessório (genoma pan)

Como discutido até aqui, a movimentação no genoma de procariotos é bastante intensa; ela inclui transferência horizontal de genes e processos de decaimento genômico. Quando analisamos o genoma de indivíduos diferentes de uma determinada espécie em procariotos, encontramos uma variabilidade que pode ser discreta, como no caso de *Bacillus anthracis*, uma vez que todos os genomas analisados de diferentes linhagens mostram um número baixo de genes específicos a cada uma; ou podemos encontrar uma variação que pode chegar a 30% do genoma completo dos organismos. *E. coli* é um exemplo desta vasta diversidade intraespecífica. Entre linhagens patogênicas e não patogênicas de *E. coli*, a linhagem K12-MG1655 não patogênica apresenta 5.288 genes preditos; a uropatogênica CFT073, 5.605; e as entero-hemorrágicas O157:H7 EDL933 e O157:H7 VT2-Sakai, 5.680 e 6.423, respectivamente. Com base nessas descobertas, foram propostas as seguintes definições:

- Genoma essencial (ou *core*): é o conjunto de genes compartilhados por todas as linhagens de uma mesma espécie bacteriana
- Genoma acessório (ou dispensável): é o conjunto de genes que estão presentes em algumas, mas não em todas, as linhagens de uma mesma espécie bacteriana
- Genoma pan (*pan*, do grego, *todo*): é o repertório global de genes de uma espécie bacteriana; ou seja, o genoma essencial mais todos os acessórios encontrados em diferentes linhagens de uma mesma espécie.

Com toda essa variação entre os genomas, tanto entre como dentro do que se conhece como uma espécie, é possível imaginar: como, afinal, se define uma espécie em procariotos? Sabe-se que o conceito de espécie para eucariotos envolve grupos filogenéticos bem-definidos com isolamento reprodutivo (ou seja, genético) – são conceitos difíceis de serem aplicados em se tratando de procariotos, uma vez que, além de toda a movimentação genômica, há ainda reprodução assexuada. As linhas de pensamento mais atuais para a definição de um conceito de espécie em procarióticos incluem, além da abordagem genética e fenotípica clássicas, as informações sobre as sequências de ribossômicos, como aquelas discutidas no início deste capítulo, e informações genômicas (genoma *core* e acessório, HGT), como apresentadas até aqui, em uma abordagem conhecida como polifásica. Muitas discussões ainda estão por vir, a fim de entender melhor esse mundo procariótico que vem se beneficiando enormemente das informações advindas da genômica.

Sequenciamento de organismos-modelo entre os eucariotos

Uma vez que é possível observar na grande diversidade de seres vivos similaridades entre seus mecanismos biológicos, alguns organismos podem ser selecionados para a realização dos estudos em laboratório, na expectativa de que os resultados obtidos possam ser multiplicados para os outros organismos. Essas espécies selecionadas para estudo são chamadas de organismos-modelo (ver detalhes no Capítulo 8).

A posição filogenética, o conhecimento genético e a facilidade de manuseio em condições de laboratório são alguns dos critérios para a escolha de um organismo como modelo. As informações derivadas de estudos nessas espécies são essenciais para a interpretação de dados de organismos não tão tratáveis quanto à manipulação, como é o caso de estudos com humanos. É fácil de entender quando pensamos em estudos sobre a regulação da expressão gênica, aspectos evolutivos, estudos de desenvolvimento ou até mesmo de doenças genéticas. Não é possível realizar experimentos em humanos. Além disso, informações obtidas a partir de organismos-modelo são fundamentais para o desenvolvimento de novas tecnologias e metodologias de estudo. De maneira geral, todo o procedimento biológico em organismos-modelo é mais simplificado.

Essa lógica se aplica também a projetos de sequenciamento de genomas completos. O primeiro genoma de um eucarioto completamente sequenciado foi o da levedura do pão *Sacharomyces cerevisiae*, em 1996, cujo título do artigo publicado é “Vida com 6.000 genes”. Desde então, para cada grupo taxonômico foram escolhidas aquelas espécies com a maior quantidade de informação biológica disponível, para que estudos funcionais pudessem ser desenvolvidos com base nos resultados do sequenciamento completo dos genomas. Observe que voltamos ao início do capítulo – informações obtidas em projetos devem ser *a posteriori* estudadas experimentalmente.

Dentre os escolhidos inicialmente, estão: *Caenorhabditis elegans*, um nematoide utilizado com frequência em estudos de desenvolvimento embrionário, porque é um organismo transparente, de ciclo de vida curto (3,5 dias) e com um número constante de células somáticas (959 nos hermafroditas e 1.031 nos machos), o que torna possível o acompanhamento do destino de cada um em apenas 16 h em uma placa de Petri; *Arabidopsis thaliana*, uma das

plantas mais estudadas em pesquisas genéticas, bioquímicas, e fisiológicas, apesar de não apresentar importância econômica direta; *Drosophila melanogaster*, a mosca-da-fruta, que possibilitou o desenvolvimento da genética moderna a partir dos experimentos de Thomas H. Morgan, principalmente no que se refere ao mapeamento e identificação de aberrações cromossômicas; camundongo e rato (*Mus musculus* e *Rattus norvergicus*), modelos animais utilizados como base para a pesquisa de doenças que acometem o ser humano, assim como para o desenvolvimento humano. Estes foram os primeiros eucariotos selecionados para projetos de sequenciamento. No entanto, atualmente, como evidenciado na Tabela 12.1, existe uma variedade bastante grande de organismos com genomas completamente sequenciados – com as mais diferentes justificativas para tal. Vale ressaltar que também foi possível o sequenciamento do DNA isolado de fósseis do homem de Neandertal (ver boxe "Genoma dos homens de Neandertal").

Os princípios discutidos no estudo dos genomas de procariotos e apresentados anteriormente se aplicam igualmente ao estudo dos genomas de eucariotos – principalmente no que se refere à genômica comparativa, sendo esta a arma mais poderosa para o entendimento dos diversos eventos celulares e moleculares. Os desafios consistem em localizar e determinar o número de genes em um dado genoma e a função de cada um deles. Neste caso, vale também comparar as sequências.

Em procariotos, o tamanho do genoma e o número de genes estão relacionados; assim, genomas maiores acompanham um maior número de genes. Ao analisar a Tabela 12.3, percebemos que a mesma regra não pode ser aplicada ao genoma de eucariotos. Há genomas com o mesmo tamanho abrigando um número muito diferente de genes. O genoma humano, por exemplo, com cerca de 20 mil genes, enquanto para o genoma do milho estima-se 32 mil genes, e cada um tem aproximadamente 3 bilhões e 2,3 bilhões de pares de bases, respectivamente.

Genoma dos homens de Neandertal

Quando os humanos anatomicamente modernos surgiram na África, outro grupo de hominídeos já vivia na Europa Ocidental e na Ásia Ocidental: os homens de Neandertal.

Um dos primeiros esqueletos de homens de Neandertal foi encontrado no vale de Neander, na Alemanha, daí o nome da espécie. De acordo com o registro fóssil, os homens de Neandertal surgiram há 230.000 anos e, provavelmente, coexistiram com os humanos anatomicamente modernos, há cerca de 80.000 anos. Em pouco tempo de coexistência (algumas dezenas de milhares de anos), no entanto, esses homínideos desapareceram, e os últimos registros fósseis datam de 30.000 anos.

Há muito interesse na comparação entre as sequências do genoma dos homens de Neandertal com a dos humanos modernos e do chimpanzé. Isso porque essa análise tornaria possível identificar as regiões no genoma que conduziram a separação desses grupos anatomicamente distintos e identificar genes que sofreram seleção positiva desde que os humanos modernos se diferenciaram dos outros grupos de hominídeos e dos primatas. Somadas ao registro fóssil, essas informações podem ajudar a revelar a origem e a evolução de nossa espécie.

O DNA ancestral é obtido a partir de amostras de ossos de indivíduos encontrados nas escavações. Cerca de 50 a 100 mg de pó de fósseis ósseos são suficientes para gerar a informação do genoma. Um dos maiores desafios, no entanto, é obter DNA de amostras fósseis livres de contaminação com DNA de humanos modernos. Apesar dos avanços na metagenômica e no sequenciamento em larga escala, as técnicas ainda não evitam a amplificação de DNA contaminante nas amostras fósseis, visto que elas foram manipuladas por humanos durante suas descobertas. Mesmo após inúmeras modificações nos protocolos para a obtenção dessas amostras, as análises comparativas mais recentes indicam que 1% das sequências de DNA geradas de amostras de fósseis de Neandertal apresenta contaminação.

Existe alguma contribuição do genoma dos homens de Neandertal ao genoma dos humanos anatomicamente modernos?

Mesmo com todas as dificuldades, as primeiras sequências de DNA de hominídeos ancestrais foram obtidas em 1997, seguidas da sequência completa de DNA mitocondrial. Em 2010, Richard Green *et al*. (2010) publicaram uma versão do genoma nuclear dos homens de Neandertal. Nessa primeira análise, o genoma mostrou-se geneticamente idêntico ao humano moderno em 99,7%; e ao chimpanzé, em 98,8%. Uma das grandes perguntas é a existência de fluxo gênico entre os homens de Neandertal e os humanos modernos, uma vez que os registros fósseis indicam que eles coexistiram há 80.000 anos. As análises comparativas entre esses genomas confirmaram essa expectativa: aproximadamente 1 a 4% do genoma dos homens de Neandertal foi encontrado no genoma dos humanos modernos, mas apenas nos de origem europeia e asiática.

Então não houve fluxo gênico entre hominídeos e os homens modernos da Oceania?

Um grupo de hominídeos distinto foi encontrado no sul da Sibéria, nas cavernas de Denisova: os homens de Denisova. Os registros fósseis indicaram que homens de Denisova apresentam uma origem comum com os homens de Neandertal, mas os registros indicam ser um grupo de hominídeos distinto. Amostras obtidas de falanges desses hominídeos foram sequenciadas, também em 2010, por David Reich *et al*., e os dados revelaram que

os homens de Denisova contribuíram com aproximadamente 4 a 6% do genoma dos humanos modernos da Oceania (Austrália, Filipinas e Melanásia), mas não foi encontrada contribuição ao genoma dos homens modernos da Eurásia.

Portanto, na época em que os humanos modernos surgiram na África, o mundo já era colonizado por pelo menos dois outros grupos de hominídeos: os homens de Neandertal e os homens de Denisova. Ainda: apesar de baixo, é possível detectar a ocorrência de fluxo gênico entre esses hominídeos e os humanos modernos. É a genômica auxiliando o registro fóssil a revelar a evolução da nossa espécie.

Então existem genes que evoluíram de maneira diferente?

A comparação entre os genomas hominídeos dos humanos modernos e do chimpanzé vem agora indicando regiões que passaram por uma evolução recente. Tais regiões contêm genes: foram encontrados genes associados ao metabolismo energético, ao desenvolvimento cognitivo, à estrutura do crânio e à morfologia da pele.

Tabela 12.3 Número de genes, tamanho do genoma em Mpb e número de cromossomos de diferentes espécies que apresentam o genoma completamente sequenciado. São apresentados valores aproximados.

Espécie	Cromossomos (n)	Tamanho (Mpb)	Genes
Humano	23	3.000	20.000
Chimpanzé	24	3.000	20.000 a 23.000
Cão	39	2.300	19.300
Galinha	39	1.000	20.000
Camundongo	20	2.700	20.000 a 30.000
Baiacu asiático	22	400	30.000 a 40.000
Mosca-da-fruta	4	120	14.000
Arroz	12	500	46.000 a 55.600
Arabidopsis	5	125	26.500
Milho	10	2.300	32.000
Caenohabitis elegans	6	97	19.700
Levedura do pão	1	12	6.000
Escherichia coli	1	4,6	4.300

A predição de genes em procariotos é mais simples que aquela em eucariotos. Como já discutido, genes de procariotos são pequenos, sem introns e distribuídos ao longo de todo o genoma. Já em eucariotos, os genes podem estar separados por milhares de pares de bases e, muitas vezes, apresentam vários introns. Assim, estimar o número de genes em genomas de eucariotos não é tarefa fácil. No entanto, nos últimos anos, o desenvolvimento dos muitos projetos de genoma, particularmente aqueles realizados em organismos-modelo, tem possibilitado estimativas realistas sobre o número de genes nos mais diferentes genomas.

De maneira geral, em eucariotos, temos que genes que codificam proteínas ocupam menos que 10% do genoma completo, e regiões não codificadoras, mais que 90%. Dentre as várias classes de regiões não codificadoras, as mais fáceis para a predição computacional são os introns e os elementos genéticos móveis. De maneira geral, os menores genomas de eucariotos apresentam também poucas sequências intrônicas, como é o caso do genoma da levedura *S. cerevisiae*. Por sua vez, genomas maiores apresentam coerentemente mais sequências intrônicas. Da mesma maneira, as sequências intergênicas ocupam mais espaço em genomas maiores que em genomas menores. Em genomas muito maiores existe um aumento significativo da presença de elementos genéticos móveis, principalmente transposons e retrotransposons. No entanto, existem variações importantes quando comparamos genomas de animais e de plantas.

Sequenciamento do genoma humano | O que nos faz humanos?

Uma primeira versão da sequência do genoma humano foi disponibilizada no início de 2001. O complemento haploide do genoma humano tem aproximadamente 3 bilhões de pares de bases. A obtenção da sequência desse genoma foi um desafio para a ciência, porque é um genoma muito maior em relação aos demais que vinham sendo sequenciados na época, além de apresentar inúmeras sequências repetitivas, difíceis de serem obtidas e montadas por programas de bioinformática. A primeira versão do genoma obteve uma cobertura de 90% e apresentava muitas lacunas não sequenciadas (*gaps*) e erros de base. Por isso, em 2004, o *International Human Genome Consortium* lançou uma nova versão do genoma, apresentando sequências de melhor qualidade, cobrindo uma maior parte do genoma (em torno de 99,7%) e com um número menor de *gaps*.

O número de genes do genoma humano ainda não é definitivo (discussão a seguir), mas fica em torno de 22 mil. Como isso seria possível? O milho apresenta 32 mil genes. Nossa complexidade biológica reside, portanto, não no número de genes, mas em nosso potencial de, alternativamente, transcrever e traduzir a informação codante do genoma. Cerca de 90% dos genes do genoma humano são capazes de sofrer processamento alternativo, e o resultado é um número maior de proteínas em relação ao número de genes.

Estimando o número de genes do genoma humano

As primeiras estimativas sobre o número de genes do genoma humano no início do projeto de sequenciamento, na década de 1990, eram de aproximadamente 100.000 genes. Desde então, esse número vem diminuindo. Na primeira publicação sobre a sequência completa, foram previstos entre 30 e 40 mil genes. Mais recentemente, agrupando todas as informações disponíveis: sequenciamento das EST (pequenos fragmentos de cDNA sequenciados, chamados de marcadores de sequências expressas – *expressed sequence tags*); alinhamento das EST com o DNA genômico; descoberta de genes por homologia com outros organismos e melhoras nos programas computacionais de predições – as estimativas estão entre 20.000 e 25.000 genes. Os modelos gênicos propostos nos bancos de dados Ensembl (EMBL) e NCBI listam 22.619 e 22.333, respectivamente, de genes codificadores de proteínas. Esses valores, obviamente, ainda não são definitivos e, com o refinamento dos métodos computacionais e a busca contínua de informações, tanto sobre os humanos quanto sobre os demais organismos, a qualidade das predições tende a melhorar.

Considerando a densidade de genes no genoma humano, chegamos à conclusão de que apenas em torno de 1% do nosso genoma codifica para proteínas. A média do tamanho de um gene em humano é de 27 mil pares de bases, e as proteínas apresentam em média cerca de 430 aminoácidos. Aproximadamente 1.300 nucleotídeos seriam, portanto, suficientes para codificar cada proteína. Os nucleotídeos restantes são regiões longas não codificadoras (introns), que interrompem sequências codificadoras relativamente pequenas (exons). No entanto, novamente, estamos falando em média, pois existem sequências gênicas tão longas quanto 293 mil pares de bases, sendo que 363 exons codificam a proteína muscular titina, com 38.138 resíduos. As regiões regulatórias responsáveis pela expressão dos genes, no caso do genoma humano, chegam a dezenas de milhares de pares de bases. Exons e sequências regulatórias compreendem menos que 2% do genoma humano. Vale lembrar da diferença quando comparado com sequências gênicas de procariotos, como discutido anteriormente: cerca de 90% desses genomas apresentam regiões codantes. Várias iniciativas acontecem em laboratórios em todo o mundo, com o objetivo de definir as funções de cada uma das proteínas codificadas pelos genes humanos. Uma delas, conhecida como CALIPHO, pretende utilizar uma combinação de ferramentas computacionais e experimentos de laboratórios para estudar sistematicamente essas funções.

Genoma humano *versus* genoma do chimpanzé

O sequenciamento do genoma do chimpanzé foi bastante aguardado, visto que, entre os animais, é o evolutivamente mais próximo dos humanos. A análise comparativa entre as sequências comuns do chimpanzé e do humano revelou 99% de identidade. Incluindo as regiões de deleções e inserções (ou seja, específicas a cada organismo), a porcentagem de identidade é ainda 96%. Em nível proteico, 29% dos genes codificam exatamente a mesma sequência de aminoácidos. O número de diferenças genéticas entre humanos e chimpanzés é aproximadamente 60 vezes menor que as diferenças entre humanos e camundongos, e cerca de 10 vezes menor que as diferenças entre camundongo e rato. As diferenças genéticas entre humano e chimpanzé são cerca de 10 vezes maiores que entre

dois humanos quaisquer. Para referência, dois indivíduos humanos diferem um do outro por uma média de um par de nucleotídeos em cada 1.000, o que possibilita a identificação de indivíduos pela análise do DNA.

Alguns dos resultados interessantes advindos de análises comparativas mais abrangentes estão associados à evolução de algumas classes de genes. Genes associados à percepção de sinais, transmissão de sinais nervosos, produção de espermatozoides e transporte celular de íons são mais variáveis em humanos e chimpanzés que em outros mamíferos já analisados.

Apesar da grande similaridade, existem diferenças importantes entre os genomas de chimpanzé e humano. Cerca de 35 milhões de pares de bases diferem em regiões comuns, 5 milhões de inserções ou deleções estão espalhadas pelos genomas, além de alguns eventos de rearranjos cromossômicos. Considerando a presença de genes específicos, 18 genes podem potencialmente ter surgido nos humanos desde a divergência dos chimpanzés. Os resultados ainda não respondem à pergunta "o que nos faz humanos?"; no entanto, restringem a busca das diferenças biológicas entre as duas espécies.

Genômica de plantas

Comparativamente aos números estudados para os genomas de humanos e outros mamíferos, o genoma de plantas parece ter um número de genes muito maior – várias das estimativas em diferentes plantas chegam a 40 mil genes. Alguns dos argumentos para o maior número de genes em plantas falam sobre o seu estilo de vida. Plantas são sésseis e precisam responder às variações ambientais, desde ataques de patógenos até condições climáticas e competição com outras plantas, produzindo milhares de componentes químicos. O processamento alternativo não parece ser a resposta para aumentar a diversidade de proteínas produzidas. Para os diferentes genomas, as estimativas giram em torno de 20% de processamento alternativo.

Uma estratégia que parece comum a todas as plantas é o grande número de famílias gênicas, ou seja, genes que se originam a partir de duplicações e que são mantidos ao longo do processo evolutivo. Plantas apresentam um número maior de famílias gênicas quando comparadas com os animais. Vale mencionar que ao longo do processo evolutivo as plantas sofreram pelo menos um evento de duplicação genômica; ou seja, passaram por processos de poliploidização. Poliploidias, ou seja, a duplicação de todo um complemento cromossômico, são comuns durante o processo de evolução de plantas.

O primeiro genoma de planta completamente sequenciado foi da angiosperma *A. thaliana*. Atualmente, estima-se em torno de 26.500 genes codificadores de proteínas e que estão documentados no banco de dados TAIR. Dentre os genomas completamente sequenciados de plantas, estão ainda o da batata, arroz, soja, videira, sorgo, populus, maçã e milho.

Paradoxo valor de C *versus* valor de G

O valor de C refere-se à quantidade de DNA do complemento haploide das células de um dado organismo. A quantidade de DNA pode ser obtida em milhões de pares de bases (Mpb). Logo, o valor de C do genoma humano é de aproximadamente 3.000 Mpb ou 3 bilhões de pares de bases (3 Gpb). O valor de G refere-se ao número de genes codificadores de proteínas do genoma haploide (Hahn e Wray, 2002).

Em procariotos, vemos que organismos de vida livre apresentam em média o tamanho do genoma em pares de bases maior que o genoma daqueles que são parasitas intracelulares. Da mesma maneira, genomas maiores apresentam maior número de genes. Analisando mais uma vez a Tabela 12.3 e considerando a complexidade biológica dos organismos, observa-se que não existe correlação entre o tamanho do genoma e o número de genes. O verme *C. elegans* tem um genoma muito menor com aproximadamente o mesmo número de genes do genoma humano, e um nível de complexidade incomparável. Sem mencionar o genoma da Ameba, com 670.000 Mpb, ou o genoma do milho, com 32 mil genes. A essa falta de associação entre complexidade biológica e tamanho do genoma chamamos paradoxo do valor de C; do mesmo modo, a falta de associação entre número de genes e a complexidade chamamos de paradoxo do valor de G.

Intuitivamente, acreditava-se que quanto mais biologicamente complexo um organismo, maior o seu genoma ou maior o número de genes presentes nele. Até certo ponto, isso é verdade: procariotos apresentam um genoma menor que de eucariotos unicelulares que, por sua vez, apresentam um genoma e um número de genes menores quando

comparados com eucariotos multicelulares. No entanto, essa regra não pode ser aplicada para eucariotos multicelulares.

A genômica revelou informações surpreendentes sobre essa percepção. Para o genoma humano, os eventos de processamento alternativo, que podem produzir duas ou mais proteínas a partir de um único RNA mensageiro, acontecem em pelo menos 90% dos nossos genes. Assim, o potencial de codificar proteínas diferentes aumenta independentemente do número total de genes.

Além disso, não passou despercebido que existe um acúmulo de sequências não codificadoras de proteínas nos genomas ao longo do processo evolutivo das espécies. Existe uma forte correlação entre o aumento na quantidade de sequências não codificadoras e a complexidade biológica (Figura 12.13). Ainda há muito para decifrar com relação aos diferentes níveis de complexidade biológica, número de proteínas necessárias para a vida e, mais recentemente, sobre o papel das regiões não codificadoras. O dogma central da Biologia Molecular estabelece que o fluxo de informação genética acontece a partir da transcrição dos genes no DNA em uma molécula de mRNA e da tradução desses mRNA, com o auxílio dos tRNA e rRNA, em proteínas que desempenham suas funções nas células. As regiões não codificadoras eram consideradas "DNA lixo", visto que não eram responsáveis por codificar proteínas.

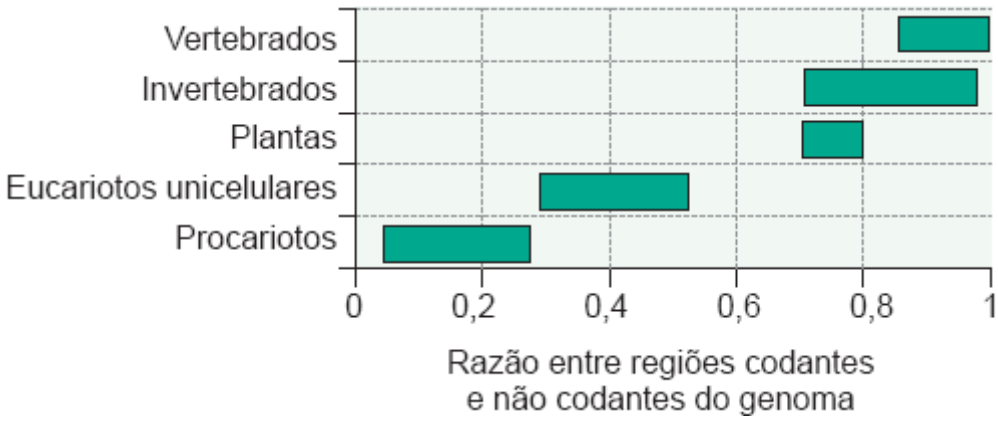


Figura 12.13 Razão entre o número de sequências de DNA não codificador pelo tamanho do genoma. Adaptada de Nunes (2011), com base nos dados de Taft e Mattick (2013).

Genômica e suas aplicações

Desenvolvimento de novas vacinas e fármacos

Os princípios da vacinação como meio de controle de agentes patogênicos foram cientificamente lançados a partir dos trabalhos pioneiros de Edward Jenner, no final do século 18, e Louis Pasteur, desenvolvido entre 1880 e 1890. Essa primeira geração de vacinas, ainda em uso, se apresenta como suspensões inativas ou ativas, mas atenuadas, do agente patogênico. Quando injetadas em um organismo, as vacinas estimulam a produção de anticorpos, imunizando-o. Em geral, os agentes imunizantes (antígenos) são proteínas presentes nas superfícies dos patógenos (p. ex., proteínas constituintes das capas virais), mas há também vacinas desenvolvidas a partir de toxinas parcialmente purificadas. Em qualquer caso, a obtenção dessas vacinas tradicionais necessita que o patógeno em questão seja cultivado em laboratório, o que exclui sua aplicação em organismos não cultiváveis. Em outros casos, ainda, os microrganismos podem ser cultivados, mas, em meio de cultura, não produzem os antígenos. As vacinas vivas atenuadas (p. ex., vacina contra a febre amarela) apresentam o adicional inconveniente da possibilidade de reverterem à forma original não atenuada, o que obviamente constitui grande problema. Assim, o desenvolvimento de vacinas por esses métodos pode levar muitos anos.

Vacinologia reversa | Vacinas desenvolvidas a partir de análises *in silico*

A disponibilização de sequências genômicas modificou radicalmente esse cenário, pois elas representam um catálogo virtual de todos os antígenos possíveis de um organismo candidatos a vacinas. Análises bioinformáticas podem auxiliar na escolha desses candidatos que são, posteriormente, analisados em laboratório, em ensaios biológicos de rotina. Ou seja, a escolha das vacinas passou a ser menos empírica e mais racional. O termo

"vacinologia reversa" é eventualmente empregado para ilustrar essa situação; ou seja, partimos de uma análise *in silico* do genoma do patógeno, e não mais *in vivo* para desenvolver uma vacina.

O primeiro teste a que um produto gênico deve ser submetido para ser considerado um candidato à vacina diz respeito à sua localização celular, devendo ser uma proteína secretada pelo patógeno ou, ainda, estar em sua superfície externa para facilitar o acesso dos anticorpos. Existem vários *softwares* que tornam possível prever a localização celular de determinado produto gênico, que examinam a presença de domínios típicos de proteínas extracelulares ou de membranas. Proteínas que ficam ancoradas na membrana celular, por exemplo, contêm resíduos aromáticos em seu término C, podendo ainda apresentar repetições do tipo Leu(Ala/Val)-Leu-Ala(Ser)-Gly(Ala)-Cys em seu término N. Outro sinal esperado em uma proteína secretada é o peptídeo sinal, que indica o destino final da proteína sintetizada.

Após a escolha de candidatos vacinais com localização celular apropriada, outras informações podem ser acrescidas para ajudar a decidir se esse candidato irá ou não para a etapa seguinte de testes biológicos. Para desenvolver uma vacina que seja eficiente contra sorotipos distintos de um mesmo patógeno, por exemplo, uma estratégia é comparar os genomas desses sorotipos e identificar genes que sejam comuns a todos. Tais genes são os candidatos nos quais se devem concentrar as atividades futuras de pesquisa, pois têm a chance de produzir uma vacina de amplo espectro.

Um exemplo de sucesso de vacinologia reversa foi a procura de candidatos à vacina contra o sorogrupo B de *Neisseria menigitidis*, bactéria que causa meningite e septicemia em crianças e jovens. Análises de bioinformática identificaram 600 genes que codificam para potenciais antígenos secretados ou de superfície. Cada um desses genes foi clonado e expresso em *E. coli*. Nem todos foram expressos (este é um problema dessa abordagem, pois muitos genes podem ser letais para o organismo hospedeiro), mas aqueles que lograram êxito (350 ao todo) tiveram seus produtos purificados e estes foram inoculados em camundongos, a fim de avaliar seus potenciais imunogênicos. Vinte e oito estimularam a produção de anticorpos bactericidas em camundongos. Análises complementares revelaram que muitos desses candidatos estão presentes não apenas em *N. meningitidis*, mas também em outras espécies de *Neisseria*, como *N. gonorrhoeae* e *N. cinerea*, que são agentes causais de gonorreia. Neste ponto, é importante salientar que os candidatos vacinais contra essa bactéria – identificados até então pelos métodos tradicionais (mais de 40 anos de pesquisas) – apresentavam o inconveniente de não garantirem proteção de amplo espectro; ou seja, contra diferentes estirpes da mesma bactéria, ao contrário dos candidatos identificados pela vacinologia reversa, que estão presentes em todas as estirpes testadas. Assim, em poucos anos, essa abordagem genômica possibilitou identificar muito mais candidatos vacinais (que agora já passaram para a fase de avaliação pré-clínica) do que aqueles identificados nas últimas quatro décadas.

Uso de antibióticos no controle de procariotos patogênicos

Além das vacinas, o emprego de antibióticos compreende a principal maneira de restringir as infecções bacterianas. Juntas, tais práticas tiveram bastante sucesso no controle de várias moléstias animais. Para ilustrar, basta citar o exemplo da penicilina, descoberta em 1928 por Alexander Fleming (descoberta que lhe renderia um Prêmio Nobel em 1945), e disponibilizada como fármaco em 1941. Em meio a Segunda Guerra Mundial, o emprego deste antibiótico foi fator decisivo para salvar a vida de milhares de soldados das tropas aliadas. O antibiótico se tornou tão crucial que o governo dos EUA recrutou as principais companhias farmacêuticas da época (totalizando 21) para a produção em massa do fármaco, de modo a garantir o suprimento do "medicamento milagroso" às tropas no *front*. Em consequência, a produção de penicilina naquele país passou de 21 bilhões de unidades em 1943 para 1.663 bilhões em 1944, e para 6,8 trilhões em 1945.

O desenvolvimento de antibióticos é um processo longo e trabalhoso, mas, conceitualmente, simples e empírico. Microrganismos (bactérias e fungos) são isolados do meio ambiente (o solo é o mais preferido) e cultivados em laboratório. Se o microrganismo produz um antibiótico, ele certamente o secreta para o meio onde é cultivado. Extratos dos meios de cultura usados para o cultivo dos microrganismos são então testados exaustivamente para ação antimicrobiana em ensaios de laboratório preestabelecidos. Um ensaio simples, por exemplo, consiste em observar o crescimento de uma bactéria patogênica em um meio sólido quando esta é exposta a diferentes doses de um extrato ou, então, observar o que acontece quando dois organismos são colocados em uma mesma placa de cultura. Se houver produção de antibiótico por parte de um deles, o crescimento do outro será inibido. Constatada a produção de antibiótico, inicia-se um longo processo de purificação e de testes clínicos, que podem levar vários

anos. Desse modo, a produção de antibióticos dá claros sinais de exaustão. Basta mencionar que, nos últimos 40 anos, apenas duas novas classes de antibióticos foram descobertas e que, dos 10 antibióticos aprovados pela agência de controle de doenças dos EUA (*Food and Drug Agency*) desde 1998, apenas dois eram inéditos. Para complicar, não há nenhuma classe de antibiótico para a qual não exista um mecanismo de resistência em bactérias. Nos EUA, por exemplo, cerca de 2 milhões de pessoas são infectadas anualmente por bactérias, e aproximadamente 70% dessas infecções resistem ao menos a um antibiótico. O rápido aparecimento de estirpes resistentes se deve à natureza dinâmica dos genomas procarióticos, como discutido anteriormente; bactérias podem adquirir grandes blocos de genes de outras bactérias e também podem perder genes por meio de mecanismos de HGT. Tal dinamismo cria oportunidades para novas combinações de genes que podem resultar em células resistentes.

Um dos grandes empecilhos à descoberta de novas classes de antibióticos está na necessidade de cultivar em laboratório o microrganismo produtor do antibiótico. Isso restringe muito o número de organismos disponíveis para o uso na prática, e o resultado é que, em 99,9% das vezes, essa abordagem acaba "descobrindo" um antibiótico que já havia sido descoberto anteriormente. A interpretação é clara: já esgotamos as possibilidades de novas descobertas usando microrganismos cultiváveis; assim, precisamos buscar antibióticos em outras fontes, mas que fontes seriam essas?

Busca por novos fármacos

Hoje se sabe, por meio da genômica, que quando se estudam somente os organismos cultivados trabalha-se com apenas 1% do total de organismos existentes – isso pelo fato de ainda não dispormos de conhecimento e de técnicas necessárias para o cultivo dos 99% restantes. Assim, a saída para a descoberta reside em explorar esse imenso microbioma ainda desconhecido. Apenas para se ter noção dessa imensidão, estimativas recentes indicam a existência de 10^6 a 10^8 espécies de procariotos no planeta. Quantos antibióticos, ou outras classes de fármacos, como fármacos antitumorais e também outros empregos industriais e agrícolas (agentes imunossupressores, inseticidas, herbicidas, inibidores de enzimas etc.), ainda estão por serem descobertos?

Antes de prosseguir, é necessário responder: como podemos avaliar o número de organismos não cultiváveis em uma amostra se não podemos cultivá-los? Para isso, pode-se aplicar a ideia de Carl Woese, discutida no início do capítulo, e combiná-la com técnicas de genômica e de biologia molecular. Em resumo, uma estratégia seria a extração do DNA da amostra (que inclui o DNA de todos os organismos aí presentes), a amplificação (via reação em cadeia da polimerase, ou PCR) e a clonagem dos genes que codificam para o RNA ribossômico presentes na amostra de plasmídeos bacterianos que são transformados em uma bactéria hospedeira – geralmente *E. coli*. A essa coleção de plasmídeos recombinantes contendo genes ribossômicos chamamos de "biblioteca" de sequências de genes ribossômicos. Posteriormente, os genes contidos nos plasmídeos recombinantes são sequenciados e comparados com as sequências depositadas em bancos de dados. Pela diversidade de sequências, infere-se a diversidade de espécies presentes na amostra, visto que as sequências desses genes são únicas para cada espécie ou grupo de espécies.

Para a amplificação dos genes, são usados oligonucleotídeos ditos universais, que reconhecem regiões relativamente conservadas desses genes. Uma abordagem alternativa seria a extração do DNA, como descrito anteriormente, seguida de sua fragmentação e construção de uma biblioteca genômica de todos os fragmentos genômicos de todas as espécies presentes (biblioteca "metagenômica") (Figura 12.14). Essa alternativa precisa analisar um número muito maior de clones; no entanto, em teoria, garante uma amostragem isenta de viés, uma vez que não utiliza o passo de amplificação dos genes via PCR antes da clonagem, que pode amplificar preferencialmente genes de alguns organismos, em detrimento de outros.

Metagenômica

Denomina-se "metagenômica" a abordagem de análise de comunidades inteiras por meio de ferramentas genômicas (e não de indivíduos isolados da comunidade). O radical "meta" dá a ideia de "transcender", no sentido de que o objeto de estudo dessa área da genômica transcende (vai além) o nível individual. Trata-se de uma ciência que tem despertado muita atenção por possibilitar a avaliação da diversidade e da funcionalidade de microbiomas inteiros, sem a necessidade de cultivo.

Como descrito na seção anterior, para avaliar a diversidade de certo microbioma, basta analisar a diversidade de sequências de genes ribossômicos presentes na amostra. Essa técnica tem sido aplicada para avaliar a diversidade

microbiológica dos mais diversos ambientes, como do Mar de Sargasso (no Caribe), do intestino de insetos e de humanos e, mais recentemente, da pele humana. Esse interessante trabalho de Grice *et al*. (2008) analisou a diversidade bacteriana em diversos pontos da pele (pontos úmidos, como debaixo do braço; secos, como o antebraço; e sebosos, como o canal auditivo e ao redor das narinas) de 10 indivíduos saudáveis e encontrou uma rica diversidade de sequências do gene ribosomal 16S, sugerindo a existência de 205 gêneros bacterianos no total, pertencentes a 19 filos. Três gêneros foram mais frequentes (*Propinibacterium*, *Staphylococcus* e *Corynebacterium*), correspondendo a 65% das sequências analisadas (de um total de 112.283). Como era de se esperar, esse método revelou um número maior de bactérias daquele previamente conhecido – oriundo de estudos com base em técnicas de isolamento. A diversidade de espécies variou tanto entre locais de uma pessoa como entre pessoas. O ponto da pele com maior diversidade bacteriana, onde foram encontrados 44 gêneros, foi a parte interna do antebraço, um ponto considerado seco. Já o local com menor diversidade foi a linha retroauricular, um ponto seboso. O estudo também revelou o grau de variação da flora da pele entre indivíduos saudáveis para, no futuro, esse resultado ser comparado com o de indivíduos portadores de diversas doenças de pele. O objetivo futuro é o de associar essas doenças a variações na população bacteriana dos locais afetados, na tentativa de identificar suas causas e possíveis métodos de tratamento.

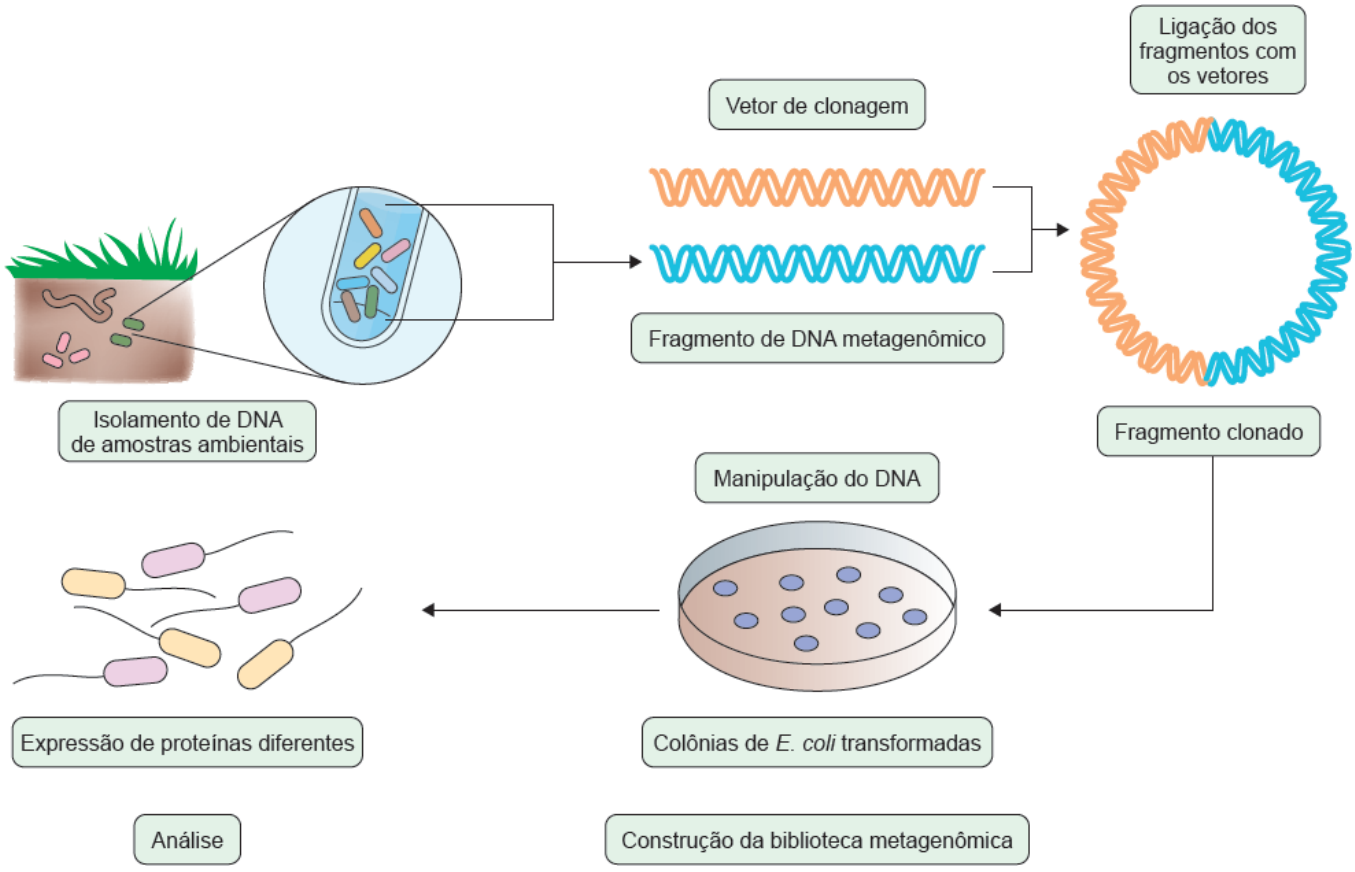


Figura 12.14 Estratégia utilizada para a construção de uma biblioteca metagenômica.

Metagenômica e a funcionalidade de microbiomas

Se alterarmos um pouco a estratégia de análise de nossa biblioteca metagenômica, poderemos usá-la para finalidades mais específicas do que a análise de diversidade de microbiomas obtida a partir da análise de seus genes ribossômicos. É de se supor que em uma biblioteca metagenômica existam plasmídeos recombinantes que carreguem genes que codificam para antibióticos, por exemplo. E aí está uma das grandes expectativas desse tipo de estudo: a de que ele possa resolver o problema da escassez de novos fármacos antibióticos, que foi discutido anteriormente. O problema agora se resume a identificar esses plasmídeos-alvo entre outros milhares que compõem a biblioteca e que carregam outros genes não alvos. Para isso, é preciso dispor de protocolos laboratoriais que possibilitem análises funcionais em *larga escala*. O processo tradicional de descoberta de novos antibióticos requer que os organismos sejam cultivados em laboratório, para só então prosseguir nos ensaios funcionais. Assim, a grande limitação é a de trabalhar apenas com o 1% cultivável. Na abordagem metagenômica, essa limitação não

existe mais. Dessa maneira, podemos explorar mais que 1% dos microrganismos, pois, em vez de cultivá-los, "cultivamos" seus genes por meio da clonagem em plasmídeos – os quais, por sua vez, são inseridos em um organismo capaz de expressar os genes clonados. Tal organismo, muito comumente, é a própria *E. coli*. Se ela receber um plasmídeo contendo um gene de outro organismo que codifica um antibiótico, *é possível* que ela produza o antibiótico *in vitro*, que pode então ser identificado por meio de ensaios laboratoriais usuais.

A abordagem metagenômica de desenvolvimento de novos fármacos já gerou dividendos com a descoberta de dois membros de uma nova classe de antibióticos denominada "turbomicina" (Gillespie *et al*., 2002). Esse antibiótico, que foi descoberto em uma biblioteca metagenômica construída de DNA extraído de solo contendo 24.546 clones, tem largo espectro contra bactérias Gram-positivas e Gram–negativas. Mais importante que a descoberta da turbomicina, o trabalho citado ilustra o grande potencial da técnica, que pode ser estendido a outras aplicações além de antibióticos, tais como enzimas que possam ser empregadas em processos industriais, como detergentes (lipases e proteases) e degradação de polímeros (celulases, quitinases, amilases), entre outras. Imagine, por exemplo, se achássemos a celulase ideal (do ponto de vista industrial), que degradasse o bagaço da cana-de-açúcar (ou outro produto vegetal, como a madeira) de maneira econômica? Estaria aí a chave para convertermos o polímero mais abundante no planeta (celulose), e do qual o Brasil é um grande produtor, em glicose e este, posteriormente, em álcool. Poderíamos substituir em larga escala os combustíveis fósseis por um combustível mais limpo. Este é, atualmente, um dos grandes desafios da genômica.

Considerações finais

Dentre os avanços obtidos a partir da ciência genômica, destacam-se:

- Alterações no conceito de gene, que incluem as regiões regulatórias: "gene pode ser então definido como uma unidade de informação genética que controla a síntese de um polipeptídeo ou de uma molécula de RNA. Inclui as regiões 5' e 3' não codificantes que estão envolvidas na regulação da transcrição e tradução e todas as regiões intrônicas daquele gene"
- As metodologias de sequenciamento avançam a passos largos e deixaram de ser limites ao avanço no conhecimento. Atualmente, esse limite é a capacidade humana de análise dos dados produzidos
- A análise de sequências possibilitou a descoberta do terceiro domínio na árvore da vida, completamente despercebido por análises clássicas. Woese e Fox identificaram o domínio Archaea
- A análise comparativa das sequências completas de genomas de bactérias torna possível identificar diferenças genéticas associadas ao habitat e papel ecológico desses organismos, os quais são mantenedores dos ciclos biogeoquímicos e de permanência da vida na terra
- Em procariotos, o tamanho dos genomas e o número de genes estão diretamente correlacionados à condição de vida parasitária ou livre desses organismos
- Em eucariotos, os paradigmas C (conteúdo de DNA) e G (tamanho do genoma e número de genes) continuam sendo um dos pontos mais importantes de estudo, uma vez que não existe uma correlação positiva entre eles e a complexidades dos organismos
- A determinação dos conceitos de genomas *pan*, essencial e assessório, revela que indivíduos de uma mesma espécie podem carregar um número diferente de genes. Esses dados são mais relevantes em bactérias, nos quais, em alguns casos, indivíduos de uma mesma espécie podem variar em até 30% na presença ou ausência de genes. Procariotos também apresentam na transferência horizontal de genes um dos *drivers* da sua evolução
- Os avanços na metodologia de sequenciamento levaram ao surgimento de estudos metagenômicos; neste caso, o sequenciamento vai além de um único organismo, mas de amostras ambientais, com o intuito de entender comunidades microbianas e seu papel ecológico
- Por fim, as aplicações imediatas da ciência genômica são inúmeras, com destaque para o desenvolvimento de vacinas e medicamentos antimicrobianos, com base na análise comparativa de sequências.

As perspectivas são inúmeras, que não podem ser todas descritas neste capítulo, mas tivemos como objetivo sedimentar uma base sólida para que o leitor possa avançar nos estudos genômicos e descobrir suas mais incríveis aplicações.

Bibliografia

Dayhoff MO, Eck RV, Chang MA, Sochard MR. Atlas of Protein Sequence and Structure. v. 1. National Biomedical Research Foundation, Silver Spring, MD; 1965.

Fleischmann RD, Adms MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science. 1995;269(5223):469-512.

Gillespie DE, Brady SF, Bettermann AD, Cianciotto NP, Liles MR, Rondon MR, et al. Isolation of antibiotics turbomycin A and B from a metagenomic library of soil microbial DNA. Appl Environ Microbiol. 2002;68:4301-6.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A draft sequence of the neandertal genome. Science. 2010;328(5979):710-22.

Grice EA, Kong HH, Renaud G, Young AC, NISC Comparative Sequencing Program, Bouffard GG, et al. A diversity profile of the human skin microbiota. Genome Research. 2008;18:1043-50.

Hahn MW, Wray GA. The g-value paradox. Evolution & Development. 2002;4(2):73-5.

Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chemistry & Biology. 1998;5(10):R245-9.

Margulies M. Egholm M, Altman WE, Attiya S, Bader JS, Bember LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005;437(7057):376-80.

Nunes FMF. Genes, Genomas, RNAs não-codificadores e a complexidade biológica. Genética na escola. 2011;6(1):80-3.

Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, et al. Nucleotide sequence of bacteriophage φX174 DNA. Nature. 1977;265(5596):687-95.

Simpson AJ, Reinach FC, Arruda P, Abreu FA, Acencio M, Alvarenga R, et al. The genome sequence of the plant pathogen Xylella fastidiosa. The Xylella fastidiosa Consortium of the Organization for Nucleotide Sequencing and Analysis. Nature. 2000;406(6792):151-9.

Taft RJ, Mattick JS. Increasing biological complexity is positively correlated with the relative genome-wide expansion of non-protein-coding DNA sequences. Genome Biology. 2003;5:P1.

Taft RJ, Pheasant M, Mattick JS. The relationship between non-protein-coding DNA and eukaryotic complexity. Bioessays. 2007;29(3):288-99.

White O, Eisen JA, Heidelberg JF, Hickey EK, Peterson JD, Dodson RJ, et al. Genome sequence of the radioresistant bacterium Deinococcus radiodurans R1. Science. 1999;286(5444):1571-7.

Whittaker R. New concepts of kingdoms of organisms. Science. 1969;163(3863):150-60.

Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci. 1977;74(11):5088-90.

13 Transcriptômica e Proteômica

Ricardo De Marco

Atualmente, o estudo dos conjuntos de moléculas de RNA e proteínas em uma célula é definido como transcriptômica e proteômica. Ambos os conjuntos são originários de um processo dinâmico nas células, derivado de programações de respostas ao ambiente, ou de processos de diferenciação celular. Neste capítulo, será abordada a importância do estudo de alterações celulares do ponto de vista da dinâmica de mudanças de transcriptomas e proteomas. São comentados métodos para identificar essas alterações do transcriptoma, desde análises de hibridação de ácidos nucleicos (p. ex., *Northern blot* e microarranjos) até técnicas modernas de sequenciamento de genes. Também são abordados métodos para determinação do conjunto de proteínas expressas em células e separação em gel 2D, incluindo análises em espectrômetros de massa para identificação de proteínas específicas e de interação proteica, através de duplos híbridos.

O que são o transcriptoma e o proteoma?

As células dos seres vivos apresentam diversos conjuntos de biomoléculas com diferentes funções. As principais classes dessas biomoléculas são carboidratos, lipídeos, proteínas e ácidos nucleicos. As proteínas compõem um conjunto muito importante de biomoléculas, pois têm grande importância na efetivação de uma série de processos celulares. Ácidos nucleicos são as moléculas responsáveis pela transmissão da informação genética na forma de DNA e intermediário para produção de proteínas por esse repositório, na forma de RNA.

Com a finalidade de sistematizar o estudo dessas biomoléculas, passou-se a designar esses conjuntos com diferentes nomes. Por exemplo, o conjunto das moléculas de DNA que compõe o repositório genético de um indivíduo é designado genoma. Por sua vez, o conjunto de moléculas de RNA de uma célula é o seu transcriptoma, e o conjunto de proteínas dessa mesma célula é designado proteoma. Não é surpreendente que esses três conjuntos (genoma, transcriptoma e proteoma) estejam intimamente relacionados, pois, de acordo com o dogma central da biologia molecular, as moléculas de DNA servem como molde para a transcrição de moléculas de RNA que, por sua vez, serão traduzidas em proteínas (Figura 13.1).

Resumidamente, sem considerar as modificações pós-traducionais de proteínas, por exemplo, toda a informação do transcriptoma estaria contida no genoma, e o proteoma, por sua vez, poderia ser deduzido a partir do transcriptoma. No entanto, tal afirmação não leva em conta o caráter dinâmico do transcriptoma e do proteoma de um organismo. Desse modo, o estudo desses dois conjuntos de moléculas não é uma mera extensão do estudo de genomas, constituindo campos próprios, que irão prover diferentes tipos de informação e com diversos desafios na obtenção e interpretação dos dados.

O transcriptoma é composto de diferentes tipos de moléculas de RNA, mas o seu componente mais estudado é o RNA mensageiro (mRNA); e isso acontece a tal ponto que, muitas vezes, o termo transcriptoma é utilizado apenas para designar o conjunto de mRNA de uma célula, ignorando os outros tipos de RNA. Tal preferência é compreensível, visto que apenas o mRNA tem função de codificação de proteínas e a maioria dos outros tipos de RNA (rRNA, tRNA, snRNA) tem função estrutural. Esses outros tipos de RNA contêm um número limitado de moléculas distintas, existindo maior importância no seu estudo individual e entendimento da sua função em alguns eventos celulares, mas não na sua caracterização enquanto grupo de moléculas, como geralmente realizado em

estudos transcriptômicos. Uma exceção é o grupo de RNA não codificantes (ncRNA) – um conjunto de RNA regulatórios (cujas propriedades serão melhor exploradas no Capítulo 16), que também são comumente alvos de estudos transcriptômicos.

O proteoma, por sua vez, é formado pelo conjunto de proteínas de um organismo. Proteínas são moléculas complexas e que apresentam comportamentos bem distintos entre si. Dependendo dos aminoácidos que constituem uma proteína, ela pode ser solúvel ou estar inserida em uma membrana, por exemplo. Além disso, diversos processos realizados por proteínas dentro da célula eucariótica ocorrem em compartimentos específicos. Desse modo, não é de se estranhar que a distribuição de uma dada proteína dentro de uma célula seja heterogênea. Por exemplo, existem proteínas responsáveis pela transcrição do DNA que são encontradas somente dentro do núcleo, ou transportadores que são encontrados somente na membrana plasmática da célula (Figura 13.2). Em adição a sua cadeia proteica, muitas proteínas sofrem modificações pós-traducionais, nas quais moléculas (como carboidratos ou fosfatos) são adicionadas às cadeias laterais de alguns aminoácidos que compõem a proteína madura. Consequentemente, o proteoma de uma célula não é resultado da mera tradução dos transcritos que se encontram no seu citoplasma, mas o resultado de uma série de processos que irão gerar proteínas maduras e distribuídas de modo desigual dentro de uma célula.

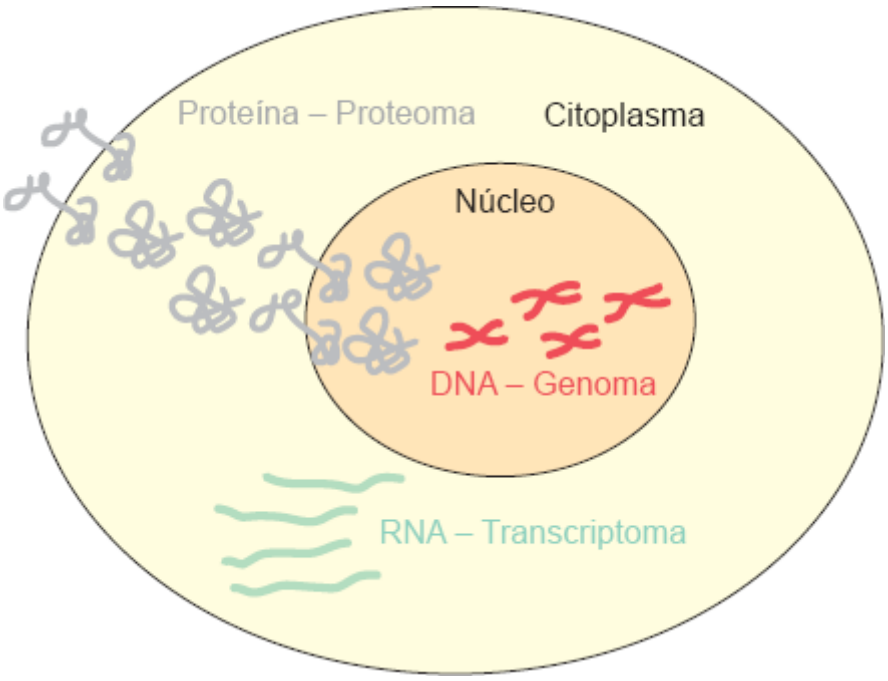


Figura 13.1 Célula eucariótica e localização de alguns dos conjuntos de biomoléculas.

Dinamicidade do transcriptoma e complexidade das respostas celulares

A princípio, toda a informação necessária para o desenvolvimento e funcionamento de uma célula encontra-se em seu genoma. No entanto, ao comparar diferentes células de um mesmo organismo, observa-se que elas apresentam características bem distintas. Por exemplo, comparando um neurônio e uma célula muscular (Figura 13.3), é possível notar que elas têm morfologias bem diversas, que são reflexo de sua função; no caso, a condução de impulsos nervosos para os neurônios e a realização de movimentos, respectivamente. Apesar de as informações contidas no genoma dessas duas células serem praticamente as mesmas, a utilização de tal informação ocorre de modo diferente. É possível notar, portanto, que cada célula do nosso corpo dispõe de uma programação genética distinta que irá determinar o seu fenótipo.

O que é, na prática, tal programação genética e como ela consegue produzir diferenças nas células? Primeiramente, é necessário lembrar-se de que o material genético da célula (o DNA) é relativamente inerte e que grande parte das funções celulares é promovida por proteínas. No entanto, um determinado tipo de proteína somente será produzido se o mRNA correspondente estiver disponível para tradução no citoplasma da célula; quanto maior a quantidade de um determinado tipo de mRNA, maior será a produção da proteína correspondente.

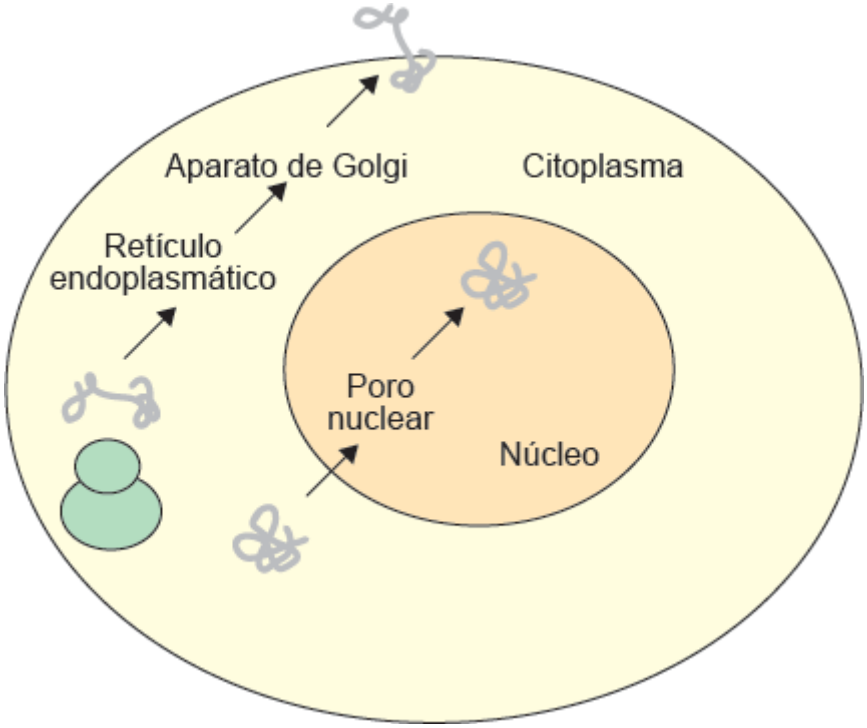


Figura 13.2 Algumas das diferentes rotas que a proteína pode percorrer em uma célula eucariótica.

Existem alguns mRNA que codificam para proteínas que realizam funções básicas nas células (chamadas, em inglês, de *housekeeping genes*), para os quais geralmente não se verificam muitas mudanças em sua concentração no citoplasma entre os diferentes tipos celulares. No entanto, há uma série de outros mRNA que codificam para proteínas com funções mais especializadas, cuja quantidade pode variar significativamente entre dois tipos celulares diferentes. Como resultado, cada célula poderá apresentar um diferente repertório de moléculas de mRNA (Figura 13.4). Assim, dois tipos celulares distintos de um indivíduo devem apresentar um genoma praticamente idêntico, mas um transcriptoma bem diferente.

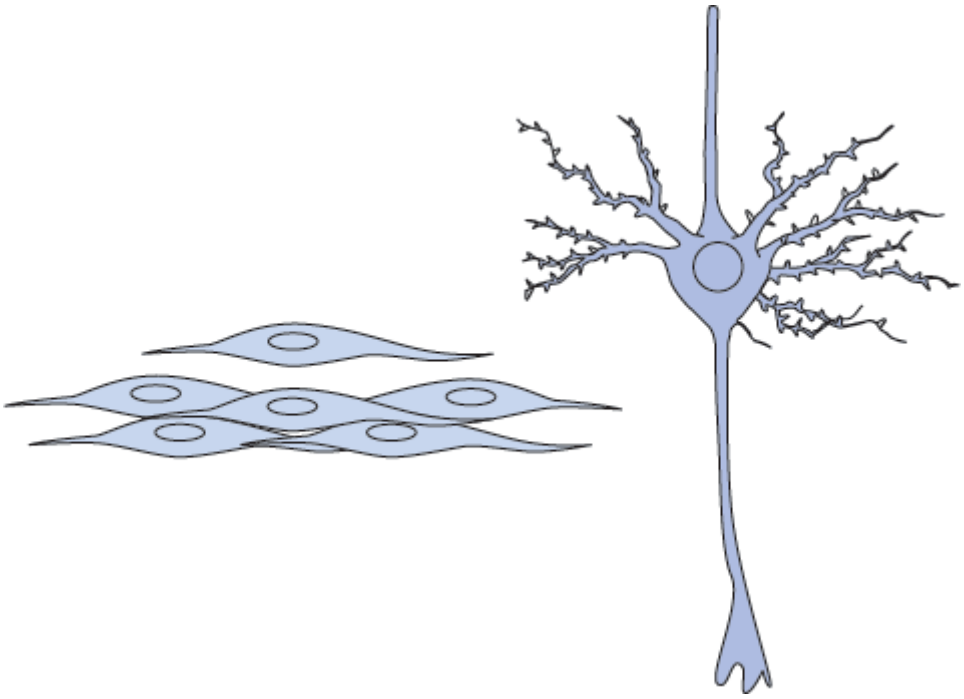


Figura 13.3 Morfologia de células musculares (**A**) e nervosas (**B**).

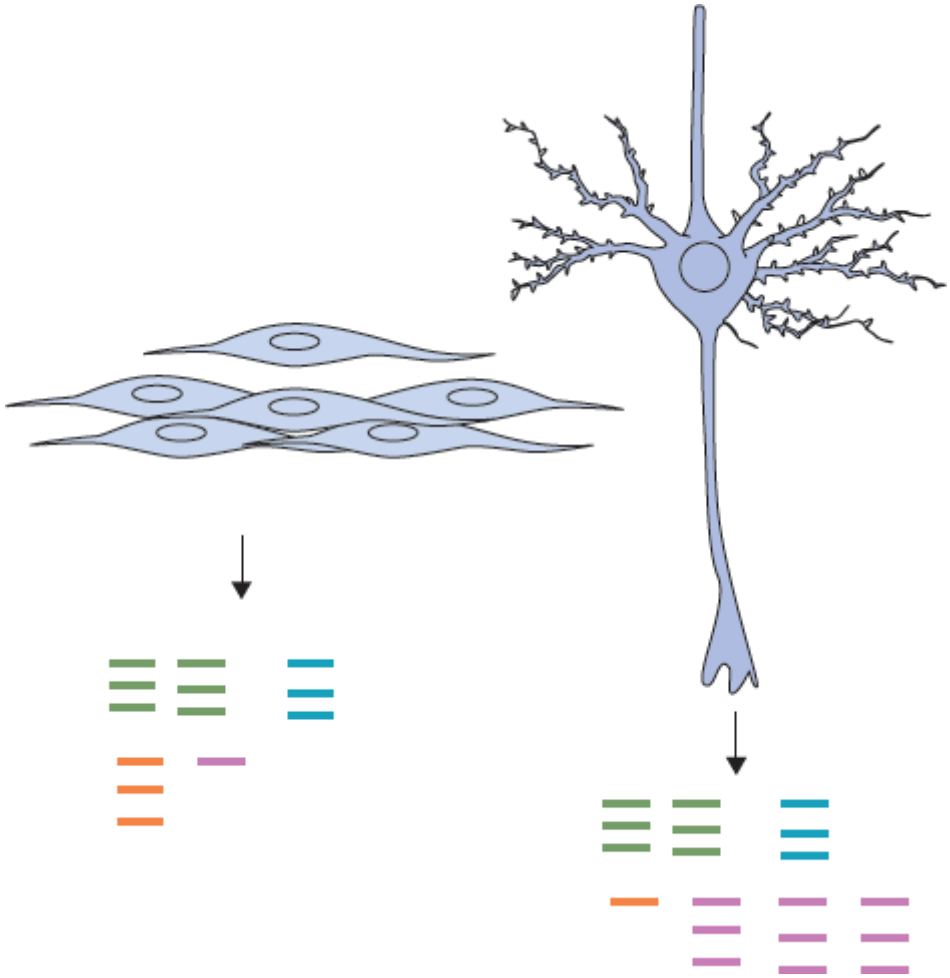


Figura 13.4 Representação dos conjuntos de transcritos distintos existentes em diferentes tipos celulares. As barras de diferentes cores representam mRNA originários de diferentes genes.

Ao analisar o transcriptoma de um dado tipo celular, estamos inferindo a programação dessa célula por meio dos níveis de mensagens de mRNA codificando para as diferentes proteínas. Por exemplo, ao analisar o transcriptoma de uma célula nervosa, certamente serão encontrados maiores níveis de mRNA codificando para proteínas de canais iônicos, responsáveis pela condução de impulsos nervosos, do que em células de epitélio, que não apresentam essa função. Embora haja alguns casos em que é óbvia a associação de função com a abundância de transcritos em uma célula, existem muitos nos quais a função da proteína é desconhecida ou o contexto celular no qual uma proteína atua não é bem entendido. Nesse caso, a verificação de maior abundância de um determinado transcrito em um determinado tipo celular pode servir como ponto de partida para tentar correlacionar a função da proteína codificada pelo transcrito com as peculiaridades do tipo celular.

Do mesmo modo, se os transcriptomas de tipos celulares idênticos de organismos diferentes forem comparados, eventuais diferenças entre eles podem refletir adaptações de cada espécie. Esse tipo de análise faz mais sentido quando comparamos duas espécies relativamente próximas evolutivamente, onde o conjunto de transcritos é praticamente o mesmo e, portanto, é relativamente fácil associar todos os transcritos de um organismo ao seu equivalente no outro organismo. Ao comparar os genomas dos seres humanos com os dos chimpanzés, encontra-se apenas 1,23% de diferença no conjunto de bases dos dois genomas. Apesar do reconhecimento que mutações que causem alterações na sequência proteica têm importância nas mudanças fenotípicas entre as duas espécies, é cada vez mais reconhecido que a alteração do padrão de transcrição das células representa um fator significativo no processo de diferenciação das duas espécies.

Tudo isso implica a necessidade de um sistema complexo de regulação das quantidades de todos os mRNA que uma célula produz. De fato, há diversas vias de sinalização cujo alvo final é uma classe de proteínas com a função de regular a produção diferencial de RNA, os chamados fatores de transcrição. Estes agem sobre a maquinaria de transcrição da célula, estimulando a transcrição a partir de certos *loci* específicos do genoma. Em razão desse complexo sistema de controle, é possível afirmar que a diversidade de transcriptomas observada entre diferentes tipos celulares ou de organismos não é fruto do acaso.

Em adição à diversidade intrínseca que o transcriptoma apresenta entre os diversos tipos celulares, o transcriptoma de uma célula também pode sofrer mudanças em decorrência de estímulos aos quais a célula é exposta. Por exemplo, a exposição de uma célula a moléculas sinalizadoras, como hormônios, pode levar a uma mudança no seu programa genético, o que pode se refletir na produção de diferentes transcritos, codificando para proteínas que exerçam funções críticas para os eventos sinalizados (Figura 13.5). Como esses sistemas de sinalização agem em cascata, é possível que a ligação de poucas moléculas de hormônios na superfície celular leve à estimulação de centenas de moléculas de fatores de transcrição que irão atuar em diferentes regiões do genoma, estimulando a produção de diversos transcritos distintos. Assim, é possível que pequenos estímulos levem a grandes mudanças no transcriptoma de uma célula. Por exemplo, determinado hormônio pode sinalizar para maior abundância de glicose, o que fará com que as células do fígado produzam maior quantidade de transcritos codificando proteínas responsáveis pela importação e armazenamento desse açúcar na forma de glicogênio.

Diversos pesquisadores se dedicam ao estudo do transcriptoma de células submetidas a diversas condições ou estímulos. São experimentos nos quais as células são expostas a diferentes tipos de estímulo e, depois, têm o seu mRNA extraído para possibilitar a sua comparação com o mRNA de células controle que não foram submetidas a este mesmo estímulo. Por meio de uma análise global dos RNA utilizando diferentes métodos, é possível comparar os dois transcriptomas e detectar transcritos cuja abundância tenha sido alterada pelo tratamento. A partir da análise desses dados, busca-se por padrões que possibilitem inferir uma programação genética específica de respostas a esses estímulos. Experimentos utilizando séries temporais, em que amostras de células submetidas a diferentes tempos de tratamento são analisadas, permitem acompanhar os diferentes passos da resposta celular ao estímulo. Esse tipo de estudo pode auxiliar no entendimento de mecanismos moleculares de adaptação das células às mais diversas condições.

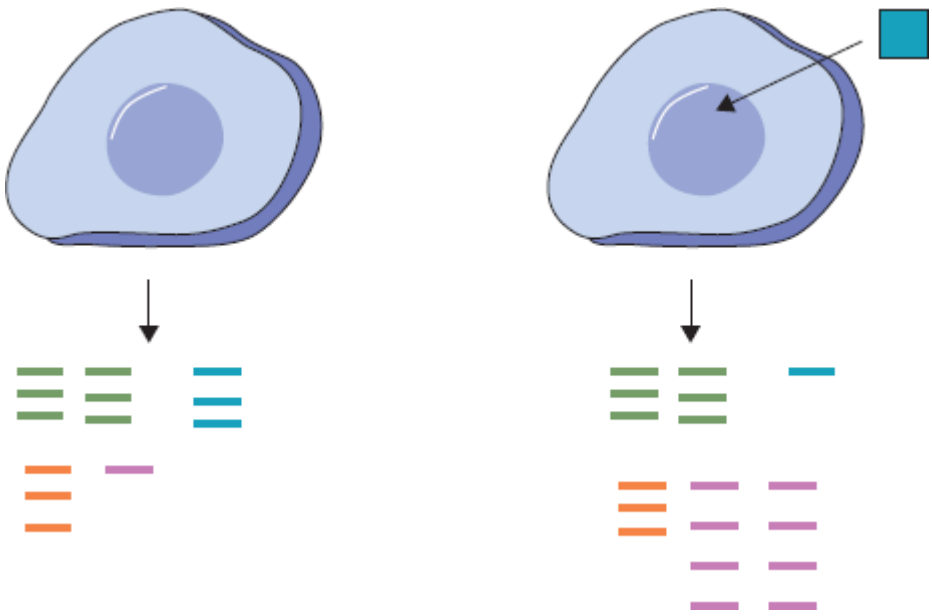


Figura 13.5 Mudança do conjunto de transcritos de uma célula, mediante estímulo de uma substância química (representada pelo quadrado).

Métodos para estudar o transcriptoma

Isolamento de RNA mensageiro e síntese de DNA complementar

Conforme descrito na seção anterior, há grande interesse no estudo do transcriptoma, pois, por meio da descrição dos transcritos e sua abundância relativa nos diferentes tipos celulares ou organismos, é possível inferir uma série de mecanismos celulares subjacentes aos processos específicos de cada célula. Em geral, o primeiro passo para estudar o transcriptoma de uma célula é pelo isolamento do seu mRNA. Há diversos métodos utilizados para separar o mRNA dos outros componentes da célula; o mais utilizado é a passagem do material biológico por uma coluna que contém polímeros de desoxitimidina (poli-dT) imobilizados em uma resina. Como resultado, as moléculas de RNA de organismos eucariotos que têm uma cauda de adeninas em sua porção 3' serão hibridadas com o poli-dT e passarão a ficar imobilizadas nesta coluna, enquanto as outras moléculas biológicas continuarão livres, podendo ser retiradas facilmente da coluna (Figura 13.6). É muito importante retirar proteínas e DNA genômico existentes no

homogenato celular das quais são extraídas as moléculas de RNA, visto que, dentre as proteínas presentes, podem estar proteínas degradadoras do RNA, e o DNA genômico pode ser um contaminante que irá interferir nas análises em algumas das técnicas normalmente utilizadas para o estudo do mRNA.

Após o isolamento do mRNA, este costuma ser convertido em uma molécula de DNA por meio do uso de uma enzima chamada transcriptase reversa. Ela tem origem viral e é utilizada por esses vírus para transformar seu material genético na forma de RNA em moléculas de DNA que podem ser integradas no genoma da célula infectada. Como o RNA serve de molde para a síntese da molécula de DNA, toda a informação contida no RNA é preservada na molécula de DNA. O motivo pelo qual convertemos a molécula de mRNA em uma molécula de DNA deve-se ao fato de que moléculas de RNA são facilmente degradadas por enzimas chamadas RNAses, produzidas em grandes quantidades por todos os organismos vivos e presentes no ambiente. Além disso, uma série de técnicas de manipulação foi desenvolvida para moléculas de DNA e, ao transformar moléculas de RNA em DNA, é possível fazer uso delas para o estudo de nossas moléculas. O DNA proveniente deste processo de transcrição reversa é denominado DNA complementar (cDNA).

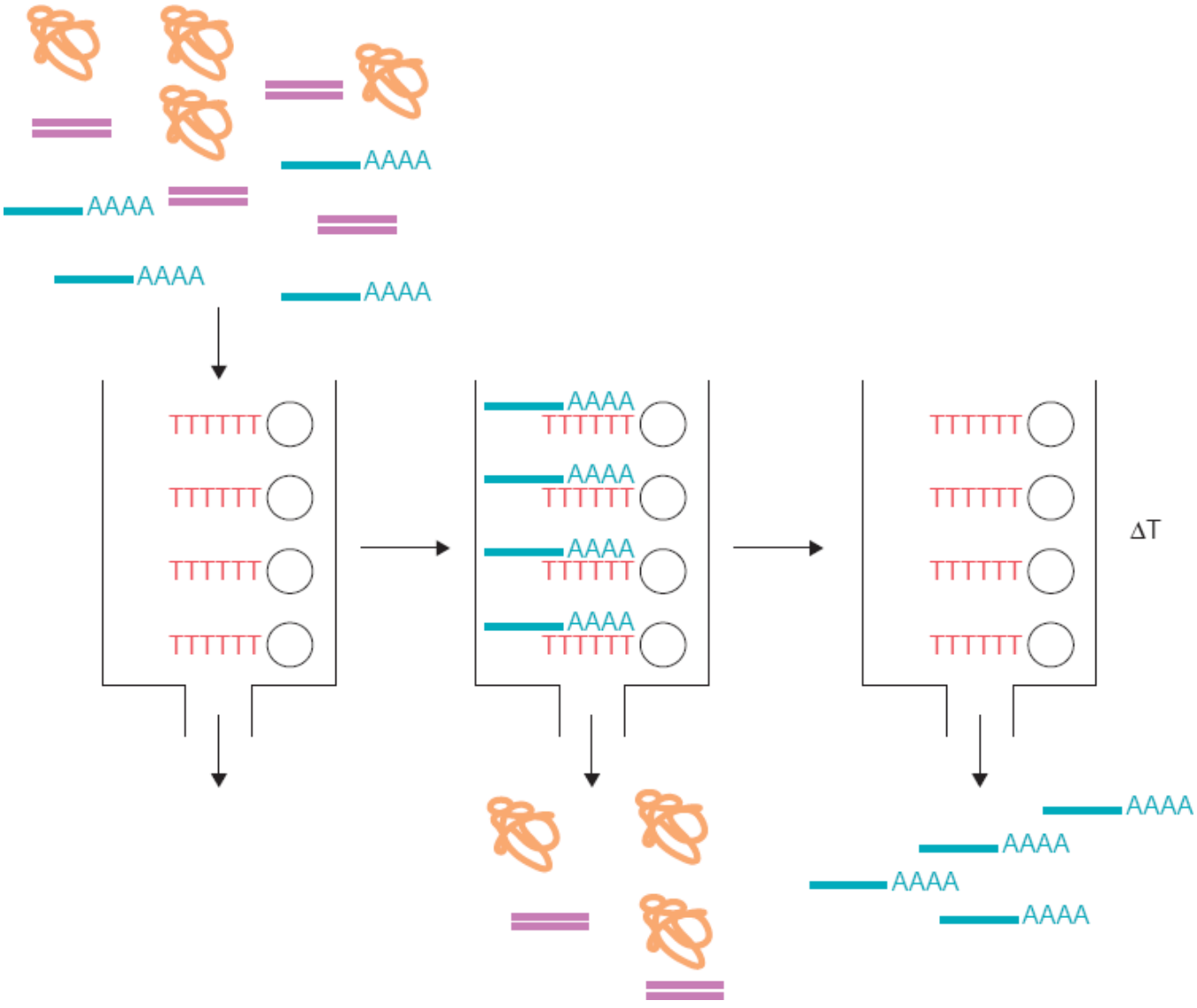


Figura 13.6 Purificação de mRNA com a utilização de colunas com oligonucleotídeos poli-dT imobilizados. Colunas contendo oligonucleotídeos poli-dT imobilizados (em vermelho) são carregadas com um homogenato celular contendo moléculas de mRNA (barras azuis), DNA genômico (barras duplas roxas) e proteínas (enovelados laranjas). Somente as moléculas de mRNA são retidas na coluna, devido à complementaridade do oligo-dT com as caudas poli-A presentes na região 3' do mRNA. Após lavagem da coluna para retirada de todo DNA e proteínas contaminantes, o mRNA puro é recuperado com o aumento da temperatura, que desestabiliza a ligação deste com o oligo-dT.

Sequenciamento de bibliotecas de DNA complementar

O modo mais simples de explorar cDNA é realizar o seu sequenciamento. Técnicas de sequenciamento de DNA foram desenvolvidas, inicialmente, na década de 1970 e possibilitam que a ordem de nucleotídeos em uma molécula

de DNA seja deduzida. Com o intuito de permitir o sequenciamento sistemático de cDNA derivadas de um transcriptoma, foram desenvolvidas técnicas de montagem das chamadas bibliotecas de cDNA, que consistem na inserção das moléculas de cDNA geradas a partir da reação de transcrição reversa em um vetor de propagação (em geral, um plasmídeo), por meio de técnicas de clonagem. A clonagem é realizada utilizando as moléculas de cDNA geradas a partir de todas as moléculas de mRNA extraídas de um tecido ou célula. Em função disso, as moléculas de cDNA da biblioteca irão representar o conjunto de moléculas de mRNA daquele tecido ou célula no momento em que ocorreu a extração de seu mRNA. A inserção em um vetor de propagação torna possível que essas moléculas de cDNA possam ser utilizadas na transformação de células bacterianas, o que permitirá que elas sejam facilmente estocadas e amplificadas. O sequenciamento desses clones possibilita que se defina a ordenação dos nucleotídeos do polímero de ácido nucleico, e a técnica predominantemente utilizada é a técnica de Sanger.

O sequenciamento em larga escala de clones de bibliotecas de cDNA utilizando a técnica de Sanger foi popularizado na década de 1990, com o maior acesso de pesquisadores a sequenciadores automáticos. Esse tipo de equipamento possibilita que a realização de um grande número de sequenciamentos de moléculas distintas (cerca de 100 sequências em máquinas comerciais modernas) ocorra de maneira paralela em um único experimento. Em geral, esse tipo de sequenciamento é realizado por escolha randômica de clones e realização de uma única reação de sequenciamento por clone. Isso faz com que cada clone geralmente não seja sequenciado em toda a sua extensão, pois uma reação de sequenciamento é capaz de determinar de 600 a 800 bases de um polímero e, em geral, cada molécula de cDNA formada a partir do mRNA contém milhares de bases.

Esse tipo de sequência parcial gerada a partir de sequências de cDNA foi denominado, em inglês, *expressed sequence tag* (EST), que significa “etiqueta de sequência expressa”. A ideia associada à produção desse tipo de sequência deriva do fato de que é extremamente trabalhoso gerar sequências completas de todos os transcritos de uma célula ou tecido, pois requer a obtenção de clones de cDNA que representem o transcrito em toda sua extensão. Isso nem sempre é comum, devido à baixa processividade da transcriptase reversa. Além disso, mesmo quando obtemos um clone completo, é necessário realizar múltiplas reações de sequenciamento iniciadas em pontos diferentes da molécula para obter uma cobertura total de sequenciamento.

A geração de sequências parciais de transcritos é relativamente mais direta, pois exige apenas que a molécula de cDNA seja clonada e a realização de uma única reação de sequenciamento iniciada em uma das pontas da molécula. Muitos pesquisadores preferem fazer um sequenciamento parcial, que possibilite a geração de sequências representando uma grande amostragem dos transcritos de uma célula, em vez de produzir sequências completas de poucos transcritos.

Apesar de esse tipo de abordagem fornecer uma informação incompleta sobre o transcriptoma, ele é bastante útil, pois, mesmo com o sequenciamento parcial de um transcrito, geralmente é possível – por meio de análises bioinformáticas – sugerir a função da proteína codificada por ele. Isso é feito pelo alinhamento da sequência parcial da proteína codificada pelo transcrito, com proteínas com função conhecida presentes em bancos públicos de dados sobre moléculas biológicas (esse alinhamento é realizado utilizando o programa BLAST). Caso haja um alto nível de similaridade detectado entre o fragmento da proteína codificada pelo transcrito e alguma proteína com função conhecida, pode-se atribuir uma possível função à proteína codificada por esse transcrito. Obviamente, a confirmação da função da proteína somente é possível a partir da obtenção da sequência completa do transcrito e a realização de experimentos bioquímicos com a proteína derivada deste transcrito. A partir de uma sequência parcial, pode-se recuperar o transcrito completo por meio do clone correspondente na biblioteca de cDNA, ou de técnicas de amplificação de pontas de cDNA (Figura 13.7).

Com a produção de um grande número de sequências parciais de transcritos de bibliotecas de cDNA, surgiu a necessidade de criar bancos de dados específicos para armazenar esse tipo de informação. Em alguns casos, o banco de dados gerado é um mero repositório onde estão disponíveis todas as sequências parciais geradas. No entanto, há casos em que análises computacionais irão agregar todas as sequências parciais produzidas para cada transcrito em uma única sequência, esperando-se obter sequências mais completas e diminuir a redundância desses bancos de dados (Figura 13.8).

Além disso, a contagem da frequência de sequenciamento de clones de um mesmo transcrito pode ser utilizada como uma medida indireta de sua abundância relativa. Isso ocorre porque a amostragem de clones da biblioteca tende a ser randômica; logo, se supormos que uma célula tem 100.000 transcritos e, dentre estes, existe um transcrito A de alta abundância com 1.000 cópias, a chance de este ser amostrado no sequenciamento de uma cópia

é de 1/100. Considerando um transcrito B de baixa abundância com apenas 10 cópias, a chance de esse transcrito ser amostrado é de apenas 1/10.000. Em um sequenciamento randômico de um número razoável de clones da biblioteca de cDNA desse exemplo hipotético, a frequência com que o transcrito A será amostrado será 100 vezes maior que o transcrito B. A partir desse princípio, é possível comparar a frequência de sequenciamento de clones em bibliotecas de cDNA originárias de diferentes amostras para inferir quais transcritos são mais abundantes em cada amostra, possibilitando levantar-se hipóteses sobre o papel desses transcritos nos diferentes processos celulares.

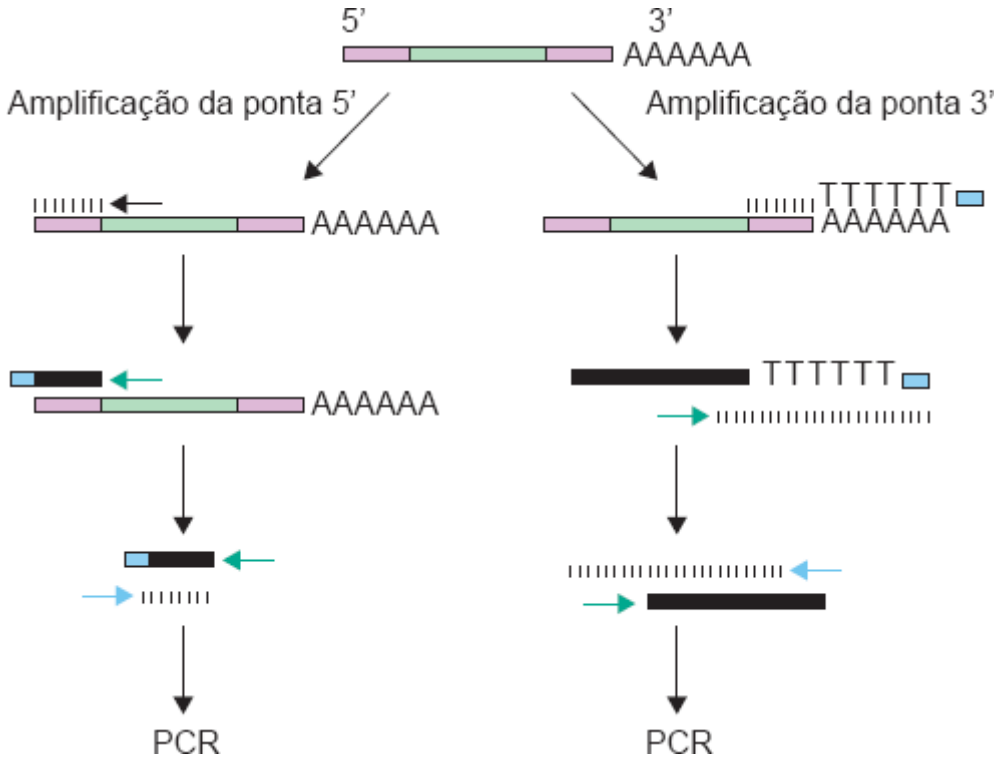


Figura 13.7 Técnica de amplificação de pontas de mRNA. A barra no topo da figura representa uma molécula de mRNA, contendo uma região central, cuja sequência é conhecida (em verde), e pontas desconhecidas (em rosa). As setas representam oligonucleotídeos iniciadores sintéticos; e a região preta em pontilhado, uma fita de cDNA recém-sintetizada. A barra azul representa um adaptador de DNA cuja sequência é conhecida e que é ligado à ponta 5', ou encontra-se adjacente a um oligonucleotídeo poli-dT. Nota-se que, no passo final, a molécula é flanqueada por dois oligonucleotídeos iniciadores, o que irá possibilitar a sua amplificação pela reação em cadeia da polimerase (PCR).

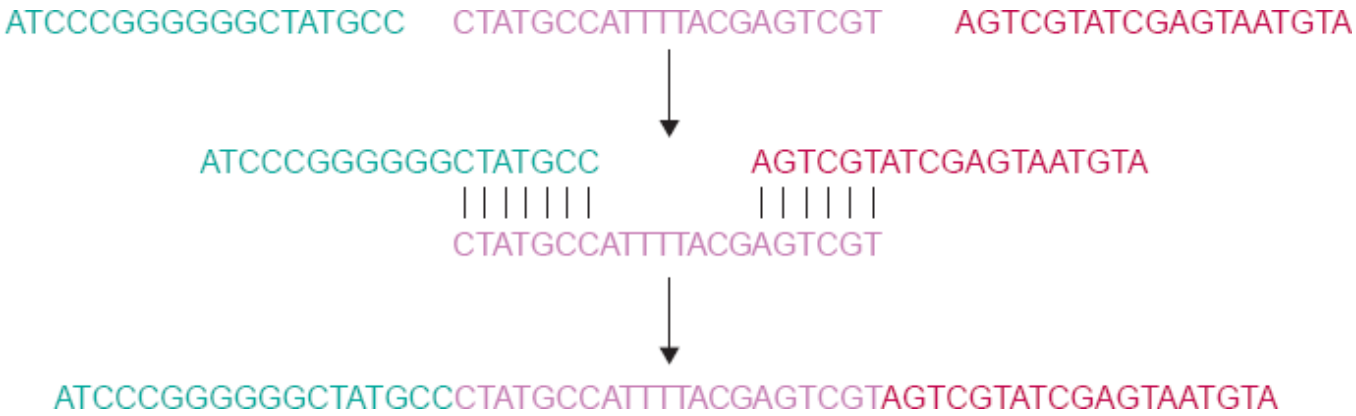


Figura 13.8 Procedimentos utilizados para agregar sequências parciais de um transcrito para a dedução de uma sequência mais completa. Textos em verde, roxo e vermelho escuro representam diferentes sequências parciais obtidas de um transcrito. A partir da determinação de regiões de sobreposição entre as pontas dessas sequências (representada por barras horizontais), é possível deduzir uma sequência final que agregue as informações de todas juntas.

Nos últimos anos, os métodos de sequenciamento de DNA evoluíram muito com o surgimento de aparelhos de sequenciamento de segunda geração (*next generation sequencing*, em inglês). O sequenciamento tradicional utilizando a metodologia de Sanger exige que cada molécula de cDNA seja isolada por técnicas de clonagem e, depois disso, uma reação de sequenciamento separada é realizada para cada clone diferente. Esse tipo de

procedimento é extremamente trabalhoso quando se pretende sequenciar milhares de clones derivados de uma mesma biblioteca. Na tecnologia de segunda geração, há um processo de paralelização das reações de sequenciamento fazendo com que ocorram simultaneamente, dentro de um único ambiente, milhares de reações. A Figura 13.9 representa uma reação em cadeia da polimerase em emulsão, que exemplifica uma das técnicas que são utilizadas para possibilitar essa paralelização de processos. Isso faz com que haja um grande aumento do número de sequências de cDNA que podem ser geradas a um custo relativamente baixo. Esse tipo de técnica está se tornando cada vez mais popular para a caracterização de transcriptomas, já que possibilita uma grande cobertura dos transcritos de uma célula ao custo de alguns poucos milhares de dólares.

Análise do transcriptoma com técnicas de hibridação

Além das técnicas de sequenciamento, existem outros tipos de técnicas que possibilitam inferir a abundância dos mRNA em uma amostra. Várias técnicas de detecção de mRNA se baseiam na hibridação entre fitas de DNA e mRNA ou entre fitas de DNA e cDNA. É possível utilizar esse tipo de técnica para detectar uma molécula específica de DNA, pois ácidos nucleicos tendem a formar duplas-fitas somente quando há duas fitas complementares. Portanto, se houver um conjunto de moléculas de ácido nucleico na forma de fita simples, é possível utilizar uma molécula fita simples específica como sonda e verificar a quantidade de moléculas no conjunto inicial de ácidos nucleicos que são capazes de se hibridar com essa sonda. Com base nesse princípio, foram desenvolvidas técnicas como o *Northern blot* e de microarranjos de DNA, que possibilitam a detecção do nível de transcrição de um ou mais transcritos e diversas amostras. Ao contrário do sequenciamento, esse tipo de técnica exige conhecimento prévio da sequência das moléculas de RNA para as quais se quer medir os níveis de transcrição, pois as sondas serão sintetizadas a partir das sequências dos transcritos que se deseja monitorar. Assim, esse tipo de técnica não serve para a primeira exploração de um transcriptoma.

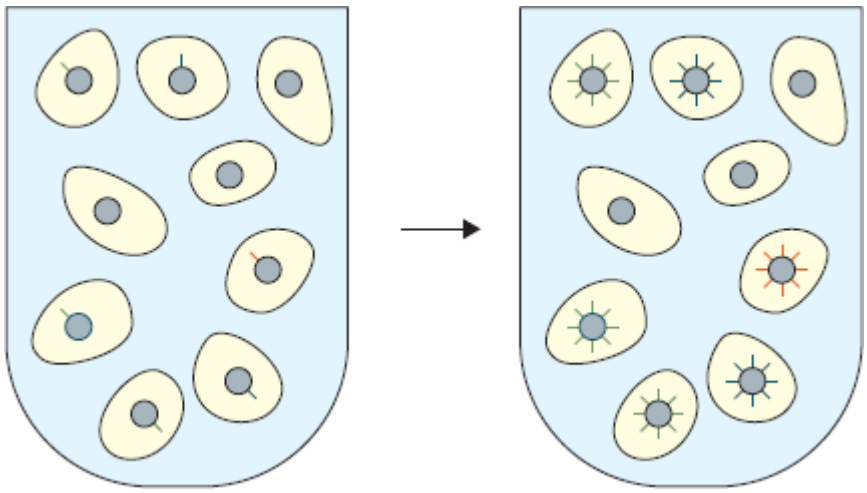


Figura 13.9 Reação de amplificação de DNA em emulsão. Devido à emulsificação da solução, cada gotícula formada funcionará como um reator em separado. As fitas de DNA amplificados se ligam a partículas especiais colocadas na solução, de modo que, após a quebra da emulsão, seja possível separar as moléculas provenientes de cada microrreator.

A técnica de *Northern blot* foi desenvolvida na década de 1970, e utiliza uma única sonda para realizar a detecção dos níveis de um único RNA em diversas amostras. Isso ocorre porque o mRNA é imobilizado em um suporte sólido que é incubado com uma solução contendo a sonda correspondente ao transcrito específico de interesse. Em geral, a sonda é marcada com isótopos radioativos ou moléculas fluorescentes, o que possibilita sua detecção. Por meio da medida da quantidade de sonda que se hibrida aos RNA imobilizados, é possível inferir a abundância de moléculas do transcrito específico na amostra. Esse tipo de experimento é muito útil quando se deseja estudar um RNA específico que esteja envolvido em um fenômeno biológico.

Uma grande vantagem dos microarranjos de DNA em relação ao *Northern blot* é que eles possibilitam monitorar ao mesmo tempo os níveis de milhares de moléculas de DNA a um custo relativamente baixo e com muito menos trabalho que a realização de múltiplos experimentos de *Northern blot*. Isso torna tal técnica popular para abordagem de problemas em que se busca propor um novo mecanismo específico subjacente a fenômenos celulares. Em um único microarranjo, é possível imobilizar dezenas de milhares de sondas, não existindo a necessidade de eleger um

número limitado de transcritos para estudo. Isso ocorre porque, com as tecnologias atuais, é possível imobilizar em uma única lâmina o número equivalente de sondas ou superior ao número de transcritos distintos de um organismo, sendo possível representar praticamente todo o transcriptoma. Uma vez sintetizadas as sondas, é possível construir centenas de lâminas que podem ser utilizados em diversos experimentos, possibilitando comparações entre diferentes amostras de modo rápido e prático. Um exemplo de comparação interessante que poderia ser realizada com esse tipo de plataforma é a comparação entre amostras de diversos pacientes em um estudo clínico, ou em um estudo temporal de efeito de uma substância em um conjunto de células. Em função da dinâmica dos transcriptomas, é esperado que cada amostra apresente um perfil diferente das outras, o que pode indicar efeitos de uma substância sobre a expressão de um determinado gene, ou até mesmo diferenças de respostas entre indivíduos (Figura 13.10). Uma dificuldade causada pela grande quantidade de dados gerados pelos experimentos de microarranjos é que o processamento e a interpretação dos dados não são triviais, exigindo testes estatísticos para verificar quais transcritos apresentam, de fato, expressão diferencial. Esse tipo de necessidade evidencia o papel cada vez mais importante que a bioinformática tem para a análise de conjunto de dados biológicos em larga escala.

Detecção e quantificação de RNA mensageiro com a utilização da reação em cadeia da polimerase (PCR)

Outro tipo de técnica utilizada para o estudo do transcriptoma é o chamado PCR em tempo real. Ela se baseia na amplificação de cDNA pela reação em cadeia da polimerase, sendo a amplificação de cada transcrito monitorada por meio de moléculas fluorescentes que passam a emitir maior quantidade de luz à medida que uma maior quantidade de cDNA é amplificada. Como nos estágios iniciais da reação de PCR a amplificação é exponencial, é possível inferir a quantidade de cDNA original a partir do número de ciclos de PCR necessários para a amostra apresentar uma concentração preestabelecida de DNA. Quanto menos ciclos forem necessários para a amostra chegar a essa concentração, maior a concentração inicial de cDNA. Esse tipo de técnica tem a vantagem de exigir uma baixa quantidade de cDNA, pois se baseia em uma técnica de amplificação. Além disso, por utilizar oligonucleotídeos específicos como iniciadores da reação de PCR, a amplificação de cDNA é bastante específica, sendo geralmente amplificado somente o DNA correspondente ao transcrito de interesse. No entanto, pelo fato de podermos realizar somente a análise de um transcrito em cada tubo, apenas um pequeno conjunto de genes pode ser monitorado por experimento.

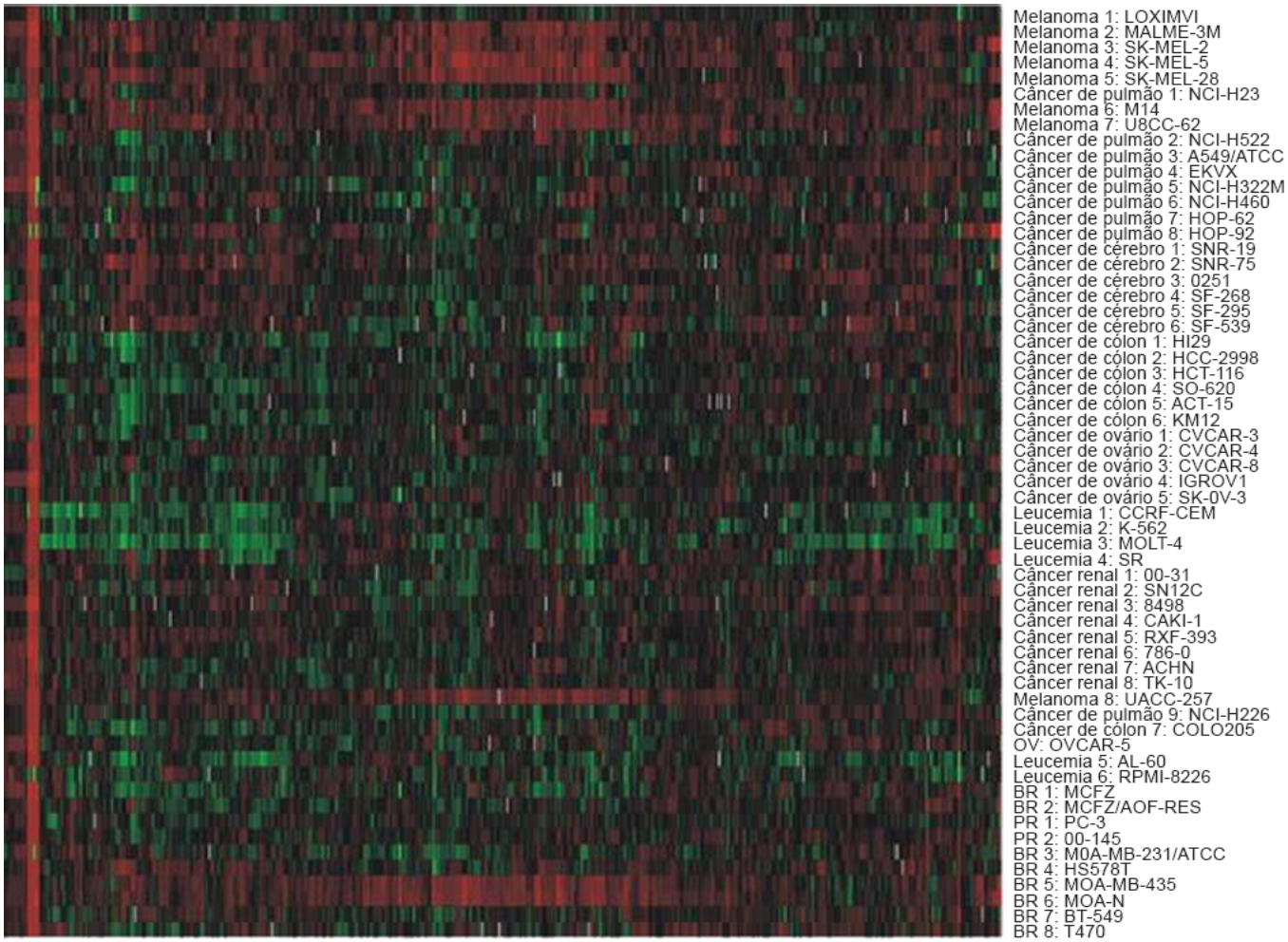


Figura 13.10 Exemplo de visualização de uma análise de dados de microarranjos. No gráfico, cada linha representa os resultados do experimento de um paciente diferente, enquanto cada coluna corresponde ao resultado da medida do nível de expressão de uma sonda específica que corresponde a um transcrito. A coloração indica uma expressão aumentada (vermelho) ou diminuída (verde) para aquele gene com relação à média dos pacientes.

Proteoma | Conjunto de moléculas efetoras da maioria dos processos celulares

As proteínas representam um grupo extremamente versátil de biomoléculas. Um dos motivos é que os 21 aminoácidos que podem compor uma proteína apresentam cadeias laterais com propriedades físico-químicas bastante distintas. Portanto, diferentes combinações de aminoácidos irão gerar proteínas com as mais diferentes conformações e propriedades. Essa característica das proteínas se traduz no fato de que elas são os biopolímeros mais versáteis dos organismos vivos, podendo atuar em diversos contextos celulares, como moléculas estruturais, enzimas, moléculas sinalizadoras, entre outras funções. Além disso, são encontradas nos diferentes ambientes celulares, podendo estar no citoplasma, associadas a membranas, no núcleo, associadas ao DNA, entre outros.

As proteínas são os grandes efetores dos diversos processos celulares e, por isso, o seu estudo é de grande interesse para os pesquisadores. Elas são essenciais para a criação e a manutenção das outras macromoléculas, sendo responsáveis pela replicação do DNA, a transcrição do RNA e a síntese de carboidratos e lipídeos. Assim, a descrição da síntese e degradação dessas diversas biomoléculas passa por uma verificação da abundância e regulação das proteínas que compõem essas rotas metabólicas.

Além disso, são sintetizadas a partir da informação contida na molécula de mRNA, que é traduzido pelo ribossomo das células para a síntese de uma proteína especifica. Cada mRNA que se encontra no citoplasma é traduzido por múltiplos ribossomos que migram em uma direção específica da molécula, transformando a informação do RNA em polímeros de aminoácidos. Continuamente, há a ligação de novos ribossomos à molécula de RNA, fazendo com que cada molécula esteja ligada a múltiplos ribossomos em diferentes estágios da tradução. Considerando esse cenário, quanto maior o número de moléculas de um mRNA codificado para uma proteína específica, maior é o número de ribossomos recrutados para a síntese dessa proteína e, portanto, resultará em sua

maior concentração final (Figura 13.11). Assim, pode-se estimar a abundância de proteínas utilizando a quantidade de mRNA como medida indireta.

No entanto, isso não leva em consideração o fato de que existem mecanismos que podem estimular a síntese proteica a partir de um mRNA, bem como mecanismos específicos de degradação de proteínas, que, certamente, podem alterar a correlação entre abundância de proteínas e correspondentes mRNA. Consequentemente, a medida direta dos níveis das diferentes proteínas em uma célula ou tecido é muito importante para avaliação completa dos sistemas moleculares atuantes.

Ademais, em muitos casos, a atividade de uma proteína pode ser alterada por modificações pós-traducionais. Tais modificações consistem na ligação de grupos químicos à cadeia de proteína após sua síntese, como a glicosilação e a fosforilação de proteínas. A glicosilação é um evento que geralmente ocorre em proteínas secretadas que são direcionadas ao retículo endoplasmático e complexo de Golgi. À proteína são adicionados diversos resíduos de açúcar que podem alterar significativamente seu peso molecular. A fosforilação, por outro lado, é uma modificação bem mais simples, que consiste na adição de um grupo fosfato em uma proteína. Apesar de ser um grupo pequeno e que, a princípio, não deveria alterar muito a estrutura da proteína, fosforilações são utilizadas como sinais para ativar/desativar uma série de proteínas envolvidas em catálise ou sinalização.

As análises do proteoma de uma célula ou tecido devem considerar todas as peculiaridades das proteínas. Logo, não se trata de um mero esforço de detectá-las, mas inclui também identificar compartimentos celulares nos quais se encontram essas proteínas e detectar eventuais alterações pós-traducionais em seus subconjuntos.

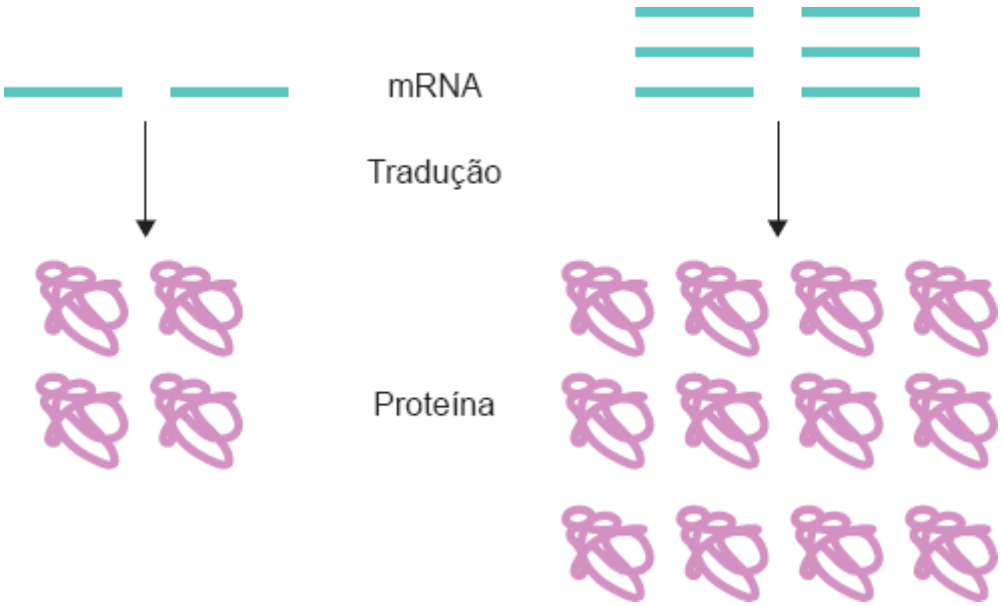


Figura 13.11 Esquema que indica a variação de quantidade de mRNA (barras horizontais) em duas situações diferentes. Em razão desse aumento, ocorre um crescimento proporcional no número de proteínas produzidas a partir da tradução dessas mensagens.

