Bioinformática: Manual do Usuário

Ilustrações cedidas pelos autores

Um guia básico e amplo sobre os diversos aspectos dessa nova ciência

Francisco Prosdocimi
Mestrando em Genética e Especialista em Bioinformática
Universidade Federal de Minas Gerais
franc@icb.ufmg.br

Gustavo Coutinho Cerqueira
Bacharel em Ciência da Computação e Especialista em Bioinformática
Universidade Federal de Minas Gerais
cerca@csr.ufmg.br

Eliseu Binneck
Doutor em Ciência e Tecnologia de Sementes e Especialista em Bioinformática
Embrapa Soja
binneck@cnpso.embrapa.br

Acácia Fernandes Silva
Mestre em Agronomia e Especialista em Bioinformática
Empresa Pernambucana de Pesquisa Agropecuária
acacia@ipa.br

Adriana Neves dos Reis
Bacharel em Informática e Especialista em Bioinformática
Universidade do Vale do Rio dos Sinos
adriana@exatas.unisinos.br

Ana Carolina Martins Junqueira
Mestre em Genética e Biologia Molecular e Especialista em Bioinformática
Universidade de Campinas
anacmj@unicamp.br

Ana Cecília Feio dos Santos
Mestranda em Genética e Biologia Molecular e Especialista em Bioinformática
Universidade Federal do Pará
cecifeio@ufpa.br

Antônio Nhani Júnior
Doutor em Bioquímica e Especialista em Bioinformática
Universidade Estadual Paulista
nhani@fcav.unesp.br

Charles I. Wust
Mestrando em Ciências da Computação e Especialista em Bioinformática
Universidade Federal de Santa Catarina
wust@inf.ufsc.br

Fernando Camargo Filho
Mestrando em Biotecnologia Vegetal e Especialista em Bioinformática
Universidade de Ribeirão preto
camargo@odin.unaerp.br

Jayme Lourenço Kessedjian
Analista de sistemas e Especialista em Bioinformática
Embrapa Agrobiologia
jayme@cnpab.embrapa.br

Jorge H. Petretski
Prof. Associado e Especialista em Bioinformática
Universidade Estadual do Norte Fluminense
jhpetretski@uenf.br

Luiz Paulo Camargo
Analista de Sistemas e Especialista em Bioinformática
Universidade de Ribeirão Preto
luizpcam@uol.com.br

Ricardo de Godoi Mattos Ferreira
Bacharel em Ciências Biológicas e Especialista em Bioinformática
Universidade de São Paulo
ricgmf@lineu.icb.usp.br

Roceli P. Lima
Mestrando em Informática e Especialista em Bioinformática
Universidade do Amazonas
rossi@horizon.com.br

Rodrigo Matheus Pereira
Mestrando em Microbiologia e Especialista em Bioinformática
Universidade Estadual Paulista
rodrigus@fcav.unesp.br

Sílvia Jardim
Mestre em Farmacologia e Especialista em Bioinformática
Embrapa Milho e Sorgo
silviajardim@yahoo.com.br

Vanderson de Souza Sampaio
Mestrando em Genética e Biologia Molecular e Especialista em Bioinformática
Universidade Federal do Pará
vander@ufpa.br

Áurea V. Folgueras-Flatschart
Doutora em Microbiologia e Especialista em Bioinformática
Universidade Federal de Minas Gerais
folguera@bol.com.br

INTRODUÇÃO

Do início até meados do século passado os geneticistas e químicos se questionaram sobre a natureza química do material genético. Das pesquisas desenvolvidas, surgiu a conclusão de que o DNA era a molécula que armazenava a informação genética e, em 1953, sua estrutura química foi desvendada no clássico trabalho de Watson e Crick. Com a posterior descoberta do código genético e do fluxo da informação biológica, dos ácidos nucléicos para as proteínas, tais polímeros passaram a constituir os principais objetos de estudo de uma nova ciência, a Biologia Molecular. Logo surgiram métodos de seqüenciamento desses polímeros, principalmente do DNA, que permitiam a investigação de suas seqüências monoméricas constituintes. Desde então, mais de 18 bilhões dessas seqüências já foram produzidas e estão disponíveis nos bancos de dados públicos.

Figura 1: O Dogma Central da Biologia Molecular

Na segunda metade da década de 90, com o surgimento dos seqüenciadores automáticos de DNA, houve uma explosão na quantidade de seqüências a serem armazenadas, exigindo recursos computacionais cada vez mais eficientes. Além do armazenamento ocorria, paralelamente, a necessidade de análise desses dados, o que tornava indispensável a utilização de plataformas computacionais eficientes para a interpretação dos resultados obtidos.

