Capítulo 11

Bioinformática

J. Miguel Ortega[1]
Fabrício R. Santos[2]

Uma grande revolução na geração de dados biológicos se deu desde o início do Projeto Genoma Humano, nos anos 1990. Até então, os dados podiam ser armazenados em qualquer unidade de disco de um computador. Nestas duas décadas houve uma crescente demanda de espaço virtual para o processamento destes dados. Assim, para acompanhar esse aumento exponencial no volume de dados moleculares, houve também a necessidade do aumento da capacidade computacional quanto a armazenamento, processamento e análise (Prosdocimi e Santos, 2004). Neste contexto, uma série de limitações foi imposta ao desenvolvimento dessa nova "ciência", por exemplo, o desenvolvimento em ritmo mais lento de plataformas computacionais apropriadas à execução da análise bioinformática dos dados moleculares. Essas limitações compreendem tanto componentes de *hardware* quanto *software*. Todavia, inúmeros computadores e ferramentas de análises foram desenvolvidos para lidar com esta quantidade massiva de dados advindos da Genômica, Proteômica, Metagenômica, Metabolômica, etc.

Neste capítulo, descrevemos o histórico e o estado atual da bioinformática aplicada ao processamento de grandes quantidades de dados gerados pelas novas ômicas.

[1] Biólogo, D.Sc. e Professor da Universidade Federal de Minas Gerais E-mail: miguel@icb.ufmg.br
[2] Biólogo, M.Sc., D.Sc. e Professor da Universidade Federal de Minas Gerais. E-mail: fsantos@icb.ufmg.br

Os Megadados das Ômicas e a Bioinformática

Dados advindos do conhecimento biológico são relativamente complexos em comparação aos provenientes de outras áreas científicas, dada a sua diversidade e ao seu inter-relacionamento, como demonstrado pelos resultados gerados pelos projetos em genômica (Figura 11.1). De acordo com o conhecimento fundamental do genoma montado a partir de sequências de DNA, objetiva-se compreender o funcionamento complexo de todo o organismo, por exemplo que genes estão relacionados com a resposta a medicamentos, uma das metas da farmacogenômica. Porém, no momento, isso somente é possível por partes, devido à grande complexidade dos dados e limitações teóricas e de bioinformática. Primeiro, busca-se entender as estruturas moleculares das proteínas e de outros produtos gênicos, como os RNAs funcionais, as interações entre várias destas moléculas sintetizadas a partir do genoma, bem como destas com as demais moléculas biológicas funcionais e estruturais (DNA, carboidratos, lipídios, etc.), as diversas vias metabólicas celulares e o papel da variabilidade genética representada pelas várias formas de cada produto gênico. Toda essa informação disponibilizada pela ciência genômica só é possível de ser organizada, analisada e interpretada com o apoio da bioinformática.

Atualmente, a bioinformática é imprescindível para a manipulação de qualquer tipo de dado biológico, principalmente os "megadados" oriundos das ômicas. A bioinformática pode ser definida como uma modalidade que abrange todos os aspectos de aquisição, processamento, armazenamento, distribuição, análise e interpretação da informação biológica. Por meio da combinação de procedimentos e técnicas de matemática, estatística e ciência da computação, são elaboradas várias ferramentas que nos auxiliam a compreender o significado biológico representado nos dados biológicos das ômicas. Além disso, mediante a criação de bancos de dados com as informações já processadas, acelera-se a investigação em outras áreas biológicas, como a medicina, a biotecnologia, a agronomia, etc.

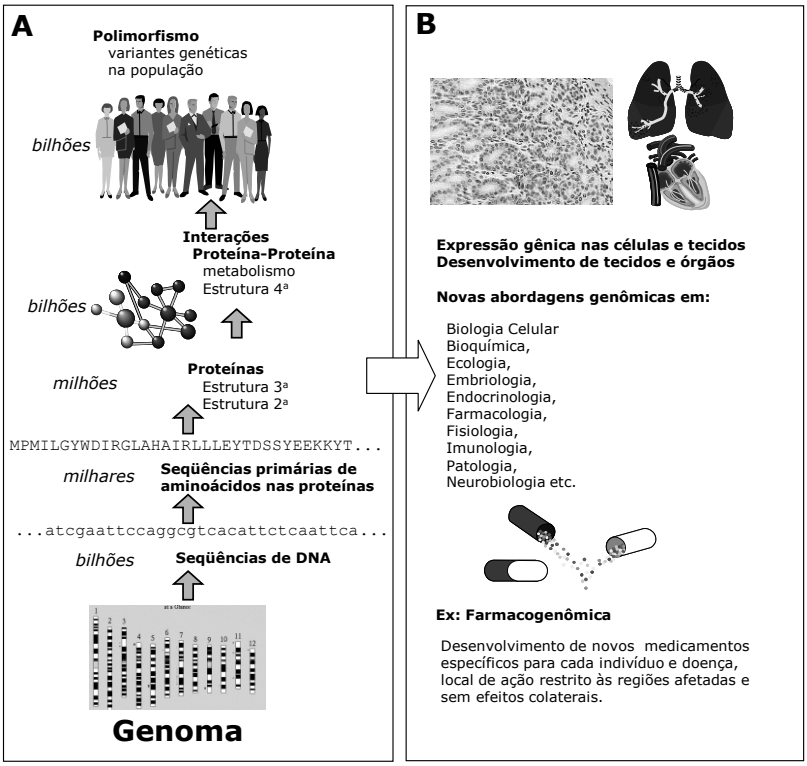

Figura 11.1. Acúmulo de dados biológicos (A) e aplicações do conhecimento genômico (B).