Métodos para estudar o proteoma

Limitações no estudo de proteomas

O estudo do transcriptoma avançou muito nas últimas décadas, em função do avanço nas técnicas de sequenciamento e hibridação. No entanto, é necessário lembrar que a principal função do mRNA é codificar proteínas que serão produzidas com o auxílio do ribossomo. Assim, a presença de determinados níveis de mRNA no citoplasma não acarretaria diretamente em nenhum processo celular específico, e somente a partir do momento em que uma proteína é produzida com base neste mRNA é que, de fato, ocorrerão mudanças no ambiente celular.

Pode-se perguntar, portanto, por que existem tantos estudos buscando quantificar as moléculas de mRNA, para utilizá-las como uma medida indireta de seus níveis, em vez de realizar a medida direta de seus níveis. Isso ocorre porque as moléculas de mRNA são facilmente convertidas em cDNA e podem ser, então, facilmente amplificadas e manipuladas por meio de técnicas de biologia molecular. Isso possibilita que mesmo porções diminutas de mRNA possam ser amplificadas e depois analisadas. O sequenciamento de DNA é uma técnica consistente e facilmente

utilizável em larga escala, viabilizando a análise simultânea de milhares de moléculas de mRNA e a detecção até mesmo das moléculas mais raras.

Em contraste, técnicas de manipulação de proteínas são relativamente limitadas, não existindo, por exemplo, técnicas que possibilitem sua amplificação. Isso ocasiona uma grande dependência da quantidade e qualidade do material biológico a partir dos quais serão extraídas as proteínas. Como consequência, o detalhamento com a qual se consegue estudar proteomas é geralmente menor, muitas vezes restringindo-se às proteínas mais abundantes do sistema estudado, ou a um subconjunto de proteínas selecionadas com base em alguma característica desejada.

Além disso, devido à grande heterogeneidade do comportamento das proteínas, muitas vezes é difícil estabelecer protocolos universais para extração e processamento de proteomas. Isso faz com que seja necessário desenvolver protocolos bastante específicos para o estudo de proteomas de determinados organismos, ou para estudo de certos subconjuntos do proteoma. Tal fato é bastante contrastante com o estudo de transcriptomas, cujo maior desafio é a obtenção de RNA puro e íntegro, sendo os protocolos posteriores praticamente universais. Assim, o estudo de proteomas passa a exigir um preparo muito mais sofisticado dos experimentos, além de dificultar o estabelecimento de rotinas de preparo de amostras que possibilitem iniciativas em larga escala de exploração de proteomas de diversos organismos.

Fracionamento de amostras biológicas

Em razão das limitações da técnica proteômica, em muitas ocasiões é necessário um processo inicial de fracionamento de amostras que possibilite um direcionamento a proteínas de interesse. Uma técnica bastante utilizada é o fracionamento celular por meio de ultracentrifugação. Essa separação se baseia no princípio de que, uma vez que a célula seja rompida e homogeneizada em procedimentos controlados, ela irá liberar suas organelas, que podem então, ser fracionadas. Tal fracionamento é realizado por centrifugação a altas velocidades (que geram forças da ordem de centenas de milhares de vezes à força da gravidade), aproveitando-se das diferentes taxas de sedimentação entre as organelas. Organelas com maior tamanho e densidade tendem a se sedimentar com maior facilidade em comparação àquelas de baixo tamanho e densidade. Assim, partículas maiores, como o núcleo, tendem a se sedimentar rapidamente quando submetidas a centrifugações a velocidades que geram forças de cerca de 1.000 vezes a gravidade, enquanto as outras organelas sedimentarão muito lentamente a essa velocidade e, na prática, continuarão em solução. Por meio de sucessivas centrifugações a diferentes velocidades, é possível separar diversos componentes celulares, viabilizando, assim, o estudo individualizado das proteínas de cada uma dessas frações.

Isolamento de complexos proteicos

Existem outras técnicas que permitem restringir um subconjunto do proteoma para análises. Um exemplo é a técnica de coimunopurificação, que consiste em utilizar um anticorpo produzido contra uma proteína específica e imobilizá-lo em uma coluna ou resina. Anticorpos são proteínas produzidas pelo sistema imune que têm a capacidade de reconhecer de modo bastante específico regiões de proteínas. Uma vez que o anticorpo se liga a uma proteína, forma-se uma interação bastante estável. Quando o anticorpo é incubado com o extrato proteico, ele irá se ligar à proteína-alvo e, por consequência, irá imobilizá-lo também na resina. Juntamente com essa proteína, serão imobilizadas todas as outras que estejam interagindo diretamente ou façam parte do mesmo complexo proteico (Figura 13.12). Uma vez que o complexo é imobilizado na resina, é fácil retirá-lo da solução, separando-o, assim, das outras proteínas do organismo.

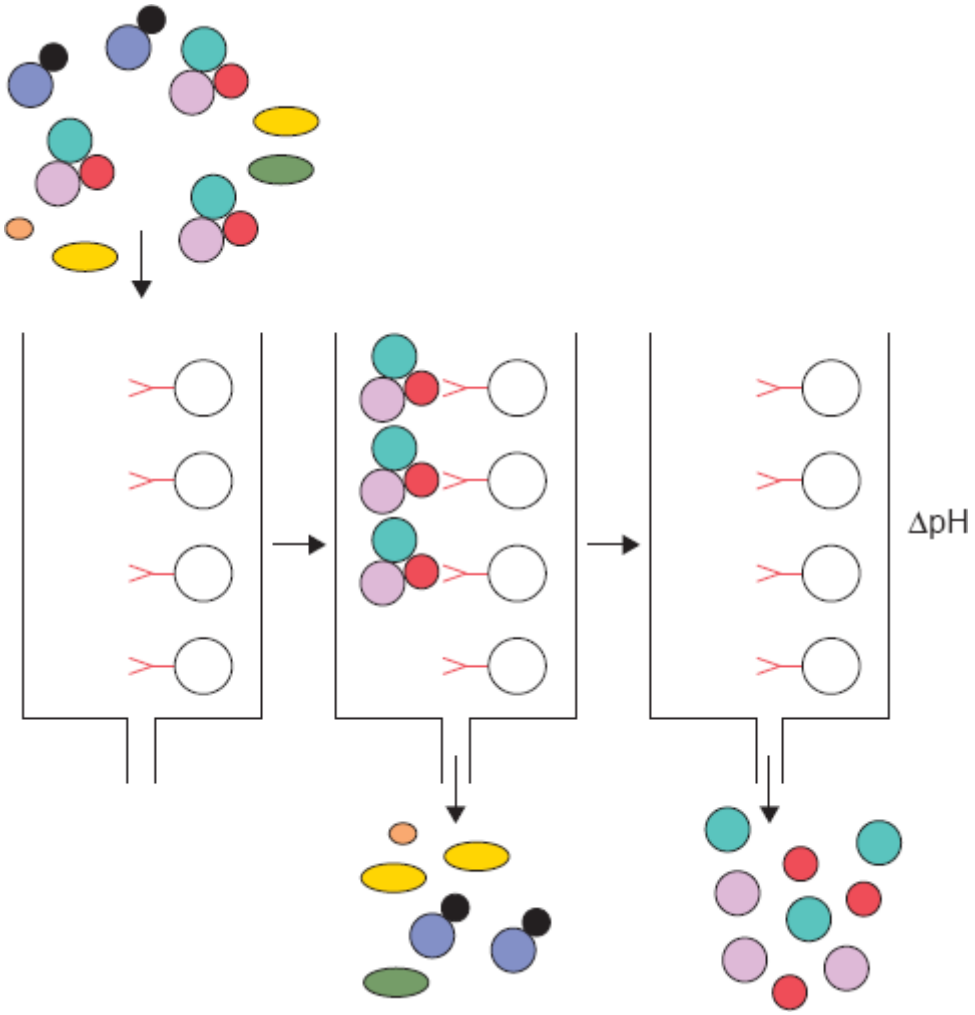


Figura 13.12 Representação esquemática de imunopurificação. Anticorpos reagentes com a proteína representada em vermelho são imobilizados em uma coluna. Após a passagem de uma solução contendo diversas proteínas (representadas por círculos), somente a proteína representada em vermelho se ligará ao anticorpo e será retida na coluna. Outras proteínas interagindo com essa proteína também serão retidas. Posteriormente, uma solução com pH muito baixo ou alto é adicionada para desestabilizar a ligação proteína-anticorpo e possibilitar a recuperação das proteínas do complexo.

Uma técnica alternativa à utilização de anticorpos para estudo de complexos proteicos é a de purificação por afinidade em tandem (TAP), que purifica o complexo proteico a partir de duas sucessivas purificações com base na afinidade de peptídeos que são adicionados à cadeia de uma das proteínas do complexo. Nesse tipo de técnica, é selecionada uma proteína "isca", com a qual se supõe que outras proteínas interajam. Uma cópia do gene codificando para essa isca é preparada contendo, em fusão, duas regiões codificando duas porções derivadas de diferentes proteínas. Em geral, essas duas porções codificam um peptídeo ligante de calmodulina e uma porção ligante de IgG da proteína A de *Staphylococcus aureus*. A calmodulina é uma proteína que se liga a outras em um processo dependente de íons cálcio. A proteína A tem a capacidade de se ligar à região invariável de anticorpos do tipo IgG. Em ambos os casos, a interação dessas proteínas com seus ligantes é bastante específica e, por isso, elas são utilizadas para a purificação a partir de afinidade. Entre as duas porções codificantes para esses peptídeos existe ainda um sítio de clivagem da TEV protease (Figura 13.13 A).

Assim, nesses estudos, o gene em análise é adicionado ao organismo de interesse, o qual passará a expressar uma cópia adicional da proteína do complexo proteico. A proteína codificada por essa nova cópia terá uma porção adicional contendo os peptídeos em fusão. Após a expressão desse gene, a proteína resultante passará a formar o complexo com as outras proteínas, do mesmo modo que a proteína nativa. A purificação dos complexos proteicos é realizada pela passagem do produto de lise celular em uma coluna contendo IgG imobilizado, na qual a porção da proteína A irá se ligar e, com isso, imobilizar o complexo proteico na coluna. Após a lavagem para retirada das proteínas que não interagem com a coluna, é realizada uma digestão com a TEV protease, que separará a porção contendo a porção da proteína A do restante da proteína, e fazendo com que o complexo proteico se desligue da coluna. Um segundo passo de purificação é realizado utilizando uma coluna contendo calmodulina imobilizada. Na

presença de cálcio, ocorre a ligação entre a calmodulina imobilizada e o peptídeo ligante de calmodulina existente na proteína do complexo. Isso possibilita um segundo passo que garantirá melhor purificação do complexo, o qual, após a lavagem, pode ser desligado da coluna com a retirada do cálcio tampão (Figura 13.13 B). A realização dessas duas purificações permite que o produto final seja o complexo proteico com alto grau de pureza, o que viabiliza análises posteriores para identificação de suas proteínas componentes. Existem casos em que os fragmentos de proteína A e peptídeo ligador de calmodulina são substituídos por outras proteínas que apresentam afinidade por outros compostos, mas o princípio de purificação continua o mesmo.

Técnicas de separação de proteínas e peptídeos

Diversas estratégias empregadas para separação de proteínas e peptídeos em experimentos de proteômica utilizam técnicas eletroforéticas ou cromatográficas. É necessário notar que, ao trabalhar com identificação de proteínas, em algum ponto do processo deve ocorrer a sua digestão com o auxílio de proteases para a formação de peptídeos. Isso é necessário porque a análise de peptídeos por meio da técnica de espectrometria de massa é mais informativa que a análise de proteínas inteiras. Portanto, dentro de uma estratégia para caracterização de proteomas, podemos encontrar, muitas vezes, passos de separação de proteínas, seguidas de uma proteólise e passos seguintes de separação e purificação de peptídeos.

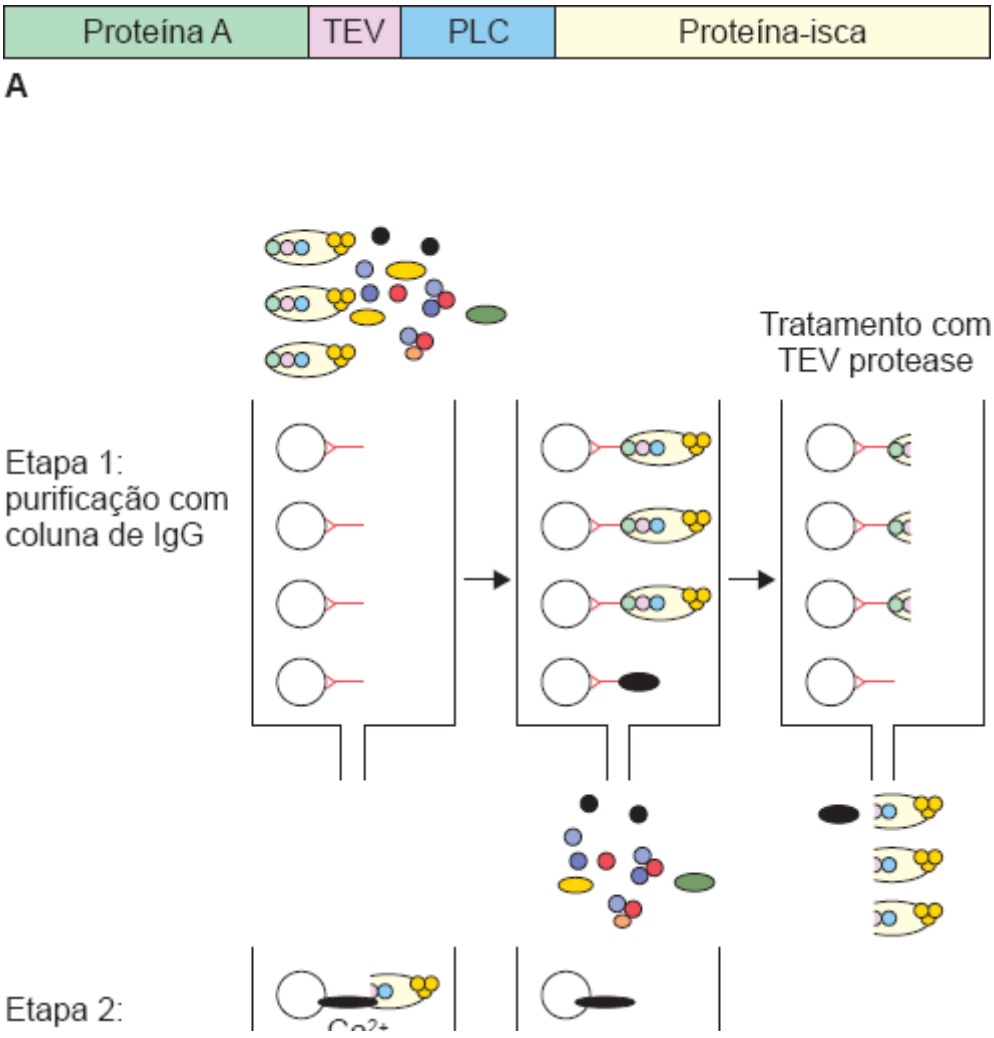


Figura 13.13 Representação esquemática de uma purificação por afinidade em tandem (TAP). **A.** Representação das diferentes regiões da proteína de fusão incluindo regiões codificando para proteína A, sítio de clivagem da TEV protease (TEV) e peptídeo ligante de calmodulina (PLC). **B.** Representação dos diferentes passos de purificação do complexo de interação envolvendo a proteína-isca.

A separação de proteínas geralmente é realizada por meio de técnicas eletroforéticas. Um tipo de técnica clássica para separação de proteínas é através de eletroforese de géis de poliacrilamida contendo o detergente sódio dodecilsulfato (SDS). Esse tipo de técnica possibilita separar proteínas com base em seu peso molecular, pois o SDS é um detergente aniônico que irá se intercalar a uma proteína linearizando-a e dando uma carga negativa

relativamente uniforme a todas as proteínas. Assim, todas as moléculas de proteínas passam a ter aproximadamente a mesma carga e formato, e, quando submetidas a um campo elétrico, o principal parâmetro que irá determinar a sua velocidade de migração é a sua massa molecular (Figura 13.14). Apesar de ser uma técnica bastante consistente, esse tipo de separação não é muito eficiente quando temos uma mistura complexa de proteínas. Isso ocorre porque, nesse tipo de mistura, são esperadas várias proteínas com massa molecular semelhante e que irão migrar até praticamente a mesma região do gel.

Outra técnica interessante para a separação de proteínas é a chamada focalização isoelétrica. As cadeias laterais das proteínas apresentam grupos capazes de serem ionizados, e a carga desses grupos irá depender do pH onde eles se encontram. A carga de uma proteína em certo pH será a somatória das cargas apresentadas pelas suas cadeias laterais. Esse tipo de técnica aproveita-se das diferentes cargas apresentadas por proteínas em um pH, e consiste em submeter as proteínas a um campo elétrico em um gradiente imobilizado de pH. Quando realizada uma focalização, proteínas com uma carga total negativa serão atraídas para o ânodo; no entanto, à medida que migram, elas irão para regiões com menor pH e, portanto, tenderão a adquirir prótons que irão neutralizar as suas cargas negativas. A migração continua até o ponto em que a protonação da proteína neutraliza as cargas negativas e a proteína passa a ser neutra, não respondendo mais ao campo elétrico. O oposto ocorre com uma proteína com carga positiva, que irá migrar para o cátodo até que seja suficientemente desprotonada para se tornar neutra; a região do gradiente de pH em que cada proteína irá se concentrar é correspondente ao seu ponto isoelétrico (pI).

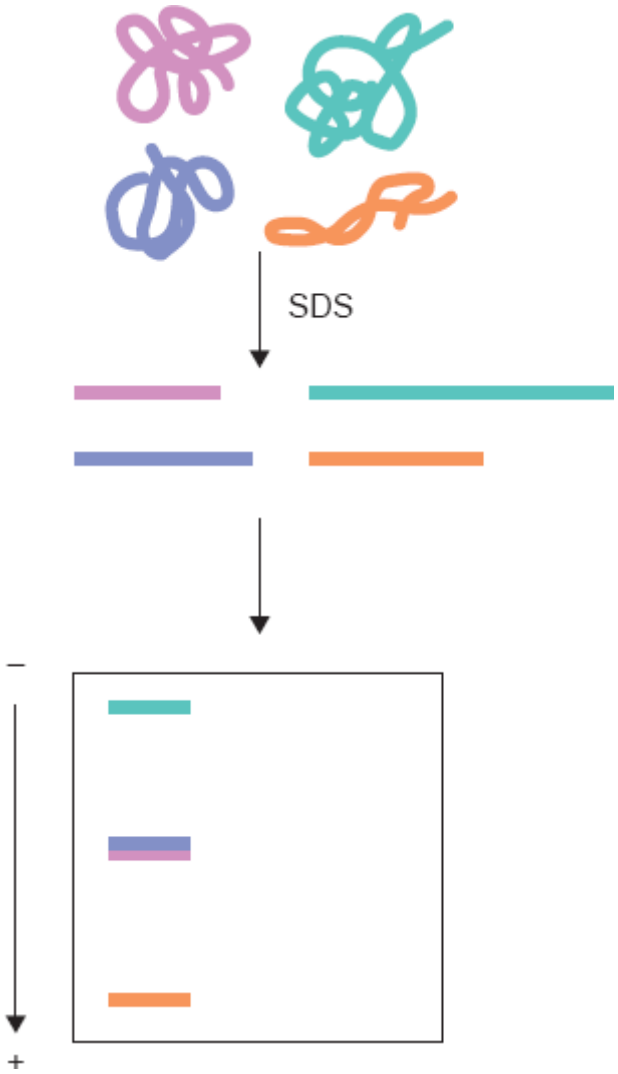


Figura 13.14 Representação esquemática de separação de proteínas com gel de poliacrilamida com SDS. Primeiramente, as proteínas enoveladas são linearizadas com a exposição ao SDS e, depois, são submetidas à eletroforese em gel de acrilamida. Durante a separação, as proteínas migrarão do polo positivo para o negativo, e as proteínas de menor peso molecular migrarão mais rapidamente. Nota-se que as proteínas roxas e azuis têm peso molecular semelhante e não são bem separadas por esse tipo de técnica.

A combinação das metodologias de isoeletrofocalização e eletroforese em gel de acrilamida com SDS define a técnica de eletroforese em bidimensional. Tal técnica foi desenvolvida na década de 1970 pelo pesquisador Patrick

O’Farrell, cujo objetivo era estabelecer um sistema de separação que caracterizasse mudanças no padrão de expressão de proteínas causada por mutações que afetavam o desenvolvimento de algas do gênero *Volvox*. Em primeiro lugar, as proteínas são separadas por isoeletrofocalização em uma fita contendo um gradiente de pH e, posteriormente, essa fita é colocada no topo de um gel de poliacrilamida contendo SDS e submetida à eletroforese (Figura 13.15). Isso favorece que as proteínas se distribuam nesse gel de acordo com o seu pI no eixo horizontal e pelo seu peso molecular no eixo vertical. Como resultado, elas serão representadas por pontos no gel após a sua coloração com reagentes específicos para detecção de proteínas, sendo o posicionamento desses pontos reflexo de tais aspectos (Figura 13.15).

As proteínas nativas são submetidas a um processo de isoeletrofocalização em um gradiente imobilizado de pH. As diferentes proteínas irão se concentrar em faixas de pH correspondentes ao seu pI. Após a focalização isoelétrica, o suporte sólido contendo as proteínas é colocado no topo de um gel de poliacrilamida com SDS, e as proteínas são desenoveladas e passam a migrar de acordo com o seu peso molecular. No gel final, elas serão separadas pelo seu ponto isoelétrico no eixo horizontal e pelo seu peso molecular no eixo vertical. Proteínas com pontos isoelétricos semelhantes (vermelha e amarela) não são bem separadas na primeira etapa; no entanto, pela diferença de peso molecular, acabam sendo separadas na segunda. As proteínas rosa e azul que não eram bem separadas na Figura 13.14 também passam a ter uma boa separação.

Aquelas separadas pela técnica de eletroforese bi-dimensional podem então ser individualmente analisadas para estabelecer as suas identidades. Com essa finalidade, as regiões correspondentes a cada ponto do gel são recortadas e passam por um processo que irá resultar na digestão de proteína em peptídeos. Essa digestão é realizada por proteases que contêm regiões específicas de corte, o que possibilita prever os peptídeos que serão formados a partir da digestão de cada proteína do organismo. A protease mais utilizada em experimentos de proteômica é a tripsina, que realiza cortes nas ligações peptídicas do grupo carboxila dos aminoácidos lisina ou arginina com o grupamento amino de qualquer outro aminoácido. Essa digestão é realizada nos próprios fragmentos de gel recortados, que são desidratados e depois reidratados com uma solução contendo a protease. Uma vez que a digestão é completada, os peptídeos formados não são mais retidos pela malha do gel, devido ao seu pequeno tamanho, e se difundem na solução na qual o gel se encontra mergulhado.

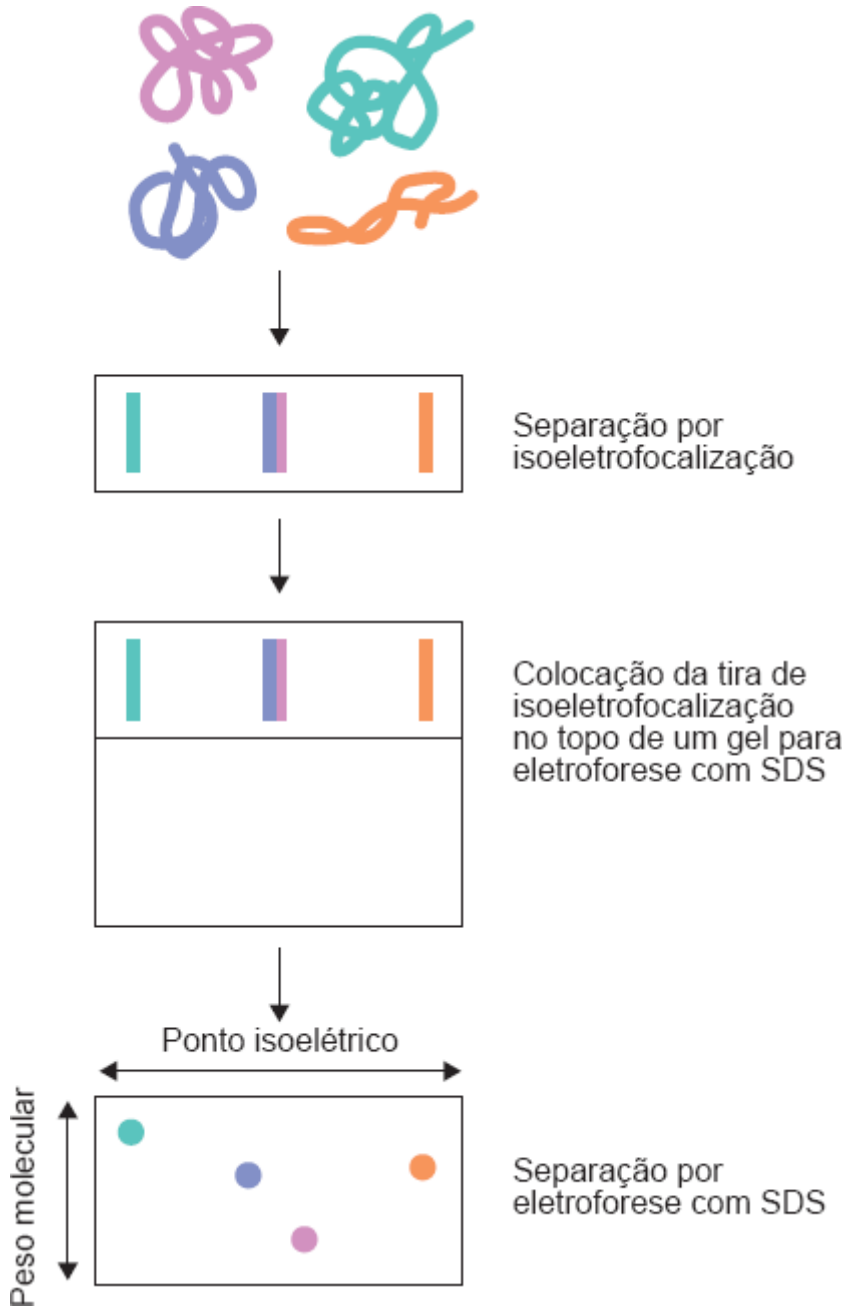


Figura 13.15 Representação esquemática da separação de proteínas por meio da técnica de eletroforese bidimensional.

Além disso, todas as ligações dissulfeto realizadas pela proteína por meio de seus resíduos de cisteína são reduzidas, e grupos químicos são ligados aos grupos SH livres formados para impedir o restabelecimento das ligações dissulfeto. Esse tipo de tratamento é realizado porque, após a tripsinização, os peptídeos formados que contenham pontes dissulfeto entre si estarão ligados covalentemente. Logo, para posterior análise e identificação desses peptídeos, seria necessário prever essas ligações por pontes dissulfeto, o que não é trivial, complicando a análise dos peptídeos.

Em adição aos métodos de separação de proteínas por eletroforese bidimensional, é possível realizar estudos de proteômica através de métodos de separação cromatográfica. Esse tipo de abordagem geralmente tem enfoque na separação de peptídeos após a digestão de proteínas. Inicialmente, uma solução contendo proteínas é digerida com tripsina ou outra protease, e será formada uma mistura complexa contendo peptídeos provenientes de todas as proteínas que se encontravam em solução. Essa mistura de peptídeos é então separada a partir de técnicas cromatográficas, com o objetivo de criar frações com uma complexidade menor que a mistura inicial (Figura 13.16). Proteínas de uma mistura são digeridas com tripsina (pontos de digestão indicado por setas). Os diferentes peptídeos resultantes são representados por barras com coloração idêntica à proteína de origem e numeradas para possibilitar a distinção entre os peptídeos gerados. Uma separação desses peptídeos em uma coluna cromatográfica é representada

com peptídeos separados em diferentes frações, de acordo com suas propriedades de interação com a coluna e o solvente, e alguns peptídeos ainda retidos na coluna neste momento do processo.

No caso de amostras bastante complexas, como a fração total de proteínas solúveis de uma célula, mais de um tipo de separação cromatográfica é normalmente utilizado no processo. A abordagem mais comum é realizar uma primeira separação em uma coluna de troca catiônica, onde os peptídeos serão separados com base em sua afinidade por uma coluna negativamente carregada, e as frações formadas a partir dessa coluna serão, em seguida, separadas por uma coluna hidrofóbica, que irá separar peptídeos com base na sua afinidade por uma coluna que tem uma cadeia hidrocarbônica alifática. Cada fração da coluna original irá ser subfracionada, criando um grande número de amostras, cada uma contendo um número limitado de peptídeos que apresentam características semelhantes em termos de interação com grupos carregados negativamente e com grupos hidrofóbicos. A divisão da amostra em várias frações é importante, pois, ao analisar uma mistura complexa em um espectrômetro de massa, somente um número limitado de peptídeos será detectado por amostra, já que eles competem no passo inicial de ionização. Ao subdividir-se a amostra, ocorre a detecção de um número muito maior de peptídeos, pois cada fração será menos complexa e, portanto, o número de peptídeos competindo pela ionização será menor.

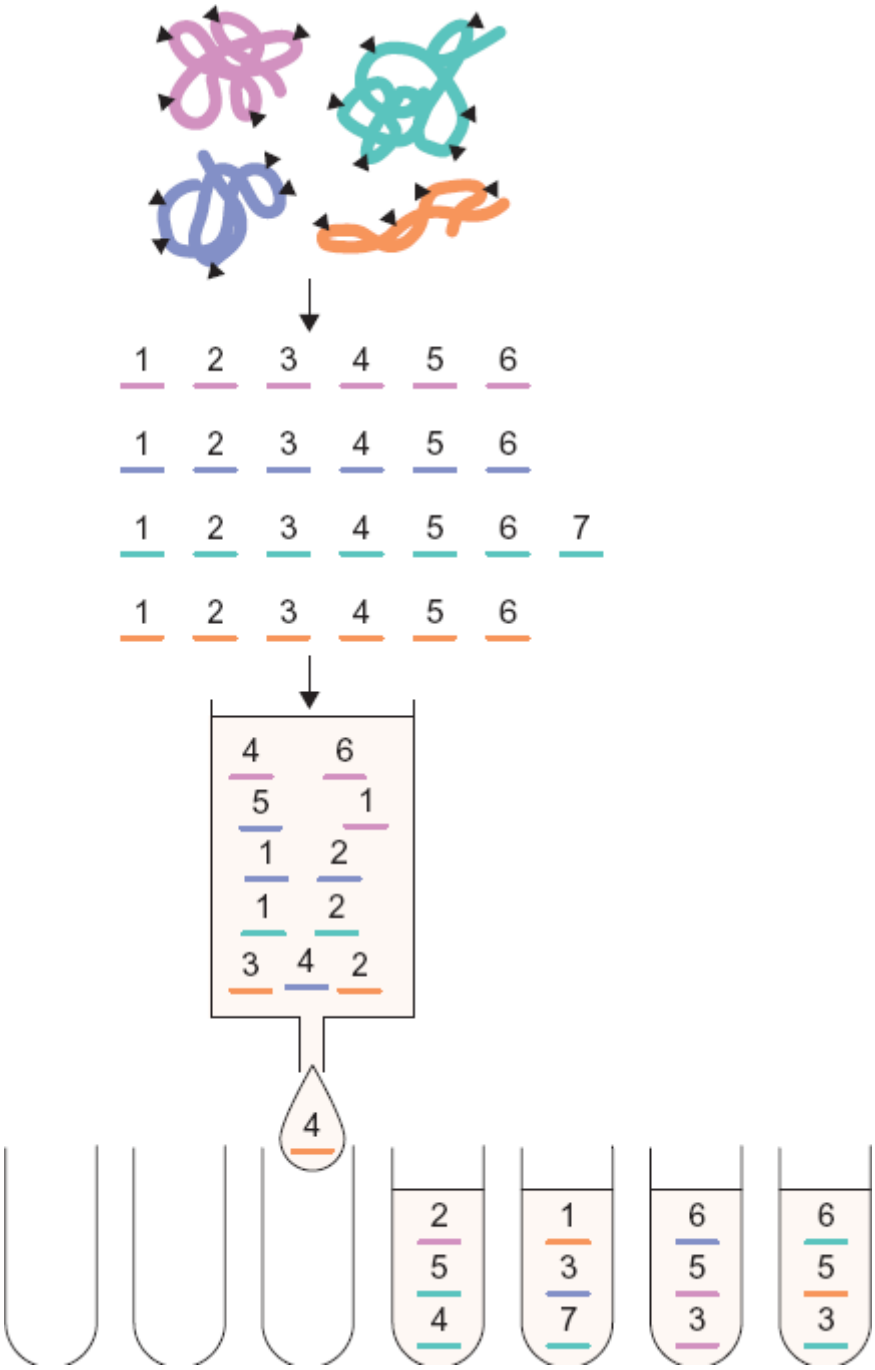


Figura 13.16 Esquema de separação de peptídeos por meio de técnicas cromatográficas.

Em vários casos, utiliza-se uma combinação de técnicas eletroforéticas e cromatográficas, a fim de obter melhor separação de peptídeos para análise com espectrometria de massa. Um exemplo é a metodologia chamada Gel-LC, que se inicia com uma separação de proteínas por meio de eletroforese em gel de poliacrilamida com SDS, que irá separar as proteínas por peso molecular. Após isso, o gel de poliacrilamida é recortado, separando-se as diversas faixas de tamanho, e cada um dos recortes é processado separadamente para obtenção dos fragmentos trípticos das proteínas contidas em cada fração. Cada uma das soluções contendo os peptídeos resultantes é submetida a um processo cromatográfico que irá separar os peptídeos componentes em frações que serão analisadas por espectrometria de massa.

Utilização de espectrometria de massa para estudo de proteomas

A análise de peptídeos em experimentos de proteômica é realizada utilizando a técnica de espectrometria de massa (ver o boxe "Desenvolvimento da espectrometria de massa"). Esse tipo de técnica possibilita medidas muito precisas da massa molecular de moléculas e pode ser utilizada para determinar a composição de polímeros biológicos, incluindo peptídeos e proteínas. Trata-se de uma técnica com alta sensibilidade, o que é muito importante quando trabalha-se com moléculas biológicas, pois a quantidade de uma proteína extraída de material biológico costuma ser limitada.

Desenvolvimento da espectrometria de massa

Espectrômetros de massa são aparelhos que utilizam campos eletromagnéticos para determinar de modo extremamente preciso a massa de íons produzidos a partir da ionização de uma amostra. É reconhecido que o desenvolvimento desse tipo de aparelho iniciou-se com os experimentos de J. J. Thomson, no início do século 20, que realizou a deflexão de gases ionizados em câmeras de vácuo, utilizando campos eletromagnéticos, e demonstrou que essa deflexão dependia da relação massa/carga do íon, estabelecendo, assim, os fundamentos para o desenvolvimento desse tipo de aparelho. Durante as três primeiras décadas do século 20, Francis Ashton (que havia sido aluno de Thomson) e outros cientistas aperfeiçoaram esse aparato inicial, melhorando a sua resolução, e forneceram evidência da existência de isótopos para uma grande parte dos elementos químicos. A partir da década de 1940, os espectrômetros de massa passaram a se popularizar, e aparelhos comerciais ficaram disponíveis e foram utilizados não somente por físicos, mas por químicos interessados em determinar a composição química de amostras. Somente na década de 1980, com o desenvolvimento de técnicas suaves de ionização, foi possível utilizar a espectrometria de massa para estudo de biomoléculas.

Pelo menos cinco cientistas que desenvolveram técnicas críticas para o desenvolvimento e aperfeiçoamento da espectrometria de massa foram ganhadores do Prêmio Nobel: Joseph John Thomson, Prêmio Nobel de Física de 1906; Francis William Aston, Prêmio Nobel de Química de 1922; Wolfgang Paul, Prêmio Nobel de Física de 1989; John Bennet Fenn e Koichi Tanaka, Prêmio Nobel de Química de 2002.

O primeiro passo necessário para a análise de peptídeos é a sua ionização. Isso é necessário porque o espectrômetro de massa trabalha com campos elétricos e, portanto, moléculas sem carga são invisíveis para esses aparelhos. Além disso, as moléculas devem passar a fase gasosa para que possam ser analisadas pelo aparelho. Existem diversas técnicas para ionizar moléculas, mas as mais utilizadas para análise de peptídeos são o *electrospray* e MALDI (ionização/dessorção de matriz assistida por *laser*), que, por serem técnicas suaves de ionização, preservam a estrutura de biomoléculas.

A técnica de ionização por *eletrospray* se baseia na passagem de líquido por uma agulha que esteja carregada eletricamente. Isso produz uma série de gotículas com cargas. Tais gotículas passarão por uma região com gás quente que levará à evaporação do solvente e, como resultado, teremos os peptídeos de interesse ionizados e em fase gasosa. Já a técnica de ionização por MALDI consiste na mistura da substância a ser analisada a uma matriz que tem a propriedade de absorver energia luminosa. Quando um *laser* é aplicado a essa matriz, ela rapidamente absorve energia, ionizando a amostra e ejetando-a para a fase gasosa.

As amostras de peptídeos ionizadas serão objeto de análise por aparatos que possibilitarão medir a sua relação de massa molecular/carga com grande exatidão. Existem diversos tipos de analisadores que farão essa medida, cada um com suas características específicas. Independentemente do analisador, o resultado final obtido será um gráfico contendo, no eixo X, as diferentes relações massa molecular/carga e, no eixo Y, a intensidade de detecção (Figura 13.17). Há uma série de metodologias que possibilitam que a carga do peptídeo detectado seja determinada, sendo possível converter esse gráfico em um que represente apenas as massas moleculares dos peptídeos detectados em seu eixo X.

Esse tipo de dado pode ajudar na identificação de uma proteína, pois cada pico irá representar um peptídeo diferente. Se a amostra injetada contiver peptídeos de uma única proteína (como no caso de um ponto recortado de um gel resultante de uma eletroforese bidimensional), pode-se considerar que esse espectro é uma espécie de impressão digital da proteína. A partir dos bancos de sequências de proteínas, são realizadas digestões virtuais com a enzima proteolítica utilizada no experimento, o que torna possível predizer os peptídeos formados por cada proteína. São então deduzidos os espectros teóricos para cada proteína do banco de dados, e estes são comparados com o espectro experimental, para determinação do espectro com maior taxa de coincidência (Figura 13.18).

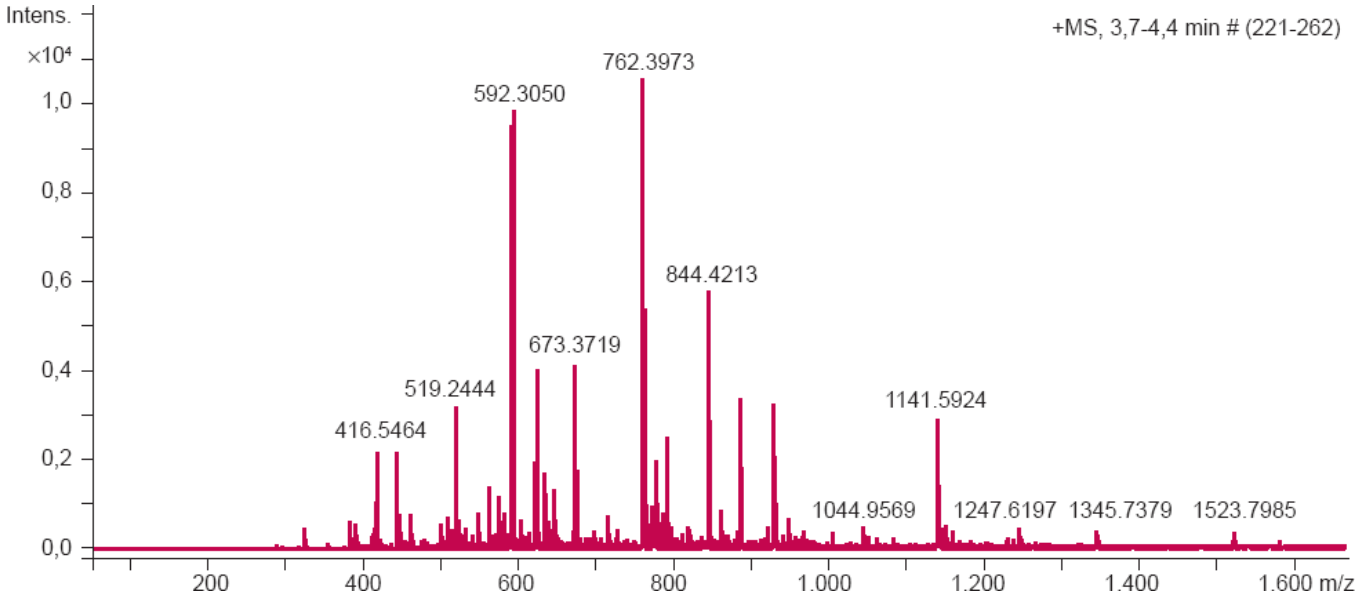


Figura 13.17 Espectro de massa de uma amostra contendo peptídeos. O eixo x indica a relação massa/carga das substâncias detectadas (m/z); o eixo y indica a intensidade de cada pico. Os números acima dos picos indicam a medida exata de m/z de alguns dos picos detectados.

Diversas sequências de um banco de dados são analisadas, possibilitando a construção de um espectro de massa teórico para cada um deles. A comparação entre o espectro experimental e os teóricos (linha pontilhada) torna possível verificar coincidências e, a partir delas, encontrar o espectro que forneça uma identificação confiável.

Não se espera que haja 100% de coincidência entre os picos do espectro teórico e o experimental de uma proteína, pois, durante o passo de ionização, nem todos os peptídeos são efetivamente ionizados e, portanto, alguns serão perdidos durante o processo. Logo, o processo de comparação entre espectros teóricos e experimentais deve ser auxiliado por uma análise estatística que indique o grau de confiança da identificação de uma proteína com base no número de picos coincidentes entre os dois espectros.

Esse tipo de abordagem apresenta algumas limitações, visto que, dependendo do número de picos de peptídeos detectados e do tamanho do banco de dados de proteínas utilizado, podem ocorrer diversas coincidências entre espectros teóricos de proteínas não relacionadas e o espectro experimental, de modo que o resultado não será confiável. Além disso, é necessário que cada amostra injetada no espectrômetro de massa contenha apenas peptídeos derivados de uma única proteína, o que exclui certas abordagens em larga escala que realizam a clivagem de várias proteínas simultaneamente e posterior separação dos peptídeos resultantes.

Devido a essas limitações, em geral, a análise de mistura de peptídeos produzidos a partir da digestão de múltiplas proteínas precisa ser avaliada por uma metodologia que forneça mais detalhes do que a simples espectrometria de massa. Isso é alcançado com a utilização da espectrometria de massa em tandem, também referenciada como MS/MS. Nessa técnica, cada uma das moléculas representadas pelos picos gerados pela espectrometria de massa tradicional passa por um processo adicional de fragmentação. Os fragmentos gerados também terão a sua relação massa/carga medida pelo espectrômetro, formando um espectro de fragmentação para cada um dos picos do espectro original.

Em geral, a fragmentação dos peptídeos no processo de MS/MS é obtida pelo choque destes com um gás inerte, geralmente o argônio. Nessas condições, os peptídeos tendem a sofrer quebras principalmente em suas ligações peptídicas (Figura 13.19). É possível selecionar um único peptídeo de uma mistura para análise de fragmentação por vez, pois os aparelhos que realizam espectrometria de massa em tandem têm uma espécie de filtro de massas

(chamado de quadrupolo), que permite selecionar íons dentro de uma faixa estreita de massa molecular/carga. Os fragmentos resultantes passam então para o analisador do aparelho, que irá registrar o espectro de massa correspondente aos novos fragmentos gerados. Na prática, isso possibilita que, ao injetar no aparelho uma mistura de peptídeos, ele seja capaz de separar cada um deles e gerar um espectro de fragmentação individual.

Nesse tipo de experimento, além da medida da massa de cada peptídeo, será obtido um registro dos fragmentos gerados a partir de quebras na sua cadeia principal. Assim, espera-se que seja formada uma série de fragmentos que tenham diferenças de massa de um aminoácido (Figura 13.20).

A medida da massa exata de cada um dos fragmentos permite determinar sequencialmente cada um dos aminoácidos de uma cadeia polipeptídica. Os únicos aminoácidos para os quais não é possível realizar uma distinção por espectrometria de massa são a leucina e isoleucina, que apresentam massas idênticas entre si. No entanto, para realizar esse tipo de sequenciamento, é necessário um espectro de fragmentação de ótima qualidade, que garanta a definição de pelo menos um fragmento para cada ligação peptídica existente no peptídeo. Infelizmente, a obtenção de espectros com a qualidade necessária não é comum, pois nem sempre os fragmentos representando a quebra de cada ligação peptídica estão presentes no espectro produzido; além disso, picos representando quebras em ligações em regiões diferentes da molécula podem complicar a análise do espectro. Assim, não é possível realizar o sequenciamento de peptídeos de modo sistemático, como acontece para ácidos nucleicos. Isso impede um sequenciamento em larga escala do proteoma com base nos métodos atuais de espectrometria de massa.

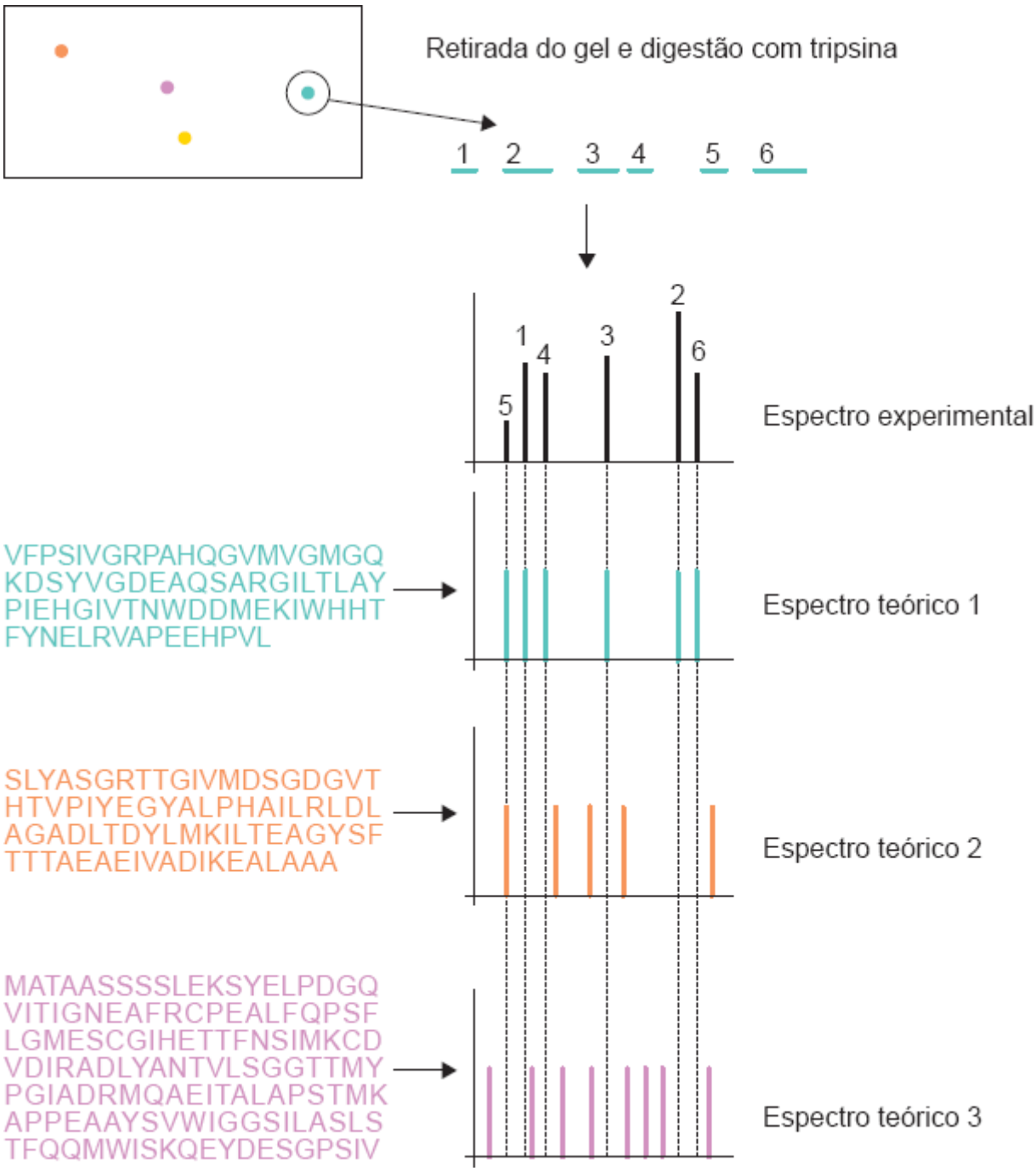


Figura 13.18 Identificação de proteínas com a utilização de espectrometria de massa. O retângulo representa um gel proveniente de uma eletroforese bidimensional de proteínas. Os pontos representando as diferentes proteínas são exibidos em diferentes cores. O ponto representando a proteína verde é retirado para digestão, resultando em seis diferentes peptídeos que, ao serem analisados por um espectrômetro de massa, resultam em um espectro com sua relação massa/carga (eixo x) e intensidade de sinal (eixo y).

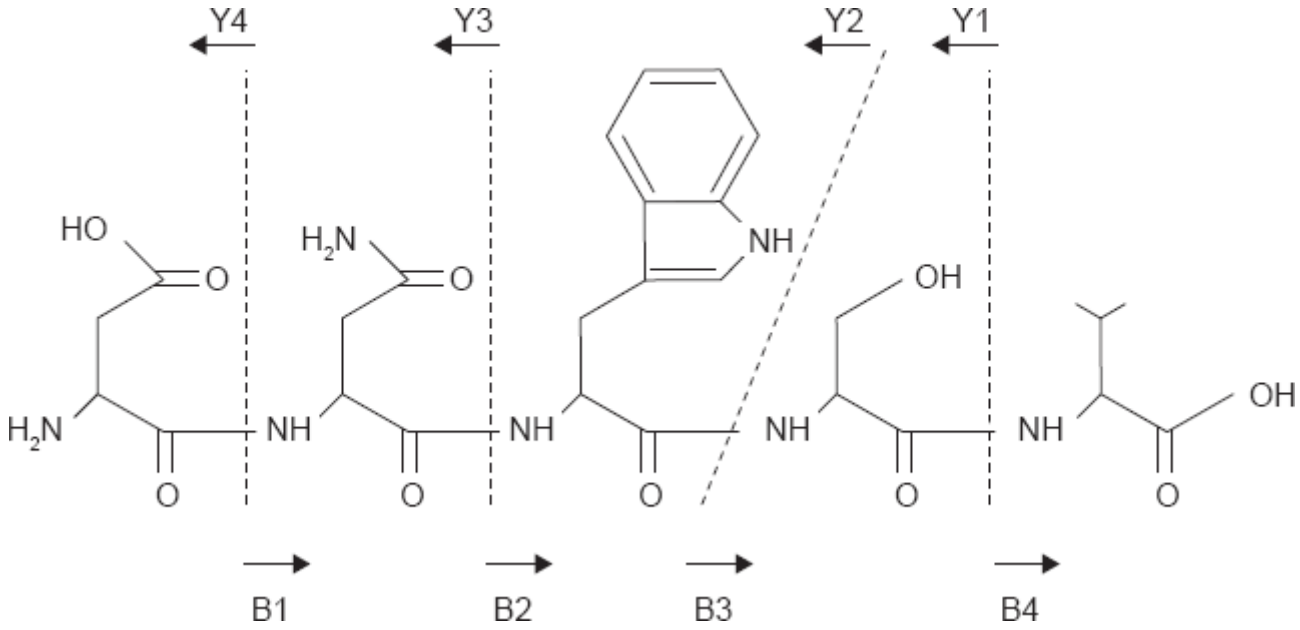


Figura 13.19 Pontos de fragmentação de um peptídeo que geram fragmentos da série Y e B, normalmente detectados em espectrometria de massa em tandem.

Apesar dessa impossibilidade, é comum realizar identificações em larga escala de proteínas por meio da comparação dos espectros de fragmentação gerados para cada peptídeo com os espectros de fragmentação teóricos calculados a partir de sequências de proteínas em um banco de dados. Essas sequências de proteínas costumam ser deduzidas a partir do quadro de leitura aberto de sequências de RNA ou DNA, permitindo que tenhamos grandes bancos de sequências para diversos organismos. Com esse banco de proteínas é realizada uma simulação computacional do processo de digestão por tripsina, resultando em um banco de peptídeos associados às proteínas existentes no banco. Tais peptídeos terão sua fragmentação simulada computacionalmente, resultando em espectros teóricos de fragmentação para cada um deles.

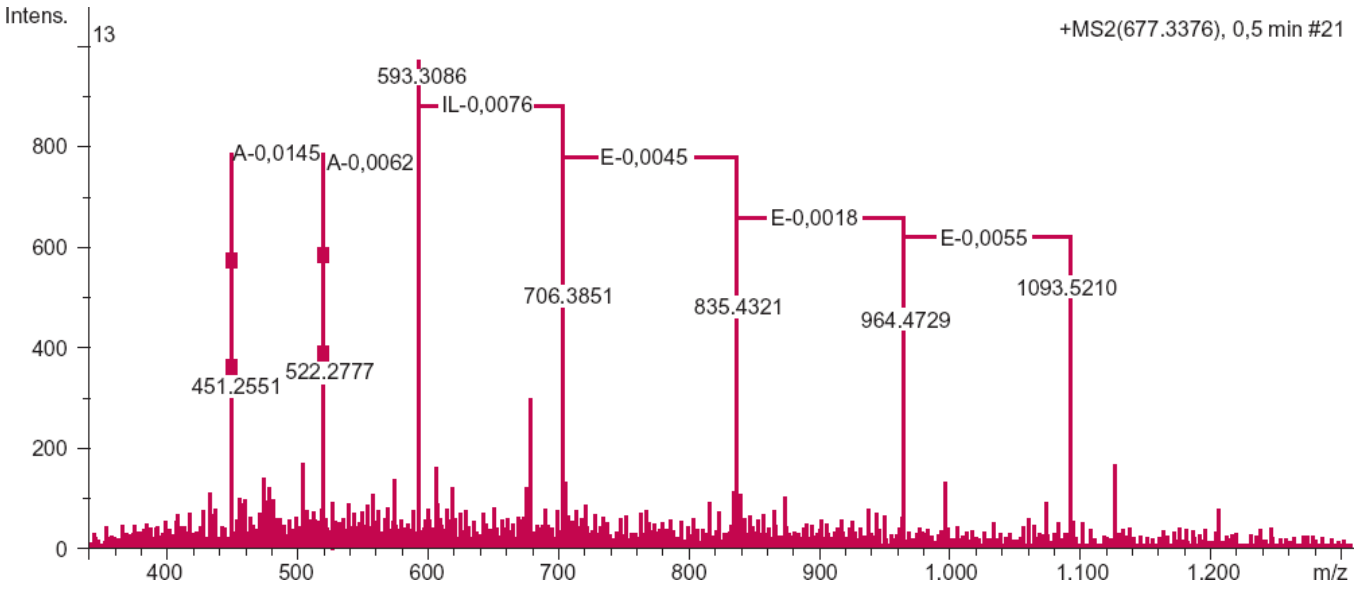


Figura 13.20 Espectro de massa da fragmentação de um peptídeo pela qual é possível determinar a sequência de aminoácidos a partir da diferença de massa entre os picos. Junto à linha, ligando os picos, existe uma letra representando um aminoácido: E: ácido glutâmico; I/L: isoleucina/leucina; A: alanina. Os números ao lado de cada letra indicam o desvio entre a massa experimental do aminoácido e a massa teórica. Nota-se que esse desvio representa cerca de 1/10.000 da massa do peptídeo, indicando que as medidas são extremamente precisas.

A análise comparativa dos espectros de fragmentação experimentais e teóricos é muito semelhante à análise de impressão digital de uma proteína com base no espectro de massa de seus peptídeos, que foi explicada anteriormente. A diferença é que, neste caso, se buscam coincidências entre espectros de fragmentação experimental e teóricos de um único peptídeo, e não mais do espectro de um conjunto de peptídeos derivados de uma mesma

proteína. Assim, cada espectro de fragmentação individual possibilitará a identificação de um peptídeo derivado de uma proteína, viabilizando, assim, identificações de proteínas em misturas complexas, sem que haja necessidade de uma separação prévia de cada proteína (Figura 13.21).

Estudos de interações proteicas e técnica de duplo híbrido

Outra promissora técnica para o estudo de interações do proteoma é a de duplo híbrido. Tal método utiliza um sistema de expressão em levedura contendo um domínio de ligação a DNA, retirado de um ativador transcricional fundido a uma proteína de outro organismo escolhida pelo pesquisador, e para qual se deseja verificar interações (chamada de isca). Essa construção é introduzida em uma célula de levedura, juntamente com diversos clones de uma biblioteca de genes do organismo em estudo, que se encontram fundidos a um domínio de ativação transcricional da levedura (chamados de presa). Quando a isca encontra-se na célula com a presa, caso ocorra interação entre as duas proteínas, haverá um efeito de aproximação entre o domínio de ligação de DNA e o domínio ativador da transcrição, fazendo com que aconteça uma ativação da transcrição de um gene repórter, que fará com que a levedura na qual ocorreu essa interação seja distinguível das outras leveduras (Figura 13.22). Um exemplo de gene repórter utilizado é o LacZ, que codifica para uma proteína metabolizadora de açúcares, a qual, quando na presença de um substrato chamado X-gal, produz coloração azulada.

Uma vez detectada uma levedura onde ocorre interação, o clone derivado da biblioteca de presas é sequenciado para determinação de sua identidade. Deve-se notar que a técnica de duplo híbrido permite uma sondagem inicial de possíveis interações, mas é preciso realizar experimentos mais específicos para confirmação da interação detectada.

Estudo do transcriptoma e do proteoma | Entendimento da biologia celular

Dinâmica do transcriptoma e complexidade em organismos

Durante o sequenciamento do genoma humano, um dos desafios que instigava os cientistas era a definição de quantos genes existiriam neste genoma. Tal era a expectativa que, no ano 2000, vários respeitados profissionais realizaram uma aposta para adivinhar o número de genes que iria surgir a partir do sequenciamento completo do genoma humano, a ser publicado dentro de 1 ano. As projeções para o número de genes humanos foram de 27.000 até 153.000 genes. Tais números eram justificados por algumas análises preliminares, além do fato de genomas de organismos bem mais simples – como a mosca *Drosophila melanogaster* e o verme *Caenorhabditis elegans* – apresentarem aproximadamente 14.000 e 19.000 genes, respectivamente. Portanto, devido à grande complexidade do ser humana comparada a esses organismos mais simples, era natural imaginar que ele teria pelo menos o dobro do número de genes. Não foi sem grande surpresa que, após o término do genoma humano, foi anunciado que o ser humano tinha apenas cerca de 30.000 genes, sendo esse número revisto posteriormente para apenas 24.000.

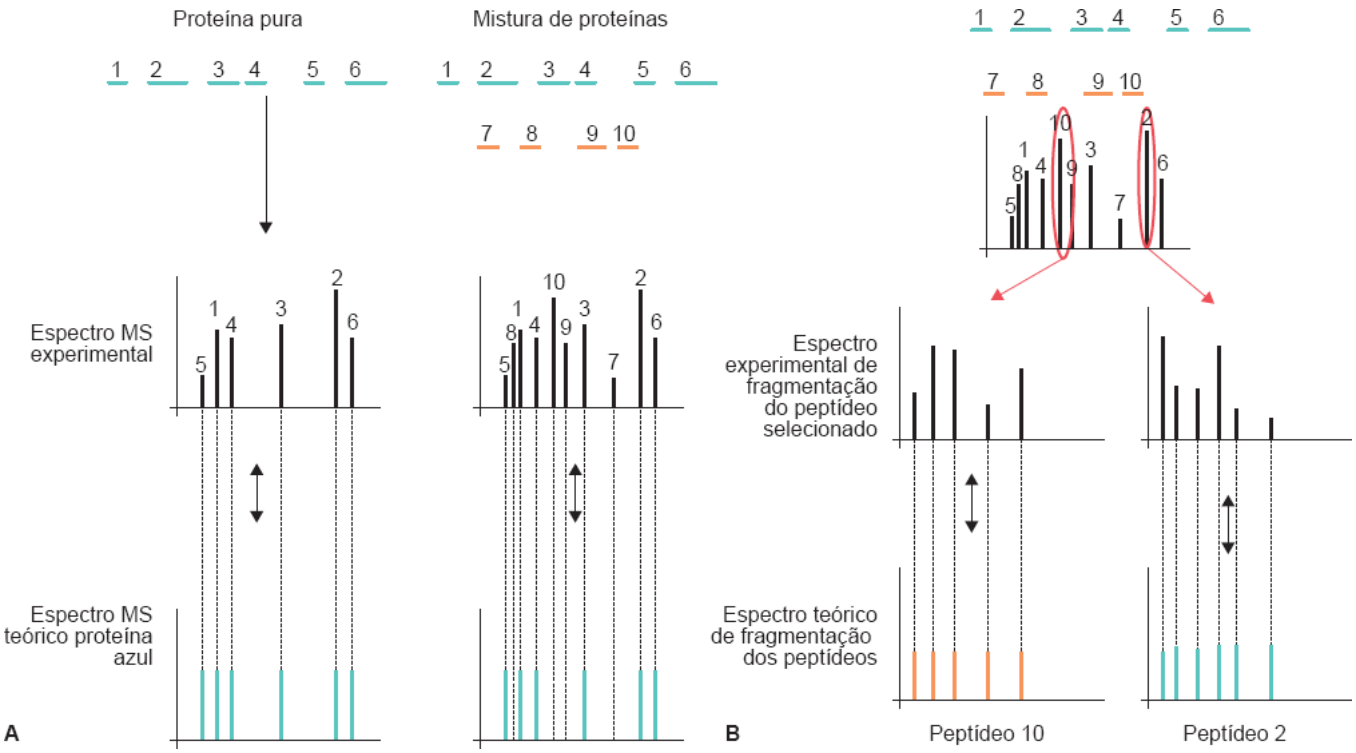


Figura 13.21 Comparação entre a identificação por impressão digital utilizando diferentes modos de espectrometria de massa. **A.** A utilização da espectrometria de massa de uma mistura de peptídeos derivados de uma única proteína (verde) possibilita a identificação a partir da comparação entre os picos dos espectros experimental e teórico. No entanto, a adição de peptídeos de uma segunda proteína (laranja) irá gerar novos picos no espectro e, portanto, haverá picos sem coincidência entre os espectros, não sendo mais possível identificar a proteína. **B.** Utilizando a espectrometria de massa em tandem, é possível selecionar picos representando peptídeos individuais (no caso da figura, os peptídeos 2 e 10) e é realizada a fragmentação destes, gerando um espectro individual de fragmentação de cada peptídeo. Comparando o espectro de fragmentação experimental com o teórico, é possível identificar os peptídeos números 2 e 10 e, consequentemente, inferir a presença das proteínas verde e laranja na amostra.

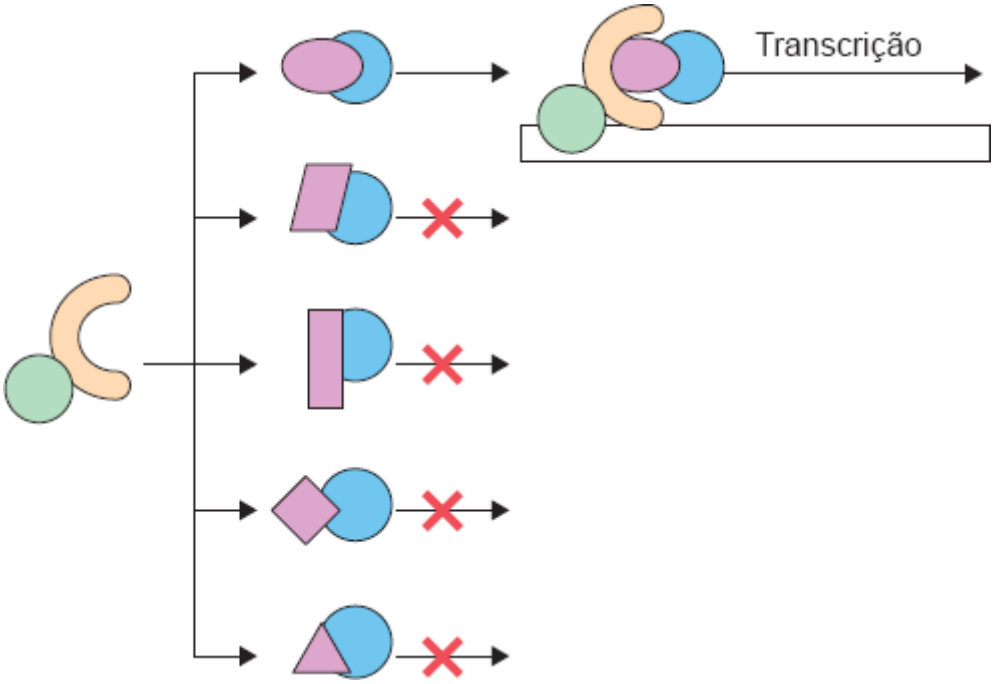


Figura 13.22 Representação esquemática da técnica de duplo híbrido. O esquema mais à direita representa a proteína-isca produzida a partir da fusão do gene de interesse (cujo produto é a porção laranja) com uma porção codificando um domínio de ligação de DNA (cujo produto é representado pelo círculo verde). Setas indicam instâncias distintas nas quais esse produto é exposto a diferentes proteínas presas, representadas em roxo e fundidas a um domínio ativador transcricional, em azul. Nos casos em que ocorre interação, o domínio ativador transcricional é recrutado à proximidade do sítio promotor reconhecido pela proteína ligante de DNA, o que possibilita a transcrição de um gene repórter.

Tais dados pareciam indicar que, em termos moleculares, não haveria uma grande diferença entre seres humanos e outros organismos multicelulares mais simples. No entanto, genes de mamíferos, quando comparados com aqueles de organismos mais simples, tendem a ser subdivididos em um maior número de exons. Isso sugere que, apesar de não haver um ganho tão considerável no número de genes, eles seriam mais complexos em mamíferos.

De fato, a partir do estudo em larga escala do transcriptoma humano, pesquisadores passaram a notar que genes de humanos tendiam a realizar com maior frequência o *splicing* alternativo de seus transcritos do que os organismos mais simples. A presença de genes com maior número de exons facilita o processo de retirada/introdução de módulos funcionais em uma proteína por meio do *splicing* alternativo, criando variantes que podem realizar funções diferentes da proteína original. Esses dados parecem apontar que a evolução de organismos mais complexos não é realizada somente com a aquisição de novos genes, mas também com a utilização criativa daqueles que já existem no organismo.

Estudos recentes comparando genomas e transcriptomas de primatas com o de humanos sugerem que grande parte das alterações fenotípicas entre as espécies é causada por variações no padrão de expressão dos genes, em vez de mutações nas suas regiões codificantes. Mutações em regiões não codificantes, como em promotores, tendem a produzir mudanças mais sutis que mutações em regiões codificantes, que levam a modificações em sequências proteicas. Tais dados indicam que a coordenação da expressão dos genes nos diversos tecidos possui um papel bastante relevante na evolução de organismos.

A visão mais tradicional do transcriptoma tinha como base o fato de que a maioria dos sítios de transcrição em nosso genoma estava associada à produção de mRNA para a produção de proteínas. Em adição a estes, existiriam os

loci para RNA estruturais, como o rRNA e o tRNA. No entanto, nos últimos anos, a exploração do transcriptoma de organismos utilizando tecnologia de microarranjos e sequenciamento em larga escala possibilitou a verificação de transcritos derivados de *loci* do genoma que não correspondiam aos exons geradores de mRNA. Utilizando esses dados, foi possível definir uma série de *loci* no genoma localizados em introns de genes ou em regiões intergênicas, onde ocorre a transcrição de RNA não codificantes. Essa classe de RNA não tem o potencial de produzir proteínas, pois não dispões de um quadro de leitura aberto em sua sequência que possibilite sua decodificação pelos ribossomos. Apesar de ainda não haver um consenso a respeito da função de tais RNA, há hipóteses de que eles realizariam a regulação da transcrição dos mRNA.

Esses dados sugerem que a dinâmica do transcriptomas de organismos eucariotos é mais complexa que o previamente imaginado. A existência de RNA não codificadores regulatórios adicionaria outro nível de regulação da abundância de transcritos, aumentando ainda mais a importância da caracterização de transcriptomas para o entendimento dos mecanismos celulares.

Eucariotos simples contêm proteomas extremamente intrincados

Com o desenvolvimento de técnicas proteômicas, foi possível realizar estudos em larga escala com o objetivo de caracterizar o conjunto de proteínas expressas por um organismo ou tecido. Foi realizado um mapeamento completo do proteoma da levedura *Saccharomyces cerevisiae*, um eucarioto unicelular, durante o seu período de crescimento em fase exponencial. Nas células estudadas, foram detectadas proteínas representando 80% dos genes desse organismo (cerca de 4.200 proteínas). A abundância individual das proteínas detectadas variava de 50 a 10^6 moléculas por célula.

Ao estudar diretamente proteomas mais complexos, como os tecidos humanos, a lista de identificação de proteínas dificilmente chega ao número de proteínas identificadas em *S. cerevisiae*. Esse aparente paradoxo foi explicado pelo fato de que, nos proteomas de organismos complexos, existe uma diferença muito maior entre as quantidades das proteínas mais abundantes em comparação com as menos abundantes (diferença na ordem de 10^{11}) do que as diferenças verificadas em células de organismos unicelulares simples (diferença da ordem de 10^4). Tais diferenças, em associação às limitações técnicas dos presentes métodos para estudo do proteoma, fazem com que, nesses organismos mais complexos, somente a fração mais abundante do proteoma possa ser detectada. Em adição, organismos mais complexos têm maior fração de proteínas que apresentam isoformas, devido a fenômenos como o *splicing* alternativo e modificações pós-traducionais, o que dificulta sobretudo a detecção dessas proteínas. Esses dados demonstram que a diferença entre proteomas de organismos de diferente complexidade não se limita ao número de proteínas presentes. Estudos proteômicos em células mais complexas devem utilizar técnicas de fracionamento celular e outras abordagens para fracionar amostras de proteínas, com o objetivo de obter uma cobertura maior desses proteomas.

Estudos proteômicos em leveduras utilizando inibidores da síntese proteica possibilitaram analisar o processo de degradação de proteínas. Uma vez que a síntese de proteína era bloqueada, a abundância das diversas proteínas da levedura era monitorada. Isso possibilitava a análise de sua taxa de degradação, pois novas proteínas não eram sintetizadas, e qualquer variação da quantidade de proteína é resultado somente da degradação das proteínas já existentes no citoplasma. Foi possível notar que as proteínas tinham, em média, uma meia-vida (tempo necessário para atingir metade da concentração original) de 43 min. No entanto, foi possível dividir as proteínas em duas populações distintas: a primeira consiste em proteínas produzidas em grandes quantidades e com grande estabilidade (esta classe é enriquecida com enzimas envolvidas na síntese de proteínas); a segunda é formada por proteínas de menor abundância que são degradadas rapidamente (esta classe é enriquecida com proteínas envolvidas no controle do ciclo celular). Aparentemente, o desenvolvimento dessas duas populações é uma adaptação das células para impedir gastos energéticos desnecessários ao evitar a degradação rápida de proteínas essenciais, mas, ao mesmo tempo, tendo uma população de proteínas de regulação com menor estabilidade, que possibilite respostas celulares rápidas através da variação da abundância dessas proteínas.

Estudos de larga escala utilizando a técnica de purificação por afinidade em tandem com proteínas da levedura *S.cerevisiae* tornaram possível caracterizar centenas de complexos proteicos, que resultam de milhares de interações em células que se apresentavam em fase de crescimento em meio de cultura. Estudos de duplo híbrido com proteínas de *S.cerevisiae* também permitiram mapear milhares de interações entre proteínas desse organismo, confirmando o alto número de interações entre as suas proteínas. Esses dados demonstram que a grande parte das

proteínas não atua de modo isolado, mas sim como parte desses complexos multiproteicos. Devido à alta complexidade de células de organismos eucarióticos multicelulares, é difícil realizar uma catalogação de todas as interações de uma célula, como foi feito para levedura. No entanto, estudos mais restritos parecem confirmar que um grande número de interações também existe nas células mais complexas.

Todos esses dados demonstram que até mesmo o proteoma de organismo mais simples apresenta uma grande heterogeneidade em relação à abundância e estabilidade de proteínas. Além disso, o entendimento atual das redes de interação das proteínas presentes no proteoma continua bastante preliminar, e ainda são necessários muitos estudos para compreender de modo satisfatório os diversos mecanismos celulares mediados pelo nosso proteoma.

Considerações finais

O avanço das tecnologias no estudo de biomoléculas vem possibilitando uma nova abordagem mais abrangente, que permite interrogar o comportamento de grandes conjuntos de moléculas, em contraste com técnicas voltadas para o estudo caso a caso. Isso torna possível o desenvolvimento de uma percepção mais global do funcionamento celular e dos desafios em seu estudo. Ainda restam enormes lacunas em nosso entendimento completo da biologia molecular da célula e, certamente, são necessárias maior acumulação e integração de dados biológicos. No entanto, se considerarmos que essas técnicas de estudo em larga escala são relativamente recentes, é possível perceber seu grande potencial, e esperar que elas favoreçam grandes avanços no entendimento da biologia celular.

Bibliografia

Beck M, Claassen M, Aebersold R. Comprehensive proteomics. Curr Opin Biotechnol. 2011;22(1):3-8.

Brady SM, Long TA, Benfey PN. Unraveling the dynamic transcriptome. Plant Cell. 2006;18(9):2101-11.

Cox J, Mann M. Is proteomics the new genomics? Cell. 2007;130(3):395-8.

Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, Dephoure N, et al. Global analysis of protein expression in yeast. Nature. 2003;425(6959):737-41.

Griffiths J. A brief history of mass spectrometry. Anal Chem. 2008;80:5678-83.

Gustincich S, Sandelin A, Plessy C, Katayama S, Simone R, Lazarevic D, et al. The complexity of the mammalian transcriptome. J Physiol. 2006;575(Pt 2):321-32.

O'Farrell PH. The pre-omics era: the early days of two-dimensional gels. Proteomics. 2008;8(23-24):4842-52.

Larance M, Lamond AI. Multidimensional proteomics for cell biology. Nat Rev Mol Cell Biol. 2015;16(5):269-80.

Völkel P, Le Faou P, Angrand PO. Interaction proteomics: characterization of protein complexes using tandem affinity purification-mass spectrometry. Biochem Soc Trans. 2010;38(4):883-7.

Wilkins MR. Hares and tortoises: the high- versus low-throughput proteomic race. Electrophoresis. 2009;30(Suppl 1):S150-5.

14 Como Desvendar Dados Biológicos com a Bioinformática

Robson Francisco de Souza, Leandro Marcio Moreira, Diego Bonatto e Pedro A. F. Galante

A ciência biológica passa por uma revolução causada por métodos de geração e análise de dados em larga escala. Transformar esses dados em informação e conhecimento é o objetivo primordial da bioinformática, uma disciplina na fronteira entre biologia, ciência da computação, estatística e medicina (entre outras áreas), que tem no desenvolvimento de base de dados, ferramentas computacionais e metodologias de análise de dados biológicos as suas principais ramificações. Neste capítulo será descrita a importância da bioinformática no estudo de dados biológicos gerados em larga escala e serão explorados os principais portais de integração, representação e análise de dados, alinhamento de sequências, estudo de genomas e a biologia de sistemas. Como a bioinformática é uma área muito dinâmica, na qual surgem diariamente novas ferramentas, metodologias e bases de dados, optou-se por focar a descrição dos tópicos mais consolidados e de uso mais frequente. Embora não sejam exaustivos, os tópicos selecionados para este capítulo possibilitarão o leitor entender como a bioinformática pode ser usada para explorar esse novo universo de dados biológicos em larga escala.

Introdução

Decorridos mais de 10 anos após a conclusão do projeto genoma humano (ver Capítulo 12), biólogos, biomédicos e médicos, assim como tantos outros setores da sociedade, assistem hoje a uma revolução do conhecimento biológico desencadeada pela introdução de métodos automáticos de geração de dados em larga escala. Um dos elementos mais importantes dessa revolução é a intensa pressão para o desenvolvimento de novos algoritmos computacionais de análise, de gerenciamento e de distribuição de dados que possibilitem processar a grande quantidade de resultados que são produzidos, sobretudo, pelas novas tecnologias de sequenciamento de e RNA e também por outras tecnologias que foram recentemente introduzidas ou adaptadas para processamento de dados biológicos em larga escala, tais como a espectroscopia de massa (ver Capítulo 13) ou novos métodos automatizados de microscopia ótica.[1] A área de intersecção entre biologia, (bio)medicina, estatística e ciência da computação no estudo, desenvolvimento de metodologias e ferramentas de análise de todos esses dados é chamada bioinformática ou biologia computacional.

Breve história da interação entre computadores, matemática e biologia

A bioinformática ganhou notoriedade como disciplina independente a partir dos anos 1990, quando o método de sequenciamento de genomas completos baseado em fragmentos aleatórios de DNA passou a ser aplicado no sequenciamento de genomas de bactérias, o que foi seguido de uma explosão no volume de dados produzidos em laboratórios de biologia molecular, fato que tornou obrigatória a participação, nesses projetos, de pesquisadores treinados em análise e processamento de dados em larga escala.[2] Porém, o emprego de computadores e de métodos computacionais em biologia, bem como o uso de métodos matemáticos para construir modelos de sistemas biológicos e de métodos estatísticos para sua análise, já havia florescido décadas antes e coincide tanto com a introdução de computadores de estado sólido (transistores) em departamentos de universidades e instituições de

pesquisa quanto com a introdução de protocolos e equipamentos para o sequenciamento e a determinação da estrutura terciária de proteínas.

Foi nesse contexto e época que Margareth Dayhoff criou o primeiro banco de dados de sequências, o Atlas de Estruturas e Sequências de Proteínas, inicialmente publicado na forma de livro – *Atlas of protein sequence and structure* (1965) –, mas convertido para formato digital quando os meios de distribuição para tal tornaram-se disponíveis. O esforço pioneiro de Dayhoff também incluiu a introdução da primeira matriz de substituição de aminoácidos, chamada PAM (do inglês, *point accepted mutation*), que foi importante para o estabelecimento dos métodos automáticos de comparação de sequências.

Como aconteceu em outras áreas nascidas da interface de disciplinas preexistentes, a bioinformática agregou e absorveu conceitos e métodos de várias dessas disciplinas, em especial métodos estatísticos para lidar com volumes enormes de dados e métodos computacionais para processar informações, no geral, em formato de "texto". A seguir estão detalhados diversos portais (interfaces para bancos de dados), acessíveis via web. Trata-se de ferramentas e metodologias da bioinformática para explorar e desvendar características de genomas de todos os organismos vivos do nosso planeta.

Principais portais integrados para análise e repositórios de dados

Com o avanço na geração de informações biológicas, a criação de base de dados para a deposição, organização e apresentação dessas informações à comunidade científica tornou-se imprescindível. A partir da década de 1980, bancos de dados biológicos, em especial os de natureza molecular, começaram a ser criados por grupos e institutos de pesquisas isolados, os quais, mais tarde, organizaram-se em consórcios e obtiveram apoio político para a formação de institutos centrados na organização de repositórios centrais. Hoje, inúmeros bancos de dados, tanto gerais como especializados, são mantidos por essas organizações e por empresas ou grupos de pesquisa de todo o mundo e podem ser acessados e utilizados via internet ou outras tecnologias. Dentre esses, destacam-se alguns dos repositórios que foram pioneiros e que resistiram à pressão do tempo, permanecendo como referência para a análise dos tipos de dados que armazenam.

NCBI

O Centro Nacional de Informações para Biotecnologia – NCBI (do inglês, National Center for Biotechnology Information) - surgiu em 1988. Sua criação se deu pela convergência de três ações políticas distintas que ocorreram entre os anos de 1984 e 1987 nos EUA, cujos temas focais envolviam o armazenamento e a divulgação de informações biotecnológicas tanto para a comunidade científica quanto para a população em geral, as quais precisavam ter acesso a informações precisas e de qualidade. Em sua essência, e após algumas modificações estruturais acarretadas pelo avanço das tecnologias biológicas e da informação, o renovado NCBI tem como metas:

- Desenhar, desenvolver, implementar e gerenciar sistemas automatizados para a recolha, armazenamento, recuperação, análise e disseminação de conhecimentos sobre biologia molecular, bioquímica e genética
- Realizar investigações a partir de métodos avançados de processamento de informações que possibilitem representar e analisar o grande número de moléculas e compostos biologicamente importantes
- Possibilitar que as pessoas envolvidas em pesquisas em biotecnologia e assistência médica usem sistemas e métodos desenvolvidos e armazenados em seus domínios sem ônus financeiro aos usuários
- Coordenar, tanto quanto possível, os esforços para reunir informações na área de biotecnologia em nível internacional.

Com essas metas e com o aporte de informações da comunidade científica internacional, o NCBI passou a ser um dos principais portais mundiais de informações biotecnológicas, disponibilizando uma série de ferramentas que agregam interesses variados aos usuários. Dentre as bases de dados e as ferramentas mais importantes disponibilizadas no portal do NCBI, merece destaque o ENTREZ.

ENTREZ

O ENTREZ está disponível no portal NCBI e compreende um sistema de busca em 40 bases de dados. Sua interface simples possibilita que o usuário, a partir de uma entrada específica (palavra ou conjunto de palavras), reúna todas as informações necessárias para embasamento de uma pesquisa científica, caracterizando-se como uma poderosa

ferramenta de organização e gestão de recursos técnico-científicos.[3] Para isso, basta que o usuário insira uma ou mais palavras-chave na barra de busca do NCBI, mantendo como padrão o ícone *All databases*. Além de fornecer resultados categorizados em temas centrais (Figura 14.1), possibilita também visualizar um breve histórico de ações de busca, otimizando o sistema de organização das informações.

Dentre as bases de dados gerenciadas pelo ENTREZ, algumas merecem destaque e serão descritas a seguir na íntegra. Já outras, embora tão importantes quanto aquelas descritas, terão suas funções apenas sumarizadas para que o leitor compreenda sua importância.

É importante conhecer: o BioProject, que agrega informações sobre os projetos de sequenciamento genômico (total ou parcial) que abastecem o NCBI com informações derivadas desses projetos; o Genome, que disponibiliza informações sobre os projetos genoma finalizados ou em andamento e dá acesso aos arquivos que contêm tais dados; e o dbSNP (do inglês, *single nucleotide polymorphism database*), que oferece informações sobre pequenas variações genéticas que podem ou não estar associadas a alterações de fenótipos em diversos organismos. Já nas bases de dados sobre informações de genes, merecem atenção dbEST, Gene e UniGene.

O dbEST (do inglês, *expressed sequence tags database*) contém informações sobre sequências transcritas oriundas de organismos eucariotos. O UniGene agrega informações sobre agrupamentos de sequências transcritas nos mais variados organismos. Nas bases de dados de Proteínas, é interessante que o usuário saiba onde estão depositadas (*Proteins*) informações sobre essas proteínas e também possa ter acesso a dados estruturais (*Structures*). Finalmente, nas bases de dados de substâncias químicas, é interessante conhecer o BioSystems, que disponibiliza informações sobre vias moleculares com *links* para genes, proteínas e produtos químicos, e o PubChem Compound, que contém informações sobre compostos químicos com *links* de acesso a outros bancos de dados pertinentes. O PubMed, o Gene e o OMIM serão descritos com mais detalhes a seguir.

PubMed

O PubMed é um repositório que compreende mais de 25 milhões de títulos científicos relacionados a trabalhos do MEDLINE, de jornais que abordam as ciências da vida e de livros *on-line*. Com um mecanismo de busca simples e eficiente, basta que usuário insira palavras-chave a respeito da temática que pretende ter acesso, e a ferramenta retornará todos os possíveis resultados dessa busca. Erros de digitação determinarão uma busca imprecisa ou nula. Como resultado, e em caso positivo de busca, a ferramenta apresentará uma lista de trabalhos relacionados à temática. Um exemplo de resultado de busca está detalhado na Figura 14.2.

OMIM

A base de dados *on-line* sobre heranças mendelianas no ser humano, ou OMIM (do inglês, *Online Mendelian Inheritance in Man*), compreende informações sobre genes humanos e fenótipos genéticos associados, e é continuamente atualizada com base nos avanços da pesquisa científica. A partir da entrada de uma ou mais palavras-chave, que podem corresponder a um gene ou uma síndrome relacionada, o OMIM retornará informações genéticas e citogenéticas. A Figura 14.3 esboça um desses resultados de busca a partir da entrada *Li-Fraumani*. O usuário poderá consultar o banco que descreve a localização citogenética do gene no cromossomo (nesse caso, cromossomo 1, braço curto “q”, *locus* 23) ou acessar as coordenadas genômicas no respectivo cromossomo. Nesse caso, terá informações sobre todos os genes que ocupam esse *locus*, sobre os mRNA associados a essa região, um grafo de alinhamento múltiplo de sequências com outros genomas de vertebrados, localização de SNP, bem como de regiões repetitivas (LINE, SINE, LTR, entre outras).

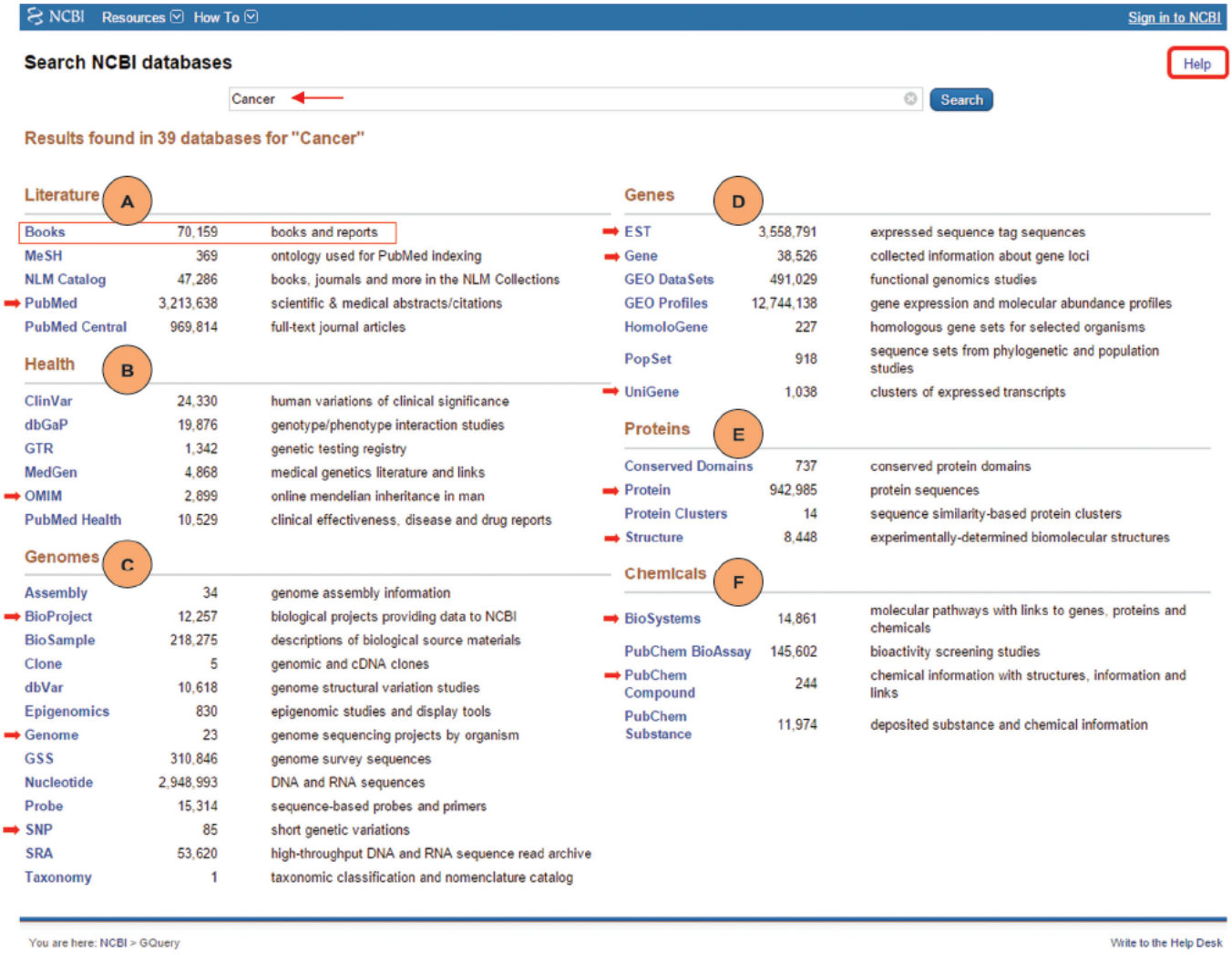


Figura 14.1 Exemplo de um resultado de busca na ferramenta ENTREZ do NCBI. As letras **A** a **F** destacam os temas centrais aos quais os bancos de dados estão relacionados. As setas vermelhas destacam algumas das mais importantes bases de dados para cada um dos temas centrais. Note que a busca para esses resultados se deu a partir da entrada com a palavra "Cancer". Em caso de dúvidas, o usuário consegue acessar um menu de ajuda (*Help*) bastante descritivo e de fácil leitura (em destaque no canto superior direito).

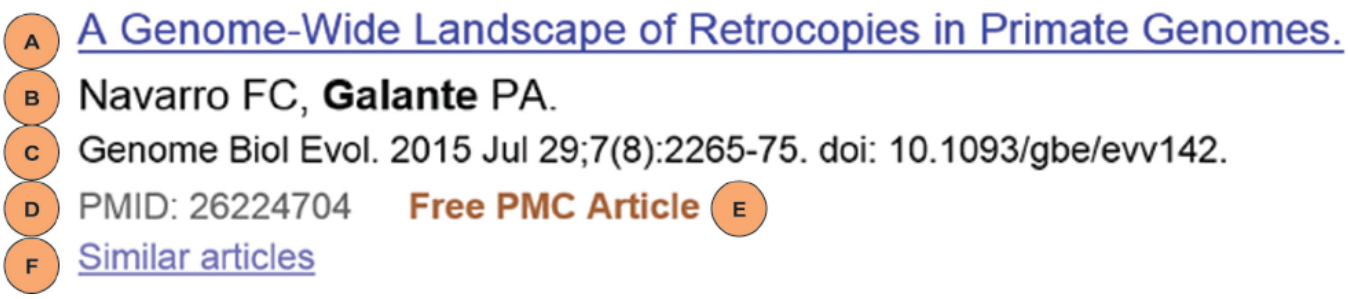


Figura 14.2 Exemplo de um resultado de busca a partir do PubAed (as letras foram inseridos para facilitar a descrição presente no corpo do texto). Para cada trabalho encontrado, será detalhado o título na íntegra (**A**), os autores envolvidos (**B**), o veículo de divulgação no qual esse trabalho foi publicado (**C**), um identificador do trabalho em caso de artigos científicos (**D**), que pode ou não estar disponível gratuitamente (**E**), e outros possíveis artigos relacionados, caso existam (**F**).

Gene

Como o próprio nome já diz, esse banco de dados reúne todas as informações associadas a um gene específico. Cabe ao usuário apenas iniciar a procura pelo nome ou GI (*gene identification*, o identificador numérico do gene no banco) que a interface web lhe retornará tudo que possa estar associado a ele. Como resultado da busca, terá acesso a uma lista de resultados (Figura 14.4 A) e, a partir de uma escolha de um dos itens dessa lista, seguirá para uma página que contém o sumário de informações sobre o gene escolhido (Figura 14.4 B), todo o contexto genômico no qual se insere (Figura 14.4 C), sua localização no respectivo cromossomo (Figura 14.4 D) e um mapa detalhado da

região genômica e dos transcritos e proteínas associadas (Figura 14.4 E e F), incluindo as variações de transcritos destacando sequências de introns e exons e possíveis SNP.

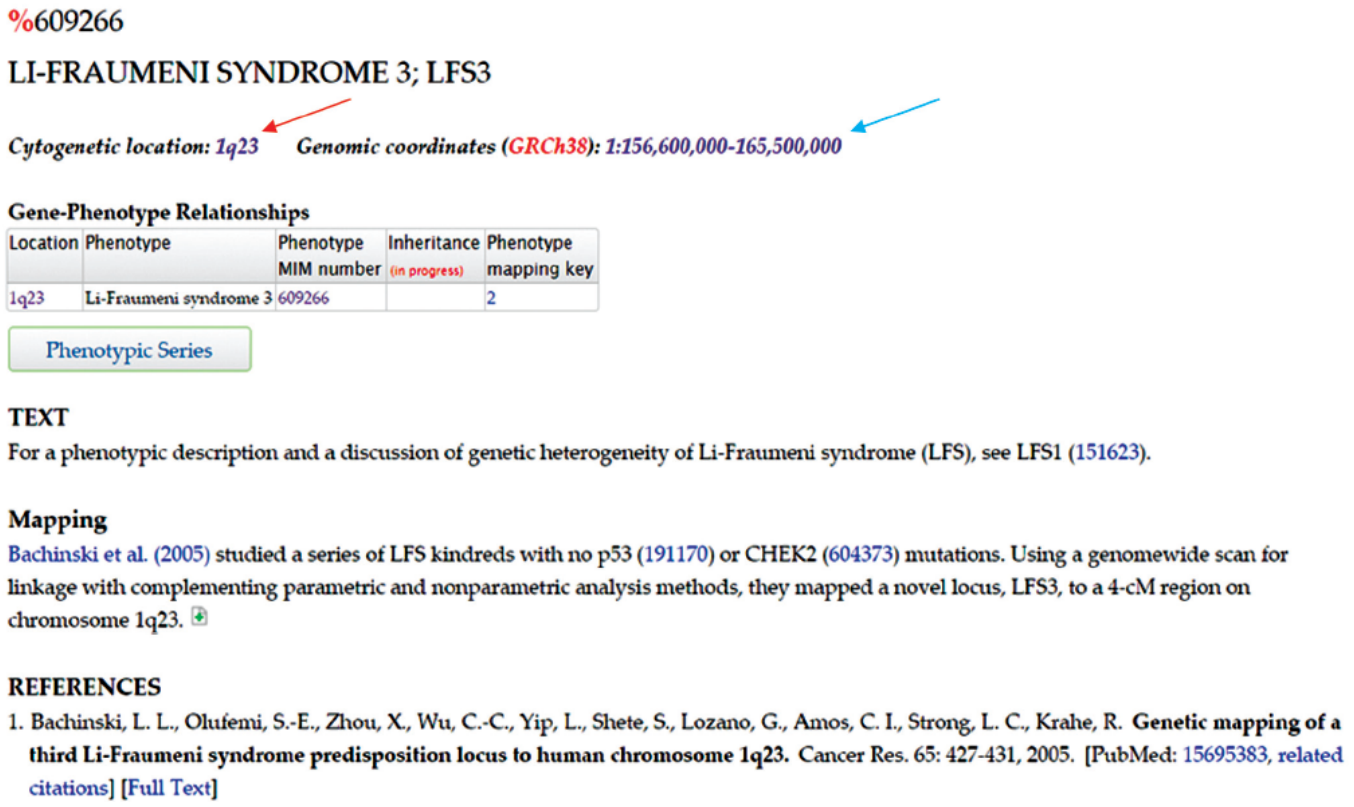
EBI

O Instituto Europeu de Bioinformática, ou EBI (do inglês, European Bioinformatics Institute), é parte integrante do Laboratório Europeu de Biologia Molecular, ou EMBL (do inglês, European Molecular Biology Laboratory). Com propósitos similares aos do NCBI, o EBI foi criado em 1993, a partir de uma iniciativa europeia; tais propósitos já vinham sendo trabalhados desde a criação do EMBL em 1980, e a missão do EBI é:

- Fornecer serviços de dados e de bioinformática gratuitamente para todos os interesses da comunidade científica
- Contribuir para o avanço da biologia por meio da investigação básica por iniciativa dos pesquisadores
- Proporcionar formação avançada em bioinformática para cientistas em diferentes níveis de conhecimento
- Ajudar no processo de divulgação das tecnologias de ponta para a indústria
- Coordenar o fornecimento de dados biológicos em toda a Europa.

Fazendo uso de uma interface simples e objetiva, basta ao usuário digitar a(s) palavra(s)-chave no campo de busca e clicar no botão *Search*, e o EBI retornará resultados correspondentes em todas as suas bases de dados (Figura 14.5). A classificação em temas centrais, como previamente descrito para o ENTREZ do NCBI, facilita a leitura e a organização dos dados correlacionados. Isso possibilita que qualquer usuário possa navegar pelas informações biológicas depositadas nessa base de dados.

Juntos, o NCBI e o EBI são certamente os principais bancos de dados para informações associadas a genes e genomas. Entretanto, alguns outros bancos são fonte primária para certos tipos de dados e apresentam interfaces muito eficientes e bem-sucedidas, o que faz com que mereçam destaque, como é o caso do KEGG e do PDB.



%609266

LI-FRAUMENI SYNDROME 3; LFS3

Cytogenetic location: 1q23 *Genomic coordinates (GRCh38): 1:156,600,000-165,500,000*

Gene-Phenotype Relationships

Location	Phenotype	Phenotype MIM number	Inheritance (in progress)	Phenotype mapping key
1q23	Li-Fraumeni syndrome 3	609266		2

Phenotypic Series

TEXT

For a phenotypic description and a discussion of genetic heterogeneity of Li-Fraumeni syndrome (LFS), see LFS1 (151623).

Mapping

Bachinski et al. (2005) studied a series of LFS kindreds with no p53 (191170) or CHEK2 (604373) mutations. Using a genomewide scan for linkage with complementing parametric and nonparametric analysis methods, they mapped a novel locus, LFS3, to a 4-cM region on chromosome 1q23.

REFERENCES

1. Bachinski, L. L., Olufemi, S.-E., Zhou, X., Wu, C.-C., Yip, L., Shete, S., Lozano, G., Amos, C. I., Strong, L. C., Krahe, R. **Genetic mapping of a third Li-Fraumeni syndrome predisposition locus to human chromosome 1q23.** Cancer Res. 65: 427-431, 2005. [PubMed: 15695383, related citations] [Full Text]

Figura 14.3 Resultado de uma busca na base de dados OMIM. Na página de resultados obtidos a partir de uma palavra-chave, o usuário poderá ter acesso a informações sobre a localização citogenética do gene no *locus* correspondente (seta vermelha), bem como coordenadas genômicas e informações moleculares associadas ao *locus* (seta azul), além de uma bibliografia completa associada a tal palavra-chave.

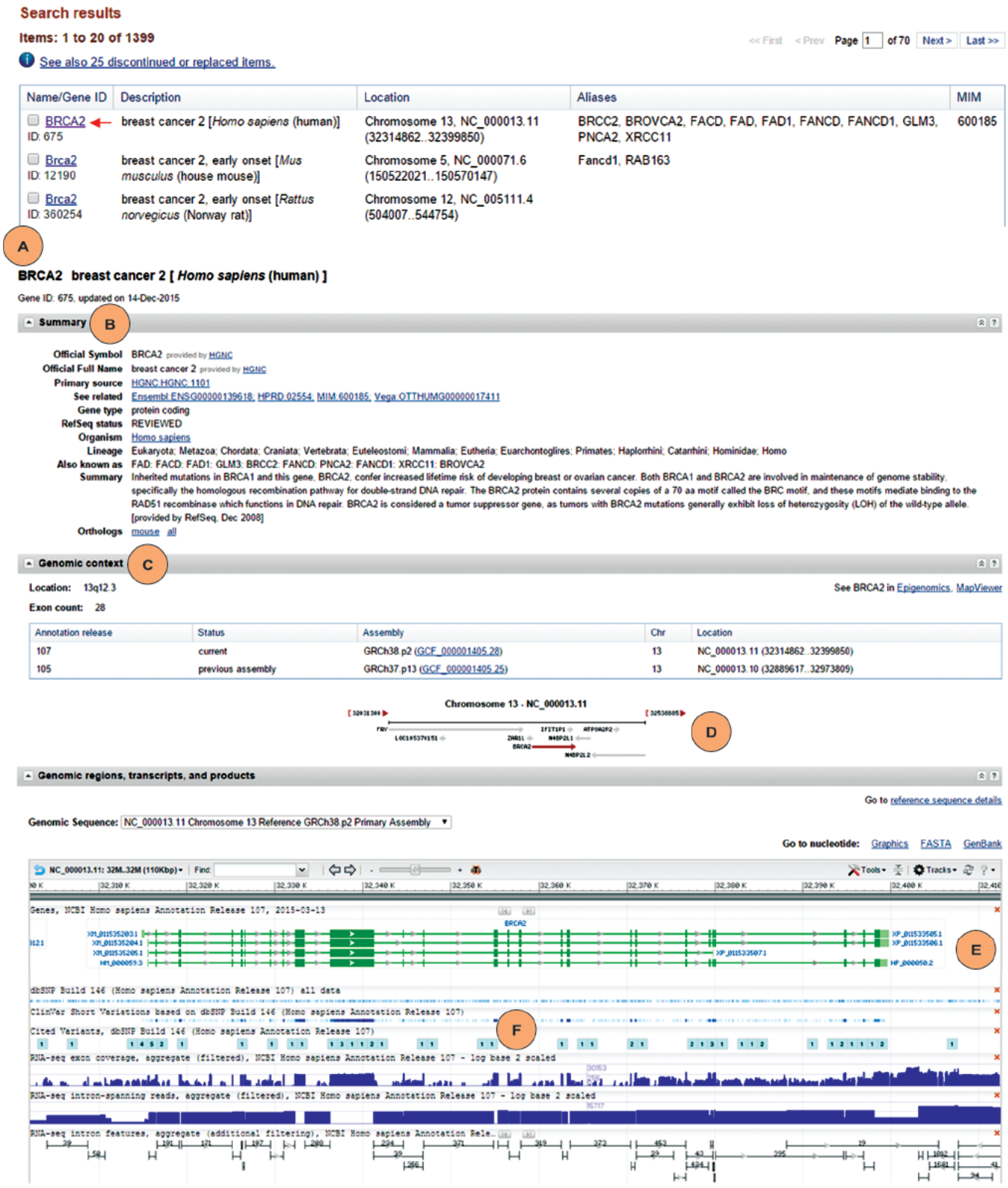


Figura 14.4 Resultados de busca na base de dados Gene. **A.** Resultado primário apresentado em uma lista que contém nome do gene, sua descrição, localização genômica, outras nomeações ao mesmo gene ou a genes parálogos e código MIM (que pode ser acessado no OMIM). **B** a **F.** Resultado completo associado à escolha do usuário em A. A seta vermelha destaca uma suposta seleção do usuário detalhada em E e F.