Assim nascia a bioinformática. Essa nova ciência envolveria a união de diversas linhas de conhecimento – a engenharia de softwares, a matemática, a estatística, a ciência da computação e a biologia molecular. Os primeiros projetos na área eram compostos por profissionais de diferentes

áreas da biologia e informática e percebia-se uma certa dificuldade de comunicação: enquanto o biólogo procurava uma solução que levasse em consideração as incertezas e erros que ocorrem na prática, o cientista da computação procurava uma solução eficiente para um problema bem definido. Assim, surgiu a necessidade de um novo profissional, que entendesse bem ambas as áreas e fizesse a ponte entre elas: o Bioinformata. Esse profissional deveria ter o conhecimento suficiente para saber quais eram os problemas biológicos reais e quais seriam as opções viáveis de desenvolvimento e abordagem computacional dos problemas em questão.

Dado o sucesso e a importância que alcançaram os projetos Genoma e seus desmembramentos, o bioinformata tem sido um profissional requisitado e raro. No exterior, podem ser encontrados pelo menos 122 cursos de formação em bioinformática, em sua grande maioria centrados na América do Norte e Europa (http://linkage. rockefeller. edu/wli/bioinfocourse/). No Brasil, entretanto, até o início deste ano, não existiam cursos que formassem tais profissionais especializados. Políticas científicas governamentais têm procurado incentivar a formação de grupos de pesquisa e de pessoal nessa área, financiando projetos e criando cursos de pós-graduação. Em 2002, foi implantado o primeiro Curso de Especialização (pós-graduação *lato sensu*) do LNCC (http://www.lncc.br/~biologia) - do qual formamos a segunda turma. Ainda neste ano foi autorizada pela CAPES a criação de dois cursos de doutorado em Bioinformática, um na USP e outro na UFMG (http://www.capes.gov.br/).

Parece-nos que cada vez mais a bioinformática vai ser necessária para a análise de dados em biologia molecular e, nesse sentido, o presente artigo foi escrito com o intuito de conter as informações mais relevantes para quem deseja começar a trabalhar na área. Assim, tentamos apresentar os principais conceitos relacionados à biologia e à computação, os softwares mais utilizados, os sites mais freqüentados e as principais áreas de interesse.

Sistemas operacionais

O sistema operacional (SO) é o principal programa de um computador. Ele é responsável pelo gerenciamento da memória, pelo acesso aos discos e também intermedeia todo acesso aos componentes físicos da máquina (*hardware*).

Os SOs mais conhecidos e utilizados são aqueles baseados no Windows, Unix e MacOS. Muitas das aplicações utilizadas em bioinformática são compiladas e distribuídas para a execução em plataformas derivadas do Unix, portanto o conhecimento desse sistema operacional é de grande importância para aqueles que desejam aprofundar-se na área. A preferência por sistemas baseados em Unix deve-se ao fato de que tais sistemas são normalmente mais confiáveis, gerenciam melhor o trabalho com grandes quantidades de dados e que algumas de suas variantes, como o Linux, possuem código aberto e distribuições gratuitas.

Linguagens de programação

Um profissional em bioinformática, além de saber utilizar os programas produzidos por outros programadores, deve também ser capaz de desenvolver programas aplicativos para lidar com os mais diversos problemas encontrados durante a análise de dados em biologia molecular. Para desenvolver, portanto, tais programas, o bioinformata deve ter conhecimento sobre algum tipo de linguagem de programação.

As Linguagens de programação foram criadas para facilitar a especificação de tarefas a um computador. Existem milhares de linguagens de programação e cada uma delas possui um conjunto de comandos específicos que criam esta interface homem-máquina. Das linguagens de programação mais utilizadas, podemos citar: basic, pascal, C, C++, java, cobol e fortran. Entretanto, a linguagem mais utilizada pelos bioinformatas é, sem sombra de dúvida, o PERL.

O PERL (*Practical Extract and Report Language*) é uma linguagem de programação, simples e muito rica, além de disponível gratuitamente. Foi criada por Larry Wall, originalmente para produzir relatórios de informações de erros, que a disponibilizou na Internet no espírito *freeware*, pensando que alguém pudesse achá-la útil. Ao longo dos anos esta linguagem conquistou milhares de adeptos e, através de várias colaborações recebidas para seu aprimoramento, o PERL é hoje conceituado como uma linguagem sofisticada, que possui como ponto forte a manipulação de texto, mas que, além disso, possui todas as características de uma linguagem de alto-nível genérica. É essa grande facilidade para a manipulação de texto que fez do PERL a linguagem mais utilizada no tratamento de dados de seqüências de DNA e proteínas.

O PERL pode ter suas funcionalidades acrescidas através de módulos, que são distribuídos gratuitamente. Existem módulos para uma gama de aplicações, desde métodos estatísticos clássicos, aplicações gráficas em 3D, até acesso a internet via programação PERL. O site CPAN (*Comprehensive Perl Archive Network* – http://www.cpan.org) é o principal ponto de distribuição de módulos e de suas respectivas documentações. Alguns destes módulos são especialmente dirigidos para aplicações em Bioinformática, destacando-se os módulos *bioperl* e *biographics*, que apresentam ferramentas bastante úteis para as mais diversas aplicações nesta área.