Hardware para a bioinformática moderna

As diversas análises bioinformáticas têm características próprias, mas uma tendência geral identificada nos últimos anos: a troca do trabalho com *software* instalado no próprio computador para a utilização de um servidor compartilhado pelos grupos de pesquisa. Mais recentemente, a execução das pesquisas é em servidores institucionais de grande porte. Com isso, a adaptação dos pesquisadores ao ambiente Linux se tornou obrigatória, sendo comum o treinamento de estudantes de graduação da área de ciências biológicas para acesso a servidores remotos. A conexão é, quase sempre, feita por meio de um programa que executa o protocolo de

conexão *SSH* (*Secure Shell*) e é curioso que o acionamento de alguns programas, o acompanhamento de sua execução ou mesmo a visualização de alguns resultados podem ser feitos com aplicativos instalados em *tablet*s ou mesmo *smartphone*s!

No Brasil, o Sistema Nacional de Computação de Alto Desempenho (*SINAPAD*) constitui-se em um importante recurso cada vez mais utilizado pelos grupos de pesquisa em bioinformática, e é bastante simples obter uma conta para desenvolvimento de projetos nos servidores dos centros que compõem o *SINAPAD*, os quais têm a denominação de *CENAPAD* (Centro Nacional de Computação de Alto Desempenho). Comumente, cada usuário executa programas que ocupam uma fração significativa da máquina, digamos cerca de 50 a 100 núcleos computacionais, o que não seria viável em um servidor local, conseguindo seu resultado em cerca de um dia a uma semana, dependendo da análise e quantidade de dados. Se mais núcleos estão disponíveis na máquina, com parcimônia é possível utilizá-los para acelerar o processamento das rotinas requeridas.

Recentemente, instituições passaram a adquirir servidores que têm como característica uma quantidade alta de memória endereçada por cada núcleo computacional, as máquinas de memória compartilhada, chegando frequentemente a 2 TB (*terabytes*). Isso facilitaria a montagem de genomas grandes e complexos como os do pau-brasil e do peixe-boi, ainda não sequenciados (estimados em 3,8 e 4,6 bilhões de bases, respectivamente), pois a montagem requer que muitas sequências pequenas, geradas em grande quantidade pelos sequenciadores de nova geração, sejam associadas umas às outras, e isso requer o endereçamento de grande quantidade de memória. Assim, o processo pode ser feito em um único passo, pois as várias possibilidades de montagem podem ser testadas simultaneamente. Máquinas com memória compartilhada de cerca de 2 TB estão se tornando acessíveis em várias instituições, incluindo o *SINAPAD*. A capacidade de estoque em discos rígidos típica nesses servidores gira em torno de 100 TB. Assim, os bioinformatas de hoje não trabalham mais nos seus próprios computadores, mas enviam os dados para serem processados remotamente em *hardware* com o formato de *cluster* computacional. Invariavelmente, esses servidores operam sistemas operacionais Linux de várias distribuições, como *RedHat Enterprise*, *CentOS* ou *OpenSUSE*. Além disso, alguns contam

também com sistemas gerenciadores de fila para distribuir as tarefas disparadas pelos usuários para muitos núcleos computacionais, como o *OpenPBS* e o *SLURM*. Muitos programas de análise atuais possuem alguma versão que permite o paralelismo, que é quando vários núcleos computacionais são acionados ao mesmo tempo para rodar rotinas, cujos resultados são reunidos por outros núcleos computacionais. Nesses casos, podem utilizar um gerenciador de paralelismo, como o *OpenMPI*, para tornar o trabalho mais simples. As requisições de tarefas nos servidores são geralmente feitas pelo protocolo *SSH*, mencionado acima, mas há uma grande tendência à migração para o que é chamado *WebService*. Esse é um novo protocolo que permite utilizar um tipo de "computação alugada na nuvem". Essa é uma boa opção para pesquisadores que não possuem servidores de alta performance e não utilizam computação de alto desempenho constantemente. Atualmente, é comum bioinformatas administrarem servidores web, de pequeno a médio porte, apresentando os resultados por meio de páginas dinâmicas, as quais acessam bancos de dados para apresentarem compilações feitas na hora, sobre resultados referentes às consultas feitas por outros pesquisadores interessados nos dados.

Software para sequenciamento genômico

Hoje a bioinformática é uma ciência que lida com a exploração da informação existente nos seres vivos, nos sistemas biológicos. Frequentemente, o primeiro passo para explorar essa informação é o tratamento de dados advindos de projetos genômicos.

Os primeiros sequenciadores automáticos forneciam leituras de dados de cerca de centenas de bases de tamanho (tipicamente 500-1000). Um detalhe despercebido nesses primórdios da genômica é que os sequenciadores não geravam sequências de DNA, mas uma sequência de picos de fluorescência, interpretados geralmente pelo programa *Phred* (http://www.phrap.org/phredphrap). O peso molecular do fluoróforo presente nas bases que estavam ligadas às didesoxirriboses, que interrompem a polimerização no método de Sanger, é diferenciado para cada base. Assim, um *software* precisava inicialmente averiguar e editar a posição dos picos de fluorescência. Mais importante que isso, o programa *Phred* calculava, pela análise do

formato do pico, a probabilidade de determinação correta da base correspondente, expressando a acurácia do sequenciamento.

Desde essa época, sequências de DNA são armazenadas conjuntamente com suas probabilidades de erro. Copiando o que fora usado para expressar a concentração de prótons como pH, determina-se "– log chance de erro", assim uma chance de erro de 1 em dez mil, ou seja, 10^{-4}, torna-se 4. Mas para não ter que salvar em disco rígido um possível ponto decimal, multiplica-se por dez, assim o valor 40 refere-se a uma chance pequena de erro na determinação da base, de 1/10.000, ou seja, 0,01%. Este valor de *Phred* é conhecido como valor de "qualidade" da determinação da base. Todos os sequenciadores automáticos de capilares e baseados no método de Sanger fornecem um valor de qualidade equivalente ao *Phred* e, às vezes, com alguma pequena diferença no cálculo, os sequenciadores de novíssima geração também expressam a chance de erro na determinação da base.