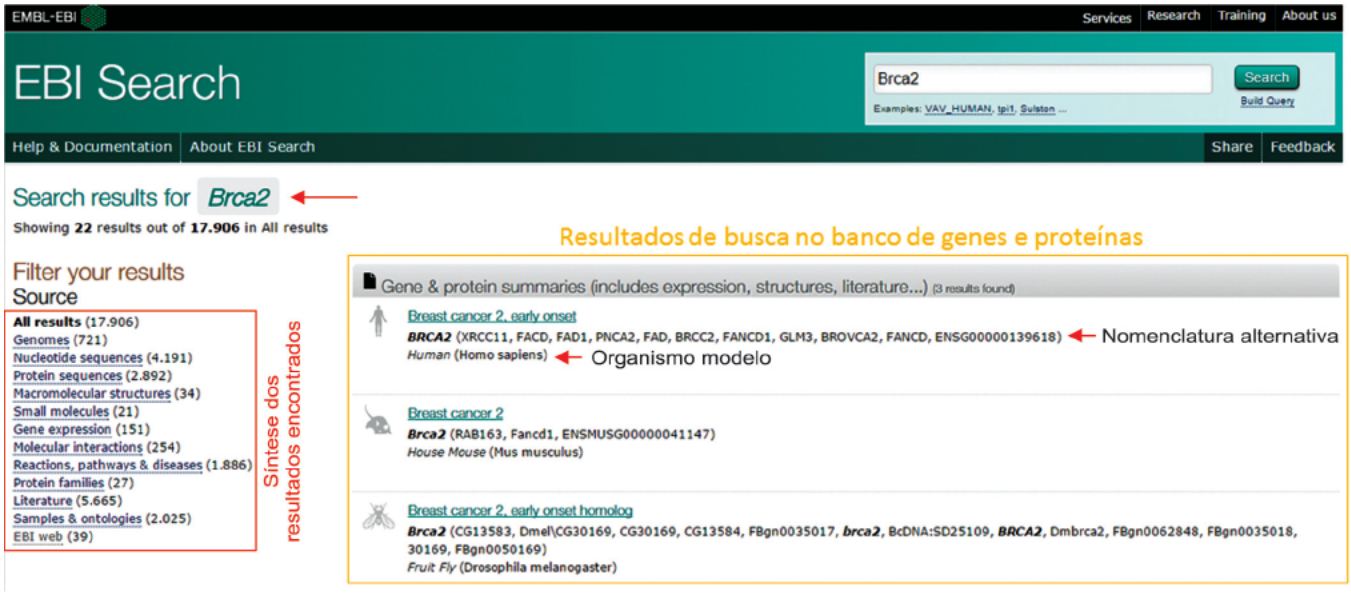


Figura 14.5 Resultado de busca a partir do EBI. À esquerda, destaca-se o menu, que apresenta uma síntese dos resultados encontrados a partir da palavra buscada – neste caso, o nome do gene relacionado com câncer de mama (*Brca2*). Na parte central da página, são apresentados os resultados da busca: os três destacaram os acrônimos do gene (nomenclatura alternativa) e o organismo-modelo de estudo (setas vermelhas).

KEGG

Assim como as bases americana e europeia, a Enciclopédia de Genes e Genomas de Kyoto (Japão), ou KEGG (do inglês, *Kyoto Encyclopedia of Genes and Genomes*), reúne uma base de dados que integra informações biológicas de fácil acesso ao usuário. Entretanto, diferentemente das anteriores, o KEGG apresenta um recurso fabuloso e indispensável que possibilita que o usuário não apenas conheça os genes e proteínas de um organismo, mas entenda qual é o papel deles no metabolismo.

A partir da página do KEGG, clicando-se na opção *KEGG Pathways* e depois selecionando um dos *links* na seção *Global and overview maps*, é possível examinar um dos mapas metabólicos para entender se um determinado organismo apresenta ou não os genes que codificam as enzimas de uma determinada via metabólica (Figura 14.6). O usuário pode selecionar um organismo que tenha o genoma disponibilizado na base de dados do KEGG e carregar os mapas metabólicos desse organismo. Por comparação do mapa para o organismo com o mapa metabólico global de referência, é possível verificar essas possíveis ausências completas de vias descritas anteriormente.

Além dessa análise global, o KEGG *Pathway* disponibiliza mapas integrativos e dinâmicos de diversas vias metabólicas – p. ex., vias associadas a processos celulares, vias associadas a doenças humanas e vias envolvidas com desenvolvimento de drogas. Para todos esses casos, o KEGG possibilita que o usuário saiba quais são as estruturas moleculares que apresentam os substratos e produtos de cada reação, os cofatores necessários às enzimas que atuam nessas vias e os genes codificadores dessas proteínas em cada um dos genomas disponíveis no banco de dados (Figura 14.7). Uma vez selecionado o organismo-modelo com base no qual essa via pretende ser analisada, o programa recarrega a via metabólica colorindo cada uma das enzimas cujos genes codificadores foram encontrados no genoma dessa espécie. Isso possibilita que o usuário determine se uma rota dentro de uma via está ou não completa e que possíveis consequências isso poderia acarretar no organismo que a apresenta ou não. Além disso, por ser uma interface dinâmica, permite que o usuário, a partir da via escolhida, migre para outras vias correlacionadas, tornando a análise ainda mais poderosa. Embora tenha uma interface de fácil compreensão, exige do usuário certo conhecimento de bioquímica para que suas informações possam ser extraídas com mais eficácia.

Durante a análise genômica comparativa de espécies aparentadas, fazer esse tipo de análise de presença ou ausência de genes/proteínas para vias metabólicas específicas possibilita que os pesquisadores proponham hipóteses sobre as distinções fenotípicas com base nas diferenças observadas nas vias metabólicas dos organismos comparados.

PDB

O Banco de Dados de Proteínas, ou PDB (do inglês, *Protein Data Bank*), serve de repositório para informações referentes à estrutura e à função de proteínas desde 1971.[4] Diferentemente das bases de dados descritas anteriormente, o PDB é especializado em proteínas e suas funções. Embora o PDB seja uma iniciativa estadunidense, atualmente há um PDB europeu (PDBe) e um japonês (PDBj).

Para que o usuário possa fazer as buscas desejadas, basta inserir as palavras-chave na barra de busca do PDB (Figura 14.8). A plataforma retornará todos os resultados associados e, em caso de resultados positivos para uma ou mais estruturas proteicas, será fornecida uma lista de todas as estruturas.

Caso selecione uma das estruturas, o usuário terá acesso ao modelo tridimensional da proteína, via Jmol ou outro programa de visualização, o que possibilitará que manipule e gire a estrutura, modificando o estilo de representação tridimensional ou inserindo o modelo de poliedro e eixos para representar a célula unitária (menor unidade assimétrica) da estrutura. O usuário também poderá salvar o arquivo ".pdb" da estrutura proteica de interesse, o qual poderá ser importado em outros programas de análise de estruturas moleculares.

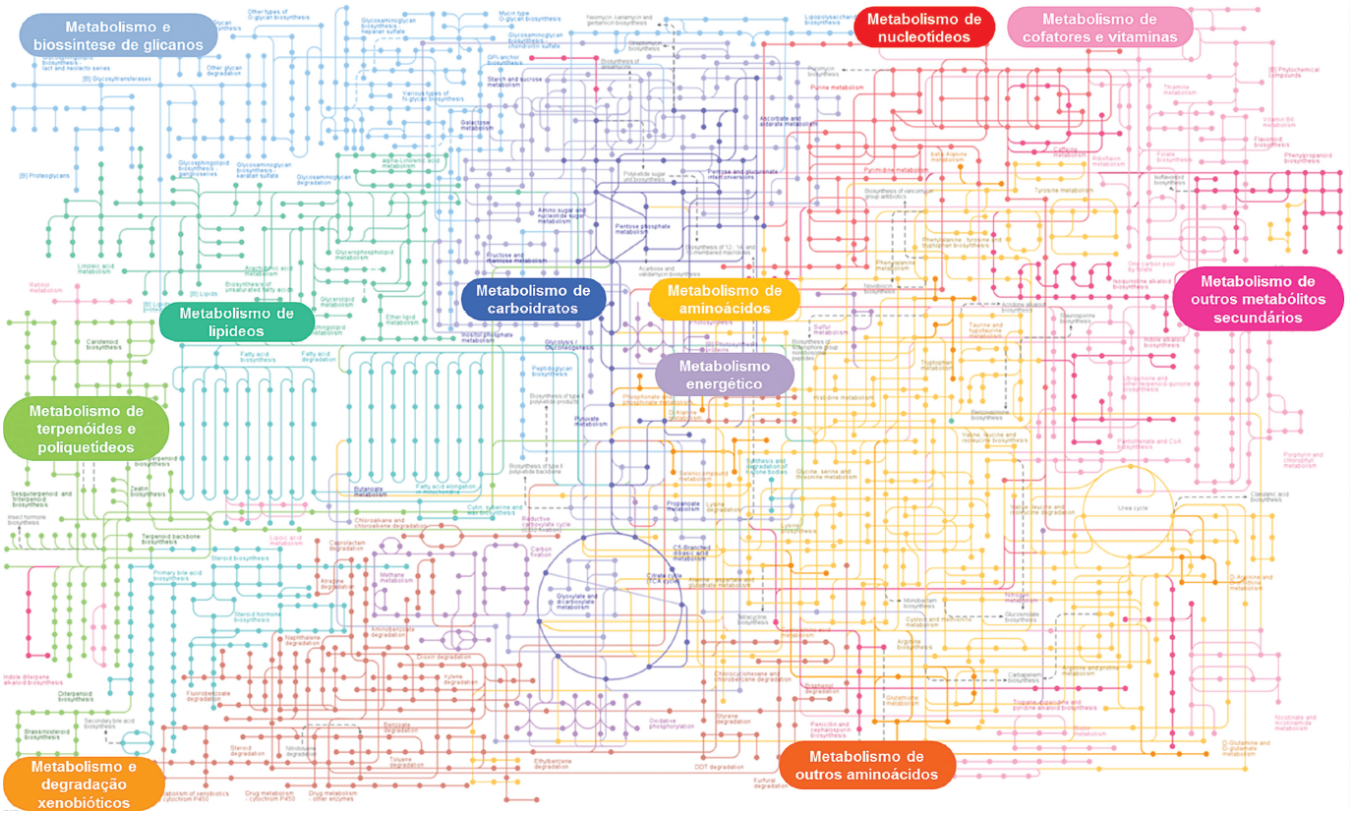


Figura 14.6 Visão global das vias metabólicas. Em virtude da complexidade de interação e do número de elementos, é possível identificar as diversas subvias que compõem o sistema metabólico apenas por meio dos seus padrões de cores. Por exemplo, em vermelho está a via de metabolismo de nucleotídeos, em azul a via de metabolismo de carboidratos e em verde escuro a via de metabolismo de lipídeos. É interessante notar que há certo grau de sobreposição entre diversas subvias. Na página do KEGG na internet é possível fazer aproximações, ver detalhes de todos os pontos dessa e de outras vias celulares.

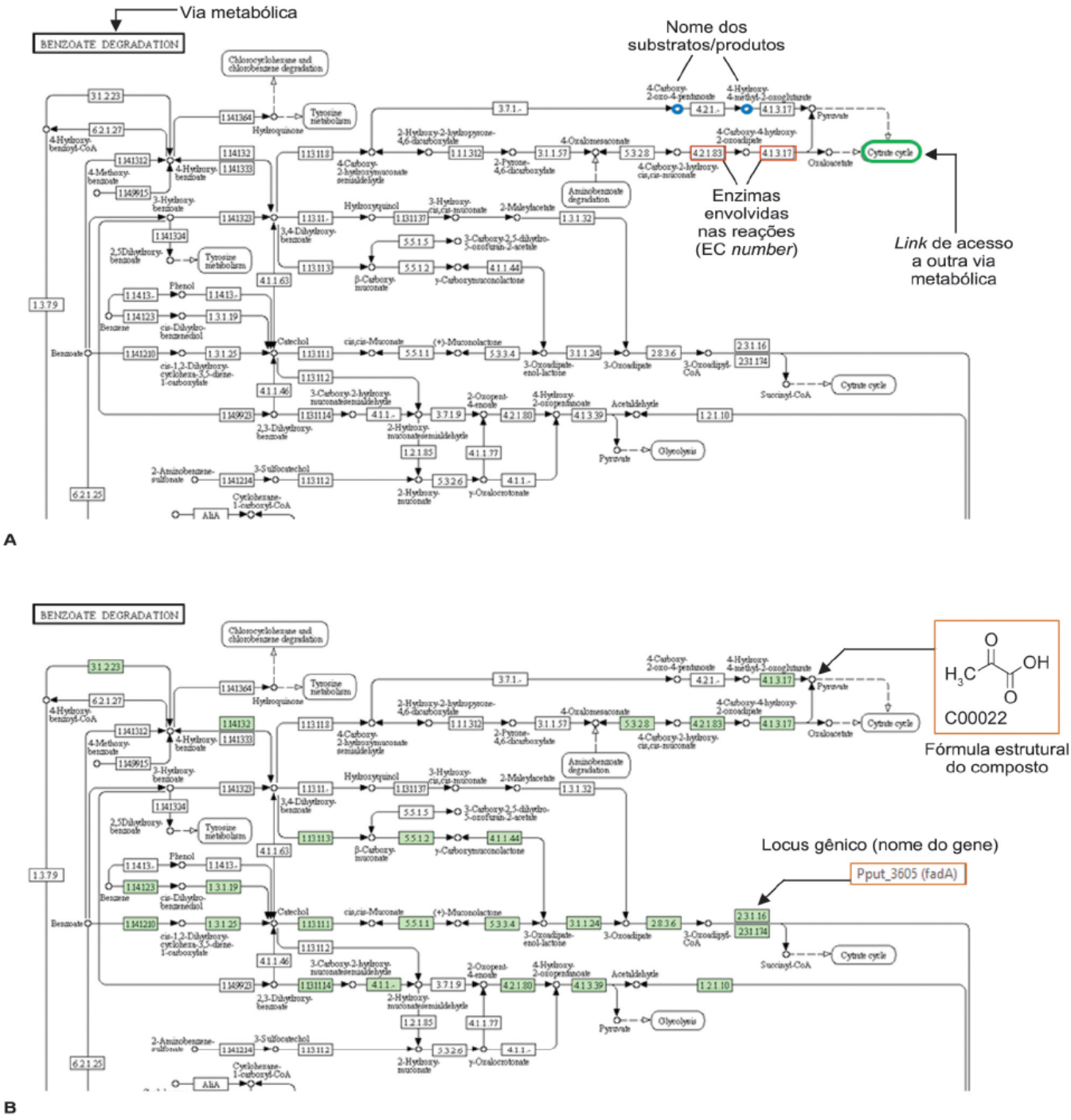


Figura 14.7 Visão de uma das vias metabólicas disponíveis no KEGG. Em **A**, destaca-se o mapa global da via associada com metabolismo de benzoato. Cada um dos pequenos círculos denota os respectivos substratos e produtos envolvidos nessa via. Os quadrados sinalizam as enzimas envolvidas nessas reações bioquímicas, assim como os números representam a classificação das respectivas enzimas de acordo com IUBMB. Em **B**, destaca-se a mesma via, selecionada agora para uma cepa de *Pseudomonas*. Em tom verde, evidenciam-se as enzimas cujos genes correspondentes foram encontrados no genoma da espécie. Caso o usuário clique nos círculos, a estrutura molecular será apresentada, para que ele compreenda o mecanismo de catálise. Caso clique nos códigos das enzimas, serão apresentados o *locus* gênico e o nome do gene no genoma da espécie selecionada.

EXPASY

Embora não se trate de um banco de dados, o EXPASY reúne uma série de ferramentas computacionais e possibilita que o usuário navegue por diversos repositórios de dados para análise de informações biológicas de acordo com seus interesses. Doze são as categorias do EXPASY que proporcionam acesso a bancos de dados e ferramentas de análise de informações biológicas: proteômica, genômica, bioinformática estrutural, biologia de sistemas, análises filogenéticas e evolução, genética de populações, transcritômica, biofísica, de busca e uso de imagens, em infraestrutura de tecnologia da informação e desenho de drogas (Figura 14.9). Para cada uma dessas opções, o

EXPASY fornecerá um conjunto de bancos de dados correspondente, bem como um conjunto de ferramentas que permitem a análise de dados biológicos nessa respectiva área de escolha.

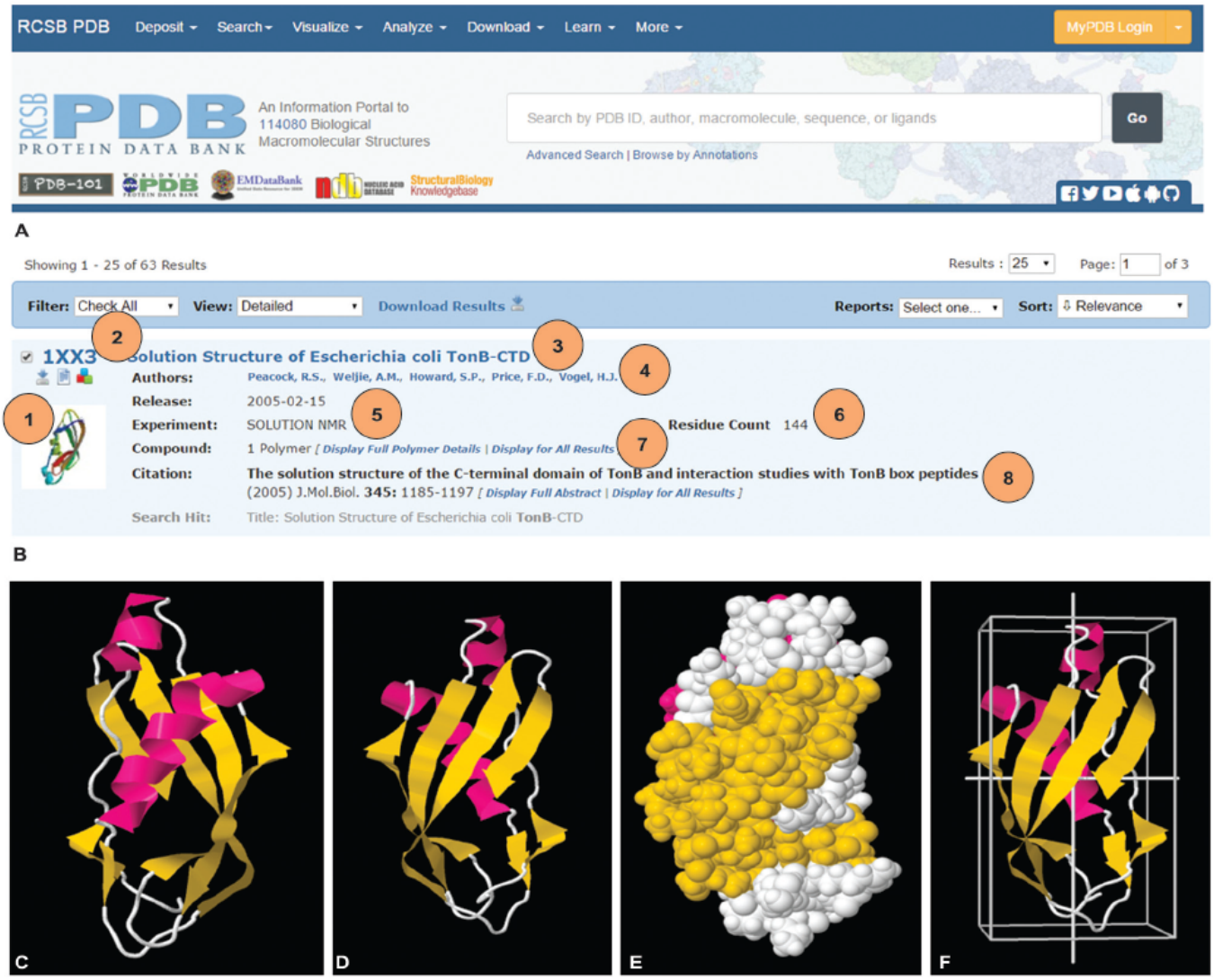


Figura 14.8 Estrutura da interface e dos principais recursos do PDB. **A.** Porção superior da ferramenta disponibiliza a interface de apresentação do PDB, com destaque para a barra de busca de informações a partir da entrada de palavras-chave. **B.** Uma vez feita a consulta, informações de destaque para um dos resultados encontrados a partir da busca serão apresentados, por exemplo, para a palavra TonB (sigla para uma proteína associada à membrana em procariotos): 1 – modelo da estrutura; 2 – código da proteína; 3 – organismo-modelo e nome da proteína; 4 – autores da pesquisa; 5 – data e metodologia usada; 6 – tamanho da proteína; 7 – informações gerais; 8 – veículo de publicação. **C** a **F.** Diferentes modelos visuais da mesma proteína, a partir da seleção de algumas opções que o PDB disponibiliza via Jmol (programa de visualização da estrutura tridimensional).

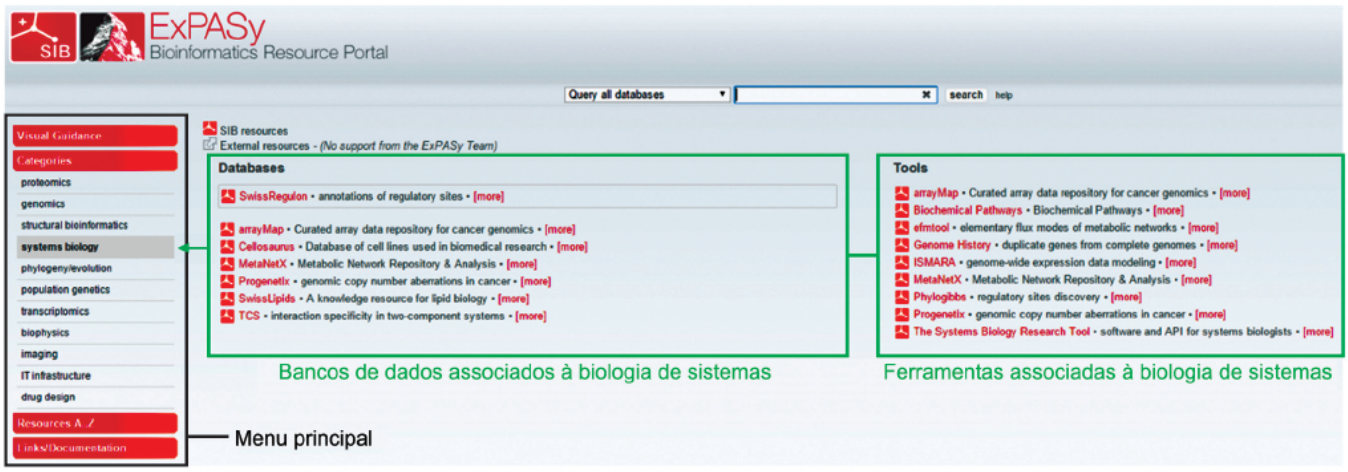


Figura 14.9 Interface do EXPASY. À esquerda, destaca-se o menu de opções com as categorias citadas no texto. Para cada uma delas, duas listas se abrem na parte central da página, uma incluindo os bancos de dados, e outra, as ferramentas associadas à temática escolhida.

Análise de genomas

Sequenciamento de genomas

Desde o seu desenvolvimento, em 1977, as metodologias de sequenciamento de DNA evoluíram muito.[1] Atualmente, o sequenciamento de genomas se tornou uma rotina em muitos laboratórios de genética e ganha cada vez mais espaço em outras áreas, tais como a medicina e biotecnologia.[2] Se, nos anos 1980, o sequenciamento de um genoma completo de um microrganismo demorava dezenas de meses ou anos para ficar pronto, hoje conseguimos sequenciar um genoma muito maior (como o genoma humano) em menos de 2 semanas (ver Capítulo 12). Concomitantemente ao aumento da velocidade de sequenciamento, ocorreu uma drástica redução no custo dessa tecnologia. No final da década de 1980, o valor para sequenciar um nucleotídeo (nt) estava por volta de US$ 1,00; hoje são sequenciados mais de 10 milhões de nucleotídeos por US$ 1,00, e esse valor tende a cair ainda mais nos próximos anos.[5] A Figura 14.10 ilustra a queda do valor do sequenciamento.

Atualmente, a dificuldade metodológica para sequenciar moléculas de DNA está superada, sendo o grande desafio armazenar, organizar e, sobretudo, processar (estudar ou analisar) as sequências geradas. Por exemplo, as metodologias de sequenciamento mais utilizadas no momento produzem leituras individuais (*reads* ou sequências) de 50 a 150 nt., tamanho considerado curto em comparação com a extensão do genoma de qualquer organismo vivo, a qual é de 3 a 9 ordens de grandeza maior do que essas sequências; esse tamanho também é considerado curto em comparação com o tamanho médio de um gene qualquer que, para humanos, tem tamanho (médio) de 1.200 nt.[6] Como consequência, organizar essas leituras geradas na sua ordem correta e reconstituir um cromossomo ou mesmo um genoma completo (um processo conhecido como montagem genômica), assim como anotar seus genes, exons e introns, são processos bastante complexos que requerem um grande poder computacional e, muitas vezes, passos demorados de intervenções manuais de especialistas. Na Figura 14.11 há um esquema simplificado do processo de sequenciamento, montagem e anotação de um genoma.

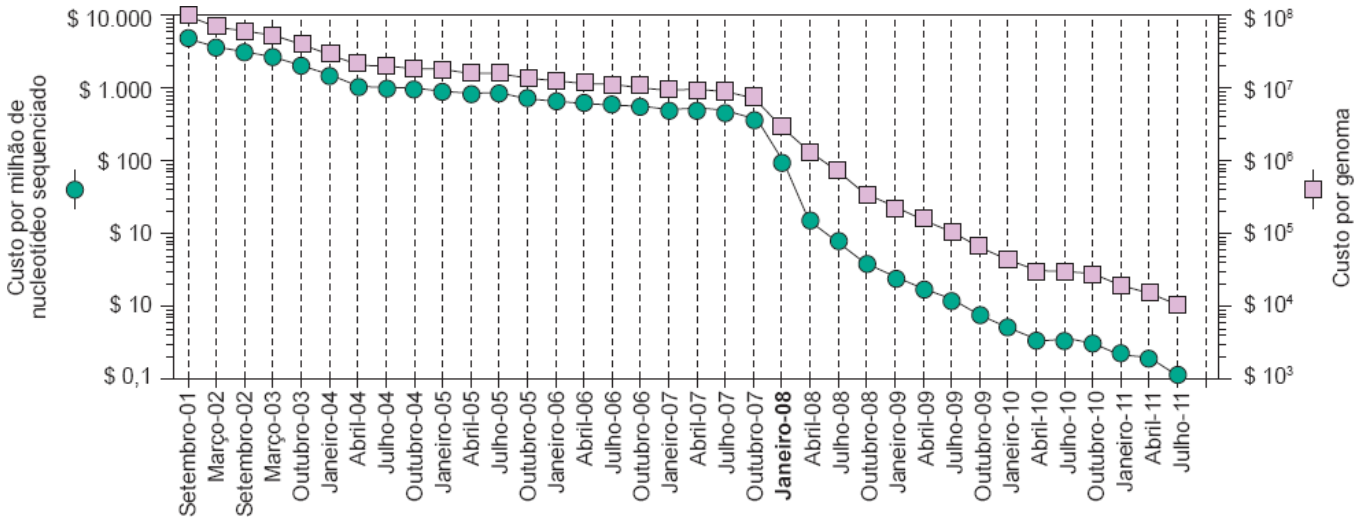


Figura 14.10 Queda no custo de sequenciamento de DNA. Em 10 anos, houve uma queda de, aproximadamente, 100 mil vezes no valor por nucleotídeo sequenciado. O custo por genoma refere-se ao valor aproximado para sequenciar o genoma humano. Adaptada de Li *et al.*, 2011.[5]

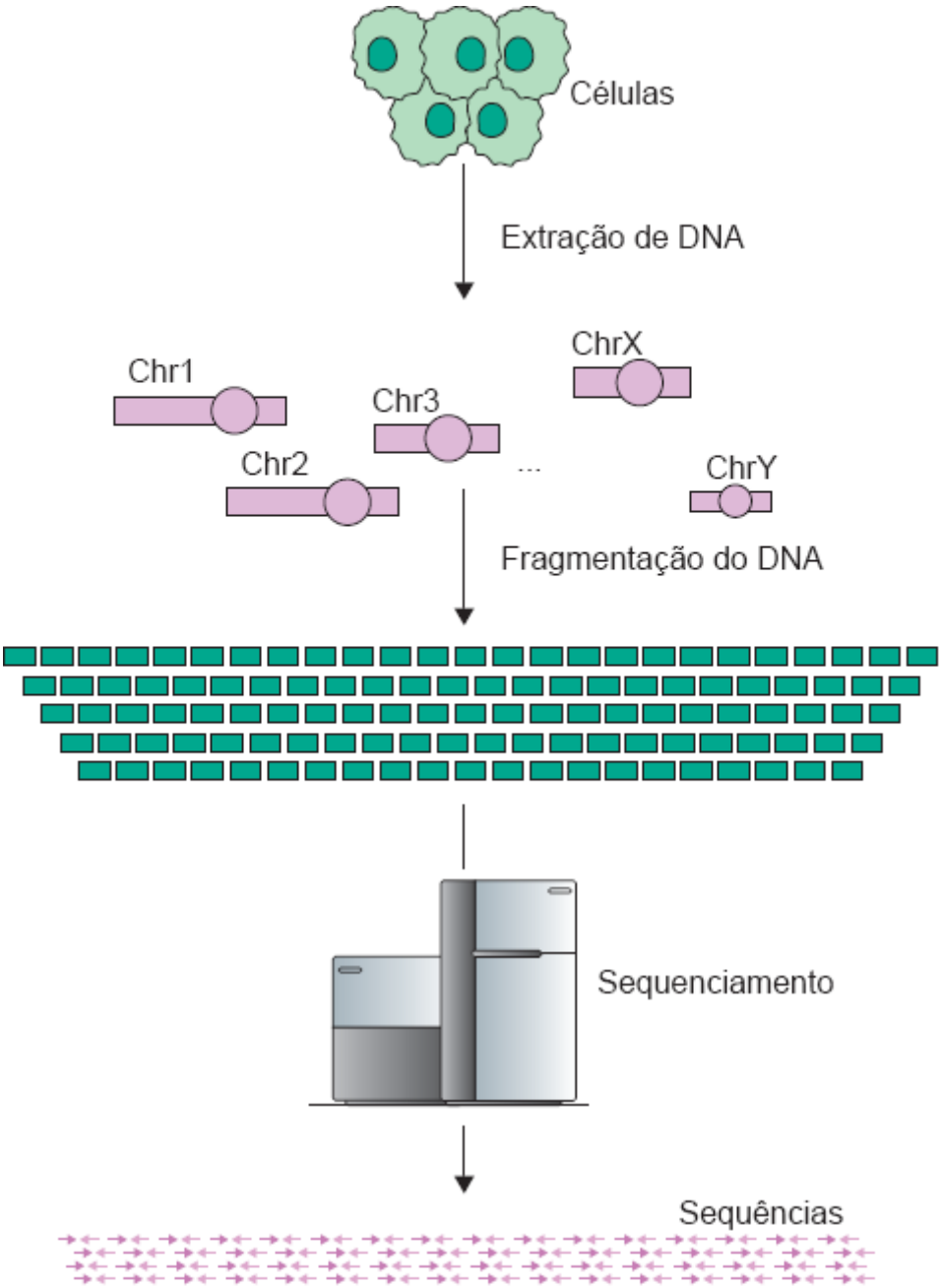
Uso de bioinformática na montagem de genomas

O maior desafio durante o processo de sequenciamento de um genoma completo é a organização linear (montagem) das leituras (sequências ou *reads*) geradas. A principal razão desse desafio é a enorme similaridade entre diferentes regiões do mesmo genoma, algo que ocorre em praticamente todos os organismos, com destaque especial para um grande conjunto de plantas e mamíferos.

Essencialmente, existem duas abordagens que podem ser usadas para realizar a montagem dos milhões (ou bilhões) de *reads* produzidos pelos sequenciadores modernos: a montagem *de novo* do genoma ou a montagem

guiada por um genoma de referência. Os métodos de montagem *de novo* são bastante custosos ("pesados") do ponto de vista computacional, e tentam produzir uma sequência-consenso com base apenas na coleção de *reads* coletados no experimento. Os métodos baseados em sequências de referência tendem a ser mais "leves" (requerem menos memória e capacidade de processamento computacional), mas dependem da disponibilidade de uma sequência genômica apropriada contra a qual os *reads* são alinhados (ancorados) e depois montados. Tal sequência de referência é, no geral, da mesma espécie ou de espécies muito próximas evolutivamente, e terá sido sequenciada e montada previamente, sem o auxílio de uma referência.

Em relação aos métodos de montagem *de novo*, os principais algoritmos de montagem baseiam-se na busca por regiões de sobreposição entre os *reads*. Para que essa sobreposição ocorra, os protocolos de sequenciamento incluem passos que fragmentam o DNA de modo aleatório, objetivando uma distribuição uniforme de fragmentos ao longo do genoma sequenciado. Por exemplo, usando *reads* de 150 pares de bases (pb), a montagem de 3 bilhões de pb vai precisar de cerca de 1 bilhão de *reads* com sobreposição para ser completada e, certamente, algumas semanas ou meses de processamento em bons computadores. A Figura 14.11 ilustra simplificadamente esse processo de fragmentação e montagem genômica. Algoritmos com o Phrap e o CAP3 funcionam bem para montagem de genomas sequenciados por métodos que geram leituras maiores (sequências de 500 a 1.000 nt). Porém, com o advento dos sequenciadores de segunda geração, a partir de 2007, o tamanho dos *reads* (de 35 a 150 nt) e a quantidade de dados gerada (na ordem de alguns milhões de *reads*) motivaram o desenvolvimento de novos algoritmos para montagem de genomas que pudessem tratar essas novas peculiaridades, tais como Velvet, EULER e ALLPATHS.



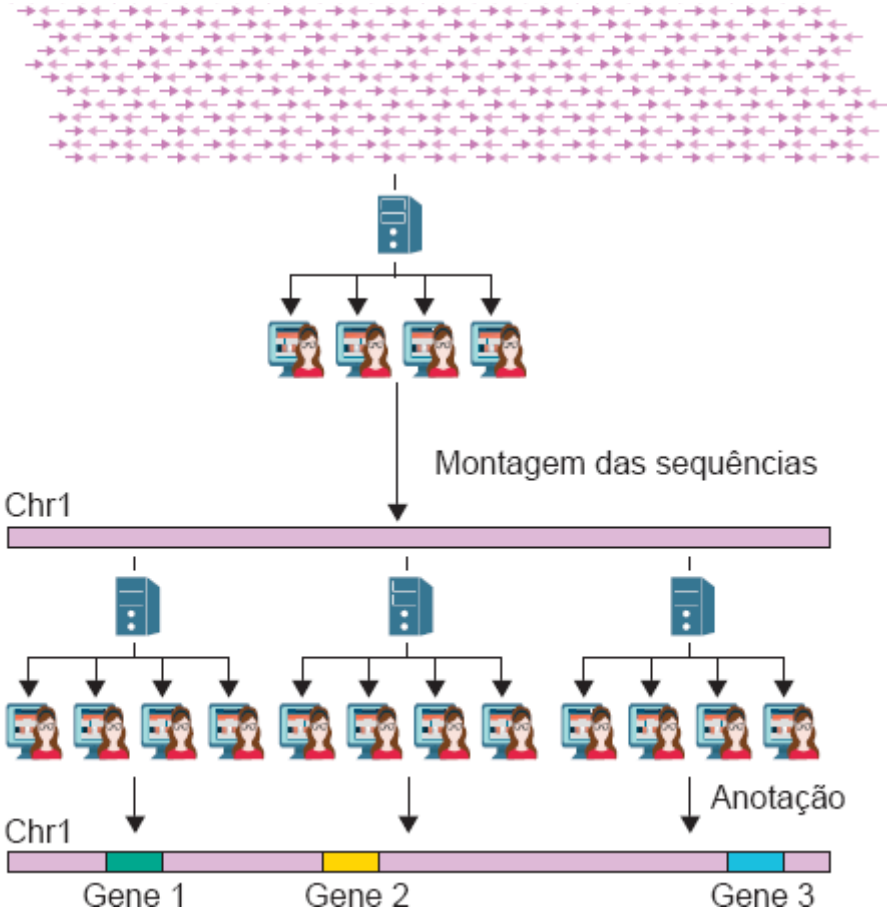


Figura 14.11 Esquema simplificado do processo de extração de DNA, sequenciamento, montagem e anotação genômica.

Atualmente, dada a grande quantidade de genomas sequenciados, montados e já disponíveis publicamente, a estratégia de fazer uma montagem guiada por um genoma de referência é bastante utilizada, em especial para genomas grandes e complexos (em termos de regiões repetitivas), como o nosso e de outros mamíferos. Nesse passo, após o sequenciamento, os *reads* gerados são todos alinhados contra o genoma de referência da espécie correspondente. Dessa maneira, o genoma de referência funciona como uma espécie de âncora, e podem ser observadas as regiões cobertas (com *reads*) e não cobertas (sem *reads*) pelas sequências geradas. No final, identificam-se as similaridades e as diferenças entre os genomas sequenciados e o genoma de referência, gerando, assim, uma montagem do genoma recém-sequenciado. Por exemplo, atualmente, os projetos de (res)sequenciamento de genoma humano (e da maioria dos mamíferos já sequenciados) majoritariamente utilizam essa estratégia de montagem guiada em aplicações como as que serão descritas nas seções a seguir.

Uso da bioinformática na anotação genômica

Após a etapa de montagem de um genoma, inicia-se, no geral, um trabalho de anotação dos seus genes, de suas regiões transcritas, regiões reguladoras e regiões polimórficas. Dependendo do organismo, o processo de anotação pode ser bastante complexo, demorado e custoso do ponto de vista computacional e de recursos humanos. Por exemplo, em princípio, localizar um determinado gene em uma sequência genômica deveria ser algo simples, pois um gene apresenta, no início de sua região codificadora (ou ORF, do inglês *open reading frame*), um códon inicial (ATG) e, no seu final, um códon de parada ou terminal (TAA, TAG ou TGA). Para a maioria dos genomas bacterianos, identificar uma ORF é uma tarefa relativamente simples, a qual diversas ferramentas computacionais, como os programas Glimmer e GeneMark, fazem muito bem, apesar da dificuldade de identificar corretamente o códon de iniciação. Entretanto, identificar tais ORF em genomas de eucariotos é um processo muito mais complicado, pois a maioria dos genes apresenta introns, que separam as ORF em diversos exons, e não há um padrão específico de tamanho gênico ou número de exons. Mesmo do genoma humano, alvo de muitos estudos, ainda não conhecemos o número exato de genes, seus exons e introns. Portanto, em genomas de eucariotos, identificar as ORF, exons e introns é um processo bastante difícil, para o qual ainda não existem algoritmos ótimos. Alguns algoritmos que funcionam muito bem são: GenScan, TWINSCAN e NSCAN e JIGSAW.[7]

Uma maneira adicional de auxiliar na anotação de um genoma é comparar suas sequências com um genoma já anotado.[8] Esse passo de anotações por similaridade se baseia em assumir que genes com sequências similares são homólogos, ou seja, descendem de um mesmo gene ancestral e, portanto, têm grandes chances de preservar a função do gene ancestral. Por exemplo, o processo de anotação do genoma de camundongo se beneficiou (e ainda se beneficia) das anotações do genoma humano, e vice-versa. Mesmo anotações de genes de organismos distantes, tais como *Drosophila melanogaster* e *C. elegans*, podem ser usadas para auxiliar a anotação do genoma de mamíferos, por exemplo. O grande problema é quando ocorrem anotações erradas. Se isso acontecer, uma anotação errada pode se propagar para diversos organismos, causando um erro em cascata, e, infelizmente, isso não é tão raro de ocorrer. Muitos dos programas que fazem predições de genes também utilizam dados de conservação entre espécies.

Uma estratégia cada dia mais utilizada é a de fazer uma anotação manual de genomas. A iniciativa mais importante é a chamada ENCODE (do inglês, *Encyclopedia of DNA Elements*). O ENCODE é um consórcio formado por diversos grupos de pesquisa que tem o objetivo de anotar todos os elementos funcionais presentes no genoma humano.[9] Para isso, os participantes utilizam diversos métodos que reportam a expressão gênica e a localização de suas regiões regulatórias, muitas ferramentas de bioinformática para integrar esses dados e um grande número de pessoas para olhar cuidadosamente cada informação e produzir uma anotação bastante precisa da parte funcional do genoma humano. Por exemplo, esse projeto afirma que cerca de 75% do genoma humano é transcrito. Por outro lado, vários autores questionaram esse e outros resultados provenientes do ENCODE, explicitando a complexidade e a dificuldade de se fazer uma boa anotação genômica.[10]

Uso da bioinformática na busca de variações genômicas

Desde a descoberta da estrutura do DNA, em 1953, houve um desenvolvimento rápido e significativo, se comparado com o de outras áreas do conhecimento, de todo o arcabouço metodológico para manipular e estudar o DNA.[6] Utilizando essas técnicas, consegue-se saber, por exemplo, a sequência, base a base, de uma região genômica ou de um genoma completo. Tais informações também podem ser comparadas e, por meio dessas comparações, é possível identificar as similaridades e as diferenças (variações genéticas) entre duas sequências de DNA. A identificação de tais variações genéticas é algo fundamental para diversas áreas, como evolução, estudo populacional e estudo de doenças ou resistência a doenças causadas por tais variações.

Nossa espécie, *Homo sapiens*, é bastante jovem (cerca de 200 mil anos de idade), o que não nos deu muito tempo, em comparação com outros organismos, para acumular variações genéticas. Com exceção de gêmeos idênticos – considerados clones naturais –, são as variações genéticas (ou genômicas, como alguns autores gostam de chamar) que determinam a unicidade de cada ser humano, assim como uma boa parte de suas características fenotípicas.

O genoma haploide humano tem aproximadamente 3 bilhões de pares de bases (na forma haploide, ou 6 bilhões na forma diploide), de modo que quaisquer dois indivíduos não relacionados apresentam cerca de 99,9% de sequências idênticas, isto é, aproximadamente 1 variação a cada 1.000 pb. Portanto, temos cerca de 6 milhões de pares de bases (já que o nosso genoma é diploide) de diferença entre dois indivíduos. O interessante é que (apenas) cerca de 15% dessas variações ocorrem entre populações; a maioria (85%) das nossas variações ocorrem dentro de uma mesma população, ressaltando a mistura genética entre indivíduos da nossa espécie. O mesmo ocorre para outras espécies, algumas com mais e outras com menos variações. Nessa estimativa, consideram-se apenas variações de uma única base, porém existem outras variações no DNA que podem ser deleções ou inserções de um conjunto de pares de base ou mesmo aberrações cromossômicas. Tais variações são menos frequentes que as variações de uma única base.

Graças às novas metodologias de sequenciamento de DNA (e as plataformas de microarranjos), identificar essas variações se tornou algo tangível, rápido e com um custo financeiro relativamente baixo.[1] Existem diversas maneiras, do ponto de vista da bioinformática, de identificar as variações genéticas. Por exemplo, partindo de um experimento de microarranjos, é necessário um determinado conjunto de programas, com taxas de identificações de falsos-positivos e negativos específicas dessa técnica e diferentes de quando se parte de uma metodologia de sequenciamento de DNA, seja de primeira ou de segunda geração. Atualmente, as metodologias de sequenciamento de DNA são o “padrão-ouro” para a identificação de variações genéticas, com baixo custo e alta sensibilidade e especificidade em comparação com quaisquer outras metodologias.

Principais ferramentas web para análise de genomas

Em geral, todo processo de estudo de genomas é feito por meio de informações armazenadas em bases de dados e ferramentas computacionais. Ferramenta computacional é o nome genérico para um algoritmo desenvolvido com o intuito de realizar uma ou mais tarefas específicas. A seguir, serão descritos detalhes de algumas muito utilizadas, porque úteis, na análise de diversos dados.

Alinhamento de sequências

O alinhamento de sequências (de DNA, RNA ou proteína) figura entre as metodologias mais utilizadas e mais importantes da bioinformática. Essencialmente, o alinhamento é uma comparação, elemento a elemento (pode ser um nucleotídeo ou um aminoácido), entre duas ou mais sequências. Por meio desse processo, podem-se identificar as similaridades e as diferenças nucleotídeo a nucleotídeo (ou aminoácido a aminoácido, por exemplo) entre sequências e inferir diversas informações, tais como possível função ou perda de função, origem evolutiva, qualidade do sequenciamento, entre outras. Como existem inúmeras ferramentas computacionais que fazem alinhamento de sequências, é essencial saber escolher a mais adequada para a questão de interesse. Neste capítulo, serão apresentados mais detalhes do BLAST, ferramenta que figura entre os *softwares* mais importantes e mais utilizados pela comunidade científica mundial. Dificilmente existem alunos de pós-graduação na área de Ciências Biológicas que nunca tenham feito ao menos um alinhamento utilizando o BLAST.

BLAST

O desenvolvimento do BLAST (do inglês, *basic local alignment search tool*) em 1990 e sua atualização em 1997 foram marcos para a Bioinformática.[11,12] Esse algoritmo possibilitou um grande aumento na velocidade das buscas em bancos de dados de sequências, as quais, até então, eram feitas por alinhadores como o Smith-Waterman e o FASTA, ambos mais lentos e com a necessidade de mais recursos computacionais do que o BLAST.[13,14] Essencialmente, o passo que torna o BLAST rápido é a sua capacidade de identificar, primeiramente, regiões curtas e altamente similares entre duas sequências e de não "gastar" recursos computacionais em regiões com baixa probabilidade de serem alinhadas.

Esse algoritmo tem três passos principais. Primeiro ele compila uma lista de regiões pequenas e similares entre ambas as sequências. Em seguida, cada uma dessas regiões é analisada com cuidado, e os seus nucleotídeos adjacentes são comparados. Por fim, é calculado um valor que reflete a qualidade do alinhamento, chamado *E-value*. Quando esse valor está abaixo do valor máximo definido pelo usuário, o alinhamento é incluído no resultado final; caso contrário, ele é descartado. Independentemente do resultado, inicia-se a comparação de uma outra região similar, caso exista, entre as duas sequências.

Existem diversos portais que realizam alinhamentos utilizando o BLAST. O mais comum e mais utilizado é o portal do NCBI (Figura 14.12), onde é possível realizar alinhamentos entre sequências de nucleotídeos, proteínas, nucleotídeos traduzidos, entre outras opções. O alinhamento de uma sequência pode ser feito contra um determinado genoma (p. ex., o genoma humano), contra um conjunto de sequências de centenas de milhares de organismos (banco NR) ou contra uma única sequência (utilizando o BLAST2 seq). Caso o leitor tenha uma sequência desconhecida (de nucleotídeo ou proteína), alinhar essas sequências contra os bancos de dados desse portal (em especial o NR) é o melhor ponto de partida para coletar mais informações sobre ela.

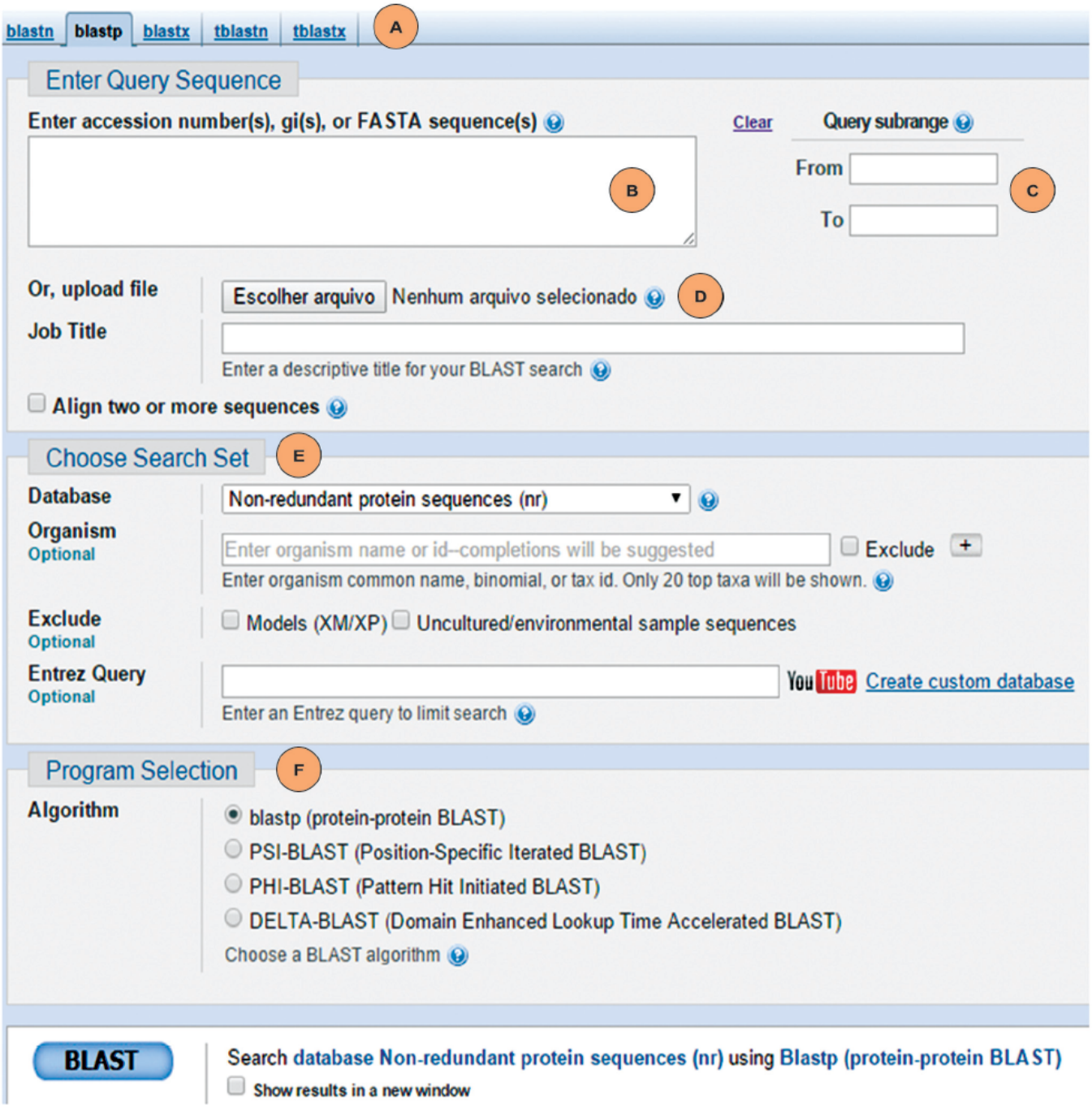


Figura 14.12 Interface da ferramenta BLAST via portal disponível no NCBI. **A.** Note que a plataforma de busca está assinalada para execução da opção *blastp*, ou seja, tanto a sequência-alvo como o banco de dados são de proteínas (ver Tabela 14.1), embora possa ser modificada para outras opções. **B.** O usuário deverá adentrar com a sequência ou identificador do genoma no campo. **C.** Em ambos os casos, o usuário poderá inserir as posições de início e fim da sequência do genoma nas opções *Query subrange* para restringir a busca àquela região. **D.** O usuário também poderá importar o arquivo que contém a sequência de interesse, no formato FASTA. **E.** Opções padronizadas para execução da ferramenta podem ser modificadas de acordo com o interesse do usuário. **F.** O mesmo é atribuído para a seleção do programa, em que variações específicas do algoritmo do BLAST (PSI-BLAST, MEGABLAST) podem ser selecionadas.

O BLAST também foi adaptado para o processamento de sequências de DNA e proteínas, bem como diferentes combinações desses alfabetos. A Tabela 14.1 lista as principais variações do método BLAST. A escolha da implementação do BLAST mais adequada para uma certa análise depende da sequência que o usuário tem em mãos e do seu banco de dados. Podem-se utilizar diferentes implementações (p. ex., BLASTX ou TBLASTX) para buscar sequências homólogas, pois, nesses casos, a busca se dá por sequências proteicas, inferidas a partir da tradução conceitual das sequências de nucleotídeos fornecidas pelo usuário, e não da própria sequência de nucleotídeos. Essa estratégia de usar, nas buscas por homólogos distantes, a tradução de uma sequência de nucleotídeos para a sequência de aminoácidos correspondente é vantajosa, pois, como muitos códons codificam o mesmo aminoácido, as sequências de nucleotídeos tendem a mudar muito mais rapidamente do que as sequências de proteína,

acumulando rapidamente mutações sinônimas, ou seja, que não alteram os aminoácidos codificados. Tais mutações sinônimas tornam as sequências de DNA, originalmente muito semelhantes entre si, tão distintas que a relação de similaridade entre elas se torna indetectável. Em tais circunstâncias, as proteínas serão capazes de revelar homólogos mais distantes por preservarem sua sequência de aminoácidos e por consistirem em um alfabeto mais extenso (20 aminoácidos contra 4 bases).

Na interface de entrada das informações, basta que o usuário insira sua sequência em formato FASTA, defina os parâmetros desejados (ver Figura 14.12) e execute o alinhamento. Como resultado, a ferramenta retornará um sumário das informações associadas à sequência de entrada (Figura 14.13 A), um perfil figurativo do alinhamento (Figura 14.13 B), uma tabela que sumariza as características do alinhamento (Figura 14.13 C). bem como a comparação base a base, ou aminoácido a aminoácido, para que o usuário consiga visualizar o grau de similaridade entres as sequências de entrada e com as quais a entrada foi comparada (Figura 14.13 D).

Na Figura 14.13 D, destaca-se a representação explícita do alinhamento, resíduo a resíduo, ressaltando as posições de início e fim para cada uma das sequências. As linhas entre as sequências *query* e *subject* representam os três resultados possíveis na comparação resíduo a resíduo: quando são idênticos os resíduos da *query* e do *subject*, o programa repete a abreviatura do resíduo entre as sequências (identidade); quando são distintos, mas apresentam cadeias laterais com propriedades semelhantes, é inserido um sinal de + entre as sequências (indicando resíduos similares); e quando são muitos distintos, é inserido um espaço em branco.

É importante destacar que, depois do alinhamento, o usuário precisa saber interpretar o resultado obtido, e esse passo não é simples, pois requer, ao menos, conhecimento sobre a nomenclatura utilizada pelo BLAST, além de outros conceitos relacionados ao alinhamento de sequências. No Quadro 14.1 estão resumidas as principais terminologias presentes em um resultado de alinhamento com o BLAST.

Clustal

Diferentemente do BLAST, que resulta sempre em uma comparação par a par (envolvendo duas sequências), o Clustal é um programa que constrói o alinhamento entre diversas (múltiplas) sequências. As comparações feitas por ele envolvem todas as sequências dadas como entrada, e o resultado é um único alinhamento múltiplo de sequências. Alinhamentos múltiplos são frequentemente usados para evidenciar padrões em sequências e são um pré-requisito para caracterizar ou agrupar possíveis famílias gênicas usando análises filogenéticas. Em outras análises, auxilia pesquisadores a predizerem estruturas secundárias de proteínas, identificarem regiões em potencial para a síntese de oligonucleotídeos iniciadores de reações de PCR ou mesmo predizerem resíduos importantes para a atividade de uma biomolécula.

Tabela 14.1 Implementações do alinhador BLAST.

Programa	Sequência-alvo	Banco de dados
BLASTP	Proteína	Proteína
BLASTX	DNA (traduzido)	Proteína
BLASTN	DNA	DNA
TBLASTN	Proteína	DNA (traduzido)
TBLASTX	DNA (traduzido)	DNA (traduzido)

O Clustal constrói alinhamentos múltiplos de sequências pelo método chamado alinhamento progressivo. Esse método é, geralmente, dividido em três etapas (Figura 14.14 A). Primeiramente, após entrada das múltiplas sequências (Figura 14.14 B), o programa gera uma matriz de distância a partir dos dados de entrada. Em cada célula dessa matriz, será adicionado um valor que reflete, para cada par de sequências, o número de substituições de resíduos necessárias para converter uma sequência na outra. Em seguida, uma árvore, denominada árvore-guia, é deduzida a partir da matriz de distâncias, pelo repetido agrupamento das sequências mais próximas (menor distância; Figura 14.14 C). Finalmente, as sequências são alinhadas na ordem em que aparecem na árvore-guia, e é constituído o alinhamento múltiplo final (Figura 14.14 D).

A qualidade do alinhamento dependerá do grau de identidade entre as sequências dadas como entrada, sendo sempre mais difícil obter um bom alinhamento múltiplo de sequências muito divergentes entre si. Na saída padrão do Clustal, toda vez que um nucleotídeo ou aminoácido (dependendo da entrada) for conservado em todas as sequências, será colocado um asterisco na linha de consenso, possibilitando que essa conservação seja rapidamente visualizada.

STRING

O STRING é uma ferramenta que possibilita que o usuário determine uma rede de possíveis interações entre proteínas e/ou genes (Figura 14.15). Tais interações são representadas como redes (também chamadas grafos; ver item Biologia de sistemas). Nessas redes, os alvos da interação (proteínas ou genes) são representados por nós, e as interações entre eles, como coexpressão ou a formação de complexos dentro da célula, são representadas por arestas (linhas que unem os nós). Essas redes de interação são determinadas pela análise de sete tipos diferentes de informação, extraídas da anotação de genomas completos, de bancos de dados externos e da literatura:

Quadro 14.1 Principais terminologias em resultado de alinhamento com BLAST.

Query: cada uma das sequências fornecidas pelo usuário para a procura por sequências semelhantes no banco de dados
Subject: cada uma das sequências semelhantes à query e identificadas no banco de dados. Um sinônimo muito usado é *hit* (resultado)
Match: a quantidade de elementos (nucleotídeos ou aminoácidos, por exemplo) compartilhados, ou seja, idênticos ou parecidos, nas duas sequências que foram alinhadas. Sinônimo: *positives* (BLAST)
Gaps: espaços inseridos entre os resíduos de uma sequência para colocar esses resíduos na mesma coluna de resíduos equivalentes de outra sequência. Tais espaços correspondem a eventos de inserção ou deleção, que são exemplos comuns de mutações
Score: uma medida do quanto é semelhante um par de sequências (Query e Subject). O score é calculado com base no número de matches e outras características do alinhamento produzido pelo BLAST
Expected value: um valor que descreve a chance de encontrar o alinhamento produzido pelo BLAST apenas por chance. Quanto menor o *E-value*, ou seja, quanto mais próximo de zero, maior a chance de o alinhamento ser distinto do que se pode obter alinhando um par de sequências qualquer

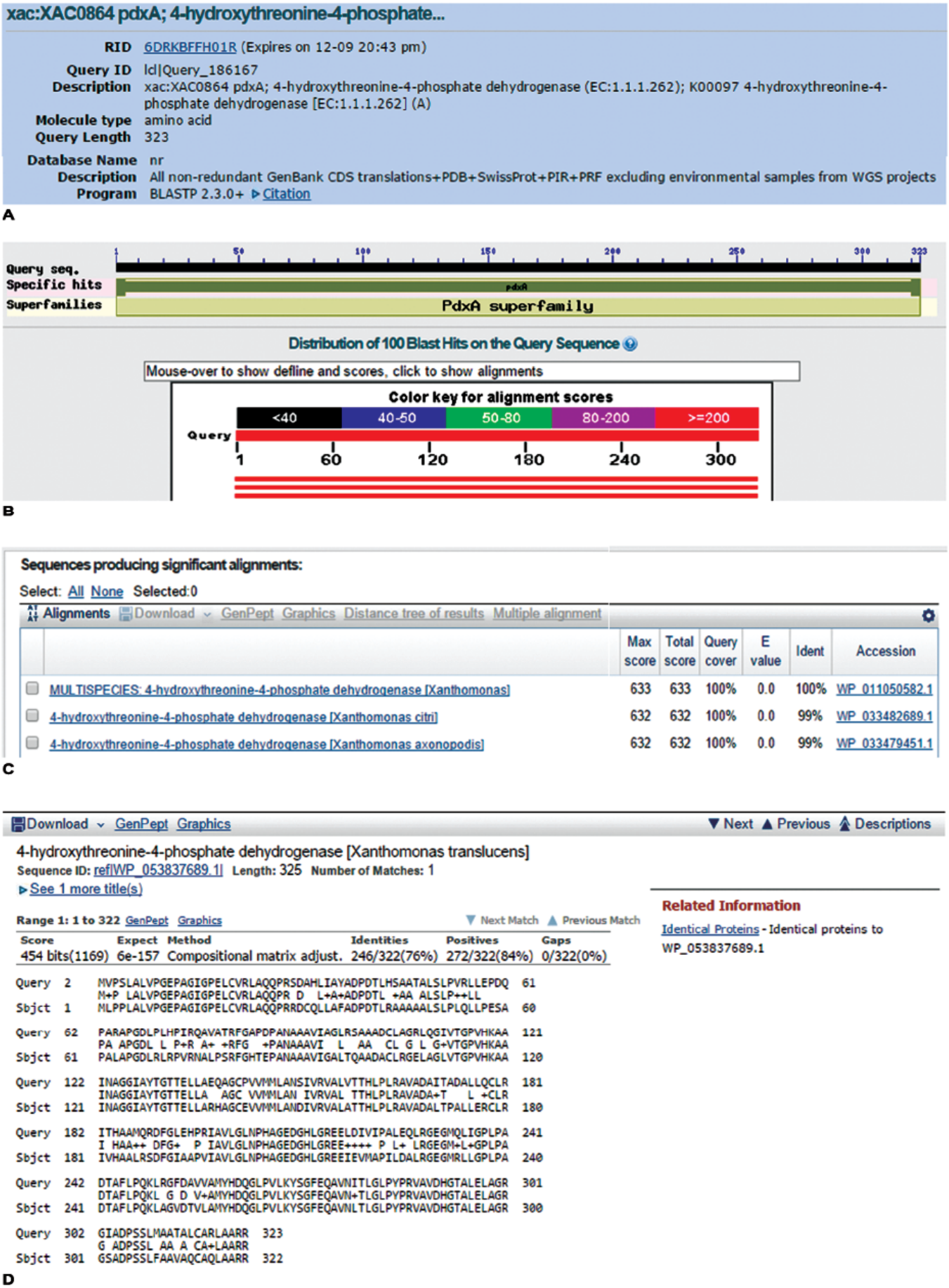


Figura 14.13 Resultados de alinhamento usando o programa BLAST, opção *blastp*. **A.** Sumário das informações de entrada, obtidas a partir da sequência FASTA. **B.** Resumo gráfico dos melhores alinhamentos. Nota-se que a sequência se alinha, com 100% de cobertura, com os melhores resultados, e estes fazem parte de uma família de proteínas já conhecida. O grau de identidade é alto, uma vez que as sequências *subject* estão representadas em tom vermelho (valores de *score* ≥ 200). **C.** Síntese das informações para cada alinhamento resultante da busca, incluindo a proporção do comprimento da sequência-alvo inserida no alinhamento (denominado *Query cover*), o número de

resíduos idênticos na região alinhada (denominados *Ident*), o *Score* e o *E-value*. **D.** Alinhamento da sequência-alvo com uma das sequências *subject*.

Figura 14.14 Clustal. **A.** Fluxograma das etapas envolvidas na geração do alinhamento múltiplo. **B.** Interface de entrada e execução do alinhamento múltiplo na versão web do programa ClustalW. 1: escolha do método aproximado (mais rápido) ou preciso (mais lento); 2: seleção dos dados de entrada, aminoácidos (proteínas) ou nucleotídeos (DNA); 3: janela de entrada das sequências em sete formatos suportados (FASTA etc.); 4: opção de carregar as sequências de um arquivo no computador do usuário. **C.** Árvore-guia gerada a partir dos dados de entrada. **D.** Alinhamento múltiplo. Nota-se que cada sequência de entrada ocupa uma linha de cada bloco de três linhas, sendo a última linha um sumário das análises. "*" denota que os resíduos da coluna são idênticos; ":" aponta que os resíduos da coluna são altamente conservados; "." significa que os resíduos da coluna são pouco conservados; " " (espaço) caracteriza que não há conservação de resíduos.

- Análise da conservação de genes vizinhos (verde)
- Fusão de genes (vermelho)
- Co-ocorrência (azul)
- Coexpressão (marrom)
- Resultados de experimentos para detecção, em larga escala, de interações físicas entre proteínas (rosa)
- Bancos de dados externos, como o KEGG, que descrevem interações entre genes e/ou vias e sistemas biológicos completos e cuja informação tenha sido trabalhada manualmente por revisores (azul-claro)
- Processamento automático de textos, cuja finalidade é detectar a recorrência de citações dos nomes dos mesmos conjuntos de genes e proteínas em um grande número de artigos científicos (bege).

Dados sobre a presença de homólogos (*Homology*, em lilás na Figura 14.15) também são exibidos, mas não são usados no cálculo da qualidade da interação (*score*).

Para que o grafo seja gerado, basta que o usuário digite o nome de um gene, seu *locus* de identificação ou mesmo a sequência de sua proteína de interesse. Quando a proteína tem alguma informação estrutural, no centro do nó será observada uma imagem representando a estrutura dessa proteína ou de um de seus domínios. É preciso

somente clicar sobre o nó para obter mais informações. Se o usuário desejar, basta aumentar ou diminuir o número de proteínas que serão incluídas no grafo, clicando em + ou -, respectivamente, no menu disponível abaixo dele. A partir dessa interface, o usuário poderá inspecionar os dados usados na inferência da rede para, por exemplo, verificar se a vizinhança do gene selecionado é conservada, se o gene que codifica a proteína de interesse aparece fusionado a outro gene em alguma espécie, se as proteínas no grafo também são observadas em outros genomas, o que diz a literatura sobre as funções e relações dos genes de interesse etc.

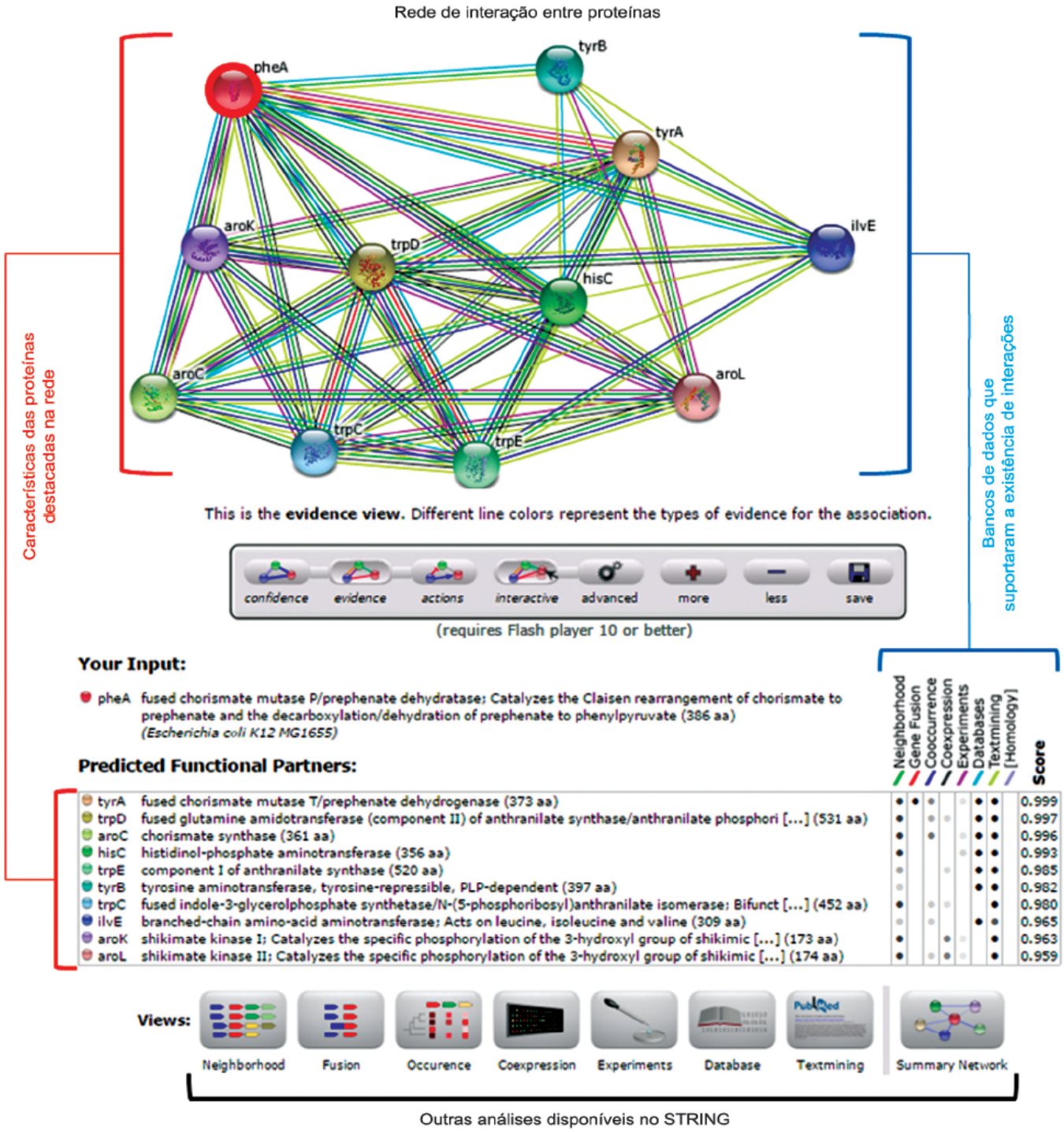


Figura 14.15 Resultados apresentados pelo STRING. A interface possibilita que o usuário identifique a informação que suporta a existência de interações entre as proteínas com base na cor da aresta (legenda à direita). Uma breve descrição da função das proteínas/genes presentes no grafo, extraída da anotação dos genomas incluídos no STRING, é exibida na tabela, e as outras ferramentas disponíveis são acessados pelos botões visualizados logo abaixo dessa tabela.

MEME

O MEME (Multiple EM for Motif Elicitation) é uma ferramenta que possibilita que o usuário faça busca por sequências curtas, chamadas motivos ou assinaturas, em conjuntos de sequências relacionadas a uma mesma função

biológica. Tais assinaturas são padrões conservados em sequências de nucleotídeos ou aminoácidos, e são preservadas e reconhecíveis por atuarem em um mesmo processo biológico, como sítios de ancoragem de fatores de transcrição ou sítios ativos em enzimas de uma determinada família. A partir de um conjunto de sequências de entrada, fornecido pelo usuário, o MEME procura padrões de sequências curtas e conservados, mas sem inserções ou deleções, e constrói um modelo que pode ser usado em análises posteriores de outras sequências para identificar o mesmo padrão. Essa ferramenta retorna um resultado em formato *WebLogo*, um gráfico onde a altura do resíduo é proporcional ao seu grau de conservação, o que possibilita que o usuário identifique os nucleotídeos ou aminoácidos (dependendo da entrada) mais conservados e que, portanto, definem a assinatura; retorna também informação sobre as posições dos motivos, o que possibilita, por exemplo, avaliar a distância entre as assinaturas e os genes-alvo em sequências promotoras (Figura 14.16).

MUMMER/MAUVE

Os programas MUMMER e MAUVE foram desenvolvidos com propósitos similares: comparar genomas e compreender como muda a ordem dos genes em um genoma. O MUMMER foi originalmente desenhado para produzir apenas alinhamentos de pares de sequências muito longas, mas não muito divergentes, como dois cromossomos bacterianos completos. Uma forma bastante útil de visualização de seus resultados é o clássico *dot-plot*, que já era usado desde as primeiras incursões no campo de alinhamento de proteínas, na década de 1970. Nesse gráfico, os eixos X e Y correspondem às sequências de entrada, se necessário linearizadas (lembre-se de que a maioria dos cromossomos e plasmídeos bacterianos é circular). A linearização é, em geral, feita com base em uma posição padrão (no caso de genomas, a origem de replicação), que é definida como posição 1, e, a partir dessa posição até o final do cromossomo, as bases são numeradas, definindo a escala em cada eixo do gráfico. O MUMMER compara os genomas a partir da detecção de sequências curtas e exatamente iguais (chamadas *n-mers*) em ambos os genomas, o que elimina a necessidade de um alinhamento explícito, como o produzido pelo BLAST. Uma vez localizados os *n-mers* em ambos os genomas, é inserido um ponto (representando o fragmento conservado) nas coordenadas correspondentes de cada genoma (Figura 14.17 A). Se os genomas apresentarem relativa conservação na ordem dos genes, denominada sintenia, o usuário visualizará uma espécie de linha (união de pontos) disposta em um ângulo de 45°. No caso de uma linha contínua, ele deverá entender que a organização gênica é a mesma ao longo de todo o genoma. Caso a comparação resulte em uma série de linhas, há, em um dos genomas, variações na ordem dos genes, que podem ser decorrentes de inserções, deleções ou rearranjos cromossômicos (*i. e.*, quando uma parte do genoma muda de lugar). Outro recurso do programa é que, quando os genes estão sintênicos e em um mesmo sentido em relação à origem, a cor é diferente da mesma sintenia observada em sentidos opostos à origem. Esse resultado enfatiza sobretudo o resultado de rearranjo estrutural cromossômico (Figura 14.17 A).

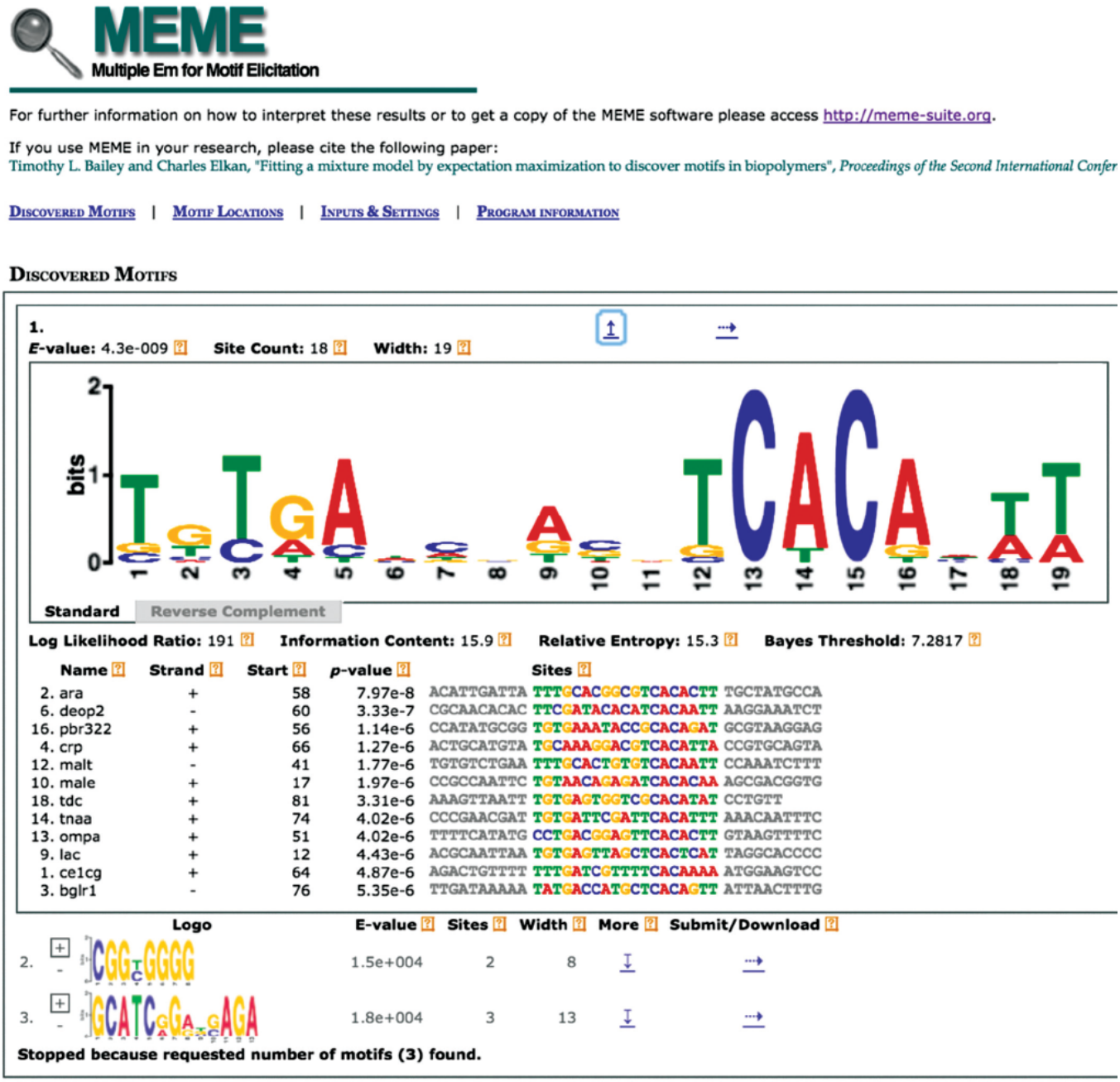


Figura 14.16 Resultados de uma busca de padrões usando o MEME.

O programa MAUVE, por sua vez, foi desenvolvido para produzir um alinhamento múltiplo de vários genomas e representa esses alinhamentos dispondo os genomas em um mesmo plano, como linhas paralelas, e interligando as regiões conservadas por linhas perpendiculares. Cada região conservada no MAUVE é retratada por um quadrado de cor específica, o que facilita a identificação visual de sintenia entre os genomas (Figura 14.17 B).

Pfam

O Pfam, ou *protein family database* [banco de dados de famílias proteicas], é mantido pelo EBI. Corresponde a um repositório de informações sobre famílias de proteínas, com foco nas regiões conservadas entre as mais diversas proteínas, ou seja, na definição das unidades evolutivas que as compõem, as quais são frequentemente chamadas domínios proteicos – muito embora essa nomenclatura não seja universal. As famílias de domínios proteicos, pela definição do Pfam, são regiões de proteínas que têm similaridade de sequência suficiente para que sejam identificadas como homólogas pelos programas do pacote HMMER, que é o *software* usado na construção do Pfam. Tais domínios podem ser encontrados em diversas combinações nas mais variadas proteínas, muitas das quais realizam funções distintas e têm origens independentes, mas compartilham aquela região (domínio), que se entende ter uma origem única. Difere, nesse sentido, de outras classificações de proteínas em famílias, que focam em proteínas ou genes inteiros em diferentes organismos (EggNOG, OrthoMCL, TreeFam) ou em uma classificação baseada primariamente na definição de domínios a partir da estrutura da própria proteína (CATH SCOP).

As famílias no Pfam são geradas manualmente por curadores ou por contribuição dos usuários para o banco de dados. Os curadores identificam as regiões conservadas correspondentes a um novo domínio e geram um alinhamento múltiplo com uma seleção de representantes da nova família. Ao serem incorporados ao Pfam, os alinhamentos múltiplos (*seed alignments*) são convertidos em matrizes numéricas chamadas modelos de Markov ocultos ou HMM (do inglês, *hidden Markov models).* Os HMM usados no Pfam são modelos que descrevem a probabilidade de se observar cada um dos aminoácidos em cada posição das sequências que pertencem à família representada pelo HMM e também a probabilidade de esses resíduos mudarem ou de corresponderem a eventos de inserção/deleção.

O *software* HMMER opera nos bastidores da página *web* do servidor de buscas do PFAM, e uma propriedade importante dos HMM do HMMER é seu comprimento variável (*profile hidden Markov model*), que viabiliza o uso do mesmo programa para criação de um HMM para cada família de proteínas, as quais necessariamente irão variar em comprimento. Os modelos do Pfam propiciam a identificação das regiões conservadas de cada família em sequências de interesse, análise que, aliada à excelente qualidade da anotação funcional no Pfam e da interface web do HMMER, possibilita que o usuário faça inferências sobre a função das proteínas mesmo em circunstâncias nas quais o BLAST não produz resultados úteis.

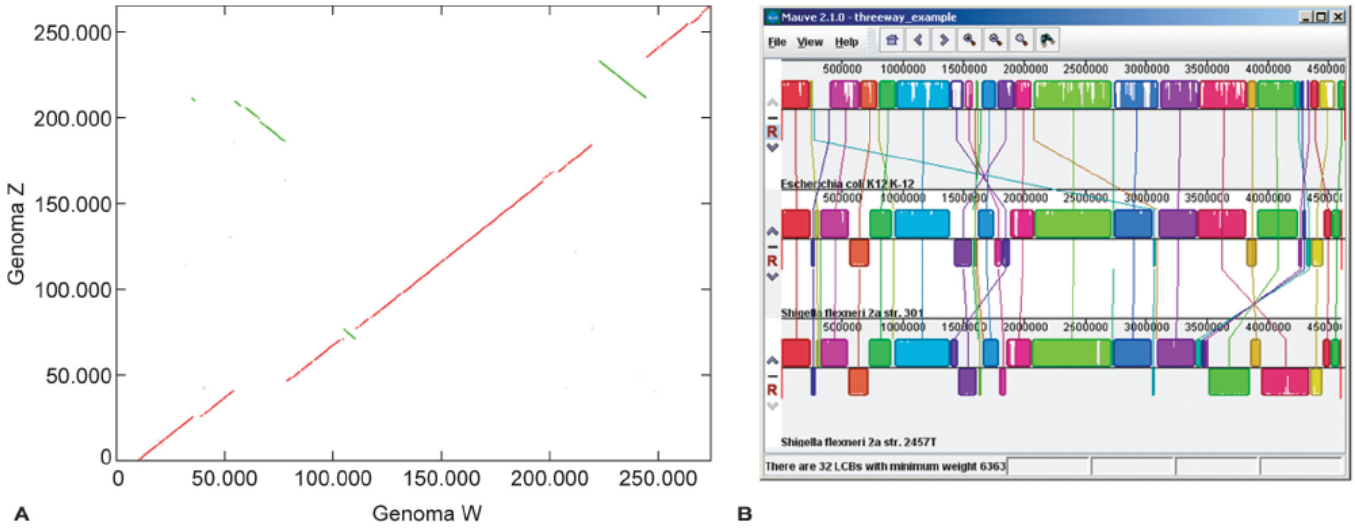


Figura 14.17 Análise de organização genômica e de genes no genoma. **A.** Resultados de comparação entre dois genomas, Z e W, pelo programa MUMMER. Os eixos representam o número de nucleotídeos a partir de uma mesma origem de replicação. Para cada fragmento (*n-mers*) conservado no genoma Z e presente no mesmo sentido no genoma W, um ponto de cor vermelha é inserido. A presença de muitos pontos vizinhos produz as linhas vermelhas, demonstrando o grau de conservação dessas regiões genômicas. As linhas verdes nada mais são do que as regiões conservadas no sentido invertido em um dos genomas. Essas inversões deixam claros os chamados rearranjos cromossômicos. **B.** Análise de uma região genômica específica em três genomas usando o programa MAUVE: os blocos coloridos representam as regiões conservadas, e as linhas verticais de mesma cor a localização correspondente do mesmo bloco nos outros genomas analisados.

Na interface do Pfam, o usuário poderá verificar a composição de domínios de sua proteína de interesse por diferentes métodos: inserindo o identificador da proteína no banco de dados UniProt no campo *View a sequence*, digitando o identificador de uma estrutura no PDB (campo *View a structure*) ou, ainda, colocando a própria sequência da proteína no campo *Sequence search*. Alternativamente, é possível inspecionar a anotação das famílias de domínios buscando por seu identificador (campo *Jump to*), por palavras-chave (campo *Keyword*) ou apenas navegando na lista de famílias (*link Browse*). Caso o usuário faça a busca usando a sequência de uma proteína e essa proteína tenha uma ou mais regiões correspondentes a domínios presentes no banco de dados, a ferramenta retornará uma página com uma representação gráfica dos domínios presentes na proteína e uma tabela com detalhes da análise, identificadores de famílias, *E-value* e as coordenadas de cada domínio na proteína. Ao clicar no desenho ou nos nomes das famílias, o usuário será conduzido à página com a anotação de cada família e os detalhes do modelo (HMM) que a representa. Essa página conterá uma descrição da importância e das funções daquela família e *links* para bibliografia relevante, além de atalhos para painéis que contêm informações sobre as proteínas de genomas completos do UniProt que apresentam regiões homólogas àquela família (Figura 14.18).

TMHMM

O *software* TMHMM é uma ferramenta que identifica domínios transmembranares, ou seja, regiões compostas de a-hélices hidrofóbicas que atravessam ou se localizam em membranas celulares. Esse *software* também utiliza um HMM para calcular a probabilidade de resíduos nas proteínas de interesse formarem hélices transmembranares. No entanto, é importante considerar que os HMM do TMHMM são diferentes daqueles usados no HMMER, motivo pelo qual precisam ser implementados em outro *software*.

O TMHMM opera de modo semelhante a outros serviços de análise de sequências: uma sequência FASTA de interesse do usuário é carregada no servidor e o programa retornará como resultado a predição de domínios transmembranares. Além de dizer se a proteína tem ou não tais domínios, o TMHMM produz um gráfico (Figura 14.19) que possibilita que o usuário infira aspectos gerais da topologia esperada da proteína uma vez que esteja inserida na membrana, ou seja, o número e a orientação relativa das hélices transmembranares. Vale ressaltar que embora a ferramenta também sinalize o local onde a proteína (N ou Carboxi-terminal) estaria no citoplasma ou no meio extracelular, essa predição deve ser interpretada com parcimônia, pois algumas vezes as regiões citoplasmáticas são preditas como extracelulares, e vice-versa, principalmente para proteínas com uma única região transmembranar.

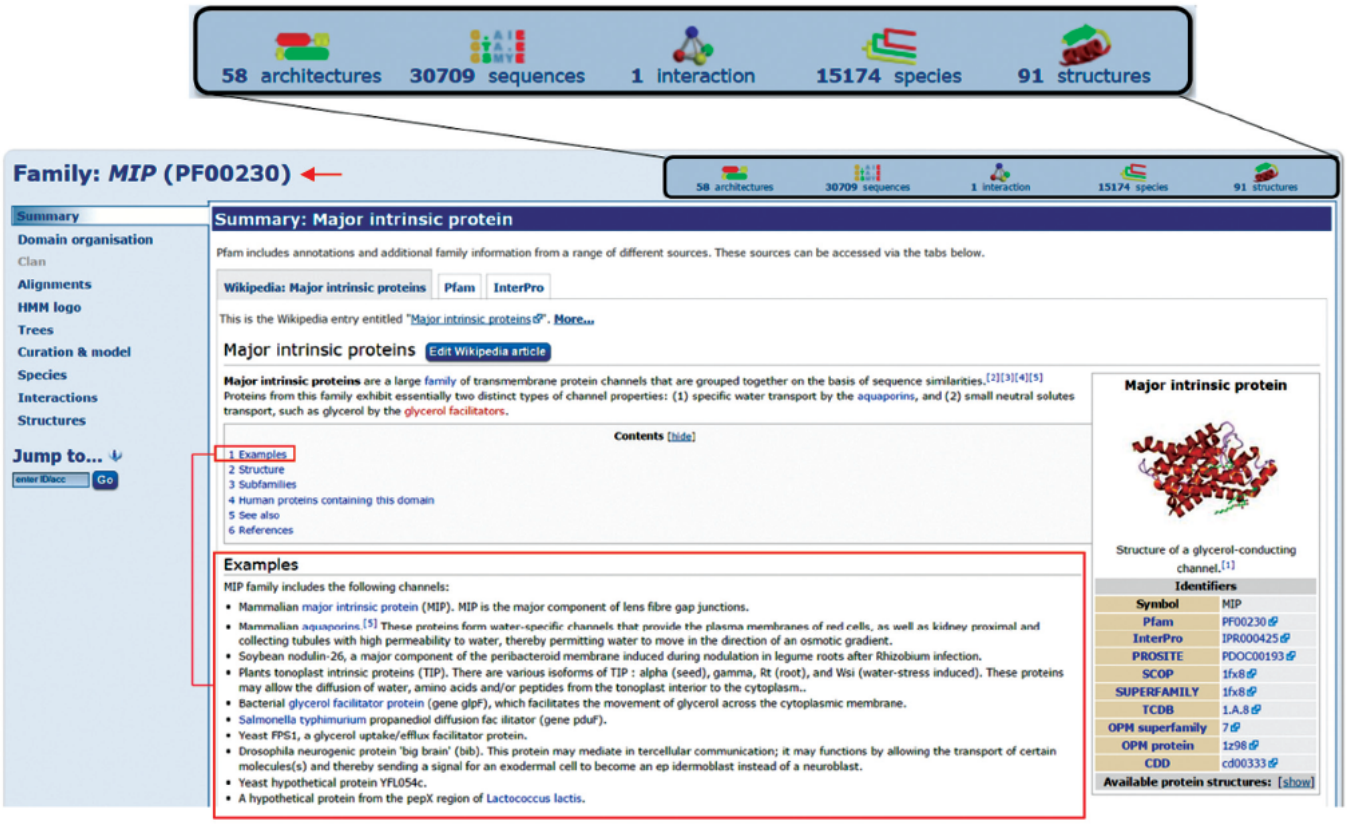


Figura 14.18 Exemplo de entrada no banco de dados Pfam (família MIP). A seta vermelha destaca os identificadores (nome e código) da família no Pfam. O boxe vermelho destaca exemplos de proteínas que são membros dessa família. Atalhos para outros bancos de dados que contêm informações sobre membros dessa família são observados abaixo da estrutura 3D. Nota-se que o painel visível nesta imagem é uma página na Wikipedia, com a qual o Pfam está integrado. Painéis equivalentes, com informações provenientes de curadores do Pfam e do Interpro, podem ser acessados clicando nos *links* de mesmo nome. O menu superior (em destaque) contém atalhos para dados sobre proteínas que são membros dessa família, com os respectivos números de resultados associados a cada um dos tipos de informação catalogados.

TMHMM result

HELP with output formats

```
# hsa:358 Length: 186
# hsa:358 Number of predicted TMHs:  3
# hsa:358 Exp number of AAs in TMHs: 59.98317
# hsa:358 Exp number, first 60 AAs:  6.53983
# hsa:358 Total prob of N-in:        0.13365
hsa:358 TMHMM2.0        outside      1    54
hsa:358 TMHMM2.0        TMhelix     55    74
hsa:358 TMHMM2.0        inside      75    80
hsa:358 TMHMM2.0        TMhelix     81   101
hsa:358 TMHMM2.0        outside    102   125
hsa:358 TMHMM2.0        TMhelix    126   148
hsa:358 TMHMM2.0        inside     149   186
```

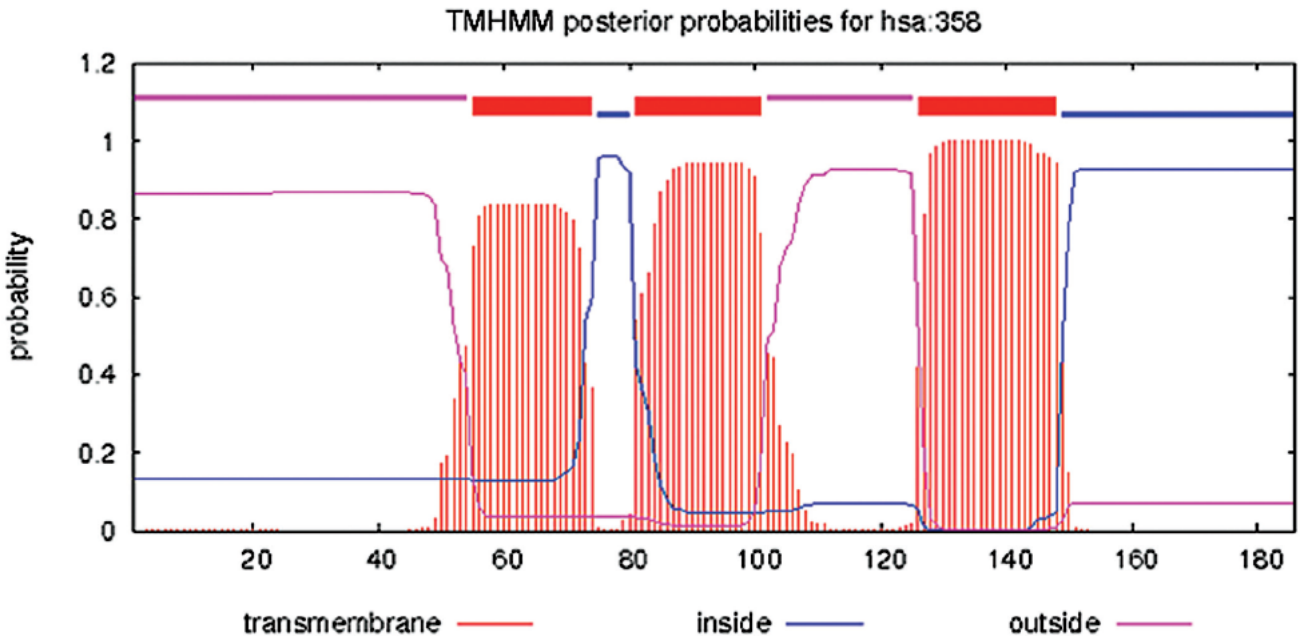


Figura 14.19 Resultado da predição de domínios transmembranares em uma proteína usando o TMHMM. O painel superior da página mostra um sumário dos dados de entrada e dos resultados encontrados. Detalhes desses resultados são apresentados no gráfico, onde as linhas vermelhas verticais destacam a probabilidade de cada resíduo pertencer a uma região transmembranar, ao passo que as linhas rosa e azul representam, respectivamente, a probabilidade de cada resíduo pertencer a uma região intra ou extracelular.

Biologia de sistemas

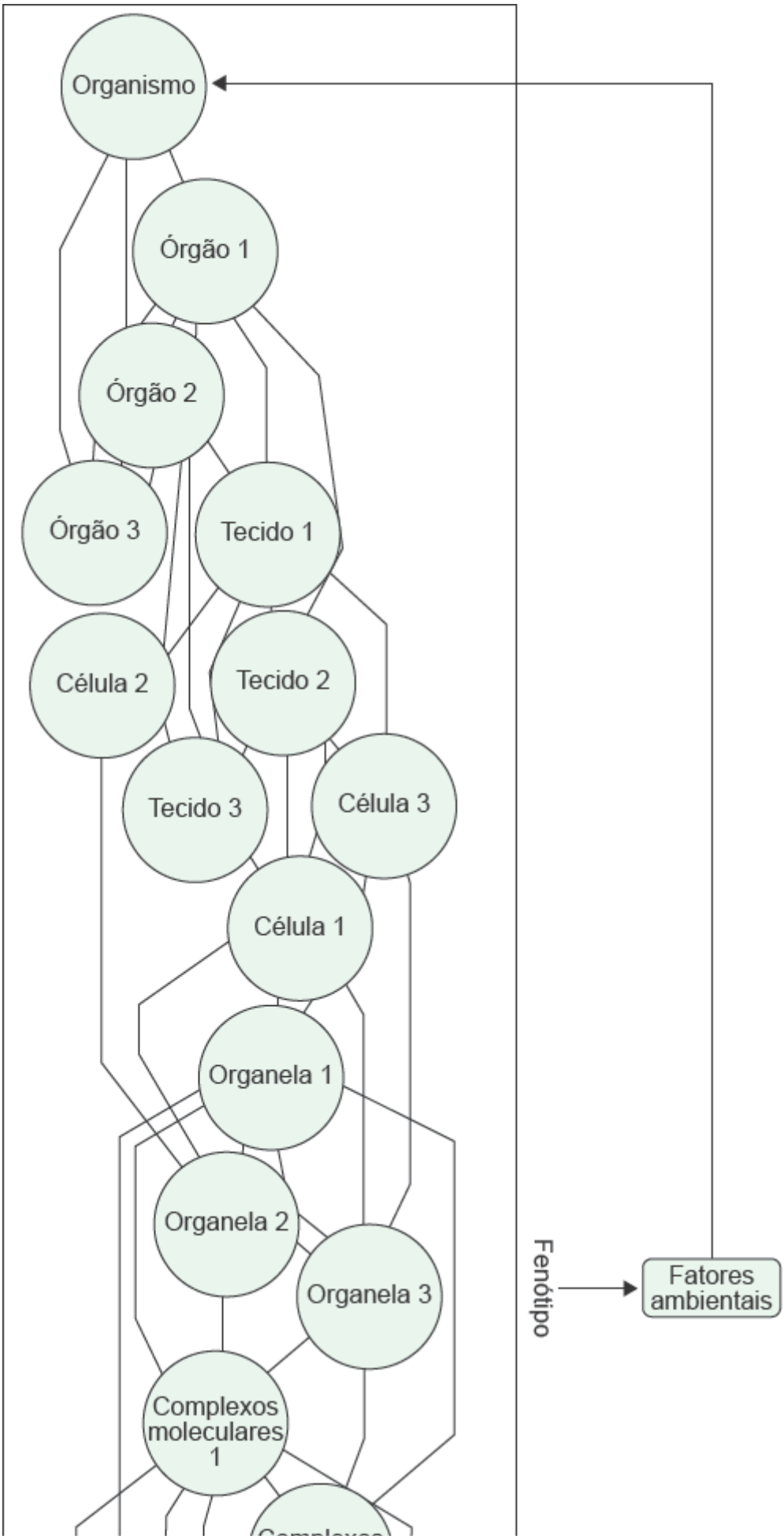
O que são os sistemas biológicos?

O final do século 20 e o início do século 21 foram marcados pelo desenvolvimento de técnicas e procedimentos de análises moleculares que proporcionaram aos pesquisadores uma visão muito mais abrangente sobre os diferentes processos bioquímicos encontrados em um organismo vivo. Esses mesmos estudos possibilitaram entender como e porque alguns mecanismos bioquímicos podem ser únicos para um conjunto específico de organismos. Entretanto, esses dados demonstraram, de maneira conclusiva, que a grande maioria dos mecanismos bioquímicos e moleculares são compartilhados por todos os organismos da Terra.

Os procedimentos moleculares de análise global, também conhecidos como projetos de natureza "ômica" (genômica, transcriptômica, proteômica, entre outros), têm mostrado que uma função biológica raramente é o resultado da ação de uma única ou de algumas poucas moléculas. Em vez disso, a maioria das características biológicas ou fenótipos surge como consequência das interações complexas que ocorrem entre as moléculas encontradas em uma célula e o ambiente que a cerca (Figura 14.20).

Muitas doenças que afligem o ser humano, como câncer e diabetes, podem ser definidas como estados particulares em mecanismos complexos de interação que são diferentes daqueles encontrados em um organismo

saudável. Diante da imensa complexidade dos organismos vivos, um interesse cada vez maior na compreensão dos sistemas biológicos por meio de métodos integrativos levou à emergência de uma área da bioinformática chamada *biologia de sistemas*.



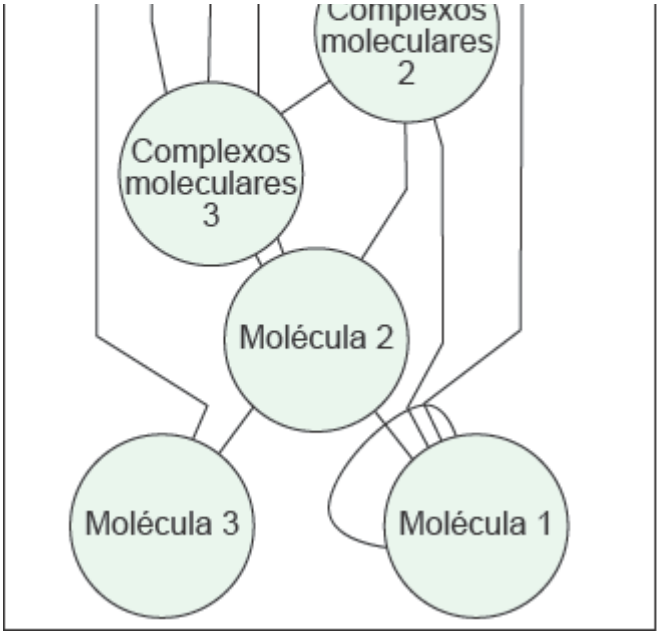


Figura 14.20 Representação simplificada dos principais componentes de um determinado organismo. As interações observadas entre os seus componentes possibilitam que ele responda a sinais ou fatores do meio ambiente, resultando no fenótipo observado para esse organismo.

A biologia de sistemas busca entender como as interações entre os elementos de um sistema biológico dão origem aos processos essenciais para a manutenção da vida. Essa visão global dos sistemas biológicos possibilita a geração de hipóteses testáveis experimentalmente.[15]

Devido à complexidade e à enorme quantidade de informações geradas pelos projetos "ômicos", um dos principais desafios da biologia de sistemas é integrar todas essas informações a fim de mapear, entender e modelar, em termos quantitativos, o comportamento celular e de um conjunto de genes, por exemplo.[16,17]

Assim, ela pode ser vista como um processo interativo em que um modelo ou um mecanismo é desenvolvido a partir da integração das características de múltiplos subsistemas e continuamente melhorado por meio da comparação constante com dados experimentais e pela integração de novas características biológicas obtidas por meio de técnicas experimentais mais pontuais, como as técnicas de biologia molecular (Figura 14.21).[18]

Fundamentos da biologia de sistemas

As informações sobre macromoléculas e suas interações funcionais (proteínas, DNA e RNA) mantidas em bancos de dados públicos constituem a fonte primordial de informações usadas na biologia de sistemas. O crescimento dramático desses bancos de dados biológicos nos últimos anos, porém, faz da sua exploração e análise um importante desafio computacional.[19]

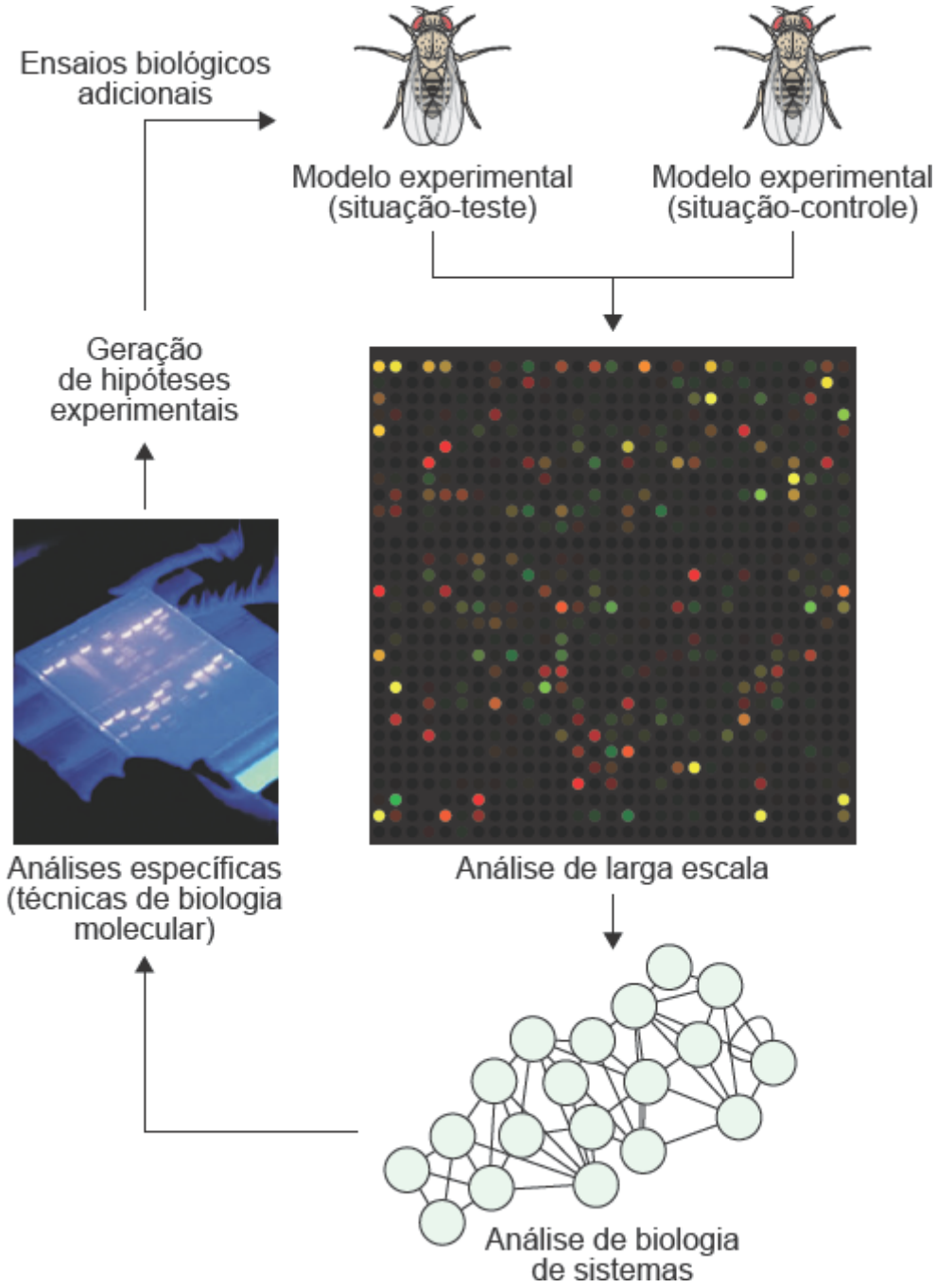


Figura 14.21 Ciclo "virtuoso" da biologia de sistemas. A partir dos dados de larga escala gerados pela análise de um modelo biológico em uma condição experimental particular, são construídas redes interatômicas. Essas redes possibilitam ensaios experimentais mais pontuais (como análise de expressão de um conjunto mínimo de genes importantes para a rede) que, então, servem de base para a geração de novas hipóteses experimentais, as quais podem ser testadas no modelo biológico e, assim, repetir todo o processo.