Uma boa interconectividade com bancos de dados é outra característica desejada em uma linguagem de programação. A linguagem PERL atende muito bem a esta demanda através da biblioteca PERL-DBI, um conjunto de módulos que fornece uma interface consistente para soluções de integração com bancos de dados.

Bancos de dados

Em conseqüência da grande quantidade de informações de seqüências de nucleotídeos e de aminoácidos que são produzidas atualmente, principalmente em projetos Genoma, Transcriptoma e Proteoma, o uso dos bancos de dados vem as-

sumindo uma importância crescente na bioinformática.

Um banco de dados pode ser considerado uma coleção de dados inter-relacionados, projetado para suprir as necessidades de um grupo específico de aplicações e usuários. Um banco de dados organiza e estrutura as informações de modo a facilitar consultas, atualizações e deleções de dados.

A grande maioria dos bancos de dados é atrelado a um sistema denominado SGBD (Sistema de Gerenciamento de Banco de Dados). Este sistema é responsável por intermediar os processos de construção, manipulação e administração do banco de dados solicitados pelos usuários ou por outras aplicações.

Existem vários sistemas de gerenciamento de banco de dados, sendo que cada sistema possui seus prós e contras. O *mysql* é um sistema muito utilizado pela comunidade acadêmica e em projetos genoma por ser gratuito, possuir código aberto e acesso veloz aos dados, mas apresenta certas limitações em suas ferramentas. O *postgreSQL* também é um SGBD gratuito, com ferramentas muito poderosas, entretanto não é muito utilizado pela dificuldade no seu gerenciamento. Os SGBD's O*racle* e *SQL Server* são robustos e sofisticados, mas devido ao alto custo de suas licenças possuem seu uso limitado às grandes empresas.

Bancos de dados públicos em bioinformática

O investimento contínuo na construção de bancos de dados públicos é um dos grandes motivos do sucesso dos projetos genoma e, em especial, do Projeto genoma Humano. Devido à magnitude do conjunto de dados produzidos torna-se fundamental a organização desses dados em bancos que permitam acesso on-line.

Os bancos de dados envolvendo seqüências de nucleotídeos, de aminoácidos ou estruturas de proteínas podem ser classificados em bancos de seqüências primários e secundários. Os primeiros são formados pela deposição direta de seqüências de nucleotídeos, aminoácidos ou estruturas protéicas, sem qualquer processamento

BOX1 - Exemplo de programa PERL para obter a fita reversa-complementar a partir de uma seqüência de DNA desejada.

```
#!/usr/bin/perl
# Seqüência que se deseja utilizar
$meuDNA = 'TTCCGAGCCAATTGTATCAGTTGCCAATAG';
# Inverte a ordem da seqüência de DNA
$RevCom = reverse $meuDNA;
# Troca as bases produzindo a fita complementar
$RevCom =~ tr/ACGT/TGCA/;
print "Minha seqüência invertida é: \n $RevCom";
```

A primeira linha é obrigatória e diz ao programa o caminho onde se encontra o interpretador PERL para que o programa possa achá-lo na hora de sua execução. As linhas seguintes que se iniciam com o sinal de "#" representam linhas de comentário. As variáveis em PERL são sempre seguidas do sinal de "$" e não precisam ser declaradas, cabe ao programador saber como e em que contexto devem ser utilizadas. Os comandos terminam sempre com ponto-e-vírgula e o sinal de "=~" está relacionado à utilização de uma expressão regular.

BOX2 - Principais Sistemas de Gerenciamento de Bancos de dados

MySQL http://www.mysql.org
Acesso livre para download do gerenciador MySQL, como também a várias ferramentas de conexão como: DBI, Java, ODBC e etc. Apresenta documentação completa.
PostgreSQL http://www.pgsql.com/
Acesso livre para download do gerenciador PostgreSQL, como também algumas ferramentas. Apresenta documentação completa.
ORACLE http://www.oracle.com
Informações comerciais sobre o banco de dados.
Microsoft SQL Server http://www.microsoft.com/sql/
Informações comerciais sobre o banco de dados.