A chance de erro de 0,01% (ou seja, 99,99% de certeza) fora estipulada como valor de precisão mínimo no sequenciamento dos genomas dos primeiros organismos modelo, como a levedura, a mosca da fruta e o homem. Na bioinformática, dizia-se que o DNA deveria ser sequenciado até que todas as bases tivessem um valor de qualidade de *Phred* igual ou superior a 40. Portanto, entende-se que a bioinformática participa desde o primeiro passo do progresso do conhecimento dos genomas. Atualmente, não é incomum bioinformatas decidirem trabalhar com chances de erro maiores que 0,01% em certos projetos, pois invariavelmente as leituras ficam mais longas, já que a chance de erro aumenta na extremidade final, quando a precisão de qualquer método vai-se perdendo, em quase todas as técnicas. Mais recentemente, para poupar espaço em disco, cada valor de qualidade foi codificado por um caractere, e costuma-se acomodar sequência e qualidade em um único arquivo. Assim, os valores de qualidade 10 e 20 passaram a ser salvos como + e 5, respectivamente (quadro 1 – comparando os formatos *Phred* e novo formato *FASTQ*).

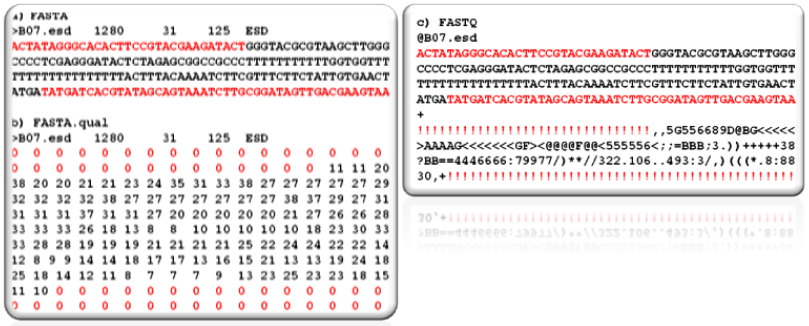

Figura 11.2. Dado de sequência de DNA no formato *FASTA* (a) e arquivo de qualidade FASTA.qual (b), ambos gerados pelo *software Phred*, em comparação ao formato *FASTQ* (c). Os arquivos *FASTQ* incorporam a sequência identificada pela "@", separada pelo símbolo "+" dos dados de qualidade codificados por diferentes caracteres, por exemplo, os caracteres "!", "+" e "5" correspondem a valores *Phred* 0, 10 e 20, respectivamente. As regiões de baixa qualidade (bases em vermelho) tiveram os valores de qualidade zerados e serão retiradas da sequência final.

Software para montagem de *Contigs*

O processo de montagem de grandes sequências de DNA parte inicialmente da busca de regiões similares que permitem gerar agrupamentos de sequências ligadas por estas regiões superpostas, que chamamos de *Contigs*. Esses, por sua vez, podem ser reconectados em *Contigs* cada vez maiores, até formar um cromossomo ou um genoma. À primeira vista, pode parecer que a junção de sequências exportadas pelos sequenciadores é necessária somente quando se trata da determinação da sequência completa de genomas. Na verdade, evidentemente, as cópias que são feitas do RNA com a enzima transcriptase reversa (cDNA) também são sequenciadas parcialmente e precisam sofrer uma montagem para gerar a sequência contínua do referido RNA. Assim, a montagem de sequências é tão útil em transcriptômica quanto em genômica.

Na transcriptômica, a montagem também produz a contagem de quantos transcritos são encontrados para cada gene, ou seja, expressa a abundância dos transcritos. Em casos nos quais o genoma do organismo já está determinado, esse trabalho é facilitado, pois, em vez de se trabalhar com a montagem a partir do zero, podem-se ancorar os transcritos ao genoma publicado. E quando não se possui o genoma do organismo, pode-se utilizar um genoma de referência, que é o termo utilizado para definir um genoma evolutivamente próximo, muitas vezes do mesmo gênero do organismo de interesse, ou até da mesma espécie. O truque de ancorar sequências pequenas em genomas de referência é comumente utilizado também em montagens de genomas novos, muito similares a algum já disponível. Esse assunto nos remete à importância estratégica de vários genomas estarem disponíveis, muito embora alguns grupos taxonômicos tenham sido negligenciados, como a ordem da barata (Blattodea), por exemplo, para a qual inexiste até o presente qualquer genoma completo.

O primeiro *software* famoso para montagem de *Contigs* foi o *Phrap* (http://www.phrap.org/phredphrap), distribuído justamente com o analisador da qualidade das sequências, o *Phred*. Essa versão já continha um *script* que automatizava a análise e era chamado *PhredPhrap*, que gerava os *Contigs* que podiam ser visualizados com o programa *Consed*, igualmente livre, distribuído à parte. Um concorrente do *Phrap* também muito usado em outros projetos era o *Cap3*. Ambos lidavam com o problema de determinar se as leituras que precisavam combinar em uma sequência consenso ou *Contig* eram de uma fita do DNA ou da outra, ou no caso de transcriptomas, se da fita "senso" ou da "anti-senso". Esses programas também ordenam as sequências, superpondo-as através das regiões idênticas e combinando duas ou mais sequências independentes em apenas uma, formando uma sequência consenso. Note que a base da montagem é o alinhamento das sequências individuais, uma técnica muito difundida em bioinformática, que foi utilizada para o desenvolvimento de vários programas para montagem de transcriptomas e genomas.

Montagem Utilizando a Teoria dos Grafos

Só mais recentemente os equipamentos sequenciadores de última geração começaram a produzir sequências individuais da ordem de centenas de bases. Até pouco tempo, as leituras eram apenas de dezenas de nucleotídeos, o que levou ao desenvolvimento de *software* para montagem de *Contigs* com uma abordagem diferente, pois era impossível lidar com alinhamentos de sobreposição tão reduzida. Todavia, em compensação, esses sequenciadores podem gerar vários milhões de sequências em uma única corrida. Assim, além da pequena superposição entre elas, não era mais possível comparar todas as sequências contra todas por técnicas de alinhamento par a par para tentar combiná-las nos *Contigs*. Felizmente, com frequência a computação dispõe de soluções que podem ser aplicadas a problemas novos. Neste caso, trata-se da Teoria dos Grafos. Essa metodologia lida com encadeamentos de elementos formando redes e já abordava problemas complexos, como determinar a melhor rota para difundir sinal telefônico por meio de subsequentes antenas, sendo também usada na internet para encontrar uma máquina pelo endereço IP, usando o caminho mais parcimonioso.