A fim de suplantar esse desafio, uma das principais abordagens computacionais utilizadas para integrar todas as informações disponíveis nos bancos de dados públicos consiste na aplicação de modelos baseados na Teoria dos Grafos, um ramo da matemática que lida com os fundamentos das chamadas redes de interação.[20,21]

As redes de interação são estruturas formadas por elementos unitários (também conhecidos como *nós*), ligados por conectores (arestas), que podem ser usadas para descrever uma ampla variedade de sistemas, sejam esses biológicos, tecnológicos ou sociais.[22] Por exemplo, um organismo é definido como uma rede formada por órgãos, tecidos, tipos celulares, complexos moleculares e pequenas moléculas conectadas por interações funcionais ou moleculares (nós e arestas, respectivamente, na Figura 14.22) que possibilitam o fluxo de informação entre esses próprios componentes.

O que é a "complexidade"?

Um *sistema* de um organismo pode ser definido como a interação observada entre os seus diferentes componentes, desde moléculas até estruturas maiores, como tecidos e órgãos (ver Figura 14.22). Dessa maneira, a integração dos diferentes sistemas permite entender o comportamento dos processos biológicos.

Essa compreensão dos diferentes níveis sistêmicos tenta responder a três questões principais:

- Quais são as partes componentes do sistema (para as células, quais são os genes e as proteínas que por eles são codificadas)?
- Como as partes trabalham juntas e em que condições?
- Como as partes trabalham juntas para alcançarem um determinado objetivo e responderem a um estímulo interno do organismo ou do meio ambiente?

Considerando a visão tradicional da biologia, a enumeração e a caracterização de todos os componentes básicos biológicos possibilitam responder à primeira questão, gerando informações para a segunda e a terceira. Por exemplo, certas proteínas têm a capacidade de ligarem-se a regiões particulares do DNA, ativando ou inibindo a transcrição do mRNA que, então, será traduzido para proteínas. Essas mesmas proteínas que controlam a transcrição do DNA, em sua grande maioria, também têm a capacidade de se ligarem a outras proteínas, formando complexos que desempenham diferentes funções. Seguindo essa linha de raciocínio e considerando as informações "ômicas" disponíveis, existem milhares (senão milhões) de diferentes tipos e estados funcionais de proteínas em um organismo vivo, de modo que o número de combinações (estados) geradas por essas interações é potencialmente gigantesco.

Complexidade e propriedades emergentes

Uma característica de todos os sistemas complexos, sejam esses biológicos ou de natureza tecnológica, é o surgimento de determinados padrões que não podem ser explicados apenas pela soma simples de seus componentes. Por exemplo, a funcionalidade de um cérebro humano não pode ser explicada apenas pelo estudo unitário de seus neurônios e células de suporte. Na verdade, o cérebro, como estrutura funcional, surge quando todas as células executam as suas funções de maneira integrada, respondendo a determinados estímulos ambientais. O mesmo vale para estruturas de natureza tecnológica, como é o caso da internet, dos sistemas de geração e transmissão de eletricidade e do transporte público. A internet é um exemplo de sistema com propriedades emergentes: trata-se de um conjunto de páginas ligadas por *hyperlinks* com características como expansão e tolerância a erros. As propriedades emergentes dos sistemas biológicos são o foco da análise em biologia de sistemas e constituem uma fonte importante de dados a respeito do estado de um organismo.

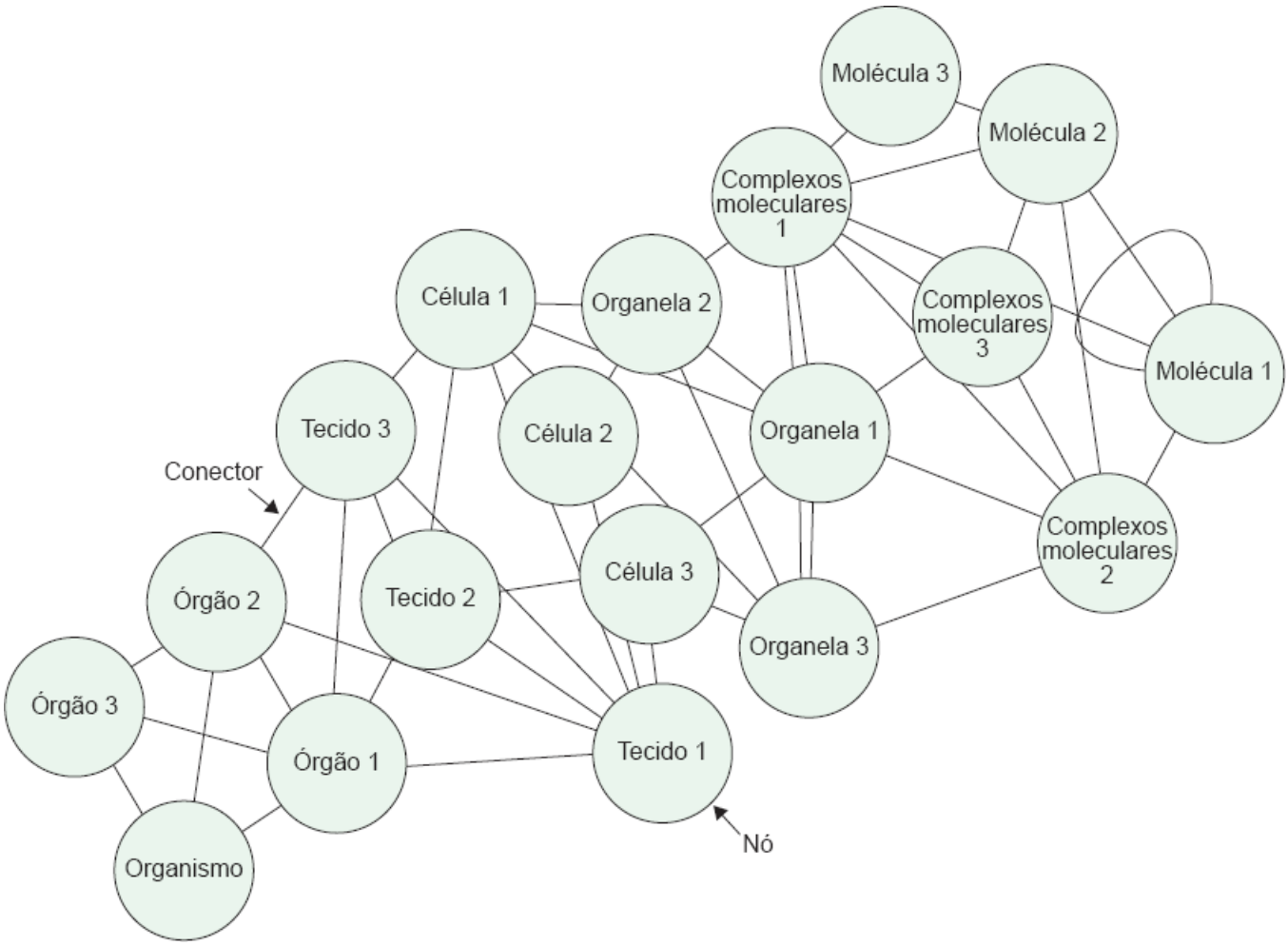


Figura 14.22 Rede de interação simplificada entre os diferentes elementos constituintes de um organismo multicelular. Os componentes de um organismo (nós) estão em constante comunicação entre si (via conectores ou arestas), o que possibilita que ele se mantenha vivo em seu ambiente.

Uso de modelos para a compreensão de sistemas biológicos

O entendimento completo de qualquer sistema biológico requer uma quantidade de dados muito maior do que as atuais tecnologias podem oferecer, mas podemos imaginar que é possível construir um modelo de um sistema biológico conhecendo todas as moléculas envolvidas, as suas concentrações, como interagem entre si, os efeitos de cada parte individual em seus vizinhos e parâmetros dinâmicos.[15,23]

Entretanto, seria necessário conhecer todos os detalhes de um processo para sermos capazes de desenvolver modelos de sistemas biológicos que sejam úteis ou preditivos? As análises de mapas de interações proteicas ou entre proteínas e DNA sugerem que, mesmo sendo os dados esparsos ou pouco conhecidos, é possível construir modelos de redes que representem de maneira realística um sistema biológico.

Um exemplo são as chamadas redes metabólicas, sistemas complexos formados por vias em que os substratos e os produtos metabólicos estão conectados entre si. Os exemplos de várias vias metabólicas incluem a glicólise, o ciclo do ácido cítrico, a via da pentose fosfato, entre outros. Dessa maneira, a construção de uma via metabólica envolve a coleção de todas as informações metabólicas relevantes de um organismo e a sua compilação em uma rede. Nesse caso, é importante que haja uma correlação entre as informações ômicas e os processos metabólicos, o que é feito por meio de banco de dados especializados, como o KEGG.

Apesar de rudimentares, as redes metabólicas são ferramentas poderosas para o estudo e a modelagem do metabolismo.

Tipos de redes biológicas

Existem vários tipos diferentes de redes biológicas que um pesquisador pode considerar em seus trabalhos com biologia de sistemas. Essas redes podem representar o estado e as interações entre metabólitos, a atividade transcricional dos genes, interação entre proteínas, pequenas moléculas e proteínas, entre outros, conforme o tipo de

dados de larga escala gerados e a análise desejada. Por exemplo, uma rede de regulação genética ou gênica descreve a expressão gênica, ou seja, a produção de proteínas a partir do genoma pela transcrição e pela tradução. A expressão de um gene pode ser controlada pela presença de outras proteínas ativadoras ou repressoras que, por sua vez, formam uma rede em que os nós representam as proteínas e os conectores indicam uma dependência da síntese de uma proteína particular na atividade ou concentração de outras proteínas. Em outras palavras, as redes de regulação genética são como interruptores elétricos ("ligado" e "desligado") de uma célula, operando em nível gênico e orquestrando dinamicamente os níveis de expressão para cada gene presente no genoma. Cada transcrito de RNA funciona como molde para a síntese de uma proteína específica pelo processo da tradução. Do mesmo modo, as redes transcricionais (regulatórias) bacterianas mostram as relações entre os fatores de transcrição e os *operons* que eles regulam. Nessas redes, cada nó representa um *operon*, um grupo de genes contíguos transcritos em uma única molécula de mRNA, e os conectores representam as interações transcricionais diretas. Finalmente, as redes de interações entre proteínas (conhecidas como interatomas) são compostas de nós que representam as proteínas e nos quais cada conector indica uma interação. Assim, uma rede com todas as proteínas de um organismo e todas as suas possíveis interações pode ser chamada rede proteômica. As interações nessas redes são importantes para a maioria dos processos biológicos, considerando que muitas dessas proteínas necessitam interagir com outras para exercerem as suas funções de maneira adequada. Do mesmo modo, a interação entre pequenas moléculas, como aquelas usadas para finalidades terapêuticas, e proteínas determina como um medicamento atua em um ou vários processos biológicos. Esse tipo de informação é de grande utilidade para o desenvolvimento de novos medicamentos ou de combinações inéditas de compostos para o tratamento de patologias, como o câncer.

Considerando os diferentes tipos de redes biológicas, a biologia de sistemas usa duas linhas de raciocínio que podem ser aplicadas para a geração de modelos biológicos. A primeira linha ou ramo de estudo é conhecida como biologia de sistemas *top-down* (uma expressão em inglês que significa do "mais complexo ao menos complexo" ou "de cima para baixo"). A segunda linha de raciocínio é denominada biologia de sistemas *bottom-up* (do "menos complexo ao mais complexo" ou "de baixo para cima"). Cada um dos ramos tem os seus pontos fortes e fracos, cabendo, então, ao pesquisador escolher aquele que melhor caracteriza o seu sistema. É importante ressaltar que as duas visões da Biologia de Sistemas não são excludentes, mas altamente complementares. O que define o uso de uma ou outra está no tipo de modelo a ser analisado pelo pesquisador.

Biologia de sistemas *top-down* e *bottom-up*

Com a introdução das ferramentas "ômicas", o método "de cima para baixo" emergiu como a ferramenta dominante, pela qual o sistema biológico é visto na sua totalidade (Figura 14.23).

Nesse tipo de método, o objetivo principal é desdobrar novos mecanismos moleculares usando um ciclo interativo no qual há obtenção de dados experimentais, análise e integração desses dados para determinar as correlações entre as partes do sistema e a formulação de novas hipóteses experimentais, as quais predizem novas correlações e propiciam um novo conjunto de experimentos. A principal característica da biologia de sistemas "de cima para baixo" é sua visão geral do sistema; por isso, ela tem grandes desafios na compreensão de bancos de dados experimentais obtidos a partir de organismos-modelo submetidos a poucas "perturbações".[24]

De modo geral, essas perturbações são definidas como todas as alterações de natureza genética (como as obtidas a partir de uma ou mais mutações ou mesmo superexpressão de proteínas), ambientais (mudanças na concentração de nutrientes, em fatores de crescimento ou níveis de estresse), induzidas por RNA de interferência, da natureza intrínseca da dinâmica celular (como o ciclo celular) ou pela administração de drogas que modificam o comportamento fisiológico de um sistema. Dessa maneira, esses estudos visam obter uma visão geral do comportamento de um sistema submetido a perturbações, tendo como objetivo a descoberta de padrões comportamentais genéricos o suficiente para realizar predições biológicas.[24]

Os modelos que usam a biologia de sistemas "de cima para baixo" são fenomenológicos, ou seja, não são baseados em mecanismos, mas sim nos fenômenos ou processos biológicos por eles gerados. Esse tipo de biologia de sistemas é usado, principalmente, para a caracterização de subsistemas celulares que ainda não foram elucidados e com detalhes que ainda precisam ser descobertos (como a elucidação das interações entre proteínas ou redes genéticas).[24] Assim, a integração dos estudos "ômicos" que, simultaneamente, analisam os dados transcriptômicos, proteômicos e metabolômicos (área que estuda o fluxo de metabólitos em um sistema) pode ajudar a biologia de sistemas "de cima para baixo" por meio da chamada genômica vertical. A genômica vertical busca traçar,

quantitativamente, as mudanças na taxa de um processo biológico, por meio da mudança nos níveis de substrato, produtos, efetores, enzimas e de mRNA. Tal desenvolvimento oferece uma nova forma de estudo para a biologia celular e permite visualizar a extensão de várias rotas regulatórias presentes internamente à rede e que podem contribuir para um determinado comportamento.[24]

Diferentemente da biologia de sistemas "de cima para baixo", a biologia de sistemas "de baixo para cima" busca deduzir as propriedades funcionais emergentes de um subsistema que foi profundamente caracterizado em termos físico-químicos (Figura 14.24).

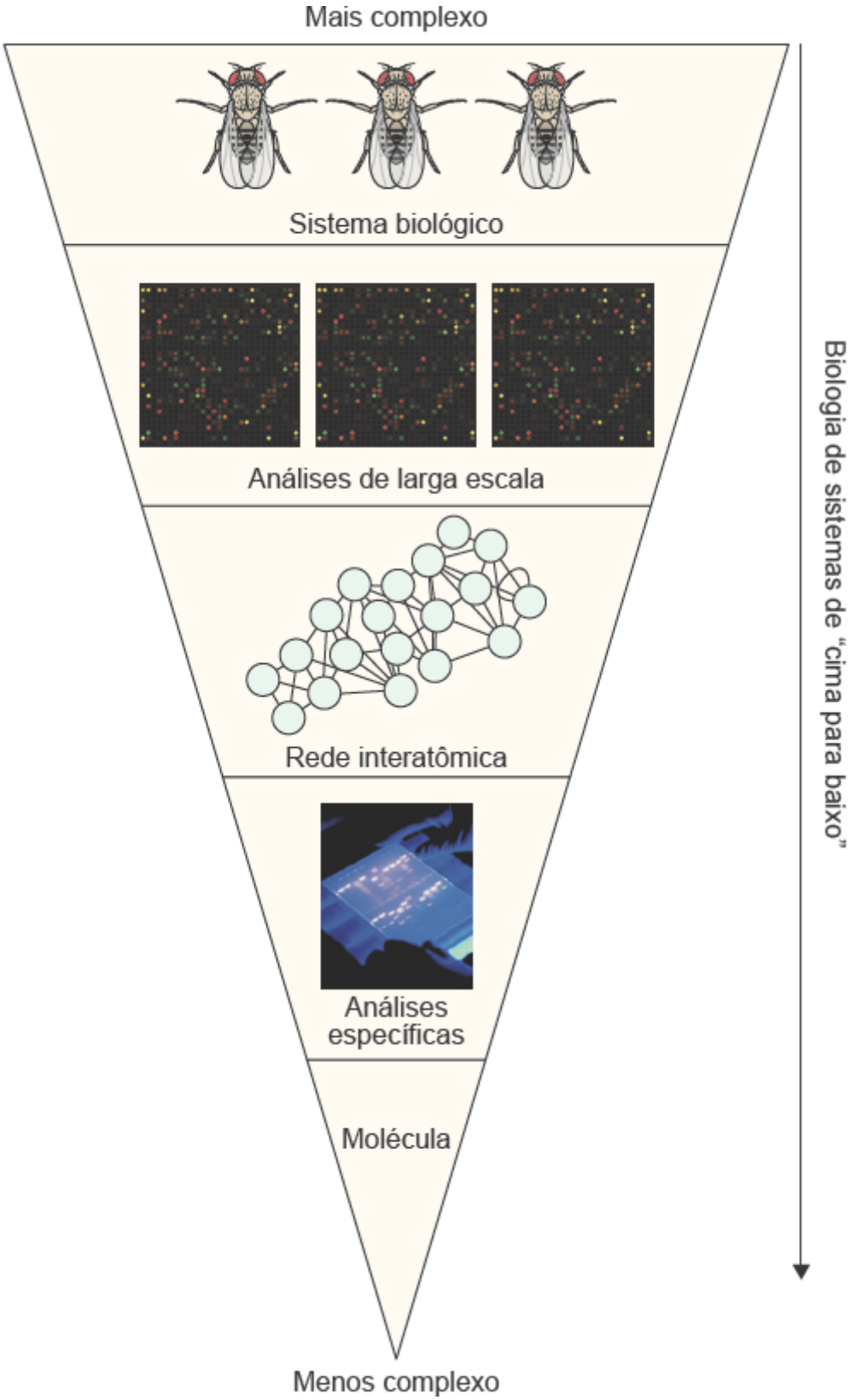


Figura 14.23 Biologia de sistemas de "cima para baixo". Os componentes mais complexos de um sistema biológico são desmembrados em seus elementos mais simples, possibilitando realizar inferências sobre a sua constituição e funcionamento.

A biologia de sistemas “de baixo para cima” começa com a enumeração dos componentes elementares de um sistema e, por meio da simulação das interações desses componentes, integra os dados conhecidos sobre o comportamento de cada um deles para predizer o comportamento do sistema como um todo. Assim, ela busca, por exemplo, integrar os modelos de várias vias enzimáticas em um único modelo. Portanto, dependerá de estudos experimentais que determinem as propriedades quantificáveis dos componentes enzimáticos (cinética enzimática, propriedades de difusão, entre outras) e da obtenção de dados referentes das respostas das redes sob estudo às perturbações introduzidas.[24] Essas informações são utilizadas para a construção de modelos detalhados do sistema e possibilitam o desenvolvimento de ferramentas computacionais para a análise de dados e simulação dos modelos.

É importante ressaltar que a determinação dos parâmetros cinéticos enzimáticos e concentrações proteicas usando as estratégias atuais de análise tem um risco inerente devido à simplificação das técnicas bioquímicas empregadas. Assim, as estratégias aplicadas para a estimação desses parâmetros cinéticos talvez não mostrem os “verdadeiros parâmetros”. Por outro lado, se é possível incorrer em erros ao medir os parâmetros cinéticos de um sistema, em alguns casos é extremamente difícil, senão impossível, medir os parâmetros cinéticos *in vitro,* como é o caso da sinalização celular e da expressão gênica eucariótica.[24] Existe a possibilidade de contornar esse problema por meio do desenvolvimento de tecnologias que medem a cinética enzimática *in vivo.*[25] Tal ferramenta está se tornando cada vez mais disponível por meio do aperfeiçoamento de técnicas de imagens de alta resolução usando fluorescência ou outros métodos que possibilitem a observação do fluxo de metabólitos dentro de um sistema ou subsistema celular.[26]

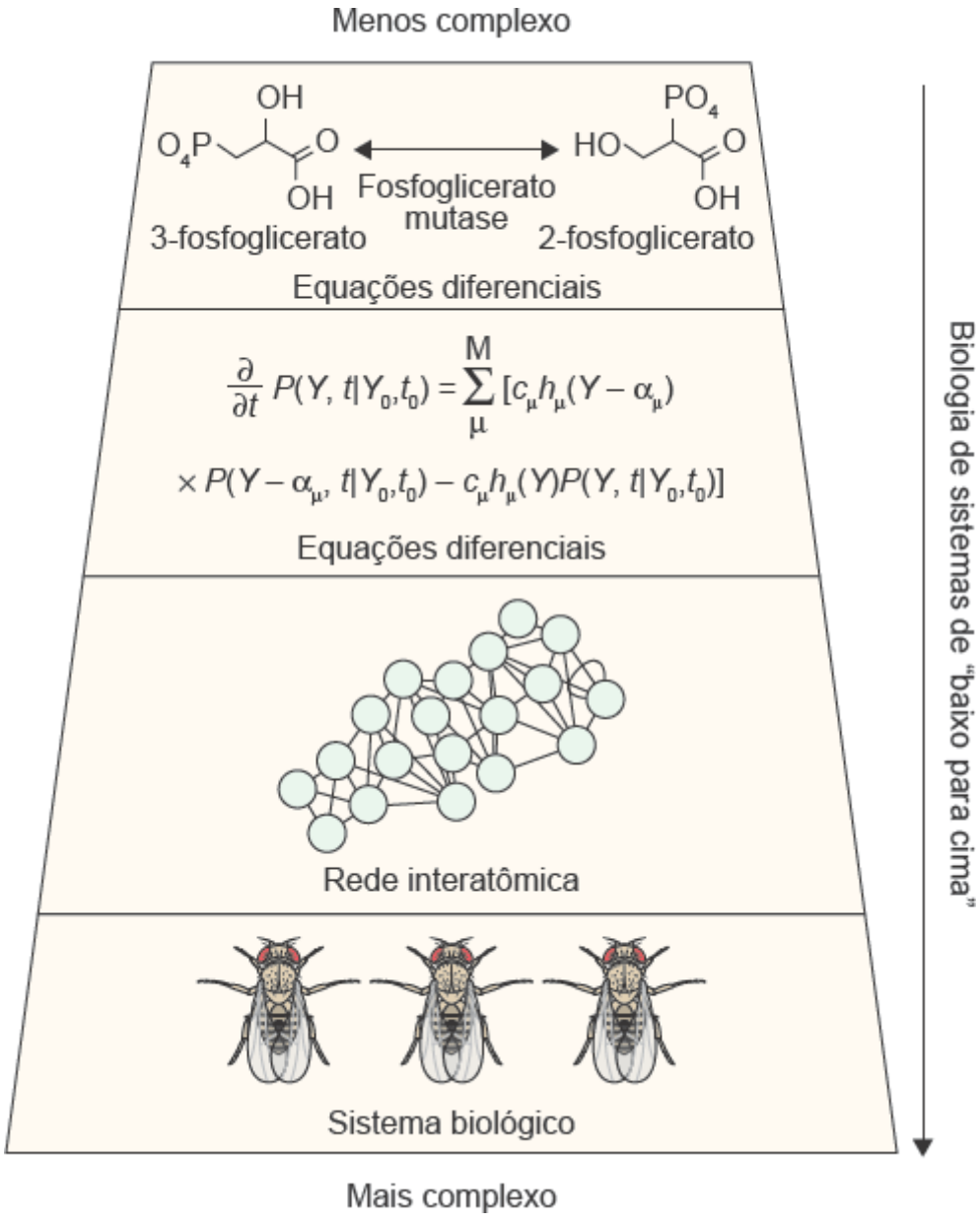


Figura 14.24 Biologia de sistemas de “baixo para cima”. A partir do estudo de uma determinada via bioquímica e do estabelecimento dos parâmetros cinéticos enzimáticos, é possível construir sistemas de equações diferenciais que

modelam o fluxo de informação e matéria em redes biológicas. Essas redes são usadas para predizer o comportamento de um sistema biológico diante de uma determinada situação experimental.

Redundância, percolação e propagação de erros em redes

O uso de modelos de rede (grafos) para análise de sistemas complexos revelou que, em muitos desses sistemas, quando um nó importante é perturbado, essa perturbação ou erro pode não se espalhar para a rede como um todo, acarretando consequências mínimas para a rede, pois o erro é tolerado; ou pode resultar no colapso de todo um sistema vivo (que corresponderia à morte desse organismo) ou de um sistema tecnológico (quebra na comunicação entre computadores como resultado de ataques de *hackers* ou desligamento do fornecimento de energia elétrica em razão de sobrecargas em sistemas de transmissão de eletricidade).

Quando um elemento ou nó de uma rede sofre um ataque ou perturbação, esta será tolerada dependendo da importância daquele nó para a rede. Nas redes livres de escala ou modulares, haverá sempre alguns poucos nós que concentram o fluxo de informações para si, seja por apresentarem-se altamente conectados com outros nós (conhecidos por *hubs*) ou por estarem servindo de "ponte" (nós do tipo "gargalo") para dois ou mais grupos (módulos) de nós que executam determinadas funções na rede. Dessa maneira, quando o erro é introduzido nesses nós importantes, a rede pode se fragmentar, de modo que os módulos perdem a comunicação entre si com a propagação do erro e eventual colapso do sistema. Para os nós considerados menos importantes, que constituem a quase totalidade da rede, um eventual erro e perda da função desse nó não acarreta maiores consequências para a rede, e o erro, então, passa a ser tolerado. Entretanto, quando muitos nós são retirados de uma rede de maneira aleatória, outra espécie de falha, denominada cascata de sobrecargas de erros, pode resultar no colapso da rede. Em termos mais simples, a cascata de sobrecargas de erros indica que os nós que não foram perturbados na rede são obrigados a lidar com uma quantidade muito mais vasta de informações do que normalmente lidariam, o que leva à sua saturação e perda de função. Para exemplificar esse ponto, com base em redes de transmissão de eletricidade, se vários postes são afetados por um vendaval e sofrem queda, toda a eletricidade que antes passava pelas linhas afetadas é redirecionada para outros pontos não afetados, resultando em um grande aumento na quantidade de energia elétrica nesses pontos, o que provoca sobrecarga (saturação) e desligamento automático dos sistemas para evitar danos maiores à estrutura. Esse tipo de propagação de erro é chamado percolação, pois ele tem a capacidade "percolar" ou difundir-se pela rede.

Felizmente, as redes de natureza biológica e a maioria das redes tecnológicas têm uma característica importante que impede o seu colapso: a redundância. A redundância se caracteriza pela presença de inúmeros nós e conexões que apresentam sobreposição de função, ou seja, dois ou mais nós diferentes são capazes de executar funções semelhantes. Assim, quando um erro é introduzido em um determinado nó, o outro é capaz de assumir a função desse nó falho e manter a rede funcionando. Em sistemas biológicos, verifica-se uma alta redundância em processos moleculares importantes, como a replicação de DNA e a tradução proteica. Nesses casos, as enzimas e proteínas estruturais têm funções altamente redundantes e, consequentemente, estão menos sujeitas à percolação de erros, seja pela sobrecarga de erros ou como resultado da perda do nó devido a mutações.

Redes de interação proteína-proteína e genes-proteína

Há muito se sabe que as interações entre as proteínas são fundamentais para todos os processos biológicos, variando desde a formação de estruturas celulares e complexos enzimáticos até a regulação de vias de sinalização ou de sistemas transcricionais.[27] As proteínas frequentemente funcionam como complexos estáveis ou transientes com outras proteínas, e as interações entre as proteínas podem servir a diferentes funções, como conferir especificidade para as interações entre enzimas e substratos em eventos de transdução de sinais, conferir proteção de proteínas em seu ambiente, facilitar a canalização de um substrato por uma via bioquímica ou construir máquinas moleculares, como o citoesqueleto. Assim, torna-se bastante óbvio que um mapa de interações entre proteínas de baixa e alta afinidade, sejam estas solúveis ou associadas a membranas, é essencial para o entendimento dos processos biológicos e dos mecanismos moleculares em termos de biologia de sistemas. Tal mapa deverá incluir pares de interações entre proteínas (também chamadas de interação binária), bem como grandes complexos proteicos. Esse conhecimento é um pré-requisito fundamental para o entendimento da maioria das funções celulares, especialmente as redes regulatórias e de sinalização.[28]

Nesse contexto, a análise e a construção de redes de interações entre proteínas se beneficiam enormemente de métodos passíveis de escalonamento em larga escala, como os ensaios de duplo-híbrido, a aplicação de *chips* de proteínas ou a purificação por afinidade de complexos estáveis seguida da identificação dos seus componentes de complexos por espectrometria de massas.[29]

No caso de redes de interação genes-proteína, as proteínas ligantes de DNA realizam uma série de funções importantes para a célula, incluindo a regulação da transcrição, a manutenção cromossômica, a replicação e a reparação de DNA. Por outro lado, as interações entre os fatores de transcrição e seus sítios de ligação ao DNA são de grande interesse, visto que essas interações controlam os passos cruciais no desenvolvimento de um organismo e na resposta a estresses de origem ambiental. Além disso, as disfunções relacionadas com os fatores de transcrição podem contribuir para a progressão de várias doenças em humanos.[30]

Outro aspecto que atrai o interesse dos pesquisadores que trabalham com controle da expressão gênica e, consequentemente, com análise de dados de interação entre proteínas e sequências de DNA é o universo de modificação pós-traducional de histonas e a localização dessas modificações na cromatina, o que pode afetar, de maneira significativa, a regulação gênica de um organismo. O estudo das interações entre proteínas e DNA tem sido facilitado pelas técnicas de microarranjos de DNA.[30] Apesar de não tão desenvolvidos quanto as técnicas associadas à análise de interações entre proteínas, os métodos de análise de interação DNA-proteínas são similares aos de interação entre proteínas. Por exemplo, a imunoprecipitação de cromatina com uma proteína de interesse seguida da detecção de fragmentos de DNA por meio de microarranjos (chamada de ChIP-chip) tem sido o método de larga escala mais usado para a identificação de sítios de ligação de fatores de transcrição.[23]

Outros métodos alternativos baseados em microarranjos também começaram a ser aplicados. Por exemplo, a fusão de proteínas que se ligam ao DNA com a enzima DNA adenina metiltransferase (Dam) possibilita identificar sítios de ligação em sequências de DNA por meio de metilação. O microarranjo ligante de proteínas é outro exemplo e propicia uma rápida caracterização (em larga escala) de proteínas ligantes de DNA.[31] Em todos os casos, deve-se salientar que os métodos em larga escala usados para a identificação de interações entre DNA e proteínas continuam na infância, e os exemplos de sua aplicabilidade na biologia de sistemas não são tão comuns quantos os de interação proteica.

Modularidade | Propriedade das redes biológicas e tecnológicas

Uma consequência inesperada da biologia de sistemas foi a observação de que as redes de interações são formadas por muitas sub-redes distintas, também chamadas de módulos.[22,32] Um módulo é definido, na área da engenharia, como uma estrutura funcional e padronizada que pode ser unida a outros módulos iguais ou a módulos com características diferentes, dando origem a um objeto mais complexo. Assim como os sistemas biológicos, módulos são encontrados em vários outros sistemas, como os círculos de amizades em redes sociais ou páginas da internet que tratam de tópicos similares.[22,32]

Estruturas altamente modulares são componentes fundamentais de sistemas de engenharia complexos, e a biologia está repleta de exemplos desse tipo de modularidade. Complexos estáveis formados por proteínas ou por proteínas e RNA constituem módulos centrais na definição de muitas funções biológicas básicas, como a síntese de ácidos nucleicos e a degradação proteica. Grupos de moléculas que são corregulados temporalmente atuam como módulos que definem as etapas do ciclo celular, a transdução de sinais externos em bactérias ou a resposta a feromônios em leveduras.

Para entender a natureza modular de uma rede e como os módulos estão relacionados a um determinado mecanismo fisiológico, ferramentas e medidas fornecidas pela Teoria dos Grafos são extremamente úteis. De uma maneira simplificada, pode-se definir um módulo em uma rede como um grupo de nós altamente interconectados.[20] Um exemplo prático da presença de módulos específicos em redes de interações proteicas pode ser observado na Figura 14.25.

Um método alternativo para o estudo de módulos envolve a identificação de grupos de nós altamente interconectados, diretamente da topologia da rede (ou seja, da estrutura da rede), seguida da correlação dessas entidades com as funções bioquímicas que são potencialmente desempenhadas por essas estruturas.

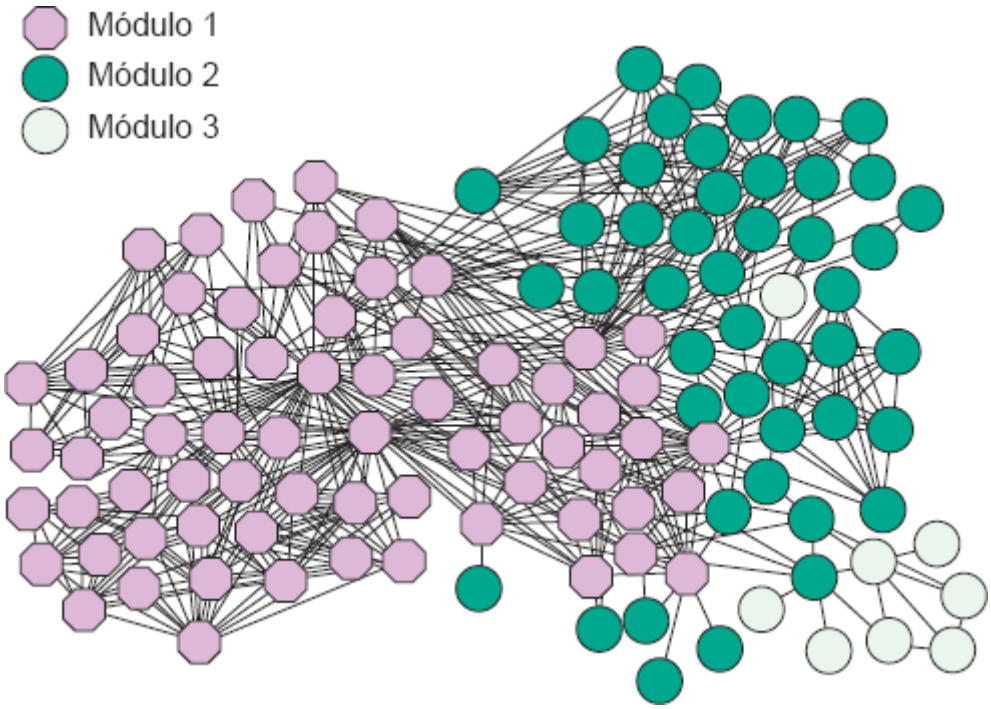


Figura 14.25 Exemplo de uma rede de interação mostrando os respectivos módulos componentes.

Linguagem matemática da biologia de sistemas

A Teoria dos Grafos oferece a linguagem e o simbolismo necessários para a descrição dos diferentes níveis de complexidade de um organismo. Como discutido anteriormente, a representação das informações sobre os componentes de um sistema biológico e sobre suas interações em modelos de grafos possibilita tanto a modelagem da dinâmica do sistema como o estudo exploratório de suas propriedades globais.

A Teoria dos Grafos se baseia no estudo de estruturas geométricas conhecidas como grafos ou redes, formadas por um conjunto de nós "V" e linhas "E" que conectam esses nós. Assim, os nós "V" são as entidades de interesse e as linhas (ou conectores) "E" representam as relações entre essas entidades.[20]

Cada elemento "E" contém um par de elementos "V". Cada conector pode ter pesos e direções ou representar diferentes tipos de dados experimentais (Figura 14.26).

Nesta seção, serão introduzidas certas medidas úteis para entender a estrutura de um grafo ou rede, como o grau ou conectividade de cada nó, a distribuição dos graus, os expoentes do grau, o caminho mais curto e o caminho médio e, por fim, o coeficiente de agrupamento. Existem outras medidas utilizadas para a análise de grafos, mas as que foram citadas aqui são as mais utilizadas e geram informações importantes a respeito do sistema em que o pesquisador está trabalhando.

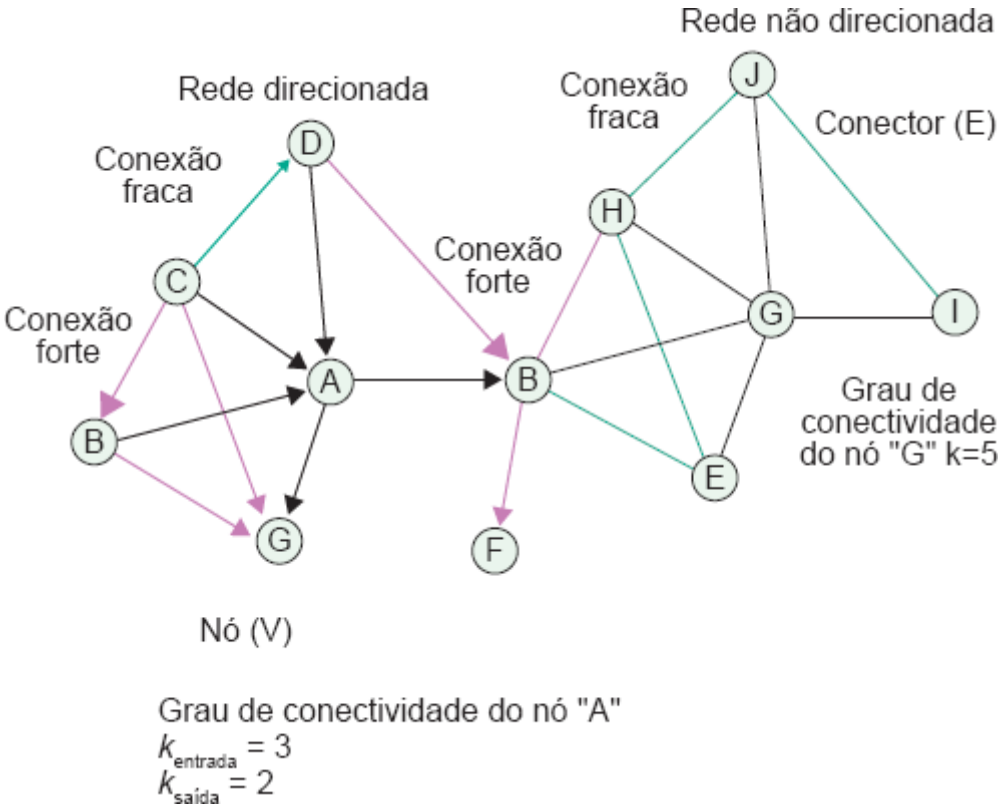


Figura 14.26 Exemplo de uma rede direcionada e uma não direcionada. Os graus de conectividade para os nós "A" e "G" estão indicados na figura.

O grau ou a conectividade é a característica mais elementar de um nó e corresponde ao número de ligações ou conectores que um dado nó tem com os outros nós. O grau de um nó é, por convenção, representado pela letra *k*.

Considerando os tipos de conexão, porém, os grafos podem ser considerados de dois tipos: não direcionados ou direcionados. No primeiro, os conectores não apresentam uma ordem específica de entrada ou saída, enquanto, no segundo, os nós podem ter conectores que entram ou que saem deles (ver Figura 14.26). Quando existe uma direção específica, o número de conectores que entram no nó é chamado $k_{entrada}$, enquanto o número de conectores que saem do nó é identificado como $k_{saída}$ (ver Figura 14.26).

Outro aspecto importante das redes não direcionadas é a possibilidade de calcular o seu grau "médio" ou $\langle k \rangle$ (na linguagem matemática, $\langle\ \rangle$ denota a média), considerando seus nós (N) e conectores (ℓ). Para tanto, a fórmula $\langle k \rangle = 2\ \ell / N$ pode ser aplicada de um modo simples e que dá uma ideia do número médio de conectores por nó.

Outra propriedade importante das redes é a chamada distribuição de graus, ou $P_{(k)}$. $P_{(k)}$, que corresponde à frequência relativa dos nós para cada valor de grau, ou seja, ao número de nós para cada valor de k, ou $N(k)$, em que $k = 1, 2, 3, ...$, seguido pela divisão pelo número total de nós N. Essa importante distribuição é usada para distinguir as diferentes classes de redes. Um bom exemplo dessa aplicação é que, nas redes nas quais existem alguns poucos nós que apresentam um grau muito elevado e que são, portanto, altamente conectados com outros nós pouco conectados, esse nós são denominados *hubs*, um termo emprestado da computação que define os centros de controle ou de distribuição de dados em uma rede qualquer. A presença de *hubs* implica a existência distinta de uma estrutura e fluxo de informações na rede do meramente aleatório.

Uma terceira propriedade matemática das redes, conhecida pelo expoente do grau, é também muito importante e relacionada com a distribuição do grau. A maioria das redes biológicas são ditas livres de escalas, pois a distribuição dos graus nelas segue uma lei de potenciação, ou $P_{(k)}$cerca de $k^{-\gamma}$, em que γ é o expoente do grau e "cerca de" significa "proporcional a".

Os valores de γ determinam as propriedades do sistema, pois, quanto menor o valor de γ, mais importante é o papel dos *hubs* na rede. Por exemplo, para valores de $\gamma > 3$, os *hubs* não são relevantes; $2 > \gamma > 3$ indica uma hierarquia de *hubs,* em que o *hub* mais conectado está em contato com uma pequena fração de todos os nós. Essa propriedade matemática possibilita definir algumas das características mais importantes das redes livres de escala, como o grau de robustez de uma rede, que está associado à resistência da rede a uma falha ou destruição de um ou vários nós.[20,22]

Um ponto essencial a respeito da abordagem matemática das redes é que, assim como para modelos matemáticos em qualquer outra área da ciência, um modelo matemático precisa ser julgado com base nos resultados experimentais obtidos. Assim, métodos de agrupamento e outros testes estatísticos foram desenvolvidos para a análise de informações "ômicas".

Os modelos baseados em mapas de auto-organização e cadeias de Markov são propostas avançadas para a análise de dados, e o uso de tais modelos possibilita o desenvolvimento de algoritmos com um bom poder preditivo.[33,34] Apesar do valor desses modelos para a reconstrução de vias bioquímicas, todos são dependentes de dados biológicos obtidos por vias experimentais e não podem ser usados para simulações, como o efeito de diferentes condições ambientais no crescimento celular. Para simular a resposta dinâmica de sistemas biológicos a diferentes estímulos ou diferentes perturbações do sistema, são empregados modelos cinéticos. Esses modelos ainda são bastante raros nas ciências biológicas, visto que dependem de uma enorme quantidade de informação proveniente de experimentos baseados em cinética enzimática ou proteica. A maioria desses modelos é construída utilizando o balanço de massas para os componentes individuais por meio de equações diferenciais. Além disso, a cinética das diferentes reações e processos é descrita por equações matemáticas bastante complexas. Alguns exemplos encontrados atualmente na literatura caracterizam a biologia de sistemas "de baixo para cima", como visto anteriormente, mas os modelos apresentados até agora são sistemas relativamente pequenos, pelo menos da perspectiva de uma célula como todo.[35] Assim, não são úteis para a integração de dados "ômicos" e são incapazes de descrever as funções gerais de uma célula. Contudo, espera-se que, em longo prazo, seja possível coletar todos os parâmetros cinéticos e definir as equações matemáticas para todas as reações e processos em uma célula, propiciando a montagem de um modelo cinético completo para ela.

Considerações finais

A utilização de métodos automáticos de geração de dados em larga escala já é uma rotina e está mudando o *status quo* dos estudos em ciências da vida. A bioinformática, ou biologia computacional, chega para transformar essa enorme quantidade de dados biológicos em informações relevantes para estudos básicos e aplicados. Tais informações estão revolucionando diversas áreas, que se estendem da biologia à medicina, passando por ciências da computação, ciências farmacêuticas, entre outras. Na bioinformática, a análise de dados biológicos depende de diversas metodologias, as quais podem utilizar portais de integração de informações, representação e análise de dados, alinhamento de sequências, estudo de genomas e biologia de sistemas, por exemplo. Essas ferramentas e metodologias computacionais integram informações geradas por diversas fontes (em grande ou pequena escala), processam-nas e podem gerar modelos refinados sobre os mecanismos relacionados com o ambiente, organismo, a patologia ou a via celular da qual os dados foram extraídos. Essa maneira holística de se estudar dados biológicos vem trazendo conhecimento e avanços em uma velocidade nunca antes registrada.

Referências bibliográficas

1. Metzker ML. Sequencing technologies: the next generation. Nat Rev Genet. 2010;11(1):31-46.
2. Chen J. Digital code of life: how bioinformatics is revolutionizing science, medicine, and business. Briefings in Bioinformatics. 2004;5(3):305-7.
3. McEntyre J. Linking up with Entrez. Trends Genet. 1998;14(1):39-40.
4. Berman HM, Kleywegt GJ, Nakamura H, Markley JL. The Protein Data Bank at 40: reflecting on the past to prepare for the future. Structure. 2012;20(3):391-6.
5. Li L, Goedegebuure P, Mardis ER, Ellis MJ, Zhang X, Herndon JM, et al. Cancer genome sequencing and its implications for personalized cancer vaccines. Cancers. 2011;3(4): 4191-211.
6. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860-921.
7. Flicek P, Birney E. Sense from sequence reads: methods for alignment and assembly. Nat Methods. 2009;6(11 Suppl):S6-12.
8. Yandell M, Ence D. A beginner's guide to eukaryotic genome annotation. Nat Rev Genet. 2012;13(5):329-42.
9. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57-74.

10. Graur D, Zheng Y, Price N, Azevedo RB, Zufall RA, Elhaik E. On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. Genome Biol Evol. 2013;5(3):578-90.
11. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Mol Biol. 1990;215(3):403-10.
12. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acid Res. 1997;25(17):3389-402.
13. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol. 1981;147(1):195-7.
14. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A. 1988;85(8):2444-8.
15. Hernandez T, Kambhampati S. Integration of biological sources: current systems and challenges ahead. SIGMOD Record. 2004;33(3):51-60.
16. You L. Toward computational systems biology. Cell Biochem Biophys. 2004;40:167-84.
17. Viswanathan GA, Seto J, Patil S, Nudelman G, Sealfon SC. Getting started in biological pathway construction and analysis. PLoS Comput Biol. 2008;4(2):e16.
18. Borodina I, Nielsen J. From genomes to in silico cells via metabolic networks. Curr Opin Biotechnol. 2005;16(3):350-5.
19. Baxevanis A. The molecular biology database collection: 2003 update. Nucleic Acid Res. 2003;31:1-12.
20. Albert R, Barabási AL. Statistical physics of complex networks. Rev Mod Phys. 2002;74:47-97.
21. Huber W, Carey VJ, Long L, Falcon S, Gentleman R. Graphs in molecular biology. BMC Bioinformatics. 2007;8,S8.
22. Barabási AL, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet. 2004;5:101-13.
23. Cary MP, Bader GD, Sander C. Pathway information for systems biology. FEBS Lett. 2005;579:1815-20.
24. Bruggeman FJ, Westerhoff HV. The nature of systems biology. Trends Microbiol. 2007;15(1):45-50.
25. Sako Y. Imaging single molecules in living cells for systems biology. Mol Syst Biol. 2006;2:56.
26. Mashanov GI, Nenasheva TA, Peckham M, Molloy JE. Cell biochemistry studied by single-molecule imaging. Biochem Soc Trans. 2006;34:983-8.
27. Levy ED, Pereira-Leal JB. Evolution and dynamics of protein interactions and networks. Curr Opin Struct Biol. 2008;18:349-57.
28. Bader S, Kühner S, Gavin AC. Interaction networks for systems biology. FEBS Lett. 2008;582:1220-4.
29. Lalonde S, Ehrhardt DW, Loqué D, Chen J, Rhee SY, Frommer WB. Molecular and cellular approaches for the detection of protein-protein interactions: latest techniques and current limitations. Plant J. 2008;53:610-35.
30. Hanlon SE, Lieb JD. Progress and challenges in profiling the dynamics of chromatin and transcription factor binding with DNA microarrays. Curr Opin Genet Dev. 2004;14:697-705.
31. Berger MF, Bulyk ML. Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. Nat Protoc. 2009;4:393-411.
32. Albert R. Scale-free networks in cell biology. J Cell Sci. 2005;118:4947-57.
33. Kohonen T. The self-organizing map. Proceedings of IEEE 78. 1464-80.
34. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 2002;30:1575-84.
35. Stephanopoulos G, Alper H, Moxley J. Exploiting biological complexity for strain improvement through systems biology. Nat Biotechnol. 2004;22:1261-7.

15 Genoma de Organelas | Cloroplastos e Mitocôndrias

Helaine Carrer, Jeferson Gross, Marcos Roberto Chiaratti e Flávio Vieira Meirelles

Cloroplastos e mitocôndrias são organelas subcelulares com seus próprios genomas e sistema genético. Ambas as organelas se originaram independentemente, pela endossimbiose de proteobactérias de vida livre, há mais de 1,4 bilhão de anos. Pela teoria endossimbiótica, as a-proteobactérias deram origem às mitocôndrias e as cianobactérias originaram os plastídeos. Ao longo da evolução, os genomas das duas organelas sofreram uma redução massiva do material genético pela deleção e transferência de genes para o núcleo da célula hospedeira. Nos plastídeos das células vegetais, ocorrem as reações da fotossíntese e síntese de aminoácidos e ácidos graxos. Nas mitocôndrias, organelas presentes em todas as células eucariotas, estão os genes componentes da respiração celular. Embora essas organelas apresentem similaridades em relação à bioenergia da célula que são essenciais para a sobrevivência dos organismos, a evolução dos genomas de cloroplastos e mitocôndrias apresenta variabilidade de formas, tamanho e organização estrutural dos genes. Esses tópicos, associados aos avanços tecnológicos de genética reversa com aplicações biotecnológicas, são abordados neste capítulo.

Introdução

Entre as principais diferenças entre células procariotas e eucariotas está a compartimentação dos componentes celulares nas células eucariotas com a presença das organelas, os cloroplastos e as mitocôndrias (Figura 15.1 A). As mitocôndrias estão presentes em todas as células eucariotas, enquanto os plastídeos estão presentes somente em plantas e algas. Essas organelas são responsáveis pela produção de energia pelas células. Nos plastídeos, ocorrem reações vitais para a célula vegetal, como fotossíntese, síntese de aminoácidos e produção de ácidos graxos e vitaminas (Figura 15.1 B), enquanto, nas mitocôndrias, ocorrem a respiração e gliconeogênese (Figura 15.1 C).

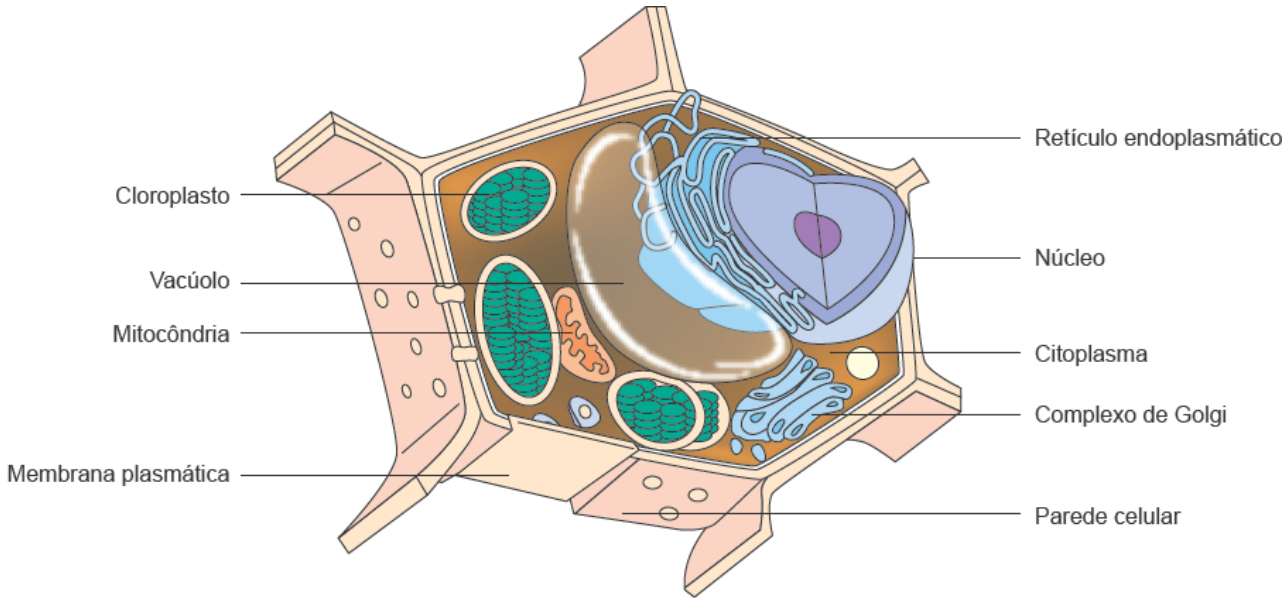
Ambas as organelas originaram-se independentemente pela endossimbiose de proteobactérias de vida livre que foram engolfadas por uma célula pré-eucariota há mais de 1,4 bilhão de anos, as quais, por algum problema enzimático, não foram digeridas pelo organismo hospedeiro (Figura 15.2 A). É considerado que as mitocôndrias se originaram antes dos plastídeos, dando origem às células eucariotas ancestrais. Assim, pela teoria endossimbiótica, as a-proteobactérias deram origem às mitocôndrias e as cianobactérias originaram os plastídeos (Figura 15.2 B). Essas hipóteses foram apresentadas pela primeira vez em 1883, por Andreas Schimper, em referência aos cloroplastos, que teriam sido originados de endossimbiontes. Pouco depois, a mesma hipótese foi apresentada para as mitocôndrias por Richard Altmann, em 1890, e, em seguida, formalizada por Konstantin Mérejkowsky, em 1905, para os cloroplastos, seguido por Ivan Wallin, em 1920, sobre as mitocôndrias. Contudo, essa teoria foi rejeitada pelos biologistas da época e levada ao esquecimento. Somente foi reintroduzida por Lynn Margulis, em 1966, após a descoberta da presença de DNA em plastídeos de *Chlamydomonas reinhardtii*, uma alga unicelular, e em mitocôndrias de frangos, utilizando microscopia eletrônica (Hagemann, 2010).

Na Figura 15.2 B, a espessura das setas é aproximadamente proporcional à quantidade de informação transferida em um dado período da evolução. As mitocôndrias surgiram pela endossimbiose de uma a-proteobactéria em uma célula pré-eucariota. A conversão da a-proteobactéria em uma mitocôndria foi acompanhada de massiva translocação de informação genética para o genoma nuclear. Em uma segunda endossimbiose, uma cianobactéria foi também engolida e gradualmente convertida em um plastídeo (cloroplasto). Essa segunda endossimbiose foi novamente seguida de transferência de informação do genoma endossimbionte para o genoma nuclear da célula

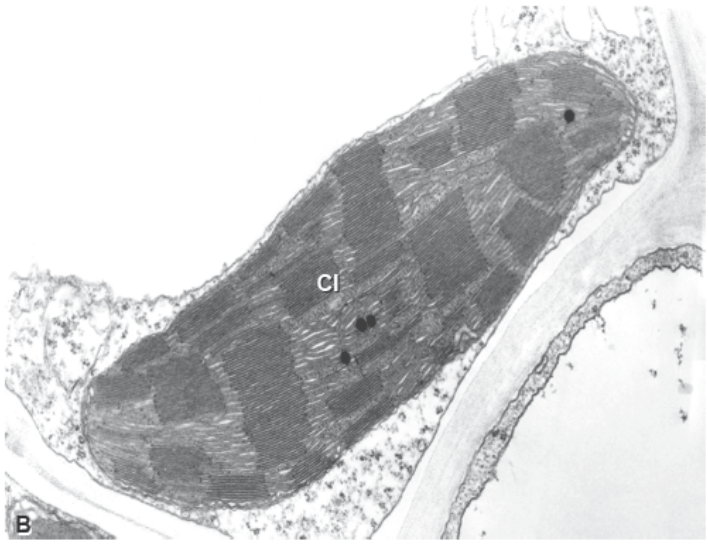
hospedeira. A transferência de genes da mitocôndria para o núcleo foi continuada em menor escala. Genes nucleares e sequências de DNA plastidial foram transferidos para o genoma mitocondrial, mas o genoma plastidial demonstra não ter a capacidade de receber sequências de DNA, uma vez que nenhuma sequência de DNA nuclear ou de mitocôndria foi identificada na organela.

Devido ao fato de que cloroplastos e mitocôndrias têm seu próprio genoma, advindo do organismo endossimbionte, ao longo da evolução houve um processo de integração gradual do endossimbionte no metabolismo da célula hospedeira, e o genoma das organelas sofreu uma dramática redução de tamanho, decorrente da massiva deleção e transferência de genes para o genoma nuclear, processo este que continua ocorrendo até a atualidade. Consequentemente, o genoma das organelas, hoje, é pequeno, contendo os genes essenciais que possibilitam a manutenção de seu próprio sistema genético e síntese de proteínas para expressar as características herdáveis da organela. Como resultado, cloroplastos e mitocôndrias são altamente dependentes de proteínas codificadas no núcleo, que são direcionadas às organelas para a manutenção de vias bioquímicas essenciais e de funções específicas, reparo e expressão gênica.

As organelas têm várias semelhanças com organismos procarióticos, uma forte evidência de que realmente tiveram um ancestral comum. Essas semelhanças incluem a localização do DNA em nucleoides, a forma circular do genoma (apesar de a presença de cópias lineares também ser observada), a organização dos genes em operons, como do gene *rpl*23, em cloroplastos semelhante ao correspondente em cianobactérias e o mecanismo procariótico de expressão gênica diferente dos eucariotos. Entretanto, na atualidade, são encontrados processos novos, sem similaridade com os organismos procarióticos nas organelas, como a presença de fatores transcricionais e pós-traducionais eucarióticos nos processos de regulação gênica. Assim, após tantas modificações durante o processo evolutivo, os plastídeos e as mitocôndrias não conseguem existir como organismos autônomos fora de uma célula eucariota.



A



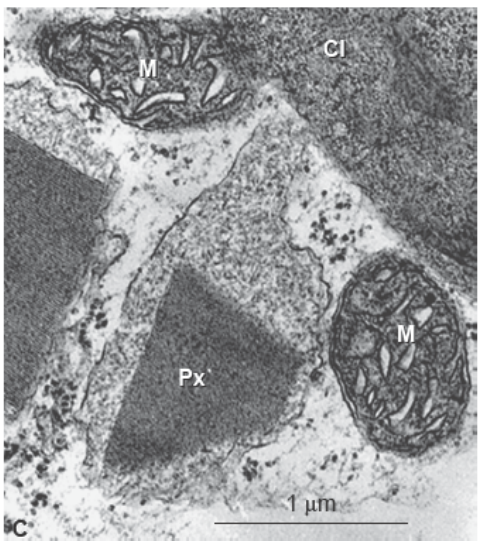


Figura 15.1 Célula vegetal e seus componentes estruturais. **A.** Esquema de uma célula vegetal. **B.** Microscopia eletrônica de células parenquimáticas de cloroplasto de cana-de-açúcar variedade SP80-3280. **C.** Microscopia eletrônica de mitocôndrias de células do parênquima de tomate. Px: peroxissoma; Cl: cloroplasto; M: mitocôndria. Imagens cedidas por Elliot Watanabe Kitajima (ESALQ/USP).

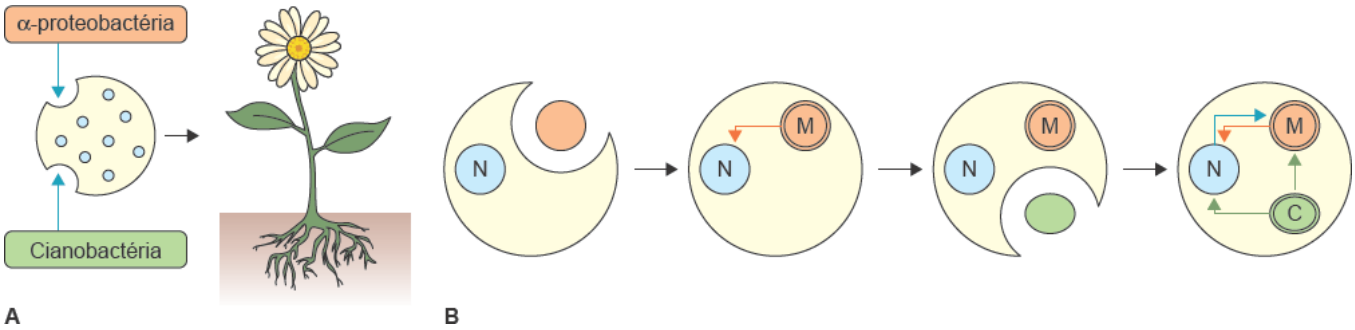


Figura 15.2 Origem das organelas em células eucariotas pela endossimbiose de proteobactérias de vida livre. **A.** As mitocôndrias resultaram do processo de endossimbiose de uma a-proteobactéria em uma célula primitiva, enquanto os cloroplastos surgiram a partir de cianobactérias. **B.** Transferência de genes entre genomas na evolução de células eucariotas. As setas pretas entre cada célula indicam a direção de transferência dos genes, e as setas coloridas correspondem à cor do compartimento do qual a transferência do material genético se originou (laranja: mitocôndria; azul: núcleo; verde: plastídeo).

Nas últimas décadas, o avanço da tecnologia de sequenciamento de DNA por meio de equipamentos de alta produtividade possibilitou que genomas de organelas fossem extensivamente sequenciados. O primeiro genoma mitocondrial de humano e camundongo foi totalmente sequenciado em 1981, e o genoma cloroplastidial da briófita *Marchantia polymorpha* e do tabaco foi sequenciado em 1986. Desde então, em pouco mais de 30 anos, milhares de outros genomas de organelas foram sequenciados. Estão disponíveis aproximadamente 5 mil genomas completos de mitocôndrias de diferentes organismos e mais de 800 genomas de cloroplastos de algas e plantas. A Figura 15.3 apresenta a distribuição de genomas completamente sequenciados de organelas de plantas e algas. As informações obtidas no sequenciamento têm contribuído profundamente para os estudos de genética, evolução e biologia molecular de organelas. A estrutura dos genomas de organelas e a organização e a expressão de genes têm sido intensamente estudadas e serão apresentadas a seguir neste capítulo.

O genoma cloroplastidial de plantas

É designado como plastoma o genoma cloroplastidial dos organismos fotossintetizantes. Em plantas superiores, o plastoma é composto de um único cromossomo de dupla-fita de DNA. É altamente poliploide, pois cada célula apresenta muitos plastídeos, e cada plastídeo, por sua vez, apresenta várias cópias do genoma plastidial. As várias cópias do DNA ficam densamente organizadas em estruturas de nucleoproteínas chamadas nucleoides. A visualização dos nucleoides por microscopia de fluorescência ocorre após serem corados com DAPI (4',6-diamino-2-fenilindol) (Figura 15.4). Foi demonstrado que, geralmente, ficam ancorados às membranas dos tilacoides e nas membranas internas dos cloroplastos. Dependendo da espécie, tecido, estágio de desenvolvimento e condições ambientais, o nível de poliploidia pode chegar a mais de 10 mil cópias do plastoma por célula. O número de cópias geralmente é maior em células fotossinteticamente ativas, nas quais os plastídeos estão diferenciados em cloroplastos (Figura 15.5). Especula-se que a grande quantidade de DNA encontrada nos cloroplastos decorre da alta demanda de proteínas envolvidas na fotossíntese, uma vez que muito pouco DNA é encontrado nos plastídeos de tecidos não fotossintetizantes, como os amiloplastos em raiz. Assim que a diferenciação do plastídeo se completa, o número de cópias do genoma é mantido constante e não se altera significativamente com a idade ou estágio de desenvolvimento da planta.

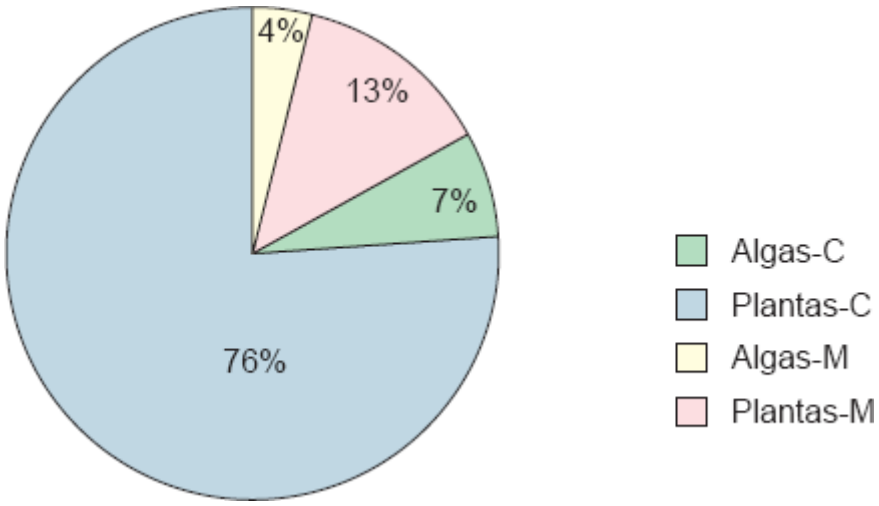


Figura 15.3 Distribuição taxonômica dos genomas mitocondriais e plastidiais completos de plantas e algas depositadas no GenBank até maio de 2017. O número total de genomas de organelas (cloroplastos e mitocôndrias) sequenciados é 7.283. O número total de organelas de algas e plantas é de 1.555. Fonte: http://www.ncbi.nlm.nih.gov/genome/browse/?report=5.

Nas células meristemáticas, a organela se apresenta na forma de proplastídeo, que tem somente um nucleoide, com poucas cópias do genoma. Durante o processo de diferenciação, os proplastídeos das células meristemáticas sofrem diferenciação de acordo com sua função. Os proplastídeos se diferenciam em etioplastos em tecidos com pequena atividade fotossintética ou em cloroplastos nas células dos tecidos fotossinteticamente ativos. Os cromoplastos, presentes nas flores e nos frutos, diferenciam-se a partir dos cloroplastos ou diretamente a partir dos proplastídeos e acumulam carotenoides. Outras formas de plastídeos especializados, como os amiloplastos (armazenam amido) e os leucoplastos (com função de armazenar lipídeos e proteínas), estão presentes em células de raiz ou tecidos específicos (Figura 15.6). O aumento do número de organelas por célula e as mudanças no número de cópias do plastoma por organela ocorrem durante o processo de diferenciação dos plastídeos no desenvolvimento da planta. Todas as cópias do plastoma são idênticas, independentemente dos tipos de diferenciação dos plastídeos.

A Figura 15.6 apresenta a diferenciação estrutural dos plastídeos. Todas as formas de plastídeos são derivadas de um plastídeo progenitor, chamado proplastídeo, presente em células meristemáticas. Os proplastídeos têm a capacidade de se converter em diferentes tipos de acordo com o tipo de molécula ou pigmentos que acumulam ou com sua propriedade bioquímica e fisiológica. Em células fotossintetizantes, os proplastídeos são convertidos em etioplastos no início da diferenciação ou diretamente em cloroplastos, com a produção de pigmentos envolvidos na fotossíntese, os quais, por sua vez, podem ser convertidos em cromoplastos em tecidos de flores ou frutos. Nos tecidos não fotossintetizantes, os proplastídeos são convertidos em leucoplastos, que se diferenciam em amiloplastos ou elaioplastos, com a capacidade de armazenar amido e lipídeos, respectivamente. Todas as formas de diferenciação dos plastídeos podem retornar à forma de proplastídeos no processo de desdiferenciação das células que se reverterem ao estágio meristemático. Cada célula meristemática tem de 5 a 20 proplastídeos. Em folhas, diferenciam-se em cloroplastos e dividem-se, resultando em 30 a 100 cloroplastos por célula. O genoma de cloroplastos e mitocôndrias é apresentado como estrutura circular, observada por microscopia eletrônica. Entretanto, estudos mais recentes, que combinaram microscopia e análise de eletroforese em gel de campo pulsado, demonstram que existe uma grande plasticidade estrutural desses genomas, sendo que somente 25 a 40% das moléculas são circulares; várias conformações, desde linear a formas multiméricas que lembram concatâmeros, são observadas, e até mesmo a forma circular se apresenta em várias conformações, mais ou menos contorcidas. A relevância funcional das diferentes conformações não está estabelecida; algumas formas demonstram ser somente intermediárias da replicação, mas, possivelmente, estão associadas à regulação da expressão gênica.

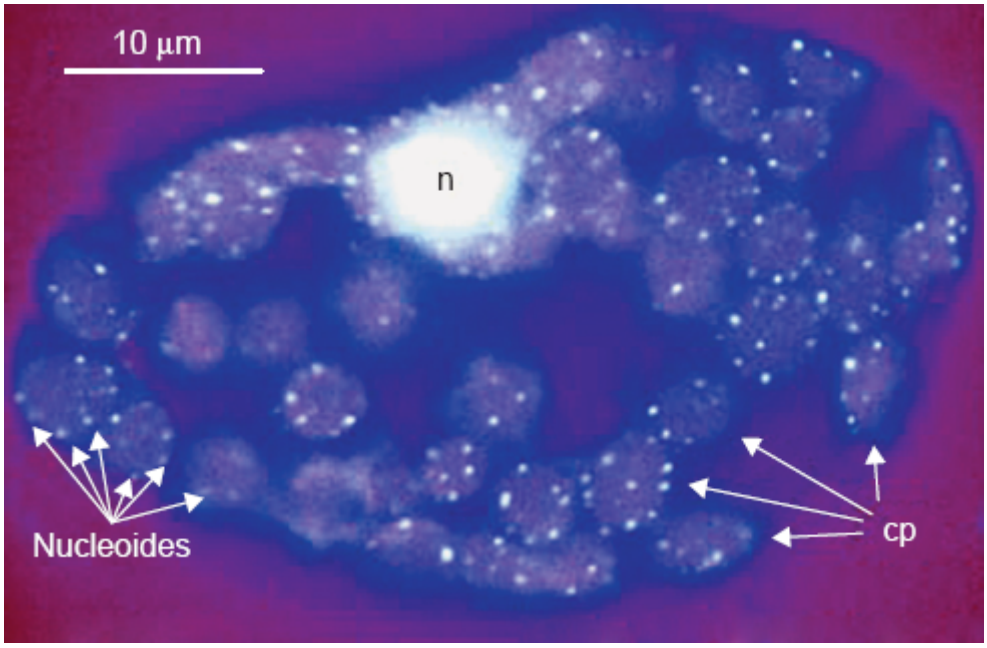


Figura 15.4 Célula do mesófilo de *Spinacea oleracea* corada com DAPI. A figura ilustra uma célula de folha mostrando grande quantidade de DNA no núcleo diploide (2n) e várias cópias de DNA nos cloroplastos (cp). Múltiplas cópias do genoma plastidial estão organizadas nos nucleoides. Uma única célula fotossinteticamente ativa apresenta várias dezenas de cloroplastos, e cada um apresenta múltiplas cópias do genoma plastidial nos nucleoides. Existe grande variação na proporção de DNA cloroplastidial entre espécies. As moléculas de clorofilas dentro dos plastídeos mostram fluorescência vermelha. As mitocôndrias não estão visíveis nessa preparação. A barra representa 10 µm. DAPI (4',6-diamidino-2-fenilindol). Fonte: Bock e Timmis, 2008.

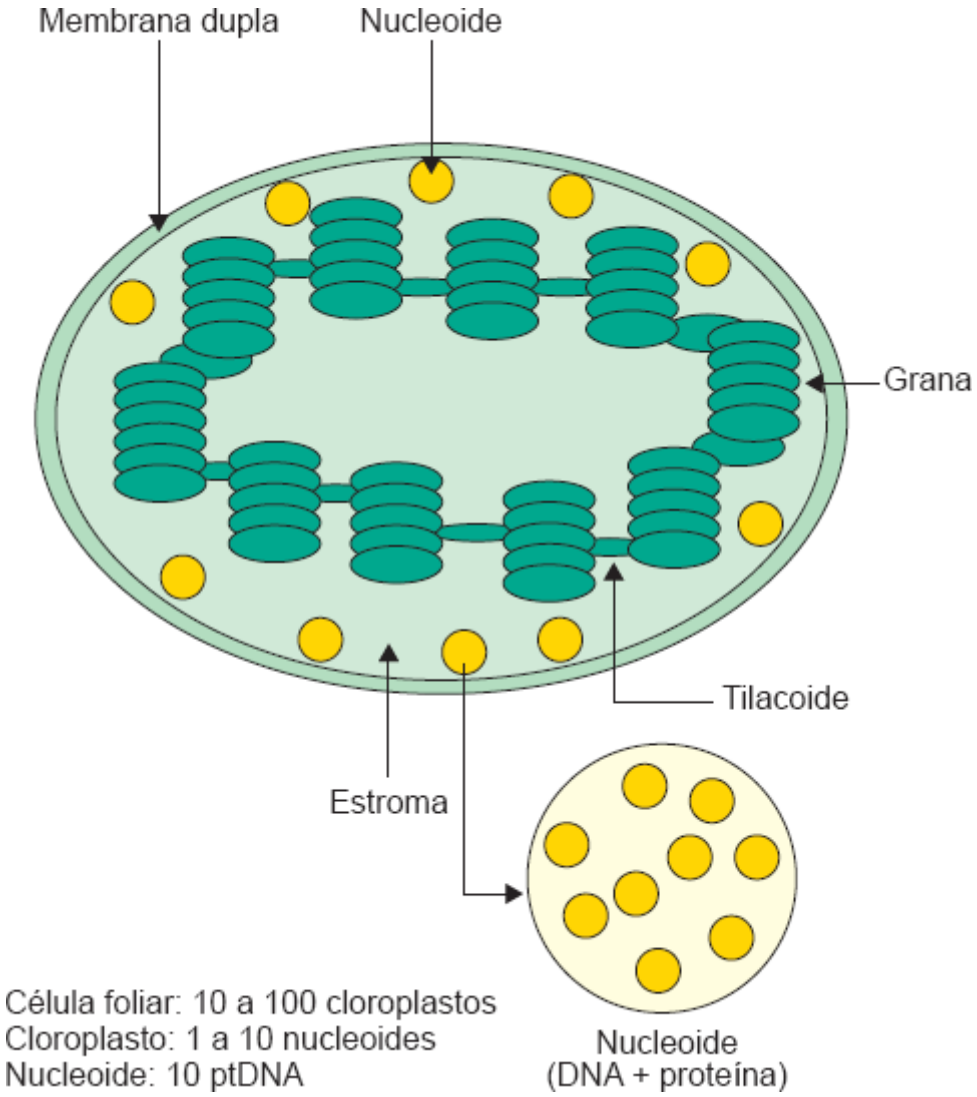


Figura 15.5 Estrutura organizacional de um cloroplasto. Em células fotossintetizantes, os plastídeos estão diferenciados em cloroplastos. Essa organela tem duas membranas lipoproteicas, sendo a membrana interna remanescente do organismo ancestral e a membrana externa resultante do processo endossimbiótico do engolfamento na célula hospedeira. O cloroplasto tem um conjunto de membranas internas, os tilacoides e o grana, onde ocorrem as reações da etapa fotoquímica da fotossíntese. No estroma (matriz cloroplastidial) estão as enzimas responsáveis pelo processo de fixação e redução do CO_2 durante a fotossíntese. Os cloroplastos são altamente poliploides. As moléculas do DNA ficam inseridas em estruturas proteicas (esse complexo de DNA e proteína é chamado nucleoide). Os nucleoides ficam associados às membranas internas, raramente estão dispersos na organela. O número de nucleoides e genomas por célula (ptDNA) é variável nos diferentes tecidos da planta e entre espécies.

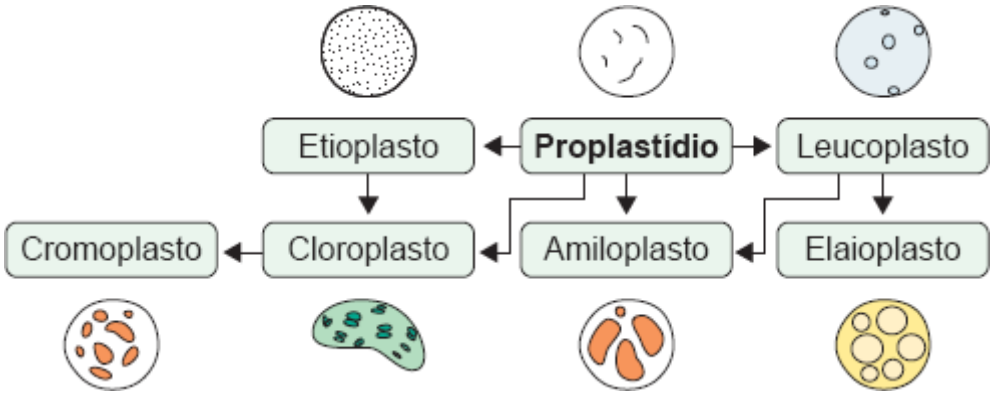


Figura 15.6 Diferenciação estrutural dos plastídeos.

Os primeiros genomas cloroplastidiais de plantas superiores sequenciados foram o do tabaco (Shinozaki *et al*., 1986) e o do arroz (Hiratsuka *et al*., 1989). Em 30 anos, mais de 800 plastomas de plantas superiores foram disponibilizados. Em geral, o genoma cloroplastidial é conservado. O tamanho varia entre 120 e 160 kb e codifica aproximadamente 100 genes. Apesar do pequeno tamanho em comparação com o genoma nuclear de plantas (no mínimo 1.000 vezes maior), o conteúdo de DNA cloroplastidial representa de 10 a 20% do conteúdo de DNA celular total. Essa alta quantidade é devida à poliploidia do genoma cloroplastidial.

Organização estrutural do genoma plastidial

Em geral, genomas de plastídeos têm baixo conteúdo GC, variando entre 30 e 40%. Nas regiões codificadoras, a presença de A e T se manifesta como preferencial sobre G ou C na terceira posição dos códons sinônimos. A predominância de A e T é ainda mais evidenciada nos espaços intergênicos não codificadores e pode chegar a 80% dos nucleotídeos.

O genoma plastidial da maioria das espécies apresenta quatro segmentos, uma região única longa (LSC) e uma região única curta (SSC), separando duas regiões repetidas invertidas (RI_A e RI_B) (Figura 15.7). As duas regiões repetidas invertidas (RI) são idênticas na sua sequência de nucleotídeos e, tipicamente, contêm um grupo de quatro genes rRNA (4,5S, 5S, 16S e 23S rRNA), cinco genes tRNA (*trnA*-UGC, *trnI*-GAU, *trnN*-GUU, *trnR*-ACG e *trnV*-GAC) e outros genes resultantes de expansões e contrações específicas da espécie. Cada gene contido nessa sequência está presente em duas cópias por genoma, diferindo somente na orientação. As sequências das bordas entre as RI e as regiões de cópias únicas são, de alguma forma, variáveis mesmo entre espécies próximas. A variação de tamanho dos genomas em plantas muitas vezes é determinada pelo tamanho das RI, que variam de 0,5 a 75 kb em comprimento. A região repetida invertida mais longa identificada em plantas ocorre no gerânio (*Pelargonium*), uma espécie de planta com o maior plastoma (217 kb) já sequenciado. Apesar de a grande maioria das espécies de plantas superiores apresentarem as regiões RI na estrutura do plastoma, coníferas e algumas leguminosas não apresentam essa estrutura, o que indica que a região RI não é essencial para a manutenção do funcionamento do plastoma (Figura 15.8).

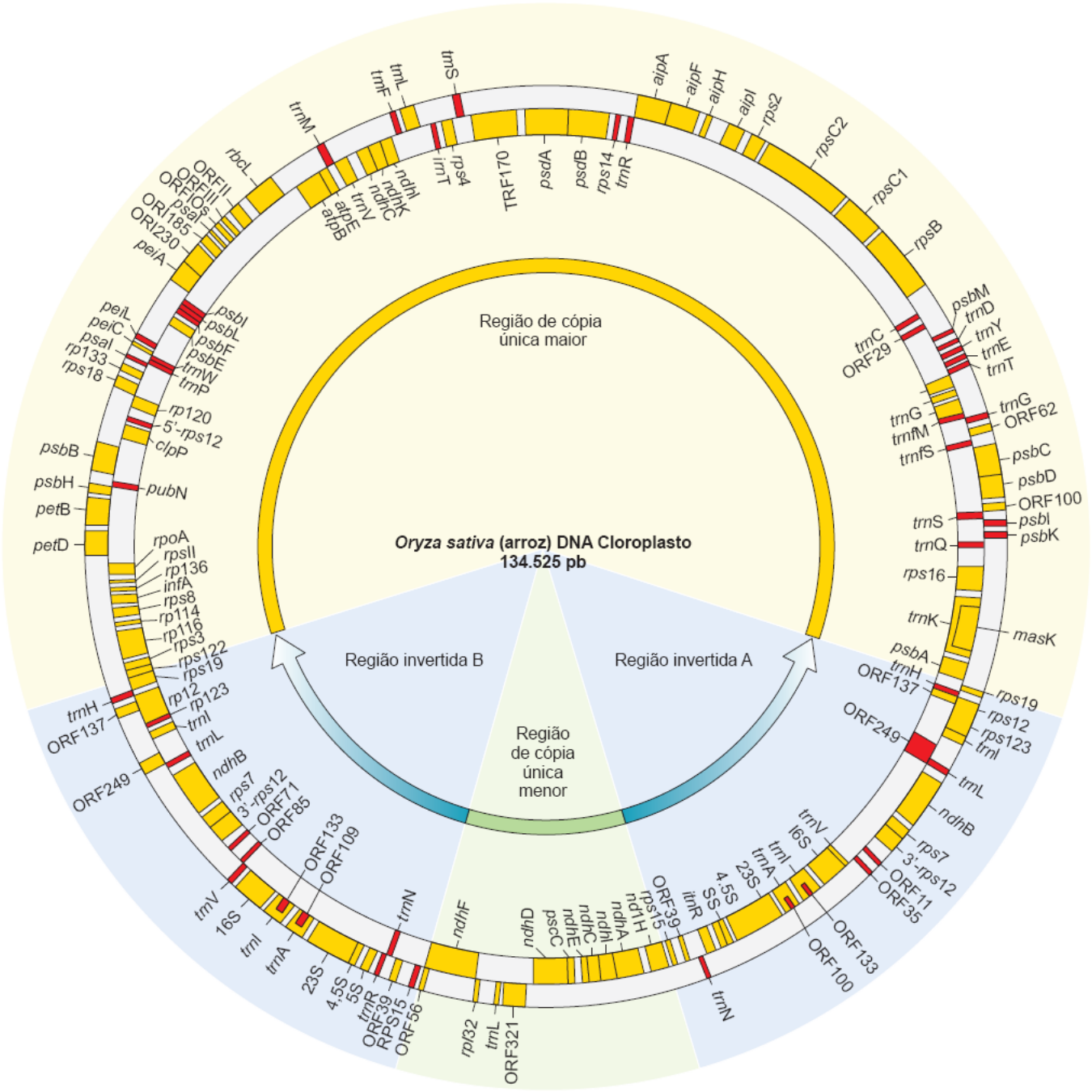


Figura 15.7 Mapa do genoma cloroplastidial do arroz (*Oryza sativa*). Os genes mostrados na parte interna do círculo são transcritos no sentido horário e os mostrados na parte externa são transcritos no sentido anti-horário. O genoma é subdividido em quatro partes: uma região de cópia única maior (LSC), uma região de cópia única menor (SSC) e duas regiões repetidas invertidas na maioria das espécies vegetais.

A função da presença das regiões RI em duas cópias não é clara; possivelmente, o aumento da quantidade de genes altamente expressos, como os genes RNA ribossômicos, e a estabilidade do genoma pela baixa frequência de mutações nas regiões RI, em comparação com os genes localizados nas regiões de cópias simples do plastoma, são considerados os principais benefícios dessas regiões duplicadas. Observa-se, ainda, que a presença das RI tem seu principal impacto na taxa de evolução das sequências dos plastomas. Tem sido demonstrado que as taxas de substituições sinônimas e não sinônimas são várias vezes menores nas RI em comparação com as regiões de cópias únicas entre as angiospermas.

Conteúdo gênico dos genomas plastidiais

O genoma plastidial remanescente evolucionário das cianobactérias sofreu expressiva redução de tamanho pela deleção ou transferência de DNA para o genoma nuclear. Em consequência, os plastomas contemporâneos têm somente uma pequena proporção dos genes dos seus ancestrais. Enquanto o genoma das cianobactérias *Synechocystis* contém mais de 3 mil genes, os genomas de plantas contêm entre 100 e 120 genes.

A limitada capacidade codificadora do plastoma é insuficiente para fornecer os milhares de componentes necessários para suportar seu sistema de expressão gênica, fotossíntese, síntese de aminoácidos, vitaminas e todas as muitas outras funções metabólicas localizadas nos plastídeos. Assim, as organelas são estritamente dependentes dos produtos de genes nucleares sintetizados por ribossomos citoplasmáticos, que, após a tradução, são transportados para as organelas. O proteoma plastidial é composto, em sua maioria, por proteínas codificadas pelo genoma nuclear. É estimado que os cloroplastos importam mais de 95% de suas proteínas do citoplasma. Consequentemente, a expressão espacial e temporal dos genes codificados pelo núcleo e organelas deve ser precisamente coordenada.

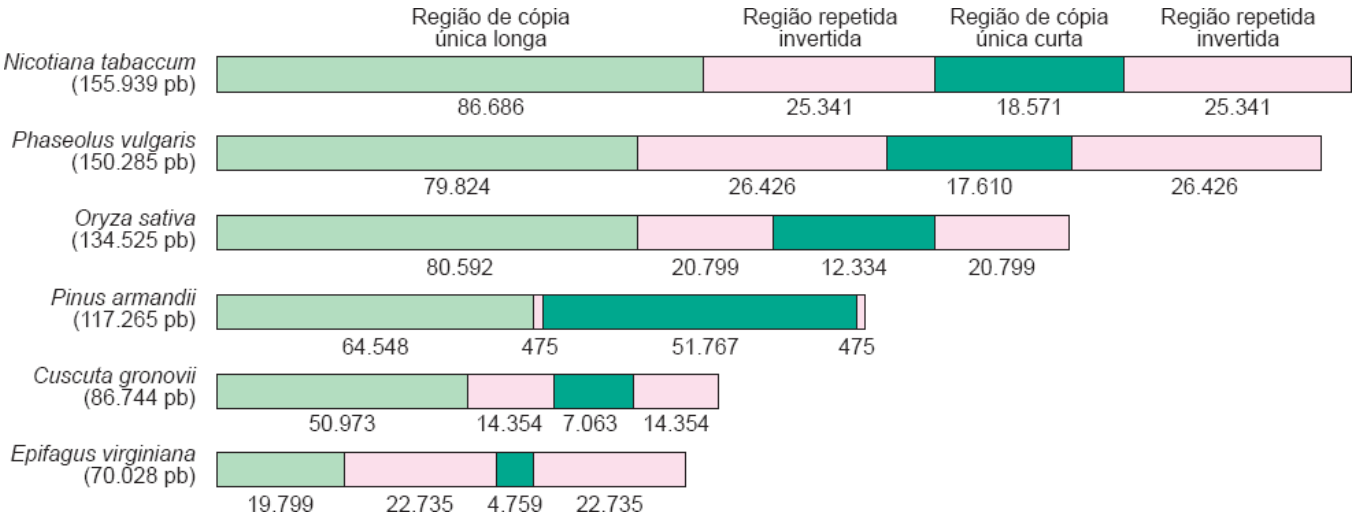


Figura 15.8 Representação esquemática de genomas plastidiais, mostrando a característica conservada da estrutura tetra dividida com duas regiões de cópias simples e uma região repetida invertida (RI). O comprimento das regiões repetidas (amarelo) pode ser muito variável entre os genomas de diferentes espécies, o que influencia no tamanho dos genomas plastidiais. A região de cópia única (azul) de genomas plastidiais de plantas parasitas não fotossintetizantes (*Epifagus* e *Cuscuta*) é menor devido à ausência de genes que codificam proteínas envolvidas no processo fotossintetizante. Plantas coníferas, representadas pelo *Pinus*, têm pequena região de cópia repetida invertida pela ausência de genes rRNA e tRNA.

Os genes codificados pelo genoma plastidial são classificados em três grupos principais: os relacionados ao sistema genético (aproximadamente 60 genes), que codificam RNA e componentes proteicos do mecanismo de expressão gênica; os relacionados à fotossíntese (aproximadamente 50 genes); e outros genes, que compreendem as orfs conservadas e as de funções desconhecidas.

Genes fotossintéticos

Cloroplasto é a organela na qual ocorre a fotossíntese, processo de conversão da energia solar em energia química. A fotossíntese consiste em duas etapas: a fase clara, ou fotoquímica, representada pelas reações de luz, quando ocorrem a síntese de ATP, $NADPH_2$ e a liberação de O_2, e a fase denominada escura, quando ocorrem as reações bioquímicas de fixação e redução do CO_2 em carboidratos (Figura 15.9). Ambas as etapas envolvem maquinarias moleculares complexas. Um número substancial de genes codificados pelo plastoma (47 genes em angiospermas) é dedicado ao aparato da fotosssíntese, sendo eles: 15 genes para as subunidades do fotossistema II, o complexo de membranas catalisando as reações de fotólise da água, outros sete genes necessários para o funcionamento do fotossistema I (FSI) e um complexo de proteínas de membranas dirigidas pela luz que catalisam a transferência de elétrons transmembrana da plastocianina para o complexo da ferredoxina-NADP. Além dos cinco genes para subunidades do complexo do FSI, também estão as orfs *ycf*3 e *ycf*4 e dois genes, cujas proteínas estão envolvidas na montagem do FSI. Outros seis genes plastidiais codificam subunidades do complexo citocromo b6 f, o complexo de proteínas que interconecta os dois fotossistemas, e outros seis genes codificam subunidades da ATP sintase cloroplastidial, a enzima que catalisa a conversão do fosfato e adenosina difosfato em adenosina trifosfato com a utilização de um gradiente de prótons atravessando as membranas do tilacoide como fonte de energia. Onze genes codificam subunidades de uma desidrogenase NAD(P)H cloroplastidial em um complexo tilacoidal, o que sugere que ela está envolvida em clororrespiração e fluxo cíclico de elétrons no FSI. Entretanto, esse complexo demonstra não ser essencial para a fotossíntese, uma vez que todos os genes para essa subunidade estão ausentes nos genomas plastidiais da gimnosperma *Pinus thunbergii* e da alga verde *Chlamydomonas reinhardtii*. Além disso, o gene *rbc*L,

que codifica a subunidade maior da ribulose-1,5-bifosfato carboxilase/oxigenase (Rubisco), está diretamente envolvido nas reações de escuro da fotossíntese; e a *ycf*10, que codifica uma proteína de membrana interna do cloroplasto, possivelmente envolvida na absorção de carbono inorgânico.

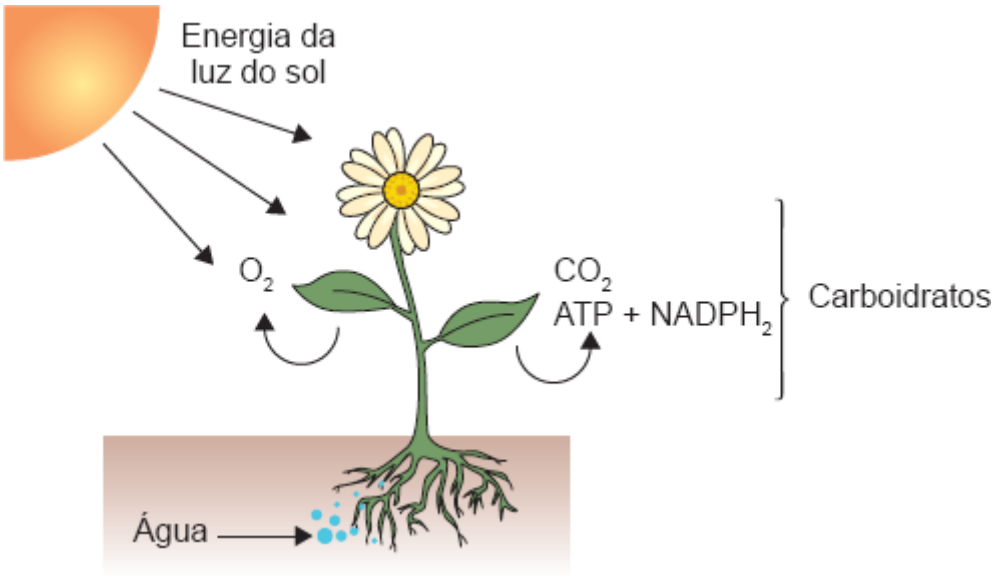


Figura 15.9 Fotossíntese em plantas. A fase fotoquímica da fotossíntese compreende a transformação da energia luminosa em energia química (ATP e $NADPH_2$) pela fotólise da água e transporte de elétrons pelas proteínas e citocromos nas membranas dos tilacoides. A fase bioquímica ocorre no estroma do cloroplasto, pela redução do CO_2, com a formação de carboidratos utilizados pelas células de diferentes organismos como fonte de carbono e energia.

Nenhum dos complexos proteicos envolvidos em fotossíntese é composto inteiramente das subunidades codificadas no plastoma. Todas elas requerem proteínas codificadas no genoma nuclear originadas de cianobactérias, mas que, ao longo da evolução, foram transferidas do genoma plastidial para o genoma nuclear. As duas subunidades da enzima Rubisco são um exemplo clássico dessa cooperação plastídeo-núcleo. Em todas as plantas, a subunidade maior da Rubisco é codificada no plastoma, enquanto a subunidade menor é codificada pelo genoma nuclear, tipicamente por uma pequena família de genes.

Genes do sistema genético

Os genes do sistema genético compreendem o maior grupo de genes localizado no plastoma de plantas vasculares (62 genes). A esse grupo pertencem todos os genes cujos produtos estão envolvidos na expressão de genes plastidiais (p. ex., transcrição, processamento do RNA, tradução, degradação de proteínas). São eles: 30 genes tRNA; quatro genes rRNA; 21 genes para proteínas ribossômicas (nove proteínas da subunidade maior e 12 proteínas da subunidade menor do ribossomo 70S); quatro genes que codificam subunidades da RNA polimerase plastidial (PEP); *mat*K, que codifica uma RNA maturase (um fator de *splicing* envolvido na remoção de um subfragmento do intron de grupo II de cloroplastos); *clp*P, que codifica uma das subunidades de uma protease de cloroplasto; e *inf*A, que codifica um fator de iniciação da tradução IF-1.

O conjunto completo de tRNA para decodificar todos os códons de genes codificadores de proteínas é composto por 32 genes de tRNA, dos quais somente 30 são codificados pelo plastoma e considerados suficientes para ler todos os códons. Enquanto os componentes do RNA na maquinaria de expressão gênica (rRNA e tRNA) são exclusivamente codificados pelo genoma plastidial, muitos dos componentes da enzima são codificados pelos genes nucleares. Por exemplo, somente um terço das proteínas ribossômicas plastidiais é codificado pelo genoma plastidial em plantas; os outros dois terços são codificados pelo genoma nuclear, traduzido no citosol e importado para o plastídeo. As quatro subunidades da RNA polimerase plastidial são como na *E.coli* (PEP- RNA polimerase codificada pelo plastídeo) e codificadas no plastoma, mas os fatores sigma, necessários para o reconhecimento do promotor, são codificados pelo genoma nuclear. Também deve ser considerada a presença de uma segunda RNA polimerase com atividade no plastídeo, a NEP (RNA polimerase codificada pelo núcleo), que é codificada por genes nucleares.

Em angiospermas, um grande grupo de genes codificados no plastoma é conservado entre espécies; entretanto, um pequeno número de genes do sistema genético, incluindo *rpl*23 e *inf*A, são notáveis exceções, pois foram transferidos para o núcleo ou substituídos por genes nucleares de origem não plastidial em algumas linhagens evolutivas. A presença, no plastoma, de remanescentes pseudogênicos dos genes transferidos para o núcleo sugere que a ocorrência desses eventos é relativamente recente na evolução. O gene *inf*A, que codifica o fator 1 de iniciação traducional plastidial, é um exemplo a ser destacado. Esse gene é funcional no plastoma de *Marchantia polymorpha* e no arroz, contudo, existe somente como pseudogene no plastoma do tabaco e outras espécies. A análise filogenética sistemática da estrutura do *inf*A em plastomas de angiospermas revelou que o gene se tornou não funcional repetidamente em aproximadamente 24 linhagens separadas na evolução de angiospermas. Cópias do gene *inf*A estão presentes no genoma nuclear, e seus produtos são direcionados para os plastídeos.

Outros genes e orfs conservadas

Um pequeno número de genes do plastoma não são diretamente envolvidos em fotossíntese ou expressão de genes. Entre eles está o gene *acc*D, que codifica a subunidade da acetil-CoA carboxilase, a enzima-chave na biossíntese de ácidos graxos. Um segundo exemplo é o gene *ccs*A (*ycf*5), cujo produto é necessário para a ligação do grupo heme do citocromo tipo-C de cloroplastos.

O plastoma de plantas também apresenta algumas orfs de funções desconhecidas. Uma orf conservada interespecífica é, geralmente, uma boa indicação que se constitui de um gene. Em plastídeos, estas orfs conservadas são referenciadas como *ycf* (orf hipotética de cloroplasto). Embora tenham sido determinadas as funções de muitas ycfs por genética reversa em *C. reinhardtii* e tabaco, ainda existem algumas sem elucidação, entre elas a *ycf*1 e *ycf*2, que são as maiores orfs dos plastomas, com produto proteico de 1901 e 2280 aminoácidos, respectivamente. Essas duas *ycf*s codificam proteínas essenciais em alguma via metabólica ou na expressão gênica, contudo não têm função estabelecida até o momento. Todos os genomas plastidiais têm um número de orfs não conservadas, a maioria delas com menos de 150 nucleotídeos. A falta de conservação entre espécies próximas é interpretada como evidência de que essas orfs não têm significância funcional.

Sistema transcricional no genoma cloroplastidial

Devido à origem endossimbiótica das organelas, genes cloroplastidiais e mitocondriais mantiveram várias características dos organismos procariotos. Entretanto, organelas em plantas adquiriram novas e específicas características durante a evolução. O controle da expressão dos genes em plantas ocorre nos níveis transcricional e pós-transcricional.

Muitos genes do genoma plastidial estão organizados em operons, que são unidades transcricionais policistrônicas com agrupamentos de dois ou mais genes transcritos pela RNA polimerase a partir de um único promotor semelhante ao operon, correspondente, no ancestral, à cianobactéria. Esses transcritos policistrônicos são processados em RNA maduros, indicando limitada unidade transcricional no plastoma. Na Tabela 15.1 são apresentadas algumas dessas principais unidades transcricionais. Observa-se que os genes estão organizados em um único operon e que nem sempre apresentam a mesma funcionalidade. Por exemplo, os genes rRNA plastidiais, na maioria das plantas, estão organizados em um único operon, que consiste dos genes 16S, 23S, 4.5S e 5S, com a presença dos *trn*L e *trn*A na unidade transcricional, que será descrita mais adiante.

Genoma plastidial de algas

Enquanto a estrutura e a capacidade codificadora do genoma plastidial se apresentam altamente conservadas entre as plantas, o plastoma de algas tem apresentado muitas mudanças arquitetônicas durante a evolução. As algas verdes (clorófitas e caróficas) são consideradas antecessores filogenéticos das plantas terrestres; assim, apresentam algumas características em comum, como a arquitetura tetrapartida e a forma circular do plastoma. Entretanto, a variação da estrutura do ptDNA entre os diversos grupos de algas é mais pronunciada do que em plantas, apresentando perdas ou rearranjos das regiões repetidas invertidas, expansões intergênicas, deleção ou fragmentação de genes e multiplicação de sequências intrônicas. As algas *Floydiella terrestris* e *Volvox carteri*, com genomas plastidiais de mais de 521 e 420 Kpb, respectivamente, são consideradas os maiores plastomas já sequenciados. Em outro extremo estão os ptDNA de algas que perderam genes associados a funções fotossintéticas e apresentam genomas reduzidos,

como em *Helicosporidium* sp., com 37,5 Kpb. Em geral, os plastomas da maioria das algas verdes variam entre 71,6 Kpb (*Ostreococcus tauri*) e 223,9 Kpb (*Stigeoclonium helveticum*) e codificam entre 88 e 138 genes.

Tabela 15.1 Unidades transcricionais em cloroplastos de plantas.

Unidades transcricionais	Funcionalidade
*atp*B-*apt*E	Os genes codificam subunidades proteicas da ATP sintase
*rps*12-*rps*7	Tradução. Codifica subunidade menor do ribossomo
psbe-psbf-psbl-psbj	Subunidades envolvidas na montagem e na estabilização do FSII
*psb*D-*psb*C (em dicotiledôneas)	Codificam proteínas do sistema antena, essenciais no FSII
psbk-psbl-psbd-psbc-orf62-trng (em monocotiledôneas)	São subunidades pequenas envolvidas na montagem e na estabilidade do FSII, *psbD* e *psbC*. Participam do centro de reação e são essenciais ao FSII. A orf62 não tem função conhecida e a *trnG* participa da tradução por ser um tRNA-glicina
*rpo*B-*rpo*C1-*rpo*C2	Transcrição. Participa da estrutura da RNA polimerase
*rpl*23-*rpl*12-*rpsl*9-*rpl*22-*rps*3-*rpl*16-*rpl*14-*rps*8-*inf*A-*rpl*36-*rps*11-*rpo*A	Tradução. Os genes *rpl* codificam a subunidade maior dos ribossomos, e os *rps* as subunidades menores, *infA* codifica o fator 1 de iniciação da tradução e aparece como um pseudogene em várias espécies. O gene *rpoA* codifica a subunidade a da RNA polimerase plastidial (PEP)
trne-trny-trnd	Tradução. Codificam as tRNA para glutamato, tirosina e aspartato, respectivamente
*clpp-rps*12-*rpl*20	O gene *clpP* é essencial para a célula. Codifica uma subunidade catalítica da proteína clp. Os genes *rps*12 e *rpl*20 codificam as subunidades menor e maior dos ribossomos, respectivamente
*orf*31-*pet*G-*psa*J-*rpl*33-*rps*18	A orf 31 não tem função conhecida, mas participa do operon com genes *pet*G, que codificam uma subunidade essencial para montagem e estabilidade do citocromo b_6 f (citb_6 f). O gene *psa*J codifica uma subunidade do FSI. Os genes *rpl*33 e *rps*18 participam das subunidades maior e menor dos ribossomos, respectivamente
*psaa-psab-rps*14	Os genes psaA e psaB codificam subunidades de proteínas essenciais do centro de reação do FSI. O gene *rps*14 codifica uma subunidade menor do ribossomo envolvido na tradução
*psa*C-*ndh*D	O gene *psa*C codifica uma subunidade do FSI e é essencial para a ligação de um cofator. O gene *ndh*D participa do processo de respiração cloroplastidial, na transferência cíclica de elétrons, e está ausente em gimnospermas
*psb*B-*psb*H-*pet*B-*pet*D	Os genes *psb*B e *psb*H codificam proteínas que participam da montagem, estabilidade e proteção do FSII. Os genes *petB* e *pet*D codificam subunidades essenciais para o citb_6 f
*psb*K-*psb*L-*trn*G	Os genes *psb*K e *psb*L participam de subunidades menores associadas ao FSII. O gene *trn*G codifica o tRNA glicina envolvido na tradução
*ndh*A-*ndhl*-*ndh*G-*ndh*E-*psa*C	Os genes *ndh* participam do processo de respiração cloroplastidial e da transferência cíclica de elétrons. Todos os genes *ndh* estão ausentes em genomas de gimnospermas. O gene *psa*C codifica uma subunidade de FSI e é essencial para a ligação de cofator
*rpl*32-*spr*A	O gene *rpl*32 codifica uma subunidade maior (L32) do ribossomo envolvido na tradução. O gene *sprA* é um pequeno RNA com função desconhecida

*rps*2-*atp*I-*atp*H-*atp*F-*atp*A	O gene *rps*2 codifica uma subunidade menor (S2) do ribossomo envolvido na tradução, enquanto os genes atp*I*, *atp*H, *atp*F e *atp*A codificam as subunidades que compõem a ATP sintase

Adaptada de Shi *et al*., 2016.