BOX3 - Bancos de Dados mais utilizados em bioinformática

Genbank http://www.ncbi.nlm.nih.gov/
Banco de dados americano de seqüências de DNA e proteínas.
EBI http://www.ebi.ac.uk/
Banco de dados europeu de seqüências de DNA.
DDBJ http://www.ddbj.nig.ac.jp/
Banco de dados japonês de seqüências de DNA.
PDB http://www.rcsb.org/pdb
Armazena estruturas tridimensionais resolvidas de proteínas.
GDB http://gdbwww.gdb.org/
Banco de dados oficial do projeto genoma humano.
TIGR Databases http://www.tigr.org/tdb/
Banco com informações de genomas de vários organismos diferentes.
PIR http://www-nbrf.georgetown.edu/
Banco de proteínas anotadas.
SWISS-PROT http://www.expasy.ch/spro/
Armazena seqüências de proteínas e suas respectivas características moleculares, anotado manualmente por uma equipe de especialistas.
INTERPRO http://www.ebi.ac.uk/interpro/
Banco de dados de famílias, domínios e assinaturas de proteínas.
KEGG http://www.genome.ad.jp/kegg/
Banco com dados de seqüências de genomas de vários organismos diferentes e informações relacionadas às suas vias metabólicas.

ou análise. Os principais bancos de dados primários são o *GenBank*, o EBI (*European Bioinformatics Institute*), o *DDBJ* (*DNA Data Bank of Japan*) e o PDB (*Protein Data Bank*). Os três primeiros bancos são membros do INSDC (*International Nucleotide Sequence Database Colaboration*) e cada um desses centros possibilita a submissão individual de seqüências de DNA. Eles trocam informações entre si diariamente, de modo que todos os três possuem informações atualizadas de todas as seqüências de DNA depositadas em todo o mundo. Apesar disso, cada centro apresenta seus dados de forma particular, apesar de bastante semelhante. Atualmente a maioria das revistas exige que as seqüências identificadas pelos laboratórios sejam submetidas a um destes bancos antes mesmo da publicação do artigo.

Os bancos de dados secundários, como o PIR (*Protein Information Resource*) ou o SWISS-PROT, são aqueles que derivam dos primários, ou seja, foram formados usando as informações depositadas nos bancos primários. Por exemplo, o SWISS-PROT é um banco de dados onde as informações sobre seqüências de proteínas foram anotadas e associadas à informações sobre função, domínios funcionais, proteínas homólogas e outros.

Os bancos de seqüências também podem ser classificados como bancos estruturais ou funcionais. Os bancos estruturais mantêm dados relativos à estrutura de proteínas. Embora a seqüência de nucleotídeos, a seqüência de aminoácidos e a estrutura de proteína sejam formas diferentes de representar o produto de um dado gene, esses aspectos apresentam informações diferentes e são tratados por projetos diferentes, que resultam em bancos específicos.

Dos bancos funcionais, o KEGG (*Kyoto Encyclopedia of Genes and Genomes*) é um dos mais utilizados. Disponibiliza *links* para mapas metabólicos de organismos com genoma completamente ou parcialmente seqüenciados a partir de seqüências e de busca através palavras-chave.

Com o crescente número de dados biológicos que vem sendo gerados, vários bancos de dados têm surgido e anualmente a revista *Nucleic Acids Research* (http://www3.oup.co.uk/nar/database/) publica uma lista atualizada com a classificação de todos os bancos de dados biológicos disponíveis.

Alinhamento de seqüências

O alinhamento de seqüências possui uma diversidade de aplicações na bioinformática, sendo considerada uma das operações mais importantes desta área. Este método de comparação procura determinar o grau de similaridade entre duas ou mais seqüências, ou a similaridade entre fragmentos destas seqüências. No caso de mais de duas seqüências o processo é denominado alinhamento múltiplo.

É bom lembrar que similaridade e homologia são conceitos diferentes. O alinhamento indica o grau de similaridade entre seqüências, já a homologia é uma hipótese de cunho evolutivo, e não possui gradação: duas seqüências são homólogas caso derivem de um ancestral comum ou, caso esta hipótese não se comprove, simplesmente não são homólogas.

Existem vários programas de computador que realizam esta tarefa e a grande maioria deles pode ser utilizado *on-line*, sem a necessidade de instalação. Como exemplo temos os programas: ClustalW, Multialin, FASTA, BLAST 2 sequences, etc.

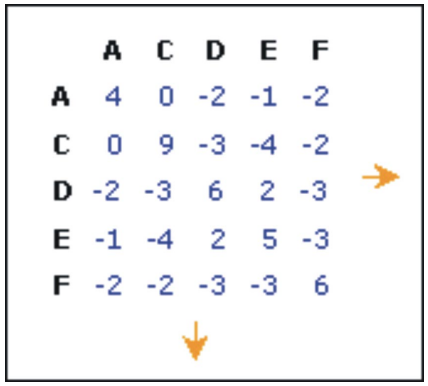

Figura 3. Parte de uma matriz de substituição BLOSUM62, utilizada em alinhamentos de seqüências de proteínas. As letras representam os aminoácidos e os números indicam os pontos a serem contabilizados na ocorrência de match (diagonal principal) ou mismatch

tas matrizes indicam os diferentes valores a serem contabilizados para cada par de unidades.

As matrizes de substituição são normalmente utilizadas no alinhamento de seqüências protéicas. Assim o valor de cada uma de suas células indica a chance da ocorrência da substituição correspondente ao par de aminoácidos deste *mismatch*.