Um *software* muito utilizado para a montagem de genomas com leituras de sequenciadores modernos é o *Velvet* (http://www.ebi.ac.uk/~zerbino/velvet), baseado na Teoria dos Grafos. Uma janela de poucas bases percorre cada uma das milhões de pequenas sequências e, ao verificar que pode conectá-la com outras pequenas sequências, vai encadeando-as em uma imensa rede. Ao final, basta determinar o caminho na rede (*Grafo de Bruijn*) que retorna o *Contig* ou genoma completo. Enquanto pode-se imaginar que DNA repetitivo causaria uma bifurcação do encadeamento, pois uma leitura apresenta várias supostas continuidades em regiões diferentes, felizmente a Teoria dos Grafos já lidava com isso e apresenta técnicas para determinar o caminho apropriado, solucionando globalmente o problema. Por exemplo, quando ocorre uma bifurcação na leitura, o caminho correto a seguir é o ramo da bifurcação que não termina abruptamente. Logicamente, a utilização da Teoria dos Grafos para lidar com milhões de leituras pequenas simultaneamente requer o uso de bastante memória, tipicamente uma

centena de GB para análises de transcriptomas e genomas de bactérias e, pelo menos, 2 TB para genomas de animais e plantas.

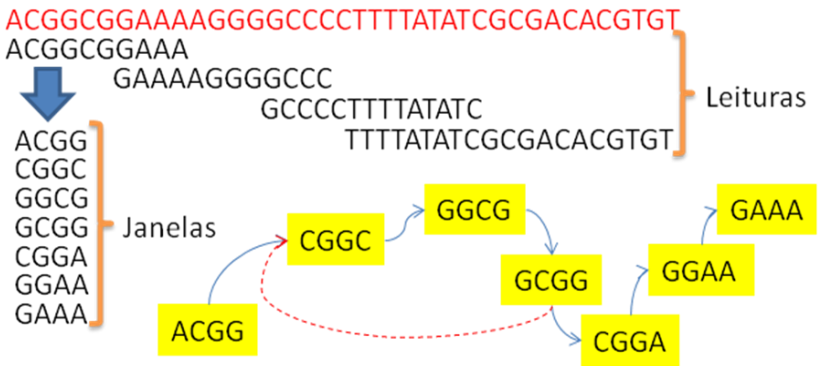

Figura 11.3. Montagem com Teoria dos Grafos. As leituras produzidas pelo sequenciamento são divididas em janelas e uma rede é formada, muito mais complexa que a representada acima. Depois se busca deduzir um caminho que passe por todos os nós da rede apenas uma vez (o caminho tracejado é eliminado). O procedimento básico para resolver este problema já existia antes de ser utilizado na montagem de genomas.

A bioinformática das novas abordagens de *RNAseq e de* sequenciamento

As tecnologias atuais de sequenciamento fazem uso da mesma estratégia utilizada no sequenciamento final do genoma humano: a determinação da sequência das duas extremidades de fragmentos de DNA de tamanho conhecido. Uma das primeiras iniciativas de sequenciamento do genoma, realizada pela empresa Celera, utilizou leituras de cerca de 500 bases provenientes de sequências de três tamanhos conhecidos, 2 kb, 10 kb e 50 kb, e a montagem foi realizada em um computador que conseguia, na época, endereçar "impressionantes" 4 GB de memória. Atualmente, além desta abordagem conhecida como "extremidades em pares" (*paired ends*), com a utilização de novas estratégias metodológicas, um grupo

químico é adicionado à extremidade de uma molécula de, digamos, 10 kb; a molécula é circularizada e fragmentada, e bioquimicamente pesca-se o grupo químico previamente adicionado. Agora, é possível sequenciar esse fragmento, o qual contém a informação de ambas as extremidades da molécula longa inicial. Essa metodologia ficou conhecida como "pares acoplados" (*mated pair*). O *software* de montagem se beneficia da informação de que, em cerca de 10 kb, deve ser encontrada na rede uma sequência *A* associada à sequência *B*. O processo, como dito acima, é muito facilitado quando as pequenas leituras geradas pelos novos sequenciadores podem ser ancoradas a um genoma de referência.

A utilização de sequenciamento de novíssima geração em estudos de transcriptomas, técnica apelidada de *RNAseq*, vem substituindo a utilização da metodologia de microarranjo, devido à facilidade de execução. Além disso, havia uma enorme pressão sobre quem coletava os dados, já que poucos pares de pontos experimentais (controle e tratado) podiam ser processados. Atualmente, um sequenciador pode produzir 150 milhões de sequências por cada uma de suas oito posições, o que permite conduzir análises em triplicata para 10 diferentes condições experimentais, com 40 milhões de sequências por biblioteca de cDNA, o que consiste em uma cobertura suficientemente alta para amostrar significativamente genes pouco expressos. Sequências de 75 bases são suficientes para determinar com precisão sua ancoragem em um genoma já sequenciado, como o genoma humano. O processamento pode ser feito com *software* livre, como *TopHat* (http://tophat.cbcb.umd.edu) e *Cufflinks* (http://cufflinks.cbcb.umd.edu), ferramentas para análise de *RNAseq*, que permitem identificar novos genes e variantes de *splicing*, bem como expressão diferencial (Trapnell et al., 2012). Todavia, assim como na era de domínio do microarranjo, a determinação de genes diferencialmente expressos continua um desafio grande para os diversos tipos de *software* disponíveis, principalmente quando a expressão é baixa. Isso invariavelmente remete o pesquisador a confirmar, por meio de análises subsequentes, se a diferença é real ou se se trata de um falso-positivo (Soneson et al., 2013).