A região repetida invertida está presente na maioria dos plastomas, sendo a que mais apresenta mudanças estruturais. Enquanto muitas algas verdes e vermelhas mantêm o operon rRNA na região repetida, algumas perderam regiões repetidas e têm somente uma cópia única do operon rRNA. É interessante observar que algumas algas verdes e vermelhas apresentam regiões repetidas diretas, em vez de invertidas, como nos plastomas das plantas superiores. Um terceiro tipo de organização de genomas plastidiais é representado pelo genoma de *Euglena gracilis*, um protista fotossintetizante que tem de 3 a 5 repetições em sequência de genes RNA ribossômicos.

O plastoma de algas também varia muito quanto a densidade de genes e ao conteúdo de informação. Os genomas plastidiais são extremamente compactos e densos em genes no plastoma da alga Cryptofita *Guillardia theta*, o qual contém 180 genes em um genoma de apenas 122 Kpb. Entretanto, a expansão pelo acúmulo de DNA não codificadores ocorreu em outras algas, como a unicelular *Chlamydomonas reinhardtii*, que tem genoma com 203 Kpb, mas contém somente 99 genes. A presença de DNA repetitivo (como pequenas repetições dispersas) em regiões intergênicas consiste em mais de 20% do tamanho do genoma.

Algas verdes dividem o mesmo ancestral comum com as plantas e talvez não seja surpresa o fato de que o conteúdo de genes de seus plastomas seja semelhante. Exceções incluem, por exemplo, o gene *pet*N (envolvido na fotossíntese) e dois genes homólogos de *E. coli* (*min*D e *min*E) envolvidos na divisão plastidial que estão presentes no plastoma das algas verdes, mas ausentes em plantas superiores e na alga *C. reinhardtii*. Diferentemente das algas verdes, o plastoma de algas não verdes tem mantido muitos genes que estão ausentes no plastoma de algas verdes e plantas. Em geral, considera-se que esses genes adicionais tenham sido transferidos para o núcleo no ancestral comum de plantas e algas verdes, que representam a estrutura mais recente da evolução. De todos os organismos que contêm plastídeos, as algas vermelhas (Rhodophyta) têm plastomas com o maior número de genes, assim como a alga vermelha *Porphyridium purpureum*: 260 em um genoma de 217 kb. A expansão de genes fotossintéticos e genes do sistema genético, como a presença de genes adicionais relacionados à biossíntese de aminoácidos, ácidos graxos, pigmentos e vitaminas, contribui para aumentar o conteúdo de genes.

Em um evento de endossimbiose secundário (Figura 15.10 A), uma alga vermelha foi engolfada e mantida no meio intracelular por um organismo unicelular eucarionte, dando origem a um novo grupo taxonômico (Cromalveolados), do qual faz parte o filo Crytophyta, um pequeno grupo de algas unicelulares com flagelos. Esse grupo de algas apresenta a característica marcante da presença de plastídeos com quatro membranas, sendo que as duas mais internas mantêm as originais do plastídeo primário da alga rodofícea, enquanto a terceira corresponde à membrana plasmática da alga capturada. O invólucro mais externo dos plastídeos secundários é contínuo à membrana do retículo endoplasmático, o que sugere que a alga rodofícea endossimbionte foi capturada por fagocitose e assimilada ao sistema endomembrana do hospedeiro. Nesse grupo de algas, o núcleo foi mantido como uma organela vestigial, o nucleomorfo, uma parte reduzida do genoma nuclear original do endossimbionte, a qual, em *Guillardia theta*, codifica 487 genes ainda necessários para manutenção do compartimento endossimbionte (Figura 15.10 B). O genoma plastidial mantém características dos plastomas de algas vermelhas. Embora a contagem gênica seja maior do que em algas verdes, é menor que a encontrada no plastoma da alga vermelha progenitora, possivelmente porque muitos genes foram transferidos para o núcleo do novo hospedeiro.

Genoma de apicoplastos

Os apicoplastos são organelas com vestígios de plastídeos presentes na maioria dos parasitas do filo Apicomplexa, que são responsáveis por importantes doenças, como malária, toxoplasmose, babesiose e coccidiose; esta última ocorrendo principalmente em aves. A malária é uma das principais doenças das regiões tropicais, pela alta temperatura e umidade, e é transmitida por um inseto vetor do gênero *Anopheles*, que transmite o parasita *Plasmodium falciparum*. A maioria dos casos ocorre na África, onde o acesso a tratamento é precário. A estimativa é de mais de 250 milhões de novos casos por ano no mundo (*World Malaria Report*, realizado pela Organização Mundial da Saúde). A essencialidade dessa organela nos parasitas não é completamente compreendida, mas torna-se um alvo importante para o desenvolvimento de medicamentos contra o *Plasmodium*, responsável pela malária, e o

Toxoplasma gondii, responsável pela toxoplasmose, causadora de infecção no cérebro e nos músculos (Van Dooren e Striepen, 2013).

Os apicoplastos tiveram origem por endossimbiose secundária, caracterizada pela presença de três ou quatro membranas (Figura 15.11). Apesar de haver dúvida se a endossimbiose secundária se originou de uma alga verde ou vermelha, evidências de análises filogenéticas dos genes da GAPDH (gliceraldeído fosfato desidrogenase) e coxII codificadas pelo genoma nuclear e a similaridade de arquitetura do genoma reforçam que os apicoplastos se originaram das algas vermelhas. Não existem fósseis de Apicomplexa, mas dados moleculares demonstram que os protistas desse filo apareceram entre 600 e 800 milhões de anos, possivelmente em uma interação parasita de Apicomplexa em invertebrado marinho antes de se estabelecer em vertebrados terrestres.

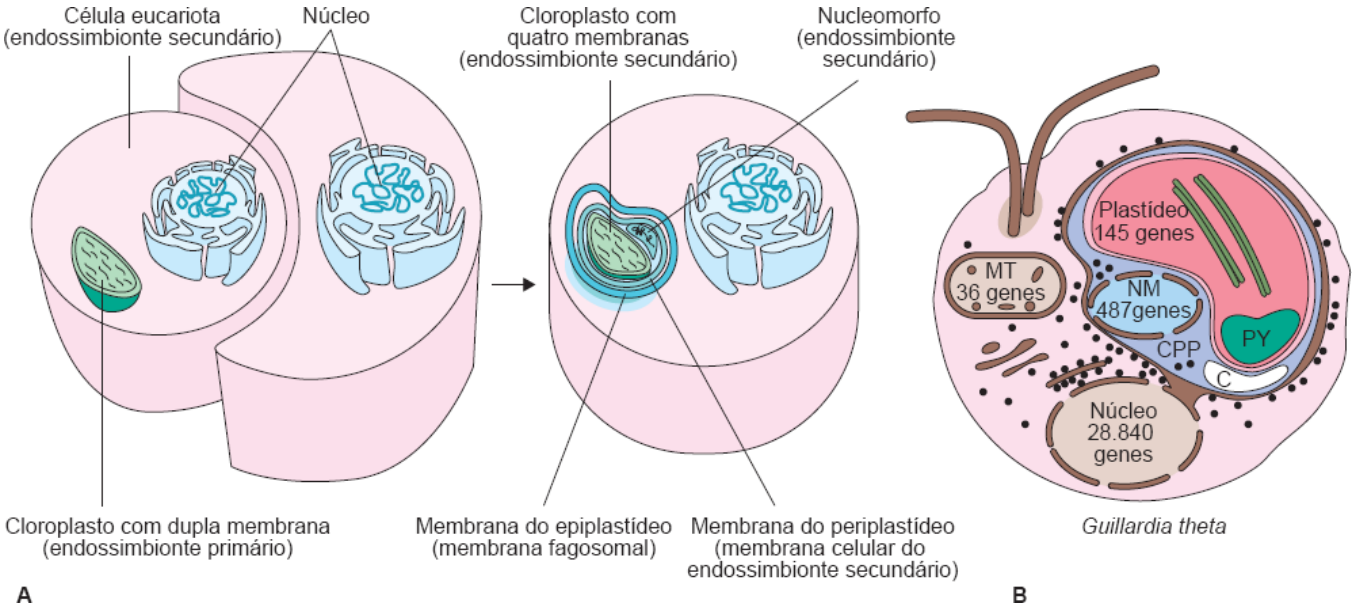


Figura 15.10 Endossimbiose secundária e estrutura de uma alga criptófita. **A.** Processo de endossimbiose secundária em que uma célula que contém um plastídeo é engolfada por uma célula eucariota; ocorre a perda de um dos núcleos, e os plastídeos adquirem mais uma ou duas membranas. **B.** O plastídeo da alga *Guillardia theta* tem quatro membranas. A membrana externa é contínua com o envelope nuclear e apresenta muitos ribossomos na sua superfície externa. Entre a membrana interna e a membrana externa está o compartimento periplastidial (CPP), onde se localiza o nucleomorfo (NM), que seria o núcleo resultante do endossimbionte eucarioto. Em cada compartimento está apresentado o número de genes codificadores de proteína. Mitocôndrias (M); Plastídeos (P); Núcleo (N) e Nucleomorfo (NM).

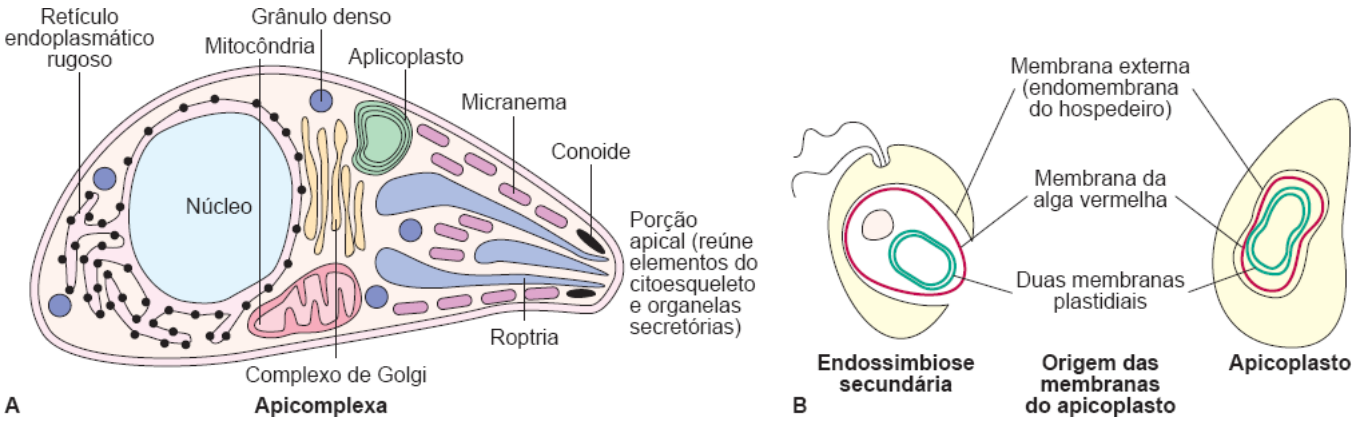


Figura 15.11 Estrutura do *Toxoplasma gondii*, protozoário agente causador da toxoplasmose. **A.** Existência do apicoplasto com quatro membranas. **B.** Endossimbiose secundária com a presença de quatro membranas, sem nucleomorfo. A membrana externa é a parte da membrana interna do hospedeiro. A segunda membrana externa (vermelho) é chamada membrana periplastial, é derivada da alga vermelha e se tornou o apicoplasto; e as duas membranas internas (verde) são homólogas às membranas internas e externas do plastídeo primário da alga vermelha.

O genoma do apicoplasto presente no parasita *Plasmodium falciparum* tem 86,9% de conteúdo de A+T, é circular e tem 35 kb. Os genes codificados por esse genoma não são fotossintéticos e incluem genes duplicados para

a subunidade maior e menor dos rRNA, 25 tRNA com nove deles duplicados, três subunidades de uma RNA polimerase, 17 proteínas ribossômicas, um fator de elongação traducional e um membro da família de chaperona *Clp*, além de uma *orf* de função desconhecida encontrada em plastídeo de alga vermelha. O genoma apresenta uma região invertida de aproximadamente um terço, onde estão os genes duplicados das subunidades maior e menor dos rRNA e os nove tRNA duplicados. Essas características altamente conservadas nos *apicoplexans* sugerem que tenham uma origem comum.

Genomas plastidiais de plantas parasitas

Enquanto a grande maioria das angiospermas tem o genoma cloroplastidial relativamente conservado entre as espécies, plantas parasitas têm o plastoma altamente diferenciado, principalmente quanto ao conteúdo gênico. Somente um pequeno número de plantas são heterotróficas, e elas utilizam fonte de carbono de outras plantas. Muitas dessas holoparasitas perderam a capacidade de conduzir a fotossíntese, perdendo também os pigmentos fotossintéticos. A habilidade de obter açúcar a partir de plantas hospedeiras faz com que essas plantas não tenham que manter os genes fotossintéticos. Consequentemente, os plastomas de plantas parasitas sofreram redução significativa de tamanho, causada pela perda dos genes fotossintéticos ou sua degradação em pseudogenes. Por exemplo, o plastoma de holoparasita de raiz *Epifagus virginiana* tem menos da metade do tamanho do plastoma das angiospermas fotoautotróficas (70.028 pb); ele contém somente 21 genes codificadores de proteínas intactos, 18 dos quais pertencem aos genes do sistema genético, e, dos restantes, três pertencem à categoria de outros genes (*acc*D, *ycf*1 e *ycf*2). Interessantemente, também alguns genes do sistema genético foram perdidos ou degradados a pseudogenes. Não está claro se esses genes faltantes foram substituídos por cópias de genes funcionais do núcleo cujas proteínas são importadas para os cloroplastos. Genes plastidiais de *Epifagus* são ativamente transcritos e seus mRNA são processados por *splicing* de introns e editamento de RNA (descritos a seguir), o que sugere que os genomas plastidiais remanescentes são expressos.

O parasitismo em plantas tem evoluído diversas vezes independentemente. É interessante que nem todas as plantas parasitas crescem exclusivamente heterotroficamente. Um número de espécies parasitas tem mantido alguma atividade fotossintética e, assim, cresce mixotroficamente. Elas fixam uma limitada quantidade de carbono por elas mesmas pela fotossíntese, enquanto também obtêm carbono, na forma de açúcar, da planta hospedeira. Essas espécies encontram-se em um estágio intermediário na evolução, em processo de perda dos genes fotossintéticos. O gênero *Cuscuta* é um exemplo de espécie nessa transição evolucionária. O cromossomo plastidial de *C. Reflexa* tem 121.521 pb em contraste, e o plastoma de *C. Gronovii* tem somente 86.744 pb, mas mantém os genes fotossintetizantes, os quais estão ausentes em *Epifagus*. Em *Cuscuta* os genes ausentes, em comparação com genomas de angiospermas, como o tabaco, estão relacionados ao aparato de expressão gênica, como os genes ribossômicos e tRNA e alguns genes fotossintetizantes *ndh*, *psa*I em *C. Gronovii* (Figura 15.12). Há mais de 150 espécies que variam bastante a atividade fotossintética e a extensão da degeneração pela deleção de genes ou decaimento a pseudogene. Essas espécies não apresentam raízes e folhas e obtêm os nutrientes orgânicos e inorgânicos e a água dos seus hospedeiros por estruturas dos haustórios; também apresentam variações quanto à estrutura e ao funcionamento dos plastídeos. Existe uma simplificação na maquinaria de expressão gênica plastidial na planta parasita em consequência do aumento de sua dependência da planta hospedeira e gradual adaptação no grau de parasitismo.

Expressão gênica

Em células eucariotas, a presença de cloroplastos e mitocôndrias com características procarióticas trazem várias implicações quanto à expressão dos genes. As organelas apresentam várias características dos organismos procariotos na estrutura dos genomas e na expressão gênica. Entre elas, é comum, mas não exclusiva, a organização dos genes em operons, que são unidades transcricionais policistrônicas (apresentam um conjunto de dois ou mais genes transcritos por uma RNA polimerase a partir de um único promotor), como apresentado na Figura 15.13, em que a fita no sentido horário tem transcrição policistrônica e a fita no sentido anti-horário tem a transcrição de um tRNA. Há semelhança das RNA polimerases (PEP) e promotores (em plastídeos) e similaridades quanto à estrutura de mRNA maduros, a presença de ribossomos 70S, entre outros. Entretanto, também há diferenças, uma vez que, nas organelas de plantas, a transcrição e a tradução não são acopladas e apresentam RNA polimerase codificadas pelo núcleo (NEP), edição do mRNA e *splicing* de transcritos primários.

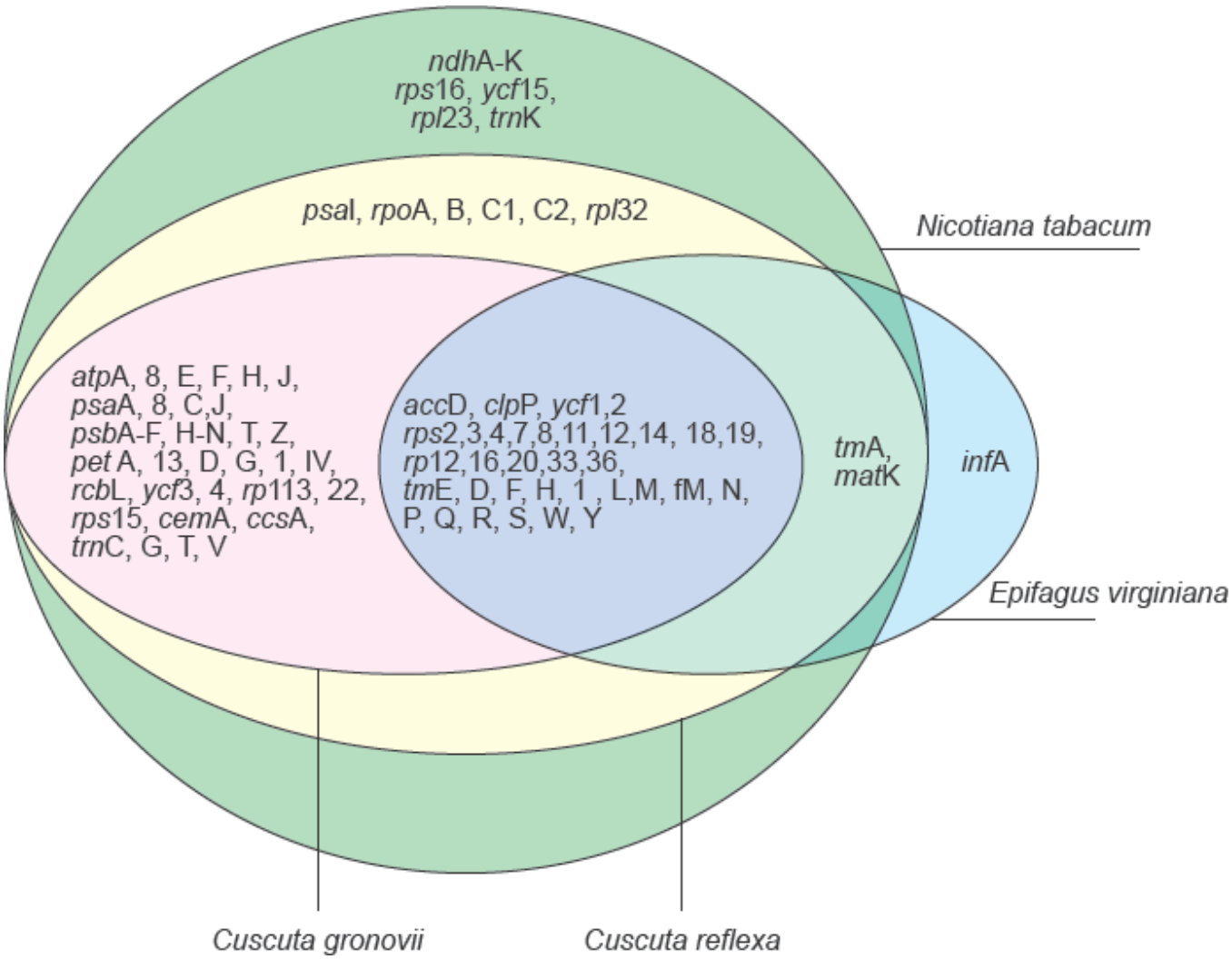


Figura 15.12 Diagrama de Venn comparando genes codificadores de proteínas em *Nicotiana tabacum* e as parasitas *Cuscuta reflexa*, *Cuscuta gronovii e Epifaus virginiana*. Todos os genes que aparecem dentro de um círculo são codificados pelo plastoma das respectivas espécies indicadas.

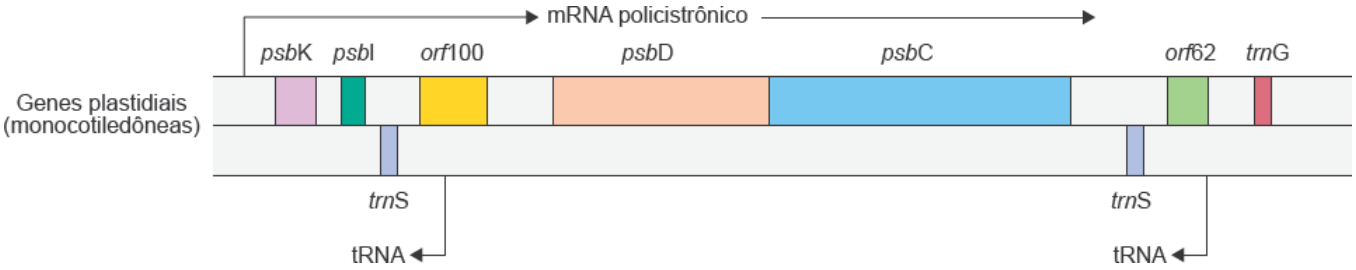


Figura 15.13 Operon *psb*K-*psb*I-*psb*D-*psb*C em monocotiledôneas (arroz). Os genes estão localizados em um segmento de 4,3 kb em uma região de cópia única longa. O gene *psb*K codifica uma proteína K pequena envolvida na estrutura do fotossistema II; o gene *psb*I codifica uma proteína no centro de reação do fotossistema II; e o gene *psb*D codifica a proteína D2 no centro de reação, a qual forma um heterodímero com a proteína D1 para ligar clorofilas, feofitina e quinona. O gene *psb*C codifica uma proteína de ligação da clorofila (CP43), que serve como uma antena de recebimento de luz. Entre os genes *psb*I e *psb*D, existe a orf100, sem função conhecida. O gene *trn*S é codificado na fita oposta como monocistrônico. Em dicotiledôneas, os genes *psb*K-*psb*I estão separados do *psb*D-*psb*C em 23 kb.

A disponibilização de dados de sequenciamento de genomas possibilitou confirmar que, ao longo da evolução, muitos genes dos organismos endossimbiontes migraram para o genoma nuclear da célula hospedeira, o que foi evidenciado pelo fato de que muitos complexos proteicos são formados por proteínas codificadas por genes nucleares e de organelas. Assim, a expressão dos genes nucleares e organelares necessariamente é corregulada pelas células para coordenar suas atividades metabólicas dentro dos tecidos nos diferentes organismos durante o desenvolvimento e em resposta às mudanças ambientais.

O mecanismo de replicação do DNA plastidial não é bem definido, pois o genoma está presente em muitas cópias nos cloroplastos, podendo chegar a até 150 cópias – mas esse número pode variar conforme o estágio de

desenvolvimento e o tipo de plastídeo. O grande número de moléculas de DNA presente em cloroplastos é justificado provavelmente pela alta demanda de proteínas envolvidas na fotossíntese, enquanto poucas cópias de DNA são encontradas em amiloplastos de raízes e outros tecidos não fotossintetizantes. É estabelecido que a replicação nas organelas é independente da replicação do DNA no núcleo e, nos plastídeos, é regulada pela frequência de iniciação da replicação em sincronismo com o ciclo celular e/ou pela diferenciação celular. Na maioria das espécies de plantas, a replicação se inicia na origem de replicação próxima ao gene 16S rRNA na região RI (repetida invertida); contudo, em espécies sem a RI, a origem de replicação se localiza próximo dos genes rRNA. Todas as enzimas e proteínas regulatórias essenciais para a replicação do DNA (topoisomerases, DNA polimerases, helicases e primase) são codificadas pelo genoma nuclear. A presença de sequências de DNA com similaridade com os genes *dna*A e *dna*B de *E.coli* no genoma de *Chlamydomonas* e *Porphyra* sugere que o mecanismo de replicação do DNA em plastídeos é semelhante ao de *E. coli* (Buchanan, 2015).

Em plastídeos, a síntese de RNA ocorre por duas enzimas, a RNA polimerase codificada pelos genes do genoma plastidial (PEP), homóloga à RNA polimerase de *E. coli* e também encontrada em cianobactérias, e a RNA polimerase codificada pelo núcleo (NEP), com homologia com RNA polimerase de fago T3 e T7. A PEP apresenta cinco subunidades: α, α, γ, β, β' codificadas pelo genoma plastidial e uma subunidade regulatória σ. Várias subunidades σ são codificadas pelo genoma nuclear. Os genes fotossintéticos são transcritos predominantemente pelas PEP, mas muitos genes plastidiais são transcritos pela NEP, assim como muitas regiões promotoras são dirigidas por ambas as polimerases. A sequência de DNA consenso de reconhecimento pela PEP é semelhante à de *E. coli* "-10 e -35" e a NEP reconhece outras sequências. Em contraste com os plastídeos, mitocôndrias de plantas têm somente a RNA polimerase codificada pelo genoma nuclear e a transcrição se inicia em diferentes locais do genoma; assim, um único gene pode ter diferentes sequências promotoras. O genoma mitocondrial de Arabidopsis apresenta a sequência de promotor YRTA como consenso, embora a transcrição possa se iniciar sem consenso reconhecido. Diferentemente, em mitocôndrias de animais, a transcrição inicia-se em um único local em cada fita do DNA, enquanto, para a terminação da transcrição, geralmente existe uma sequência repetida invertida que forma a estrutura de grampo na região 3'-UTR dos transcritos em plastídeos e mitocôndrias envolvida na terminação da transcrição ou maturação do RNA.

O controle da expressão gênica em organelas de plantas ocorre ao nível transcricional e pós-transcricional, incluindo, neste último, a regulação da maturação da transcrição, estabilidade e tradução. Antes da síntese de proteínas, o transcrito é clivado intercistronicamente e as terminações 5' e 3' passam por maturação. A posição 5' de cloroplastos não é processada e tem dois ou três fosfatos ou, quando é processada, tem uma 5'hidroxila. A terminação da transcrição 3' é ineficiente, sendo frequentes o ligamento de um complexo de alto peso molecular a jusante da estrutura do grampo ou a ligação de proteínas PPR (pentatricopeptídeo repetida) para estabilização do transcrito.

A tradução está relacionada à transcrição em eubactérias com a participação de fatores de iniciação, rRNA, tRNA e ribossomos tipo 70S. Os rRNA e tRNA e algumas proteínas ribossômicas são codificados pelo genoma plastidial, enquanto outros componentes são codificados pelo genoma nuclear e importados para os cloroplastos. Da mesma maneira, o genoma mitocondrial de plantas precisa importar vários componentes da maquinaria traducional, como aminoacil-tRNA sintetases e tRNA para expressar em torno de 30 mRNA, uma vez que somente poucas proteínas ribossômicas e alguns rRNA e tRNA são codificados pelo genoma da organela.

A sequência Shine-Dalgarno para o reconhecimento do início da síntese de proteína presente nas bactérias também se encontra na grande maioria dos genes plastidiais, além de fatores regulatórios em cis e trans responsáveis pelo correto início da tradução nessas organelas. Esses processos não foram determinados em mitocôndrias, o que sugere a função das proteínas PPR na tradução em mitocôndrias de plantas. O códon UAG nem sempre é o códon de iniciação em organelas de plantas, apesar de ser preferencial; alguns códons alternativos foram encontrados. Genes mitocondriais de plantas podem ser expressos a partir de mRNA sem códons de terminação e sem evidências de terem sofrido modificações pós-transcrionais por edição do mRNA ou poliadenilação. O fato de o processo de tradução não ser acoplado à transcrição possibilita diferentes níveis de regulação na organela.

Presença de introns em genes de organelas

A transcrição é um processo complexo e altamente regulado para a expressão de um gene, e representa somente uma das etapas necessárias para produzir um RNA funcional. O RNA pode ser modificado extensivamente, mas a

maioria dos transcritos primários em células eucariotas são processados para produzir o RNA maduro para dar início à síntese de proteínas. A região codificadora de muitos genes eucariotos é interrompida por sequências que não aparecem no RNA maduro. Os transcritos primários sintetizados pelas RNA polimerases incluem, além de segmentos de RNA que fazem parte do produto gênico (os exons), sequências não codificadoras que devem ser removidas (introns) para produzir a molécula do mRNA que codifica diretamente para uma proteína. O processo em que esses introns são excisados do pré-RNA e os exons religados é chamado "*splicing* do RNA", e resulta na sequência do RNA maduro que será traduzido em proteína.

Os introns estão presentes na maioria das classes de genes nos genomas do núcleo, mitocôndria e cloroplastos. A presença de introns em genes de organelas causa certa surpresa, pois introns não são comuns em genes cujos ancestrais deram origem aos cloroplastos e mitocôndrias. O número e a frequência nos genes variam entre as espécies de plantas e organelas. A *Euglena* apresenta 155 introns, que correspondem a 40% do genoma plastidial, enquanto, no genoma plastidial da alga vermelha *Porphyra*, nenhum intron foi identificado. A grande maioria dos genes codificadores de proteínas no núcleo de plantas contém um ou mais introns. Os genes ribossômicos 5.8 e 5S e snRNA também não apresentam introns.

Em cloroplastos e mitocôndrias são encontrados introns do grupo I e do grupo II, os quais são classificados de acordo com sua estrutura primária e secundária. Introns do grupo I (Figura 15.14 A) são ribozimas (moléculas de RNA com atividade catalítica) e são encontrados em genes de mitocôndrias de leveduras e outros fungos, rRNA nucleares de eucariotos unicelulares (*Tetrahymena*) e em organelas de cloroplastos. Até 1994, nenhum intron do grupo I havia sido encontrado em genes mitocondriais de angiospermas, e acreditava-se que eles poderiam estar completamente ausentes nesses genomas. Porém, o gene *cox*I de *Peperomia polybotrya* (peperomia) revelou a presença de um intron do grupo I, com similaridade com o intron do gene *cox*I de levedura, e codifica uma endonuclease específica que possibilita ao intron funcionar como uma ribozima durante sua autoexcisão. Em algumas linhagens de leveduras, um mesmo gene mitocondrial pode ter um intron em uma linhagem, mas não em outra, e eles podem até mesmo se mover para dentro ou fora do genoma como elementos de transposição; o mesmo processo é encontrado no gene 23S rRNA de *Chlamydomonas*.

Introns do grupo II são os mais frequentemente encontrados em genes plastidiais e mitocondriais de fungos e plantas. Essa classe de introns tem algumas características em comum com o grupo de intron pré-mRNA nuclear, o que indica que têm uma origem evolucionária comum e dividem o mesmo mecanismo de *splicing*, seguindo a regra GU-AG (Figura 15.14 B). Cloroplastos e mitocôndrias de plantas apresentam, além do processo de *cis-splicing* entre exons de um único transcrito, o processo de *trans-splicing*, que envolve duas ou mais moléculas de RNA. O gene rps12 de cloroplasto de tabaco é organizado em dois *loci* separados no genoma (Figura 15.15 A). Cada *locus* produz um transcrito independente: um codifica o exon 1, e o outro codifica o produto do RNA exon 2-intron-exon 3. Em seguida à transcrição, o processo de *trans-splicing* liga os exons 1, 2 e 3. A ligação dos exons 2 e 3 ocorre em *cis* após a remoção do intron. O gene *psa*A do cloroplasto de *Chlamydomonas* é codificado em três *loci* separados no genoma (Figura 15.15). Nesse caso, dois eventos trans-*splicing* originam o mRNA maduro de três transcritos diferentes. Em mitocôndrias de angiospermas, os genes *nad*1, *nad*2 e *nad*5 são descontínuos, e seus transcritos passam por trans-*splicing* para produzir mRNA funcionais. Os genes *nad*2 e *nad*5 no genoma de pteridófitas são contínuos, com os introns ligados por *cis-splicing*. Em algumas espécies, como em *Beta vulgaris* (beterraba), os introns só ocorrem em genes codantes de proteínas, sendo que os genes rRNA e tRNA não apresentam introns. Nessa espécie, o gene *rps*3 não tem introns, o que contrasta com o fato de que ele tem um único intron em outras espécies.

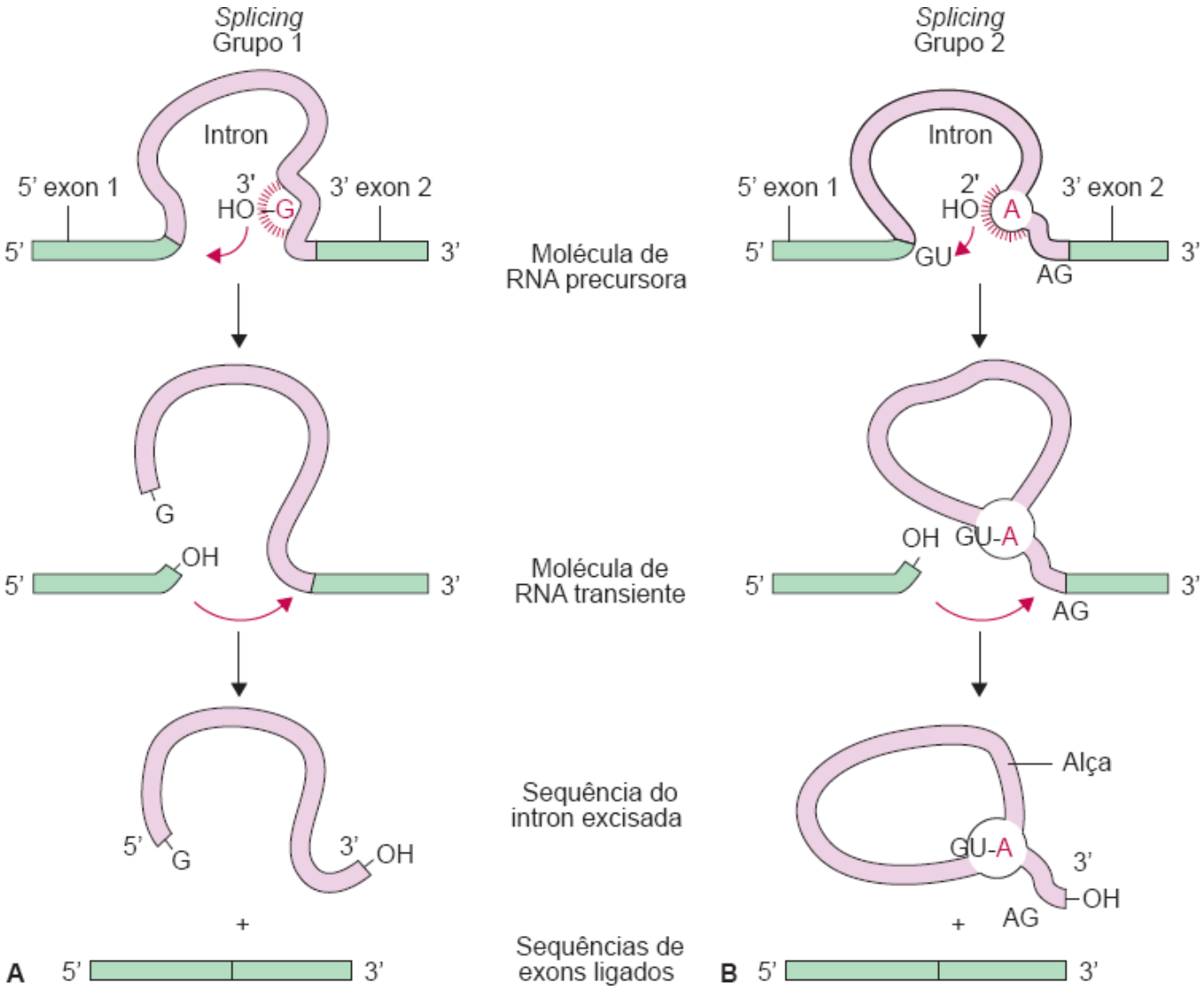


Figura 15.14 Auto-*splicing* tipo I e II. **A.** *Splicing* tipo I: requer um nucleotídeo de Guanosina (G), que se liga a uma sequência específica no intron que age na posição 5' do intron, promovendo uma clivagem. O grupo 3'OH do exon 1 reage com o fosfato do sítio de *splicing* do exon 3', ligando os dois exons, e libera o intron linear. **B.** *Splicing* tipo II: a clivagem no sítio 5' é mediada por um grupo hidroxila de um nucleotídeo adenina que é parte da sequência do intron. Após a clivagem, a terminação 5' do intron forma uma ligação com essa adenina, produzindo uma alça.

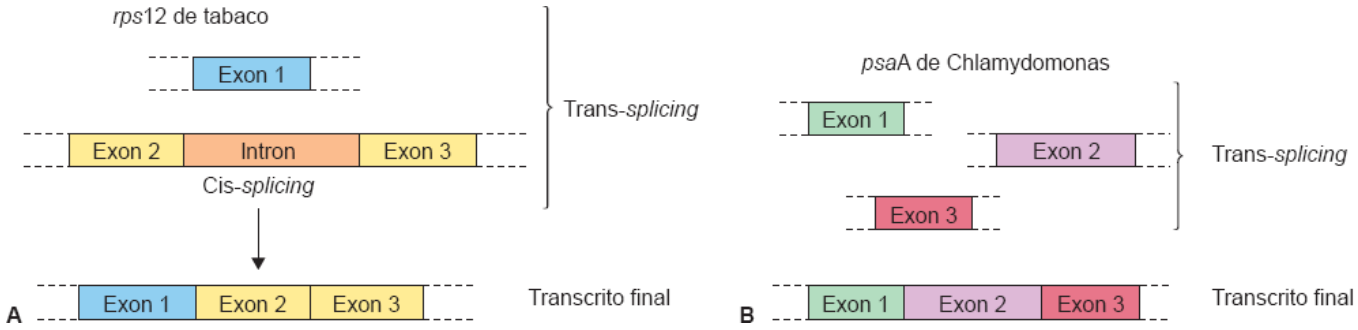


Figura 15.15 Reações de *splicing* intramolecular em cis e trans. **A.** O gene rps12 de tabaco (*Nicotiana tabacum*) de cloroplasto está organizado em dois *loci* separados no genoma; cada *locus* produz um transcrito independente: um codifica o exon 1, e o outro codifica o produto cis-*splicing* do exon 2 e exon 3, cujo produto irá se ligar por trans-*splicing* ao exon 1. O transcrito maduro tem o exon 1, exon 2 e exon 3. **B.** O gene *psb*A de *Chlamydomonas* é descontínuo, codificado por três *loci* separados no genoma. O transcrito maduro é resultado de dois eventos trans-*splicing* entre os exons.

Os genes codificados no genoma mitocondrial de Arabidopsis são interrompidos por 23 introns do grupo II com tamanho que varia de 485 a aproximadamente 4.000 nucleotídeos. Alguns genes são interrompidos por mais de um intron, como o gene *nad*7, que tem quatro introns. Todos os quatro genes ribossômicos são codificados como uma unidade transcricional policistrônica que inclui também dois genes tRNA, Ile e Ala, cada um dos quais com um

intron. Após o processamento com enzimas específicas, o RNA precursor é clivado em 16S, 23S, 4.5S e 5S (Figura 15.16) e, em seguida, ocorre o *splicing* para produzir os tRNA funcionais $tRNA^{Ile}$ e $tRNA^{Ala}$.

Edição do RNA de organelas

Edição do RNA é um termo usado para processos que alteram a sequência das moléculas de RNA e, assim, a diferenciam da sequência correspondente do DNA, sendo definida como uma modificação pós-transcricional. Esse tipo de processamento foi descoberto em transcritos de mitocôndria (cinetoplasto) de tripanosoma (Benne *et al*., 1986) e ocorre em muitos organismos, sendo mais frequente em transcritos de organelas de organismos eucariotos.

Em organelas eucariotas ocorre ao menos dois tipos de edição do mRNA: conversão de nucleotídeo (pela substituição ou modificação de nucleotídeo), geralmente C por U; inserção ou deleção de nucleotídeo. A conversão de nucleotídeos ocorre em mitocôndrias e cloroplastos de plantas, mamíferos e vírus, e resulta em mudanças menos dramáticas na sequência do mRNA. A frequência de edição do RNA em organela e em genes específicos e a distribuição parecem ser ao acaso. Em mitocôndrias, a conversão de C para U ocorre em aproximadamente todas as proteínas que codificam transcritos. Em mitocôndrias de plantas foi estimado um número em torno de 500 sítios de edição, e aproximadamente 40 em cloroplastos. A edição por inserção e deleção de nucleotídeo foi inicialmente observada em mitocôndrias de tripanossoma, na qual o transcrito é editado pela inserção de segmentos de Us presentes em pequenas moléculas de RNA, chamadas RNA guia (gRNA), as quais contêm a informação necessária para a edição. O RNA guia é complementário à parte editada do mRNA maduro e tem uma pequena sequência poli-U na posição terminal 3', que é inserida durante a edição. A principal consequência metabólica desse tipo de edição de mRNA em tripanossoma é a criação de novos quadros de leitura, incluindo, em alguns casos, a criação de códons de iniciação e terminação.

O genoma mitocondrial humano é bastante compacto, com um gene seguido do outro, com poucos ou nenhum nucleotídeo entre eles. Dos 13 genes que codificam proteínas, sete têm códons de terminação que não estão codificados no DNA, mas foram criados pós-transcricionalmente por edição do mRNA; seis dos sete genes terminam em U ou UA e são imediatamente seguidos por genes de tRNA. Quando o último intron é clivado do transcrito primário, a poliadenilação do mRNA precedente cria um códon de terminação UAA no quadro de leitura. Duas proteínas codificadoras de genes (*atp*ase 6 e *cox* III) têm códon de terminação e iniciação sobrepostos, por exemplo: UAAUG. A quebra entre os dois resíduos de A destrói o códon de terminação da ATPase 6, que é regenerada pela poliadenilação 3'.

Em plantas, o mecanismo de edição pela substituição de um C por U foi reconhecido pela primeira vez em mitocôndrias em 1989, por Covello e Gray, e 3 anos depois, em cloroplastos (Hoch *et al*., 1991). Algumas angiospermas sem sementes, como as criptógamas (samambaias e musgos), apresentam a alteração de U para C e de C para U nos sítios de edição. Até o momento não foi observada edição do RNA em RNA citoplasmático de plantas, somente nas organelas.

A maioria dos eventos de edição do mRNA ocorre na primeira ou segunda posição dos códons, com forte preferência para a segunda posição do códon; assim, geralmente altera o códon original no genoma e o transcrito no precursor do mRNA. Consequentemente, o RNA maduro especifica um aminoácido diferente do que o codificado pelo DNA genômico. A edição de C para U pode alterar qualquer códon que contenha um nucleotídeo C, incluindo a criação de códons de iniciação, por mudar um ACG para AUG, e a introdução de sinais de terminação, alterando CGA para UGA ou CAA para UAA (Figura 15.17). A alteração de um U para C em criptógamas pode converter um códon de terminação, como um UAA, em um CAA, que codifica o aminoácido glicina. Desse modo, uma sequência codificadora pode não somente se tornar diferente da sequência genômica original, como também ser estendida ou encurtada pela alteração de um códon de terminação ou pela adição deste.

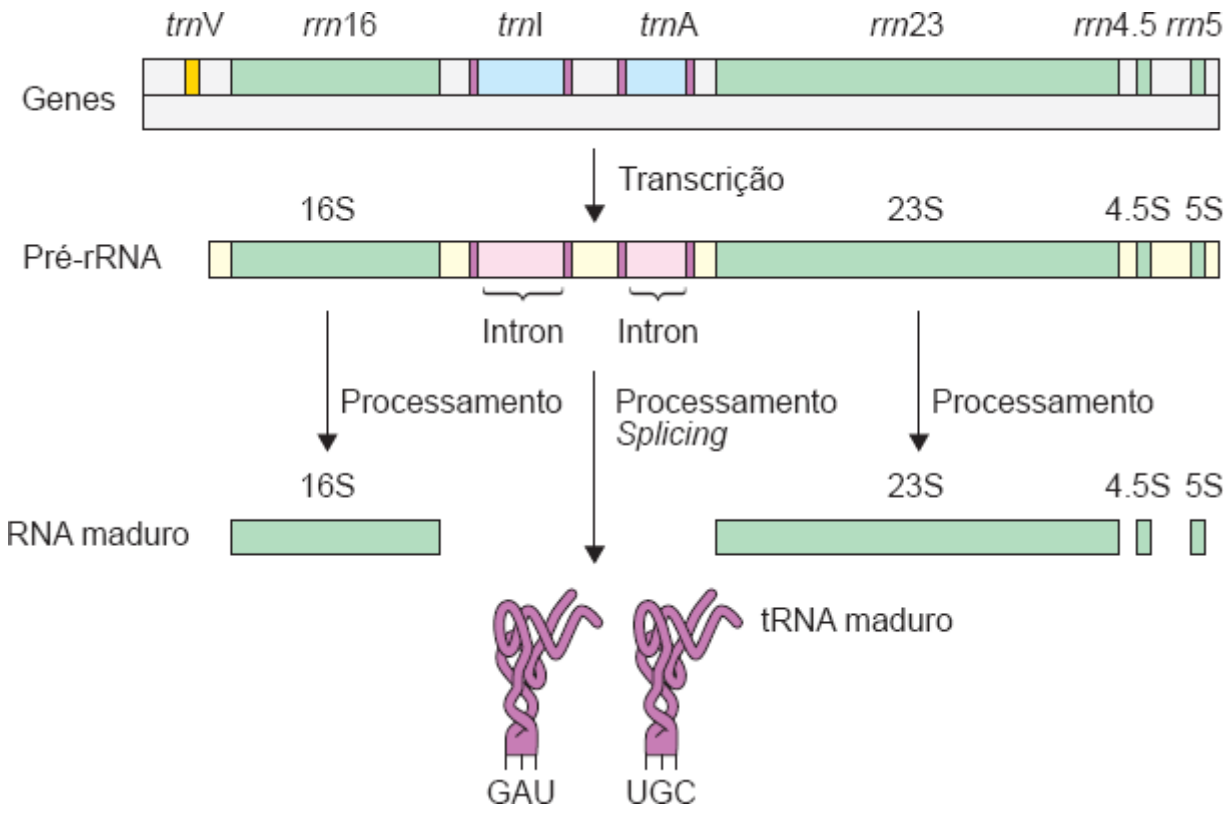


Figura 15.16 Processamento do pré-rRNA de cloroplastos de plantas. O operon rRNA de cloroplastos também codifica dois tRNA nas regiões entre genes os 16S e 23S. Esses tRNA são interrompidos por introns e requerem processamento e *splicing* para produzir os tRNA finais.

As alterações de edição do RNA ocorrem não somente em mRNA, mas também em tRNA, introns e sequências não traduzidas 5' e 3' em plastídeos e mitocôndrias de plantas. Parecem estar ausentes em RNA ribossômicos.

O nucleotídeo, para ser editado, precisa ser reconhecido dentro do universo dos nucleotídeos citocinas em uma população de moléculas de RNA. Foram identificadas sequências de 20 a 25 nucleotídeos na posição 5' antes do nucleotídeo citocina, que, uma vez reconhecida por fatores *trans*, induz à atividade de edição. Esses fatores são codificados no núcleo e pertencem à família de proteínas PPR (pentatricopeptídeo repetida). Essas proteínas são transportadas para as mitocôndrias e para os plastídeos, onde estão envolvidas em várias etapas de processamento do RNA, incluindo *splicing* de introns, estabilidade do RNA e acesso à tradução (Figura 15.18).

As principais consequências da edição do RNA em organelas de plantas são: introdução de códons de iniciação e terminação; mudanças na especificidade do códon, levando à conservação das sequências; melhora no sítio de ligação dos ribossomos e na estrutura do RNA, para ocorrer o *splicing*; e favorecimento da transferência de genes para o núcleo.

Herança das organelas

Organelas que contêm DNA, como cloroplastos e mitocôndrias, são herdadas de maneira não mendeliana em todos os organismos eucariotos. Em muitos organismos, o genoma organelar é herdado somente de um dos pais, sendo a herança maternal a mais comum (presente em 80% das angiospermas). Entretanto, plastídeos e seu DNA podem ter herança paterna ou ser biparentais. Em plantas superiores, genomas plastidiais normalmente não sofrem recombinação sexual até mesmo quando ocorre herança biparental. Isso significa que, em raras exceções (que podem ser considerados acidentes), não ocorrem fusão de cloroplasto e recombinação genética.

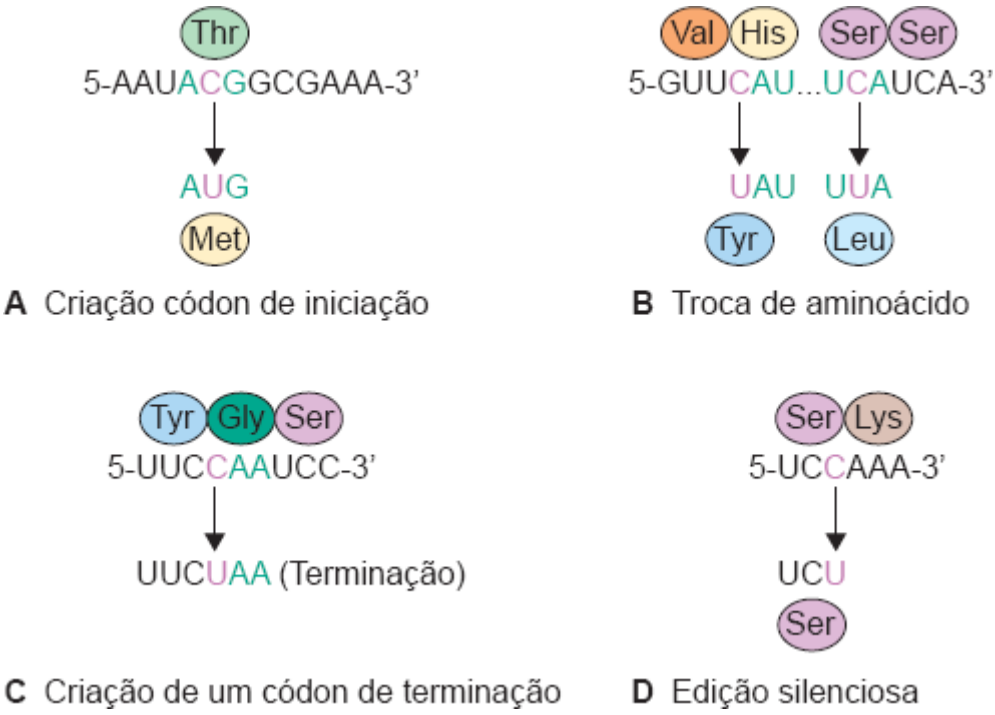


Figura 15.17 Diferentes tipos de edição do mRNA envolvendo substituição de base que pode alterar o polipeptídeo codificado pela sequência de RNA. **A.** Substituição resulta na criação de um códon de iniciação. **B.** Substituição resulta em troca de aminoácido que pode ou não interferir na função da proteína. **C.** Substituição resulta na criação de um códon de terminação e provavelmente estabelece a terminação prematura da tradução. **D.** Substituição não altera a sequência de aminoácidos.

Herança uniparental tem sido considerada uma regra, embora a primeira exceção tenha sido observada em *Pelargonium* (gerânio), que apresenta herança biparental, descrita por Baur em 1909, e foi descoberta simultaneamente com a existência da herança maternal de plastídeos em espécies de angiospermas. Sabe-se que a transmissão é controlada por genes nucleares, mas os fatores envolvidos não são conhecidos.

Herança maternal cloroplastidial

A grande maioria das angiospermas apresenta modo de herança maternal e, assim, não transmite plastídeos e genes plastidiais por meio do pólen. Diferentes espécies podem utilizar diferentes mecanismos para eliminar plastídeos paternais e/ou genomas plastidiais paternos em algum momento entre a formação do grão de pólen (micrósporo) e a fertilização do ovo por um dos dois núcleos espermáticos. O mecanismo citológico de herança maternal deve ser determinado em espécie por espécie.

De acordo com o mecanismo de eliminação paternal de plastídeo, existem ao menos quatro subtipos diferentes de herança maternal que podem ser distinguidos (Hagemann e Schröder, 1989 Greiner *et al*., 2014). Essa classificação é amplamente baseada em pesquisas com microscopia eletrônica, com observações do destino dos plastídeos durante o desenvolvimento do gametófito masculino (Figura 15.19).

Os subtipos são nomeados com base na primeira espécie em que o mecanismo foi descrito.

Tipo *Lycopersicon*

Em angiospermas, exemplificado pelo tomate, o desenvolvimento do gametófito masculino inicia com divisões mitóticas nas células germinativas da célula-mãe, originando quatro micrósporos haploides. Subsequentemente, esses micrósporos sofrem duas divisões mitóticas, referenciadas como mitoses do pólen. A primeira mitose do pólen envolve uma divisão assimétrica do micrósporo haploide, resultando em uma célula vegetativa grande e uma célula germinativa pequena. A célula vegetativa recebe a maior parte do citoplasma do micrósporo. A célula germinativa produz duas células espermáticas por outra divisão mitótica (segunda mitose do pólen). Esse tipo maternal de herança plastidial envolve a exclusão plastidial durante a primeira mitose do pólen. A divisão assimétrica do micrósporo resulta em uma célula vegetativa que contém todos os plastídeos e uma célula germinativa que é livre de plastídeos. Assim, ambas as células espermáticas não apresentam plastídeos. É geralmente assumido que herança plastidial, na maioria das espécies de angiospermas, segue esse mecanismo de exclusão do pólen.

Tipo *Solanum*

Em um número de espécies de angiospermas, a herança maternal ocorre pela degradação do plastídeo paternal. Durante o desenvolvimento do gametófito masculino em espécies pertencentes ao tipo *Solanum* de herança maternal, os plastídeos nas células germinativas são seletivamente destruídos, enquanto aqueles nas células vegetativas permanecem intactos. Assim, como no modelo *Lycopersicon*, as duas células espermáticas não apresentam plastídeos. É interessante que mesmo as espécies altamente próximas diferem no mecanismo citológico, conferindo herança maternal; por exemplo, tomate (*Solanum lycopersicum*) e batata (*Solanum tuberosum*) pertencem ao mesmo gênero, mas diferem no modo de herança maternal dos plastídeos.

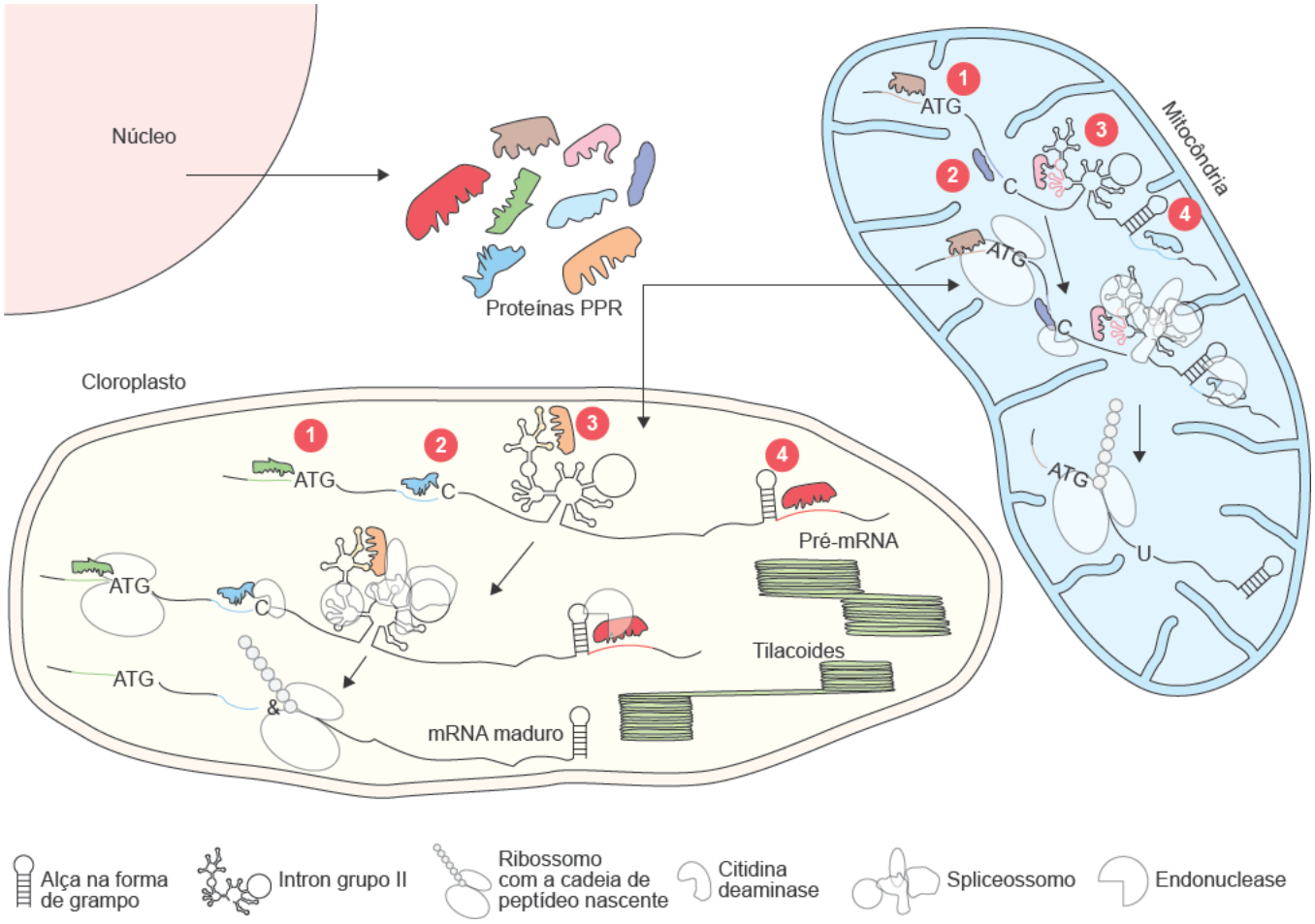


Figura 15.18 Funções moleculares da PPR (proteínas pentatricopeptídeo repetidas). Essas proteínas são definidas pela presença de uma sequência de 35 aminoácidos repetidos em sequência até 30 vezes. Elas garantem o correto processamento pela participação dos seguintes processos: 1) síntese de proteínas: liga-se a sequências específicas na região 5'-UTR de mRNA; 2) edição do mRNA: liga-se a pequenos elementos *cis* imediatamente a montante do sítio de edição do RNA, necessário para o processamento de C para U; 3) *splicing*: presente nas sequências dos introns do grupo II; 4) estabilidade do RNA: tem função na clivagem e estabilidade, e a ligação de PPR poderia recrutar endonucleases para clivagem do RNA ou se ligar à sequência, evitando a clivagem por endonucleases e, assim, mantendo a estabilidade do RNA. Possivelmente a ligação ocorre nas proximidades da estrutura secundária, (estrutura de grampo) que é alvo para endonucleases. Adaptada de Schmitz-Linneweber e Small, 2008.

Tipo *triticum*

Esse tipo de herança maternal é restrito às gramíneas. Plastídeos são encontrados em células germinativas e espermáticas, entretanto, essas espécies não transmitem plastídeos paternalmente no zigoto. Durante a fertilização da célula-ovo por uma das células espermáticas, os plastídeos e as mitocôndrias são excluídos com o conteúdo do citoplasma e, dessa maneira, não entram no zigoto junto com o núcleo da célula espermática.

Herança biparental

Um pequeno grupo de plantas transmite os plastídeos biparentalmente. Erwin Baur e Carl Correns, ao estudarem *Pelargonium* (gerânio), *Melandrium*, *Mirabilis*, *Antirrhinum* e *Aquilegia*, há mais de 100 anos, notaram que os

plastídeos apresentavam herança diferente entre as espécies. Enquanto Correns observou que mutantes de *Melandrium*, *Antirrhinum* e *Aquilegia* transmitiam as cores de folhas alteradas maternalmente, Erwin Baur observou que características semelhantes eram transmitidas via pólen em *Pelargonium*, e concluiu que os plastídeos eram herdados biparentalmente, porém de maneira não nuclear e não mendeliana, estabelecendo que o plastídeo era o responsável pelo modo de herança extranuclear.

Herança biparental de cloroplastos foi encontrada em diferentes espécies, incluindo *Oenothera* (estrela-da-tarde), *Hypericum* (erva-de-são-joão) e *Medicago* (alfafa). Muitos estudos sobre a herança maternal e paternal dos plastídeos dessas espécies revelaram diferenças interessantes, como demonstrado em *Oenothera* e *Hypericum*, que apresentam baixa taxa de transmissão paterna, verificada por cruzamentos entre plastomas mutantes brancos e plantas verdes selvagens. Quando o plastoma mutante servia como mãe, obtinham muitas plantas brancas e variegadas e poucas verdes; em contraste, quando o plastoma mutante era o pai (doador de polén), as plantas obtidas eram verdes ou variegadas, poucas eram brancas. Em gerânio e alfafa, a contribuição paterna é bem maior.

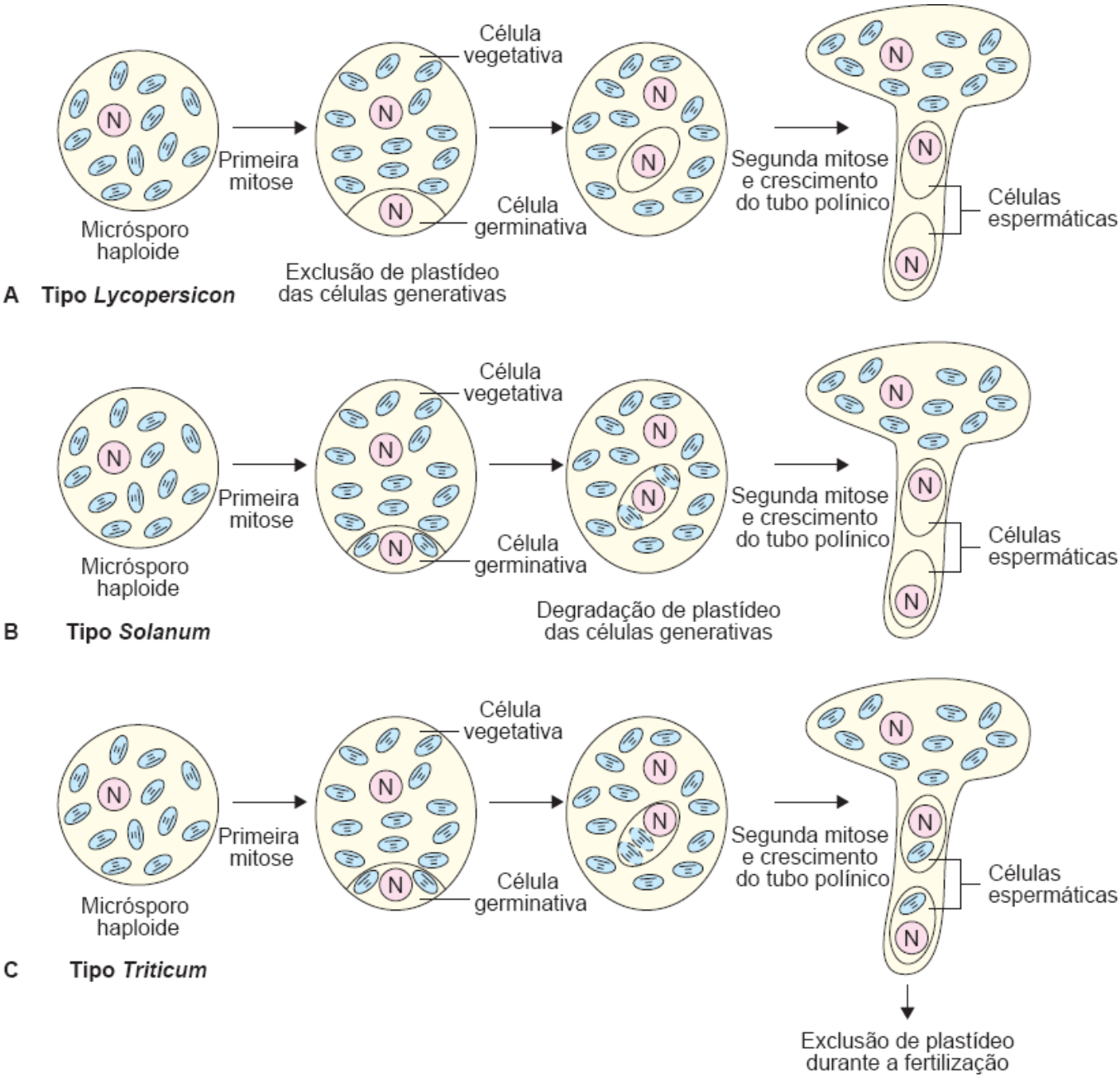


Figura 15.19 Mecanismos de herança maternal em cloroplastos de angiospermas. Espécies que pertencem ao tipo *Lycopersicon* excluem os plastídeos do polén na mitose I, resultando na divisão celular desigual em que as células generativas recebem todos os plastídeos. O tipo *Solanum* exclui os plastídeos após a mitose I do pólen, e suas células contêm poucos plastídeos, os quais são seletivamente degradados (por mecanismo não conhecido) antes da divisão das células generativas para as duas células espermáticas na mitose II no pólen. O tipo *Triticum* produz células espermáticas que contêm plastídeos, contudo são excluídos na fertilização e não entram no citoplasma da

célula-ovo. Outros mecanismos possibilitam que plastídeos paternos entrem na célula-ovo, mas não contribuem para o embrião. Adaptada de Bock, 2007.

Tipo *Chlamydomonas*

A alga verde *C. reinhardtii* tem um único e grande cloroplasto por célula. Existem algas "macho" e "fêmea", que são morfologicamente não distintas e referenciadas como tipo + (mt+, "femea") e tipo – (mt-, "macho"). A herança de organelas em *Chlamydomonas* apresenta várias características interessantes. Primeira – os genomas de cloroplastos e mitocôndrias são herdados de modos opostos – o DNA cloroplastidial é herdado maternalmente, enquanto o DNA mitocondrial é transmitido paternalmente. Segunda – durante a singamia, o cloroplasto materno se funde com o cloroplasto paterno. Terceira – o DNA cloroplastidial em mt+ é metilado pela DNA metiltransferase, convertendo citosina em 5-metilcitosina. Em contraste, acredita-se que o DNA cloroplastidial de plantas não seja metilado, ao menos em tecidos somáticos, apesar de a metilação da citosina ter sido observada.

Se dois cloroplastos parentais se fundem por cruzamento, como ocorre a herança maternal no DNA da alga? É possível que, após a singamia, o programa de maturação zigótico leve à destruição do DNA do cloroplasto parental mt-, enquanto o genoma cloroplastidial mt+ sobrevive. A degradação do genoma paterno ocorre por endonucleases antes da fusão dos dois parentais, resultando na herança uniparental do DNA plastidial maternal. Existe especulação sobre a diferença na metilação do DNA, a qual poderia ser responsável pela degradação seletiva; por analogia com o sistema de restriçãometilação em eubactéria, esse modelo poderia significar que a metilação protegeria o genoma plastidial mt+ da degradação por endonucleases. Contudo, aparentemente, a metilação do DNA não necessariamente protege o genoma plastidial mt+ em estágios iniciais do zigoto, mas pode afetar a replicação do genoma plastidial nas células mt+ e mt-. Assim, o mecanismo ainda não está totalmente estabelecido.

Apesar da herança maternal do DNA plastidial, moléculas de DNA plastidial paternas podem sobreviver até a fusão de cloroplastos ocorrer e recombinar com o genoma plastidial maternal. São chamados de "zigotos excepcionais" e ocorrem espontaneamente, à baixa frequência de 1 a 9%, dependendo do genótipo da linhagem da alga e nas condições experimentais.

Herança paternal

É predominante nas gimnospermas e, entre as angiospermas, somente o kiwi (*Actinidia deliciosa*) apresenta herança paternal. O fato de que dificilmente há marcas fenotípicas visíveis limita a análise de progênies. Análises de microscopia eletrônica confirmam a ausência de plastídeos em células-ovo e a presença em células espermáticas em espécies de gimnospermas (*Pinus taeda*, *Sequoia sempervirens*, *Larix decidua* e outras).

Aplicações biotecnológicas dos cloroplastos

O conhecimento do genoma cloroplastidial tornou essa organela um importante alvo para manipulações genéticas com aplicações biotecnológicas. As plantas geneticamente modificadas têm revolucionado a agricultura nas últimas décadas pela introdução de variedades resistentes a insetos e patógenos, tolerantes a herbicidas, com maior teor de vitaminas e de rápido crescimento, que contribuem para um marcante aumento de produtividade e qualidade dos alimentos. Todas as plantas modificadas geneticamente disponíveis comercialmente são resultantes da modificação no genoma nuclear, mas, alternativamente, a metodologia de introdução de genes no genoma cloroplastidial se apresenta muito atrativa, por apresentar algumas características relacionadas à organela, entre elas:

- O genoma é poliploide, com mais de 10 mil cópias em cada célula, e, por isso, capaz de acumular altos níveis de proteínas (até 70% de proteína solúvel total), enquanto inserção no genoma nuclear resulta em baixo nível de expressão dos transgenes, em geral menor que 2%
- O fluxo de transgenes pelo pólen é reduzido ou inexistente, em razão da ausência de organelas no pólen na maioria das espécies cultivadas (ver sobre herança das organelas)
- São possíveis a inserção e a expressão de múltiplos genes simultaneamente, em um único evento de transformação, porque as organelas apresentam sistema de organização de genes em operons
- É possível expressar proteínas que possam causar risco de danificar ou matar a célula, devido ao fato de as proteínas expressas ficarem compartimentalizadas na organela, protegidas pelas duas membranas que a delimitam, herdadas do processo de endossimbiose já descrito.

A transformação do genoma plastidial foi obtida primeiro em *Chlaymydomonas reinhardtii*, em 1988 (Boynton *et al*., 1988), e em plantas em tabaco em 1990, pelo grupo do Prof. Pal Maliga (Svab *et al*., 1990). Desde então, essa metodologia tem sido implementada em diversas espécies de plantas, com expressão de diferentes genes com aplicações biotecnológicas e de pesquisa básica. O processo de integração de genes no genoma cloroplastidial ocorre exclusivamente por recombinação homóloga. Assim, na construção de vetores de transformação, é essencial que o gene de interesse seja flanqueado por sequências de genes cloroplastidiais vizinhos à região intergênica, onde será introduzido o DNA, e, assim, a integração do transgene nesse local seja direcionada (Figura 15.20 A). Como o genoma plastidial é bastante compacto, com sequências codificadoras de genes em ambas as fitas, a inserção do transgene na região intergênica é um requisito essencial para não interromper a sequência de genes nativos do cloroplasto. Por isso, é necessário conhecer a sequência do genoma da espécie que será manipulada para que seja possível fornecer as sequências flanqueadoras do transgene (cassete de expressão) e as sequências regulatórias.

O processo de inserção do transgene no interior dos cloroplastos é via biolística. Essa tecnologia envolve a adesão do DNA (vetor de transformação) a micropartículas de ouro ou tungstênio, que são bombardeadas em um tecido vegetal a alta pressão e velocidade, obtidas pelo uso de um equipamento específico (Figura 15.20 B), o acelerador de partículas. Como o processo é aleatório, o transgene pode ser dirigido ao núcleo, cloroplastos, mitocôndrias ou citoplasma e, possivelmente, integrar-se a qualquer dos genomas presentes na célula; contudo, ocorrerá a expressão do transgene nos cloroplastos ou no núcleo definido pelas sequências regulatórias. A transformação de mitocôndrias de plantas ainda não é possível, mas sucesso foi obtido na alga *Chlamydomonas* e na levedura *Saccharomyces.* Algumas características e dificuldades dessa tecnologia em mitocôndrias são apresentadas adiante neste capítulo.

Um dos desafios para obter transformação estável de cloroplastos é alterar uniformemente as centenas ou milhares de organelas em uma célula vegetal. Inicialmente, poucas cópias do genoma plastidial são transformadas, as quais são seletivamente amplificadas enquanto as células crescem em cultura de tecidos na presença do agente seletivo. A seleção dos genomas plastidiais transformados é essencial e ocorre pela expressão de um gene marcador seletivo no vetor de transformação, geralmente o *aad*A, que codifica a proteína aminoglicosídeo 3”-adeniltransferase e confere resistência aos antibióticos espectinomicina e estreptomicina. O modo de ação desse antibiótico é pela ligação específica nos ribossomos dos cloroplastos (tipo 70S, como em procariotos), impedindo a síntese de proteínas sem interferir no processo de tradução eucariótica que tem ribossomos 80S no citoplasma. Outros antibióticos, como canamicina e clorofenicol, também podem ser utilizados como agentes seletivos, mas com menor eficiência. Após o processo de inserção do transgene, o tecido vegetal é colocado no meio de cultura contendo nutrientes, reguladores vegetais para induzir a regeneração de plantas e o antibiótico para selecionar os transgênicos, que, depois de 3 a 5 semanas, iniciam a regeneração a partir de uma única célula (Figura 15.20 C).

Além do uso da tecnologia para inserir novas características de interesse agronômico ou industrial em plantas, a transformação tem sido bem-sucedida quando usada para introduzir mudanças nos genes endógenos para estudos da funcionalidade de genes plastidiais por meio de genética reversa. Essa possibilidade tem contribuído significativamente para a fotossíntese, identificação de genes hipotéticos conservados com funcionalidade desconhecida e mecanismos moleculares de expressão de genes, como a contribuição do “*wobbling*” na leitura do código genético e a essencialidade de alguns genes, que codificam tRNA, rRNA e *acc*D (única proteína codificada pelo cloroplasto envolvida na biossíntese de ácidos graxos). A Figura 15.21 apresenta algumas estratégias na genética reversa para interromper um quadro de leitura (orf) pela inserção de um gene marcador seletivo (Figura 15.21 A) ou somente introduzir uma pequena alteração (uma mutação pontual, deleções ou inserções) (Figura 15.21 B).

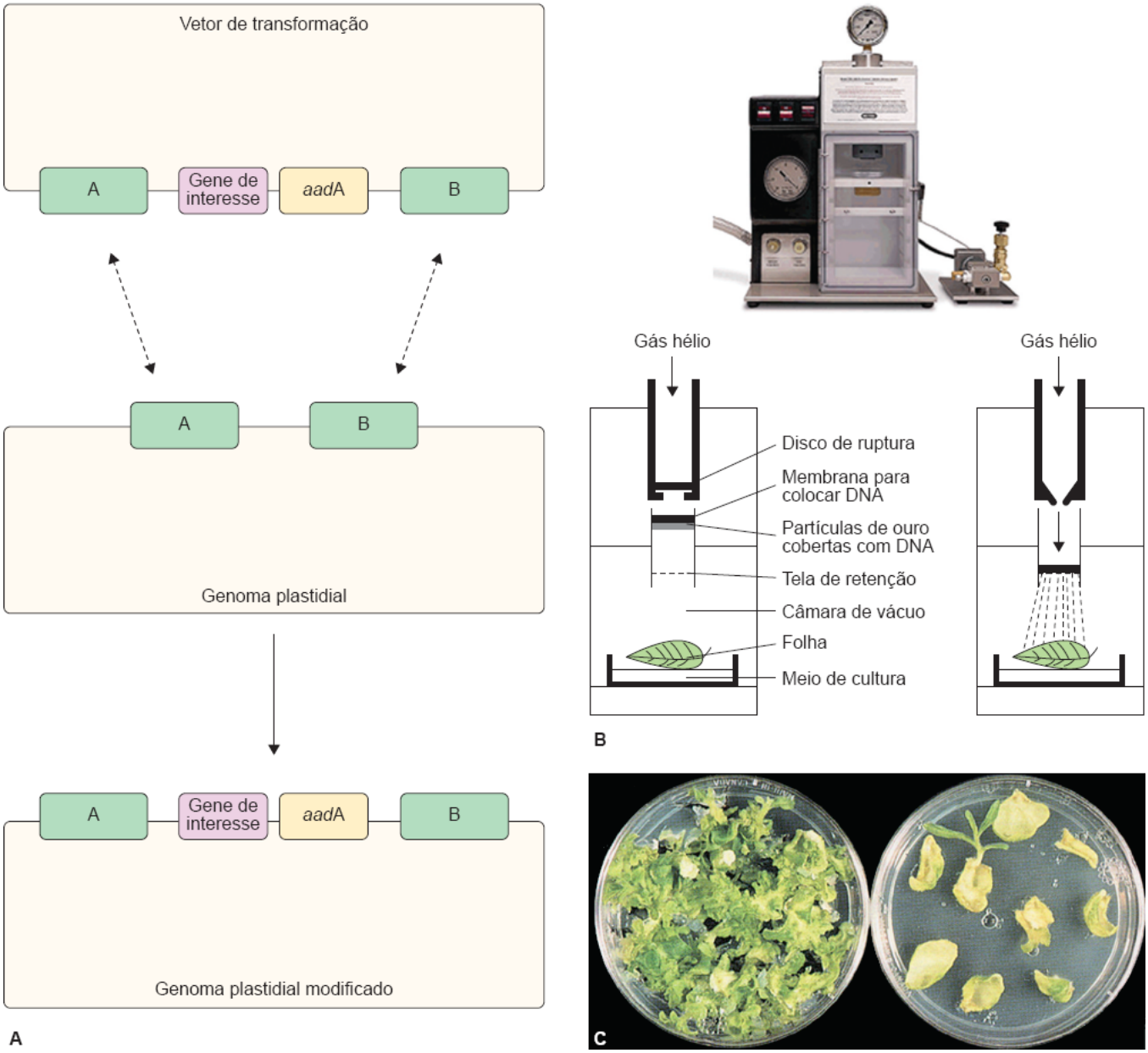


Figura 15.20 Metodologia de transformação genética de cloroplastos. **A.** Representação esquemática do processo de recombinação homóloga para inserção do gene de interesse no genoma plastidial. O vetor de transformação tem sequências de genes cloroplastidiais (A e B) que flanqueiam a região intergênica. Apresenta o cassete de expressão com o gene de interesse e o gene marcador seletivo, geralmente o *aad*A, que confere resistência aos antibióticos espectinomicina e estreptomicina. **B.** Acelerador de partículas para inserir DNA no interior da célula por alta pressão e velocidade. **C.** Após o bombardeamento de DNA nas folhas, estas são cortadas em pequenos fragmentos, para crescimento em meio de cultura, com reguladores de crescimento e antibiótico para induzir a regeneração de plantas a partir de células que receberam o vetor de transformação. A placa à esquerda demonstra a regeneração de plantas em meio sem antibiótico, e a placa à direita contém antibiótico (espectinomicina), para que somente as plantas resistentes (pela presença do gene marcador seletivo -*aad*A) sejam capazes de se desenvolver.

As primeiras plantas transplastômicas (transgênicas cloroplastidiais) apresentaram características de tolerância a herbicidas e resistência a insetos-praga. Entretanto, a tecnologia tem grande potencial para produzir proteínas recombinantes para fins terapêuticos, como vacinas, hormônios de crescimento e endolisinas, que potencialmente podem produzir nova geração de antibióticos e alterar vias metabólicas endógenas para aumentar o valor nutricional de plantas, uma vez que é no cloroplasto onde ocorrem algumas vias biossintéticas de vitaminas e aminoácidos, com exemplos já demonstrados com a vitamina E e o aminoácido triptofano. Além dos produtos gênicos com fins agronômicos e medicinais, a expressão de enzimas industriais tem se tornado uma nova área de interesse na biotecnologia de cloroplastos. Com o interesse crescente por fontes renováveis de energia, enzimas para a produção de biocombustíveis têm chamado atenção. Alguns estudos demonstram que muitas das enzimas que potencialmente podem ser usadas para converter biomassa celulósica em açúcar fermentável, como as celulases, xilanases, glicosidases, pectinases e cutinases, podem ser expressas pelo genoma plastidial em altos níveis. Alguns desses

genes foram isolados de organismos termofílicos, o que é útil para o processamento de biomassa celulósica em escala industrial que utiliza temperaturas altas no processo de produção. Embora, por enquanto, o pré-tratamento de biomassa lignocelulósica permaneça como o principal obstáculo para a produção economicamente viável de etanol celulósico, ter uma fonte barata de enzimas que degradem celulose e hemicelulose pode representar uma parte importante da solução.

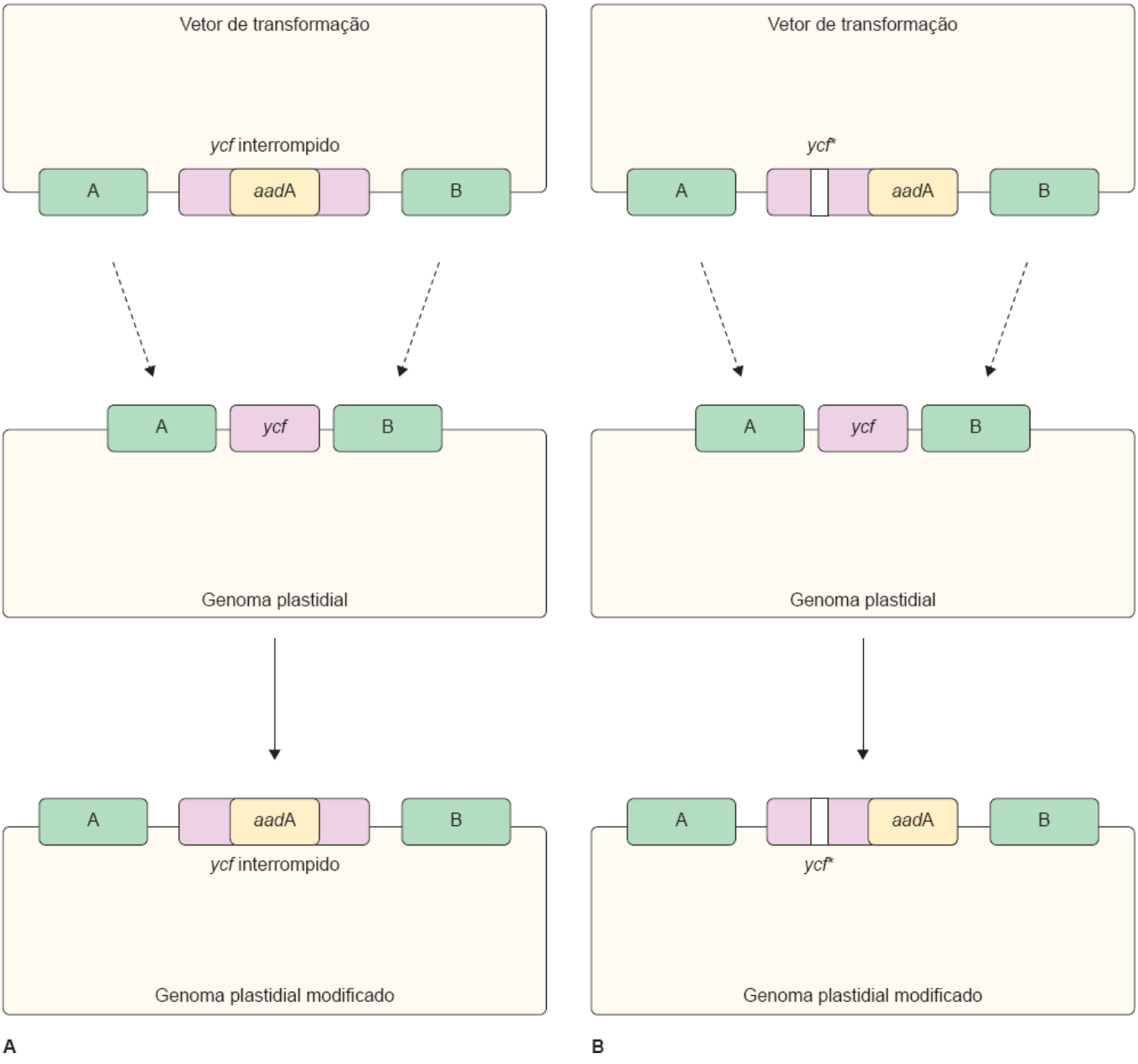


Figura 15.21 Metodologia para genética reversa em cloroplastos. **A.** Uma sequência de quadro aberto de leitura com função desconhecida (*ycf*) pode ser interrompida pela inserção de um cassete de expressão de um gene marcador seletivo; por exemplo, o *aad*A, que confere resistência ao antibiótico espectinomicina. A substituição da sequência *ycf* endógena pela sequência interrompida (ycf^{del}) ocorre por dois eventos de recombinação homóloga, demonstrados pelas setas tracejadas. **B.** Mutações pontuais podem ser introduzidas em um gene endógeno (neste caso, uma *ycf*), o qual será substituído pelo gene alterado (*ycf**), e o gene marcador seletivo (*aad*A) pode ser introduzido da mesma maneira que apresentado anteriormente. Eventos de recombinação homóloga dirigem a substituição do gene alterado (*ycf**) e a inserção do gene marcador seletivo. Nesse caso, é possível ocorrer recombinação homóloga entre o gene *aad*A e o *ycf** e produzir plantas com o gene *aad*A sem o *ycf**.

Existe também interesse na utilização de plantas transgênicas cloroplastidiais, como em fábricas, para produção da chamada "química verde", que são materiais para a indústria química. A recente demonstração da produção de poli-hidroxibutirato (PHB), um bioplástico renovável para mais de 18% da massa seca de plantas de tabaco, representa um exemplo relevante de sucesso para o redirecionamento do metabolismo de plantas em direção à síntese massiva de um novo composto. A alta expressão de PHB provocou o crescimento reduzido das plantas, que pode ser solucionado com a utilização de um operon induzido de PHB; contudo, a tecnologia apresenta alguns

desafios. Desde o estabelecimento da transformação de cloroplastos em tabaco, são poucas as espécies de plantas com protocolo bem-sucedido. Houve progressos na transferência da tecnologia desenvolvida em fumo para espécies como tomate, batata, soja, alface, repolho, berinjela e cenoura, mas a transferência para Arabidopsis, uma planta-modelo para estudos básicos de biologia molecular, e gramíneas (arroz, milho, cana-de-açúcar, sorgo e outros), que representam importantes espécies de interesse econômico, ainda desafia os cientistas da área. Considerando o enorme potencial dessa metodologia, especialmente para os setores farmacêutico, de química verde e de bioenergia, essa tecnologia demonstra ser promissora, com grande expectativa pelo surgimento de novos produtos biotecnológicos.

Genomas mitocondriais

A existência de um genoma próprio na mitocôndria (DNA mitocondrial – mtDNA) está atrelada à origem endossimbiótica dessa organela. Ao longo da evolução, a grande maioria dos genes originalmente codificados pelo mtDNA foram transferidos para o núcleo, como evidenciado pela presença de genes ortólogos no genoma mitocondrial de algumas espécies e no genoma nuclear de outras. Além disso, moléculas de mtDNA de diferentes tamanhos podem ser encontradas em diferentes organismos. Em animais são encontrados genomas mitocondriais de menor tamanho, variando de 16 a 19 Kb. Já em plantas, incluindo algas, o mtDNA varia de 16 a 2.400 Kb, podendo chegar a 11 Mb, como em *Silene conica*, uma angiosperma. Em protistas, o mtDNA pode variar entre 7,7 e 76,5 kb e, em fungos, entre 17 e 176 kb. Essas diferenças são indicativos de diferentes momentos evolutivos da transferência gênica de mitocôndria para o núcleo. Além da grande maioria dos genes que codificam proteínas envolvidas na maquinaria mitocondrial, foram transferidos genes que codificam proteínas de replicação, transcrição e tradução do mtDNA, tornando a função mitocondrial dependente do núcleo. Nesse sentido, é surpreendente que os eucariotos tenham mantido somente alguns genes na mitocôndria, uma vez que isso implica em um significativo gasto energético. Uma possível explicação é que os genes remanescentes codificam proteínas altamente hidrofóbicas, o que dificulta que elas sejam importadas e corretamente posicionadas na mitocôndria. Uma explicação alternativa sugere que o genoma mitocondrial foi mantido em razão de diferenças nos códigos de síntese proteica mitocondrial e nuclear, o que limitaria a transferência dos demais genes. Finalmente, uma terceira teoria sugere que a manutenção do genoma mitocondrial é importante para o controle metabólico em eucariotos. Independentemente do motivo, a manutenção de um genoma próprio e sua expressão em níveis adequados na mitocôndria parecem ser vitais para os organismos, por exemplo em mamíferos, como exemplificado pela manifestação de sérias patologias em humanos quando da ocorrência de mutações no mtDNA.

Genoma mitocondrial de mamíferos

Organização e replicação do genoma mitocondrial

O mtDNA de mamíferos contém somente cerca de 16.500 pares de bases (podendo variar de 6.000 a 300.000 pares de bases em outras espécies), tem formato circular e está organizado em um complexo núcleo-proteico denominado nucleoide. Múltiplas cópias de mtDNA estão presentes em um mesmo nucleoide, em um arranjo com proteínas envolvidas na replicação e transcrição do mtDNA. A mais comum dessas proteínas é o fator de transcrição mitocondrial A (TFAM), que parece ser importante no empacotamento do mtDNA. Os nucleoides também contêm proteínas que sabidamente residem na membrana mitocondrial interna, como o translocador de nucleotídeos adenina 1 (SLC25A4), o que sugere que o mtDNA está preso à membrana. Treze proteínas essenciais para a produção de adenosina trifosfato (ATP) via fosforilação oxidativa são codificadas pelo mtDNA. As outras cerca de 70 proteínas que compõem a cadeia respiratória mitocondrial são codificadas por genes nucleares e devem ser importadas para a mitocôndria para interagirem com as proteínas codificadas na organela. O genoma mitocondrial também codifica 22 RNA transportadores (tRNA) e 2 RNA ribossômicos (rRNA).

As duas fitas que compõem o mtDNA são comumente denominadas leve (L) e pesada (H), uma vez que podem ser distinguidas com base nas suas densidades quando separadas em gradiente de cloreto de césio desnaturante. A maior parte da informação codificada pelo mtDNA está localizada na fita H, que contém os dois genes codificantes de rRNA, 14 genes codificantes de tRNA e 12 genes codificantes de mRNA. A fita L por sua vez, codifica oito tRNA e um único mRNA. O mtDNA é uma molécula extremamente compacta, já que seus genes não apresentam introns e, com exceção de uma região regulatória (D-loop), regiões intergênicas são inexistentes ou limitadas a poucos nucleotídeos. Além disso, alguns dos genes codificantes de mRNA se sobrepõem, parte dos códons de

terminação são criados pós-transcricionalmente por poliadenilação dos mRNA e os genes de rRNA e tRNA são excepcionalmente compactos (Figura 15.22).

Em mamíferos, uma fita simples de DNA de cerca de 700 nucleotídeos, complementar à fita L, é comumente encontrada na região D-loop. Essa fita é produzida pelo término prematuro da replicação da fita H e permanece pareada à fita L, formando uma tripla-fita de DNA que caracteriza a região. A D-loop constitui o principal sítio de controle da replicação e transcrição dos genes codificados pelo mtDNA e contém a origem de replicação da fita H e os principais promotores de transcrição das fitas L e H (Figura 15.22).

A origem de replicação da fita H está posicionada logo depois do promotor de transcrição da fita L, uma vez que o transcrito inicial dessa fita é usado como iniciador de replicação do mtDNA. Alternativamente, o transcrito inicial pode servir para transcrição dos genes da fita L, o que parece ser determinado pela formação de G-quadruplexes entre o RNA recém-sintetizado e a fita H. No caso de o transcrito inicial ser utilizado para a replicação do mtDNA, o processamento do transcrito fica a cargo de uma endonuclease mitocondrial (RNase MRP). Ainda, a decisão de utilizar a fita nascente de RNA como iniciador não implica, necessariamente, a replicação do genoma mitocondrial, já que, na maior parte das vezes, a replicação é prematuramente abortada, originando a tripla-fita de DNA que caracteriza a região D-loop. Não se sabe o que determina a continuação ou não da replicação, mas fortes evidências sugerem que esse mecanismo de terminação prematura sirva como um importante regulador do número de cópias de mtDNA na célula. Em muitos casos, a replicação do mtDNA é abortada prematuramente, em cerca de 95% da vezes, sendo que, quando necessário, a frequência de aborto pode ser diminuída para rapidamente aumentar ou restituir o número de cópias de mtDNA na célula. Uma vez que a forquilha de replicação ultrapassa a sequência associada à terminação (TAS), a síntese da fita H prossegue unidirecionalmente ao longo da fita molde (Figura 15.22). Uma DNA polimerase específica da mitocôndria (POLG) é necessária para replicação do mtDNA, juntamente com uma helicase (TWINKLE) e proteínas que se ligam à fita simples de DNA (mtSSB), todas codificadas no núcleo e importadas pela mitocôndria. Somente depois de já ter replicado dois terços da fita H, a forquilha de replicação encontra a origem de replicação da fita L, desencadeando o início da replicação dessa fita em sentido oposto. Esse modelo de replicação do mtDNA foi o primeiro a ser proposto e era o mais aceito, porém, mais recentemente, foram fornecidas evidências de um modo alternativo de replicação baseado na existência de múltiplas origens e replicação simultânea de ambas as fitas. Além disso, tem sido proposto que uma origem de replicação alternativa é responsável pela manutenção do mtDNA em condições normais, enquanto a síntese a partir da origem de replicação discutida anteriormente é responsável pela reposição de cópias nos casos de depleção e replicação acelerada em resposta a demandas fisiológicas.

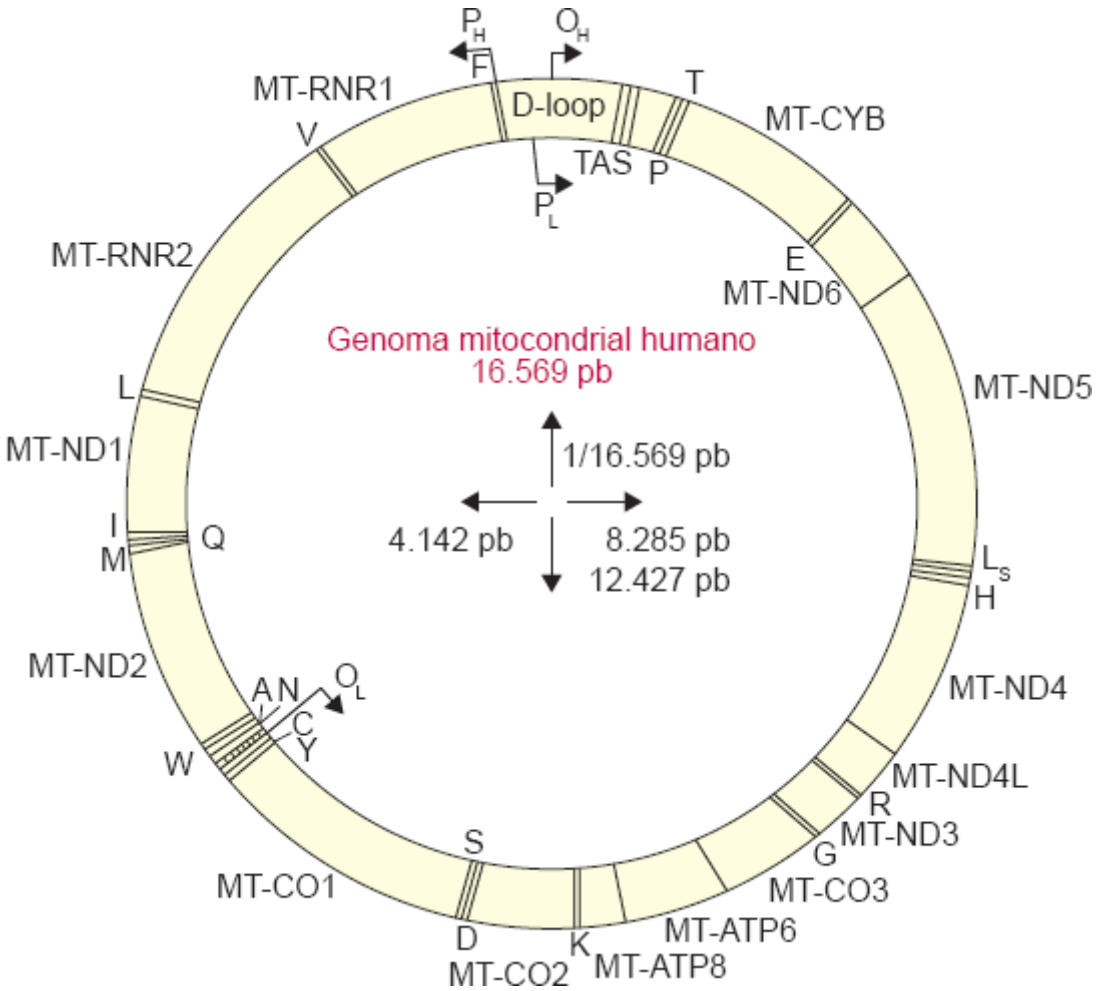


Figura 15.22 Mapa esquemático do genoma mitocondrial humano. O círculo de fora representa a fita pesada (H), que contém a maior parte dos genes; o círculo interno representa a fita leve (L). As origens de replicação da fita pesada (O_H) e da leve (O_L) e a direção da síntese de DNA são indicadas pelas setas na região regulatória D-loop. Da mesma maneira, estão indicados a sequência associada à terminação (TAS) da replicação, os promotores de transcrição de ambas as fitas (P_H e P_L) e o sentido da síntese de RNA. Os 22 genes codificadores de tRNA são indicados pelo código de letras individuais correspondentes aos seus respectivos aminoácidos. Os dois genes codificadores de rRNA e os 13 genes codificadores de mRNA são indicados pelos seus símbolos oficiais. As quatro setas no centro dos círculos indicam o número de pares de bases (pb) em quatro porções diferentes da molécula de mtDNA. Adaptada de Taanman, 1999.

Diferentemente do DNA nuclear, existem de centenas a milhares de cópias de mtDNA em cada célula e o genoma mitocondrial é continuamente renovado e replicado ao longo do ciclo celular. Portanto, embora a replicação do mtDNA esteja sob o controle do núcleo, ela não ocorre em um momento específico do ciclo celular e pode sofrer influência do metabolismo e do tipo celular. Dessa maneira, células de tecidos com elevada exigência energética, como as do tecido nervoso, muscular e hepático, contêm muito mais cópias do que células de tecidos com menor exigência energética. A existência de múltiplas cópias, até mesmo dentro de uma mesma organela, minimiza o efeito de mutações sobre a função mitocondrial. Esse fenômeno, conhecido por complementação mitocondrial, possibilita que cópias selvagens codifiquem proteínas normais, que compensam o defeito causado por cópias mutantes. A complementação mitocondrial é favorecida pela ocorrência constante de fissão e fusão mitocondrial, que possibilita a troca de mtDNA, RNA, proteínas e outros componentes entre as organelas de uma mesma célula. No entanto, isso somente é possível quando cópias mutantes e selvagens coexistem em uma célula (heteroplasmia), não sendo possível nos casos de homoplasmia (p. ex., presença de um único tipo de mtDNA). Além disso, quando a porcentagem de moléculas mutantes ultrapassa um limiar, a complementação é incapaz de impedir o dano sobre a função mitocondrial, o que resulta na manifestação de sérias patologias. Esse limiar é diferente dependendo da mutação no mtDNA e do tipo celular, variando de 50 a 90% nos casos de deleção e ficando acima de 90% para mutações em tRNA.

A herança mitocondrial em mamíferos é exclusivamente materna, uma vez que o óvulo contribui com uma quantidade muito maior de mtDNA que o espermatozoide. Acreditava-se que isso se devia à eliminação das

mitocôndrias paternas por autofagia logo após a fecundação. No entanto, trabalhos mais recentes têm contestado essa versão em várias espécies (inclusive em camundongos), fornecendo evidências de que as mitocôndrias presentes nos espermatozoides no momento da fecundação são vacuolares e destituídas de mtDNA. Independentemente do mecanismo, a eliminação do mtDNA paterno é evolutivamente vantajosa, uma vez que ele pode ser danificado pelas espécies reativas de oxigênio envolvidas com a fecundação, o que pode resultar em sérias doenças mitocondriais. No entanto, mesmo com a eliminação das mitocôndrias paternas, mutações no mtDNA estão presentes em cerca de 0,5% da população, devido, entre outros fatores, à suscetibilidade aumentada de ocorrência de mutações no mtDNA em células somáticas e germinativas. Quando presente na linhagem germinativa, a mutação no mtDNA pode ser transmitida para a geração seguinte, mas a transmissão depende da quantidade de moléculas mutantes. Em razão do gargalo genético mitocondrial, somente uma fração muito pequena das cerca de 250.000 cópias de mtDNA existentes no oócito é transmitida para a próxima geração. Como resultado, a frequência de uma mutação sofre drásticas alterações de uma geração para outra e tende ao restabelecimento da homoplasmia. Além disso, mesmo em tecidos somáticos, a porcentagem de cópias mutantes pode variar ao longo do tempo. Antes da divisão celular, as mitocôndrias se fragmentam e se distribuem espacialmente por toda a célula para assegurar sua distribuição igualitária entre as células-filhas. No entanto, esse mecanismo não garante que a distribuição se dará por igual em termos de mtDNA com ou sem mutações. Além disso, nenhum mecanismo assegura a replicação de todas as moléculas de mtDNA de uma célula uma única vez por ciclo, o que resulta em um padrão estocástico de segregação que tende à homoplasmia. Apesar disso, é sabido que algumas mutações no mtDNA, tanto em humanos quanto em modelos animais, seguem um padrão bem definido de herança em tecidos somáticos, podendo aumentar ou diminuir em frequência dependendo do tempo e da mutação.

Expressão gênica e regulação em mitocôndrias

Uma vez que a função mitocondrial depende de polipeptídeos codificados tanto pelo DNA nuclear como pelo mtDNA, a expressão de subunidades codificadas por ambos os genomas deve ser estritamente coordenada. Diversos fatores regulam o metabolismo mitocondrial, mas o mais conhecido é o coativador PPAR gama 1 (PPARGC1), codificado pelo DNA nuclear. No núcleo, o PPARGC1 regula a expressão dos fatores de respiração nuclear 1 e 2 (NRF1 e NRF2). Por sua vez, o NRF1 e o NRF2 coordenam, também no núcleo, a expressão de genes codificantes de proteínas da cadeia transportadora de elétrons, bem como de fatores que regulam a replicação, transcrição e tradução do mtDNA. Em nível mitocondrial, tanto a replicação como a transcrição do mtDNA são reguladas por fatores importados pela mitocôndria que interagem com ele na região da D-loop. Proteínas como o TFAM, fatores de transcrição mitocondrial B1 e B2 (TFB1M e TFB2M) e uma RNA polimerase mitocondrial (POLRMT) interagem com os promotores de transcrição de ambas as fitas e são responsáveis por iniciar a transcrição bidirecional do mtDNA. Esses mesmos fatores também regulam a replicação do mtDNA, já que, como discutido anteriormente, o transcrito inicial da fita L é usado como iniciador da replicação. O TFAM constitui um fator-chave na transcrição e na replicação, uma vez que ele se liga ao mtDNA, abre a dupla-fita e torce a molécula para expor os promotores de transcrição (Schon *et al*., 2012). Isso possibilita o acesso da POLRMT, do TFB1M e do TFB2M aos promotores e o subsequente início da síntese de RNA. Ambas as fitas do mtDNA são transcritas em duas fitas de RNA policistrônico, as quais, posteriormente, são processadas para originar moléculas individuais de mRNA, tRNA e rRNA. Visto que o mtDNA é uma molécula compacta, com poucas regiões regulatórias, a expressão dos genes mitocondriais é regulada principalmente por controle do número de cópias. Assim, um aumento repentino da necessidade energética, como nos casos de estímulo muscular prolongado, é seguido por aumento do número de cópias de mtDNA e subsequente aumento da expressão dos genes mitocondriais.

