Manuscrito para capítulo do Biowork IV

Bioinformática aplicada à Genômica

Fabrício R. Santos[1] e José Miguel Ortega[2]

1 Departamento de Biologia Geral e 2 Departamento de Bioquímica e Imunologia da Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brasil.

Autor para correspondência: Prof. Fabrício R. Santos
Departamento de Biologia Geral, ICB, UFMG,
Av. Antônio Carlos 6627, CP 486
31270-010, Belo Horizonte, MG, Brasil.
Tel: +55 31 3499-2581. Fax: +55 31 3499-2570
e-mail: fsantos@mono.icb.ufmg.br

Introdução

Com o início do Projeto Genoma Humano em 1990 e subseqüente disponibilização de seqüenciadores automáticos de DNA capazes de gerar dados genômicos em grande escala, os bancos de dados e ferramentas de análise tiveram de se adaptar a este volume crescente de informações. Seqüências de nucleotídios são adicionadas aos bancos de dados (como o GenBank) na ordem de milhares de pares de bases (pb) por segundo todos os dias. Nos serviços de bioinformática de projetos genoma, essas inúmeras seqüências individuais, cada uma portando geralmente entre 400 a 1000 pb, devem ser montadas em seqüências cada vez maiores, os *contigs*, através de ferramentas que avaliam a qualidade das seqüências individuais e a superposição destas, para que finalmente sejam disponibilizados segmentos cromossômicos inteiros de alta qualidade. Para a cobertura total de um genoma com boa qualidade estima-se que este deva ser seqüenciado ao equivalente a dez vezes seu tamanho em pares de bases. O dito "rascunho de trabalho" do genoma humano contém cerca de 20% da informação assim tratada e o restante com uma cobertura de cinco vezes, o que inclusive demandou um esforço bioinformático ainda maior para sua montagem. Espera-se para 2003 o mapa completo de alta qualidade com 24 segmentos de cada tipo de cromossomo humano (1-22, X e Y). No mapa físico de seqüências, as diferentes regiões devem ser interpretadas com respeito à sua função, através de um processo denominado anotação genômica. A homologia existente entre genes presentes em diversos organismos é utilizada na anotação de função; assim um gene caracterizado numa levedura pode ajudar na identificação funcional do gene com a mesma função - denominado ortólogo, no homem, por exemplo. Vários algoritmos distintos foram desenvolvidos para facilitar o processo de anotação nas suas várias etapas. Neste processo são identificados os vários tipos de seqüências repetitivas (transposons, micro e minissatélites, etc.), seqüências estruturais (centrômeros, telômeros, heterocromatina, satélites, etc.), seqüências regulatórias (promotores, *enhancers*, etc.) e regiões transcritas que correspondem aos genes de cada organismo. Vale a pena notar que a presença dos íntrons nos organismos que os contém , como no homem, dificultam em muito a anotação do genoma, sendo nestes casos muito importante a existência de

projetos de seqüenciamento do transcriptoma. Este pode ser definido como o conjunto de seqüências expressas de um genoma na forma de mRNA, que pode ser seqüenciado a partir de bibliotecas de cDNAs preparadas com o auxílio da enzima transcriptase reversa que converte RNA em DNA. Diferentemente do seqüenciamento do genoma, a análise do transcriptoma exige a investigação de várias células e tecidos diferentes, bem como de distintos estágios do desenvolvimento, para que se detecte o maior número possível de genes. Com essas seqüências em mãos, é facilitada a procura de genes no DNA genômico, proporcionando também a correta identificação dos íntrons.

Dados biológicos advindos do conhecimento genômico são relativamente complexos em comparação aos provenientes de outras áreas científicas, dada a sua diversidade e ao seu inter-relacionamento (figura 1). A partir do conhecimento fundamental do genoma objetiva-se compreender o conjunto de peças que atuam no funcionamento complexo de todo o organismo. Porém, no momento, isso somente é possível por partes. Busca-se entender as estruturas moleculares das proteínas, as interações entre várias proteínas, bem como destas com as demais moléculas biológicas (DNA, carboidratos, lipídios, etc), as diversas vias metabólicas celulares e o papel da variabilidade genética representada pelas várias formas de cada proteína. Toda essa informação disponibilizada pela ciência genômica (figura 1) só é possível de ser organizada, analisada e interpretada com o apoio da informática. Um novo projeto ambicioso denominado *Genomes to life* foi recentemente lançado pelo Departamento de Energia dos EUA (o mesmo que lançou a idéia do Projeto Genoma, em 1987) e objetiva chegar a uma compreensão fundamental e sistemática sobre a vida, através dos genomas que estão sendo descritos. Uma das idéias deste mega-projeto é reconstituir *in-silico* (no computador) o funcionamento de um microorganismo com todas suas funções biológicas.