Figura 2 – Alinhamento de duas seqüências de proteínas

O processo consiste em introduzir espaços (*gaps*) entre os monômeros de uma ou mais seqüências a fim de obter o melhor alinhamento possível. A qualidade de um alinhamento é determinada pela soma dos pontos obtidos por cada unidade pareada (*match*) menos as penalidades pela introdução de *gaps* e posições não pareadas (*mismatch*).

Matrizes de substituição

Matrizes de substituição são uma alternativa aos valores fixos de pontuação para *matches* e *mismatches*. Es-

As matrizes de substituição mais utilizadas são aquelas pertencentes às famílias de matrizes PAM (*Point Accepted Mutation*) e BLOSUM. A matriz PAM1 foi construída através da análise de mutações entre proteínas homólogas com 1% de divergência (1% dos aminoácidos diferentes). As outras matrizes, PAM50, PAM100, PAM250 são extrapolações da matriz PAM1. As matrizes BLOSUM foram construídas tendo como base os alinhamentos do banco de motivos BLOCKS. Uma matriz BLOSUM62 é definida através da análise das substituições nas seqüências de BLOCKS que possuem menos

que 62% de similaridade. As seqüências que ultrapassam este limite são mescladas, e participam da definição da matriz como se fossem uma única seqüência.

Alinhamento global e local

Quanto à região analisada, o alinhamento de seqüências pode ser grosseiramente classificado em dois tipos, o alinhamento global e o alinhamento local. No alinhamento global, as seqüências envolvidas devem ser alinhadas de um extremo ao outro, dando origem a apenas um resultado. Já no alinhamento local, procura-se alinhar apenas as regiões mais conservadas, independente da localização relativa de cada região em sua seqüência. Consequentemente, este alinhamento tem como resultado uma ou mais regiões conservadas entre as seqüências.

O alinhamento global é freqüentemente utilizado para determinar regiões mais conservadas de seqüências homólogas. Exemplo de programas que utilizam este alinhamento são ClustalW e Multialin. O alinhamento local é geralmente utilizado na procura por seqüências homólogas ou análogas (funcionalmente semelhantes) em banco de dados. O algoritmo utilizado pelo programa BLAST (*Basic Local Alignment Search Tool*) realiza este tipo de alinhamento.

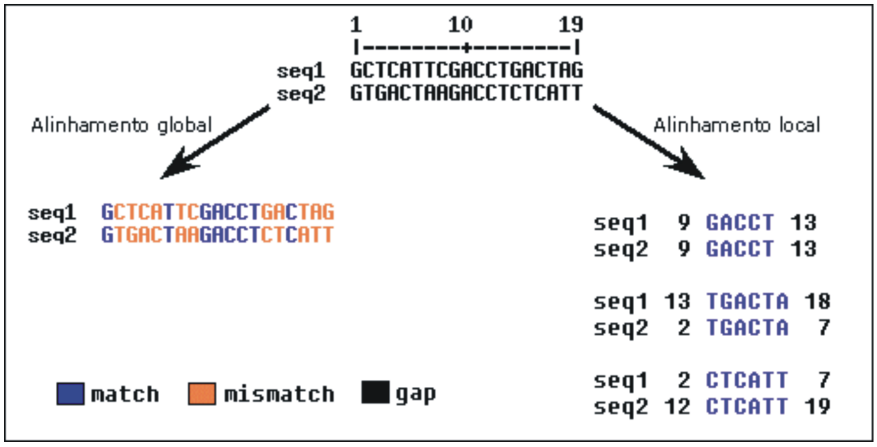

Figura 4: Exemplos de alinhamento global e local. No alinhamento global as seqüências são alinhadas do início ao fim, já no alinhamento local alinha-se as subseqüências conservadas

Projetos genoma e transcriptoma

Grande parte dos bioinformatas modernos trabalha com dados de projetos genoma ou transcriptoma. Em projetos genoma adota-se a abordagem de fragmentar todo o genoma de um organismo em pequenos pedaços e de seqüenciar tais pedaços, utilizando programas computacionais para montá-los e reconstituir a informação genômica inicial. Essa estratégia é adotada principalmente devido à restrição do tamanho da seqüência que pode ser lida nos seqüenciadores. Mesmo os mais modernos conseguem ler apenas cerca de 1000 pares de base em cada corrida.

Em projetos genomas de procariotos, normalmente realiza-se a quebra do DNA inteiro do organismo desejado em fragmentos pequenos (através da técnica de *shotgun*) que são clonados em vetores plasmidiais que serão seqüenciados em suas extremidades. Após uma primeira etapa de montagem desse genoma, fragmentos maiores são clonados em cosmídeos e seqüenciados. Essa segunda etapa é importante para a montagem do genoma completo do organismo, já que a primeira normalmente produz uma seqüência incompleta, apresentando alguns buracos de seqüência (*gaps*).