Uma vez que regiões intrônicas não estão presentes no mtDNA, o processamento dos mensageiros policistrônicos é relativamente simples e requer o envolvimento de poucas enzimas. Cada gene codificante de proteína ou rRNA é imediatamente flanqueado por, no mínimo, um gene codificante de tRNA (Figura 15.22). Depois de transcritos, os tRNA se organizam em uma estrutura tridimensional típica, que facilita sua excisão para a produção de mRNA e rRNA maduros. As três formas de RNA devem sofrer uma série de modificações catalisadas por enzimas codificadas pelo núcleo antes de serem utilizadas na síntese proteica mitocondrial.

Os tRNA e rRNA codificados pelo mtDNA são utilizados exclusivamente na tradução dos mRNA produzidos no interior da mitocôndria. Os ribossomos mitocondriais de mamíferos caracterizam-se pelo reduzido conteúdo de rRNA, o que é compensado por uma quantidade elevada de proteínas ribossômicas codificadas pelo DNA nuclear.

Os ribossomos mitocondriais têm coeficiente de sedimentação de 55S e são formados por duas subunidades de 39S e 28S, representativas dos rRNA de 16S (MT-RNR2) e 12S (MT-RNR1), respectivamente. Uma característica surpreendente é que quatro códons na mitocôndria diferem do código genético universal, que foi extremamente conservado durante a evolução. Os códons mitocondriais também variam entre espécies, com a existência, inclusive, de organismos em que o código mitocondrial é idêntico ao nuclear, o que sugere que a degeneração de códons é um evento posterior ao surgimento da mitocôndria. O sistema genético mitocondrial também utiliza um mecanismo de decodificação simplificado, possibilitando a tradução de todos os códons com somente 22 tRNA (em contraste com os 32 tRNA normalmente necessários para tradução dos genes nucleares). Isso é possível porque as regras de pareamento códon-anticódon são mais flexíveis e qualquer um dos quatro nucleotídeos na terceira posição pode ser reconhecido pela maioria dos tRNA mitocondriais. Os ribossomos mitocondriais também são sensíveis a antibióticos antibacterianos e a síntese de proteínas inicia-se com N-formil metionina (não com metionina, como no citoplasma), o que reitera a origem endossimbiótica da organela.

Relógio molecular

O mtDNA é comumente utilizado em estudos evolutivos como um marcador de diversidade molecular, considerando que a taxa de substituição de nucleotídeos durante a evolução foi 10 vezes maior no genoma mitocondrial do que no nuclear. A facilidade de amplificação do mtDNA *in vitro* e a conservação entre espécies do conteúdo gênico mitocondrial, com poucas ocorrências de duplicação, ausência de introns e regiões intergênicas bastante curtas, são fatores que também contribuíram para o emprego do mtDNA como um relógio molecular. Uma vez que as mitocôndrias constituem um importante sítio de geração de espécies reativas de oxigênio, o mtDNA está mais sujeito a mutações na sua estrutura do que o DNA nuclear. Além disso, os processos de replicação e reparo do mtDNA não são tão eficientes quanto aqueles do núcleo, o que pode ter contribuído para a elevada taxa de mutação. A pressão de seleção também é menor para o genoma mitocondrial, visto que múltiplas cópias desse genoma estão presentes em uma mesma célula e podem minimizar o efeito de mutações por complementação. O reduzido tamanho do mtDNA e, possivelmente, a reduzida fidelidade dos tRNA e ribossomos durante a síntese proteica são fatores que também contribuem para a propagação de mtDNA mutantes. Como consequência, o mtDNA tem sido utilizado como uma espécie de código de barras molecular, visando à análise da biodiversidade e à identificação de espécies. Com esse propósito, uma sequência de cerca de 600 pares de bases correspondente ao gene mitocondrial MT-CO1 é sequenciada utilizando DNA do organismo em estudo e comparada com a sequência do mesmo gene de outros organismos para cálculo da distância filogenética e classificação taxonômica. Embora o mtDNA seja bastante utilizado com essa finalidade, o seu emprego em estudos de biodiversidade e identificação de espécies tem sido bastante questionado no que se refere ao modo como esse procedimento é realizado e à confiabilidade dos resultados gerados.

Doenças mitocondriais

As doenças mitocondriais relacionadas ao mtDNA podem ser definidas como desordens que surgem diretamente de uma mutação (tanto pontuais como rearranjos) no mtDNA e comprometem diretamente a fosforilação oxidativa e, consequentemente, a função mitocondrial. Até o momento, mais de 270 mutações pontuais já foram descritas, afetando cada gene do mtDNA. Dentre as mutações pontuais, as mais comuns são: uma transição de A para G no nucleotídeo 3.243 do gene MT-TL1, a qual causa encefalomiopatia mitocondrial, acidose láctica e episódios do tipo acidente vascular cerebral (MELAS); uma transição de A para G no nucleotídeo 8.344 do gene MT-TK, que causa epilepsia mioclônica das fibras vermelhas gastas (MERRF); e uma transversão de T para G no nucleotídeo 8.993 do gene MT-ATP6, responsável por neuropatia periférica, ataxia, retinite pigmentosa, convulsões e demência (NARP) e síndrome de Leigh herdada maternalmente (MILS). Com relação a rearranjos no mtDNA, os mais comuns são deleções parciais de grande escala, que removem de 2 a 10 kb (denominado Δ-mtDNA). Essas sequências são flanqueadas por curtas sequências repetidas (5 a 13 pares de bases). Tais deleções podem ocorrer em praticamente qualquer porção da molécula de mtDNA, mas todas elas causam uma de três desordens patogeneticamente similares: síndrome de Kearns-Sayre (KSS), oftalmoplegia externa crônica progressiva (CPEO) e síndrome de Pearson.

Surpreendentemente, mais da metade das mutações pontuais estão localizadas em genes que codificam tRNA, embora tRNA compreendam somente 10% da capacidade codificante do mtDNA (ver Figura 15.22). Os genes que

codificam proteínas, os quais compreendem quase 70% do genoma mitocondrial, respondem por somente 40% das mutações, e os dois genes codificadores de rRNA (15% da capacidade codificante do mtDNA) respondem por somente 2% das mutações. Provavelmente, essa distribuição inesperada se deve ao fato de a maioria dos pacientes serem heteroplásmicos, ou seja, apresentarem uma mistura de mtDNA selvagens e mutantes. Além disso, com uma única exceção, todas as mutações no mtDNA são "recessivas": uma quantidade elevada da forma mutante é necessária para a manifestação do fenótipo clínico. Por motivos desconhecidos, essa quantidade varia entre tipos mutantes e costuma ser maior para mutações em genes codificadores de tRNA (ao redor de 90%) do que em genes codificadores de proteínas (ao redor de 70 a 80%). Assim, embora a prevalência de doenças causadas por mutações no mtDNA seja da ordem de 1 para 4.300, a frequência populacional das dez mutações patogênicas mais comuns no mtDNA é muito maior – aproximadamente 1 em 200 –, o que sugere que muitos indivíduos normais são portadores de níveis subclínicos de mutações no mtDNA potencialmente patogênicas.

Aplicações biotecnológicas

A presença de um genoma próprio e da maquinaria necessária para a tradução dos genes codificados pelo mtDNA possibilita que a mitocôndria desempenhe suas funções de maneira semiautônoma na célula. Tais funções impactam diretamente a célula, o tecido e o indivíduo, não somente por seu papel no metabolismo energético, como também por seu papel em outras formas de metabolismo, na regulação dos níveis de cálcio e sinalização em neurônios, no estabelecimento de grupamentos ferro-enxofre importantes para as reações de oxidação-redução, na morte celular programada e na imunidade inata. Como consequência, a genética mitocondrial pode ser considerada um importante componente genético, capaz de influenciar significativamente o fenótipo de um indivíduo ou de uma população. Assim, se, por um lado, a elevada taxa de variação a que o mtDNA está sujeito pode ser prejudicial para a função da mitocôndria, podendo resultar em patologias sérias, por outro, mutações no mtDNA podem eventualmente resultar em incrementos de função, que são passíveis de seleção.

Nesse sentido, o fato de a herança mitocondrial ser exclusivamente materna facilita a seleção de possíveis variantes por meio de linhagens matrilineares contendo as mais diversas variantes de mtDNA encontradas em uma mesma espécie. De fato, essa seleção é possível e já foi realizada em várias espécies, notadamente pelo cruzamento de animais de subespécies distintas, o que inclui camundongos (*Mus musculus domesticus* × *M. m. molocinus*), bovinos (*Bos taurus taurus* × *B. t. indicus*), suínos (animais de origem europeia × asiática), equídeos (equinos × muares), entre outros. Além disso, biotécnicas da reprodução podem ser utilizadas para transferir mtDNA entre animais de espécies diferentes ou até mesmo de gêneros diferentes. No entanto, apesar da facilidade de transferência de uma variante de mtDNA, não há muitos relatos que atestem sua eficácia sobre a produtividade ou adaptação em animais. Além disso, uma possível incompatibilidade núcleo-mitocôndria parece inviabilizar a transferência de mitocôndrias entre espécies geneticamente distantes.

Um modelo animal emblemático no Brasil, decorrente do modo como os bovinos foram introduzidos no país, possibilita ilustrar de maneira clara essa situação. Durante o início da colonização das Américas, os europeus trouxeram bovinos de raças europeias da subespécie *B. t. taurus* (taurinos). Esses animais foram criados e se multiplicaram ao longo dos anos, até que, no fim do século 19 até meados do século 20, foram introduzidos animais zebuínos de origem asiática (*B. t. indicus*). Essa introdução aconteceu majoritariamente mediante o uso de touros zebuínos, que foram cruzados com as fêmeas taurinas, resultando na formação da população de zebuínos do país (p. ex., Nelore, Gir e Guzerá). No entanto, o uso de fêmeas zebuínas importadas também contribuiu para a formação dessas raças. Depois de muitas gerações de cruzamentos, não é mais possível distinguir visualmente animais zebuínos produzidos por cruzamento absorvente ou mediante multiplicação dos descendentes das fêmeas zebuínas importadas. Entretanto, com um simples teste molecular é possível identificar tais linhagens com base na presença de mtDNA zebuíno ou taurino.

Diferentemente da incerteza que há a respeito da contribuição de variantes mitocondriais sobre características produtivas em animais, é inquestionável que mutações deletérias no mtDNA podem causar sérias patologias. Nesse sentido, algumas biotécnicas vêm sendo aplicadas com o objetivo de evitar a transmissão dessas patologias. Não há tratamentos efetivos para doenças mitocondriais; a única maneira de evitar o seu desenvolvimento é a prevenção da transmissão de mutações deletérias. Isso é possível com base em aconselhamento genético, em que se investigam, na mãe, possíveis mutações deletérias no mtDNA e a probabilidade de elas serem transmitidas para a progênie. O

aconselhamento genético pode ainda ser aplicado em associação com o diagnóstico pré-implantacional, por meio do qual embriões com níveis reduzidos da mutação são selecionados para implantação no útero materno.

Nos casos em que não é possível selecionar embriões livres da mutação ou com baixos níveis, algumas alternativas, ainda em desenvolvimento, seriam a eliminação seletiva de mutações no mtDNA por edição genômica e a substituição das mitocôndrias mutantes por selvagens. Esse último procedimento é possível por meio da transferência do DNA nuclear (pró-núcleos) desses embriões para embriões doados que tenham tido seu DNA nuclear removido e descartado (Figura 15.23). Os embriões receptores devem ser derivados de uma outra mãe, que não tenha mutações no mtDNA, originando, assim, embriões livres da mutação mitocondrial. No entanto, esse procedimento causa um problema ético e legal, já que a criança que se desenvolver a partir desse procedimento será formada por material genético derivado de três indivíduos diferentes (tripaternidade). Um casal deve contribuir com o DNA nuclear e uma terceira pessoa com o mtDNA. Além desse entrave, há riscos reais associados com esse procedimento, como a ocorrência de aneuploidia e anormalidades epigenéticas. Isso, no entanto, não impediu que esse procedimento fosse aprovado para teste em seres humanos no Reino Unido e ele se encontra, neste momento, em fase inicial de implementação.

Outra biotécnica associada à manipulação de mitocôndrias em oócitos e embriões se refere à terapia de substituição de mitocôndrias utilizada para tratar infertilidade em mulheres. Sabe-se que há uma forte associação entre envelhecimento e declínio da função mitocondrial, o que parece prejudicar a fertilidade de mulheres mais velhas, devido à importância das mitocôndrias para a qualidade do oócito. Além disso, mesmo nos casos não associados à idade, defeitos mitocondriais também são apontados como um importante fator de infertilidade em mulheres. Nesse sentido, a introdução de mitocôndrias “saudáveis”, derivadas de oócitos férteis, em oócitos de baixa qualidade foi proposta como uma possível forma de tratamento nesses casos. Essa tecnologia foi inicialmente implementada em seres humanos na década de 1990 e, de fato, comprovou-se bastante eficaz, resultando no nascimento de vários bebês. No entanto, alguns destes apresentaram anomalias cromossômicas, o que resultou no banimento da técnica em vários países. Mais recentemente, uma derivação dessa técnica passou a ser utilizada com o mesmo propósito. Nesse caso, em vez de as mitocôndrias introduzidas serem derivadas de oócitos doados por uma outra mulher, são derivadas de células precursoras do oócito oriundas da própria paciente. Células precursoras são células-tronco unipotentes com a função, ainda controversa, de renovarem o estoque de oócitos no ovário. As mitocôndrias presentes nessas células se assemelham às dos oócitos, exceto pelo fato de que continuam “saudáveis” com o avançar da idade, enquanto as mitocôndrias dos oócitos acumulam uma série de defeitos. Assim, as mitocôndrias podem ser isoladas das células precursoras a partir de uma biopsia ovariana e microinjetadas nos oócitos de baixa qualidade. Os resultados obtidos até o momento são surpreendentes e indicam um significativo ganho de fertilidade. Em comparação com a transferência de citoplasma utilizada no passado, esse procedimento tem a vantagem de introduzir mitocôndrias purificadas (destituídas de outros componentes do citoplasma) derivadas da própria paciente. No entanto, mais uma vez essa técnica está sendo implementada sem os devidos estudos que assegurem sua eficiência e segurança, motivo pelo qual ela é proibida na maioria dos países. Além disso, o uso desse tipo de tecnologia tem o potencial de resgatar o desenvolvimento de oócitos que estavam fadados a serem destruídos por conterem defeitos irreversíveis (p. ex., defeitos cromossômicos), o que pode levar ao nascimento de bebês com anormalidades cromossômicas, síndromes metabólicas, entre outras doenças.

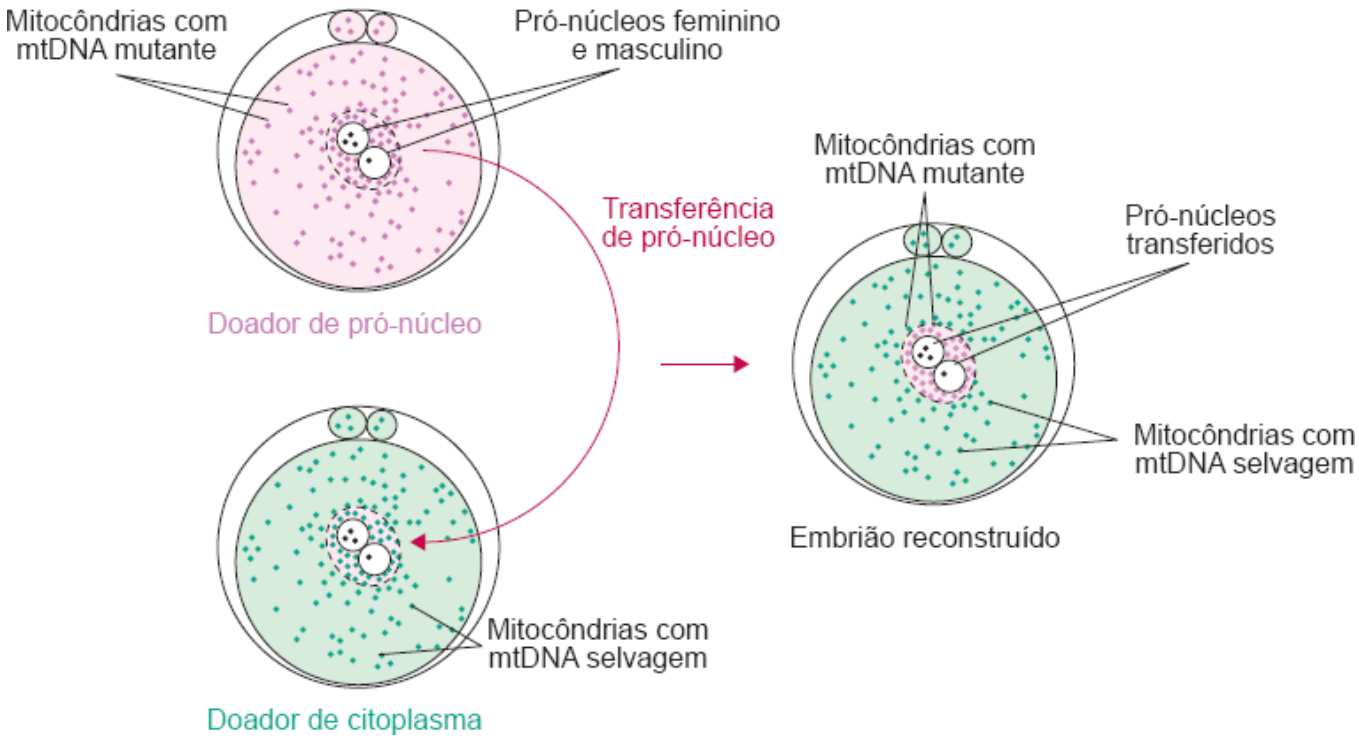


Figura 15.23 Representação esquemática do procedimento de transferência de pró-núcleos. Zigotos contendo mitocôndrias portadoras de mtDNA mutante (doador de pró-núcleo) devem ter seus dois pró-núcleos (masculino e feminino) transferidos por micromanipulação para zigotos que tiveram seus pró-núcleos previamente removidos (doador de citoplasma). Uma vez que no estágio de zigoto pró-nuclear as mitocôndrias se organizam ao redor dos pró-núcleos, algumas mitocôndrias mutantes acabam sendo transferidas juntamente com os pró-núcleos. Como resultado, o embrião reconstruído pode conter baixos níveis da mutação no mtDNA. Outras técnicas semelhantes a essa vêm sendo aprimoradas e têm a vantagem de resultarem na transferência de menores quantidades da mutação durante o procedimento de micromanipulação. Adaptado de Wolf *et al.*, 2015.

A importância das mitocôndrias para a qualidade dos oócitos também é revelada por uma série de trabalhos que indicam que, para serem férteis, os oócitos devem conter uma qualidade mínima de mtDNA. Paradoxalmente, os embriões mais viáveis são aqueles com menor conteúdo de mtDNA por célula. Essa aparente contradição é explicada pela observação de que, em condições normais, o mtDNA não é replicado durante o desenvolvimento pré-implantação. Como consequência, após cada divisão celular a quantidade de mtDNA por célula diminui pela metade. Nesse caso, se a quantidade de mtDNA herdada do oócito for baixa, o número de cópias remanescentes no embrião após várias divisões celulares será insuficiente para atender à demanda mitocondrial, havendo a necessidade de reativar prematuramente a replicação do mtDNA. Embora a reativação da replicação seja capaz de restituir o conteúdo de mtDNA, há uma correlação inversa entre quantidade de mtDNA no embrião e taxas de gestação e desenvolvimento a termo. Esses resultados são tão consistentes que tem sido proposta, em clínicas de reprodução assistida de humanos, a seleção dos embriões mais viáveis com base na quantidade de mtDNA.

Finalmente, existe um grande potencial biotecnológico na mitocôndria, que inclui uma eventual possibilidade de incrementar sua função como motor da célula ou introdução de novas funções. Com o advento de novas tecnologias de edição de genomas, como a CRISPR, a perspectiva de edição do mtDNA tornou-se uma possibilidade real e promissora para a área biotecnológica.

Genomas mitocondriais de algas e plantas

O tamanho grande do genoma mitocondrial de plantas, com aproximadamente 200 kb em *Brassica*, 367 kb em *Arabidopsis* e até 11 Mb em *Silene conica*, pode ser explicado com base em três fenômenos específicos: a mitocôndria de plantas codifica duas a três vezes mais genes; as frequentes recombinações e rearranjos de sequências levaram à formação de pseudogenes, os quais se mantêm no genoma; há integração de sequências de outros compartimentos celulares, tais como sequências do núcleo e do cloroplasto (ptDNA). Dentre o grande número de polipeptídeos envolvidos no funcionamento mitocondrial, em torno de 300, estima-se que somente 5% sejam codificados pelo genoma mitocondrial; os outros são codificados pelo núcleo, sintetizados no citoplasma e importados para a organela. Os polipeptídeos codificados por genes mitocondriais são sintetizados dentro da

mitocôndria, pois ela tem seu próprio sistema de síntese proteica. O mtDNA de angiospermas, entre elas o milho, que apresenta 700 kb, codifica menos genes que o genoma mitocondrial de mamíferos, apesar de o tamanho do genoma ser, no mínimo, 10 vezes maior em plantas (Figura 15.24). Comparativamente, o genoma mitocondrial da briófita *Marchantia polymorpha* tem 186 kb, e ambos codificam três RNA ribossômicos (26S, 5S e 18S), vários tRNA e 48 e 66 proteínas, respectivamente. A recombinação está ausente em *M. polymorpha*.

A estrutura do genoma mitocondrial, na maioria das plantas vasculares, é constituída por um conjunto de moléculas subgenômicas circulares, que originam-se a partir de um cromossomo principal circular (cromossomo *master*). Esse genoma seria representativo do genoma completo e, por meio de recombinações intra e intermoleculares via sequências repetidas diretas e invertidas, daria origem a moléculas subgenômicas ou formas isômeras do círculo principal. Os círculos que representam somente uma pequena fração do mtDNA em muitos eucariotos são formas opcionais da estrutura do genoma mitocondrial, sem importância para a função mitocondrial. Sequências repetidas em orientação direta e ativas em recombinações levariam à formação de várias moléculas subgenômicas redundantes e de diferentes tamanhos (Figura 15.25). Já as repetições de pequenos tamanhos efetuariam recombinações intragênicas, que poderiam resultar na formação de novas orfs (Mackenzie e McIntosh, 1999). A grande atividade recombinatória detectada no mtDNA de plantas resulta em grande complexidade estrutural do genoma. Talvez uma vantagem disso para a organela seja a manutenção da variação, o que resultaria na presença de diferentes subpopulações fenotípicas.

A alga *Chlamydomonas reinhardti* foi o primeiro organismo fotossintético a ter o DNA mitocondrial sequenciado, e seu genoma linear tem aproximadamente 16 kb. Até o momento, são mais de 48 genomas mitocondriais de algas e 178 plantas disponíveis. Embora os mtDNA das plantas sejam complexos e variáveis em tamanho, eles contêm essencialmente a mesma informação genética de outros genomas mitocondriais. O genoma não codifica para muitos genes e as enzimas necessárias para a replicação do DNA, transcrição e tradução são codificadas no núcleo e transportadas até a organela. Os produtos dos genes codificados no genoma mitocondrial participam predominantemente em processos oxidativos da respiração (genes *cox*, *cob*, *nad*), síntese de ATP (gene *atp*), biogênese do citocromo c (ccb) e genes ribossômicos envolvidos na síntese de proteínas (genes *rrn18, rrn26 e rrn5*), proteínas ribossômicas (genes *rps*, *rpl*) e alguns tRNA (genes *trn*). O genoma plastidial contém genes que codificam todos os tRNA necessários à síntese proteica, entretanto, não foi encontrado um grupo completo de tRNA nos genomas mitocondriais; dessa maneira, os tRNA necessários são supridos pelo núcleo. Existe uma tendência evolutiva de perdas de genes mitocondriais que codificam tRNA em plantas. A briófita *M. Polymorpha* contém 29 genes que codificam tRNA, enquanto o girassol tem 11 (sendo que somente 10 são expressados), e o milho somente 10.

O sequenciamento de DNA de organelas revelou que o genoma mitocondrial de plantas contém sequências de nucleotídeos homólogas a sequências no cloroplasto. São muitos os genomas mitocondriais que apresentam abundantes sequências adquiridas horizontalmente, o que não ocorre em genomas plastidiais. A comparação entre a sequência do genoma cloroplastidial e a do mitocondrial do arroz (*Oryza sativa*) demonstrou 16 sequências entre 32 pb a 6.8 kb em comprimento do genoma plastidial presente no genoma mitocondrial, que corresponde a 6% do genoma mitocondrial e 19% do genoma cloroplastidial. Essas transferências ocorreram em momentos diferentes, antes e depois da divergência do arroz e do milho (Figura 15.26). O mtDNA do arroz tem nove genes que codificam tRNA oriundos do genoma plastidial – seis deles são transcritos e corretamente processados e três são pseudogenes. Como sequências de DNA de cloroplastos e mitocôndrias também são encontradas no DNA nuclear, elas são consideradas “promíscuas”, por estarem presentes em mais de um dos três compartimentos celulares: genoma do núcleo, cloroplasto e mitocôndria. Sequências de DNA promíscuas são vestígios de extensiva transferência de DNA entre as organelas e o núcleo. Em mitocôndrias, cerca de 13% (aproximadamente 130 kb) do genoma mtDNA da abobrinha e da melancia é representado por sequências derivadas de plastídeos e do núcleo, sendo sequências não codificadoras. Outras angiospermas, como o genbibre e o *Plantago*, apresentam mosaicos de genes mitocondriais formados pela conversão de genes homólogos nativos e externos.

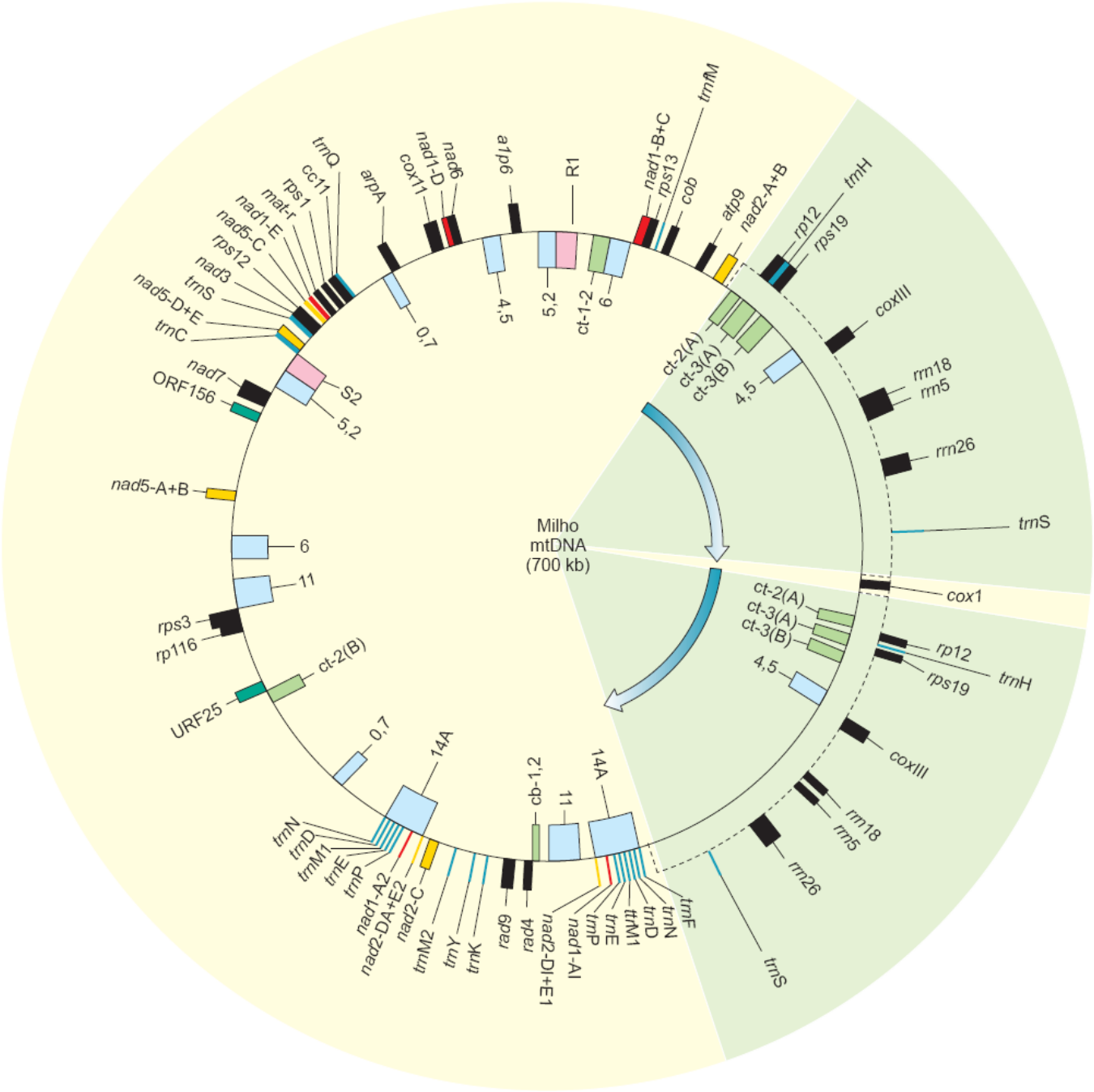


Figura 15.24 Mapa do genoma mitocondrial de milho (*Zea mays*). O genoma é baseado em uma molécula hipotética circular completa contendo todos os genes mitocondriais. Embora o genoma mitocondrial do milho seja maior que o genoma cloroplastidial (140.384 pb), ele contém menos genes. Muitas regiões repetidas invertidas e diretas (mostradas em azul, verde e retângulos lilás no círculo interno) foram identificadas nos eventos de recombinação e produzem moléculas de DNA circular subgenômicas.

Como a mitocôndria tem a mesma função em todos os eucariotos que a contém, o mtDNA de plantas superiores exibe características estruturais únicas, como grande tamanho, recombinações frequentes e inserções de DNA do núcleo e cloroplasto. Além disso, apresenta modos específicos de expressão com introns do grupo I ou grupo II (p. ex., "cis e trans *splicing*"), edição no RNA e vários sítios de iniciação da transcrição.

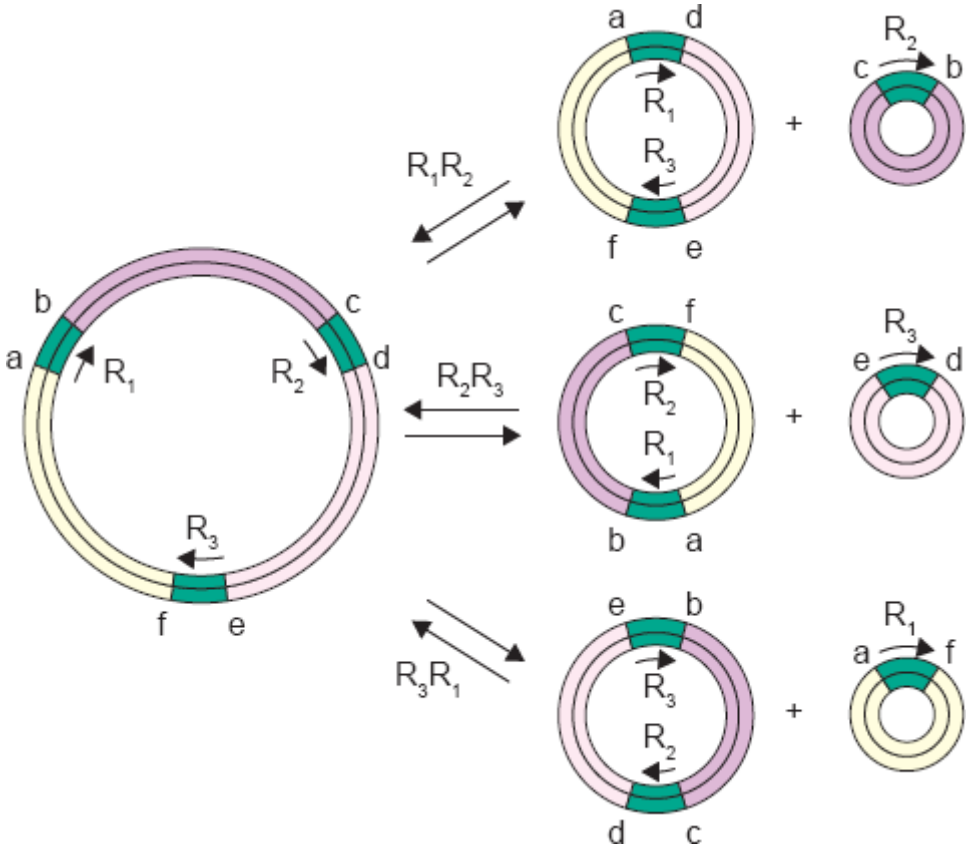


Figura 15.25 Recombinação homóloga ocorre nas sequências de DNA repetidas no genoma mitocondrial de plantas. Três repetições diretas presentes no genoma central podem produzir três eventos de recombinação distintos. Cada possível evento de recombinação produz um par de círculos subgenômicos diferente.

Transformação genética de mitocôndrias

Mitocôndrias e cloroplastos são organelas semiautônomas que têm seu próprio genoma e são consideradas passíveis de serem modificadas geneticamente. Entretanto, tem havido sucesso somente na transformação de cloroplastos, conforme apresentado anteriormente neste capítulo, e em mitocôndrias de dois organismos unicelulares, a levedura *Saccharomyces cerevisae* (Johnston *et al*., 1988) e a alga verde *Chlamydomonas reinhardtii* (Remacle *et al*., 2006). Apesar de muitos estudos, a transformação genética de mitocôndria de células animais e de plantas ainda não é possível.

A transformação mitocondrial da levedura *S. cerevisae* e da alga *Chlamydomonas* foi obtida por biolística (bombardeamento de micropartículas) com DNA linear ou plasmídeo. As células recipientes da linhagem de levedura utilizada não apresenta parte do *rho* ou todos os *rho* mitocondriais, replicam independentemente, não exigem sequências específicas na origem de replicação e apresentam ativa maquinaria de recombinação homóloga. O método de seleção por marcadores auxotróficos que correspondem a genes nucleares que codificam enzimas ativas na organela é atrativo para a transformação; um exemplo é o gene *arg*8, que, integrado no genoma mitocondrial, possibilita que mutantes nucleares *arg8* que não sintetizam arginina cresçam. A proteína ARG8 é sintetizada no citoplasma, transportada para a mitocôndria onde é expressa. Quando inserida no mtDNA, repõe a falta de arginina, fator que proporcionou o sucesso da transformação nessa espécie. Em *Chlamydomonas*, a transformação foi obtida em uma linhagem uniparental mutante *dum1*, que teve a região do telômero esquerdo e o gene *cob*1 deletados e vive em condições de escuro. Mutantes que perderam essa capacidade de viver em condições heterotróficas com suplementação de acetato como fonte de carbono no escuro, pela perda da via do citocromo, foram restauradas pela introdução de fragmentos de mtDNA de *C. reinhardtii* ou *C. smithii*. Essa tecnologia possibilitou introduzir mutações pontuais, que demonstraram que modificações no códon geralmente usado resultam em reduzida atividade de tradução mitocondrial, o que compromete o complexo respiratório e outras atividades. Também possibilitou estudos sobre a maquinaria de importação de tRNA na síntese de proteínas na organela.

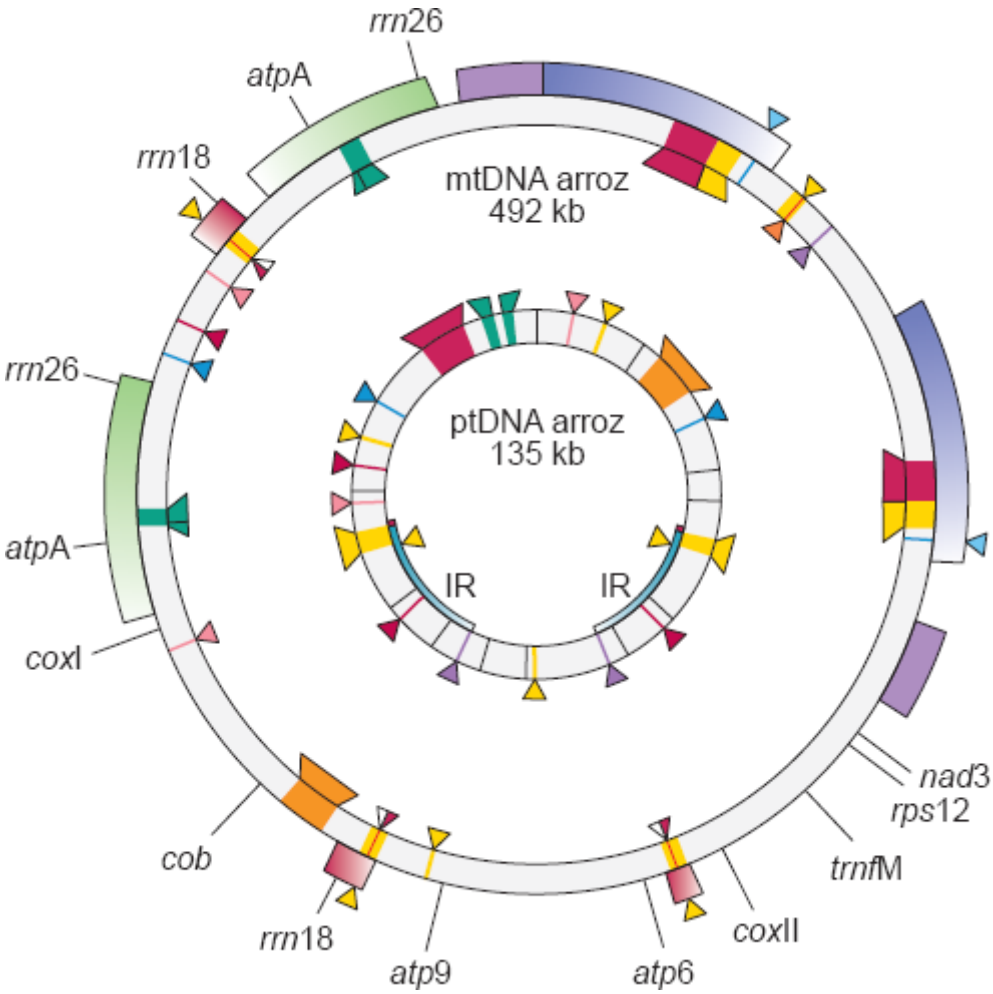


Figura 15.26 Comparação entre sequências dos genomas mitocondrial e plastidial do arroz (*Oryza sativa*) revela a presença de DNA "promíscuo". Os triângulos coloridos e os retângulos mostrados no genoma plastidial (círculo interno) representam sequências do DNA que foram transferidas para a mitocôndria e inseridas em locais diferentes no genoma mitocondrial (círculo externo). Observa-se que sequências de DNA mitocondrial não são encontradas no genoma plastidial, o que sugere que a transferência de DNA pode ser unidirecional ou o cloroplasto pode ter um mecanismo para se proteger da introdução de moléculas de DNA.

Em mamíferos, entre eles os humanos, mutações nas subunidades proteicas dos complexos da cadeia respiratória, tRNA e rRNA do mtDNA, são responsáveis por vários problemas clínicos relacionados a distúrbios do metabolismo energético e por doenças relacionadas ao envelhecimento, como a diabetes. Uma vez que o genoma mitocondrial está presente em múltiplas cópias, algumas mutações ocorrem em todas as cópias (homoplásmicas), enquanto outras ocorrem em algumas (heteroplásmicas), as quais nem sempre são detectadas. Também é possível que doenças mitocondriais sejam decorrentes de mutações em genes nucleares cujas proteínas estejam envolvidas no metabolismo mitocondrial. A transformação do genoma mitocondrial certamente seria uma importante ferramenta molecular para estudar as mutações que ocorrem na organela. Várias tentativas de introdução de DNA exógeno na forma de oligonucleotídeos, DNA fita dupla e plasmídeo em mitocôndrias de células de mamíferos foram realizadas por eletroporação, embora sem sucesso. Algum sucesso foi obtido utilizando a transferência de mitocôndrias isoladas de células somáticas de camundongo em células *in vitro* por microinjeção. O uso de vírus adenoassociado para expressar genes mitocondriais demonstrou a expressão e a tradução do gene presente em uma construção gênica com uma mutação responsável pela perda de visão, mas não foi detectada a integração no genoma mitocondrial (Larosa e Remacle, 2013).

Em plantas, disfunções resultam em fenótipos visíveis de machoesterilidade citoplasmática, como observado em fumo (*Nicotiana sylvestris*), e de folhas listradas em genótipos de milho. Apesar de esses genótipos serem bons candidatos para estudos de transformação genética mitocondrial, até o momento não houve sucesso. Os possíveis motivos estão principalmente na instabilidade do genoma e na falta de um marcador seletivo apropriado que possa ser mantido durante o desenvolvimento da planta adulta após a formação do zigoto e na formação de subgenomas. Como visto, essas estruturas ocorrem devido à alta frequência de recombinações inter e intramoleculares de regiões repetidas, formando genomas subdivididos, cada um com somente uma porção da informação genética, com

variação no número de cópias que resulta em alta complexidade heteroplásmica. Estudos sugerem que o silenciamento de genes que controlam essas recombinações, como o gene da proteína de reparo de DNA MSH1, poderia contribuir para a transformação de mitocôndrias. Algumas tecnologias, como as descritas anteriormente, em genomas de mamíferos, com isolamento de mitocôndrias, análises microscópicas, sequenciamento de genomas e biolística, estão sendo introduzidas para tornar possível a transformação mitocondrial de plantas e humanos.

Conclusões e perspectivas

O fato de cloroplastos e mitocôndrias terem se tornado organelas celulares, advindas da endossimbiose com proteobactérias de vida livre, teve várias implicações sobre a biologia, a evolução e a diversidade dos organismos atuais. Estudos sobre organelas têm sido bastante beneficiados pelo advento da tecnologia de sequenciamento de DNA de alta produtividade, que já disponibilizou alguns milhares de genomas, propiciando marcante ampliação do conhecimento da genética, da genômica e de aspectos biotecnológicos de organelas. Alguns dos principais desafios, perspectivas e avanços no conhecimento estão apresentados a seguir.

O genoma das organelas sofreu uma dramática redução de tamanho em comparação com seus ancestrais, devido à massiva deleção e transferência de genes para o genoma nuclear durante o processo de integração do endossimbionte no metabolismo da célula hospedeira. Assim, as organelas são dependentes dos genes codificados no núcleo e transportados para elas para a preservação de vias metabólicas, mecanismos de reparo e funções relacionadas à expressão gênica.

O genoma de cloroplastos e mitocôndrias apresenta muitas características em comum: geralmente são circulares, poliploides e localizam-se em estruturas chamadas nucleoides, que ficam ancorados nas membranas internas das organelas. A estrutura organizacional de muitos genes em operons e o mecanismo de processamento são semelhantes ao dos organismos procariotos que originaram as organelas.

A maior parte dos organismos tem herança maternal de transmissão das organelas. Apesar de raro, é possível herança paternal ou biparental de cloroplasto em algumas espécies de plantas. As algas *Chlamydomonas* apresentam herança maternal de maneira diferenciada, pois têm um único cloroplasto. Em mitocôndrias, a herança é exclusivamente materna.

O genoma cloroplastidial de plantas superiores codifica entre 100 e 120 genes relacionados principalmente ao sistema genético que codificam componentes proteicos do mecanismo de expressão gênica, como subunidades da RNA polimerase, rRNA e tRNA, além de genes relacionados à fotossíntese e orfs conservadas com funções desconhecidas.

A expressão gênica dos genes plastidiais é altamente corregulada por genes nucleares cuja atividade ocorre na organela. Desse modo, é essencial a interação entre os compartimentos celulares para coordenar as atividades metabólicas e as respostas da planta durante o crescimento e resposta às mudanças ambientais. Na replicação do DNA plastidial, todas as enzimas envolvidas são codificadas pelo genoma nuclear, e a síntese do RNA ocorre por duas RNA polimerases, a PEP (codificada por genes plastidiais) e a NEP (codificada no núcleo). O controle da expressão gênica ocorre ao nível transcricional e pós-transcricional. Alguns genes cloroplastidiais apresentam introns do tipo I e do tipo II e ocorre edição do mRNA.

No processamento pós-transcricional, a edição do mRNA ocorre em cloroplastos e mitocôndrias de plantas, fungos e animais. Geralmente, a edição ocorre pela substituição de uma citosina (C) por uma uracila (U) e resulta em alteração no transcrito. Entre as alterações mais dramáticas está a criação de um códon de iniciação ou de terminação.

O genoma plastidial de algas é muito variável quanto ao tamanho e aos genes que codifica. Algumas algas, como as Cryptofitas, são resultado de processo de endossimbiose secundária e têm três a quatro membranas e o nucleomorfo, estrutura remanescente, mas reduzida, do genoma nuclear do endossimbionte original com genes que codificam proteínas somente para a manutenção desse compartimento na célula.

Os apicoplastos são vestígios de organelas plastidiais adquiridos por endossimbiose secundária, apresentam quatro membranas e não têm nucleomorfo e genes fotossintetizantes. Essas organelas são responsáveis por importantes doenças, como malária, toxoplasmoses, babesioses e coccidioses, sendo importante alvo de estudos para o desenvolvimento de medicamentos para controle dessas patologias.

O genoma mitocondrial apresenta grande variação de tamanho dependendo dos organismos. Mitocôndrias em células animais têm os menores genomas (10 a 20 kb), e as células de plantas de algas apresentam genomas grandes,

entre 16 e 2400 kb, podendo chegar a 11 Mpb (*Silene conica*). Em plantas, a maioria do genoma mitocondrial é composta de sequências não codificadoras e apresenta moléculas subgenômicas.

O genoma mitocondrial de mamíferos codifica genes essenciais para os componentes da cadeia transportadora de elétrons, 2 rRNA e 22 tRNA e 13 genes codificantes de mRNA. Os genes mitocondriais não apresentam introns, de modo que o processamento dos mRNA policistrônicos é relativamente simples, com a participação de poucas enzimas.

O DNA mitocondrial de animais é utilizado em estudos evolutivos como marcador de diversidade molecular, em razão de a taxa de substituição de nucleotídeos durante a evolução ter sido maior que no genoma nuclear, da ausência de introns, de pequenas regiões intergênicas e da conservação do conteúdo gênico entre as espécies. O DNA mitocondrial tem sido utilizado como código de barras molecular na análise de biodiversidade e identificação das espécies.

Algumas doenças em humanos e animais são decorrentes de mutações pontuais, deleções ou rearranjos no genoma mitocondrial. A grande maioria das mutações ocorre em genes de tRNA, mas algumas mutações pontuais são responsáveis por doenças neurológicas. Em plantas, a machoesterilidade está relacionada a alteração no genoma mitocondrial.

Assim, conforme apresentado neste capítulo, a disponibilidade das sequências dos genomas de organelas, associada à tecnologia de transformação genética, trouxe avanços relevantes no conhecimento da biologia molecular de organelas e abre perspectivas de desenvolvimento de produtos biotecnológicos. Entretanto, ainda são muitos os desafios para os cientistas da área, pela grande diversidade dos organismos e pelas dificuldades em manipular geneticamente mitocôndrias de plantas e cloroplastos de cereais.

Bibliografia

Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. Biologia molecular da célula. 5. ed. Porto Alegre: Artmed; 2010.

Benne R, Van Den Burg J, Brakenhoff JP, Sloof P, Van Boom JH, Tromp MC. Major transcript of the frameshift coxII from trypanosome mitochrondria contains four nucleotides that are not encoded in the DNA. Cell, Washington. 1986;46:819-26.

Bock R. Structure, function, and inheritance of plastid genomes. In: Bock R. Topics in Current Genetics. v. 19. Berlin: Springer-Verlag; 2007. p. 29-63.

Bock R, Knoop V, editors. Genomics of chloroplasts and mitochondria. Springer Dordrecht Heidelberg; 2012.

Bock R, Timmis JN. Reconstructing evolution: gene transfer from plastids to the nucleus. BioEssays. 2008:30:556-66.

Bogorad L, Vasil IK. The molecular biology of plastids. San Diego-CA, USA: Academic Press; 1991.

Boynton JE, Gilham NW, Harris EH, Hosler JP, Johnson AM, Jones AR, et al. Chloroplast transformation in chlamydomonas with high velocity microprojectiles. Science. 1988;240:1534-8.

Buchanan BB, Gruissem W, Jones RL. Biochemistry & Molecular Biology of Plants. 2. ed. West Sussex, UK: American Society of Plant Biologists; 2015.

Covello PS, Gray MW. RNA editing in plant mitochondria. Nature. 1989;341:662-6.

Falkenberg M, Larsson N-G, Gustafsson CM. DNA replication and transcription in mammalian mitochondria. Annu Rev Biochem. 2007;76:679-99.

Greiner S, Sobanski J, Bock R. Why are most organelle genomes transmitted maternally? Bioessays, UK. 2014;37:80-94.

Hagemann R, Schröder MB. The cytological basis of the plastid inheritance in angiosperms. Protoplasma. 1989;152(2):57-64.

Hagemann R. The foundation of extranuclear inheritance: plastid and mitochondrial genetics. Molecular Genetics and Genomics. 2010;283(3):199-209.

Herrmann RG. Cell Organelles. Springer-Verlag Wien; 1992.

Hiratsuka J, Shimada H, Whittier R, Ishibashi T, Sakamoto M, Mori M, et al. The complete sequence of the rice (Oryza sativa) Chloroplast genome: Intermolecular recombination between distinct tRNA genes accounts for a major plastid DNA inversion during the evolution of cereals. Mol Gen Genet. 1989;217:185-94.

Hoch B, Maier RM, Appel K, Igloi GL, Kossel H. Editing of a chloroplast mRNA by creation of an initiation codon. Nature. 1991;353:178-80.

Johnston SA, Anziano PQ, Shark K, Sanford JC, Butow RA. Mitochondrial transformation in yeast by bombardment with microprojectiles. Science. 1988;240(4850):1538-41.

Larosa V, Remacle C. Transformation of the mitochondrial genome. International Journal Developmental Biology. 2013;57:659-65.

Mackenzie S, McIntosh L. Higher plant mitochondria. Plant All. 1999; 11(4):571-86.

Maliga P. Chloroplast Biotechnology: Methods and Protocols. New York: Springer; 2014.

Remacle C, Cardol P, Coosemans N, Gaisne M, Bonnefoy N. High-efficiency biolistic transformation of Chlamydomonas mitochondria can be used to insert mutations in complex I genes. Proceedings of the National Academy of Sciences of the United States of America. 2006;103(12):4771-6.

Schmitz-Linneweber C, Small I. Pentatricopeptide repeat proteins: a socket set for organelle gene expression. Trends Plant Sci. 2008;13(12):663-70.

Schon EA, DiMauro S, Hirano M. Human mitochondrial DNA: roles of inherited and somatic mutations. Nat Rev Genet. 2012;13(12):878-90.

Shi C, Wang S, Xia E-H, Jiang J-J, Zeng F-C, Gao L-Z. Full transcrition of the chloroplast genome in photosynthetic eukaryotes. Scientific Reports. 2016;6:30135.

Shinozaki K, Ohme M, Tanaka M, Wakasugi T, Hayashida N, Matsubayashi T, et al. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. EMBO Journal, Heidelberg. 1986;5(9):2043-9.

Svab Z, Harper EC, Jones JD, Maliga P. Aminoglycoside-300-adenyltransferase confers resistance to spectinomycin and streptomycin in Nicotiana tabacum. Plant Molecular Biology. 1990;14:197-205.

Taanman JW. The mitochondrial genome: structure, transcription, translation and replication. Biochim Biophys Acta. 1999;9(1410-2):103-23.

Van Dooren GG, Striepen B. The algal Past and Parasite Present of the Apicoplast. Ann Rev Microbiol. 2013;67:271-89.

Wolf DP, Mitalipov N, Mitalipov S. Mitochondrial replacement therapy in reproductive medicine. Trends Mol Med. 2015;21(2):68-76.

16 Mundo dos RNA Não Codificantes

Franceli Rodrigues Kulcheski, Júlio César de Lima, Guilherme Loss de Morais e Rogerio Margis

Nos últimos anos, evidenciou-se a ampla participação dos pequenos RNA não codificadores (ncRNA) nos mais diversos mecanismos regulatórios das células eucarióticas. Novas classes de ncRNA, bem como seus processos de biogêneses e modos de ação, têm sido continuamente descritos, e novas propriedades de ncRNA já estabelecidas ainda estão sendo descobertas. Essas moléculas apresentam ampla gama de regulação da expressão gênica, podendo atuar a níveis de transcrição, processamento de RNA e tradução. Neste capítulo, é apresentada uma compilação das principais classes de pequenos ncRNA descritos na literatura, abordando sua biogênese e meios de atuação. Adicionalmente, serão discutidas algumas aplicações biotecnológicas, bem como metodologias de identificação e caracterização dessas moléculas.

Introdução

Algumas teorias evolutivas têm sugerido que o início da vida na Terra teve como fonte moléculas de RNA, as quais carregavam não apenas informação biológica, mas também uma capacidade autorreplicativa. Ao longo do processo evolutivo, a função de portador da informação genética foi transferida para uma molécula mais estável – o DNA; enquanto as funções catalíticas foram substituídas por proteínas versáteis. Assim, acredita-se que o RNA foi o primeiro elo entre genes e proteínas. No entanto, não menos importante que o DNA, compreende-se, atualmente, que os RNA são cruciais na biologia da célula. Além da descoberta dos processamentos alternativos e de genes sobrepostos que causou retificação na relação "um gene, uma proteína", o conceito de gene mudou bastante a partir da descoberta de que, no genoma, existem sequências que são transcritas, mas não traduzidas. Certamente, os genes que dão origem aos pequenos ncRNA são motivo de intensas discussões na comunidade científica.

Há aproximadamente duas décadas, um tipo de regulação da expressão gênica, mediado por pequenos RNA que ocasionam diminuição nos níveis de RNA-alvo, foi denominado silenciamento gênico. As primeiras evidências para o mecanismo-chave do silenciamento gênico foram reveladas no trabalho de Napoli *et al*., com plantas de petúnia transgênicas, em meados da década de 1990 (ver boxe "Primeira evidência de iRNA | Cossupressão"). As plantas apresentavam diferentes fenótipos quanto à cor das pétalas, variando de brancas (ausência de pigmentos) até pétalas com padrões variados de coloração roxa. Esses fenótipos foram ocasionados pelos diferentes níveis de silenciamento gênico do transgene codificante para a enzima chalcona sintase, envolvida na pigmentação das pétalas. Esse fenômeno biológico foi inicialmente denominado cossupressão. Até então, muito pouco se sabia sobre a presença de pequenos RNA envolvidos na regulação da expressão gênica.

Primeira evidência de iRNA | Cossupressão

Em 1990, Napoli *et al*. evidenciaram pela primeria vez o fenômeno que mais tarde seria conhecido como RNA de interferência (iRNA). Naquela época, eles tentavam determinar se a chalcona sintase (CHS), uma enzima-chave na biossíntese de antocianinas, era taxa-limitante na síntese desse flavonoide. O acúmulo de antocianinas em pétalas de petúnia é responsável pela coloração violeta ou roxa nessas flores. Visando a um aumento na intensidade da cor roxa, eles superexpressaram a CHS em petúnias, o que resultou inesperadamente em 42% das plantas transgênicas com flores totalmente brancas e/ou variegadas, com regiões roxas e brancas. A análise de RNA presente nas pétalas das flores brancas mostrou que,

embora o tempo de desenvolvimento da expressão de mRNA do gene endógeno CHS não tenha sido alterada, o nível do mRNA produzido por esse gene foi reduzido 50 vezes em relação aos níveis desse transcrito em plantas não transformadas. Os resultados desses experimentos mostraram que a introdução do gene quimérico bloqueava a biossíntese de antocianina por meio da inibição tanto do gene endógeno quanto da cópia introduzida da CHS. Entretanto, a reversão somática das flores brancas para o fenótipo parental com flores violetas foi associada a um aumento coordenado dos níveis de mRNA produzidos por ambos os genes endógenos e introduzidos da CHS. A partir da observação das flores brancas alteradas, Napoli *et al*. puderam concluir que a expressão de ambos os genes (endógeno e quimérico) foi suprimida de forma coordenada, indicando que a expressão do gene introduzido da CHS não era suficiente por si só para a supressão de níveis dos transcritos endógenos. Dessa maneira, o fenômeno descoberto foi denominado de "cossupressão", e posteriormente o termo *post-transcriptional gene silencing* (PTGS) foi cunhado para designar o mecanismo de silenciamento gênico observado em plantas.

Em 1998, Andrew Fire *et al*., em outro trabalho pioneiro envolvendo estudos com o nematoide modelo *Caenorhabditis elegans*, revelaram que RNA de dupla fita (dsRNA) transgênicos levam à potencialização do silenciamento gênico, em contrapartida aos RNA de fita simples antissenso. O resultado mostrou uma inibição da expressão gênica maior quando os animais eram injetados com os dsRNA. Uma série de animais mutantes contendo diferentes construções gênicas foi gerada para se demonstrar que, de fato, os dsRNA geravam um sinal muito maior no silenciamento gênico em relação aos RNA antissenso. Além da potencialização, os autores desse trabalho demonstraram que o silenciamento gênico era direcionado e tinha alta especificidade, não alterando a expressão de outros genes. Outra característica importante foi a de que as gerações posteriores dos animais utilizados nos experimentos não perderam a capacidade de inibir a transcrição dos genes-alvo. Ou seja, os fenótipos mutantes silenciados permaneceram ainda nas gerações filhas, demonstrando a estabilidade não apenas da transmissão genética, mas também da manutenção do mecanismo de inibição da expressão gênica. Pela primeira vez, demonstrou-se, em detalhes, como o mecanismo de silenciamento poderia afetar a regulação positiva da expressão dos genes. Em 2006, os coautores Andrew Fire e Craig Mello foram laureados com o Nobel de Fisiologia ou Medicina pelo seu feito científico (ver boxe "Andrew Fire e Craig Mello").

Andrew Fire e Craig Mello

Em 2006, Andrew Fire e Craig Mello receberam o Prêmio Nobel de Fisiologia ou Medicina pela descoberta do mecanismo de RNA de interferência e silenciamento gênico por RNA dupla fita. O trabalho desenvolvido juntamente com seus colegas SiQun Xu, Mary Montgomery, Stephen Kostas e Sam Driver já havia sido publicado em 1998 na revista *Nature*. A pesquisa de Fire e Mello, realizada no Instituto Carnegie para Ciência e na Escola de Medicina da Universidade de Massachusetts, mostrou que, de fato, o RNA desempenha um papel fundamental na regulação dos genes. Utilizando o verme *Caenorhabditis elegans*, eles comprovaram a hipótese de que não era um RNA fita simples (ssRNA) que desencadeava o silenciamento gênico, mas sim um RNA dupla fita (dsRNA). O gene-alvo escolhido para demonstrar essa teoria foi o *unc-22*, que codifica uma proteína de miofilamento. Milhares de transcritos de *unc-22* estão presentes em cada célula muscular estriada de *C. elegans*. Existe uma correlação semiquantitativa entre a atividade da proteína UNC-22 e o fenótipo do verme. Quando ocorre a diminuição na atividade de UNC-22, o verme apresenta espasmos graves, enquanto a perda completa da função dessa proteína resulta em defeitos estruturais musculares e mobilidade prejudicada.

Fire e Mello purificaram extensivamente RNA fita simples (ssRNA) senso e antissenso, e então compararam seus efeitos com o RNA dupla fita (dsRNA) sobre o gene *unc-22*. Os ssRNA senso e antissenso foram consistentemente 10 e 100 vezes, respectivamente, menos eficientes do que o dsRNA no silenciamento do mesmo mRNA-alvo. De fato, o ssRNA foi eficiente apenas quando a fita senso era injetada nos animais, seguida pela injeção da fita antissenso ou vice-versa, sugerindo a ocorrência da hibridização dos ssRNA formando dsRNA *in vivo*. Investigações subsequentes mostraram que a injeção de um dsRNA específico pode silenciar qualquer gene, e que a presença de poucas moléculas desse RNA é suficiente para fazê-lo. Consequentemente, a área de iRNA explodiu nos últimos anos como uma nova técnica poderosa para o estudo da função gênica e como base para terapias de novas doenças.

Esses trabalhos começaram a revelar parte do "quebra-cabeça molecular", que a célula naturalmente utiliza para regular a expressão gênica por meio dos pequenos ncRNA. A partir de então, inúmeros grupos de pesquisa, espalhados pelo mundo, investigam quais os mecanismos que estão por trás dessa alternativa de regular a expressão de um gene.

Além dos RNA mensageiros ou codificantes (mRNA) que são utilizados como molde para a síntese de proteínas, já foram identificadas e caracterizadas diversas classes de ncRNA. Estes não são utilizados como molde para a síntese de proteínas, mas têm papel fundamental na biologia molecular estrutural da célula (RNA estruturais), como também papéis regulatórios (RNA reguladores) (Figura 16.1). Nas últimas décadas, os RNA vêm ocupando lugar de destaque na ciência, devido à percepção de que desempenham papéis diversos e fundamentais no ambiente celular e no contexto funcional de qualquer organismo onde estejam presentes. Este capítulo versará sobre os pequenos ncRNA (Figura 16.1), que englobam desde os RNA que desempenham papel no desenvolvimento do organismo, como também no âmbito da regulação da expressão gênica em diferentes situações biológicas.

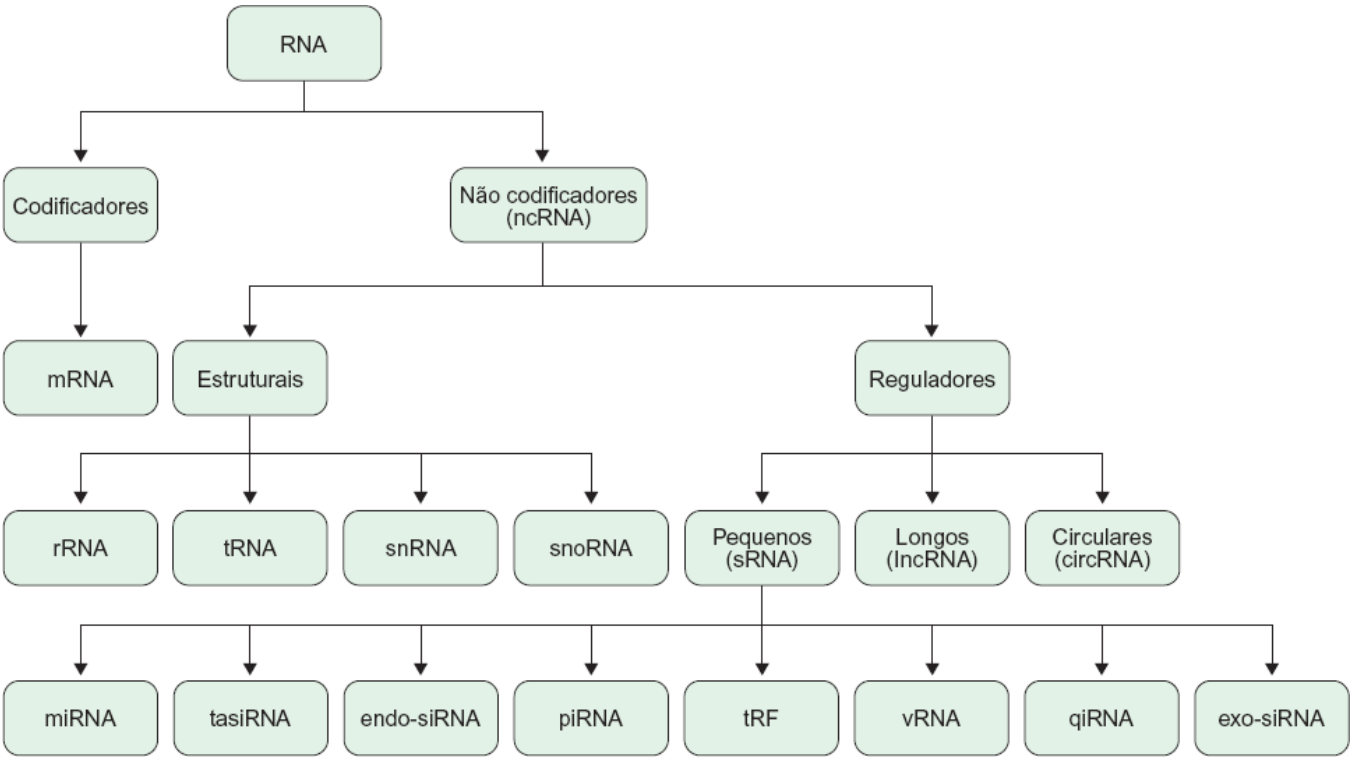


Figura 16.1 Diferentes classes de RNA detectadas em eucariotos.

Pequenos RNA não codificantes envolvidos na regulação da expressão gênica

microRNA

Os microRNA (miRNA) constituem a principal classe de pequenos RNA envolvidos na regulação da expressão gênica, atuando em uma série de processos biológicos, tais como crescimento, desenvolvimento e resposta a diversos estresses. O primeiro miRNA, *lin-4* (do inglês, *lineage-deficient-4*), foi descoberto em 1993 e identificado como regulador pós-transcricional do gene *lin-14*, que está envolvido no controle do desenvolvimento do nematoda *C. elegans*. Em plantas, a primeira identificação de miRNA foi feita em *Arabidopsis thaliana*, e descrita em 2002 por dois grupos distintos. Atualmente, os miRNA já foram descritos em animais, plantas, fungos e até mesmo a partir de sequências virais; todas as sequências já identificadas de miRNA estão disponíveis em um banco de dados público, denominado miRBase. O aumento na identificação de miRNA tem possibilitado demonstrar suas relações com diversos processos celulares. Inúmeros miRNA identificados em plantas e animais, organismos nos quais os miRNA tiveram melhor estudo até o momento, têm sido associados a variados processos biológicos. Em plantas, por exemplo, muitos miRNA foram detectados em processos relacionados com o desenvolvimento, sinalização hormonal, floração, determinação sexual, bem como a respostas a estresses bióticos e abióticos. Já em animais, além de seu envolvimento na determinação de estágios de desenvolvimento, os miRNA têm sido associados a inúmeras doenças, tais como diferentes tipos de câncer.

Biogênese dos miRNA

Os miRNA são pequenos RNA com 20 a 24 nucleotídeos (nt) de tamanho. Assim como os genes codificantes de proteínas, os miRNA têm sido organizados em famílias, principalmente de acordo com a similaridade da sequência madura. Essas famílias podem ser simples ou complexas, compreendendo de um a poucos membros até dezenas de

miRNA. Os ncRNA derivam de genes de miRNA ou de introns produzidos do processamento de genes codificantes transcritos inicialmente pela RNA polimerase II (RNA Pol II).

Em plantas, genes de miRNA são transcritos por uma enzima, a RNA polimerase do tipo II (RNA Pol II); no entanto, alguns miRNA podem ser transcritos pela RNA Pol III. Os transcritos iniciais de miRNA são chamados de miRNA primários (pri-miRNA) (Figura 16.2). RNA Pol II gera pri-miRNA capeados e poliadenilados, tanto em plantas quanto em animais. Após a transcrição dos pri-miRNA, a proteína Dawdle (DDL) liga-se a eles, protegendo-os contra degradação. O pri-miRNA irá adquirir uma estrutura em formato de grampo (*hairpin structure*) imperfeita, a qual será processada em um pré-miRNA (precursor do miRNA), também conhecido como estrutura *stem-loop*. Em plantas, ambos – pri e pré-miRNA – são processados no núcleo por uma enzima do tipo RNase III, denominada Dicer-Like 1 (DCL1).

O processamento do pri-miRNA em pré-miRNA ocorre em um centro de processamento nuclear conhecido como corpo-D (do inglês, *D-body*, ou SmD3/SmB-*body*). Embora não exista um tamanho padrão para o comprimento efetivo dos pri-miRNA, sabe-se que todos podem dobrar-se e originar uma estrutura secundária em formato de grampo. Conter estruturas *hairpin* ou "grampo" é uma importante característica de miRNA. Em plantas, os pré-miRNA, ainda no núcleo, são clivados pela DCL1 e pela proteína HYL1 (*hyponastic leaves* 1), originando uma dupla miRNA:miRNA* (miRNA e miRNA estrela). As duplas miRNA:miRNA* recentemente processadas contêm 2 nt não pareados na região terminal 3' de cada fita, podendo ser facilmente degradadas por uma classe de exonucleases, conhecida por SDN (*small RNA degradading nuclease*). Para estabilizar a dupla fita miRNA:miRNA*, uma metiltransferase de pequenos RNA, HEN1 (*Hua Enhancer*), imediatamente metila os nt no terminal 3' de cada fita, evitando assim sua uridilação e subsequente direcionamento para a degradação. Em mutantes *hen1* observa-se redução na quantidade de miRNA, indicando o papel de HEN1 na proteção de miRNA e evitando sua degradação. O *duplex* miRNA/miRNA* é transportado do núcleo para o citoplasma pela proteína de membrana HASTY. No citoplasma, a dupla miRNA/miRNA* é separada, sendo uma das fitas do miRNA incorporada ao complexo de silenciamento induzido por RNA ou RISC (*RNA-induced silencing complex*), formando, assim, o complexo miR-RISC (Figura 16.2).

Esse complexo contém a proteína AGO1 (*Argonaute* 1), que irá clivar o mRNA-alvo na região mediada da dupla mRNA-miRNA. A outra fita, designada como miRNA*, será degradada. No entanto, alguns estudos têm demonstrado que miRNA* também podem funcionar como uma sequência efetiva de miRNA e controlar a expressão de genes-alvo específicos.

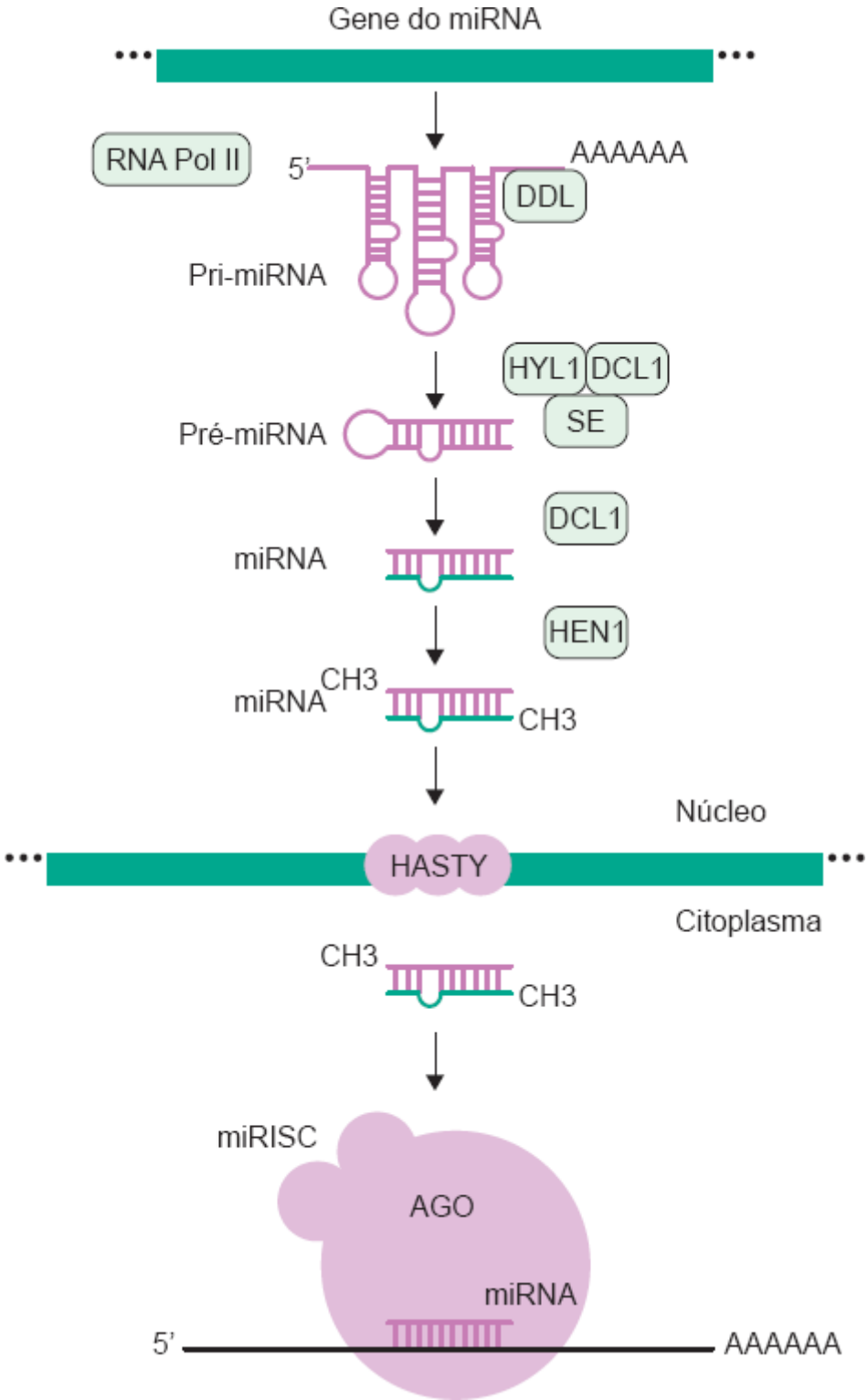


Figura 16.2 Biogênese dos miRNA em plantas. Os pri-miRNA são transcritos por uma RNA Pol II, formando uma estrutura em formato de grampo poliadenilada em sua extremidade 3'. A essa estrutura liga-se a proteína Dawdle (DDL), responsável por estabilizar o pri-miRNA até a sua conversão em pré-miRNA. Este processo conta com a ação combinada das proteínas: Serrate (SE), Hyponastic Leaves 1 (HYL1) e Dicer-like 1 (DCL1). O pré-miRNA é então clivado novamente por uma DCL1, gerando um *duplex* miRNA/miRNA* com extremidades 3' salientes, que será exportado para o citoplasma pela proteína de membrana HASTY. Ambas as sequências miRNA/miRNA* serão metiladas por uma metilase HEN1, protegendo as sequências da degradação via nucleases degradadoras de pequenos RNA (SDN). A fita guia de miRNA é então incorporada a proteínas AGO componentes do complexo RISC, formando o complexo miR-RISC, agindo no silenciamento gênico.

Em animais, os genes de miRNA também são transcritos por uma RNA Pol II em um pri-miRNA capeado no terminal 5' e poliadenilado na região 3' (Figura 16.3). Ainda no núcleo, o pri-miRNA é clivado por um complexo proteico formado por uma ribonuclease RNase III denominada Drosha e sua proteína parceira, designada Pasha em invertebrados, e DGCR8 em vertebrados. A clivagem resulta em um pré-miRNA, que pode variar entre 70 e 300 nt de comprimento, o qual forma uma estrutura em formato de grampo. O pré-miRNA é então exportado para o

citoplasma pela Exportina-5, onde será clivado pela nuclease Dicer. Tal clivagem resulta em um *duplex* miRNA:miRNA* com aproximadamente 22 nt de comprimento. Diferentemente dos miRNA de plantas, em animais, os miRNA maduros não são metilados na extremidade 3'. Ambas as fitas do *duplex* de miRNA podem atuar com fita guia, mas geralmente apenas uma delas é incorporada ao complexo RISC, onde o miRNA e seu mRNA-alvo interagem.

Modo de ação e papel biológico dos miRNA

Os miRNA foram descobertos há aproximadamente duas décadas. Desde então, esses pequenos RNA têm sido apontados como alguns dos mais importantes dentre os reguladores gênicos. Investigações indicam que miRNA controlam a expressão gênica de aproximadamente 30% dos genes codificantes de proteínas em humanos. Em plantas, ainda não foram realizados estudos determinando esse percentual, mas acredita-se que eles estejam envolvidos em uma parcela significativa da regulação gênica vegetal.

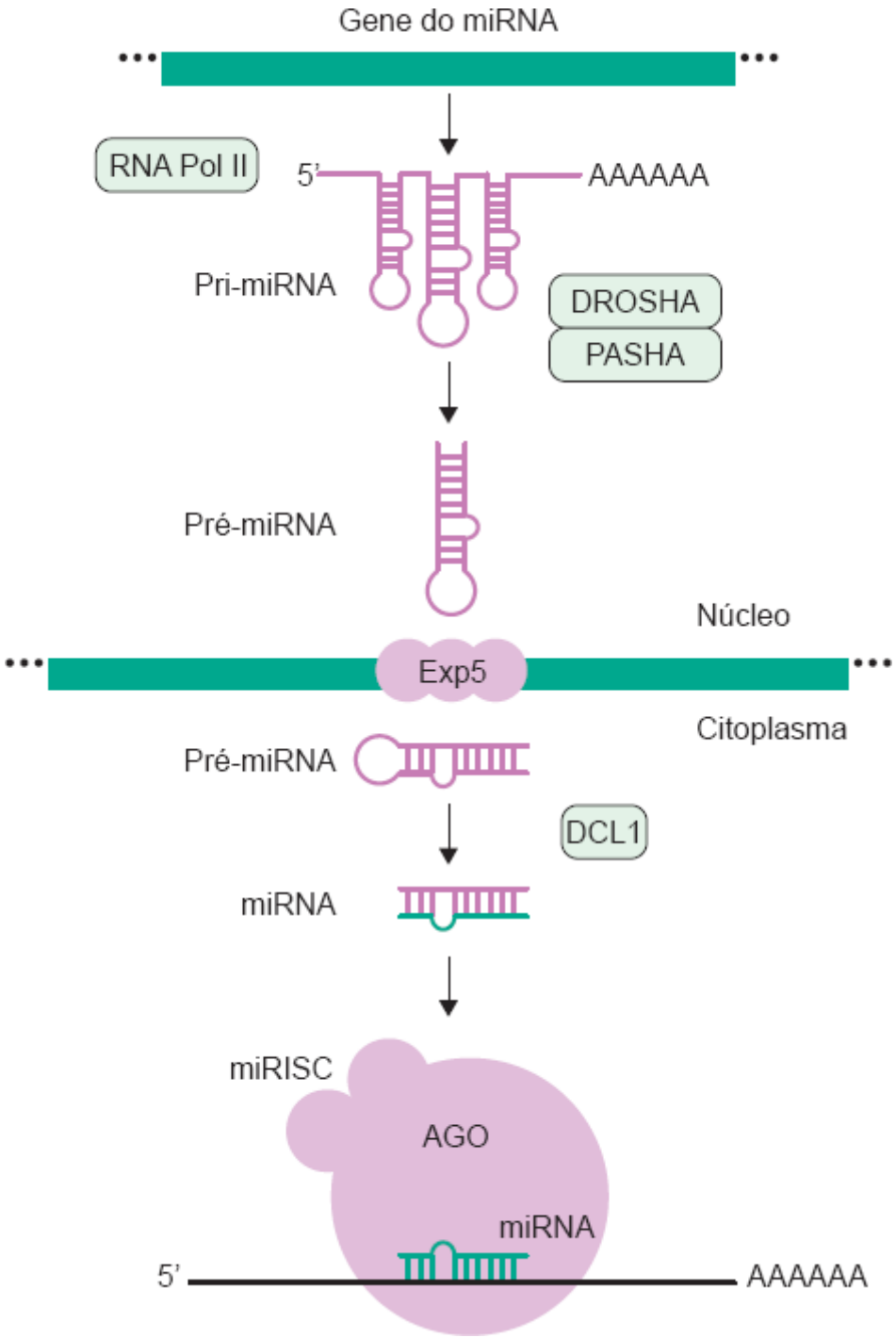


Figura 16.3 Biogênese dos miRNA em animais. Os genes de miRNA são transcritos por uma RNA Pol II em um pri-miRNA poliadenilado que, ainda no núcleo, é clivado por um complexo proteico formado por uma RNAse III (Drosha) e uma proteína Pasha (invertebrados) ou DGCR8 (vertebrados). A primeira clivagem resulta em um pré-miRNA, que

será exportado ao citoplasma pela ação de uma Exportina 5 (Exp5). No citoplasma ocorre o segundo processamento, dando origem ao *duplex* miRNA:miRNA*.

Uma vez que o complexo funcional miR-RISC é formado, o miRNA conduz a maquinaria proteica RISC (à qual está associado) até seu alvo cognato. Tanto em plantas quanto em animais, a ação dos miRNA pode se dar de duas maneiras distintas: via clivagem ou pela inibição da tradução dos mRNA-alvo (Figura 16.4). Quando o pareamento entre a fita guia e o mRNA envolve a totalidade ou quase totalidade das bases, gerando um pareamento completo, este mRNA será clivado pela ação da enzima AGO, uma das subunidades enzimáticas do complexo RISC. Quando o pareamento entre a fita guia e o mRNA-alvo ocorre de maneira parcial, o complexo RISC não promove a clivagem do mRNA, mas atua inibindo o processo de tradução deste, promovendo um impedimento espacial para a maquinaria ribossômica de tradução de proteínas. Até pouco tempo, acreditava-se que os miRNA agiam especialmente no nível de regulação da tradução em animais, inibindo o processo traducional, e no nível do controle pós-transcricional em plantas, degradando mRNA-alvo. No entanto, estudos mais recentes sugerem que ambos os mecanismos atuam em animais e plantas.

Contudo, a regulação da expressão gênica via miRNA apresenta algumas diferenças entre plantas e animais. Em geral, em animais, a repressão de expressão gênica é mediada pela atenuação da tradução por meio do pareamento parcial do miRNA a seu sítio-alvo (único ou múltiplos), localizado na região 3' não traduzida (3'UTR) do mRNA-alvo. Em plantas, devido ao pareamento completo entre o miRNA e o mRNA, a regulação dos alvos é preferencialmente direcionada pela clivagem em sítios únicos, localizados na região codificadora (CDS) dos mRNA (Figura 16.5).

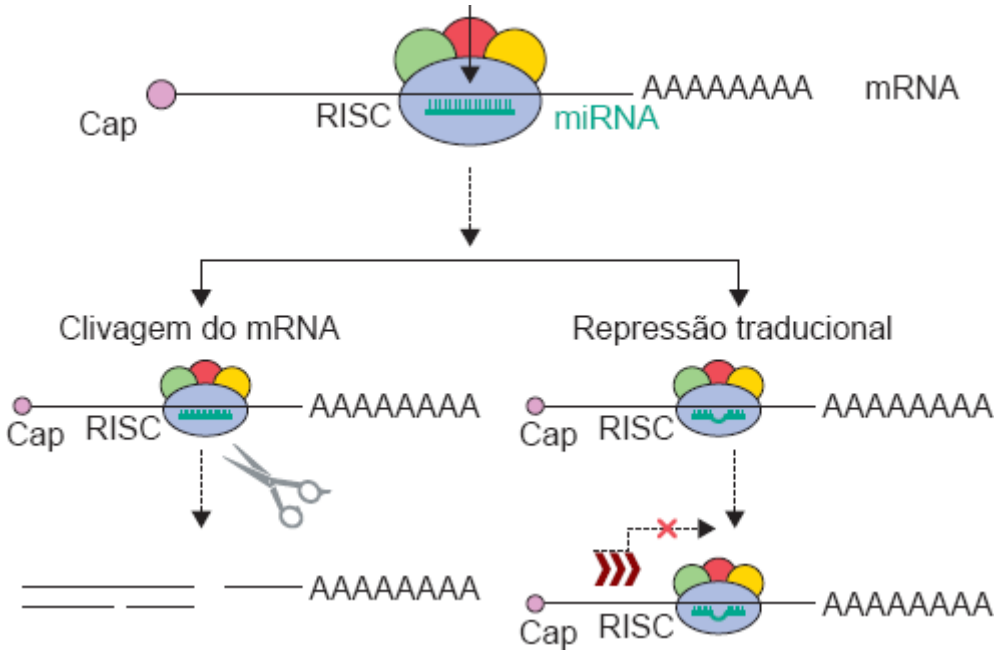


Figura 16.4 Mecanismos de ação dos miRNA. Uma vez que o complexo funcional miR-RISC é formado, o miRNA conduz a maquinaria proteica RISC (à qual está associado) até seu alvo cognato. Tanto em plantas quanto em animais, a ação dos miRNA pode ser via clivagem ou inibição da tradução dos mRNA-alvo. Quando o pareamento entre a fita guia e o mRNA envolve diversas bases, levando a um pareamento completo, o mRNA será clivado pela ação da enzima Argonauta, uma das subunidades do complexo RISC. Quando o pareamento entre a fita guia e o mRNA-alvo ocorre de maneira parcial, RISC não promove a clivagem do mRNA, mas atua inibindo o processo de tradução deste, promovendo um impedimento espacial para a maquinaria ribossômica de tradução de proteínas.

Os miRNA têm importante papel em diversos processos biológicos no nível de regulação da expressão gênica, tais como desenvolvimento de animais e plantas, proliferação e diferenciação celular, apoptose, metabolismo e repostas a múltiplos estresses. Em plantas, miR156 e miR172 são exemplos da importância desses pequenos RNA na transição de fases durante o desenvolvimento (Figura 16.6). Os miRNA participam da transição da fase juvenil para a adulta. O miR156 silencia fatores de transcrição SPL (*squamosa promoter binding protein like*), que são responsáveis pela manutenção das características fenotípicas juvenis. O miR156 é abundante em plântulas e sua expressão diminui conforme a planta chega à fase adulta; já o miR172 exibe um padrão oposto. Os fatores de

transcrição SPL ativam a expressão do miR172, o qual regula repressores da floração AP2 (*Apetala* 2). Por isso, durante a fase reprodutiva, os níveis do miRNA são aumentados e seu alvo é diminuído.

Já em mamíferos, um miRNA bem conhecido é o miR103/107, que participa da via de sinalização de insulina (Figura 16.7). O alvo desse miRNA é o transcrito codificador da proteína caveolina-1 (CAV-1), que é um regulador dos níveis do receptor de insulina na membrana plasmática de hepatócitos e adipócitos. Quando ocorre aumento da expressão do miR103/107 nessas células, CAV-1 é reduzido, levando à redução na resposta à insulina e, consequentemente, desequilíbrio na homeostase da glicose.

Figura 16.5 Sítios preferenciais de ancoramento de miRNA em mRNA de plantas e animais. A regulação da expressão gênica via miRNA apresenta algumas diferenças entre plantas e animais.

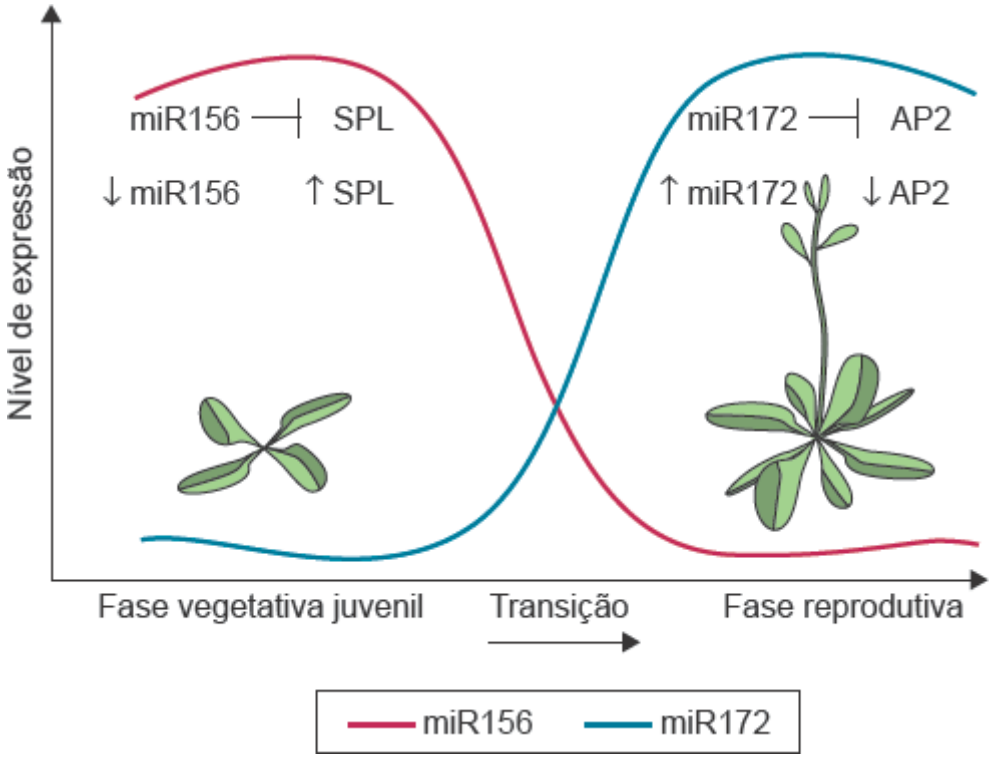


Figura 16.6 Impacto da ação regulatória do miR156 e miR172 no desenvolvimento de plantas. Esses miRNA participam na determinação da transição da fase juvenil para a adulta. O miR156, o qual silencia fatores de transcrição SPL, é altamente expresso em plântulas, e sua expressão diminui na fase adulta. Os fatores de transcrição SPL ativam a expressão do miR172, o qual reduz os níveis do repressor da floração AP2, que terão seus níveis de expressão reduzidos durante a fase reprodutiva.

miRtrons

Uma via alternativa na biogênese de miRNA foi descrita primeiramente em invertebrados, depois em mamíferos e, mais recentemente, em plantas. Neste caso, o pré-miRNA é processado a partir de um subproduto da via de *splicing* através do espliceossoma, e não por Drosha ou DCL (Figura 16.8). Após a transcrição, o mecanismo de *splicing* libera sequências que servem como precursoras de miRNA. Essas sequências também originam estruturas em formato de grampo, similares aos clássicos pré-miRNA, os quais serão processados em miRNA maduros, e assim, denominados miRtrons.

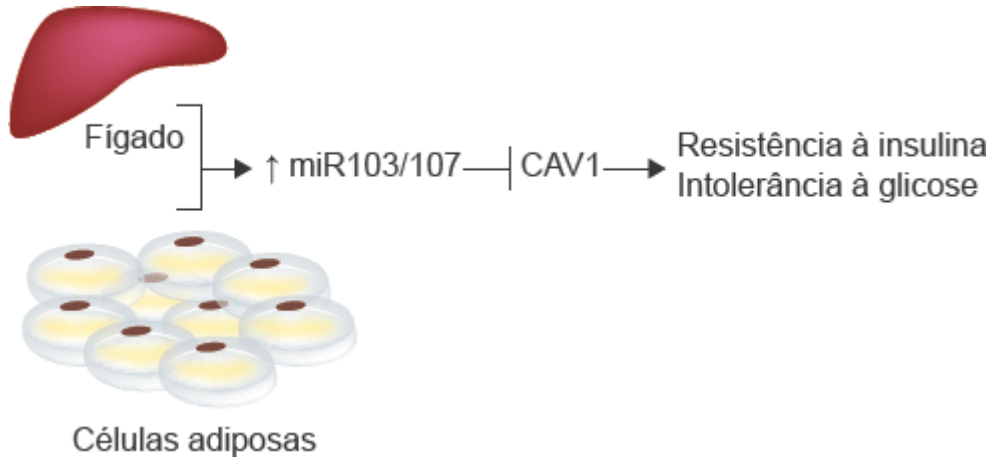


Figura 16.7 Ação regulatória dos miRNA 103/107 na via de sinalização da insulina em mamíferos. Um aumento da expressão do miR103/107 no fígado e em tecido adiposo leva a um desequilíbrio na homeostase da glicose. A proteína caveolina-1 (CAV-1), um regulador dos níveis do receptor de insulina na membrana plasmática, é o alvo desses miRNA. Portanto, com o aumento de miR103/107, CAV-1 é reduzida e, consequentemente, ocorre redução na resposta à insulina.

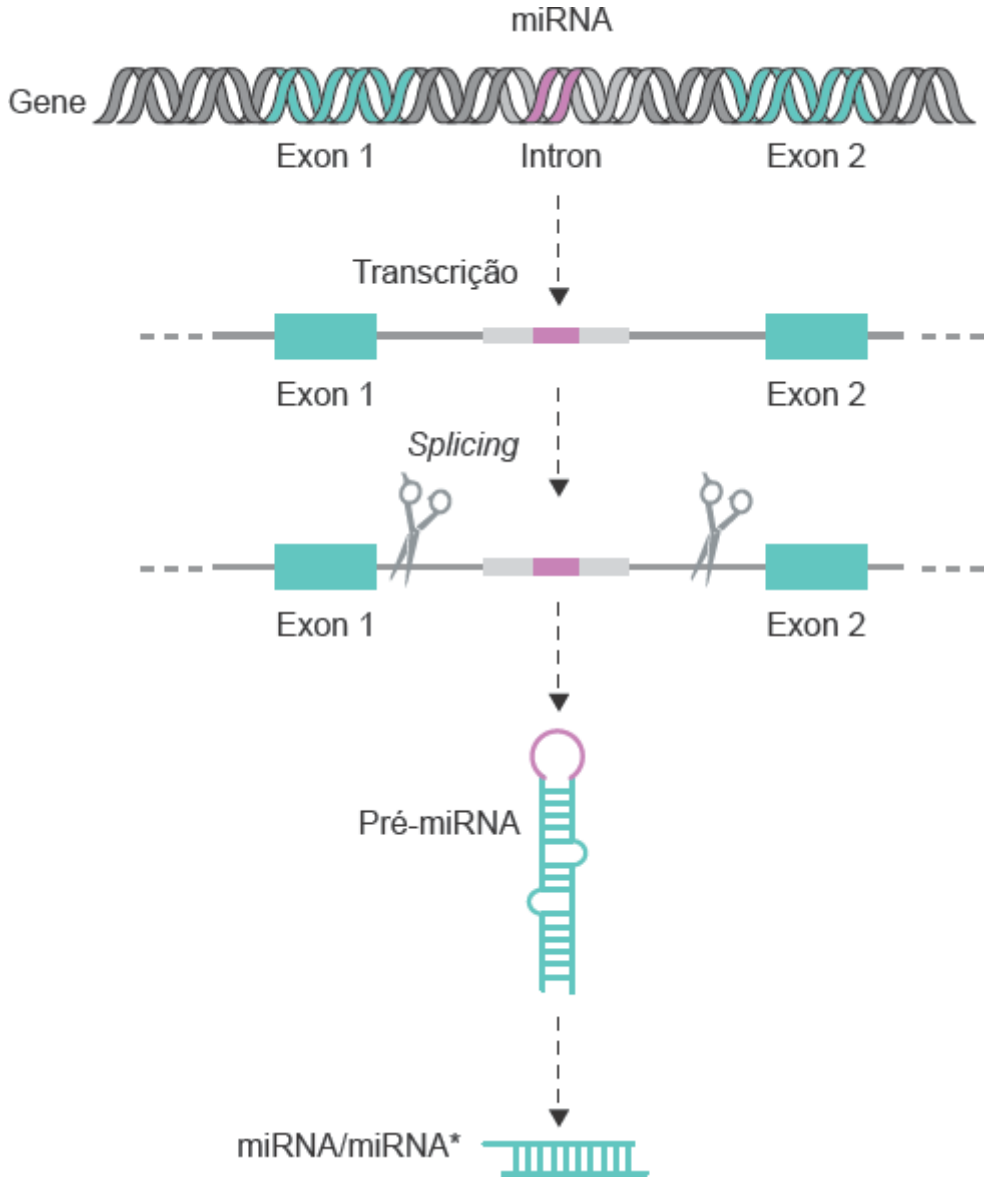


Figura 16.8 Os miRNA originados de introns: miRtrons. Os miRtrons derivam de regiões intrônicas curtas. Após a transcrição, o mecanismo de *splicing* libera sequências que servem como precursoras de miRNA. Essas sequências originam estruturas em grampo, similares aos clássicos pré-miRNA, os quais serão processados em miRNA maduros.

Aproximadamente 40% dos miRNA de animais são codificados a partir de introns de genes codificantes de proteínas e, desse modo, muitos dos transcritos de miRNA primários são simultaneamente precursores de mRNA. Em *Drosophila*, muitos introns têm cerca de 60 nt, o que é similar ao tamanho dos pré-miRNA nesta espécie. Por muito tempo, acreditou-se que os miRtrons eram apenas produzidos em animais; no entanto, em 2012, a subclasse de miRNA foi também identificada em *Arabidopsis* e arroz.

Outros pequenos RNA de interferência endógenos

Existem diferentes rotas endógenas de formação de diferentes endo-siRNA em plantas e animais. A seguir, de modo geral, serão descritas tais rotas.

Biogênese e papel biológico dos tasiRNA

Em plantas, existem genes *TAS*, os quais são transcritos por uma RNA polimerase II em longos RNA denominados tasiRNA primários (Figura 16.9). Tais transcritos são reconhecidos e clivados por miRNA associados a proteínas AGO1 ou AGO7 (dependendo do miRNA). No entanto, em vez de serem silenciados, esses transcritos de ncRNA clivados são usados como molde para a síntese de dsRNA por meio da ação de uma RNA polimerase dependente de RNA (RdRp-RDR6), com o auxílio de uma proteína supressora de silenciamento gênico (SGS3). Os dsRNA são reconhecidos e clivados por enzimas Dicer-*like* 4 (DCL4) e processados em pequenos RNA de 21 nt, os quais serão denominados tasiRNA. Os tasiRNA vão associar-se ao complexo RISC e silenciar seus alvos.

Existem quatro famílias de tasiRNA. As famílias recebem seus nomes conforme a localização dos genes *TAS* a partir dos quais estes foram transcritos. As famílias *TAS1*, *TAS2* e *TAS4* codificam mRNA com um sítio de clivagem, enquanto *TAS3* contém dois sítios. Os transcritos TAS1 e TAS2 são alvos do miR173, e TAS4 do miR828, e todos são clivados por AGO1. Já TAS3 é reconhecido por miR390 em dois sítios distintos, e a clivagem é mediada por AGO7.

A ação regulatória mais proeminente dos tasiRNA é o controle de fatores transcricionais responsivos à auxina (ARF), que participam da via de sinalização do hormônio auxina, o qual desempenha papel central no desenvolvimento das plantas.

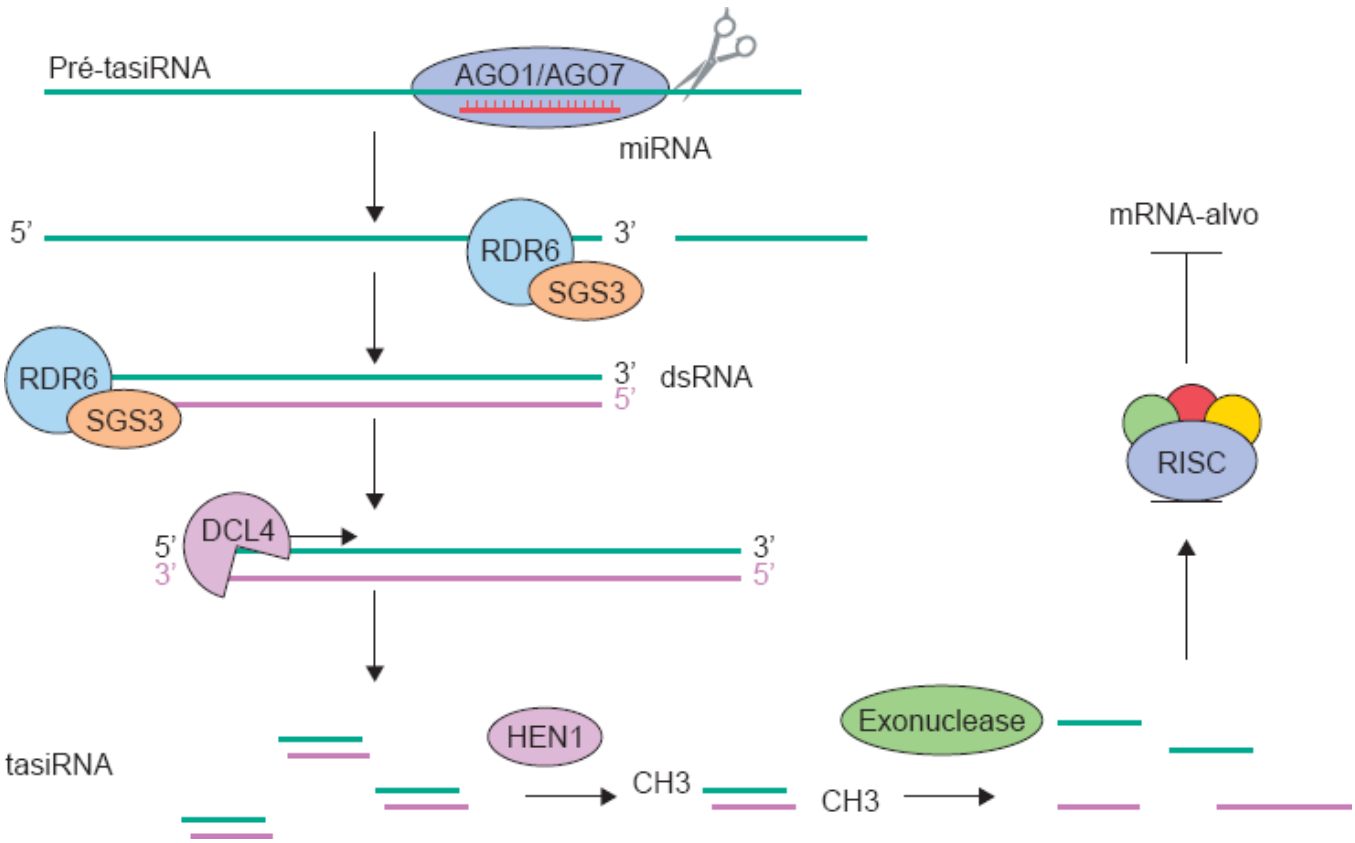


Figura 16.9 Geração de tasiRNA. Transcritos gerados a partir de *loci TAS* são reconhecidos e clivados por miRNA associados a proteínas Argonauta (AGO1 ou AGO7). Os fragmentos originados são estabilizados pelas proteínas SGS3 e utilizados como molde por uma RNA polimerase dependente de RNA (RdRp-RDR6) para a síntese de

dsRNA. Os dsRNA serão reconhecidos e clivados por uma DCL4, originando tasiRNA que podem associar-se ao complexo RISC, levando à regulação de seus alvos.

Biogênese e papel dos endo-siRNA

Existem siRNA de 24 nt de tamanho envolvidos no silenciamento transcricional de sequências repetitivas e/ou elementos de transposição em plantas. Os siRNA são inicialmente transcritos em RNA longos pela RNA polimerase IV, e convertidos em dsRNA pela proteína RDR2. Os dsRNA são reconhecidos pela DCL3 e processados em siRNA de 24 nt de tamanho. Os siRNA são carregados em um complexo multiproteico composto pelas proteínas AGO4, NRDP1b, metilases DRM2 e DRD1. Este complexo siRNA-AGO4-NRPD1b-DRM2-DRD1 será responsável pela metilação de citosinas no nível de DNA e o consequente silenciamento transcricional dos *loci* metilados.

Em *Drosophila*, são produzidos siRNA de 21 nt de tamanho, diferentemente dos siRNA de 24 nt gerados em plantas. Apesar da diferença de tamanho, os siRNA de *Drosophila* têm papel similar, pois estão envolvidos no silenciamento de elementos de transposição. Eles são gerados pela ação da enzima Dicer2 sobre transcritos de dsRNA; após serem carregados em complexos proteicos contendo AGO2, medeiam o silenciamento transcricional no nível de DNA (Figura 16.10).

Em *C. elegans*, os siRNA são transcritos inicialmente em dsRNA derivados de genes sobrepostos, RNA longos em formato de grampo (*hairpins*) e elementos de transposição. Os dsRNA são reconhecidos por um complexo proteico que contém a enzima essencial Dicer1, responsável pela clivagem dos RNA longos em pequenos siRNA. Os endo-siRNA estão relacionados com a manutenção e a estabilidade do genoma por meio do silenciamento de elementos transponíveis.

Em mamíferos, os siRNA são gerados a partir de transcritos de regiões invertidas no genoma, como os elementos SINE (*short interspersed elements* – elementos intercalares curtos). A biogênese dos siRNA em mamíferos depende da proteína Dicer. Os endo-siRNA estão relacionados com a manutenção e a estabilidade do genoma por meio do silenciamento de elementos transponíveis. Em humanos, por exemplo, há em torno de 1 milhão de sequências *Alu* (um tipo de SINE). A maquinaria de silenciamento de elementos de transposição desempenha papel crucial na manutenção das sequências *Alu* por meio da produção de siRNA contra elas. Aproximadamente 45% do genoma humano é composto de elementos transponíveis. Portanto, a presença de siRNA na célula tem importante papel biológico no silenciamento e manutenção desta grande quantidade desses elementos.