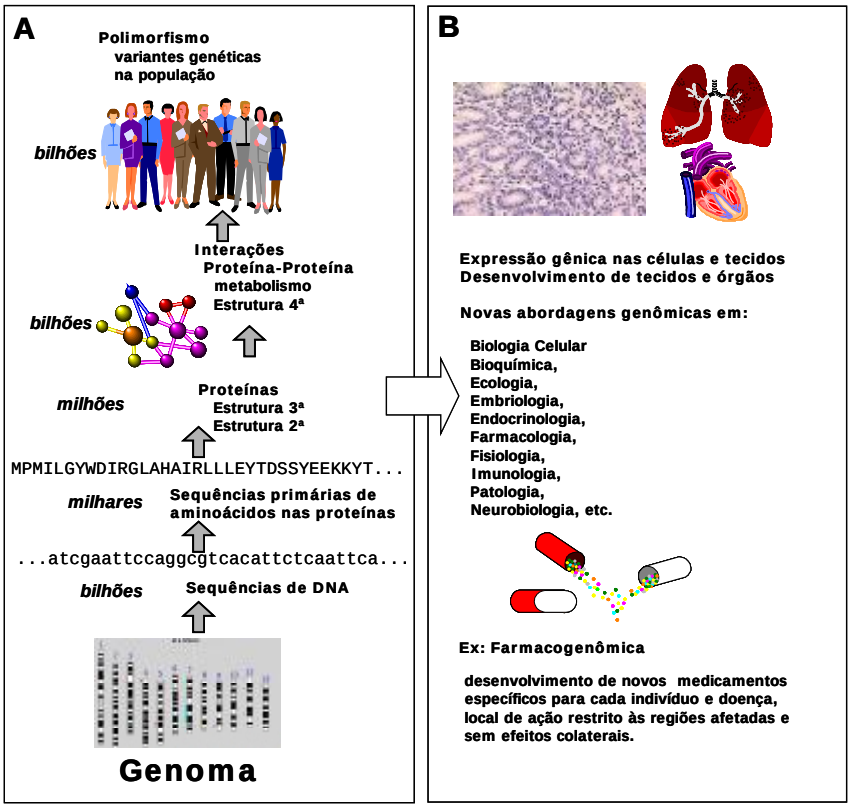


Figura 1 – Acúmulo de dados biológicos (A) e aplicações do conhecimento genômico (B).

Atualmente a bioinformática é imprescindível para a manipulação dos dados biológicos. Ela pode ser definida como uma modalidade que abrange todos os aspectos de aquisição, processamento, armazenamento, distribuição, análise e interpretação da informação biológica. Através da combinação de procedimentos e técnicas da matemática, estatística e ciência da computação são elaboradas várias ferramentas que nos auxiliam a compreender o significado biológico representado nos dados genômicos. Além disso, através da criação de bancos de dados com as informações já processadas, acelera a investigação em outras áreas como a medicina, a biotecnologia, a agronomia, etc (Borém e Santos, 2001).

Bancos de dados Genômicos

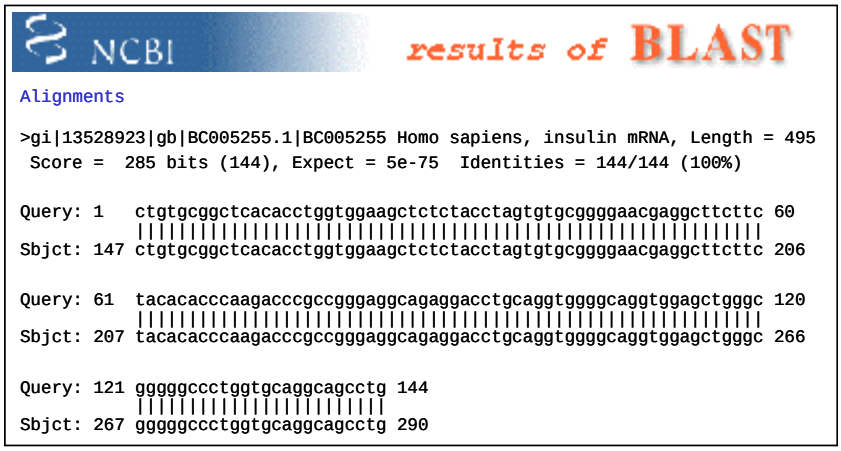
Devido a essa imensa quantidade de dados gerados em inúmeros laboratórios de todo o mundo, faz-se necessário organizá-los de maneira acessível, de modo a evitar

redundância na pesquisa científica e possibilitar a análise por um maior número possível de cientistas. A construção de bancos de dados para armazenamento de informações de seqüências de DNA e genomas inteiros, proteínas e suas estruturas tridimensionais, bem como vários outros produtos da era genômica, tem sido um grande desafio, mas simultaneamente extremamente importante.

O NCBI, ou Centro Nacional para Informação Biotecnológica dos EUA, é considerado o banco de dados central sobre informações genômicas. Vários outros bancos de dados similares estão distribuídos por países da Europa e Japão, mas todos trocam dados em um intervalo de 24 horas com o NCBI. O GenBank é o principal banco de dados do NCBI e armazena todas seqüências disponíveis publicamente de DNA (de seqüências pequenas a genomas inteiros), RNA e proteínas. Além do GenBank, que coleta todas as entradas de seqüências, outros bancos do NCBI apresentam as informações organizadas de diferentes maneiras. Por exemplo, o UniGene agrupa todas as seqüências parciais do transcriptoma de um organismo em aglomerados ou *clusters*, onde cada aglomerado representa a seqüência consenso de um gene. Também no NCBI, o banco de dados RefSeq reúne somente as seqüências de referência, ou seja, a mais representativa seqüência de um transcrito, editada e inspecionada por um curador. É, freqüentemente, o melhor banco de dados para se evitar a redundância natural num universo com tantas informações. Para acesso ao RefSeq e outros bancos de seqüências curadas foi desenvolvida a ferramenta LocusLink no NCBI. Outros bancos são específicos de um organismo, tal como o OMIM (*Online Mendelian Inheritance in Man*) que foi criado para catalogar todos genes e alelos relacionados a doenças e outras características humanas, bem como proporcionar um detalhamento técnico e bibliografia referente a cada característica. A existência destes bancos de dados, ditos secundários, têm sido tão importante quanto preservar os dados originais no GenBank.