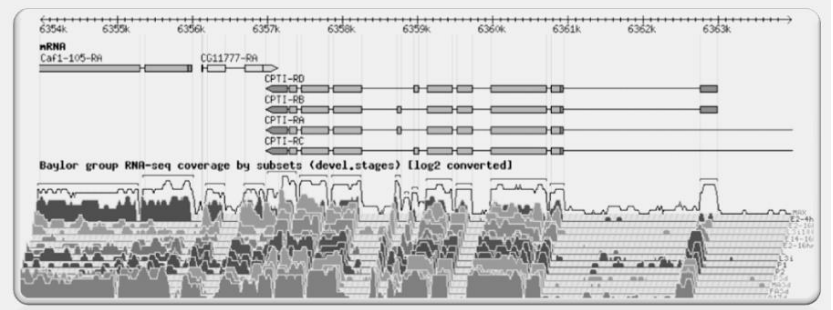

Figura 11.4. Perfis de expressão baseados em RNAseq. Extraída do banco de dados *FlyBase* (flybase.org), com a representação do genoma (escala superior), transcritos (éxons representados por caixas e íntrons por linhas) e cobertura das diferentes regiões por sequências ancoradas, geradas em sequenciador de novíssima geração.

Bancos de dados, identificação de sequências homólogas e anotação funcional

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 maior número possível de cientistas. A construção de bancos de dados para armazenamento de informações de sequências de DNA e genomas inteiros, proteínas e suas estruturas tridimensionais, redes de interações de proteínas, metabolômica, bem como vários outros resultados complexos das diferentes ômicas, tem sido um grande desafio, mas simultânea e extremamente importante.

Um dos primeiros bancos de dados biológicos está no NCBI, ou Centro Nacional para Informação Biotecnológica dos EUA, que é considerado o banco de dados central sobre informações genômicas. Vários outros bancos de dados similares estão distribuídos por países da Europa e no Japão, mas todos trocam dados em um intervalo de 24 horas com o NCBI (http://www.ncbi.nih.gov). O *GenBank*

(http://www.ncbi.nih.gov/genbank) é o principal banco de dados do NCBI e armazena todas sequências disponíveis publicamente de DNA (de sequências pequenas a genomas inteiros), RNA e proteínas. Além do *GenBank*, que coleta todas as entradas de sequências, outros bancos do NCBI apresentam as informações organizadas de diferentes maneiras. Por exemplo, o *UniGene* (http://www.ncbi.nih.gov/unigene) agrupa todas as sequências parciais do transcriptoma de um organismo em aglomerados ou *clusters*, onde cada aglomerado representa a sequência consenso de um gene, ao passo que no *GEO* (http://www.ncbi.nih.gov/geo) é possível analisar a expressão de um dado gene em todos os dados de microarranjo públicos. Também no NCBI, o banco de dados *Gene* (http://www.ncbi.nih.gov/gene) reúne somente as sequências de referência, ou seja, a mais representativa sequência de um transcrito, editada e inspecionada por um curador e ancorada ao genoma. É frequentemente o melhor banco de dados para se evitar a redundância natural num universo com tantas informações. Outros bancos são específicos de um organismo, tal como o *OMIM* (*Online Mendelian Inheritance in Man*, http://www.ncbi.nih.gov/omim), que foi criado para catalogar todos os genes e alelos relacionados a doenças e outras características humanas, bem como proporcionar um detalhamento técnico e bibliografia referente a cada característica. A existência desses bancos de dados, ditos secundários, tem sido tão importante quanto preservar os dados originais no *GenBank*.

Já há algum tempo bancos de dados que congregam rotas metabólicas estão disponíveis, como o *Kegg* (http://www.genome.jp/kegg), um banco disponibilizado pela GenomeNet do Japão (http://www.genomenet.jp). Essa base acopla rotas metabólicas à informação sobre quais organismos em que estas ocorrem, ou não. As consequências da ausência de vias, como a biossíntese dos aminoácidos essenciais (Guedes et al., 2011), são possíveis de serem investigadas somente com bancos de dados, agora disponíveis, de genomas completos.

Várias ferramentas desenvolvidas pela bioinformática permitem o acesso e a análise dos bancos de dados. A ferramenta mais popular de comparação de sequências de DNA com os bancos de

dados de sequências é o *BLAST* (http://www.ncbi.nih.gov/blast) ou *Basic Local Alignment Search Tool* (Altschul et al., 1990). Por meio deste algoritmo podemos comparar uma sequência de DNA ou proteína em busca (*Query*) qualquer com todas as sequências 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 sequência suficientemente parecida com aquela pesquisada. Descarta, assim, rapidamente, os resultados não produtivos e estende a vizinhança da região de similaridade detectada até onde for possível. O resultado dessa busca retorna, dentre as sequências depositadas (DNA, RNA ou proteínas), aquelas com maior pontuação nas medidas de similaridade local. Dessa forma, várias regiões de DNA podem ser anotadas por meio do *BLAST*, cujo resultado pode servir para sugerir ou atribuir uma função a qualquer segmento de DNA, devido ao fato de a similaridade observada ser muito alta em comparação com o que, se esperaria por acaso. É interessante observar que se utilizássemos um dinucleotídeo, "AT", por exemplo, para pesquisar sequências do *Genbank*, o número esperado de alvos seria altíssimo, pois se espera encontrar aleatoriamente vários desses dinucleotídeos em inúmeras sequências depositadas. Se a nossa sequência pesquisada fosse mais complexa, por exemplo, 144 bases, a chance de encontrarmos ao acaso outra sequência idêntica de 140 bases seria infinitamente pequena. O valor de "E" (*E-value*), 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. Nas buscas que retornam sequências ligeiramente diferentes, mas com *E-value* muito pequeno, sugere-se a hipótese alternativa de que as sequências tiveram origem comum e depois sofreram mutações ao longo da evolução, que podem ter ou não importância funcional.