Já em projetos genomas de organismos eucariotos, que possuem freqüentemente uma enorme quantidade de DNA, normalmente prefere-se adotar uma técnica conhecida como *shotgun* hierárquico. Nessa técnica, o DNA inteiro do organismo é primeiramente inserido em grandes vetores de clonagem, como cromossomos artificiais de bactérias (BACs) ou de leveduras (YACs). Depois então é realizado um *shotgun* desses grandes fragmentos dos vetores, gerando fragmentos menores que são agora clonados em vetores plasmidiais para o sequenciamento. Portanto, tais projetos consistem de duas etapas, a montagem de cada um dos grandes fragmentos clonados nos BACs e YACs e a montagem final que reunirá as seqüências completas dos BACs e YACs montados para a reconstituição da informação genômica inicial.

BOX4 - Softwares mais utilizados para o alinhamento de seqüências

ClustalW http://www.ebi.ac.uk/clustalw/index.html
Versão web de um dos programas de alinhamento múltiplo mais utilizados (Clustal). Fornece ao usuário uma grande quantidade de parâmetros e de saídas diferentes. Possui interface gráfica onde os alinhamentos podem ser visualizados de forma agradável e alterados.
Multialin http://prodes.toulouse.inra.fr/multalin/multalin.html
Programa de alinhamento múltiplo bastante conhecido. Fácil e rápido.
Fasta http://www.ebi.ac.uk/fasta33/
Precursor dos programas de alinhamento.
Promove serviço de busca em banco de dados de ácidos nucléicos e proteínas.
BLAST, BLAST2sequences http://www.ncbi.nlm.nih.gov/BLAST/
BLAST é o programa de alinhamento mais utilizado no mundo. Realiza a busca por seqüências homólogas em banco de dados de ácidos nucléicos e proteínas. O programa *BLAST 2 sequences* consiste no algoritmo BLAST para alinhamento de duas seqüências.

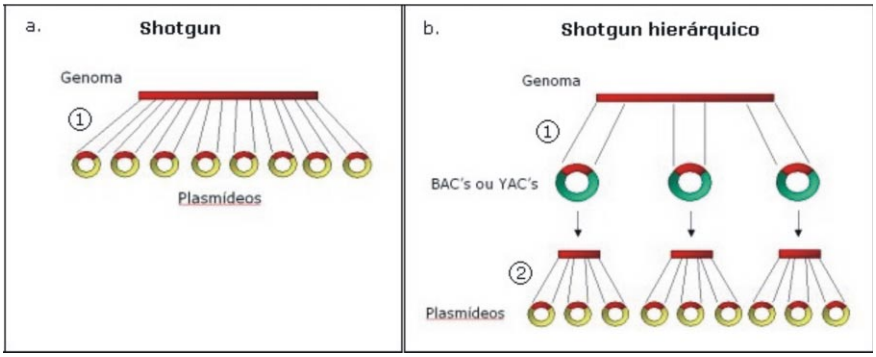

Figura 5. a) Na estratégia de shotgun*, todo o DNA genômico de um organismo é fragmentado em pequenos pedaços (1), que são clonados em vetores de pequeno porte, como plasmídeos, para o posterior seqüenciamento. b) Na estratégia de* shotgun *hierárquico, normalmente utilizada para grandes genomas, realizam-se dois passos. (1) Primeiramente fragmenta-se o genoma em grandes pedaços, que são clonados em vetores de grande porte, como BACs ou YACs. (2) Posteriormente realiza-se uma segunda etapa de shotgun, onde as seqüências contidas nesses vetores são fragmentadas em pequenos pedaços e clonadas em vetores de pequeno porte, que serão sequenciados*

Muitas vezes, ao invés de ser realizado o seqüenciamento genômico de um organismo eucarioto, prefere-se realizar o seqüenciamento só das regiões gênicas, utilizando informações oriundas de RNA mensageiro (mRNA). Dessa forma é realizada uma biblioteca de cDNA, representando o conjunto de mRNAs de uma célula, que são clonados em vetores plasmidiais. Os insertos de cDNA presentes em tais vetores são então seqüenciados a partir de suas extremidades 5' ou 3', produzindo pequenas seqüências que irão representar pedaços dos genes expressos no momento da extração do mRNA da célula em questão. Esses pedaços seqüenciados representam etiquetas de genes expressos, ou ESTs (*Expressed Sequence Tags*) e uma análise dos genes expressos é uma abordagem bastante utilizada na tentativa de entender o funcionamento do metabolismo dos mais diversos organismos. Como exemplo, no Brasil abordagens transcriptômicas já foram utilizadas em larga escala no projeto da cana-de-açúcar e vêm sendo utilizados em organismos parasitas, como é o caso dos projetos de seqüenciamento de ESTs de *Schistosoma mansoni* em São Paulo e em Minas Gerais.