Várias ferramentas desenvolvidas pela bioinformática permitem o acesso e análise dos dados no GenBank. A ferramenta mais popular de comparação de seqüências de DNA com os bancos de dados genômicos é o BLAST ou *Basic Local Alignment Search Tool*. Através deste algoritmo podemos comparar uma seqüência de DNA ou proteína (*Query*) qualquer com todas seqüências genômicas de domínio público. É importante notar que o

programa BLAST não procura conduzir uma comparação da extensão total das moléculas comparadas, mas apenas identificar, no banco de dados, a presença de uma seqüência suficientemente parecida com a pesquisada. Descarta, assim, rapidamente, os resultados não produtivos e estende a vizinhança da região de homologia detectada até não mais conseguir. O resultado desta busca, que é feita no GenBank ou em várias de suas subdivisões (pode-se facilmente limitar a pesquisa a seqüências de um dado organismo, por exemplo), retorna aquelas seqüências (DNA ou proteínas) depositadas (*Subject*) com maior homologia. Desta forma várias regiões de DNA podem ser anotadas através do BLAST, cujo resultado pode servir para atribuir uma função a qualquer segmento de DNA que apresenta homologia significativa a outras seqüências de DNA ou proteínas previamente depositadas no GenBank com função conhecida experimentalmente (figura 2). É interessante verificar que se utilizássemos um nucleotídio, "A" por exemplo, para pesquisar seqüências humanas, a chance de encontrarmos uma região homóloga seria igual a 1 (100%). Se a nossa seqüência pesquisada fosse mais complexa, 144 bases por exemplo, a chance de encontrarmos uma seqüência perfeitamente idêntica seria pequena. O valor de "E" , um parâmetro calculado pelo BLAST, expressa essa dificuldade e, quanto menor seu valor, menor a chance de tal comparação ter sido encontrada por pura coincidência.


NCBI results of BLAST

Alignments

```
>gi|13528923|gb|BC005255.1|BC005255 Homo sapiens, insulin mRNA, Length = 495
 Score =  285 bits (144), Expect = 5e-75  Identities = 144/144 (100%)

Query: 1   ctgtgcggctcacacctggtggaagctctctacctagtgtgcggggaacgaggcttcttc 60
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 147 ctgtgcggctcacacctggtggaagctctctacctagtgtgcggggaacgaggcttcttc 206

Query: 61  tacacacccaagacccgccgggaggcagaggacctgcaggtggggcaggtggagctgggc 120
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 207 tacacacccaagacccgccgggaggcagaggacctgcaggtggggcaggtggagctgggc 266

Query: 121 gggggccctggtgcaggcagcctg 144
           |||||||||||||||||||||||
Sbjct: 267 gggggccctggtgcaggcagcctg 290
```

Figura 2 - Resultado da busca por similaridade com o programa BLAST. O segmento de DNA seqüenciado (*Query*) demonstrou alta homologia (100%) com o gene da Insulina humana (*Sbjct*).

Há várias modalidades de BLAST. A mais curiosa e de grande importância na descoberta gênica é aquela onde tanto a *Query* como a base de dados (*Subject*) são seqüências de nucleotídios. Neste programa, antes de verificar a homologia, são feitas as seis traduções possíveis de cada seqüência de nucleotídios, ou seja, tanto a seqüência pesquisada quanto cada uma das presentes na base de dados são transformadas em seis proteínas (iniciando pela base 1, 2 ou 3 de cada fita). Essa modalidade, denominada tBLASTx, permite que seja retornado o par proteína *Query* - proteína *Subject* e é muito válida pois as proteínas de dois organismos são mais parecidas entre si que os nucleotídios que as codificam. Nesta análise, apenas uma das seis leituras é de significado biológico, as demais geram resultados que são desprezados. O tBLASTx foi utilizado em descoberta gênica inúmeras vezes, como por exemplo na identificação da subunidade catalítica da telomerase humana assim que tal enzima foi identificada no protozoário *Euplotes* (Meyerson et al. 1997). Outras modalidades buscam homologia entre seqüências de nucleotídios (BLASTn), seqüências de proteínas (BLASTp) ou entre seqüências de nucleotídios e proteínas (BLASTx). Uma outra variedade de BLAST é o PSI-BLAST, que em uma primeira busca encontra as proteínas mais homólogas à pesquisada - *Query*; procede identificando as regiões conservadas dentre os melhores resultados da pesquisa e, em buscas subseqüentes, mascara as regiões não conservadas da *Query* e pesquisa levando em conta apenas as regiões conservadas.

Nos bancos de dados há também uma grande variedade de informações sobre estruturas moleculares, expressão gênica diferencial, diversidade genética, evolução, etc. que podem ser extraídas pela bioinformática. Um dos grandes desafios é o desenvolvimento de procedimentos pelos quais esses dados podem ser “inseridos” e "extraídos" em bancos de dados secundários, pelos pesquisadores. Há várias ferramentas que se encontram disponíveis no próprio NCBI e em outros centros, mas há muito campo para o desenvolvimento de procedimentos específicos. Ferramentas desenvolvidas recentemente incluem bancos de genes classificados de acordo com sua história evolutiva (COG-NCBI), algoritmos de comparação de genomas inteiros (ACT - *Artemis Comparison Tool*), ferramentas de busca de similaridade estrutural de proteínas, independentemente da seqüência primária (VAST-NCBI), etc.

À medida que é feito o seqüenciamento do genoma de muitas espécies, a genômica comparativa assume grande importância e procedimentos computacionais para correlação entre organismos no nível molecular tornam-se essenciais. Pesquisas comparativas têm sido utilizadas para estudos funcionais do genoma, por exemplo da análise dos genes de bactérias *E. coli* patogênicas e não-patogênicas (Perna et al. 2001), para identificação de genes relacionados às doenças que estes provocam (Jimenez-Sanchez et al. 2001), para identificar seqüências de DNA e proteínas que possam ser responsáveis por diferenças entre espécies, tal como entre homem e chimpanzé (Ebersberger et al. 2002). Dentre os procariotos foi demonstrado por genômica comparada que na história evolutiva vários segmentos de DNA foram trocados entre distintas espécies, num processo de transferência horizontal. Outras aplicações das análises comparativas entre genomas estão emergindo: desenvolvimento de tecidos e órgãos, base da resistência a doenças infecciosas, prognóstico de câncer, etc. Para cada um desses propósitos, novas ferramentas de bioinformática são construídas e muitas delas são disponibilizadas via servidores www na *Internet*.