Há várias modalidades de *BLAST* (Figura 11.5). A mais curiosa e de grande importância na descoberta gênica é aquela onde tanto a *Query* como a base de dados (*Subject*) são sequências de nucleotídeos, mas as comparações são feitas entre os aminoácidos codificados por estas sequências. Neste programa, antes de verificar a

similaridade, são feitas as seis traduções possíveis de cada sequência de nucleotídeos, ou seja, tanto a sequência pesquisada quanto cada uma das presentes na base de dados são transformadas em seis proteínas (iniciando a tradução pela base 1, 2 ou 3 de cada fita, a fita "+" e a 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 geralmente mais parecidas entre si que as sequências de nucleotídeos 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 por similaridade da subunidade catalítica da Telomerase humana (Figura 11.5), assim que tal enzima do protozoário *Euplotes* foi clonada (Meyerson et al., 1997). Outras modalidades buscam homologia entre sequências de nucleotídeos (*BLASTn*), sequências de proteínas (*BLASTp*) ou entre sequências de nucleotídeos *versus* proteínas (*BLASTx*). Outra variedade de *BLAST* é o *PSI-BLAST*, que em uma primeira busca encontra as proteínas mais similares à pesquisada - *Query*; prossegue identificando as regiões conservadas dentre os melhores resultados da pesquisa e, em buscas subsequentes, mascara as regiões não conservadas da *Query* e executa a pesquisa levando em conta apenas as regiões conservadas.

Nos bancos de dados, há também 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 possam 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 sequência primária (*VAST*-NCBI), etc.

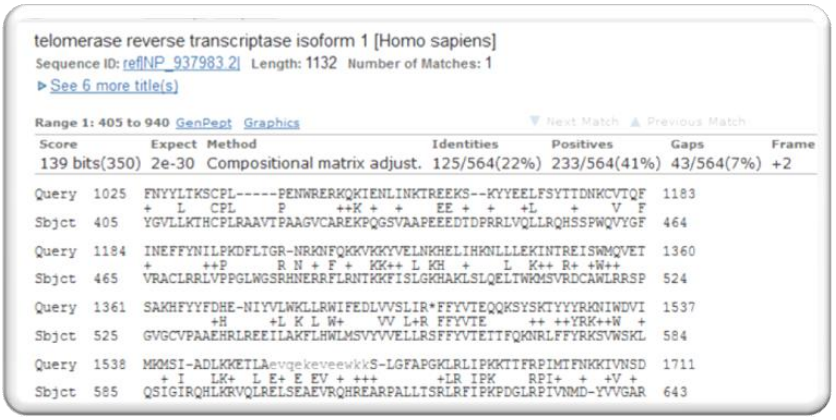

Figura 11.5. Resultado de busca por similaridade com o programa *BLAST*. A primeira clonagem da subunidade catalítica de uma Telomerase foi do protozoário *Euplotes*. Curiosamente, a busca de similaridade nucleotídica (*BLASTn*) entre o gene desse ciliado e genes humanos não encontra nenhum alvo suficientemente similar. No entanto, a modalidade *BLASTx* realiza a tradução da sequência nucleotídica de *Euplotes* em busca (*Query*) em seis sequências de aminoácidos possíveis (três a partir da fita "+" e três da fita "-"). A Figura mostra que a segunda fase de tradução possível (*Frame* +2) alinha com Telomerase humana (*Subject* da busca, ou Subjct). O número de alinhamentos esperados ao acaso é baixo, $2e^{-30}$. Deve-se a buscas por alinhamento local desse tipo a descoberta da Telomerase humana, em 1997.

À medida que é feito o sequenciamento 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, a análise diferencial dos genes de bactérias *E. coli* patogênicas e não-patogênicas (Perna et al., 2001) permitiu a identificação daqueles genes relacionados às causas da doença bacteriana (Jimenez-Sanchez et al., 2001). Outros estudos permitem identificar sequências de DNA e elementos funcionais responsáveis por diferenças marcantes entre

espécies, tal como entre homem e chimpanzé (Ebersberger et al., 2002). Foi demonstrado por genômica comparativa que, na história evolutiva dos procariotos, 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, na Internet.

Uma disciplina derivada da genômica, 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 interindividuais em humanos para a localização de genes relacionados à susceptibilidade ou resistência a doenças ou fármacos. Algumas empresas possuem bancos de dados privados contendo essas variações genéticas, na maior parte do tipo SNPs (*Single Nucleotide Polymorphisms*), que correspondem a diferenças em uma única posição nucleotídica. O NCBI possui um banco de dados público de SNPs de diferentes organismos, sendo que na espécie humana são mais de quatro milhões catalogados. Outros grupos de pesquisa e empresas investiram fortemente na identificação de SNPs de organismos modelo, como o camundongo, para aplicações na farmacogenômica. A partir das coleções de SNPs podem-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 personalizados e tratamentos mais precisos e sem efeitos colaterais.

A organização do conhecimento atual em bancos de dados secundários é muito interessante, pois, mesmo antes de um organismo ser completamente sequenciado, muito da análise de funções moleculares e processos biológicos presentes pode ser prontamente obtido por comparação. Uma iniciativa importante foi a criação de termos específicos de Ontologia Gênica, pelo consórcio *Gene Ontology* (http://www.geneontology.org). Os termos GO (pronuncia-se como o verbo "ir" em inglês) têm relações hierárquicas no formato de uma árvore, na qual as folhas especificam as funções ou processos gerais. Paralelamente ao progresso da construção de ontologias, outro

consórcio, GOA (*Gene Ontology Annotation*), atribui os termos a sequências. Surgiram então abordagens como o *BLAST2GO* (Conesa et al., 2005), que classifica populações de sequências, como as de diferentes transcriptomas, segundo as ocorrências de termos GO, e pode-se perceber o enriquecimento de transcritos relacionados a determinadas funções ou processos, em resposta a um desafio.