Como já foi mencionado anteriormente, normalmente adota-se a estratégia de seqüenciamento genômico em organismos cujo genoma é pequeno e que contém baixa quantidade de seqüências repetitivas. Entretanto, a estratégia de seqüenciamento do transcriptoma, ou a produção de ESTs, não é utilizada apenas quando o genoma do organismo é muito grande. Essa estratégia é importante também para estudar o desenvolvimento dos organismos, produzindo bibliotecas de diferentes fases de desenvolvimento e observando quais genes são expressos em cada momento. Tal abordagem também é importante para estudarmos como ocorre a expressão diferencial de genes em diferentes órgãos de um mesmo organismo, para que possamos entender a função desses órgãos ou como eles realizam funções conhecidas. Portanto podemos dizer que as estratégias de seqüenciamento de genomas e transcriptomas são complementares e ambas devem ser realizadas, quando possível, para que possamos obter informações relevantes sobre os organismos que estamos estudando.

Base calling

Os dados brutos provenientes do seqüenciador de DNA são normalmente submetidos diretamente a algum programa de *base calling*. O *base calling* consiste no processo de leitura dos dados do seqüenciador e identificação da seqüência de DNA gerada, atribuindo ainda um valor de qualidade para cada posição nucleotídica identificada. Normalmente cada seqüenciador apresenta um programa de *base calling* associado. Entretanto, o programa mais utilizado nessa etapa é o PHRED.

O PHRED reconhece dados de seqüências a partir de arquivos *SCF* (*Standard Chomatogram Format*), arquivos de cromatograma dos analisadores automáticos de DNA *ABI* e arquivos *MegaBACE ESD*. Este *software* reconhece a seqüência de nucleotídeos a partir do arquivo de dados brutos do seqüenciador, atribui valores de qualidade às bases constituintes da seqüência nucleotídica e gera arquivos de saída contendo informações sobre o *base call* e os valores de qualidade. O valor de qualidade das seqüências analisadas pode ser encontrado nos arquivos FASTA e PHD.

De acordo com Ewing *et al* (1998) as atribuições seguras de valores às seqüências nucleotídicas são proporcionadas pela implantação de um algoritmo que tem como base os métodos de Análise de Fourier. O algoritmo analisa as quatro bases e prediz a provável região central dos picos e as distâncias relativas entre os picos da seqüência de DNA. O valor de qualidade atribuído a cada base é obtido pela fórmula a seguir, que calcula a probabilidade de erro no *base call*, onde o *Pe* é a probabilidade de uma base estar errada.

$$\text{PHRED Quality} = -10 \log (Pe)$$

As pontuações inseridas nos arquivos de saída do PHRED representam a probabilidade logarítmica negativa em escala de erro de um *base call*; portanto, quanto maior o valor de qualidade do PHRED, menor a probabilidade de ter ocorrido um erro. Só como exemplo, um valor de PHRED 20 para uma determinada posição nucleotídica significa que ela apresenta uma chance em 100 de estar errada. Já um valor de PHRED 30 significa que determinada base apresenta uma chance em 1000 de ter havido um erro no *base calling*. Esses valores são importantes para determinar se uma região precisa ser resseqüenciada.

Mascaramento de vetores

A estratégia freqüentemente adotada após a realização do *base calling* é a

procura por regiões de contaminantes na seqüência produzida. Regiões contaminantes são partes da seqüência obtida que não representam o DNA ou o cDNA que se deseja analisar. Tais regiões representam, normalmente, partes dos vetores de clonagem onde as seqüências de interesse foram inseridas ou pedaços de DNA adaptadores utilizados durante a construção das bibliotecas. Como essas regiões não representam as seqüências que se deseja analisar, elas devem ser retiradas ou mascaradas por um programa. E aqui, o programa mais utilizado é o Cross_match. Esse é, na verdade, um programa para a comparação de duas seqüências e é preciso utilizar como entrada um arquivo apresentando a seqüência dos vetores que se deseja mascarar. O que o Cross_match faz é comparar a seqüência desejada com o arquivo de seqüências de vetores e, onde o programa encontrar similaridade entre as seqüências, ele irá mascarar (acrescentando letras X) a seqüência de entrada. Assim, os nucleotídeos das seqüências de entrada similares a regiões de vetores de clonagem serão alterados para X e não atrapalharão os processos posteriores de análise computacional.