Uma nova disciplina, a farmacogenômica, já possui investimentos pesados de várias empresas para desenvolvimento de novos medicamentos a partir de análises genômicas. Grande parte da pesquisa em farmacogenômica depende da identificação de variações inter-individuais em humanos para a localização de genes relacionados à susceptibilidade ou resistência a doenças ou fármacos. Algumas empresas, tal como a *Orchid BioSciences*, possuem bancos de dados privados contendo estas variações genéticas, na maior parte do tipo SNPs (*Single Nucleotide Polymorphisms*) que correspondem a variantes em uma única posição nucleotídica. O NCBI possui um banco de dados de SNPs de diferentes organismos, sendo que na espécie humano são mais de 4 milhões catalogados. A Celera investiu fortemente na identificação de SNPs de camundongo para aplicações na farmacogenômica. A partir das coleções de SNPs pode-se estudar com métodos de biologia molecular e ferramentas bioinformáticas as associações entre os distintos alelos e características importantes para o desenvolvimento de novos medicamentos e tratamentos mais precisos e sem efeitos colaterais.

Mapas Genômicos

Em 1995, o primeiro genoma de um organismo celular foi decifrado por meio do seqüenciamento da bactéria *Haemophilus influenzae* utilizando uma metodologia de “tiro no escuro" (do inglês *shotgun*). Esta estratégia envolve o seqüenciamento totalmente ao acaso, para posterior montagem numa seqüência contígua, ou *contig* (figura 3) e tem-se mostrado extremamente útil para o seqüenciamento de genomas simples, como o de bactérias e, mais recentemente em genomas complexos, como o da drosófila (Adams et al. 2000) e do homem (Venter et al. 2001). Na verdade, o seqüenciamento executado pelo consórcio público do genoma humano também teve uma porção *shotgun* (figura 3). Fragmentos grandes de DNA clonados em BAC (cromossomos artificial de bactéria), de cerca de 150 mil pb, previamente mapeados em lugares específicos dos cromossomos, eram enviados para centros de seqüenciamento ao redor do planeta e, em cada centro, fragmentos pequenos eram gerados por quebras físicas e seqüenciados no escuro, com uma cobertura de até dez vezes. *Contigs* eram montados e geravam a sequëncia do grande fragmento e a informação era devolvida para a montagem final do genoma. Mas o processo *Shotgun* utilizado pela empresa Celera foi diferente: o genoma era fragmentado em pedaços de 2 mil, 10 mil e 50 mil nucleotídios, que após serem clonados, seqüenciavam-se as extremidades destas moléculas. Cada extremidade seqüenciada encontrava alguma sobreposição com alguma outra seqüência da coleção, mas sabendo-se a seqüência das outras extremidades destas duas moléculas, era possível conferir duplamente o resultado das sobreposições. As moléculas longas funcionam como âncoras, onde as extremidades das moléculas maiores podem ser utilizadas não apenas para comprovar a montagem mas também para ligar e ordenar *contigs* bem como direcionar o seqüenciamento para algumas áreas de descontinuidade entre os *contigs*.

A par do procedimento inteiramente *shotgun*, as metodologias convencionais de seqüenciamento (figura 3) utilizam várias etapas de subclonagens (hierarquia de clonagens) que dependem de mapeamentos diversos para ordenamento das seqüências feitas a partir de clones em plasmídios que são montados em segmentos contínuos de DNA (*contigs*), em pedaços cada vez maiores, até o cromossomo completo.

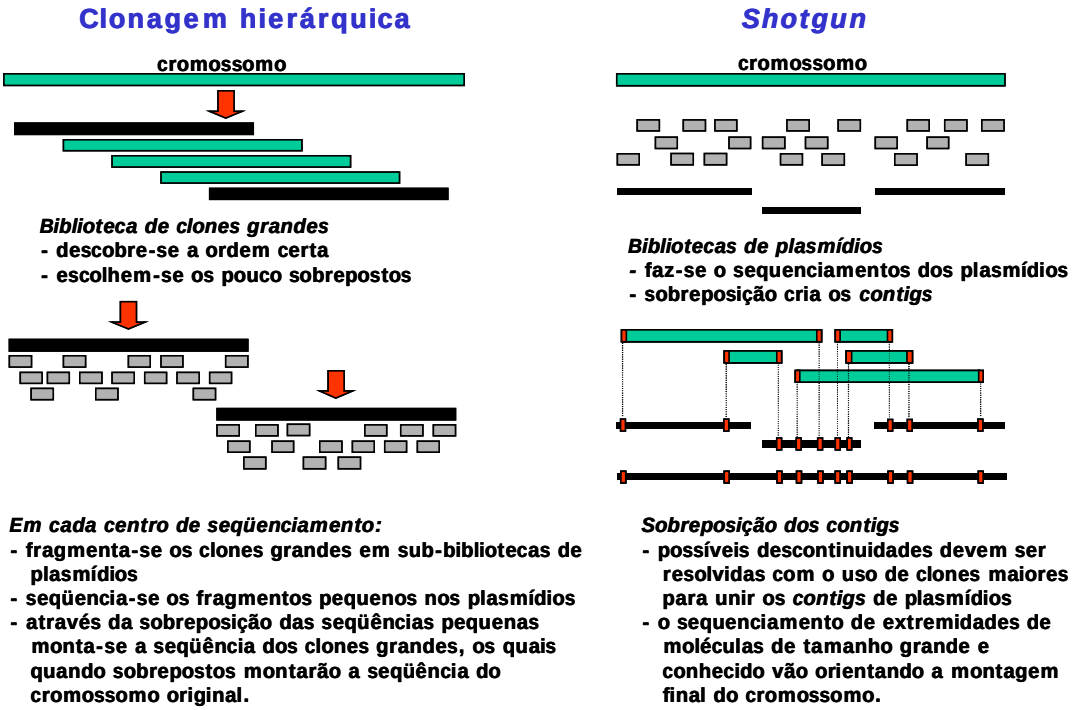


Figura 3 - Seqüenciamento por clonagem hierárquica e por tiro no escuro (*shotgun*)