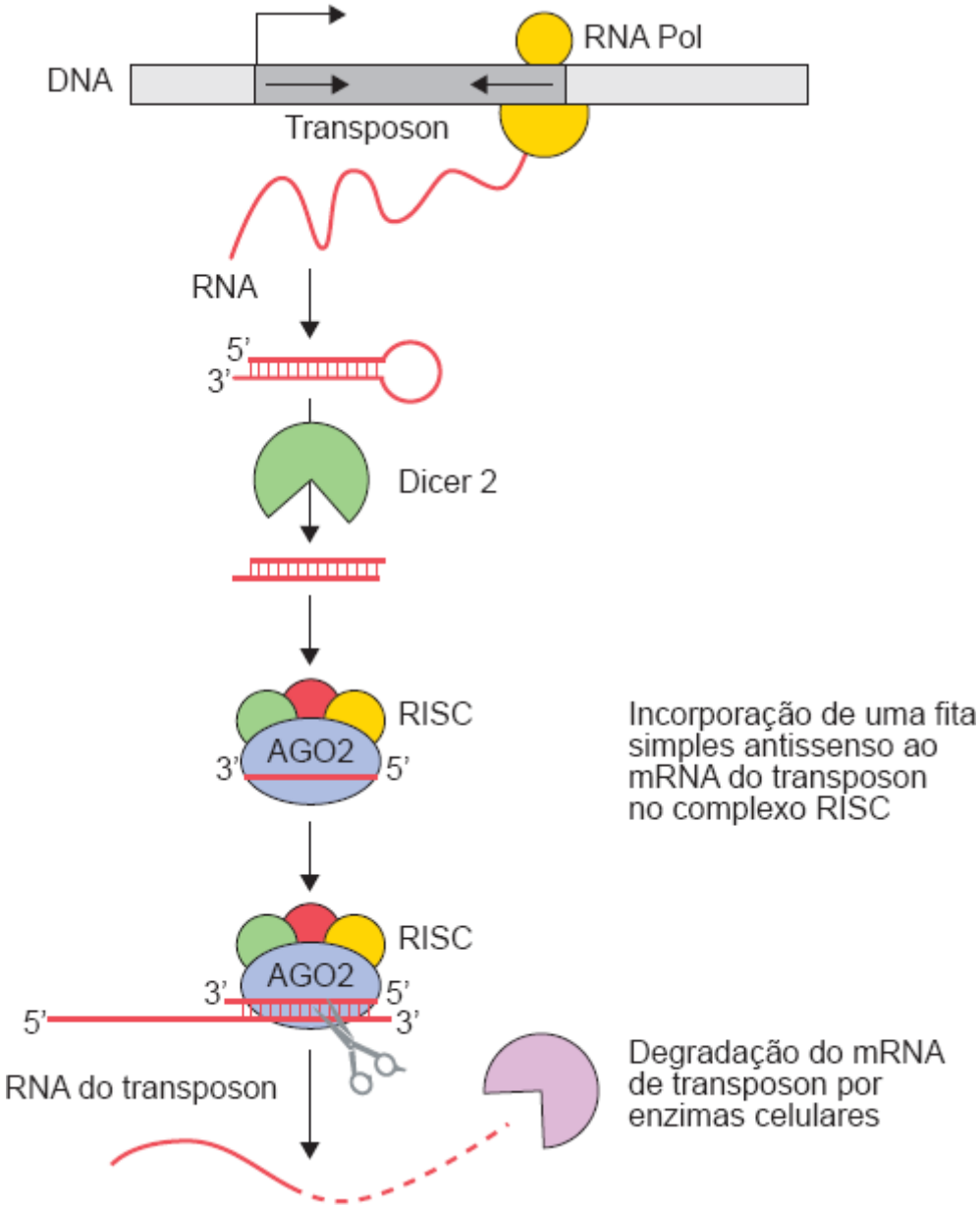


Figura 16.10 Produção de pequenos RNA de interferência endógenos (endo-siRNA) a partir de transposons de DNA em *Drosophila melanogaster*. O transcrito primário de transposons de DNA forma uma estrutura em grampo, a qual é clivada por uma enzima Dicer-2, originando um dsRNA. O dsRNA será processado originando pequenos RNA endógenos de fita simples com aproximadamente 21 nucleotídeos (endo-siRNA). Esses endo-siRNA serão incorporados ao complexo RISC, desencadeando o silenciamento dos transcritos dos mesmos transposons dos quais foram originados.

Pequenos RNA associados à proteína PIWI (piRNA)

Outra classe bastante conhecida de pequenos RNA endógenos são os PIWI-*interacting* RNA, ou piRNA. Estes pequenos RNA são assim classificados em decorrência de sua origem e sua interação com as proteínas PIWI, uma subfamília das proteínas Argonauta. Os piRNA são gerados em um sistema chamado pingue-pongue a partir de um precursor de fita simples, em um processo independente de enzimas RNase III, originando piRNA maduros com uma uridila no terminal 5', os quais podem conter de 24 a 32 nt, caracterizando-se por serem um pouco maiores que os miRNA e outros endo-siRNA (Figura 16.11).

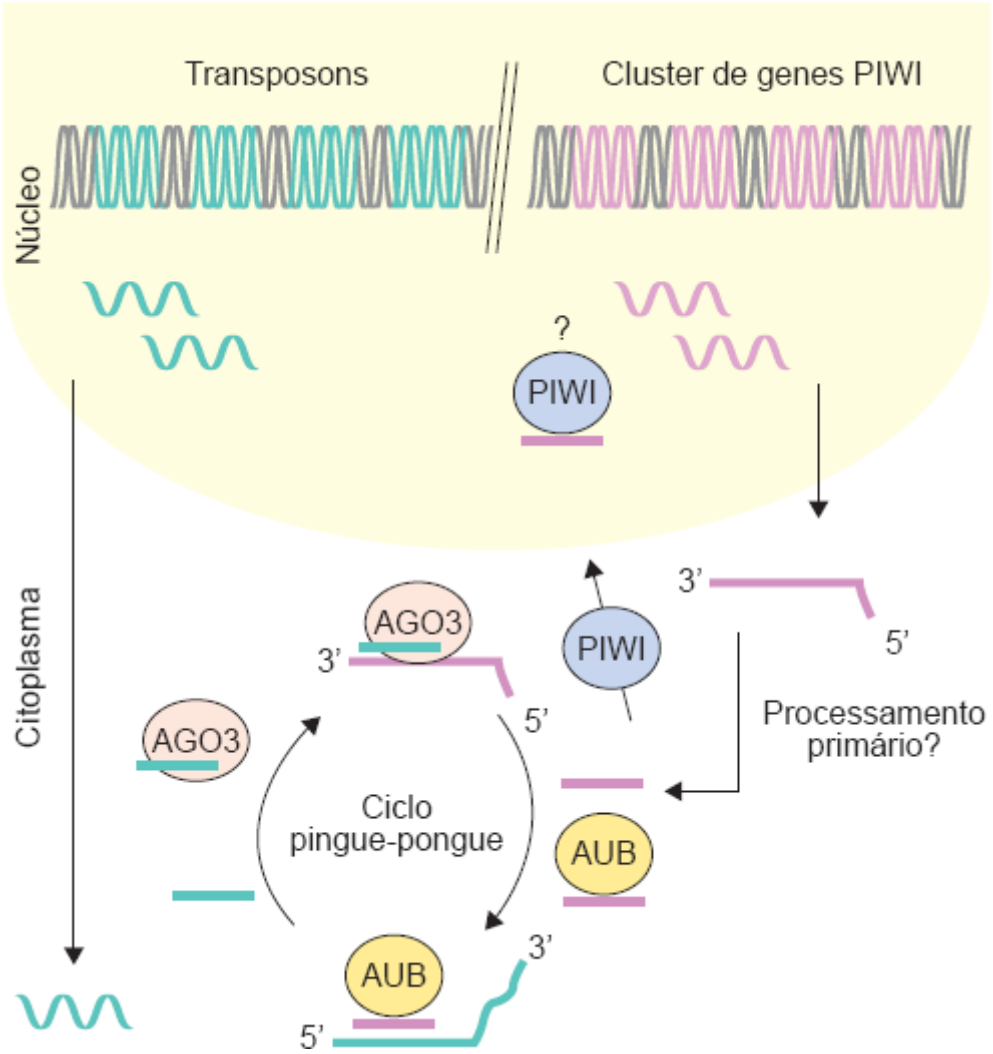


Figura 16.11 Ciclo pingue-pongue para a geração de piRNA e repressão de transposons. Os piRNA são pequenos RNA primariamente processados a partir de transcritos de *clusters* de genes *PIWI*. O processamento primário ocorre no citoplasma e, assim que formados, os piRNA ligam-se às proteínas Aubergine (AUB) e PIWI, e clivam transcritos antissenso provenientes de transposons. A clivagem origina novas sequências de piRNA em um novo sentido, as quais se associam a uma Argonauta AGO3. Desse modo, AGO3 associada aos novos piRNA mediará a clivagem de novos transcritos antissenso de PIWI, direcionando uma amplificação do mecanismo.

Os piRNA são originados de elementos repetitivos intergênicos, chamados de *clusters PIWI*. Tais *clusters* responsáveis pelos piRNA abrangem uma grande região do genoma, ocasionalmente, consistindo em mais de 100.000 bases; a maioria é composta de vários elementos transponíveis de DNA ou restos remanescentes destes. Assim, a maior parte dos piRNA tem orientação antissenso aos transcritos de transposons, o que pode induzir o silenciamento desses elementos. Mutações com perda de função em piRNA e seu cofatores, como as proteínas PIWI, levam a um aumento no número e nas realocações de transposons ao longo do genoma, causando defeitos no desenvolvimento gonadal e infertilidade em *Drosophila*.

Os transcritos originados de *clusters PIWI* são processados em sequências de piRNA, as quais são associadas a proteínas PIWI. Os fatores envolvidos na transcrição de *clusters PIWI* e sua regulação são ainda desconhecidos. O modelo atual propõe que o terminal 5' da sequência do piRNA é determinado antes da associação à proteína PIWI; enquanto o processamento do terminal 3' ocorre somente após a associação a PIWI. Assim, o comprimento das sequências maduras de piRNA é determinado após a ligação destas às proteínas PIWI. Após a maturação, a extremidade 3' do piRNA é metilada pela ação de uma proteína homóloga a HEN1. Tal modificação mantém a estabilidade dos piRNA, e é observada também em miRNA de plantas, mas não em animais. Os piRNA primários serão apenas associados a proteínas PIWI específicas. Em *Drosophila*, há três proteínas PIWI: Aubergine (AUB), PIWI e Argonauta 3 (AGO3). Os piRNA primários apenas se associam à AUB e PIWI; o mecanismo desta seleção ainda não foi esclarecido.

Em células germinativas de moscas, os piRNA primários estão sob controle de um sistema de amplificação que força uma alta expressão deles mesmos. No sistema pingue-pongue, a proteína AUB ligada a piRNA primários cliva

o RNA-alvo através da sua atividade de endonuclease. Esse processo gera piRNA secundários que irão se ligar a uma AGO3. Então, AGO3 associada a piRNA secundários cliva seus RNA-alvo e origina outros piRNA que se ligarão a AUB, desencadeando assim um ciclo autodependente. Em geral, quando piRNA primários ligam-se a proteínas PIWI, estes são translocados para o núcleo e, desta maneira, raramente PIWI contribui para a formação de piRNA secundários. Até o momento, é desconhecida a função do complexo piRNA-PIWI no núcleo.

No ciclo pingue-pongue, os transcritos de transposons são clivados por piRNA-AUB; isso significa que tal ciclo realiza duas funções simultaneamente – ou seja, produz piRNA secundários e silencia transposons pela clivagem dos seus transcritos.

Pequenos RNA derivados de RNA transportador (tRF)

Os RNA transportadores (tRNA) são uma classe de RNA não codificante, onipresente e necessário para a leitura ou decodificação do código genético e subsequente síntese proteica. Os tRNA são transcritos por RNA polimerase III em pré-tRNA, os quais são processados por uma RNase P, que cliva sua sequência 5' *leader*, e por uma RNase Z, que retira sua sequência 3' *trailer*. Após esse processamento, ocorrem modificações de bases, tais como adição de pseudouridina e adição de um trinucleotídio "CCA" à extremidade 3', a qual é posteriormente utilizada para acoplamento do aminoácido, resultando no tRNA ativado ou maduro.

Recentemente, uma nova classe de pequenos RNA, contendo de 18 a 22 nt, foi identificada a partir de culturas celulares humanas, e denominada fragmentos derivados de tRNA (*transfer RNA-derived fragments* ou tRF) (Figura 16.12). Um tRNA pode originar até três tipos de tRF, os quais são nominados conforme a posição de origem no tRNA, podendo ser tRF-1, originados pela liberação da extremidade 3' do pré-tRNA pela RNase Z; tRF-5', originados da extremidade 5' do tRNA e tRF-3', produzidos a partir da extremidade 3' do tRNA maduro – isto é, após a adição do trinucleotídio CCA.

As enzimas que clivam os tRNA, originando essa classe de pequenos RNA, são RNases P e Z, mas também RNases do tipo III, denominadas Dicer, as mesmas envolvidas no processo de clivagem dos miRNA. Experimentos de imunoprecipitação têm demonstrado que a classe de pequenos ncRNA também se associa às proteínas Argonauta, indicando seu envolvimento em processos de regulação pós-transcricional. A atuação no silenciamento gênico via iRNA foi demonstrada com tRF silenciando *in vitro* o gene repórter da proteína verde-fluorescente (GFP).

Os tRF possivelmente apresentam um papel de reguladores da replicação viral de retrovírus endógenos humanos (HERV), os quais compreendem 7% do genoma na forma de retrotransposons de repetições terminais longas – LTR. Apesar de o mecanismo de regulação dos HERV por tRF não ser ainda conhecido, estes apresentam alta complementaridade de sequência a LTR, maior que a encontrada em relação a outras classes de retrotransposons, como os LINES e os SINES.

Tais fatos indicam que os fragmentos derivados de RNA transportador contêm não somente uma possível função de reguladores pós-transcricionais da expressão gênica, como os miRNA, mas também inibidores de replicação de retrotransposons endógenos.

Pequenos RNA derivados de RNA vault (svRNA ou svtRNA)

Os vaults são grandes complexos ribonucleoproteicos de formato ovoide e localização citoplasmática, que foram descobertos em 1986 em tecidos do fígado de ratos. Foram identificados em uma ampla variedade de eucariotos, incluindo mamíferos, aves, peixes, anfíbios, equinoides e mixomicetos. A presença difundida e a ampla conservação evolutiva dos RNA vault (vRNA), entre distintos filos, sugere que essas partículas exerçam papéis biológicos importantes. Embora as funções dos vaults não tenham sido completamente elucidadas, alguns trabalhos demonstram que essas ribonucleoproteínas estão envolvidas no transporte celular e na resistência a múltiplos fármacos.

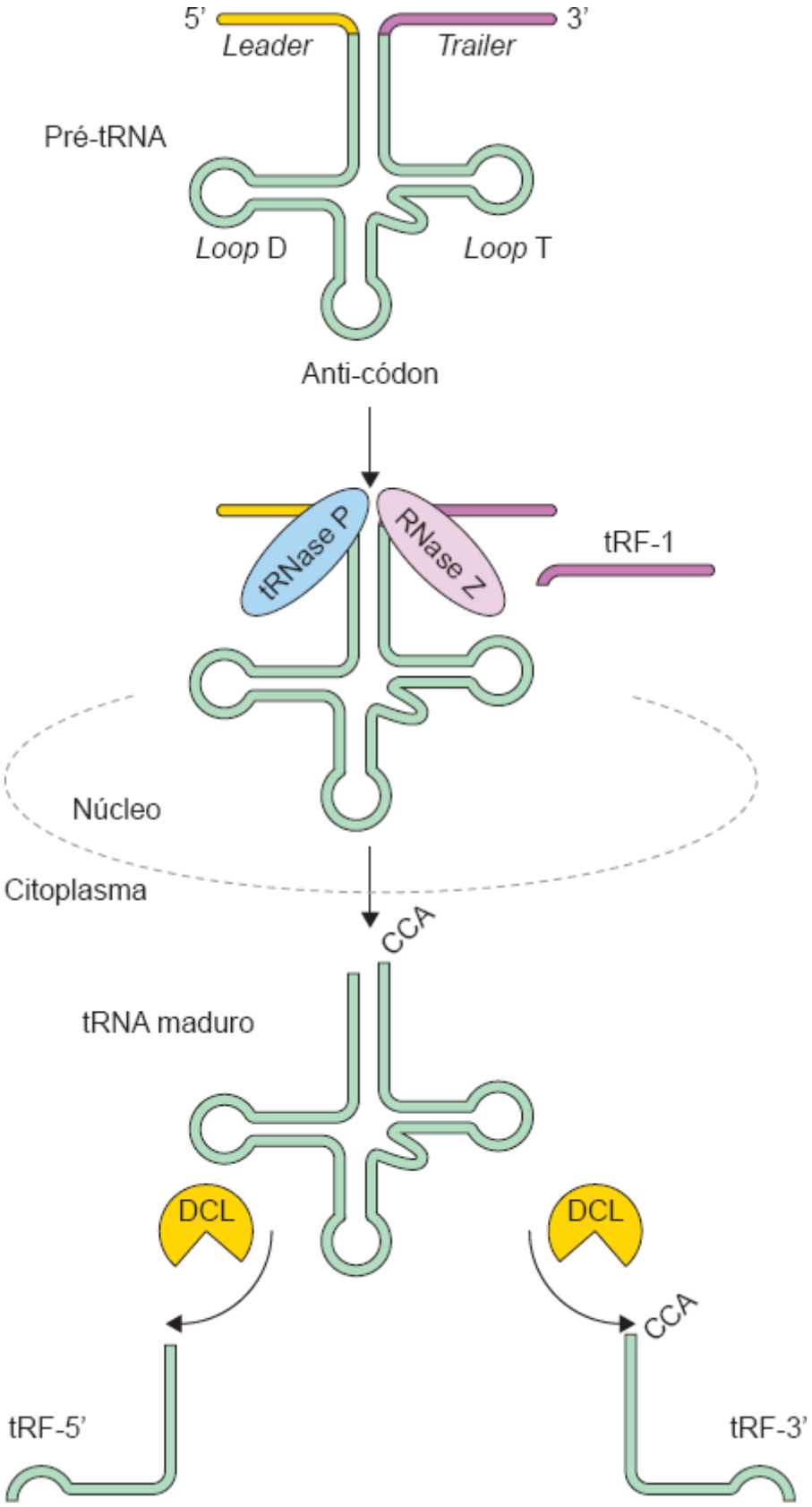


Figura 16.12 Biogênese de pequenos RNA derivados de RNA transportadores: tRF. O tRNA é transcrito por uma RNA polimerase do tipo III em um precursor (pré-tRNA), contendo uma sequência *leader* no terminal 5', e uma *trailer* na extremidade 3'. Ambas as sequências são removidas por ação de uma RNase P e Z, respectivamente. A sequência *trailer*, liberada pela RNase Z, consiste em um tRF do tipo I (tRF-1). Após esse processamento, ocorre a adição de um CCA na extremidade 3', finalizando a maturação do tRNA. O tRNA maduro é então exportado para o citoplasma, onde poderá sofrer clivagens por enzimas do tipo Dicer na porção do Loop T, liberando o tFR-3', ou no Loop D, originando tFR-5'. Os tRF podem variar de 18 a 22 nt de comprimento.

Análises da função biológica dos complexos vaults demonstram que eles são formados por três proteínas: uma proteína vault principal (MVP), uma polimerase vault poli-ADP-ribose (vPARP) e uma proteína associada à telomerase (TEP1). Além disso, foi observado que alguns vRNA também produzem pequenos RNA (svRNA) com

aproximadamente 23 nt, por meio de um mecanismo diferente da biogênese de miRNA (Figura 16.13). As sequências de vRNA podem se dobrar e originar estruturas que se assemelham a precursores de miRNA. O braço 5' do vRNA em grampo codifica dois svRNA de 23 nt que se sobrepõem, e são nominados svRNAa e svRNAb. Outros dois svRNA são derivados do braço oposto (3') e pareiam com os svRNA mencionados anteriormente. Uma vez que estes svRNA originados do braço 3' são menos abundantes que os outros, recebem o nome de svRNAa* e svRNAb*, por se assemelharem sequências-estrela (*) dos miRNA. Os svRNA também se associam a proteínas Argonauta, o que demonstra que esses pequenos RNA também utilizam a clivagem de sequências específicas ou o bloqueio da tradução do transcrito-alvo para a regulação da expressão gênica, de modo semelhante aos miRNA. Em seres humanos, foi caracterizado que o svRNAb reprime a enzima CYP3A4, uma enzima-chave no metabolismo de detoxificação de fármacos.

Pequenos RNA associados às proteínas QDE-2 (qiRNA)

Pequenos RNA associados à proteína *quelling deficiente* (QDE-2), ou qiRNA, são um grupo de pequenos RNA endógenos identificados em *Neurospora crassa*, produzidos após exposição deste fungo a agentes danosos ao DNA. A estrutura básica e a biogênese dos qiRNA são similares às de outros siRNA endógenos, os quais são observados interagindo com proteínas Argonautas.

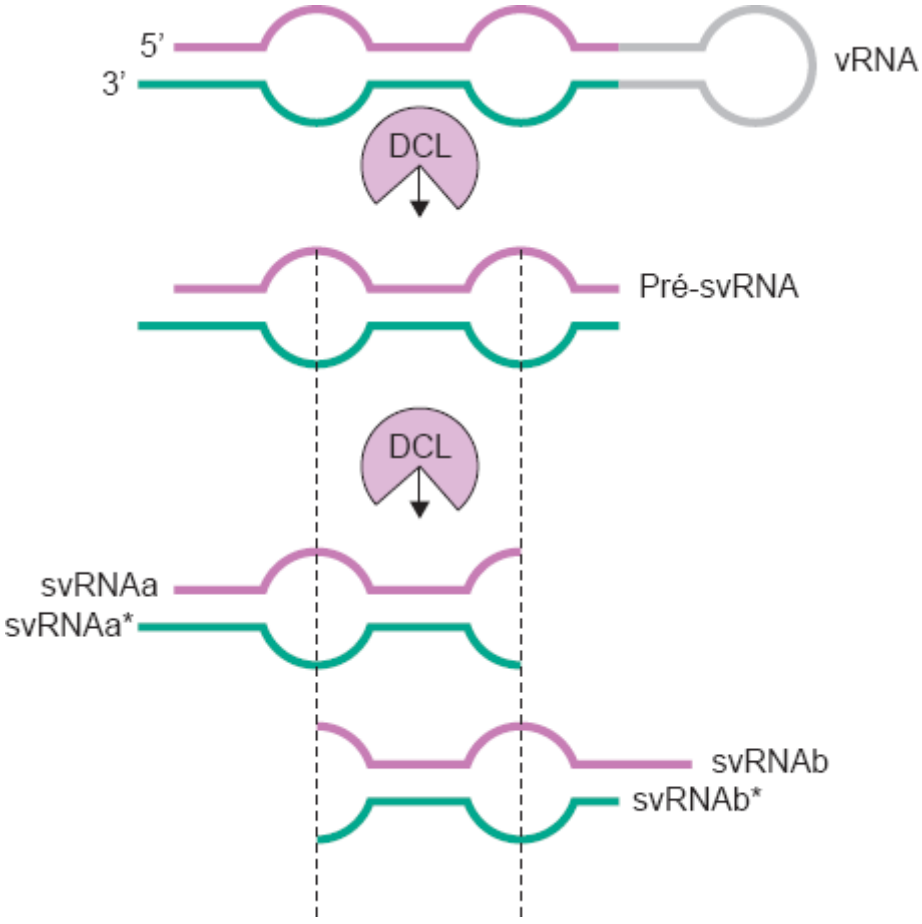


Figura 16.13 Biogênese de pequenos RNA derivados dos RNA vault: svRNA. São componentes da partícula vault, os quais são complexos ribonucleoproteicos associados à resistência a fármacos. Os svRNA são sintetizados por uma via dependente de Dicer, e atuam de modo similar aos miRNA no controle da expressão gênica.

Os qiRNA têm de 21 a 23 nt e também requerem proteínas RNA polimerase dependente de RNA (RdRp), Dicer e Argonauta para sua biogênese, bem como uma helicase QDE-3 para seu processamento. As sequências maduras de qiRNA, assim como outros siRNA, geralmente apresentam uma uridila na extremidade 5'.

A característica mais marcante dos qiRNA é o fato de sua produção ocorrer a partir de sequências repetitivas de DNA ribossômico (rDNA) em resposta a danos que ocorrem nessas mesmas sequências. Assim, sugeriu-se que o sinal que desencadeia a produção de qiRNA é a síntese de RNA aberrantes a partir das sequências de DNA danificado.

O modelo proposto sugere que danos ao DNA promovem recombinações no DNA repetitivo, o qual é reconhecido por proteínas QDE-3 e QDE-1 que irão produzir RNA aberrantes e dsRNA. A molécula de dsRNA será processada por proteínas do tipo Dicer (DCL) produzindo qiRNA maduros, os quais irão se associar a proteínas Argonautas QDE2. Acredita-se que a função dos qiRNA seja interferir na tradução proteica de transcritos originados a partir de regiões danificadas do DNA (Figura 16.14).

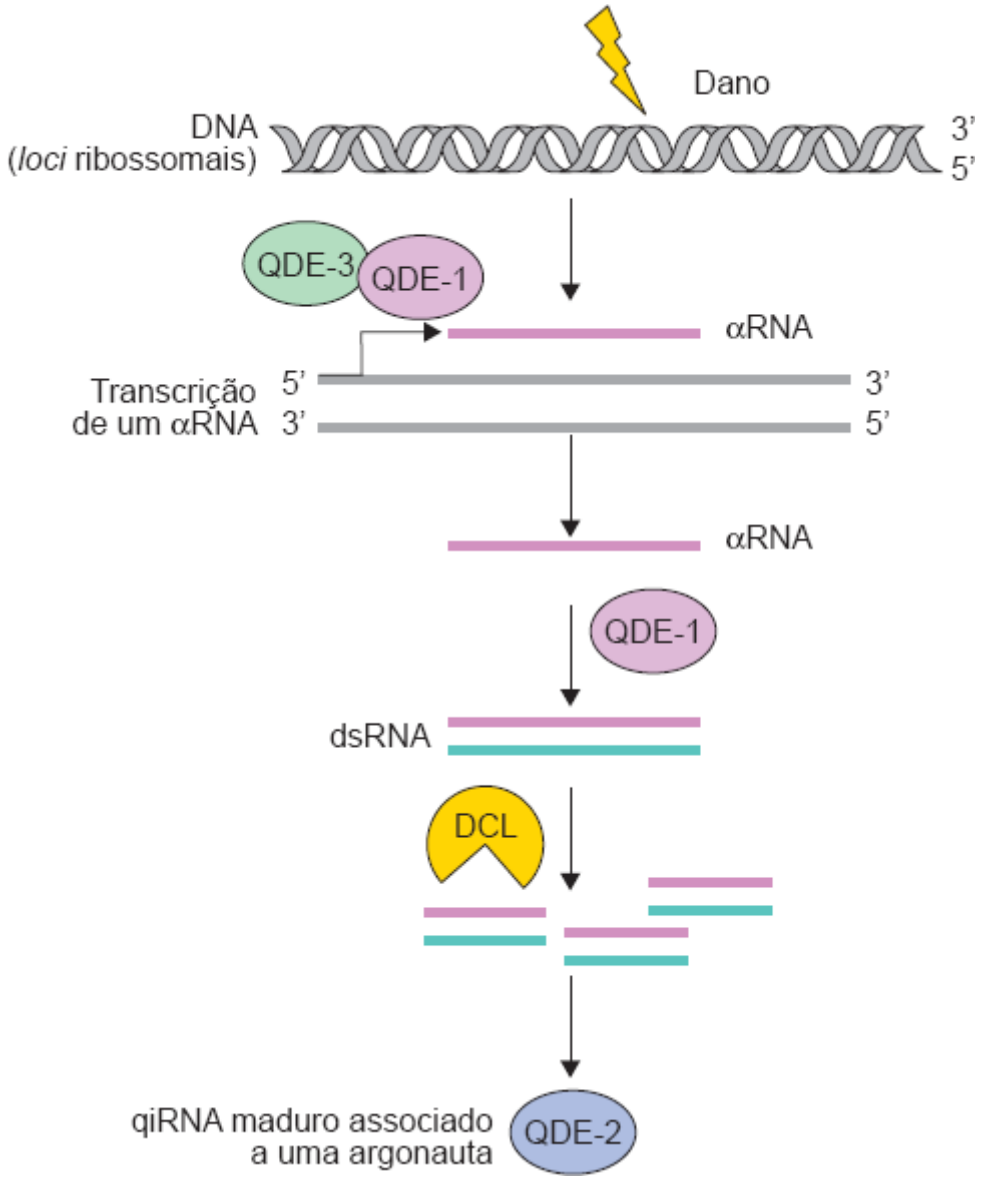


Figura 16.14 Biogênese de qiRNA em *Neurospora crassa*. Danos ocasionados no DNA, em genes ribossomais, induzem a produção de pequenos RNA de 20 a 21 nucleotídeos de tamanho. Inicialmente, um RNA de fita simples aberrante (aRNA) é produzido pela ação de uma RNA polimerase dependente de RNA (RdRp) QDE-1 e de um homólogo de DNA helicase QDE-3. Em seguida, um intermediário de dsRNA é produzido pela ação sequencial da RdRp-QDE-1. A molécula de dsRNA será processada por proteínas do tipo Dicer (DCL) produzindo qiRNA maduros, os quais irão se associar a proteínas argonautas QDE-2.

Pequenos RNA de interferência exógenos (exo-siRNA)

Pequenos RNA exógenos podem ser originários da interação de plantas com vírus e a partir de transgenes em animais e plantas (Figura 16.15). O mecanismo que origina os siRNA é denominado silenciamento gênico pós-transcricional (do inglês, *post-transcriptional gene silencing* – PTGS) em plantas e iRNA em animais. Em plantas, há diferentes mecanismos de geração de siRNA via PTGS: S-PTGS (silenciamento de transgenes), VIGS ou VI-PTGS (silenciamento induzido por vírus) e IR-PTGS (silenciamento induzido por sequências invertidas). Os dsRNA oriundos de vírus são reconhecidos pela maquinaria de silenciamento da planta hospedeira. Esses RNA são processados em siRNA derivados do vírus. Dependendo da via de geração de siRNA, se apresentarem 24 nt de tamanho pela enzima DCL3 ou 21 nt de tamanho pela enzima DCL4, eles podem estar relacionados com o remodelamento da cromatina e ou silenciamento viral, respectivamente. Em *C. elegans*, a geração de exo-siRNA

pode ocorrer a partir da interação com vírus. Similar ao mecanismo de plantas, este fenômeno denomina-se VI-iRNA (silenciamento induzido por vírus). Os dsRNA de origem viral são reconhecidos pela maquinaria de silenciamento, originando os siRNA, que são responsáveis pela defesa antiviral. De outro modo, como parte do mecanismo de coevolução, os vírus também contêm proteínas de inibição do mecanismo de silenciamento disparado pelo organismo hospedeiro. Portanto, os exo-siRNA estão relacionados com a defesa antiviral, pois os siRNA produzidos levam ao silenciamento do vírus. Além disso, o silenciamento de transgenes ocorre pela geração de siRNA a partir dos dsRNA transcritos dos transgenes inseridos de forma estável no genoma hospedeiro e/ou transiente.

Os exo-siRNA já foram detectados em plantas, vermes, *Drosophila* e fungos, mas não em *Saccharomyces*.

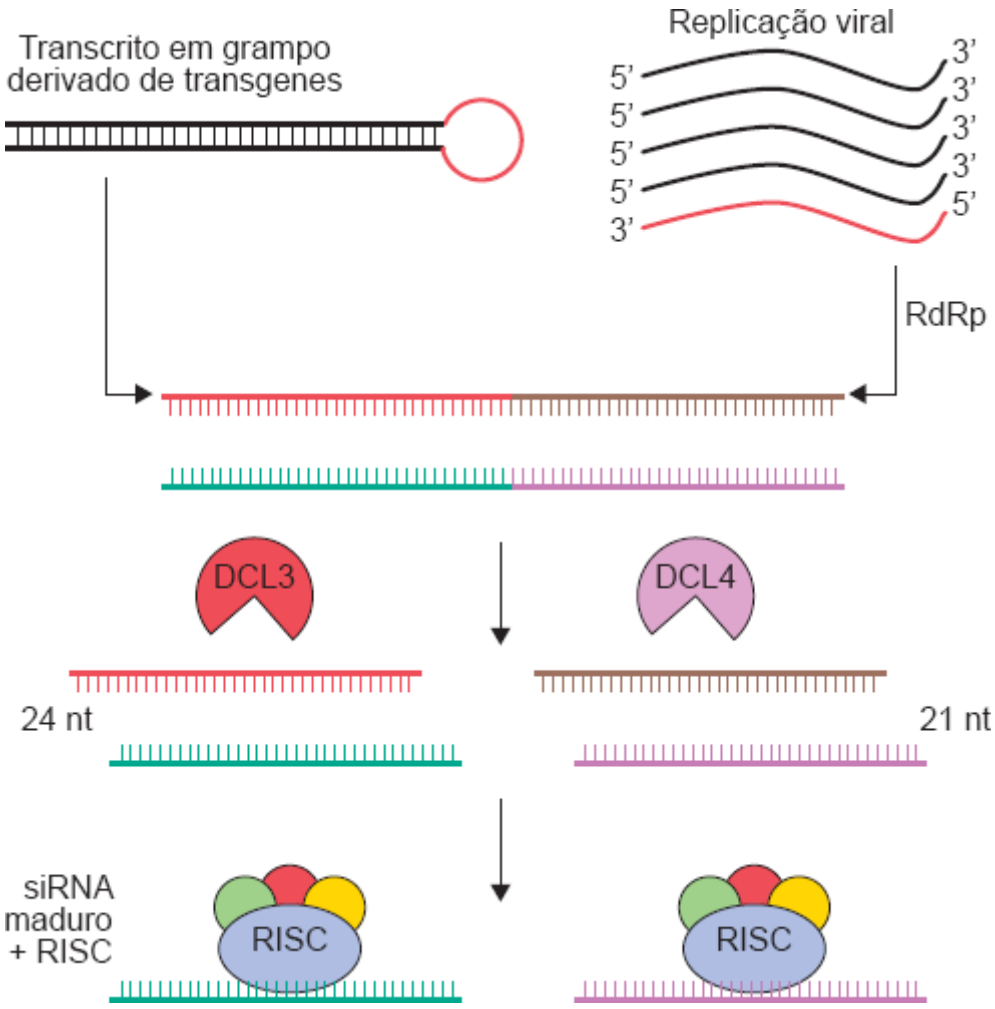


Figura 16.15 Representação esquemática da formação de siRNA exógenos (exo-siRNA). Os dsRNA originados de diversas fontes, como RNA replicativos produzidos durante uma infecção viral ou transcritos artificiais expressos em organismos transgênicos, serão processados por enzimas Dicer (DCL) gerando siRNA primários. Conforme a enzima DCL que atuar sobre o dsRNA, tamanhos diferentes de siRNA serão produzidos. Os siRNA serão acoplados ao complexo RISC, que promoverá reconhecimento e clivagem de um mRNA-alvo.

Pequenos RNA virais de interferência (vsiRNA)

Essa classe de pequenos ncRNA é originada a partir do material genético de um vírus, o qual transfectou uma célula hospedeira. Tanto o vírus de DNA quanto o de RNA são capazes de ativar silenciamento gênico por meio da formação de um DNA ou dsRNA, o qual é reconhecido pela maquinaria de iRNA, produzindo pequenos RNA de 21 a 24 nt de comprimento.

Os vsiRNA são posteriormente amplificados por uma enzima RNA polimerase dependente de RNA (RdRp), que utiliza uma das fitas como molde para síntese de uma fita complementar. O dsRNA resultante é novamente utilizado pela maquinaria de iRNA, resultando em vsiRNA secundários.

Os vsiRNA secundários podem ser mobilizados com diferentes proteínas Argonautas, possibilitando não somente o silenciamento do transcrito-alvo, mas também modificações no DNA e histonas. Assim, a função dos vsiRNA é facilitar a infecção do vírus no organismo hospedeiro, usufruindo da mesma maquinaria de iRNA do

hospedeiro, a qual também é utilizada para combatê-lo, a fim de silenciar genes do hospedeiro que inibam o desenvolvimento viral.

Aplicações biotecnológicas dos sRNA

Pequenos RNA de interferência (siRNA)

Os pequenos RNA de interferência (siRNA) podem ser utilizados para estudos da função gênica ou obtenção de organismos com características especiais com base no silenciamento de genes de interesse. A estratégia básica é caracterizada pela geração de organismos geneticamente modificados que contêm transgenes na posição senso e antissenso. Esses transgenes são capazes de gerar dsRNA, os quais serão clivados por enzimas Dicer ou DCL e irão gerar pequenos fragmentos de RNA de fita simples, os chamados siRNA. A sequência antissenso de siRNA é acoplada ao complexo RISC e identificará o mRNA-alvo, levando a um silenciamento deste transcrito.

O uso de siRNA é uma ferramenta bastante utilizada para a investigação da função de genes (genômica funcional), uma vez que dsRNA sintéticos ou siRNA ou transgenes podem ser introduzidos nas células para induzir o silenciamento seletivo de genes específicos de interesse. Por outro lado, a exploração racional de siRNA também é uma ferramenta promissora em biotecnologia e medicina.

Uma possibilidade é a aplicação de siRNA em terapias antivirais, incluindo a inibição da expressão de genes virais responsáveis por doenças como AIDS, hepatite, gripe, dentre outras. O uso de RNA interferentes também é considerado uma alternativa promissora para o tratamento do câncer, por meio do silenciamento de oncogenes. Na Tabela 16.1 são apresentados alguns dos fármacos com base na tecnologia de siRNA que já estão em testes pré-clínicos e clínicos.

O silenciamento via RNA de interferência pode também ser aplicado na manipulação de plantas com baixos níveis de toxinas e redução dos níveis de alergênios, bem como resistência a vírus, patógenos e pragas (Figura 16.16) que acometem inúmeras plantas de interesse econômico. Na Tabela 16.2 são listadas algumas das aplicações de iRNA em plantas, assim como os locais onde o trabalho está sendo desenvolvido.

Tabela 16.1 Ensaios clínicos empregando terapia com base em siRNA.

Cenário clínico	Fármaco	Indicações	Transcrito-alvo
Doenças oculares	TD101	Paquioníquia congênita	Mutante da queratina 6A N171 K
	QPI-1007	Neuropatia óptica isquêmica anterior não arterítica	Caspase 2
	AGN211745	Degeneração macular relacionada com a idade	VEGF-R1
	PF-655	Edema macular diabético	RTP801
	SYL040012	Glaucoma	Receptor β2 adrenérgico
	Bevasiranib	Degeneração macular	VEGF
Câncer	CEQ508	Polipose adenomatosa	β-catenina
	ALN-PLK1	Tumores do fígado	PLK1
	FANG	Tumores sólidos	Furina
	CALAA-01	Tumores sólidos	RRM2
	SPC2996	Leucemia mieloide crônica	BCL-2
	ALN-VSP02	Tumores sólidos	VEGF
	NCT00672542	Melanoma metastático	LMP(2)(7) e MECL1
	Atu027	Tumores malignos sólidos	PKN3

Doenças renais	QPI-1002/15NP	Lesão renal aguda	p53
Redução de LDL	TKM-ApoB	Hipercolesterolemia	APOB
Antivirais	SPC3649	Vírus da hepatite C	miR-122
	pHIV7-shI-TAR-CCR5Z	HIV	Proteínas HIV Tat, RNA HIV TAR
	ALN-RSV01	Pacientes com RSV	Nucleocapsídio do RSV

APOB (apoliproteína B); BCL-2 (células B de linfoma); HIV (vírus da imunodeficiência humana); LDL (lipoproteínas de baixa densidade); LMP(2)(7) (proteassoma subunidade beta tipo 2 e tipo 7); MECL1 (proteassoma subunidade beta tipo 10); PKN3 (proteinoquinase N3); PLK1 (polo-like quinase 1); RRM2 (ribonucleotídeo redutase); RSV (vírus sincicial respiratório humano); RTP801 (proteína relacionada com dano ao DNA); VEGF (fator de crescimento vascular endotelial).

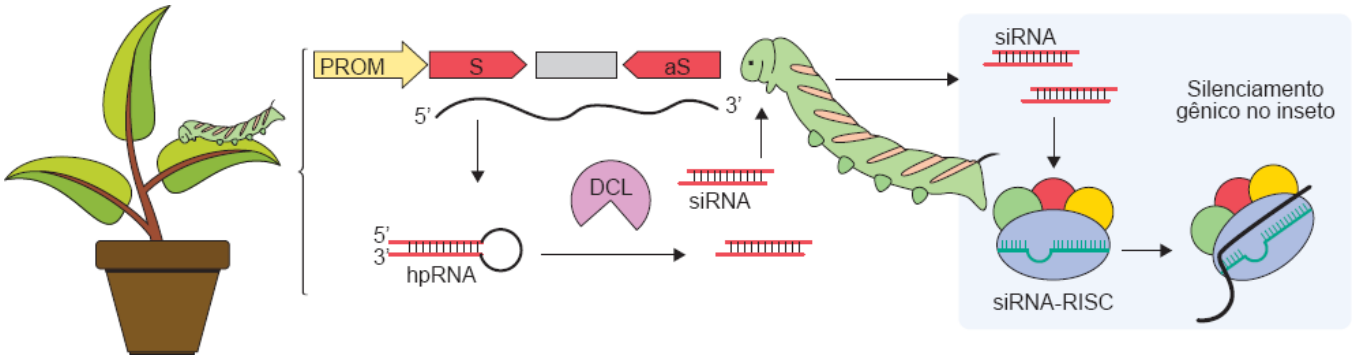


Figura 16.16 Aplicação biotecnológica utilizando pequenos RNA de interferência (siRNA). Produção de uma planta transgênica contendo uma construção formada por uma mesma sequência de interesse na posição senso (S) e antissenso (aS), separadas por uma sequência espaçadora, sob o controle de um promotor (PROM). Os transcritos produzidos a partir da expressão do transgene irão se organizar em estruturas no formato de um grampo (hpRNA), que serão processadas pela maquinaria de biogênese de siRNA da planta (DCL), produzindo siRNA. No exemplo apresentado, foi escolhida uma sequência expressa durante o desenvolvimento de uma lagarta. Ao se alimentar da planta transgênica, a lagarta ingere os siRNA produzidos na planta, que foram projetados com sequências homólogas a genes vitais do inseto, resultando no silenciamento do gene-alvo da lagarta e consequente resistência da planta ao inseto praga.

amiRNA | miRNA artificiais

A construção de cassetes de DNA para expressão de miRNA artificiais (amiRNA) têm sido utilizada como um meio alternativo e de maior especificidade pelas sequências-alvo. Tanto em plantas quanto em animais e seres humanos, esta abordagem tem se mostrado mais eficiente nos estudos de função gênica em plantas e/ou em terapias moleculares, visando ao silenciamento específico em células animais.

Essa metodologia baseia-se na produção de sequências de amiRNA com 21 nt, os quais podem atuar como silenciadores específicos de um ou de múltiplos genes-alvo (Figura 16.17). Uma vez selecionado o gene-alvo a ser silenciado, sequências complementares reversas serão escolhidas e introduzidas em estruturas de precursores de miRNA endógenos ao organismo de interesse. Quando o precursor (contendo as sequências de amiRNA) for introduzido no organismo, a produção dos amiRNA seguirá a via canônica de processamento de miRNA. Tal mecanismo tem sido amplamente utilizado em plantas para estudos de genética funcional (Figura 16.18), mas também apresenta potencial para uso na agricultura (p. ex., no controle de viroses e regulação da repressão de genes de interesse).

Tabela 16.2 Aplicação da tecnologia de iRNA em plantas.

Cultura	Aplicação	Instituição
Algodão	Redução do nível de gossipol na semente do algodoeiro	Texas University (EUA)
Arroz	Redução do nível de glutelina	National Institute of Agrobiological Sciences (Japão)

Banana	Resistência ao vírus do mosaico das brácteas da bananeira	Queensland University (Austrália)
Café	Redução do nível de cafeína	Nara Institute of Science and Technology (Japão)
Cevada	Resistência ao vírus do nanismo amarelo da cevada	CSIRO Plant Industry (Austrália)
Feijão	Resistência ao vírus do mosaico dourado do feijoeiro	EMBRAPA CENARGEM (Brasil)
Milho	Aumento do nível de lisina e melhora na germinação de sementes	Weizmann Institute (Israel)
Tomate	Aumento dos níveis de carotenoides e flavonoides	Stazione Zoologica (Itália)

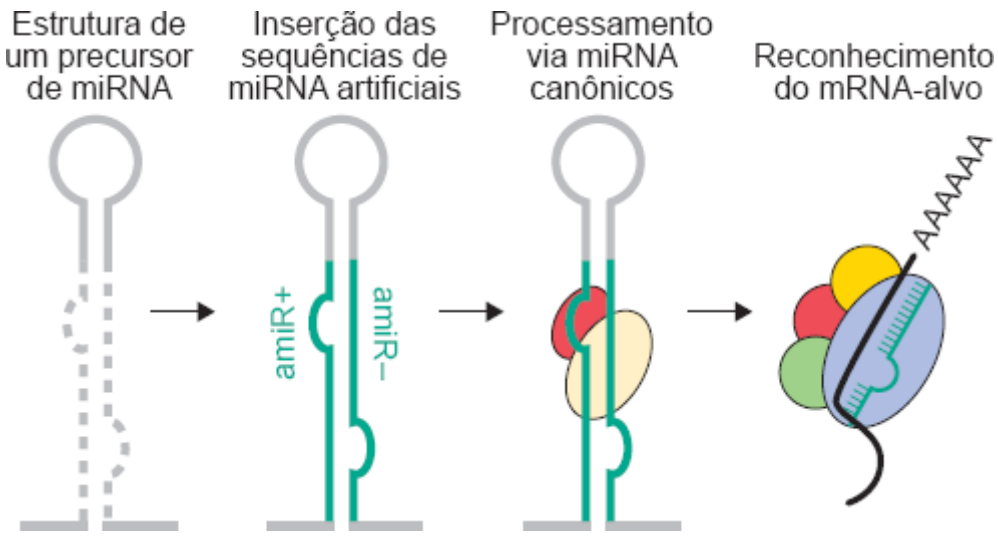


Figura 16.17 Estratégia de produção de miRNA artificiais: amiRNA. Os amiRNA são pequenos RNA de 21 nucleotídeos, os quais podem ser geneticamente modificados e atuar como silenciadores específicos de um ou de múltiplos genes. Uma vez selecionado o gene-alvo a ser silenciado, sequências complementares reversas de 21 nucleotídeos serão escolhidas e introduzidas em estruturas de precursores de miRNA endógenos ao organismo de interesse. Quando o precursor (contendo as sequências de amiRNA) for introduzido no organismo, a produção dos amiRNA seguirá a via canônica de processamento de miRNA.

Mimetização de alvos

Esta estratégia baseia-se em um mecanismo natural de regulação da expressão gênica encontrado em plantas. Por exemplo, há certos miRNA que contêm dois tipos de transcritos-alvo: um com um sítio totalmente complementar à sequência do miRNA, o que leva à clivagem e degradação deste; e um segundo alvo que não pareia perfeitamente com a sequência do miRNA, ocasionando apenas uma hibridização sem clivagem do mRNA. Este segundo mRNA foi denominado mimetizador de alvo canônico ou, do inglês, *target mimicry* (Figura 16.19).

O primeiro caso de *target mimicry* foi descrito em 2007, em *A. thaliana*, onde foi possível observar que certos miRNA continham alvos que eram degradados e outros aos quais estes miRNA apenas se ligavam e lá permaneciam sequestrados. Um exemplo é o caso do transcrito IPS1 (*induced by phosphate starvation1*), que contém um sítio complementar ao miRNA miR399 com uma incompatibilidade (erro) na região central (11-13 nt). Embora isso possibilite que o miR399 se ligue a IPS1, ele é incapaz de clivar este alvo, mas, ao mesmo tempo, miR399 está colapsado a IPS1. Assim, PHO2 (*phosphate 2*), um alvo genuíno do miR399, acaba escapando da supressão por miR399 até mesmo nos casos em que este miRNA está presente em altos níveis. Quando a sequência de IPS1 foi mutada, possibilitando que miR399 se ligasse com perfeita complementaridade, observou-se a degradação de IPS1.

Figura 16.18 Exemplo de silenciamento gênico utilizando miRNA artificiais (amiRNA). Na planta modelo *Arabidopsis thaliana*, foram introduzidos amiRNA visando silenciar o transcrito do gene codificante da proteína indutora da floração (*Flowering locus T* ou FT). À esquerda, está representada uma planta de *A. thaliana* selvagem (WT); à direita, uma linhagem silenciada (amiRNA-FT). Observa-se total inibição na indução do eixo de inflorescência na planta transgênica expressando o amiRNA-FT.

O mecanismo de mimetização de alvos apoia-se na ideia de que quando vários alvos estão presentes, alguns serão favorecidos em detrimento de outros, e tal fenômeno pode ser governado pela força com que o miRNA se liga ao seu alvo.

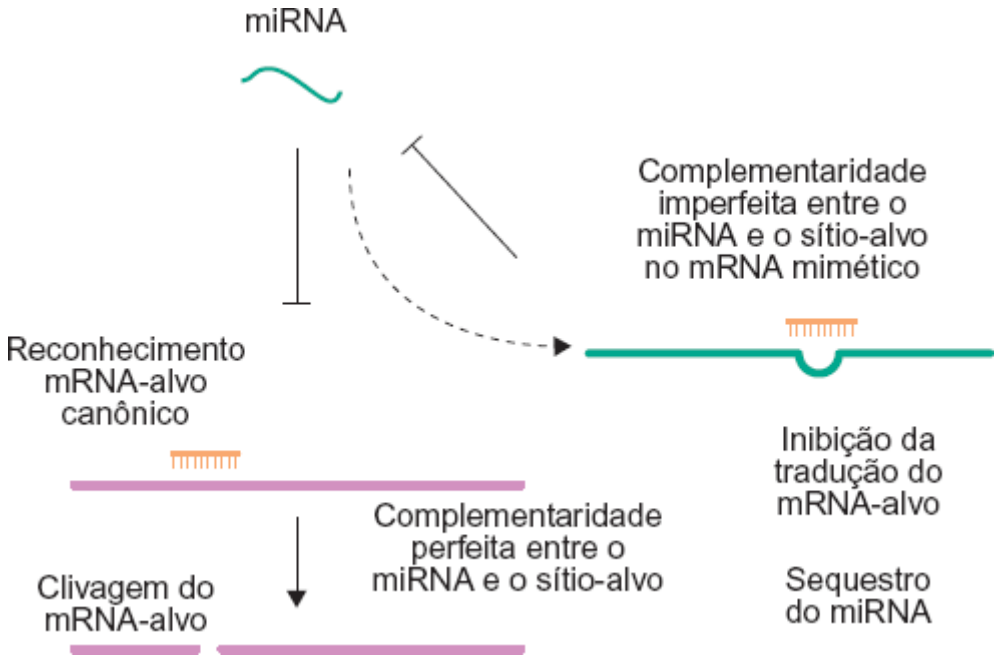


Figura 16.19 Ferramenta de mimetização de um alvo para miRNA: *target mimicry*. Esta metodologia consiste em criar e expressar sequências de RNA que contenham sítios homólogos ao RNA-alvo canônico reconhecidos por um miRNA específico. No entanto, esse RNA mimético irá apresentar alguns polimorfismos nos nucleotídeos correspondentes ao sítio de clivagem, impedindo que o mRNA mimético seja clivado. Consequentemente, esta armadilha de miRNA ocasiona o sequestro e a inativação funcional dos mesmos.

Metodologias de identificação e caracterização de ncRNA

A existência dos pequenos RNA já é conhecida há bastante tempo; no entanto, eles sempre foram desconsiderados em estudos, sendo relacionados com produtos espúrios de degradação do RNA.

Análises mais criteriosas desses possíveis produtos de degradação demonstraram que havia um viés quanto à origem desses pequenos RNA; ou seja, para assumir que um pequeno RNA é um produto de degradação, espera-se que seja encontrada uma distribuição homogênea ao longo do mRNA. No entanto, em vários casos, isso não ocorria, sendo sempre identificada uma região onde havia maior abundância de pequenos RNA.

Somente com a caracterização de pequenos RNA em *C. elegans*, aproximadamente 20 anos atrás, como reguladores de desenvolvimento celular, foi que estratégias e metodologias de estudo começaram a ser adaptadas ou desenvolvidas.

A utilização de sequenciamento de ácidos nucleicos é a maneira mais popular e eficaz de identificar um pequeno RNA. É possível utilizar técnicas convencionais de sequenciamento como a de Sanger, mas com algumas adaptações. Primeiramente, deve-se enriquecer a amostra de RNA a ser sequenciada para a faixa de tamanho dos pequenos RNA. Isso pode ser efetuado por meio de migração da amostra de RNA em gel de poliacrilamida. Em seguida, deve-se concatemerizar os pequenos RNA purificados resultantes, e cloná-los em um vetor plasmidial, que posteriormente poderá ser multiplicado em *Escherichia coli*, purificado e sequenciado. Existem *kits* comerciais que proporcionam esse tipo de análise.

Contudo, essa abordagem é laboriosa, com baixo poder de identificação de pequenos RNA (apenas aqueles mais abundantes serão clonados) e com baixa cobertura (poucos pequenos RNA são identificados por sequenciamento).

Somente com o surgimento de novas tecnologias de sequenciamento, ditas de larga escala, a identificação de um abundante número de novas classes de pequenos RNA foi possível. Independentemente da plataforma de sequenciamento, essas novas gerações de sequenciamento de ácidos nucleicos possibilitaram identificar milhares de sequências. Com isso, a problemática de abundância e cobertura de pequenos RNA foi resolvida.

No entanto, a grande quantidade de dados resultantes do sequenciamento de larga escala necessitou do desenvolvimento de algoritmos e ferramentas de bioinformática específicas para lidar com sequências pequenas e abundantes. Atualmente, já existem várias ferramentas de bioinformática disponíveis para caracterizar pequenos RNA – a maioria destinada à análise de miRNA. Dentre as mais conhecidas, existe o conjunto de ferramentas de bioinformática, UE sRNA *tools*.

A análise de expressão dos pequenos RNA pode auxiliar na sua caracterização funcional. Para isso, é possível utilizar os dados de sequenciamento de larga escala ou técnicas mais usuais de análise de expressão de RNA, tais como microarranjos, *Northern blot*, RT-PCR e RT-qPCR.

Os dados oriundos do sequenciamento de última geração podem ser utilizados para inferir a expressão de pequenos RNA, mesmo aqueles ainda não identificados. Contudo, tais metodologias ainda apresentam custo elevado, o que pode dificultar a montagem de experimentos biológicos que propiciem a obtenção de mais detalhes e informações, diminuindo a robustez da análise de expressão dos pequenos RNA. As metodologias de normalização dos dados de expressão e os possíveis vieses das técnicas de sequenciamento de larga escala seguem evoluindo.

A metodologia de microarranjo, bastante estabelecida na análise de expressão gênica, também pode ser utilizada para quantificação de pequenos RNA. Atualmente, já existem microarranjos para siRNA e miRNA humanos disponíveis. Contudo, esta metodologia também depende de replicatas biológicas, e a analise de expressão é delimitada somente aos pequenos RNA contidos no microarranjo.

A metodologia de *Northern blot*, muito popular na análise de expressão gênica, também é utilizada para inferir sobre a quantificação de pequenos RNA. A partir de uma sonda radioativa ou fluorescente, complementar aos pequenos RNA, obtém-se um sinal proporcional à expressão destes. A técnica é bastante utilizada para analisar a expressão de pequenos RNA como os tRF e miRNA, mas ela exige uma quantidade relativamente grande de RNA para o experimento, e é bastante laboriosa.

Metodologias com base em reações de cadeia de polimerase (PCR) para análise de expressão de modo semiquantitativo (no caso de RT-PCR) ou quantitativo (no caso de RT-qPCR) também são utilizadas para análise da expressão de pequenos RNA. Ambas as técnicas exigem uma etapa anterior, na qual a enzima transcriptase reversa sintetiza a fita complementar reversa ao pequeno RNA. Em razão do tamanho diminuto dos pequenos RNA, em geral, é utilizado um adaptador denominado *stem-loop*. A metodologia de RT-qPCR, também denominada PCR em tempo real, é bastante utilizada para o estudo da expressão gênica de um pequeno RNA específico; tal técnica apresenta boa reprodutibilidade e especificidade.

A análise de localização tecidual de pequenos RNA é realizada pela metodologia de hibridização *in situ,* na qual se utilizam sondas contendo ácidos nucleicos bloqueados (LNA). Essas sondas fluorescentes são complementares aos pequenos RNA de interesse, e emitem um sinal fluorescente na região tecidual em que este é expresso.

Considerações finais e perspectivas

Os avanços tecnológicos promoveram a descoberta de várias classes de pequenos ncRNA regulatórios, os quais estão constantemente revelando um novo mundo associado às moléculas de RNA. Muitos estudos demonstram que os RNA não são apenas moléculas intermediárias entre genes e proteínas, mas também que eles evoluíram, ao lado de DNA e proteínas. Tal evolução levou à formação de diversas classes de RNA, que assumiram formas específicas de acordo com a função desempenhada. Por exemplo, os RNA transportadores (tRNA) adicionam aminoácidos a sítios específicos do RNA mensageiro (mRNA); o RNA ribossômico (rRNA) catalisa a formação da cadeia peptídica; o mRNA serve de molde para a tradução da proteína; e os ncRNA regulatórios estão envolvidos em uma complexa regulação da expressão gênica. Os ncRNA regulatórios exercem suas funções por vários modos especiais de ação, e variam bastante entre diferentes espécies. Por exemplo, em torno de 98% do genoma humano é composto por DNA não codificador, enquanto cerca de apenas 2% de um genoma bacteriano corresponde a DNA não codificador. Com base nesses dados, as regiões não codificadoras dos genomas parecem ser as grandes responsáveis pela diversidade entre os mais variados organismos presentes no planeta.

Os tipos e a quantidade de ncRNA reguladores estão em constante expansão, de acordo com o aumento da complexidade fisiológica e o desenvolvimento de cada espécie. Neste capítulo, foram abordadas as classes mais proeminentes dos pequenos ncRNA. Assim, foram discutidos oito diferentes tipos de RNA, com suas peculiaridades e capacidade de criar uma grande complexidade funcional nas vias onde estão atuando. Adicionalmente, ainda há diferenças dentro das várias classes de ncRNA, o que os possibilita atuar de acordo com os organismos em que ocorrem. Um exemplo são os miRNA – em plantas, eles se ligam (na grande maioria) à região codificadora do seu transcrito-alvo; enquanto, em animais, é mais frequente a ligação na região 3’ não traduzida do mRNA.

A grande capacidade regulatória dos pequenos ncRNA, os quais podem controlar direta ou indiretamente diversos processos celulares, é que despertou o interesse nessas moléculas em estudos de funções gênicas ou até mesmo no desenvolvimento de possíveis aplicações biotecnológicas. As informações acumuladas sobre os meios de

atuação desses pequenos RNA têm levado ao desenvolvimento de diversos fármacos, com base em tecnologia de iRNA, que trata a expressão inapropriada de certos genes (p. ex., oncogenes). Outro ponto é a aplicação desta tecnologia no desenvolvimento de plantas com pequenos ncRNA que tenham por alvos alguns transcritos vitais dos patógenos, ou pragas, ou até mesmo vírus que provoquem riscos para determinada cultura. Uma vez que inúmeros grupos de pesquisas (públicos ou privados) estão se dedicando ao tratamento de várias lesões com o uso de iRNA, talvez, em um futuro não muito distante, esses fármacos cheguem até o mercado. Possivelmente, em poucos anos, estarão sendo comercializadas terapias de RNA-injetável para tratamento de doenças humanas e animais. Adicionalmente, a sociedade terá ao alcance alimentos livres de substâncias alergênicas ou tóxicas. A ampla modulação dos mais diversos genes por mecanismo de siRNA é um claro incentivo de que há muito a aprender sobre esses pequenos ncRNA e sobre os possíveis papéis que eles realizam na função regulatória de genes. A ampliação desse conhecimento terá como consequência um significativo aperfeiçoamento das tecnologias que venham a empregar o mecanismo de siRNA.

Bibliografia

Bartel DP. MicroRNAs: genomics, biogenesis, mechanism and function. Cell. 2004;116:281-97.

Bartel DP. MicroRNAs: target recognition and regulatory functions. Cell. 2009;136:215-33.

Baulcombe D. Unwinding RNA silencing. Science. 2000;290:1108-9.

Carrington JC, Ambros V. Role of microRNAs in plant and animal development. Science. 2003;301:336-8.

Chen X. Small RNAs and Their Roles in Plant Development. Annu Rev Cell Dev Biol. 2009;35:21-44.

Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature. 1998;391:806-11.

Ghildival M, Zamore PD. Small silencing RNAs: an expanding universe. Nature Reviews Genetics. 2009;10:94-108.

Napoli C, Lemieux C, Jorgensen R. Introduction of a chimeric chalcone synthase gene into petunia results in reversible cosupression of homologous genes in trans. Plant Cell. 1990;2:279-89.

Okamura K, Lai EC. Endogenous small interfering RNAs in animals. Nature Reviews. 2008;9:673-8.

Pereira TC. Introdução à técnica de interferência por RNA-RNAi. Sociedade Brasileira de Genética. 2013;1:46-63.

17 Genoma Móvel | Mecanismos de Transposição e Impacto Evolutivo

Érika Maria de Jesus, Danielle Maluf Quintanilha, Nathalia de Setta Costa e Marie-Anne Van Sluys

Os elementos de transposição são sequências de DNA com a habilidade de se movimentar de uma posição para outra no genoma que habitam. Eles apresentam diferentes estruturas, modos de mobilização e histórias evolutivas. Em razão da ampla variedade observada, foi proposta uma classificação desses elementos, a fim de facilitar sua identificação e estudo. A capacidade de mobilização pode gerar mutações e rearranjos cromossômicos, com consequências diversas, adaptativas, neutras ou deletérias. Já foram descritos diversos exemplos de elementos de transposição contribuindo para a funcionalidade dos genomas de eucariotos e procariotos. Para controlar essa capacidade de gerar variabilidade genética, os genomas contêm mecanismos de silenciamento dos elementos de transposição, o que possibilita que genomas e elementos evoluam de maneira análoga a uma relação de mutualismo. Neste capítulo, serão apresentadas as mais diferentes facetas dos elementos de transposição; dentre elas, sua estrutura e classificação, mecanismos de controle pelo genoma, impacto da mobilização e padrões evolutivos.

Introdução

Os seres vivos transmitem suas características para as próximas gerações por meio da informação genética, mantendo a integridade das espécies. Em função disso, pode-se imaginar que os cromossomos são estruturas estáticas e imutáveis, de modo a manter as características que são passadas para a geração seguinte. De fato, por muito tempo, foi o que se imaginou. No entanto, conforme as descobertas científicas foram avançando, foi possível notar que havia genes capazes de se mover de uma posição a outra dentro do genoma de uma célula. Tais genes, ou entidades genéticas, foram então denominados elementos de transposição (TE, do inglês, *transposable elements*).

Os TE estão presentes em praticamente todos os seres vivos, incluindo bactérias, algas, fungos, plantas e animais. Uma das únicas exceções conhecidas é o protista parasita *Plasmodium falciparum*, no qual, em seu genoma, não foram encontradas sequências relacionadas aos TE. Sua ampla distribuição é uma evidência de que se tratam de elementos muito antigos associados aos genomas; eles podem constituir uma porcentagem considerável dos genomas. A porcentagem de TE presente nos genomas pode variar bastante de espécie para espécie, assim como a proporção dos diferentes tipos de TE (transposons e retrotransposons, cujas características serão abordadas no item *Existem diversos tipos de TE*, a seguir; Figura 17.1). No ser humano, por exemplo, aproximadamente 45% do genoma é composto de TE, com predomínio dos retrotransposons. Em algumas gramíneas, como o milho, essa porcentagem pode chegar a surpreendentes 80%, também com predominância dos retrotransposons. Por outro lado, há organismos com menor quantidade, como é o caso da planta *Arabidopsis thaliana*, cujo conteúdo de TE no genoma é de apenas 14%. O nematódeo *Caenorhabditis elegans* também apresenta um conteúdo de TE relativamente pequeno em seu genoma, de apenas 13%, e com predomínio dos transposons.

A descoberta desses elementos ocorreu no início da década de 1940, quando a pesquisadora Barbara McClintock observou em seus estudos uma correlação entre a movimentação de um *locus* do genoma, que gerava quebras cromossômicas, e a origem de mutações espontâneas e instáveis em milho. Devido à propriedade de alterar

características fenotípicas da planta, modificando o funcionamento dos genes aos quais os TE estão associados, ela os denominou, inicialmente, de "elementos controladores dos genes". Apesar de o seu trabalho ter sido recebido com grande ceticismo pela comunidade científica na época, com o passar dos anos, várias descobertas foram corroborando as observações de McClintock. Em 1983, finalmente, veio o reconhecimento, e ela foi laureada com o Prêmio Nobel de Fisiologia e Medicina pela descoberta dos TE em seus estudos com o milho.

As tecnologias atuais possibilitam o sequenciamento de genomas inteiros muito rapidamente e com baixo custo. Com a sequência completa de um genoma em mãos, é possível identificar praticamente a totalidade dos TE ali presentes. Mobiloma é o termo recentemente criado para designar o conjunto dos TE presentes em um genoma. No entanto, os primeiros TE descritos foram descobertos por originarem novos fenótipos, ou seja, alguma característica mutante. A investigação sobre a causa do aparecimento dessas mutações naturais acabou levando à identificação dos TE inicialmente descritos. Um exemplo é o elemento *Tnt1*, retrotransposon que foi descoberto quando pesquisadores estudavam células de tabaco em cultura, deficientes para a enzima nitrato redutase. Ao sequenciarem o gene que codifica a nitrato redutase em mutantes, descobriram que havia uma cópia de *Tnt1* em seu interior. Como o gene estava interrompido pela inserção do TE, a enzima não era produzida.

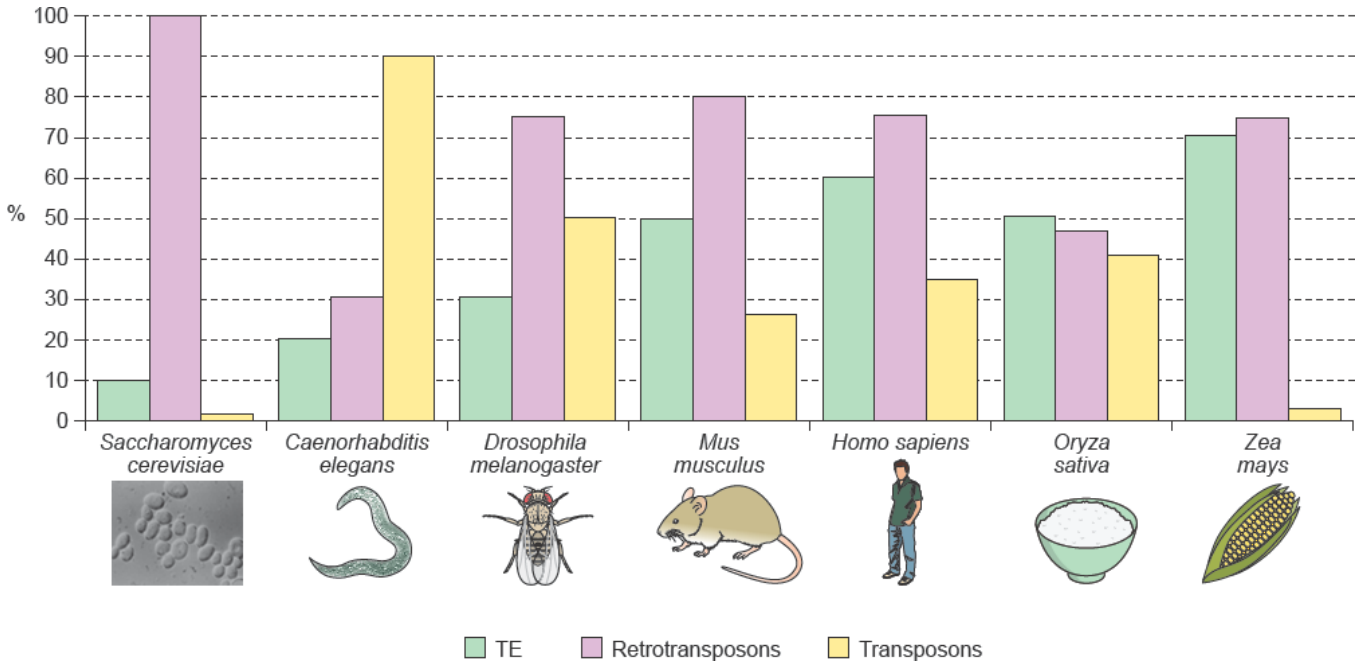


Figura 17.1 Porcentagem total de TE e porcentagem de cada classe de TE nos genomas de levedura (*Saccharomyces cerevisiae*), verme (*Caenorhabditis elegans*), mosca-da-fruta (*Drosophila melanogaster*), camundongo (*Mus musculus*), homem (*Homo sapiens*), arroz (*Oryza sativa*) e milho (*Zea mays*). Adaptada de Feschotte e Pritham, 2007; Chénais *et al*., 2012.

O elemento *Ac*, descrito por McClintock, foi descoberto graças às quebras cromossômicas que geravam nas células do milho, resultando em fenótipos alterados de coloração dos grãos. Essas quebras ocorriam no braço curto do cromossomo 9, uma região que continha genes cujas características poderiam ser facilmente observadas nos grãos. Um dos genes dessa região cromossômica tem papel na rota de síntese do pigmento aleurona, conferindo ao grão de milho a coloração avermelhada. Uma mesma espiga podia apresentar grãos vermelhos (devido à presença do pigmento), grãos totalmente brancos (sem o pigmento) ou grãos com pigmentação na forma de mosaico, mesclando setores vermelhos e brancos (fenótipo variegado). McClintock acabou constatando que os grãos brancos tinham o elemento interrompendo o gene da rota de síntese do pigmento, enquanto os grãos vermelhos tinham o gene intacto. Os grãos variegados, por sua vez, tinham o elemento inserido no gene inicialmente, gerando células sem a pigmentação. Na medida em que o grão ia se desenvolvendo, o elemento se excisava do gene em apenas algumas células, que tinham sua função restaurada. As células originadas a partir destas geravam os setores vermelhos em mosaico do grão (Figura 17.2). Ao investigar o surgimento dos fenótipos, McClintock propôs a movimentação de um *locus* dentro dos cromossomos do milho. De acordo com a própria pesquisadora, essas observações somente poderiam ter ocorrido de maneira acidental, pois se tratava de uma descoberta visionária para a época.

Este capítulo visa apresentar os TE, discutindo os tipos de elementos existentes, seu modo de mobilização e evolução. Além disso, serão abordados temas como as consequências da presença e atividade desses elementos nos genomas, assim como as tecnologias desenvolvidas a partir dos TE, dando exemplos de como esses elementos podem ser utilizados como ferramentas biotecnológicas.

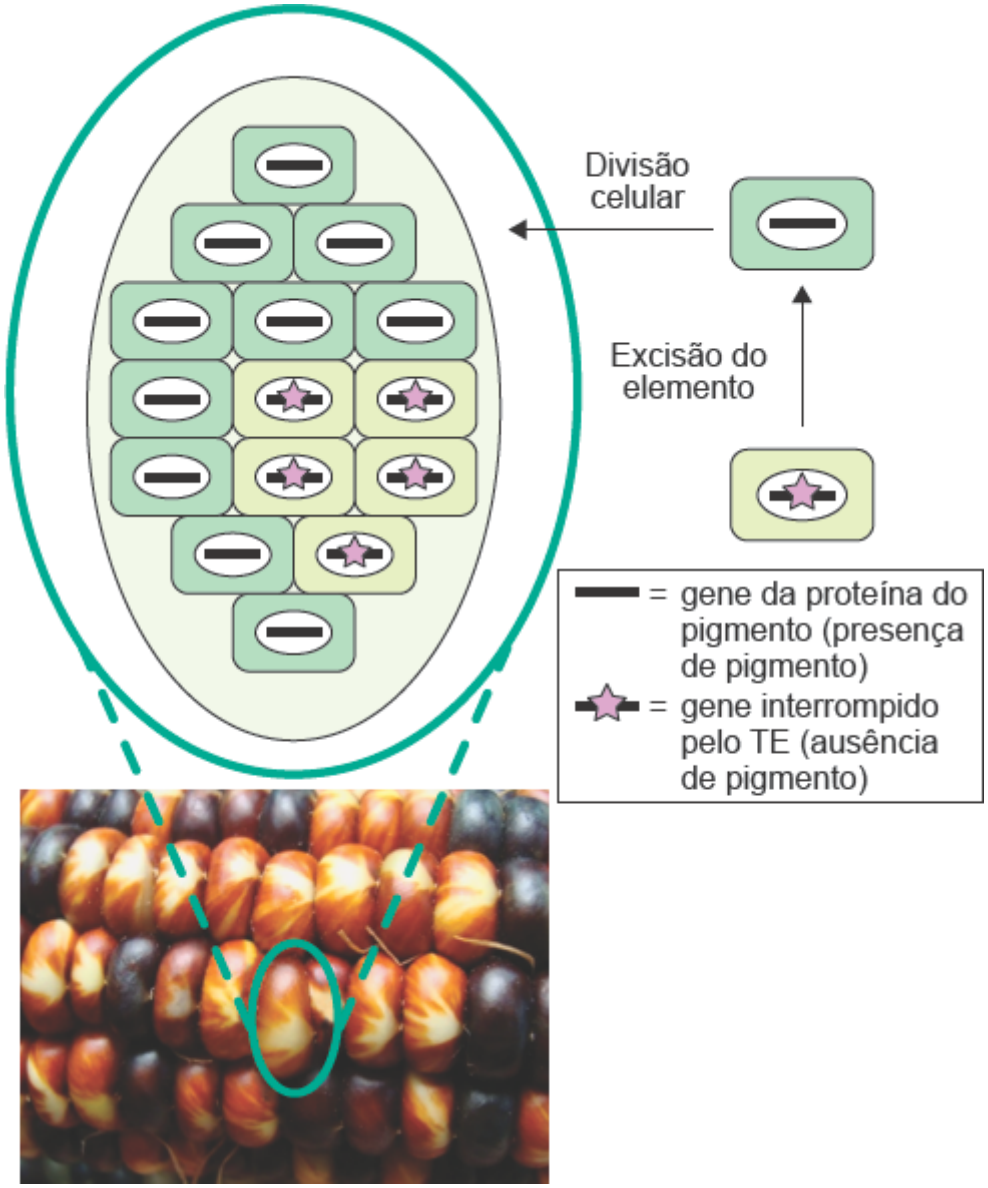


Figura 17.2 Mecanismo da formação de fenótipo variegado. A inserção de um TE em um gene da rota de síntese do pigmento leva ao fenótipo mutante sem pigmento. A excisão do TE do alelo mutado implica na reversão para o fenótipo selvagem, com pigmentação. A excisão precoce no desenvolvimento do grão dá origem à formação de um setor grande de fenótipo revertente, enquanto a reversão tardia provoca a formação de pontos pequenos. A ocorrência da excisão em células germinais leva à formação de um grão revertente por completo. Adaptada de Feschotte *et al.*, 2002.

Existem diversos tipos de TE

Os TE apresentam grande diversidade em sua estrutura. Uma de suas características é que carregam, em sua sequência nucleotídica, genes que codificam as proteínas necessárias à sua transposição. A maioria deles apresenta também terminações repetidas, as quais podem ser diretas ou invertidas, contendo as regiões promotoras completas ou parciais do elemento. São regiões de reconhecimento importantes no momento da transposição. Ao se inserirem em um determinado sítio do genoma, os TE geram uma duplicação do sítio de inserção (TSD, do inglês, *target site duplication*), que fica como uma cicatriz que permanece mesmo que o elemento se excise (Figura 17.3).

Os TE estão presentes nos genomas como elementos autônomos e não autônomos. Enquanto os autônomos têm a capacidade de codificar enzimas necessárias para se mobilizarem, os não autônomos são defectivos, ou seja, apresentam mutações na sua região codificante e, consequentemente, suas enzimas não são funcionais. No entanto,

um elemento não autônomo pode ser mobilizado pela ação *in trans* da enzima produzida por um elemento autônomo, por meio do reconhecimento de suas terminações repetidas.

Os TE também podem ser classificados em superfamílias, e estas subdivididas em famílias. São considerados da mesma família os TE que compartilharem suas terminações repetidas, ou seja, que a enzima que promover a mobilização de um elemento possa reconhecer e mobilizar outro. Famílias também podem ser classificadas pela regra 80-80-80, em que são considerados de uma mesma família TE maiores que 80 pb, que compartilharem ao menos 80% de identidade de nucleotídeos, em ao menos 80% da sua sequência. O grande número de famílias de TE já descritas na literatura revela sua extensa diversidade estrutural.

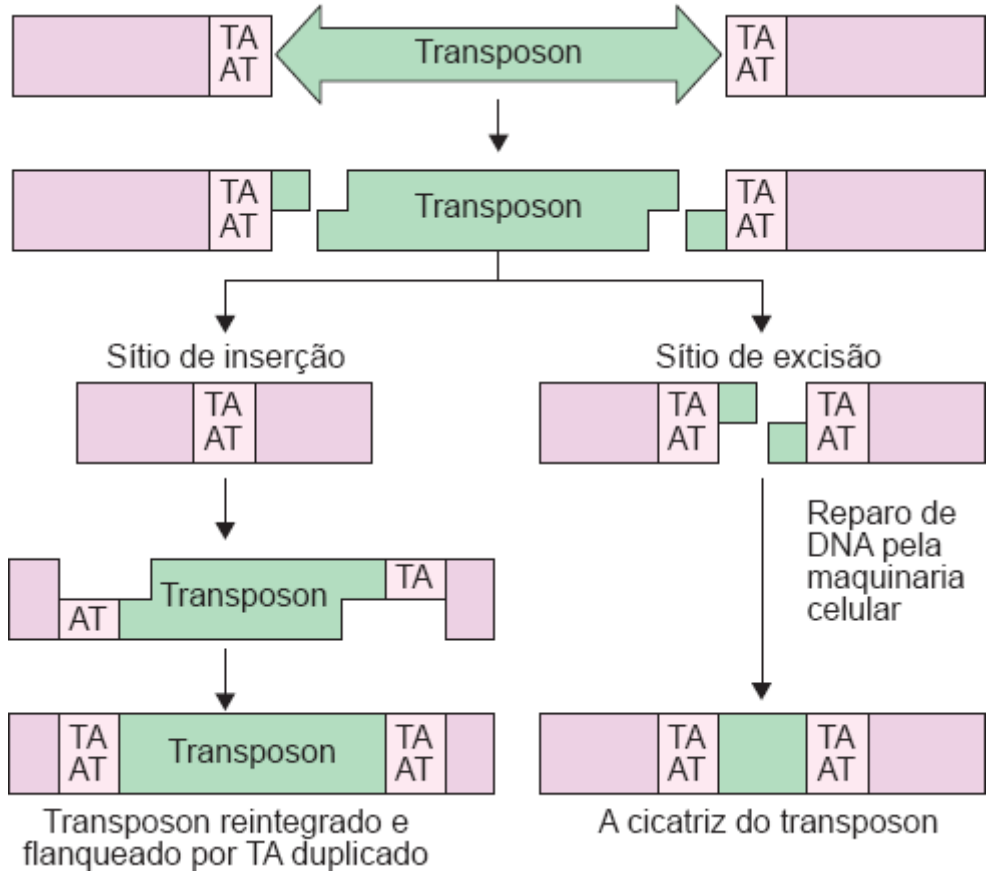


Figura 17.3 Duplicação do sítio de inserção de um transposon e a cicatriz deixada no sítio de excisão. O transposon é excisado de um sítio e reinserido em outro sítio-alvo que contém, por exemplo, a sequência dinucleotídica TA. Quando o transposon se insere em um novo sítio, o dinucleotídeo TA é duplicado, o que leva à duplicação do sítio-alvo (TSD). A maquinaria celular repara o sítio de excisão do transposon. Se esse reparo for feito por junção de extremidades não homólogas (NHEJ, do inglês, n*onhomologous end-joining*), uma cicatriz é formada. Adaptada de Muñoz-López e García-Pérez, 2010.

Os TE são divididos em duas grandes classes, de acordo com seu modo de transposição: os de classe I ou retroelementos, que transpõem-se por meio de um RNA intermediário, usado como molde. Já os de classe II, os transposons de DNA (ou somente transposons), movem-se diretamente a partir de uma molécula de DNA.

Retroelementos têm intermediários de RNA e dependem de transcriptase reversa

O grupo dos retroelementos é composto por retrovírus, retrotransposons com LTR (do inglês, *long terminal repeats*) e retrotransposons sem LTR. Todos esses elementos têm em comum a característica de se mobilizarem por meio da transcrição de um mRNA, que é então reversotranscrito em uma molécula de cDNA pela enzima transcriptase reversa. Essa molécula, que é uma cópia do elemento original que foi transcrito, se insere em um novo sítio do genoma, assim completando a retrotransposição.

Elementos de classe I se transpõem por meio de um mecanismo de "copia e cola"

Um fenômeno importante que ocorre tanto para os retrotransposons com LTR quanto para os sem LTR é que, a cada ciclo de transposição, novas cópias do elemento são sintetizadas e se inserem no genoma. Por isso, é comum referir-

se a esse tipo de transposição como “copia e cola”. Esse ciclo é replicativo, pois aumenta o número de cópias do elemento. A principal diferença na transposição dos retrotransposons com LTR e sem LTR é que, no primeiro caso, a cópia de cDNA é sintetizada dentro de um capsídeo no citoplasma e, no segundo caso, dentro do núcleo, com o complexo RNA-proteína já ligado ao DNA.

Os retrotransposons com LTR são transcritos pela maquinaria celular em um único mRNA policistrônico que contém duas ORF (do inglês, *open reading frames*) – uma que codifica a proteína GAG e outra que codifica uma poliproteína com os demais domínios proteicos (POL). Uma parte dos transcritos é direcionada para os ribossomos, nos quais são traduzidos, dando origem às proteínas GAG e às poliproteínas; dentre elas, uma transcriptase reversa, uma RNAse H, uma protease e uma integrase. A protease é responsável pela autoclivagem da poliproteína, liberando as proteínas funcionais que estão prontas para realizar a transposição. Parte dos transcritos é empacotada junto com transcriptase reversa, a RNase H e a integrase dentro do capsídeo, formado pelas proteínas codificadas pela GAG. Dentro do capsídeo, os transcritos servirão de molde para a transcriptase reversa e a RNase H produzirem novas cópias do elemento, as quais serão inseridas no genoma da célula hospedeira pela ação da integrase (Figura 17.4).

O tamanho desses elementos pode variar de centenas de pares de base até dezenas de quilobases. As LTR são diretamente repetidas, ou seja, são iguais em ambas as extremidades do elemento, e podem variar de centenas de pares de base (pb) até mais de 5 mil pares de base (Kpb). Elas não codificam proteínas, mas contêm as regiões promotoras e terminadoras associadas à transcrição do mensageiro. As regiões promotoras presentes em ambas LTR são funcionais. Cada LTR se divide em três partes: U3 (do inglês, *unique 3’*), R (do inglês, *repeated RNA*) e U5 (do inglês, *unique 5’*). Logo após a LTR 5’ há o PBS (do inglês, *primer binding site*) e, logo antes da LTR 3’, há o PPT (do inglês, *polypurine tract*). A transcrição do elemento ocorre a partir da região R da LTR 5’ e vai até R da LTR 3’ (Figura 17.5).

Se o mRNA intermediário vai de R a R, como a nova cópia de cDNA sintetizada pela transcriptase reversa contém as regiões U3 na extremidade 5’ e U5 na extremidade 3’, de modo a compor o elemento completo que vai se inserir em um novo sítio? A resposta está exatamente no modo como a transcrição reversa ocorre (Figura 17.6):

- Um tRNA se liga ao PBS do mRNA, atuando como um iniciador para a síntese do cDNA pela transcriptase reversa
- A síntese ocorre no sentido 5’ para 3’, de modo que o PBS e as regiões U5 e R da LTR 5’ do mRNA são reversotranscritas
- Esse pedaço de cDNA já sintetizado se desassocia da porção 5’ do mRNA e se liga na porção 3’, sobre a região R. Essa ligação é possível em função da complementaridade de bases, uma vez que a região R está presente nas duas extremidades do mRNA
- A partir daí, a síntese prossegue, iniciando na região U3 adjacente ao R da extremidade 3’ e avançando até o R da extremidade 5’
- Em seguida, a RNase H degrada o mRNA que serviu de molde para a síntese do cDNA, deixando apenas um fragmento correspondente ao PPT, que vai servir de iniciador para a síntese da segunda fita de DNA
- A transcriptase reversa sintetiza, então, a segunda fita de DNA correspondente à LTR 3’ completa
- Esse fragmento de segunda fita se move, ligando-se à LTR 5’
- A partir daí, a transcriptase reversa promove a síntese na direção 5’ para 3’ em ambas as fitas, de modo a completar a dupla-fita de DNA contendo o retrotransposon em sua sequência integral
- A enzima integrase insere a cópia recém-sintetizada em um novo sítio do genoma.

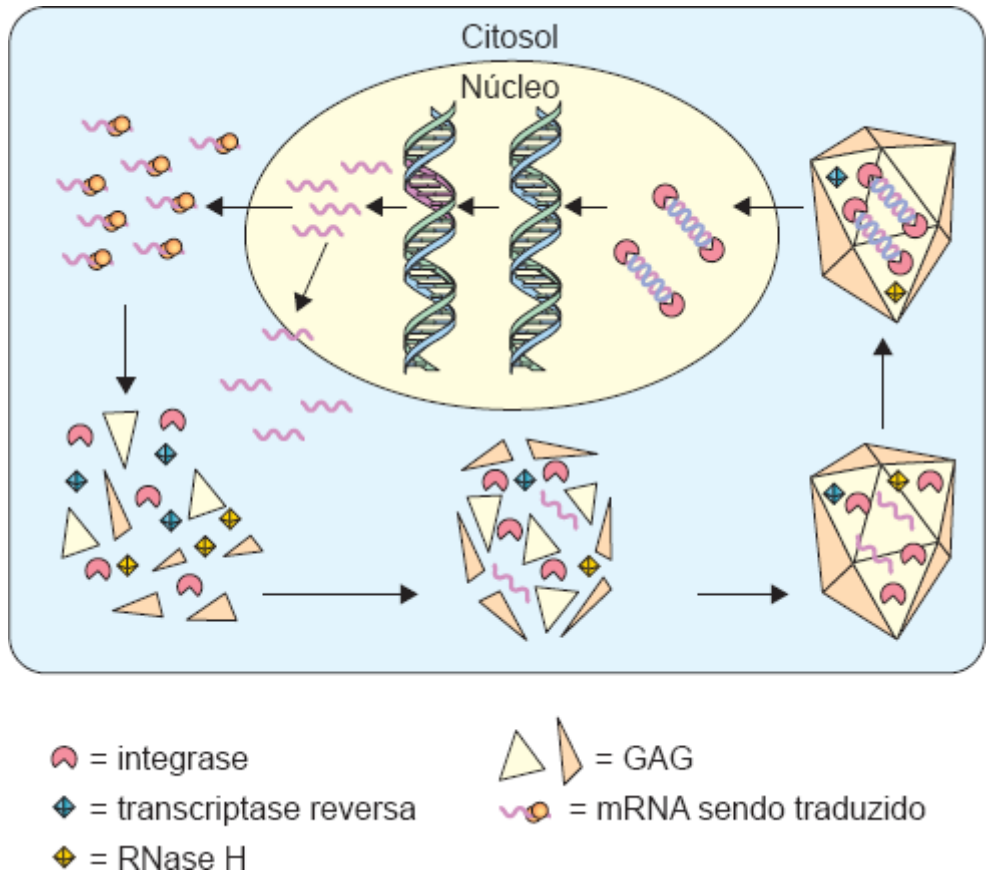


Figura 17.4 Replicação dos retrotransposons com LTR. Adaptada de Sabot e Schulman, 2006.

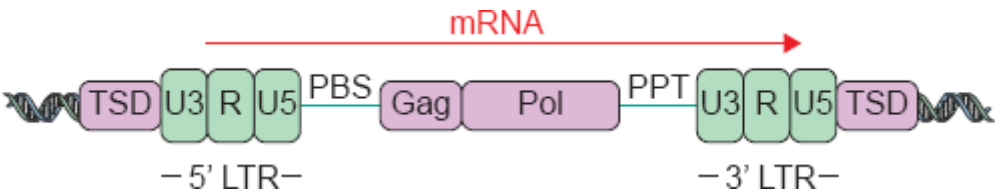


Figura 17.5 Estrutura de um retroelemento com LTR.

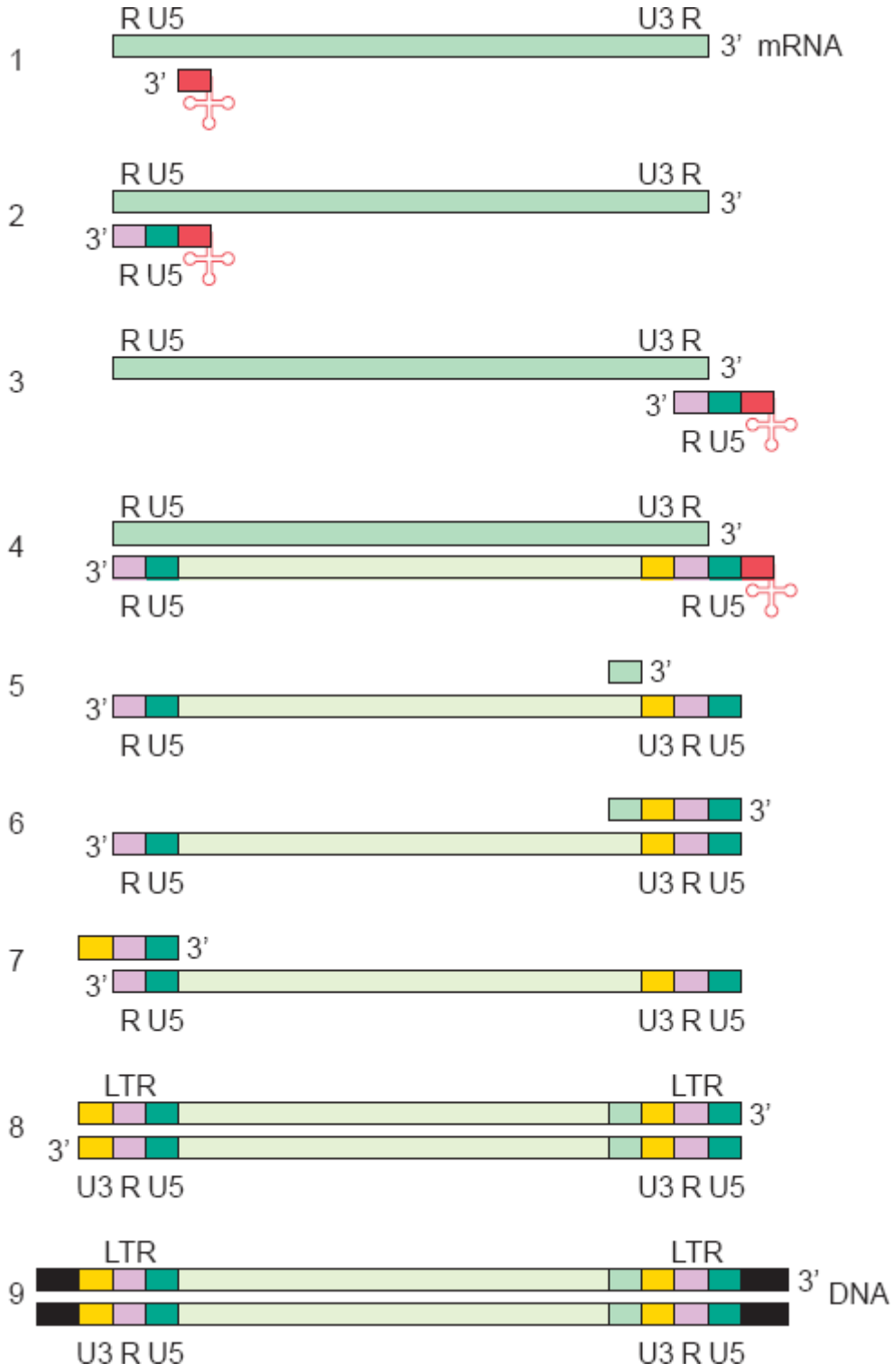


Figura 17.6 Síntese da cópia de DNA para retrotransposons com LTR. Adaptada de Finnegan, 2012.

A replicação dos retrotransposons sem LTR também se inicia com a transcrição do elemento em um mRNA. Parte dos transcritos se liga aos ribossomos, nos quais são traduzidos em proteínas que promovem a replicação, incluindo a transcriptase reversa e a endonuclease. A outra parte dos transcritos servirá de molde para a transcrição reversa. Ainda no citoplasma, as proteínas se ligam aos transcritos, formando um complexo RNA-proteína. Esse complexo é direcionado para o núcleo da célula e se liga ao DNA. A transcriptase reversa sintetiza a nova cópia do elemento, que vai ser inserida no genoma (Figura 17.7). A síntese da nova cópia depende do complexo formado pelo mRNA intermediário (transcritos) e pelas enzimas transcriptase reversa e endonuclease. Esse complexo se associa ao DNA, em um sítio rico em A/T, onde o TE será inserido. A endonuclease promove a quebra de uma única fita do DNA cromossômico, em uma região adjacente a uma sequência de timinas (Figura 17.8 A). A cauda poli-A do transcrito se liga a essas timinas por complementaridade de bases (Figura 17.8 B). A transcriptase reversa inicia a transcrição da primeira fita de cDNA, usando a extremidade 3' da poli-T como iniciador (Figura 17.8 C). A endonuclease quebra então a fita oposta de DNA, a alguns nucleotídeos de distância da primeira quebra (Figura 17.8

D). O mRNA que serviu de molde para a síntese do cDNA é degradado pela RNase H (Figura 17.8 E), que deixa apenas uma ponta 3' livre para servir de iniciador para a síntese da segunda fita de DNA (Figura 17.8 F). As enzimas de reparo da célula se incumbem de realizar a finalização da integração da nova cópia (Figura 17.8 G).

Retrotransposons com LTR

Os retrotransposons com LTR subdividem-se em duas principais superfamílias, Gypsy e Copia, também denominados Ty3/gypsy e Ty1/copia. Esses nomes se devem aos primeiros elementos descritos para cada uma destas superfamílias: *copia* e *gypsy* de *Drosophila melanogaster* e os elementos *Ty1* e *Ty3* de levedura. Essas duas superfamílias são distinguíveis pela organização interna de seus domínios proteicos da transcriptase reversa, da RNase H e da Integrase. Uma terceira superfamília dentro dos retrotransposons com LTR é Bel-Pao. Apresenta a estrutura e a ordem dos domínios proteicos de modo semelhante à superfamília Gypsy. No entanto, é classificada como uma superfamília à parte, visto que, em uma filogenia com base em sua transcriptase reversa, agrupa-se em um clado distinto das outras duas superfamílias. Acredita-se que os retrovírus estão evolutivamente relacionados aos retrotransposons da superfamília Gypsy, pois apresentam estrutura e domínios proteicos semelhantes. No entanto, os retrovírus têm um domínio que codifica proteínas do envelope viral (env), que lhes confere a propriedade de infecção intercelular. Contudo, como toda regra tem exceções, existem elementos Gypsy com o domínio env em insetos e plantas. A Figura 17.9 apresenta as superfamílias de retrotransposons e as respectivas estruturas.

No que diz respeito à localização cromossômica, os retrotransposons com LTR apresentam uma distribuição bastante heterogênea. Na cana-de-açúcar, por exemplo, já foram descritas famílias de elementos com distribuições distintas, podendo ter tendência de localização centromérica, pericentromérica ou dispersa ao longo dos cromossomos.

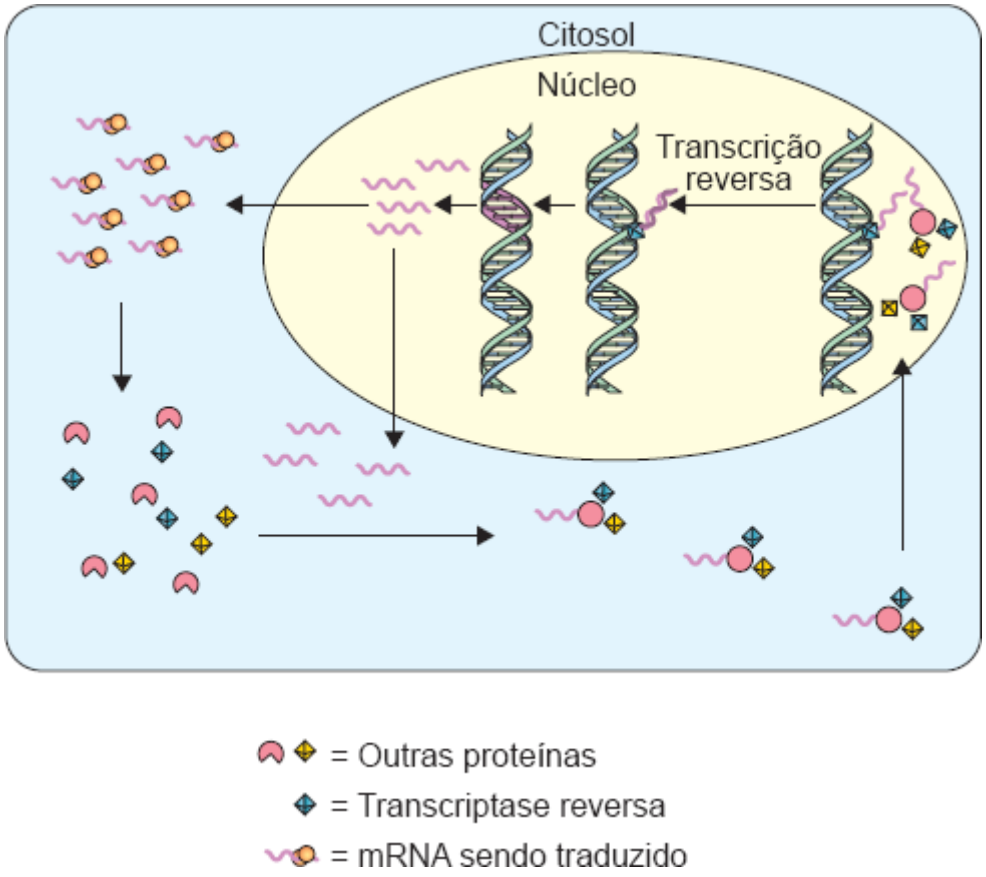


Figura 17.7 Replicação dos retrotransposons sem LTR. Adaptada de Schaack *et al.*, 2010.

DIR

Os DIR apresentam em suas extremidades LTR invertidas e, em vez da integrase, têm um domínio proteico tirosina recombinase. Com isso, pode-se deduzir que seu modo de integração é distinto dos outros retrotransposons com LTR. Em decorrência de um mecanismo diferenciado de integração no genoma, não geram TSD nos sítios onde se inserem. Apesar de características tão peculiares, os DIR são classificados como elementos de classe 1, em função

de sua transcriptase reversa. Outros elementos que também apresentam um domínio tirosina recombinase, em vez de uma integrase, são Ngaro e VIPER. Eles são semelhantes aos DIR em sua estrutura, mas apresentam duas sequências repetidas flanqueando suas LTR 3' (Figura 17.9 B).

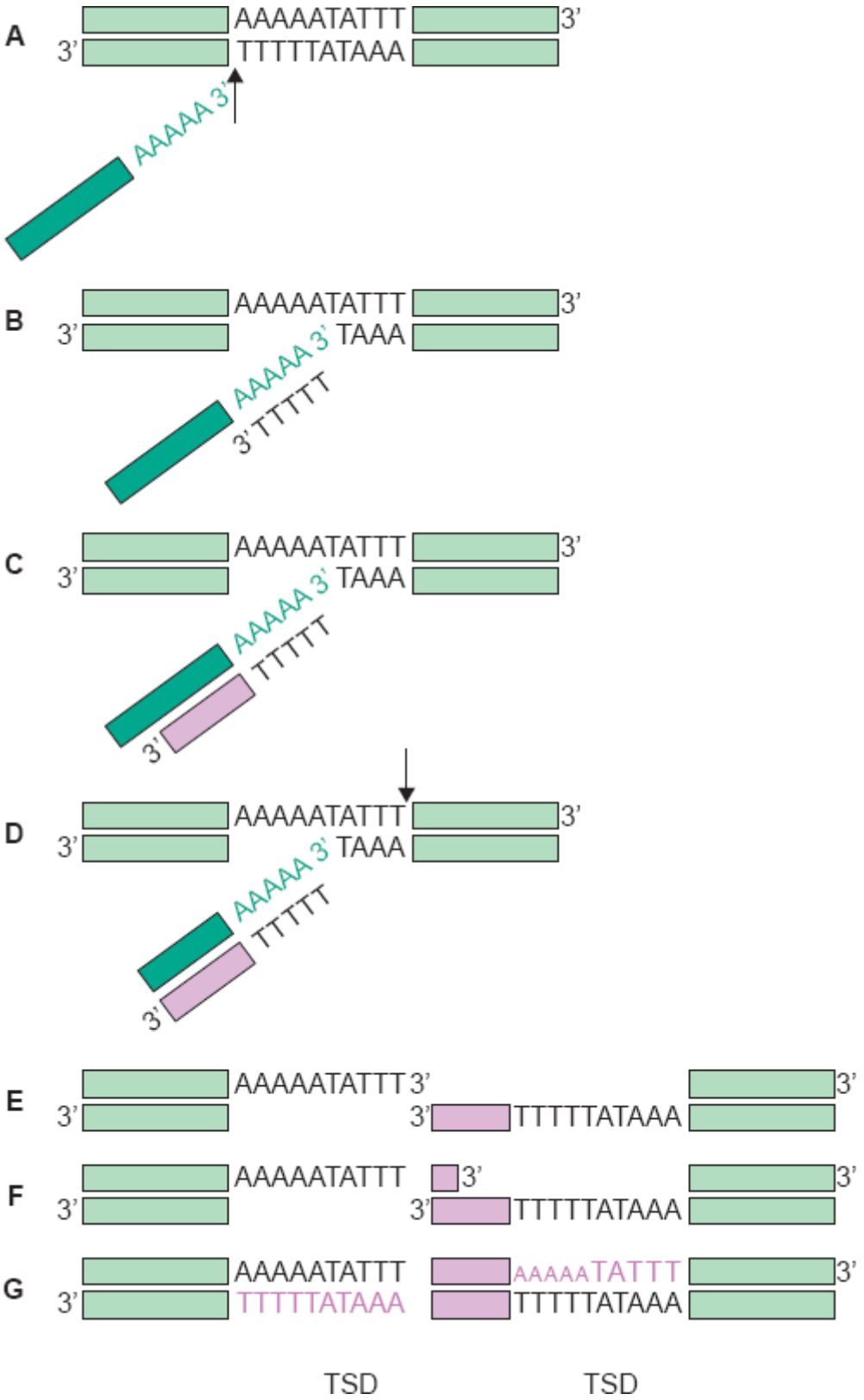


Figura 17.8 A a **G.** Síntese da cópia de DNA e inserção no genoma para retrotransposons sem LTR. As setas pretas indicam as regiões de quebra da fita de DNA. Adaptada de Finnegan, 2012.