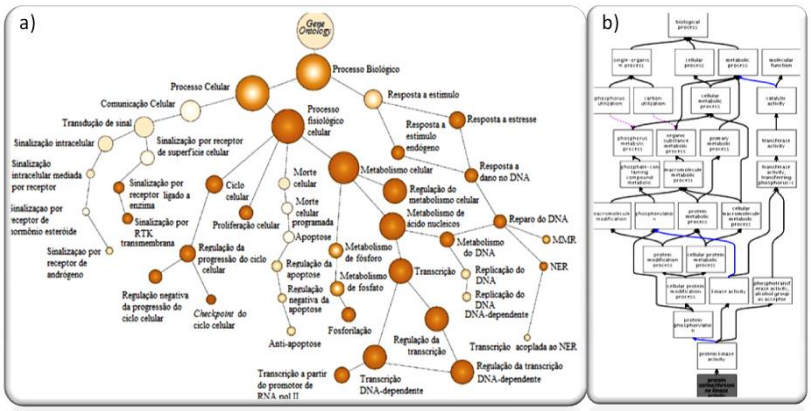

Figura 11.6. Ontologia Gênica. Termos GO para processos envolvidos com câncer (a), em que o tamanho dos círculos representa o número de genes relacionados com o dado processo e as cores mais escuras o suporte estatístico maior. Em (b), a hierarquia de termos GO associada a uma proteína controladora da proliferação celular, *cdk1*.

A anotação de um genoma completo

O domínio do *software BLAST*, explicado acima, permanece na análise dos dados ao longo do progresso da bioinformática. Todavia, algumas proteínas são exclusivas de determinados grupos taxonômicos, ou não foram ainda caracterizadas. Uma maneira bem simples de identificar uma possível região codificadora de proteínas é a ausência inesperada das trincas TAA, TAG e TGA no DNA, pois estas são transcritas em UAA, UAG e UGA, códons de terminação. Sua ausência em trechos significativamente longos é um bom indício da presença de

codificação de proteínas, naquele trecho do DNA. Outros parâmetros, como frequência de dinucleotídeos e características que às vezes são extraídas por inteligência artificial permitem sugerir a presença de proteínas no DNA, que são então chamadas de proteínas "preditas" por *software*. Interessantemente, embora ninguém as tenha estudado em detalhe, essas proteínas podem ser encontradas em vários organismos, o que é outra evidência a favor de sua existência funcional, e nesse ponto passam a ser chamadas de proteínas hipotéticas. Assim, quando um operador utiliza um *software* de anotação de genomas como o *Artemis* (Figura 11.7), ele foca sua atenção em trechos sem códons de parada e, com o auxílio de um *software* de predição gênica (*Glimmer*, por exemplo) e de um *software* de alinhamento com sequências de outros organismos, como o *BLAST*, consegue identificar regiões codificadoras de proteínas, sejam de funções conhecidas ou hipotéticas. A anotação de um genoma com íntrons é um pouco mais complexa, pois nela se adiciona a dificuldade de determinar corretamente o modelo gênico, ou seja, as regiões onde se localizam os éxons e os íntrons, além de que frequentemente existem várias isoformas de processamento do RNA possíveis (Figura 11.4).

Sistema Computacional Gerenciador de Tarefas Encadeadas (*Workflow*)

Recentemente, tornou-se possível a integração de tarefas bioinformáticas por um sistema computacional gerenciador. Esse sistema não somente coordena as tarefas, mas também facilita a utilização do resultado de um *software* por outro, integrando os formatos de dados produzidos. Assim, ele facilita a execução de uma sequência de passos conectados (*workflow*). Geralmente estes sistemas são instalados em *clusters* computacionais. O sistema mais conhecido e utilizado com frequência em centros de bioinformática é o *Galaxy* (Schatz, 2010). Foi inicialmente desenvolvido para genômica, mas já é utilizado para diversas aplicações. A simplicidade de seu uso reside na utilização de ambientes gráficos visualizados em um navegador web, com painéis de controle para submissão de tarefas, escolha da

sequência de *software* desejada e gerenciamento do projeto. Os resultados obtidos com várias análises são todos integrados, fazendo referência uns aos outros. E, o que é mais importante, possibilita que a instalação de um novo *software* integrante do sistema seja muito amigável, integrando-o aos demais automaticamente.

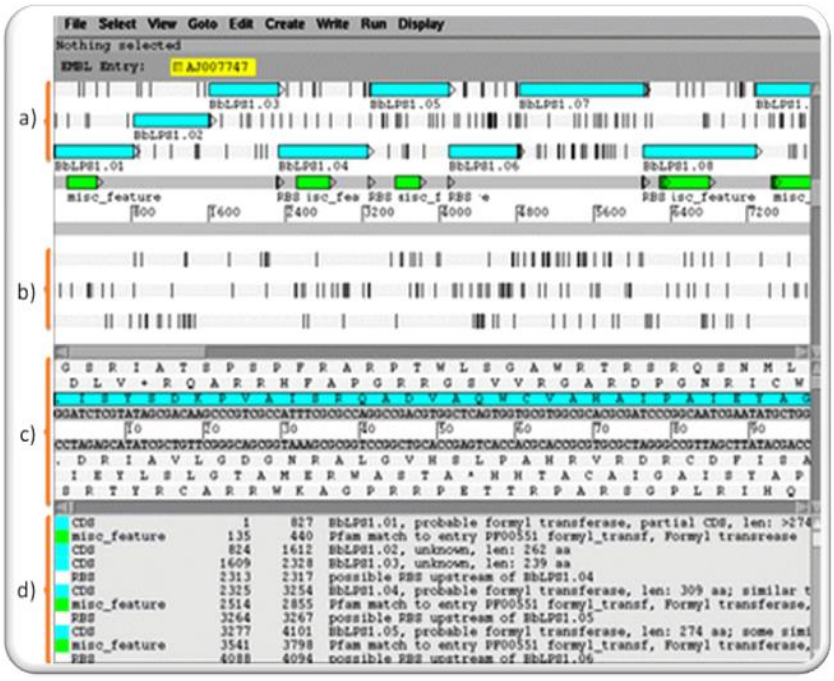

Figura 11.7. Visão de janela do *software Artemis*. Este *software* permite a marcação de regiões em três fases de leitura das duas fitas (a e b) onde não ocorrem códons de parada (riscos verticais) muito próximos, prováveis regiões gênicas (ciano) e a visão da sequência de aminoácidos daquela com a qual se trabalha (c). As anotações baseadas em *BLAST* vão sendo editadas em (d). Figura extraída do website do *software* (http://www.sanger.ac.uk/resources/software/artemis).