Para executar essa montagem feita pela superposição das seqüências dos vários clones, novas ferramentas da bioinformática foram construídas. Na figura 4, observa-se que duas dessas ferramentas, o PHRED e o PHRAP, possibilitam a análise das milhares de seqüências de DNA geradas pelo seqüenciador automático. O PHRED verifica a qualidade do seqüenciamento de cada base das várias seqüências e junto ao PHRAP faz o alinhamento de todos os clones, construindo uma seqüência contínua, ou *contig*. No final, vários *contigs* irão compor um grande *contig* que pode ser a fita de DNA completa de um cromossomo de bactéria, que é na maioria dos casos o seu genoma completo. Para a montagem final várias outras ferramentas foram desenvolvidas para manipulação e ordenamento de grandes *contigs*, bem como a visualização do mapa final com toda a anotação funcional (Ex: *Mummy* e *Assembler* do TIGR). Nos eucariotos, cada cromossomo possui uma molécula de DNA e, como humanos têm 24 tipos de cromossomos (1 a 22, X e Y), deve-se seqüenciar completamente 24 dessas moléculas, avançando-se muitas vezes por longos trechos de DNA repetitivo, que são praticamente impossíveis de seqüenciar com perfeição.

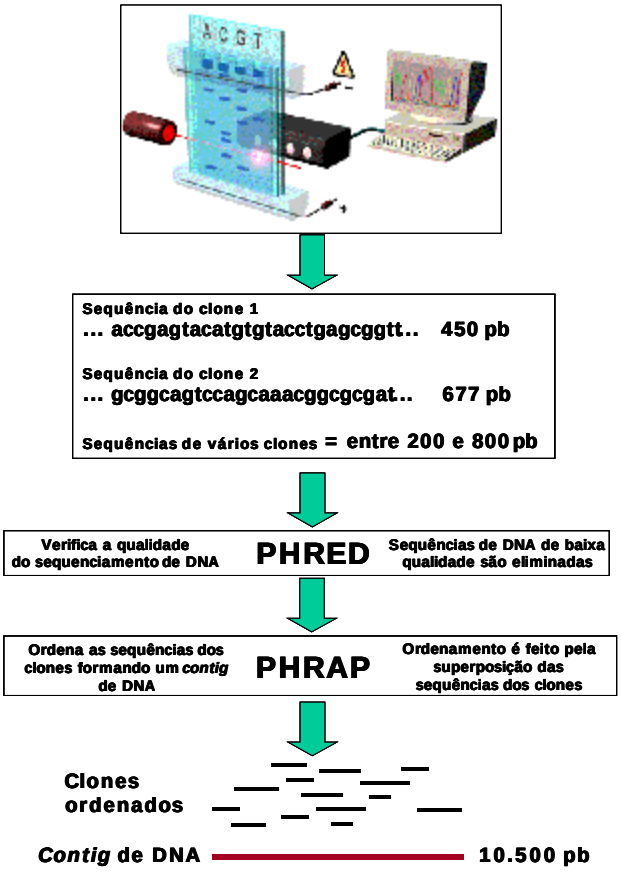


Figura 4 – Montagem de um *contig* pelo PHRED e o PHRAP.

Anotação Genômica e Predição de Genes

O processo de anotação genômica envolve a atribuição de funções e identificação de padrões e de genes na seqüência linear do DNA obtida do seqüenciamento. Toda esta informação está disponível nas diferentes ordens e arranjos das seqüências de DNA.

Encontrar os genes é a principal tarefa da anotação genômica. Para se fazer a predição de genes, vários parâmetros podem ser avaliados tais como a existência de seqüências no DNA que possam funcionar como promotores seguidas por seqüências que possam gerar uma proteína funcional, ou que tenham similaridade com genes conhecidos, etc. Diferentes algoritmos (Ex: GenScan) empregam processos estatísticos diversos para se fazer a busca por ORFs (*Open Reading Frames*) ou fases de leitura aberta do código genético, identificadas por um códon iniciador e um terminador, que correspondem a

seqüências com possíveis regiões codificadoras. Vale notar que a ocorrência no genoma de ORFs superiores a 100 bases é um evento raro, já que um dos 64 códons (ATG) abre a fase de leitura e três são os terminadores (TAA, TAG e TGA), sendo que estes últimos apareceriam com alta probabilidade (3/64), a não ser quando se trata de uma região codificadora. Há também vários programas que detectam o uso não aleatório de códons (*codon usage*), o qual é típico para cada organismo. Nos projetos de análise do transcriptoma (ver abaixo) freqüentemente o códon iniciador não está presente e programas de análise do *codon usage* podem auxiliar no reconhecimento da fase de leitura da porção codificadora. O programa ESTScan é um dos mais usados para esses fins.

O conhecimento prévio da proteína e a sua função em qualquer outra espécie facilita bastante o processo de anotação de genes. No entanto, atualmente, grande parte dos genes são ainda hipotéticos, isto é, não se conhece a função biológica destas seqüências. Por exemplo, na bactéria *Escherichia coli*, na planta *Arabidopsis thaliana* e na mosca das frutas, *Drosophila melanogaster*, entre 40 e 60% dos genes anotados não possuem produto gênico ou função conhecida.