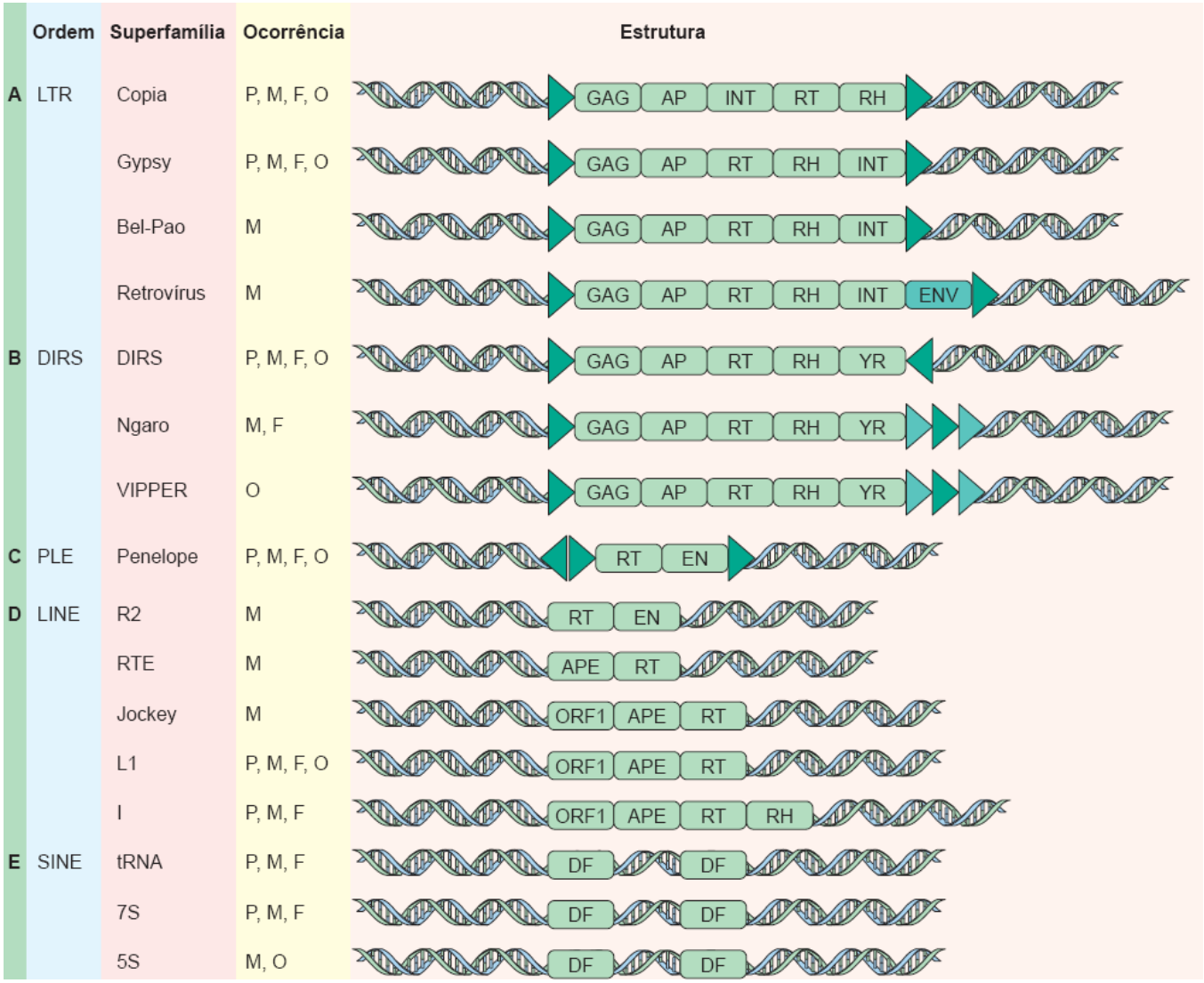


Figura 17.9 A a **E.** Distribuição, classificação e estrutura dos retrotransposons. Adaptada de Wicker *et al.*, 2007.

Penelopes

Penelope é uma superfamília de retrotransposons com LTR, descritos primeiramente em *Drosophila virilis*. Até o presente momento, foram encontrados apenas elementos transposicionalmente ativos desta superfamília nessa espécie. Os Penelopes são responsáveis pela ocorrência da disgenesia híbrida em *D. virilis*. Trata-se de uma síndrome caracterizada pela malformação das gônadas, consequente esterilidade e redução do *fitness*, devido a diversas alterações cromossômicas causadas, na maioria das vezes, por TE.

O elemento *Penelope* e os demais PLE (do inglês, *Penelope-like elements*) apresentam em sua região codificante um domínio da transcriptase reversa e um domínio endonuclease do tipo GIY-YIG, não encontrado em outros retrotransposons (ver Figura 17.9 C). Acredita-se que esse domínio endonuclease atua com a função de integrase, responsável pela inserção do elemento em um novo sítio do genoma. Sua transcriptase reversa também é peculiar, e apresenta similaridade maior com telomerases do que com as transcriptases reversas de outros retrotransposons. Alguns PLE têm a capacidade de reter introns, o que é incomum em elementos que se transpõe por uma molécula de mRNA, uma vez que esses transcritos são processados. Todas as características já citadas, junto ao fato de que os PLE não se agrupam filogeneticamente nem aos elementos de classe I nem aos de classe II, fazem desses elementos uma provável nova superfamília dentro dos retroelementos eucarióticos.

LINE são retrotransposons sem LTR, abundantes no genoma humano

Os LINE (do inglês, *long interspersed elements*) dividem-se em cinco principais superfamílias: R2, RTE, Jockey, L1 e I. Um LINE autônomo codifica pelo menos uma transcriptase reversa e uma proteína com função de endonuclease, como é o caso das superfamílias R2 e RTE. Jockey e L1, além desses domínios já citados, apresentam uma ORF adicional, ORF1, que é similar ao domínio GAG dos retrotransposons com LTR, mas cuja função

continua indeterminada. A superfamília I contém ainda um domínio da RNase H (ver Figura 17.9 D). Esses elementos geralmente apresentam em sua extremidade 3' uma cauda poli-A. Estudos evolutivos sugerem que os LINE são os mais antigos retrotransposons, e que os retrotransposons com LTR originaram-se da aquisição de LTR por um LINE.

Os LINE predominam nos genomas animais, principalmente em humanos. O elemento *L1* é o principal LINE presente no genoma humano, representado por cerca de meio milhão de cópias. No entanto, a maioria é incompleta e estima-se que apenas em torno de 100 cópias sejam ativas no que diz respeito à transposição. A transcrição de *L1* ocorre a partir de um promotor atípico para a RNA polimerase II, de modo que a região promotora também é transcrita. Dessa maneira, a região promotora, presente nos transcritos, pode ser transmitida para as novas cópias que vão se inserir em outros sítios do genoma, preservando a capacidade de *L1* de se mobilizar. Há evidências de que essa família é a única que mantém a capacidade de se transpor no genoma humano, gerando novas inserções. Já foi documentado que *L1* pode se transpor em alta frequência em neurônios de vertebrados, gerando polimorfismos nas células do sistema nervoso central.

SINE são as versões não autônomas dos LINE

Os SINE (do inglês, *short interspersed elements*) são sequências pequenas, de 100 a 300 pb, que não codificam as proteínas necessárias para a transposição. Todos os SINE descritos até o momento são caracterizados por apresentarem um promotor interno da RNA polimerase III, o que lhes confere a habilidade de serem eficientemente transcritos pela maquinaria enzimática do hospedeiro. Como elementos não autônomos, sua transcrição reversa e a integração no genoma dependem da atividade das enzimas codificadas pelos seus respectivos elementos autônomos, os LINE.

Os SINE podem ser divididos em três superfamílias, determinadas pelo tipo de promotor da RNA polimerase III presente em sua sequência. Essas superfamílias são tRNA, 7S e 5S, e os seus nomes indicam também a origem do SINE (ver Figura 17.9 E). Esses elementos estão amplamente distribuídos nos genomas de mamíferos. Em seres humanos, aproximadamente 13% do genoma é composto por SINE. Junto com o LINE autônomo *L1*, os SINE *Alu* e *SVA* são os TE mais abundantes no genoma humano. Os elementos *Alu* apresentam tamanho aproximado de 300 pb, são transcritos pela RNA polimerase III e não têm capacidade codificante, dependendo de *L1* para sua transposição. Os *Alu* são considerados SINE bem-sucedidos evolutivamente devido ao seu alto número de cópias no genoma humano, que é de aproximadamente 1 milhão, o dobro do número de cópias de *L1*. Já foi documentado que os *Alu* são responsáveis pela maior parte das doenças causadas por mutações em humanos. Já os *SVA* (SINE/VNTR/Alu) são elementos peculiares, cujas sequências são constituídas na sua extremidade 5' da repetição de motivos CCCTCT, seguida por duas sequências parciais de *Alu* em orientação reversa, e por um número variável de sequências *VNTR* (do inglês, *variable number tandem repeat*) repetidas em *tandem*, e sequências do retrovírus humano endógeno-K (HERV-K) em sua porção 3'. A classificação dos *SVA* como SINE é controversa tanto pelo seu tamanho anormalmente grande de 4 Kpb, quanto pelo seu mecanismo de amplificação ainda não elucidado. No entanto, eles continuam a ser classificados desta maneira em razão de sua dependência do elemento *L1* para realizarem sua transposição.

Os transposons se mobilizam diretamente por meio da molécula de DNA

Os elementos de classe II, também conhecidos como transposons de DNA, se transpõem diretamente por meio de uma molécula de DNA, não havendo síntese de RNA intermediário. Esse mecanismo de transposição é chamado de "recorta e cola". O transposon é transcrito pela maquinaria celular e seus transcritos são direcionados para os ribossomos no citoplasma. Ali são traduzidos, dando origem às enzimas conhecidas como transposases. Estas são importadas para o núcleo, onde reconhecem os elementos da mesma família, ligam-se a esses e promovem a quebra da dupla-fita de DNA, liberando-os. O elemento é então inserido em um novo sítio do genoma. Portanto, pode-se concluir que esse modo de transposição não aumenta necessariamente o número de cópias do elemento a cada ciclo, como ocorre com os retrotransposons (Figura 17.10). No entanto, o número de cópias do transposon pode aumentar se ele se mobilizar, logo após a passagem da forquilha de replicação do DNA para um sítio logo antes da forquilha, de modo que a nova fita de DNA sintetizada vai apresentar as duas cópias do elemento.

A transposição desses elementos depende do reconhecimento das TIR (do inglês, *terminal inverted repeats*) e de repetições subterminais pela transposase. Ao se inserirem, os transposons causam a duplicação do sítio de inserção,

gerando TSD. O centro catalítico da transposase, responsável por essa atividade, é caracterizado por um domínio proteico que apresenta a tríade de aminoácidos DDE ou DDD. O domínio DDE/D já foi identificado em todas as superfamílias de transposons da ordem TIR, o que sugere uma origem evolutiva comum para seu o mecanismo de transposição.

Os transposons de DNA estão amplamente distribuídos nos genomas de eucariotos, e geralmente em um número de cópias menor que os retrotransposons. Subdividem-se em duas subclasses, I e II. A subclasse I é composta pelos transposons de DNA clássicos e os MITE (do inglês, *miniature inverted repeat transposable elements*). Os transposons apresentam em sua estrutura uma ou mais ORF, e codificam uma transposase. A região da ORF é flanqueada em suas extremidades pelas TIR, que podem variar entre 10 e 50 pb, são reversocomplementares entre si e contêm as regiões promotoras do elemento. Os elementos que apresentam TIR são classificados dentro da ordem que recebe esse mesmo nome (Figura 17.11 A). Os transposons também são encontrados nos genomas como elementos autônomos e não autônomos.

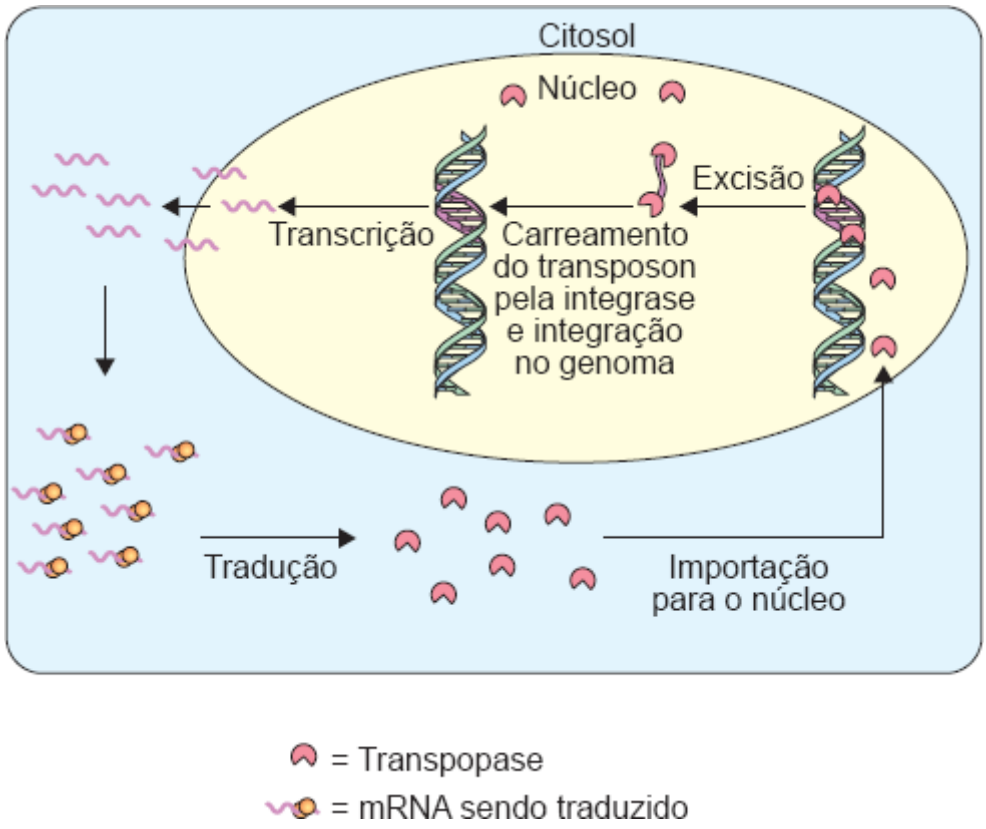


Figura 17.10 Replicação dos transposons.

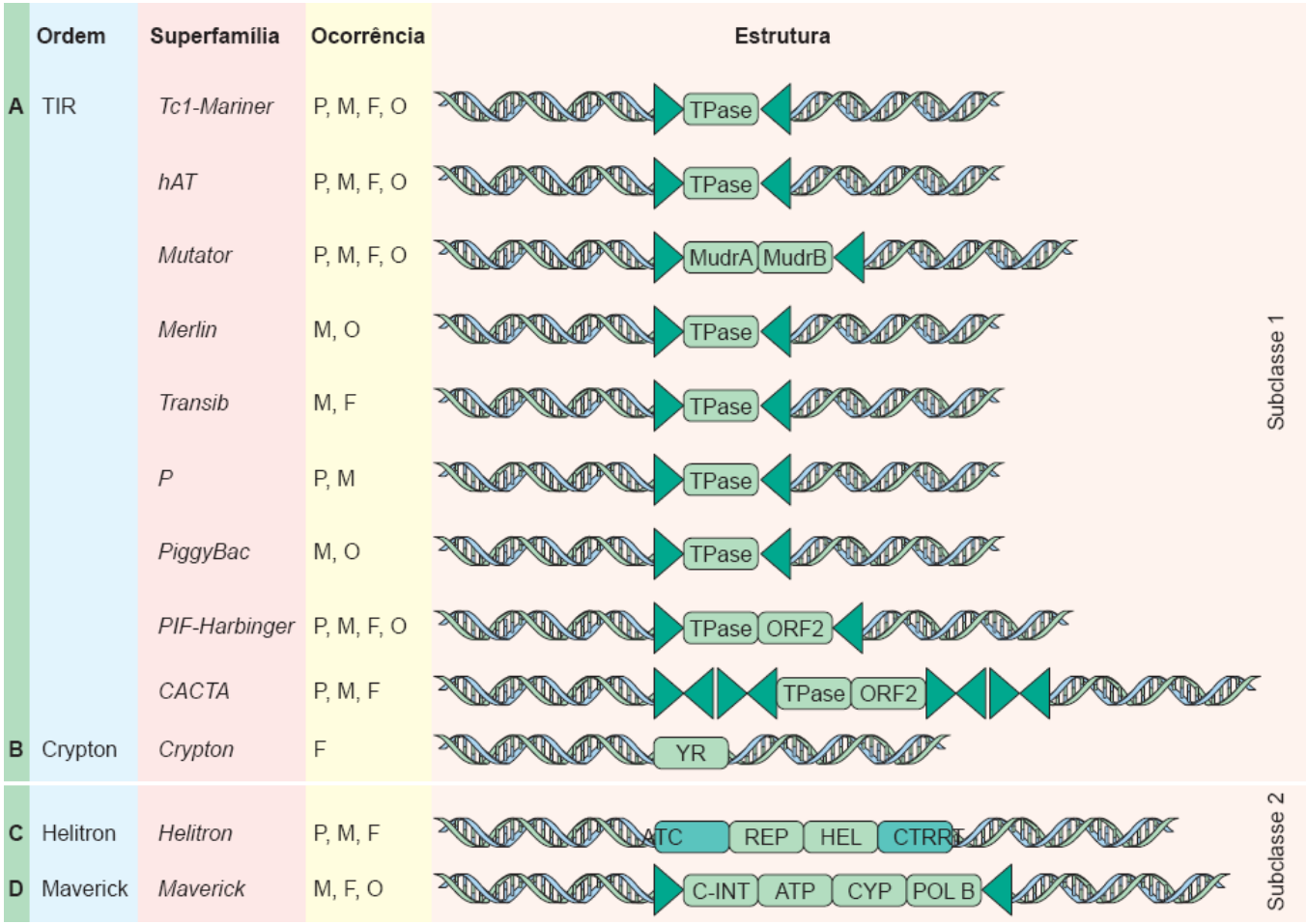


Figura 17.11 A a **D.** Distribuição, classificação e estrutura dos transposons. Adaptada de Wicker *et al.*, 2007.

Um dos tipos mais comuns de transposons de DNA não autônomos são os MITE. Eles estão presentes nos genomas eucarióticos em grande número, principalmente nos genomas de plantas. Suas sequências são curtas, entre 100 e 600 pb e não apresentam ORF, consistindo praticamente da justaposição de duas TIR, flanqueadas pelos TSD. Essas características os correlacionam com os transposons clássicos (subclasse I). A existência de transposons que potencialmente podem codificar transposases funcionais e que compartilham similaridade de sequência com alguns MITE levou à hipótese de que os MITE podem ser sequências remanescentes, derivadas de deleções de transposons, que são mobilizadas em *trans* pelas transposases codificadas por eles.

As principais superfamílias de elementos da ordem TIR autônomos são Tc1-Mariner, hAT, Mutator, Merlin, Transtb, P, Piggy-Bac, PIF-Harbinger e CACTA. Tc1-Mariner é uma superfamília assim denominada com base em seus principais membros. O elemento *Tc1* foi descrito no genoma de *C. elegans*. Já o elemento *Mariner* foi identificado em *Drosophila mauritiana,* por estar inserido em um *locus* relativo à cor dos olhos, causando mutações instáveis no gene que expressa essa característica. A caracterização do elemento revelou um gene, que codifica a transposase, de 1.286 pb, com TIR de 28 pb e TSD de 2 pb. Elementos dessa superfamília estão presentes nos genomas de praticamente todos os animais, como vertebrados, artrópodes, nematódeos, e também em genomas de fungos e ciliados. Tc1-Mariner tem grande importância biotecnológica (ver boxe “Os elementos de transposição podem ter aplicações biotecnológicas”).

A superfamília *hAT* de transposons tem como ícones o elemento *hobo* de *D. melanogaster*, *Ac* do milho e *Tam3* de *Antirrhinum majus*, popularmente conhecida como boca-de-leão. Suas principais características são a presença de três domínios proteicos conservados, localizados na porção carboxiterminal de suas transposases, TIR de 11 pb e duplicação do sítio de inserção de 8 pb. Outros transposons de diversos organismos têm sido descritos dentro dessa superfamília, como os elementos *Crypt1* do fungo *Cryptonectria parasítica* e *DodoPi*, do fungo fitopatógeno *Phytophthora infestans*, o elemento *Hermes* de *Musca domestica*, a dupla de elementos autônomo e não autônomo *Dart* e *nDart* do arroz e o elemento *SChAT* da cana-de-açúcar. Essa superfamília tem grande importância histórica, visto que *Ac* foi o elemento observado nos primeiros estudos de Barbara McClintock.

O sistema *Mutator*, como essa superfamília também é conhecida, foi primeiramente descrito em milho por Robertson, em 1978. Posteriormente, os MULE (do inglês, *Mutator-like elements*) foram identificados em diversos outros organismos, como na eudicotiledônea *A. thaliana*, e nas monocotiledôneas arroz e cana-de-açúcar. O elemento *Mutator* autônomo do milho apresenta TIR de 220 pb flanqueando duas ORF, *mudrA* e *mudrB*, que podem ser transcritas de maneira independente e codificam para as proteínas MURA e MURB, respectivamente. Enquanto MURA corresponde à transposase e é essencial para a mobilização do elemento, a função de MURB ainda não foi completamente elucidada. Sabe-se que é necessária para a integração dos MULE. É importante notar que a ORF *mudrB*, até o momento, foi encontrada somente nos elementos *Mutator* de milho, sendo ausente em outras espécies. Esses elementos apresentam alta frequência de transposição, sendo considerados os transposons mais ativos já identificados, altamente mutagênicos. Essa propriedade possibilitou que fossem utilizados como ferramentas de *transposon tagging*, para clonagem e identificação de genes em milho. Outros interessantes membros dessa superfamília são os *Pack-MULE* – sequências gênicas do hospedeiro, capturadas pelas TIR dos *MULE*, que passam a se comportar como transposons, podendo ser mobilizados pela ação da transposase de um elemento autônomo.

Os elementos de transposição podem ter aplicações biotecnológicas

Um dos grandes desafios atuais no estudo dos TE é como utilizar sua capacidade de mobilização e sua característica repetitiva para o desenvolvimento de tecnologias. São diversas linhas de pesquisa que focam a aplicação desses elementos como ferramentas de engenharia genética nas áreas de pesquisa básica, saúde e produção de alimentos, nos mais diversos organismos como, por exemplo, bactérias, plantas cultivadas, animais e até mesmo no ser humano. Os TE podem ser usados como marcas genéticas em programas de melhoramento vegetal e animal, informando relações de parentesco e a presença de genótipos de interesse durante o processo de cruzamento entre linhagens. Outra aplicação dos TE seria sua utilização como ferramenta de mutagênese, para a geração de bancos de mutantes com fins de caracterização genética (Figura 17.12 A). Essa aplicação envolve a manipulação genética de organismos de interesse pela introdução de TE exógenos hiperativos. Esses elementos terão altas taxas de mobilização e de geração de mutações no novo genoma hospedeiro. A caracterização dessas mutações possibilitará a correlação entre genótipos e fenótipos de interesse nas mais diversas áreas de pesquisa.

Uma das vertentes biotecnológicas de amplo interesse público e que utiliza TE com aparente sucesso é a terapia gênica, ou seja, o uso de genes como agentes no tratamento de doenças. Entre os elementos já adaptados para a terapia gênica em mamíferos está o *Sleeping Beauty* (SB), uma apropriada referência à fábula da Bela Adormecida (Figura 17.12 B). SB é um conjunto de elementos sintéticos desenvolvidos a partir de um transposon inativo da superfamília Tc1-Mariner de peixes salmonídeos. Ele é composto basicamente de dois vetores plasmidiais associados a diferentes versões do elemento. O primeiro vetor apresenta as TIR flanqueando o gene de interesse a ser mobilizado, sem o gene da transposase. O segundo vetor contém o elemento com gene da transposase hiperativa, mas não apresenta as TIR do elemento, sendo incapaz de mobilização. Desse modo, quando ambos os vetores são administrados às células a serem tratadas, o gene de interesse será integrado no genoma hospedeiro, utilizando a transposase produzida *in trans* pelo primeiro vetor e as TIR do segundo vetor. Esse processo torna possível a introdução de sequências de DNA de interesse em genomas de mamíferos de maneira eficiente e aleatória. Já foi utilizado para a transformação de células murinas no tratamento da hemofilia B, diabetes e doença de Huntington. Além disso, SB apresentou-se estável durante a transformação de células-tronco hematopoiéticas.

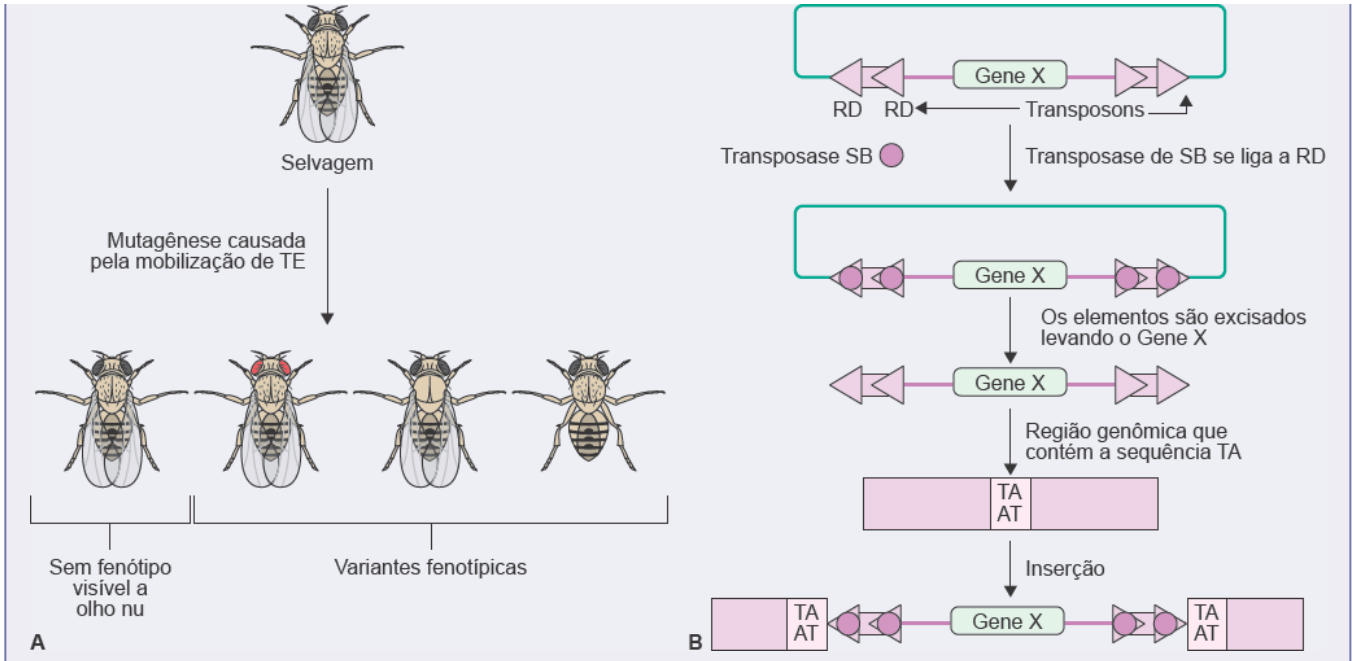


Figura 17.12 Aplicações biotecnológicas dos TE. **A.** Mutagênese. Os TE podem ser utilizados como geradores de variabilidade genética, que pode ser convertida em fenótipos visíveis e caracterizáveis macro e microscopicamente. Quando um elemento é inserido dentro de um gene ou em sua vizinhança, ele pode interromper o gene ou modificar seu padrão de expressão. Essa alteração molecular pode ser traduzida em alteração macroscópica, tornando os TE interessantes ferramentas para a geração de bancos de mutantes. **B.** Terapia gênica. O elemento SB é usado como ferramenta para a transformação genética. Para tanto, é necessário que se forneça ao tecido a ser transformado o vetor contendo o gene de interesse (gene X) entre dois elementos SB e também a transposase ativa, na forma de um vetor capaz de sintetizá-la. As transposases excisarão os elementos SB, levando com eles o gene X de interesse. Esse cassete formado pelos elementos SB e o gene X será inserido no genoma em uma região que contenha o dinucleotídeo TA. Adaptada de Geurts *et al.*, 2003.

Os transposons da superfamília Merlin foram descritos pela primeira vez em nematódeos. Apresentam similaridade com o elemento IS1016 da bactéria *Haemophilus influenzae*. Esses elementos foram encontrados apenas nos genomas de animais e bactérias. Um elemento autônomo pode ter uma ORF de até 10 Kpb. Suas TIR apresentam tamanho bastante variado, podendo ter de 24 até 462 pb, flanqueadas por TSD de 8 ou 9 pb.

A superfamília *Transib* foi identificada inicialmente na mosca *D. melanogaster* e no mosquito *Anopheles gambiae*. O elemento *Transib* autônomo tem tamanho de 3 a 4 Kpb, codificando uma transposase de aproximadamente 700 aminoácidos. É caracterizado por TIR de 43 a 45 pb e TSD de 5 pb. A transposase de *Transib* apresenta similaridade com as proteínas RAG1 do sistema de recombinação V(D)J, que é responsável pela variabilidade do sistema imunológico dos vertebrados. Cerca de 600 aminoácidos do centro catalítico de RAG1 apresentam similaridade com a transposase dessa superfamília, incluindo o domínio DDE, que é conservado entre elas. RSS (do inglês, *recombination signal sequence*) são os alvos de ligação do complexo RAG1/RAG2 no DNA, a partir do qual se inicia a clivagem e o processo de recombinação. Curiosamente, os RSS apresentam padrão e estrutura conservados em relação às TIR de *Transib*. Além disso, RAG1 é capaz de catalisar uma reação de transposição *in vitro*. Essa reação gera TSD semelhantes aos produzidos pela transposição de *Transib*. Essas evidências sugerem que transposons *Transib* tiveram importante papel na origem da maquinaria V(D)J.

O elemento *P* foi descoberto quando pesquisadores investigavam a causa da disgenesia híbrida em *D. melanogaster*. O primeiro elemento *P* autônomo molecularmente caracterizado apresentava 2,9 Kpb, com TIR de 31 pb e TSD de 8 pb. Outro aspecto interessante dessa superfamília é a recente colonização do genoma de *D. melanogaster*, em meados do século 20, por meio de um evento de transferência horizontal. Desde então, ele se disseminou entre populações naturais de *D. melanogaster*. Devido a sua alta taxa de transposição, os elementos *P* foram utilizados como ferramentas moleculares para *transposon tagging* e vetores de transformação tanto para linhagens germinativas quanto para culturas de células.

Elementos da superfamília PiggyBac estão presentes nos genomas de fungos, plantas e animais, incluindo insetos, crustáceos, urocordados, anfíbios, peixes e mamíferos. Ao se transpor, inserem-se preferencialmente em sítios TTAA. São amplamente utilizados como ferramentas biotecnológicas como, por exemplo, em vetores para

expressão gênica, mutagênese insercional para a descoberta de genes relacionados com o câncer em ratos, bem como para transformação de culturas de células humanas.

A superfamília PIF-Harbinger foi inicialmente identificada em plantas, com o elemento *PIF* em milho e *Harbinger* em *A. thaliana*; já foi descrita em diversos organismos, tais como fungos, plantas e animais. Apresenta como características marcantes a presença de duas ORF, uma que codifica a transposase e outra que codifica uma proteína de ligação a DNA, e TSD de 3 pb; além disso, há forte preferência pelo sítio de inserção TAA. Pertencem também a essa superfamília os elementos da família *Tourist* como o MITE *mPing*, e os elementos autônomos cuja transposase é capaz de realizar sua transposição, *Ping* e *Pong*.

Nos elementos da superfamília CACTA, as TIR flanqueiam complexas sequências subterminais de centenas de nucleotídeos, contendo repetições tanto diretas como invertidas de motivos de aproximadamente 15 pb. Apresentam também duas ORF, uma codificando a transposase e uma proteína regulatória. Em plantas, as TIR apresentam o motivo CACTA/CACTG altamente conservado, enquanto em animais e fungos esse motivo é substituído por CCC. O elemento ícone dessa superfamília, *Spm*, foi identificado por Barbara McClintock em seus estudos com milho, na década de 1950.

Crypton é uma segunda ordem dentro da subclasse I, descoberta no genoma de fungos patogênicos. São transposons incomuns, pois apresentam uma ORF que codifica não uma transposase, mas sim uma tirosina recombinase (ver Figura 17.11 B). Essa característica pode sugerir uma relação com os retrotransposons do tipo DIR e Ngaro. No entanto, os Cryptons não apresentam qualquer gene relacionado com uma transcriptase reversa, o que indica que devem se transpor diretamente via DNA. Apesar de seu mecanismo de transposição ainda não ter sido elucidado, é classificado na classe II. Outra peculiaridade dos Cryptons é que não apresentam TIR, mas são flanqueados por pequenas repetições diretas que podem ser resultado da sua integração no genoma hospedeiro por meio de recombinação.

Helitrons e Mavericks são transposons de DNA que se mobilizam de maneira atípica

Os transposons de DNA da subclasse II têm um modo de transposição diferenciado do "copia e cola", mas são incluídos nessa classe, pois se mobilizam sem a necessidade de um intermediário de RNA. Esses elementos pertencem a duas superfamílias, Helitrons e Mavericks, os últimos também conhecidos como Polintons.

Os Helitrons foram recentemente descobertos em *A. thaliana*, arroz e no nematódeo *C. elegans*. Atualmente, eles já foram identificados em quase todos os grandes grupos eucarióticos, dentre eles, animais, plantas, fungos e protistas. Sua estrutura é bem diversa dos outros transposons de DNA. Helitrons podem ser bem longos, chegando a mais de 50 Kpb, não produzem sítios de inserção no genoma hospedeiro, nem repetições terminais. As características que possibilitam sua identificação são o fato de que se inserem entre nucleotídeos A e T, têm extremidades conservadas que apresentam os nucleotídeos TC na região 5' e CTRR (R significa nucleotídeos A ou G) na região 3' e apresentam uma sequência de nucleotídeos capaz de formar um grampo de aproximadamente 18 pb próxima a região 3'. O conteúdo gênico dos Helitrons é muito variável, mas sua organização mais básica consiste em um gene com dois domínios proteicos, a transposase RepHel (do inglês, *replication protein*) (ver Figura 17.11 C). O domínio Rep da transposase RepHel tem função de clivagem, transferência e ligação da fita de DNA durante a transposição. O domínio Hel tem função helicase, desfazendo a dupla-fita de DNA durante a replicação. Outros genes encontrados nos Helitrons são endonucleases, proteases e diferentes tipos de RPA (do inglês, *replication protein A*), mas a presença de cada um desses genes está restrita a algumas famílias de Helitrons. Outra característica interessante dos Helitrons é a capacidade de capturar genes do hospedeiro. Acredita-se que *Helitrons* ancestrais tenham capturado as enzimas RPA do genoma hospedeiro. Um dos principais impactos da captura de genes por TE é sua duplicação a cada evento de mobilização do elemento.

Similar aos Helitrons, os Polintons foram recentemente identificados, por mais de um grupo de pesquisadores, em diversas espécies eucarióticas, dentre elas, animais (com exceção dos mamíferos), fungos e protozoários. Assim, na literatura, os Polintons podem ser também referidos como Mavericks e Tlrs (ver Figura 17.11 D). Os Polintons autônomos contêm entre 15 e 20 Kpb e codificam diversas proteínas, como uma integrase semelhante à de retrovírus (C-INT), uma ATPase (ATP), uma protease semelhante à dos adenovírus (CYP) e uma DNA polimerase B (POL B). Outra característica dos Polintons são suas longas TIR, que podem alcançar até 700 pb. Os pesquisadores que identificaram os Polintons acreditam que esses elementos podem ser um elo evolutivo entre

plasmídeos mitocondriais e citoplasmáticos, mas aspectos relacionados com estrutura, mobilização e origem ainda precisam ser mais bem investigados.

O modo de transposição tanto dos Helitrons quanto dos Mavericks ainda não foi comprovado experimentalmente, mas acredita-se que seja diferenciado entre as duas superfamílias e envolva a capacidade de quebra da simples ou dupla-fita de DNA, transposição mediada por recombinação, replicação e reparo de DNA (síntese de DNA), promovendo a integração em uma nova posição do genoma. Esses modos de mobilização estariam evolutivamente relacionados com a replicação de geminivírus, no caso dos Helitrons, e de bacteriófagos, plasmídeos lineares e adenovírus, no caso dos Mavericks.

A transposição tem mecanismos complexos de regulação

A atividade dos TE pode ser mutagênica, trazendo efeitos deletérios para os organismos. Assim, não é difícil imaginar que existem mecanismos de controle da transposição que possam silenciar os elementos na maior parte do tempo, possibilitando sua atividade apenas em momentos específicos. No entanto, há evidências da ativação dos TE em situações de estresse, e isso poderia ser interessante do ponto de vista evolutivo. Quando uma população se encontra sob pressão seletiva, a variação genética adaptativa pode gerar indivíduos capazes de superar tais pressões.

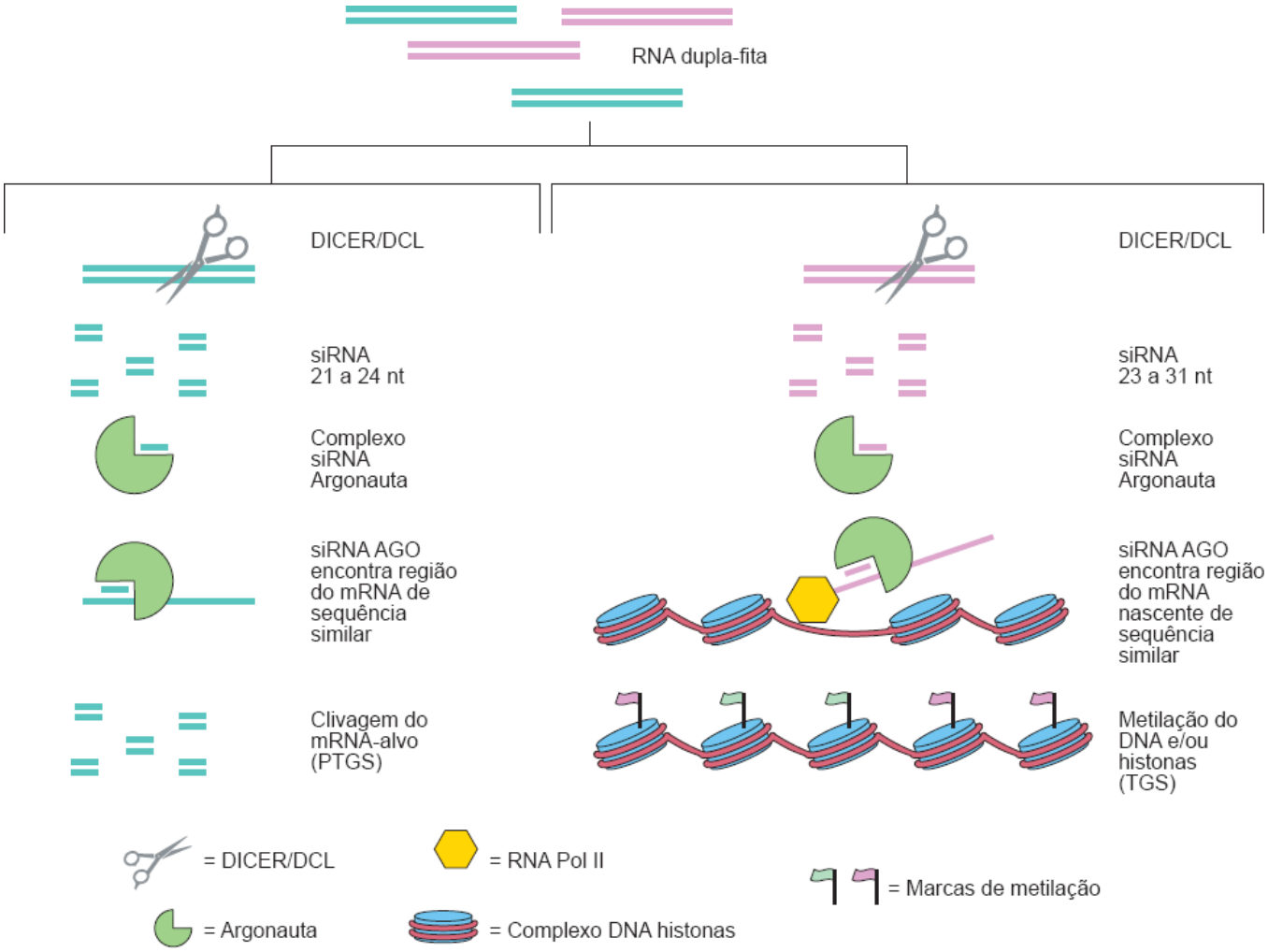


Figura 17.13 Mecanismo de controle dos elementos de transposição por pequenos RNA em plantas e *Drosophila*. Adaptada de Castel e Martienssen, 2013.

TE são alvo de regulação epigenética

Existem dois níveis de regulação e silenciamento dos TE: um atua em nível transcricional e outro em nível pós-transcricional, conhecidos como TGS e PTGS (do inglês, *transcriptional gene silencing* e *post transcriptional gene silencing*), respectivamente (Figura 17.13). Tanto TGS como PTGS se iniciam com a presença de uma dupla-fita de RNA (dsRNA, do inglês, *double strand RNA*), que é clivada gerando siRNA (do inglês, *small interference RNA*). A origem dos dsRNA difere entre espécies. Na planta *A. thaliana* e na mosca *D. melanogaster*, por exemplo, eles se originam da transcrição concomitante de transcritos complementares, que se ligam formando a dupla fita de RNA.

Esses transcritos podem vir de *loci* separados, direcionados em senso e antissenso, ou de *loci* estruturados para transcrever fragmentos senso e antissenso do TE em um mesmo transcrito. Em arabidopsis, adicionalmente, os dsRNA podem ser gerados por meio da síntese de uma fita complementar de RNA sobre o transcrito de um TE, pela ação de uma RNA polimerase dependente de RNA. O processo segue com o reconhecimento do dsRNA pela enzima DICER. Ela promove a sua fragmentação em siRNA, que pode ser de duas faixas de tamanho, de 21 a 24 nucleotídeos e de 23 a 31 nucleotídeos. Os siRNA de 21 a 24 nucleotídeos ligam-se ao complexo enzimático RISC (do inglês, *RNA-induced silencing complex*), capaz de reconhecer os mRNA com identidade de sequência nucleotídica aos siRNA. Uma vez reconhecido, o mRNA-alvo é degradado pela ação do centro catalítico de RISC, conhecido como Argonauta (AGO). Esse tipo de silenciamento é chamado de pós-transcricional, pois silencia o gene por meio da degradação de seus transcritos.

Diferentemente do silenciamento pós-transcricional, os siRNA de 23 a 31 nucleotídeos se acoplam ao complexo RITSC (do inglês, *RNA-induced transcriptional silencing complex*), que também tem um domínio catalítico do tipo Argonauta. Esse complexo direciona-se para a dupla-fita de DNA, reconhecendo os mRNA-alvo recém-transcritos pela RNA polimerase II e acoplando-se a esses transcritos. O complexo então recruta para a região uma maquinaria enzimática que promove a metilação do DNA ou das histonas correspondentes à região-alvo, heterocromatinizando a região e impedindo a sua transcrição. Assim, esse mecanismo é conhecido como silenciamento transcricional.

Outra classe de pequenos RNA regulatórios são os piRNA. Os piRNA ocorrem em animais e têm como função o silenciamento de TE em suas células germinativas. A importância desse mecanismo está em evitar que eventuais mutações geradas pela ação dos TE sejam incorporadas, por meio de sua passagem para as gerações seguintes. Um exemplo da atuação de piRNA ocorre em *D. melanogaster*. Os precursores primários dos piRNA são gerados pela transcrição de *loci* específicos localizados na heterocromatina, que apresenta sequências de TE degradadas e direcionadas tanto no sentido senso como antissenso, geralmente chamadas de cemitérios de TE. No citoplasma, esses precursores são clivados em fragmentos de 24 a 30 nucleotídeos, gerando os piRNA em direção antissenso; eles se acoplam ao complexo proteico PIWI/Aubergine. O complexo PIWI/piRNA antissenso volta para o núcleo e então se liga aos transcritos senso gerados a partir dos TE ativos, promovendo a sua clivagem e gerando piRNA no sentido senso. Os piRNA senso, por sua vez, se ligam a AGO3, que reconhece os precursores primários dos piRNA antissenso e os cliva gerando novos piRNA antissenso. Esses vão novamente se ligar a proteínas PIWI, dando continuidade a esse ciclo conhecido como pingue-pongue. As proteínas PIWI acopladas aos piRNA antissenso promovem, portanto, o silenciamento pós-transcricional dos TE. Por outro lado, PIWI também pode levar ao silenciamento transcricional da região-alvo, recrutando enzimas que promovem a metilação tanto do DNA quanto de histonas a ele ligadas.

Os piRNA evitam a ocorrência da disgenesia híbrida em *D. melanogaster*. Fêmeas que contêm o elemento *P* produzem e acumulam piRNA relacionados com esse elemento em seu citoplasma durante a formação dos oócitos. Esses piRNA presentes no citoplasma dos oócitos são passados para os zigotos, uma vez que essa é uma herança citoplasmática materna, pela qual induzem o silenciamento epigenético do elemento *P*. Dessa maneira, o elemento fica sob controle e os indivíduos que se desenvolvem a partir desses embriões são normais. A disgenesia híbrida ocorre quando uma fêmea de uma linhagem que não contém o elemento *P* é cruzada com um macho de uma linhagem que contém o elemento em seu genoma. Como não são passados piRNA para o zigoto, não há controle sobre o elemento *P*. O elemento escapa do silenciamento e se transpõe nas células germinativas do embrião, levando à esterilidade dependente de temperatura, altas taxas de mutação e aumento das taxas de recombinação e rearranjos cromossômicos (Figura 17.14).

Os efeitos da transposição podem ser prejudiciais se passados para linhagens celulares germinativas, pois isso possibilitaria sua perpetuação para as gerações seguintes. Recentemente, foi descrito que a regulação epigenética tem papel importante na formação tanto do gametófito masculino como do feminino em *A. thaliana*. Assim, esses mecanismos garantem a manutenção da herança epigenética, a fim de proteger as células germinativas da ação dos TE.

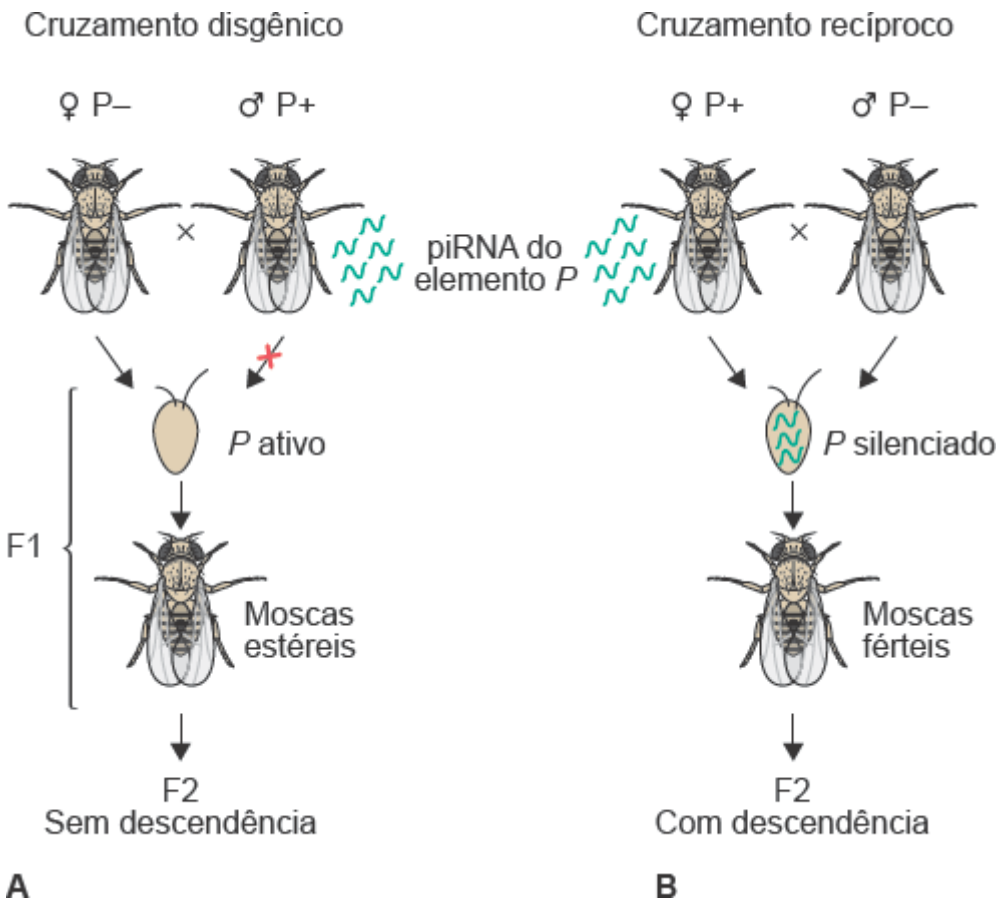


Figura 17.14 Esquema de cruzamentos disgênicos entre moscas *D. melanogaster*. **A.** Fêmeas que não contêm o elemento *P* em seu genoma são cruzadas com machos contendo elemento *P*. **B.** Fêmeas contendo o elemento *P* são cruzadas com machos que não contêm o elemento *P*. Adaptada de Siomi *et al*., 2011.

Presença de TE e impacto em genomas

Os TE podem causar mutações e levar à variabilidade genética. Essa capacidade mutagênica está relacionada com a sua transposição (ou seja, com a inserção do próprio elemento ou de uma cópia em um novo sítio do genoma) e com eventos de recombinação de regiões não homólogas, assim como ao seu envolvimento em mecanismos de regulação epigenética. A manutenção da integridade dos genomas depende, portanto, de mecanismos de regulação da atividade dos TE, de modo a evitar possíveis danos decorrentes da sua atividade.

Por outro lado, eventos causados pela atividade dos TE podem ser selecionados positivamente ao longo da evolução, e incorporados na estrutura e funcionamento dos genomas. Estudaremos agora, em detalhes, as consequências da mobilização dos TE sobre os genomas.

A transposição pode alterar o tamanho dos genomas

Ao contrário do que o senso comum pode sugerir, o tamanho dos genomas não está diretamente correlacionado com o grau de complexidade morfofisiológica ou evolução dos organismos. O genoma de algumas gramíneas, por exemplo, é maior que o da espécie humana. Esse fenômeno é chamado de paradoxo do valor C, e pode, em parte, ser explicado pela presença dos TE. Uma classe de TE responsável por esse processo em particular são os retrotransposons, cuja transposição é replicativa, aumentando o número de cópias do elemento. Há evidências de que a diferença no tamanho do genoma das gramíneas milho e sorgo, o primeiro com genoma de 2.160 Mpb e o segundo com 740 Mpb, se deve ao grande acúmulo de retrotransposons no milho. A comparação da região do gene *adh1* (desidrogenase do álcool) do milho e do sorgo demonstrou que o milho contém retrotransposons nessa região, enquanto o sorgo, não. Isso sugere que os retrotransposons se multiplicaram no milho após a divergência entre essas duas espécies. No caso de duas espécies de arroz, *O. sativa* e *O. australiensis*, a variação no número de cópias do elemento RIRE1 pode explicar a diferença de ⅓ no tamanho de seus genomas.

A inserção de TE pode alterar a estrutura e a expressão gênica

Os TE podem alterar os níveis de expressão de genes por meio da sua inserção dentro ou próxima deles. A inserção dentro de um gene pode tanto causar sua disrupção (ou seja, levar à perda da funcionalidade do gene) (Figura 17.15 A) quanto levar à exonização do TE (Figura 17.15 B). Neste caso, o TE é incorporado ao gene, em um fenômeno conhecido como domesticação, que será discutido mais adiante. Outra possibilidade é a inserção do TE ocorrer em uma região intrônica. Aqui, a proteína codificada pelo gene pode não ser comprometida. No entanto, se as regiões promotoras do TE levarem à transcrição no sentido oposto ao do gene, podem ser gerados transcritos antissenso aos transcritos do gene. Os transcritos antissenso, na presença dos transcritos senso, podem formar um RNA de dupla-fita e acionar o mecanismo de silenciamento por RNAi, levando ao silenciamento do gene (Figura 17.15 C).

Ao se inserir na região promotora de um determinado gene, o TE pode influenciar na sua expressão, tanto positivamente quanto negativamente. A repressão da transcrição pode ocorrer no caso de o TE interromper alguma parte fundamental da região promotora. Por outro lado, a própria região promotora do TE pode induzir a expressão do gene, aumentando a sua transcrição (Figura 17.15 D).

A inserção de um TE pode ainda levar à reorganização de estruturas do gene em que se insere. O TE pode, por exemplo, introduzir um novo sítio de início de transcrição, alterando os transcritos do gene. A inserção do TE pode ainda ocorrer próxima à fronteira de regiões intrônicas, modificando-as de modo a gerar transcritos processados alternativamente (Figura 17.15 E). Finalmente, os TE podem também carregar marcas epigenéticas para os genes em que se inserem, ou que se localizam em regiões vizinhas às de inserção. Desse modo, o estado de metilação desses genes pode ser modificado, levando ao seu silenciamento (Figura 17.15 F).

Um exemplo que ilustra o impacto causado pela inserção de TE é a diferença entre variedades de uva. A antocianina é um pigmento que confere a cor vermelha aos frutos. Algumas variedades de pele branca, como Chardonnay, apresentam um TE inserido na região promotora de um dos genes que regulam a síntese da antocianina. Trata-se do retrotransposon com LTR *Gret1* (*grapevine retrotransposon 1*). A inserção do TE no promotor teve como consequência a perda da função do gene, inibindo a produção de antocianina, o que resultou em frutos de pele branca. Uvas com coloração, como a Cabernet, não apresentam a inserção do TE. Por outro lado, um rearranjo decorrente da recombinação das LTR de *Gret1* levou ao restabelecimento da função do gene, de modo a gerar uvas com coloração vermelha, como a Ruby Okuyama (Figura 17.16).

Os TE podem participar na formação de redes regulatórias

Os TE podem dispersar motivos regulatórios ao longo do genoma, levando à formação de novas redes de regulação gênica. A formação das redes pode ocorrer das seguintes maneiras:

- Ao se dispersar pelo genoma, o TE pode se posicionar a montante de diversos genes. Se o TE apresentar um motivo ao qual se liga uma PLD (proteína de ligação a DNA), e essa PLD for um fator de transcrição, ela pode se ligar ao motivo e ativar a transcrição dos genes vizinhos à inserção do TE. Assim, genes que anteriormente não tinham relação passam a se ativar de maneira coordenada, participando de uma rede de regulação transcricional (Figura 17.17 A)
- Se um TE se dispersar pelo genoma e passar a ser cotranscrito com os genes vizinhos, os transcritos desses diversos genes vão carregar uma parte em comum, que contém a sequência do TE. Se o TE apresentar um motivo de ligação de uma PLR (proteína de ligação ao RNA), a proteína em questão vai se ligar aos transcritos e realizar o encaminhamento deles para uma mesma rede de regulação pós-transcricional (Figura 17.17 B).

Devido a essas características, os TE podem ser considerados importantes agentes na evolução dos genomas. Enquanto efeitos deletérios são negativamente selecionados, mudanças vantajosas podem ser incorporadas. A análise de genomas complexos tem demonstrado o grande impacto que os TE tiveram na evolução dos genes. Já foi comprovado, por exemplo, que mais de 10.000 componentes não exônicos do genoma humano, que evoluíram sob pressão seletiva, são derivados de TE. Esses componentes estão associados a fatores de transcrição ou genes relacionados com o desenvolvimento, podendo atuar como elementos regulatórios. Em humanos, também foi demonstrado que a regulação epigenética sofrida pelos TE pode ter consequências na regulação de genes que contêm sequências de TE em suas regiões promotoras. Estudos recentes demonstraram que a fração do genoma humano responsável por sua regulação é significativamente maior que a fração que codifica proteínas, agregando indícios de que os TE podem estar relacionados com funções regulatórias. A implicação dos TE na regulação

epigenética dos genomas, bem como na formação de redes regulatórias, é um assunto bastante atual e tem sido extensivamente discutido, levando-se em conta o conhecimento adquirido para diversos grupos de organismos.

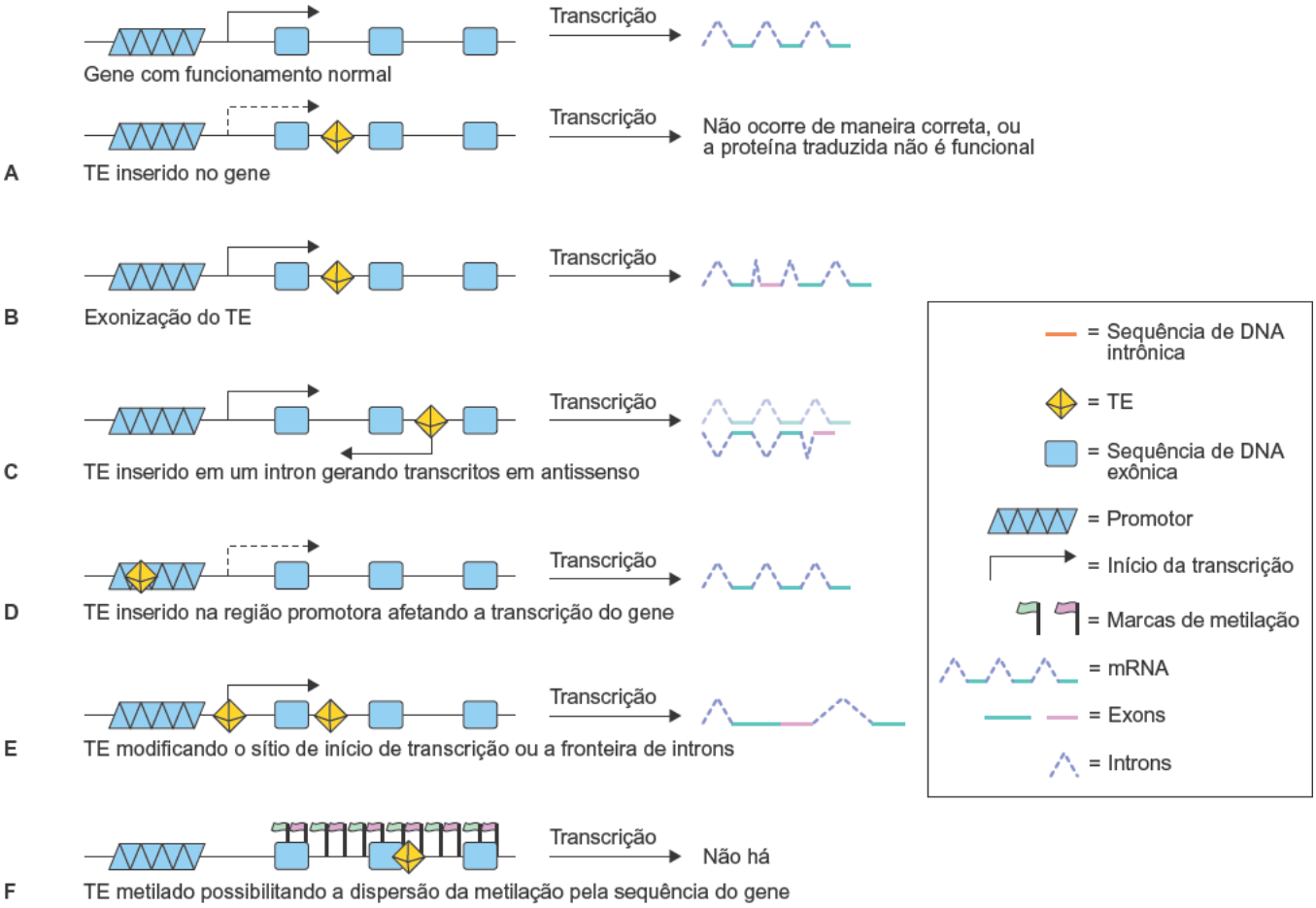


Figura 17.15 A a **F.** Mudanças estruturais e funcionais causadas pela inserção de um TE. Adaptada de Feschotte, 2008.

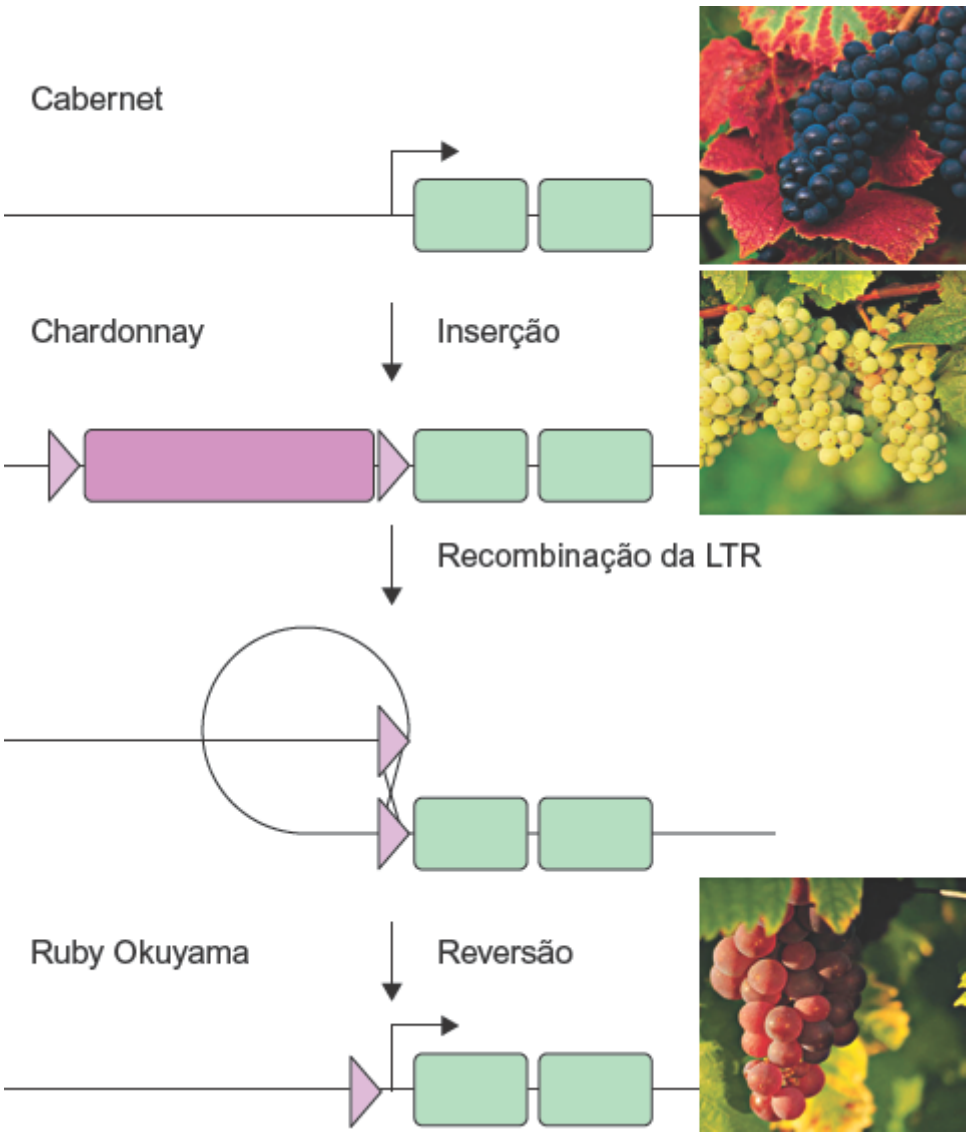


Figura 17.16 A inserção/remoção de elementos de transposição gerando fenótipos em uvas. Adaptada de Lisch, 2013.

Rearranjos cromossômicos

Pela sua característica repetitiva, os TE podem servir como agentes de recombinação não homóloga ou ilegítima, provocando uma variedade de mudanças estruturais, entre as quais, deleções, inserções, inversões, translocações e duplicações. A cada evento de transposição replicativa são gerados substratos potenciais para recombinação. Por exemplo, a recombinação entre LTR de diferentes cópias de um retrotransposon inseridas na mesma fita do DNA pode levar a deleções substanciais de material genético. No entanto, também existem relatos de recombinação associada a eventos de expansão do genoma. Em geral, esses eventos de expansão estão associados ao aumento da ploidia, visto que o aumento do número de cromossomos homólogos pode possibilitar maior reorganização estrutural. Isso demonstra o papel dos TE na plasticidade dos genomas.

Exemplos de TE promovendo recombinação nos genomas hospedeiros são observados facilmente nos cromossomos politênicos de glândulas salivares de *Drosophila*. Cromossomos politênicos são excepcionalmente grandes em função de repetidos eventos de replicação de cromossomos homólogos interfásicos, sem a separação dos filamentos duplicados. Nesses casos, a recombinação de TE inseridos na mesma fita de DNA gera inversões, que são facilmente visualizadas se estiverem em estado heterozigoto, devido à formação de uma alça de inversão cromossômica (Figura 17.18).

Os TE podem ser doadores de domínios proteicos para a formação de novos genes

Os domínios proteicos dos TE têm propriedades como ligação ao DNA, clivagem e integração. Essas funções podem ser bastante interessantes, do ponto de vista evolutivo, na formação de novos genes. O recrutamento de domínios funcionais de TE na formação de novos genes exemplifica a importância desses elementos como agentes

ativos na evolução dos genomas. O mecanismo pelo qual partes de um TE podem ser cooptadas pelo genoma para a formação de um novo gene é conhecido como domesticação. Uma característica marcante da domesticação é que o TE perde sua capacidade de transposição. Uma mutação em regiões essenciais para a transposição (p. ex., nas LTR) pode levar à fixação do TE no sítio em que se encontra no genoma. Outra característica é a aquisição de uma função na célula. Por exemplo, o TE pode perder sua função de transposase, mas manter o domínio proteico de ligação ao DNA, podendo adquirir uma nova função como fator de transcrição. A perda da capacidade de se mobilizar pode ocorrer antes da domesticação em si (Figura 17.19 A), ou a domesticação pode ocorrer primeiro. Neste último caso, o TE ainda pode se transpor, mesmo já tendo adquirido uma nova função celular, e a fixação ocorre somente depois (Figura 17.19 B). Enquanto os TE se encontram nos genomas como famílias multicópia, podendo apresentar entre si mutações, inserções ou deleções, os TE domesticados costumam se apresentar em uma ou poucas cópias, conservadas entre diferentes organismos.

Um exemplo da reciclagem de material genético dos TE na formação de novos genes é o gene que codifica para a proteína SETMAR. Ele se formou a partir da fusão do gene da histona metiltransferase SET com a transposase do transposon Hsmar1. SET era preexistente no genoma de vertebrados, enquanto Hsmar1 se inseriu na sua vizinhança na linhagem dos primatas. Um evento de exonização capturou a transposase, que passou a fazer parte da proteína, conferindo-lhe a função adicional de ligação ao DNA, mais especificamente às TIR de Hsmar1. No genoma humano, existem aproximadamente 1.500 cópias de Hsmar com TIR conservadas. Esses sítios representam potenciais alvos de ligação ao DNA pela SETMAR, que podem possibilitar uma ampliação da rede de ação da histona metiltransferase.

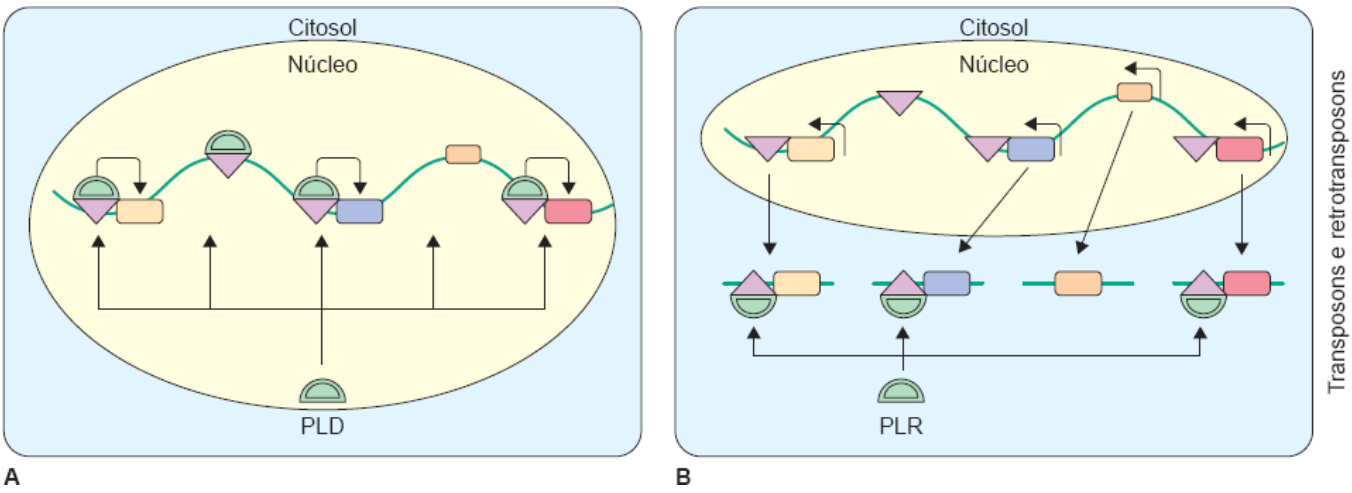


Figura 17.17 Elementos de transposição criando redes regulatórias no genoma. **A.** Rede de regulação transcricional criada pela distribuição de um elemento de transposição. O elemento de transposição (triângulo roxo) está disperso pelo genoma, às vezes localizado a montante de alguns genes (retângulos laranja, azul e vermelho). Se o elemento contiver motivos de ligação aos quais uma proteína de ligação a DNA (PLD, meio-círculo em verde) se ligue, então ele será capaz de interferir na transcrição dos genes a jusante. **B.** Rede de regulação pós-transcricional criada pela distribuição de um elemento de transposição. Se o elemento contiver motivos de ligação aos quais uma proteína de ligação a RNA (PLR, meio-círculo em verde) se ligue, ele será capaz de interferir na tradução ou destino do RNA mensageiro a jusante. Adaptada de Feschotte, 2008.

Figura 17.18 A e **B.** Exemplos de inversões em cromossomos politênicos de glândulas salivares de *Drosophila willistoni*. A seta indica a base da inversão causada pela recombinação do DNA. Adaptada de Rohde e Valente, 2012.

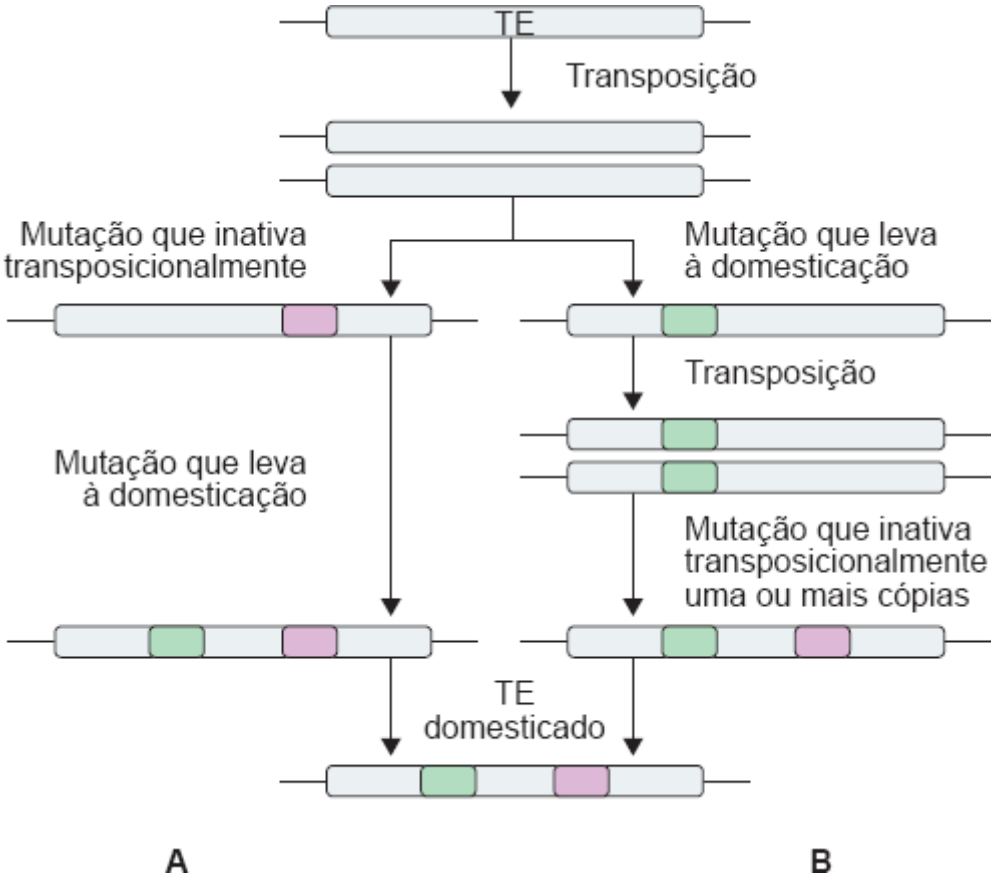


Figura 17.19 A e **B.** Eventos relacionados com a domesticação de um elemento de transposição. Adaptada de Volff, 2006.

Outro exemplo é a transposase domesticada DAYSLEEPER. Ela foi descrita como um fator de transcrição que se liga a motivos presentes em regiões promotoras de determinados genes, incluindo genes de reparo de DNA. O *knockout* desse gene demonstrou que ele é essencial para o desenvolvimento da planta *A. thaliana*, que apresentou graves deformidades nas partes reprodutivas na falta de DAYSLEEPER.

Instabilidades do genoma devido à atividade dos TE podem gerar doenças

Apesar da existência dos mecanismos de regulação, não é surpreendente que a atividade dos TE cause instabilidade nos genomas. Tal instabilidade pode ser muito valiosa do ponto de vista evolutivo, pois possibilita a inovação

genética; por outro lado, a atividade mutagênica dos TE pode causar danos à estrutura e funcionamento dos genomas, que se refletem no aparecimento de diversas doenças.

No genoma humano, há indícios de que os retrotransposons sem LTR *L1*, *Alu* e *SVA* se mantêm ativos transposicionalmente há milhares de anos, aumentando seu número de cópias. Estudos demonstraram correlação entre a hipometilação da região promotora do elemento *L1* e novos eventos de transposição em certos tipos de tumores. É vasta a lista de casos relacionando mutações geradas pelos TE com determinados tipos de câncer – entre eles, câncer de mama, de colo e retinoblastoma. No entanto, sabe-se que o câncer é uma doença complexa, causada por múltiplos fatores. Ainda que os TE possam gerar mutações que aumentam a propensão para o desenvolvimento da doença, eles não podem ser apontados como sua única causa. Além do câncer, diversas doenças genéticas hereditárias também já foram relacionadas com mutações causadas pela atividade de TE, tais como a distrofia muscular de Duchenne, síndrome autoimune linfoproliferativa, hemofilia A e B, fibrose cística, entre outras.

Quando um evento de transposição ocorre nas linhagens germinativas, ele pode ser passado para as gerações futuras. No entanto, a transposição pode também ocorrer em tecidos somáticos, não afetando as linhagens germinativas. Nesse caso, é denominada transposição somática. Apesar de não serem passadas para as gerações seguintes, essas novas inserções podem desencadear um mosaicismo de expressão gênica nos tecidos. Em humanos, este mosaicismo pode estar relacionado a algumas síndromes neurológicas, assim como com alguns tipos de tumor. Um exemplo de mutação somática que levou ao desencadeamento de doença é o caso de uma paciente com carcinoma de mama. Comparando-se o tecido mamário normal da paciente com o tecido do carcinoma, foram identificadas mudanças em dois *loci* relativos ao proto-oncogene c-myc. Um *locus* apresentou um rearranjo; o outro, uma inserção do retrotransposon sem LTR *L1*.

Doenças complexas relacionadas a distúrbios mentais também já foram associadas a TE. A neurofibromatose tipo 1 (NF1) é frequente em pacientes com problemas neurodesenvolvimentais e autismo. Alguns pacientes autistas com polimorfismo de quatro nucleotídeos em uma cópia de um elemento *Alu*, localizado em um intron do gene NF, apresentam uma forma mais grave da doença. Outros distúrbios neurodesenvolvimentais (p. ex., síndrome de Rett e ataxia-telangiectasia) são patologias complexas que apresentam como um de seus aspectos o aumento na retrotransposição do elemento *L1*.

A evolução dos TE ocorre em ciclos

A manutenção de uma família de TE em uma espécie é garantida por uma combinação de fatores intrínsecos dos elementos e da espécie. Entre os fatores intrínsecos do TE, podemos citar a taxa de mobilização e de integração nos cromossomos. Como fatores da espécie, é possível mencionar seleção natural, deriva genética e taxa de mutações. O entendimento de como esses fatores atuam na evolução dos TE levou à proposição de um ciclo evolutivo, o que é dividido basicamente em três etapas, como descrito a seguir (Figura 17.20).

Origem

Propõem-se dois mecanismos para a emergência de uma linhagem de TE em um genoma. Esses mecanismos são a reativação de elementos preexistentes, que estavam silenciados por mecanismos regulatórios, e o fenômeno de transferência horizontal (TH). A reativação de uma família de TE é um evento difícil de acompanhar experimentalmente; por isso, a maior parte das discussões sobre essa forma de origem de TE permanece no campo teórico. Já a TH tem sido demonstrada experimentalmente e aceita pelos pesquisadores da área. Um dos motivos é a disponibilidade cada vez maior de genomas completos, que possibilita a comparação dos mobilomas de diversas espécies de forma rápida e eficiente.

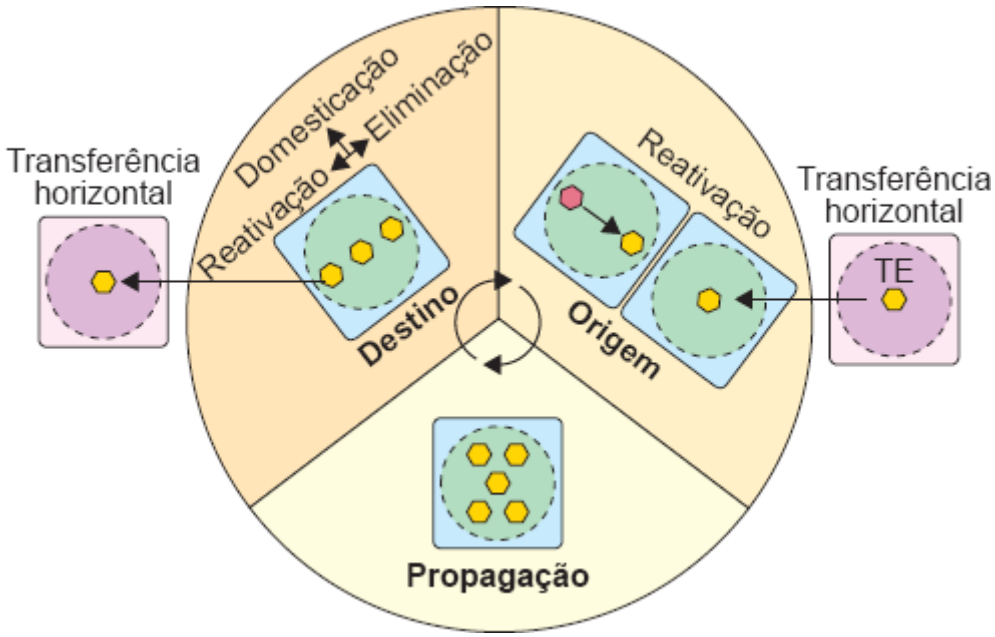


Figura 17.20 Ciclo evolutivo dos elementos de transposição. Adaptada de Almeida e Carareto, 2005.

Transferência horizontal

Um dos conceitos de TH mais simples se refere à troca de material genético entre genomas, sem o envolvimento de herança vertical que envolve gerações. Até recentemente, acreditava-se que a TH era muito frequente, mas exclusiva de procariotos, e acontecia durante os processos de conjugação, transformação e transfecção. Nas últimas duas décadas, vários estudos demonstraram que, embora menos frequente que em procariotos (ver boxe "Os procariotos também contêm TE em seus genomas"), esse mecanismo acontece nos eucariotos e diversas linhagens de transposons e de retrotransposons colonizaram novas espécies por meio da TH.

A transferência horizontal de TE já foi detectada em três reinos eucariotos, com diferentes taxas. Ela é mais frequente em *Animalia* (94%), seguido por *Plantae* (4%) e *Fungi* (2%). Acredita-se que essa grande discrepância entre reinos possa ser explicada por um desvio no estudo de TH em espécies de modelos animais, entre elas, as das moscas-da-fruta do gênero *Drosophila*, para as quais 178 dos 330 eventos de TH de animais foram propostos. Independentemente da espécie hospedeira, a detecção de eventos de TH não é um estudo trivial. Em geral, o primeiro indício de TH é uma incongruência entre a árvore filogenética de um TE e de seus hospedeiros. No entanto, antes de se propor um evento de TH, é necessário realizar vários testes para excluir hipóteses alternativas; entre as quais, perda estocástica por deriva genética, domesticação, diferentes taxas evolutivas entre hospedeiros e polimorfismo ancestral.

Outro ponto a ser considerado é como um TE consegue romper a barreira específica e colonizar uma nova espécie; isto é, quais seriam os vetores responsáveis pela transmissão do TE. Existem diversas possibilidades de transmissão, desde transmissão direta até o envolvimento de vetores dos mais variados tipos. É importante ressaltar que cada uma dessas possibilidades pode ocorrer de modo independente, mas que combinações de diferentes vetores também podem ter ocorrido durante a transmissão de alguns TE.

Os procariotos também contêm TE em seus genomas

As sequências de inserção (IS – do inglês, *insertion sequences*) são TE que habitam os genomas procarióticos, tanto de bactérias como de arqueias (Figura 17.21 A). São elementos equivalentes aos transposons de DNA dos eucariotos, mas mais compactos, geralmente menores que 2,5 Kpb. Apresentam em sua estrutura uma ou mais ORF flanqueadas por repetições invertidas (IR – do inglês, *inverted repeats*). Enquanto a ORF codifica uma transposase, as IR são as sequências reconhecidas pela transposase durante a transposição. Ao se inserirem, geram as TSD, como ocorre com a maioria dos TE eucarióticos. Os IS podem ser encontrados tanto nos cromossomos bacteroides como nos plasmídeos. Esses elementos já foram descritos em quase todas as bactérias já estudadas e dividem-se em famílias que são designadas com base na similaridade de sequência de sua transposase e na sua estrutura, como número e disposição de ORF, similaridade e tamanho das IR e TSD.

Além dos IS, transposons de DNA e fagos também são encontrados nos genomas procarióticos. Esses elementos são considerados importantes agentes de plasticidade desses genomas. Assim como nos genomas eucarióticos, sua atividade pode gerar alterações tanto estruturais quanto

funcionais, seja pela inativação de um gene por inserção ou pela modificação da expressão de um gene vizinho. São também responsáveis por rearranjos, que ocorrem através de recombinação homóloga. Há evidências de que esses eventos mutagênicos são responsáveis por um embaralhamento dos genes, capaz de levar à formação de agrupamentos de genes especializados em diversas funções, tais como virulência, simbiose ou até mesmo o estabelecimento de novas rotas metabólicas. Esses agrupamentos de genes, contendo TE em suas extremidades, também são conhecidos como ilhas gênicas, e podem se movimentar dentro de genomas e plasmídeos como se fossem um só TE, ou, ainda, serem transmitidos entre organismos, por meio de transferência horizontal.

Esses eventos de transferência horizontal têm um impacto especialmente significativo sobre os genomas bacterianos, promovendo uma troca de material genético que pode mudar características de resposta ambiental ou patogênicas, por exemplo, gerando especiação (Figura 17.21 B). Comparando-se três espécies de arqueias do gênero *Sulfolobus*, observou-se que elas apresentam um conjunto de genes conservados em comum; no entanto, a ordem desses genes não era conservada. A principal diferença entre esses três genomas são os TE ali presentes, IS e MITE, assim como rearranjos decorrentes de sua atividade. Já foi relatado que, para bactérias do gênero *Xanthomonas*, que pertencem à classe *Gamma-Proteobacteria*, aproximadamente 40% dos genes foram adquiridos por transferência horizontal, uma vez que apresentam maior similaridade de sequência com genes de bactérias não classificadas nessa classe. Esses exemplos ilustram como a presença de TE nos genomas bacterianos e sua associação a ilhas gênicas está diretamente ligada à diversificação desses microrganismos, cumprindo um importante papel em sua evolução.

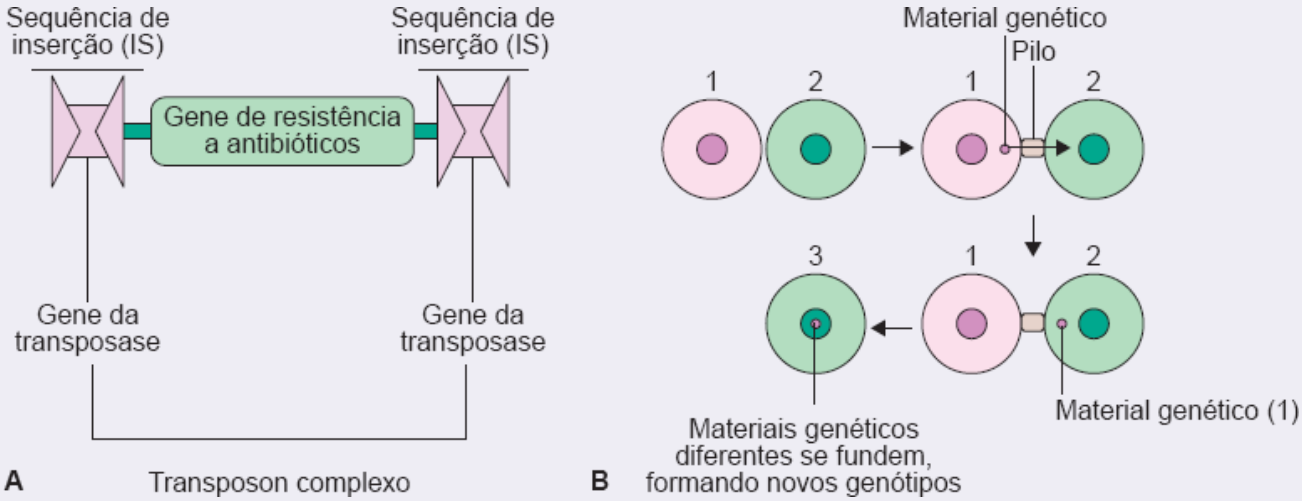


Figura 17.21 A. Ilustração de dois IS e de um transposon complexo, típicos de bactérias. Cada IS é flanqueado por sequências invertidas repetidas, e codifica para uma transposase responsável pela sua mobilização. Um transposon complexo é composto por dois IS em cada extremidade e parte do genoma hospedeiro na região intermediária. Essa região pode conter genes codificadores de proteínas, como genes para resistência a antibióticos. **B.** Transferência horizontal. Duas bactérias (1 e 2) têm material genético distinto. Por meio da formação de um pilo entre as bactérias, elas são capazes de trocar material genético (p. ex., IS). A bactéria recipiente do material genético incorpora esse novo DNA ao seu próprio.

Nos casos mais simples, os TE poderiam invadir uma nova espécie sem a necessidade de qualquer vetor. Neste caso, retrotransposons e transposons teriam diferentes modos de transferência, ambos envolvendo o processo de mobilização. Para os retrotransposons, a transmissão aconteceria durante o período em que o cDNA recém-sintetizado fica encapsulado na VLP (do inglês, *virus-like particle*), que é codificada pelo gene GAG do TE. Já foi demonstrado que os retrotransposons *Gypsy* e *Copia* de *Drosophila* são capazes de produzir VLP. Para os retrotransposons com LTR que contêm o gene ENV, o processo de transferência seria ainda mais fácil, pois o próprio TE seria capaz de produzir todos os componentes necessários para invasão de novas células. Por outro lado, na TH direta dos transposons de DNA, a transmissão ocorreria pela formação de um episssomo estável, uma molécula de DNA circular originada por recombinação e que pode ser passada de uma célula para outra.

Uma segunda possibilidade seria o TE "pegar uma carona" na estrutura infecciosa de um vírus capaz de infectar mais de uma espécie. Um estudo de sequenciamento do material genético contido no capsídeo do *flockhouse virus* de insetos mostrou que mais de 5% do conteúdo do RNA era composto de material genético derivado de TE; entre os quais, retrotransposons com e sem LTR, e transposons de DNA.

Parasitas, parasitoides e simbiontes também poderiam atuar como vetores de TH. Acredita-se que o primeiro caso bem documentado de TH aconteceu na década de 1950 e envolveu o transposon de DNA *P*, a espécie receptora *D. melanogaster*, a espécie doadora *D. willistoni* e um dos parasitas mais comuns de drosofilídeos, o ácaro *Proctolaelaps regalis*. Esse ácaro se alimenta sugando ovos dessas moscas, mas não necessariamente mata cada um

dos ovos de que se alimenta. Assim, pode ter ocorrido troca de material genético entre diferentes espécies de *Drosophila* através dos ovos. Fortes evidências dessa TH são o fato de que os elementos dessas espécies – que divergiram há pelo menos 30 milhões de anos – são quase 99,97% idênticos, e que populações naturais de *D. melanogaster* coletadas antes da década de 1950 e mantidas em laboratório não contêm o elemento *P*.

Propagação

A etapa de propagação é uma das menos compreendidas no ciclo evolutivo dos TE. Acredita-se que logo após a origem de uma nova família de TE, esses elementos consigam escapar dos mecanismos de regulação. Isso aconteceria porque a taxa da transposição é principalmente regulada pelos mecanismos de iRNA, que dependem da complementaridade de pequenos RNA preexistentes no citoplasma e dos transcritos produzidos pelo elemento. A ausência ou até mesmo a baixa quantidade desses pequenos RNA acarretaria uma alta taxa de mobilização e uma consequente amplificação do número de cópias dos novos elementos. A amplificação de TE também tem consequências para o genoma, pois, durante sua mobilização, podem acontecer rearranjos cromossômicos e todo tipo de mutação por inserção. Como consequência do aumento do número de cópias, ocorreria uma retroalimentação dos mecanismos geradores de pequenos RNA, que culminaria na regulação e diminuição da taxa de mobilização desse elemento.

TE cumprem o papel das telomerases em *Drosophila*

Telômeros são estruturas localizadas nas extremidades dos cromossomos eucarióticos, com a função de proteger suas pontas da degradação e fusão, de modo a garantir a replicação fiel do material genético, mantendo a estabilidade e a integridade do genoma. São formados por sequências de DNA repetidas, ligadas a uma série de proteínas. A telomerase é um complexo ribonucleoproteico, com a função de polimerase dependente de RNA, responsável pela manutenção dos telômeros por meio da adição de repetições de nucleotídeos às pontas dos cromossomos. O complexo da telomerase é formado basicamente pela transcriptase reversa TERT (do inglês, *telomeric reverse transcriptase*) e por um RNA que serve de molde para a transcrição reversa TERC (do inglês, *telomeric RNA component*). TERT sintetiza os novos fragmentos de DNA usando TERC como molde. A adição das repetições nucleotídicas serve para compensar a perda de sequências da extremidade dos cromossomos, que ocorre a cada ciclo de replicação.

Moscas do gênero *Drosophila*, no entanto, perderam a telomerase. Quem cumpre esse importante papel na manutenção dos telômeros são três retrotransposons especializados para essa função: HeT-A, TART e TAHRE (Figura 17.22). Seus telômeros são compostos por longas sequências de cópias *in tandem* desses TE. Sucessivos eventos de retrotransposição nas extremidades dos cromossomos garantem a adição de repetições desses TE, fazendo cópias de si mesmos. Essas repetições são análogas às repetições adicionadas pelas telomerases em outros organismos, ainda que sejam mais irregulares.

HeT-A, TART e TAHRE são retrotransposons sem LTR da superfamília Jockey. São os únicos membros dessa superfamília a compor telômeros; eles se transpõem apenas nas regiões teloméricas, inserindo-se na extremidade 5′ de outros elementos já presentes ali. Talvez não tenha sido por acaso que esses elementos assumiram esse papel. O modo de funcionamento das telomerases se assemelha muito à transposição dos retrotransposons sem LTR, com uma quebra da fita simples de DNA expondo sua extremidade 3′, que servirá de iniciador para a síntese de DNA pela transcriptase reversa, tendo como molde a fita de RNA intermediário. Esses três TE se transpõem nos telômeros de maneira randômica, o que sugere papéis equivalentes na adição de sequências aos telômeros. Sua inserção ocorre de forma polarizada, de modo que a sua extremidade 5′ fica sempre voltada para a ponta do cromossomo. Em muitos casos, a extremidade 5′ é truncada. Há evidências evolutivas de que esses TE estão presentes nos telômeros de *Drosophila* há pelo menos 60 milhões de anos, uma data anterior à separação das espécies existentes no gênero.

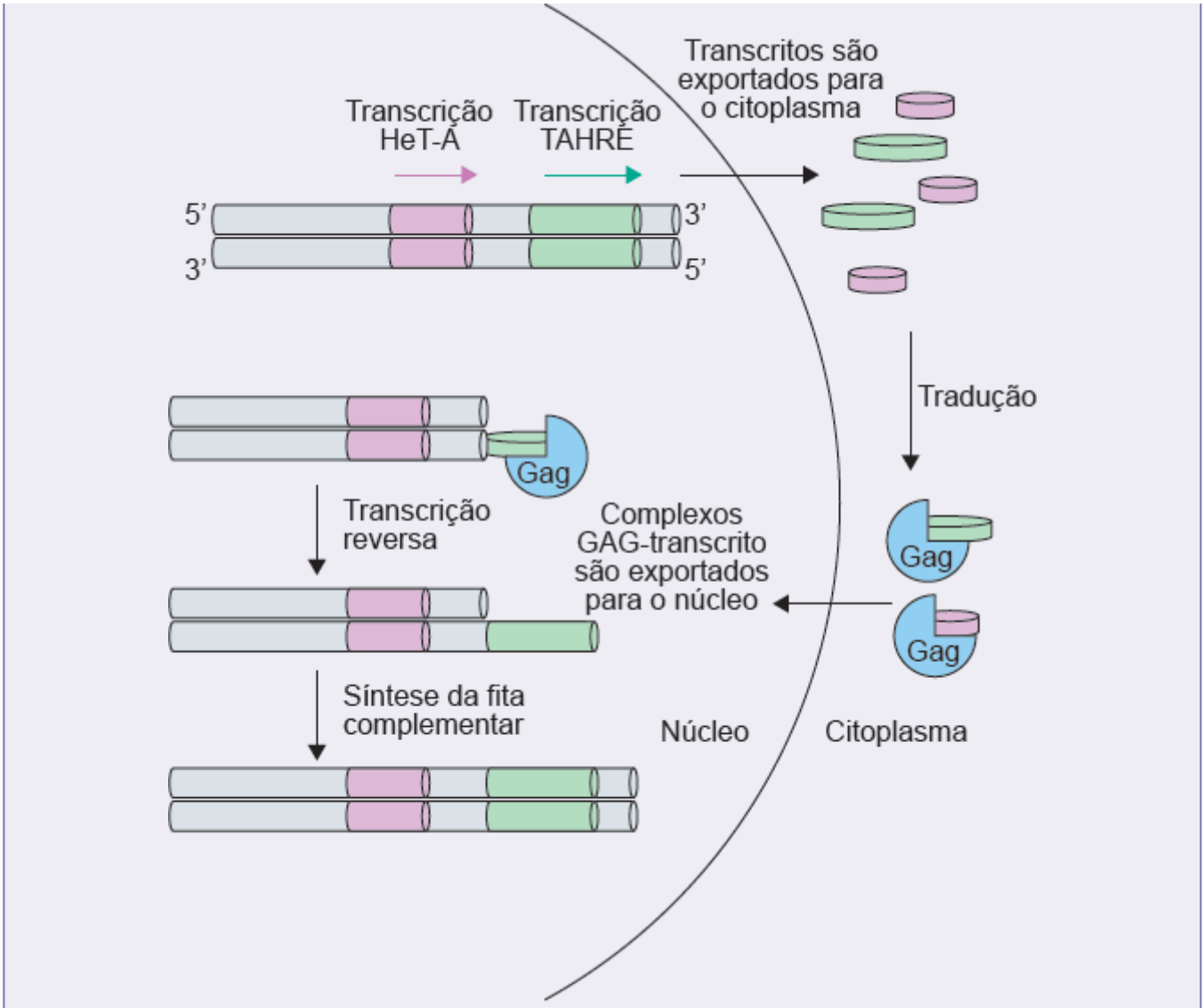


Figura 17.22 Retrotransposons de *Drosophila* reconstituem os telômeros na ausência de telomerase. Inicialmente, os elementos HeT-A e TAHRE são transcritos, e a transcrição é promovida pela região promotora contida a 3' do elemento. Os transcritos são enviados ao citoplasma, onde são traduzidos. Os retrotransposons codificam para enzimas como a GAG, que "sequestram" uma fita de transcrito do elemento que a codificou. O complexo GAG-transcrito é então enviado de volta ao núcleo. Lá, o complexo é acoplado à extremidade final do cromossomo. A enzima transcriptase reversa sintetiza a primeira fita de DNA com base no transcrito acoplado à GAG. Aí, então, é feita a síntese da segunda fita de DNA, completando a adição de uma nova cópia do retrotransposon, assim estendendo as extremidades cromossômicas. Adaptada de Mason *et al.*[18]

Destino

As diversas cópias de um TE produzidas durante a fase de propagação podem ter destinos diferenciados. Cópias de elementos de uma mesma família de TE tendem a acumular mutações, derivadas tanto de erros no momento da síntese dos elementos durante a mobilização quanto de erros durante a replicação do DNA na divisão celular de seus hospedeiros. Essas mutações podem levar à inativação dos genes do elemento, impedindo a produção das proteínas necessárias à mobilização. Ao mesmo tempo, pode ocorrer a degradação das sequências de nucleotídeos sinalizadoras da mobilização (p. ex., dos sítios de reconhecimento da transposase para os transposons de DNA, ou da integrase para os retrotransposons). Assim, os elementos podem acumular mutações ao ponto de se tornarem inativos ou até mesmo irreconhecíveis como TE por meio da sua sequência de DNA. Esse processo pode chegar à eliminação total de um TE, por deriva genética ou seleção natural.

De modo alternativo, os TE seriam a domesticação ou a reativação. A domesticação já foi discutida no item "Os TE podem ser doadores de domínios proteicos para a formação de novos genes" deste capítulo, e é caracterizada pelo recrutamento de elementos completos ou partes de elementos no desempenho de novas funções, diretamente relacionadas ao hospedeiro (ver boxe "TE cumprem o papel das telomerases em *Drosophila*"). Finalmente, existe a

possibilidade de reativação de um TE pelo acúmulo de mutações, possibilidade que levaria à origem de novos elementos e reinício do ciclo evolutivo dessas sequências (ver Figura 17.20).

Conclusões e perspectivas | Revolução do mobiloma

Este capítulo abordou os principais temas relacionados aos TE. O estudo desses elementos sofreu uma grande propulsão na última década, graças ao advento do sequenciamento de genomas em grande escala. As novas descobertas a respeito desses intrigantes componentes dos genomas têm demonstrado sua grande importância na evolução dos seres vivos. No entanto, a compreensão dos mecanismos pelos quais os TE atuam sobre os genomas, moldando-os tanto estrutural quanto funcionalmente, ainda constitui um campo novo dentro da Biologia.

No que diz respeito ao conceito sobre os TE, muito já se avançou desde as primeiras observações feitas por Barbara McClintock até os dias atuais. Inicialmente, a comunidade científica teve grande dificuldade em aceitar que um elemento do genoma poderia ser móvel, uma vez que o DNA é o portador da herança genética e deveria se manter o mais estável possível, a fim de garantir a integridade da herança passada para as gerações futuras. A evolução das descobertas científicas veio a confirmar a característica móvel dos TE e corroborar a plasticidade dos genomas.

Por não codificarem proteínas com um papel conhecido no funcionamento da célula e manutenção dos organismos, os TE foram inicialmente considerados como DNA lixo. Dentro desse conceito, esses elementos não teriam utilidade nenhuma para os genomas e organismos. Sendo considerados entidades gênicas, que apenas usam a maquinaria celular para perpetuar suas cópias dentro do genoma, ao modo dos vírus, esses elementos foram rotulados parasitas dos genomas. No entanto, as crescentes descobertas a respeito dos TE modificaram esse conceito. A grande diversidade e a representatividade desses elementos dentro dos genomas, bem como o significativo impacto que causam sobre o seu funcionamento e evolução, tornam possível afirmar que a relação estabelecida entre os TE e os genomas não é o parasitismo, mas sim um tipo de mutualismo. No primeiro caso, apenas o TE se beneficiaria da relação com o genoma, perpetuando sua existência e causando prejuízos, como consequência de sua atividade mutagênica. Já no segundo caso, admite-se que os TE adquiriram um papel tanto de agente como de substrato para a evolução dentro dos genomas. Apesar de haver um custo para a célula em manter a atividade dos TE regulada, a coexistência entre eles e os genomas tem sido vantajosa para ambos ao longo da evolução. São elementos intrínsecos aos genomas, e importantes para a manutenção da vida como a conhecemos.

Bibliografia

Almeida LM, Carareto CMA. Origem, proliferação e extinção de elementos transponíveis: qual seria a importância da transferência horizontal na manutenção desses ciclos? Ribeirão Preto: Sociedade Brasileira de Genética; 2005.

Castel SE, Martienssen RA. RNA interference in the nucleus: roles for small RNAs in transcription, epigenetics and beyond. Nature Reviews Genetics. 2013;14(2):100-12.

Chenais B, Caruso A, Hiard S, Casse N. The impact of transposable elements on eukaryotic genomes: from genome size increase to genetic adaptation to stressful environments. Gene. 2012;509(1):7-15.

Fedoroff NV. McClintock's challenge in the 21 st century. Proceedings of the National Academy of Sciences. 2012;109(50):20200-3.

Feschotte C, Jiang N, Wessler SR. Plant transposable elements: where genetics meets genomics. Nature Reviews Genetics. 2002;3(5):329-41.

Feschotte C, Pritham EJ. DNA transposons and the evolution of eukaryotic genomes. Annual Review of Genetics. 2007;41:331-68.

Feschotte C. The contribution of transposable elements to the evolution of regulatory networks. Nature Review Genetics. 2008;9(5):397-405.

Finnegan DJ. Retrotransposons. Current Biology. 2012;22(11):R432-R437.

Geurts AM, Yang Y, Clark KJ, Liu G, Cui Z, Dupuy AJ, Bell JB, Largaespada DA, Hackett PB. Gene transfer into genomes of human cells by the sleeping beauty transposon system. Molecular Therapy. 2003;8(1):108-17.

Grandbastien MA, Audeon C, Bonnivard E, Casacuberta JM, Chalhoub B, Costa APP, et al. Stress activation and genomic impact of Tnt retrotransposons in solanaceae. Cytogenetic and Genome Research. 2005;110:229-41.

Kumar A, Bennetzen JL. Plant retrotransposons. Annual Review of Genetics. 1999;33:479-532.

Lisch D. How important are transposons for plant evolution? Nature Review Genetics. 2013;14(1):49-61.

Mason JM, Frydrychova RC, Biessmann H. Drosophila telomeres: an exception providing new insights. Bioessays. 2008;30(1):25-37.

McClintock B. The significance of responses of the genome to challenge. Science. 1984;226:792-801.

Munoz-Lopez M, Garcia-Perez JL. DNA transposons: nature and applications in genomics. Current Genomics. 2010;11(2):115-28.

Rohde C, Valente VLS. Three decades of studies on chromosomal polymorphism of Drosophila willistoni and description of fifty different rearrangements. Genetics and Molecular Biology. 2012;35(4):966-79.

Sabot F, Schulman AH. Parasitism and the retrotransposon life cycle in plants: a hitchhiker's guide to the genome. Heredity. 2006;97(6):381-8.

Schaack S, Gilbert C, Feschotte C. Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. Trends in Ecology and Evolution. 2010;25(9):537-46.

Siomi MC, Sato K, Pezic D, Aravin AA. PIWI-interacting small RNAs: the vanguard of genome defence. Nature Reviews Molecular Cell Biology. 2011;12(4):246-58.

Slotkin RK, Martienssen R. Transposable elements and the epigenetic regulation of the genome. Nature Review Genetics. 2007;8(4):272-85.

Volff JN. Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. Bioessays. 2006;28(9):913-22.

Wallau GL, Ortiz MF, Loreto EL. Horizontal transposon transfer in eukarya: detection, bias, and perspectives. Genome Biology and Evolution. 2012;4(8):689-99.

Wessler SR. Transposable elements and the evolution of eukaryotic genomes. Proceedings of the National Academy of Sciences. 2006;103(47):17600-1.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. Nature Review Genetics. 2007;8(12):973-82.

Parte 6

Genética Médica