Aplicações para estudos em plantas

A Figura 11.8 mostra a grande quantidade de informação dentro do reino Viridiplantae (plantas verdes). São milhões de sequências nucleotídicas e de proteínas, e milhares de estruturas proteicas com estrutura 3D determinada. Milhares de experimentos de microarranjos podem ser analisados por estarem depositados em bases de dados públicas como a *GEO* (Barrett et al., 2013). Entradas editadas, sem redundância, são encontradas na base de dados *Gene*, que já passam de meio milhão. Há informações sobre sequências de quase 150 mil organismos. O conhecimento acumulado até o momento facilita muito a análise de dados de projetos com novos organismos e situações. Isso acelera cada vez mais a análise sistêmica dos dados, na qual a bioinformática tem papel estratégico.

Base de Dados	Entradas
Nucleotídeos	13.369.290
ESTs	24.811.263
Proteínas	2.574.090
Estruturas	2.715
Genomas	826
SNPs	24.665.666
Domínios	1.576
Microarranjos	5.526
Genes em Microarranjo	3.191.196
Genes	551.544
Homólogos	12.488
Seq. Nova Geração	11.282
Grupos Taxonomicos	141.158

Figura 11.8. Entradas em bases de dados referentes a organismos do reino Viridiplantae (plantas verdes). Consulta feita à base de dados Taxonomy do NCBI em abril de 2013 (http://www.ncbi.nih.gov/taxonomy).

Considerações Finais

O século 19 foi palco de uma revolução na biologia com a Teoria da Evolução, que tornou possível investigar e compreender o funcionamento do organismo, que a partir do final do século 20 se materializou nos grandes avanços das abordagens genômicas e suas derivadas, resultando na geração de dados biológicos em larga escala. No século 21, observamos o avanço da bioinformática permitindo a análise e o armazenamento de toda esta informação advinda das ômicas. A bioinformática depende de um esforço da mente humana na elaboração de algoritmos e rotinas que permitem utilizar a capacidade lógica dos computadores para processar e classificar os megadados das ômicas. No entanto, existe atualmente uma grande demanda pela ampliação do tratamento lógico dos dados, acoplada à possibilidade de elaboração, pelo cientista, de hipóteses agora sistêmicas, o que coloca a bioinformática em uma situação central na exploração das Ômicas. Ainda há muito que se desenvolver em termos de *software*, mas cada vez mais estes estão sendo dirigidos para testar e elaborar hipóteses e descobrir o que está por trás dos processos biológicos mais complexos.

Referências

Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. 1990. Basic local alignment search tool. J Mol Biol. 215(3):403-10.

Barrett, T.; Wilhite, S.E.; Ledoux, P.; Evangelista, C.; Kim, I.F.; Tomashevsky, M.; Marshall, K.A.; Phillippy, K.H.; Sherman, P.M.; Holko, M.; Yefanov, A.; Lee, H.; Zhang, N.; Robertson, C.L.; Serova, N.; Davis, S.; Soboleva, A. 2013. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. 41(D1): D991–D995. Published online 2012 November 26. doi: 10.1093/nar/gks1193 PMCID: PMC3531084.

Conesa, A.; Götz, S.; García-Gómez, J.M.; Terol, J.; Talón, M.; Robles, M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 15; 21(18):3674-6. Epub 2005 Aug 4. PubMed PMID: 16081474.

Ebersberger, I.; Metzler, D.; Schwarz, C.; Pääbo, S. 2002. Genome-wide comparison of DNA sequences between humans and chimpanzees. Am J Hum Genet. 70(6):1490-7. Epub 2002 Apr 30. PubMed PMID: 11992255; PubMed Central PMCID: PMC379137.

Guedes RL, Prosdocimi F, Fernandes GR, Moura LK, Ribeiro HA, Ortega JM. 2011 Amino acids biosynthesis and nitrogen assimilation pathways: a great genomic deletion during eukaryotes evolution. BMC Genomics. Dec 22;12 Suppl 4:S2. Epub 2011 Dec 22. PubMed PMID: 22369087; PubMed Central PMCID: PMC3287585

Jimenez-Sanchez, G.; Childs, B.; Valle, D. 2001. Human disease genes. Nature. 15; 409(6822):853-5. PubMed PMID: 11237009.

Perna, N.T.; Mayhew, G.F.; Pósfai, G.; Elliott, S.; Donnenberg, M.S.; Kaper, J.B.; Blattner, FR. 1998. Molecular evolution of a pathogenicity island from enterohemorrhagic *Escherichia coli* O157:H7. Infect Immun. 66(8):3810-7. PubMed PMID: 9673266; PubMed Central PMCID: PMC108423.

Prosdocimi, F.; Santos, F.R. 2004. Sobre bioinformática, genoma e ciência. Ciência Hoje. 35 (209):54-57.

Schatz, M.C. 2010. The missing graphical user interface for genomics. Genome Biol.11(8):128. doi: 10.1186/gb-2010-11-8-128. Epub 2010 Aug 25. PubMed PMID: 20804568; PubMed Central PMCID: PMC2945776.

Soneson, C.; Delorenzi, M. 2013. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics. 9;14:91. doi: 10.1186/1471-2105-14-91. PubMed PMID: 23497356; PubMed Central PMCID: PMC3608160.

Trapnell, C.; Roberts, A.; Goff, L.; Pertea, G.; Kim, D.; Kelley, D.R.; Pimentel, H.; Salzberg, S.L.; Rinn, J.L.; Pachter, L. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. Mar 1;7(3):562-78. doi: 10.1038/nprot.2012.016. PubMed PMID: 22383036; PubMed Central PMCID: PMC3334321.