Provavelmente, muitos dos supostos "genes hipotéticos" serão futuramente descartados enquanto outros segmentos gênicos serão identificados após terem passado desapercebidos pelos atuais algoritmos de predição gênica. Este aparente paradoxo resulta do fato de que não existe uma identificação inequívoca de um gene. Por esta razão, várias estimativas do número de genes em diferentes espécies têm sido amplamente divulgadas e freqüentemente apresentavam resultados discordantes. Para o genoma humano acreditava-se até bem pouco tempo em um número estimado ao redor de 70-100 mil genes que foi reduzido para 30-40 mil genes com a publicação dos primeiros rascunhos de nosso genoma em 2001 (Lander et al. 2001 e Venter et al. 2001). Para facilitar a identificação e classificação funcional dos genes foi criado o consórcio *Gene Ontology* que pretende fornecer um vocabulário padronizado para a descrição dos produtos gênicos.

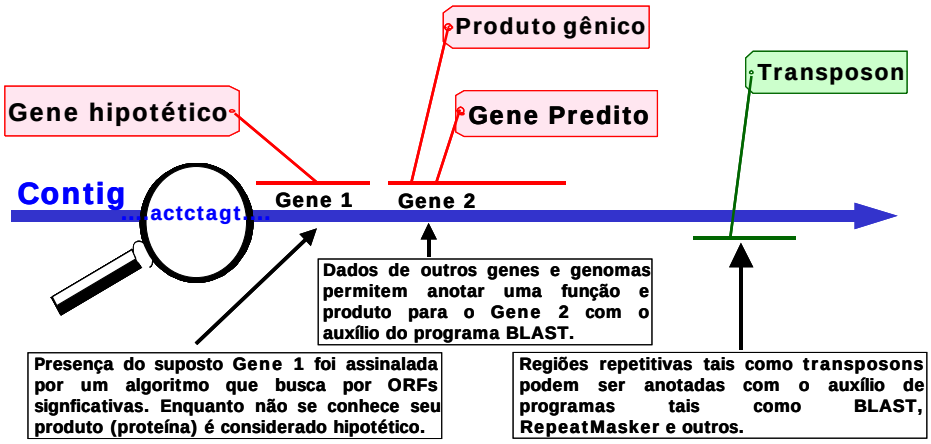


Figura 5 – Processo de anotação de genes

Análise de Transcriptomas

O estudo do transcriptoma de cada organismo é de grande importância para a identificação de genes, mas também incorpora informações sobre o funcionamento do seu genoma. As seqüências produzidas pelos projetos de seqüenciamento do transcriptoma constituem-se em evidência direta da existência de genes com sua determinada ordem de éxons. Por outro lado, a análise de transcriptomas de diferentes espécies, inclusive a humana, tem evidenciado uma altíssima freqüência de processamentos (*splicing*) diferenciais dos transcritos primários. Neste caso, um gene pode apresentar uma grande variação funcional devido simplesmente ao sorteio de éxons promovido pelo processamento diferencial.

Para se estudar o transcriptoma não é necessário seqüenciar completamente todos os genes de um tecido ou organismo. Grande parte dos genes podem ser identificados através da análise de pequenas seqüências que funcionam como etiquetas. Estas seqüências chamadas ESTs, ou *Expressed Seqüence Tags*, são resultado do seqüenciamento parcial de cDNAs (figura 6). O objetivo das ESTs é identificar a presença de genes expressos em um transcriptoma, associando a etiqueta ao gene (e sua função) através um programa tal como o BLAST que faz busca por homologias. Freqüentemente as seqüências parciais (ESTs) se originam de ambas as extremidades do cDNA, embora

alguns projetos prefiram a extremidade 3' por facilitar a geração de seqüências consenso através do agrupamento de vários ESTs, enquanto outros escolhem a extremidade 5' por estar mais próxima da região codificadora da proteína, o que facilita a identificação por homologia. Todavia, uma tecnologia recentemente desenvolvida no Brasil (Dias-Neto et al. 2000) permite o seqüenciamento da região central dos mRNAs. A tecnologia, denominada ORESTES, de *Open Reading frames ESTs* (figura 6) baseia-se na amplificação de cDNAs por PCR aleatório cujos produtos são utilizados para gerar uma biblioteca. O seqüenciamento desta biblioteca, contendo fragmentos aleatórios derivados de diferentes regiões de cada mRNA, favorece o reconhecimento da função do transcrito por pesquisa de homologia, pois incorpora mais freqüentemente a ORF no transcrito do que as ESTs convencionais (figura 6). Os ORESTES foram responsáveis pela identificação de 219 novos genes no cromossomo 22 humano (Souza et al. 2000) que não haviam sido detectados previamente por outras análises bioinformáticas. Além disto, o agrupamento de seqüências para geração de consensos é facilitado quando são utilizados ESTs convencionais associados aos ORESTES. Estes consensos são importantes pois muitas vezes contêm toda região codificadora facilitando o processo de anotação gênica em eucariotos.

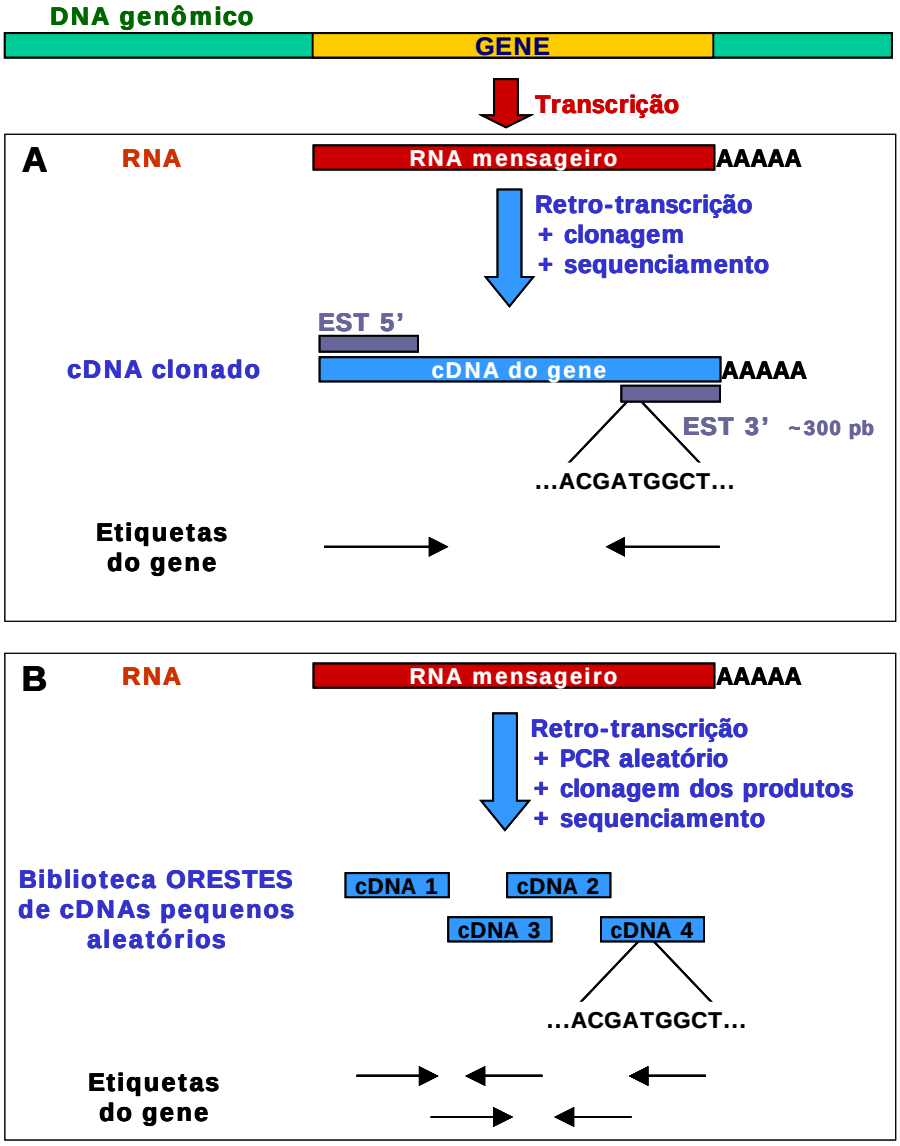


Figura 6 – ESTs (A) e ORESTES (B) utilizados nos projetos transcriptomas

O transcriptoma pode revelar padrões distintos de expressão gênica. Uma das maneiras de se evidenciar a expressão gênica diferencial é analisar a freqüência de ocorrência de um determinado transcrito numa preparação de cDNA de um tecido ou fase de desenvolvimento. Apesar da construção de bibliotecas de cDNA sempre trazer um viés, incorrendo na redundância de alguns transcritos, a análise de várias bibliotecas permite alguma aproximação do padrão de expressão de um tecido ou fase de desenvolvimento de um organismo. Todavia, nada se compara à inversão introduzida pelos microarranjos (*microarrays* ou *biochips*) na análise da expressão gênica. Em uma lâmina de microscópio

podem ser depositados por um robô cerca de 10 a 100 mil seqüências de genes conhecidos. Sondas com fluorescências distintas podem ser preparadas a partir de mRNA isolado de duas populações de células, normais ou transformadas por exemplo, e através da análise da intensidade de hibridização pode-se comparar a expressão gênica diferencial desses múltiplos genes em um tempo extremamente reduzido. Ferramentas bioinformáticas, principalmente voltadas ao processamento de imagens em uma escala micro e nanométrica, estão surgindo para analisar a expressão conjunta de genes, detectadas em microarranjos.

Uma metodologia recente incorpora um nova técnica de biologia molecular e ferramentas de bioinformática para análise de expressão gênica diferencial. O SAGE, ou *Serial Analysis of Gene Expression* (Velculescu et al. 1995), se baseia no uso de pequenas seqüências chamadas *tags* (10 a 14 pb), únicas de cada gene, que são obtidas por etapas de clivagens e ligações com o cDNA e posteriormente co-amplificadas por PCR, formando um concatâmero de *tags*. A quantificação da expressão gênica se dá pela análise do seqüenciamento dos concatâmeros através ferramentas específicas de bioinformática. Desta forma puderam ser identificados vários genes provavelmente relacionados ao processo de transformação celular nos tumores.

Bioinformática no Brasil

No Brasil, o Laboratório de Bioinformática da Unicamp é pioneiro nesta área, desenvolvendo e aplicando várias ferramentas à pesquisa genômica. Este laboratório foi responsável pela montagem, no computador, do genoma do primeiro organismo seqüenciado no País em 2000, a bactéria *Xyllela fastidiosa* (Simpson et al. 2000), causadora da doença do amarelinho-da-laranja.

Vários outros centros de bioinformática têm aflorado no Brasil com a criação de redes nacionais e regionais de seqüenciamento de genomas. No Laboratório Nacional de Computação Científica (LNCC) em Petrópolis, RJ, funciona o Centro de Bioinformática do Projeto Genoma Brasileiro (figura 7), formado por iniciativa do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). Há vários projetos de análise de

transcriptoma em andamento tal como o projeto Genoma Humano do Câncer da FAPESP e o projeto transcriptoma do parasita humano *Schistosoma mansoni* executado pela Rede Genoma de Minas Gerais. O progresso dos vários projetos de genomas no Brasil pode ser acompanhado nesses bancos de dados dos centros de bioinformática que são disponibilizados via *Internet*.

Figura 7 – Logotipo do Projeto Genoma Brasileiro, uma rede nacional de seqüenciamento de DNA financiada pelo CNPq, órgão de fomento à pesquisa do Ministério de Ciência e Tecnologia do Brasil.

Em 2001 a bioinformática foi considerada pela CAPES, órgão brasileiro que coordena o ensino superior, como área prioritária para incentivo de formação na pós-graduação. Em 2002 foi lançado um edital para criação de cursos de pós-graduação nesta área no Brasil, dentro do qual foram selecionados dois programas, um da USP e outro da UFMG. O objetivo inicial seria de formar ao redor de 50 doutores até 2007, refletindo a necessidade crescente destes profissionais nas universidades e institutos de pesquisa.

Referências online

Projetos Genomas

Bancos de dados de genomas

- http://www.ncbi.nlm.nih.gov/Genomes

Projeto Genoma Brasileiro

- http://brgene.lncc.br

Projetos Genomas da FAPESP

- http://watson.fapesp.br/onsa/Genoma3.htm

Projeto Genoma Humano

- http://www.ncbi.nlm.nih.gov/genome/guide/human

Projeto Genomes to Life

- http://doegenomestolife.org

Recursos de Bioinformática

Bancos de dados e ferramentas do NCBI

- http://www.ncbi.nlm.nih.gov

BLAST - ferramenta de busca de homologia por alinhamento local

- http://www.ncbi.nlm.nih.gov/BLAST

Phred, Phrap e Consed - ferramentas para análise da qualidade de seqüências e para montagem e visualização de *contigs*

- http://www.phrap.org

COG - Cluster of Ortolog Groups - Bancos de dados filogeneticamente referenciado.

- http://www.ncbi.nlm.nih.gov

UniGene - Agrupamento de seqüências em consensos de genes.

- http://www.ncbi.nlm.nih.gov/UniGene

LocusLink - ferramenta para recuperação de seqüências funcionais curadas.

- § http://www.ncbi.nlm.nih.gov/LocusLink

Gene Ontology Consortium - banco de dados genômicos para categorização dos genes de acordo com suas funções moleculares, processos biológicos e componentes celulares.

- § http://www.geneontology.org

Orchid BioSciences - empresa da área farmacogenômica

- § http://www.orchid.com

Celera - mega-empresa da área genômica

- § http://www.celera.com

ACT - Artemis Comparison Tool - comparação de genomas inteiros

- § http://www.sanger.ac.uk/Software/ACT

National Center for Genome Research (USA) - ferramentas de anotação

- § http://www.ncgr.org

European Bioinformatics Institute - ferramentas e bancos de dados

- § http://www.ebi.ac.uk

The Biocomputing Service Group - várias ferramentas de análise genômica e anotação

- § http://genome.dkfz-heidelberg.de

TIGR - ferramentas para anotação gênica e montagem final e visualização de genomas

- § http://www.tigr.org/software

GenScan - programa para predição de ORFs em um segmento genômico

- § http://genes.mit.edu/GENSCAN.html

ESTScan - programa para identificação de fase de leitura através do *codon usage*

- http://www.ch.embnet.org/software/ESTScan.html

Laboratório de Bioinformática da Unicamp

- http://www.lbi.ic.unicamp.br

Núcleo de Bioinformática da UFMG - ferramentas simples de análise

- http://www.icb.ufmg.br/~infobio

Referências Bibliográficas

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, et al. (2000) The genome seqüence of Drosophila melanogaster. Science. 287: 2185-2195

Dias Neto E, Garcia Correa R, Verjovski-Almeida S, Briones MR, Nagai MA, et al. (2000) Shotgun seqüencing of the human transcriptome with ORF expressed seqüence tags. Proc Natl Acad Sci U S A. 97: 3491-3496

Ebersberger I, Metzler D, Schwarz C e Paabo S. (2002) Genomewide comparison of DNA seqüences between humans and chimpanzees. Am J Hum Genet. 70: 1490-1497

Jimenez-Sanchez G, Childs B e Valle D. (2001) Human disease genes. Nature. 409:853-855

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. (2001) Initial seqüencing and analysis of the human genome. Nature. 409: 860-921

Meyerson M, Counter CM, Eaton EN, Ellisen LW, Steiner P, Caddle SD, Ziaugra L,Beijersbergen RL, et al. (1997) hEST2, the putative human telomerase catalytic subunit gene, is up-regulated intumor cells and during immortalization. Cell. 90: 785-795

Perna NT, Plunkett G 3rd, Burland V, Mau B, Glasner JD, Rose DJ, Mayhew GF, et al. (2001) Genome seqüence of enterohaemorrhagic Escherichia coli O157:H7. Nature. 409: 529-533

Borém A e Santos FR (2001) Biotecnologia Simplificada. Editora Suprema. Viçosa, MG.

Simpson AJ, Reinach FC, Arruda P, Abreu FA, Acencio M, Alvarenga R, Alves LM, et al. (2000) The genome seqüence of the plant pathogen Xylella fastidiosa. Nature 406: 151-157

Souza SJ, Camargo AA, Briones MR, Costa FF, Nagai MA, Verjovski-Almeida S, et al. (2000) Identification of human chromosome 22 transcribed seqüences with ORF expressed seqüence tags. Proc Natl Acad Sci U S A. 97: 12690-12693.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, et al. (2001) The seqüence of the human genome. Science. 291: 1304-1351

Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. (1995). Serial Analysis Of Gene Expression. Science. 270: 484-487.