Agrupamento de seqüências

Após a geração de arquivos sem contaminantes, contendo a identificação das bases e a qualidade, todas essas informações são repassadas a um *software* de montagem como o PHRAP, o CAP3 ou o TIGR Assembler. O *software* mais utilizado nessa etapa, o PHRAP (*Phragment Assembly Program*) é o programa responsável pela leitura das informações do *base call* e montagem dos pequenos fragmentos de DNA seqüenciados em seqüências maiores, os contíguos (*contigs*). Este programa possui diversos pontos chaves para a obtenção de resultado final satisfatório, como: construção de seqüência do contíguo através de um mosaico de partes das seqüências com alta qualidade; utilização de informações da qualidade dos dados computados internamente e de implementações feitas pelos usuários para aumentar a qualidade da montagem; apresenta extensivas informações sobre a montagem realizada (incluindo valores de qualidades para a seqüência dos contíguos). Em projetos genoma espera-se obter, na saída do PHRAP, a seqüência montada do contíguo genômico. Já em projetos trancriptoma esperamos obter as seqüências de cada dos genes expressos após a execução deste software de montagem.

A visualização e edição das seqüências geradas após a montagem são realizadas normalmente através do programa Phrapview ou Consed.

O processo de anotação gênica

Uma vez obtidos os dados do seqüenciamento das moléculas de DNA é preciso saber o que representa cada uma das seqüências nucleotídicas produzidas. A anotação consiste simplesmente no processo de identificação dessas seqüências. Em projetos genoma, este processo normalmente é realizado em três etapas: anotação de seqüências de nucleotídeos, de seqüências protéicas e de processos biológicos.

Figura 6: Interface do programa Consed

BOX5 - Programas mais utilizados em projetos genoma e transcriptoma

PHRED http://www.phrap.org
Software para a realização do *base calling* e a produção do cromatograma processado.
CROSS-MATCH http://www.phrap.org
Software para a comparação entre duas seqüências de DNA. Normalmente utilizado para o mascaramento de regiões representando vetores em seqüências genômicas ou de cDNA. Distribuído juntamente com o PHRAP.
PHRAP http://www.phrap.org
Software mais utilizado para a realização do agrupamento de seqüências (*clustering analysis*) e montagem de contíguos genômicos.
CAP3 http://genome.cs.mtu.edu/cap/cap3.html
Software utilizado para o agrupamento de seqüências e montagem de contíguos genômicos. Utiliza um algoritmo diferente do PHRAP.
CONSED http://www.phrap.org
Software mais utilizado para a visualização dos resultados obtidos por softwares de agrupamento de seqüências. Permite a edição das bases seqüenciadas, além de diversos outros recursos.

Figura 7: Etapas da anotação em projetos genoma e as perguntas que se deseja responder em cada uma delas

A partir da anotação de seqüências nucleotídicas procura-se, primeiramente, identificar a natureza de uma determinada seqüência. Devemos descobrir se tal seqüência está inserida em uma região gênica, se representa uma molécula de RNA transportador ou RNA ribossômico, se pertence a algum tipo de região repetitiva já descrita ou se apresenta algum marcador genético conhecido em seu interior. O principal objetivo dessa etapa é construir um mapa do genoma do organismo, posicionando cada um dos possíveis genes e caracterizando as regiões não-gênicas. Nesta fase, alguns programas de predição gênica são usados para a localização de possíveis genes nas seqüências de DNA. A procura por elementos como o códon de iniciação de proteínas (a trinca de nucleotídeos ATG) e códons de terminação na mesma fase de leitura são utilizados por alguns desses programas. O tamanho delimitado por esta janela de leitura é freqüentemente utilizado para definir uma determinada região como sendo gênica ou não. Alguns outros programas são capazes de identificar, dependendo do genoma analisado, regiões gênicas codificadoras (éxons) e não codificadoras (íntrons). Alguns exemplos são o GenomeScan e o GenScan. Em projetos de trancriptômica, onde se utiliza a abordagem de seqüenciamento de ESTs, essa etapa não é realizada, uma vez que todas as seqüências produzidas se restringem a regiões gênicas.

Mapeados os genes, a etapa seguinte consiste em identificar quais proteínas são codificadas, e nisso consiste o processo de anotação das seqüências protéicas. Nessa etapa, procura-se montar um catálogo dos genes presentes no organismo estudado, dando-lhes nomes e associando-os a prováveis funções. No caso de projetos genoma, deseja-se identificar o número total de genes presentes no organismo seqüenciado, já que há informação da seqüência de DNA de todo o genoma. Já em projetos transcriptoma, a tarefa consiste em identificar os genes expressos no organismo em uma determinada condição. Apesar de não ser capaz de identificar todos os genes de um determinado organismo, os projetos de transcriptômica podem permitir a identificação de genes expressos em diferentes tecidos e fases de desenvolvimento, além de permitir a observação daqueles que apresentam variantes de *splicing*. Portanto, nessa etapa da anotação, o principal objetivo é identificar e caracterizar cada uma das proteínas codificadas pelos mRNAs presentes no organismo estudado em determinada condição.

A parte mais interessante e desafiadora dos processos de anotação gênica é relacionar, finalmente, a genômica com os processos biológicos, e essa é a etapa de anotação dos processos biológicos. Essa etapa é comum a projetos genoma e transcriptoma. Identificados os genes, devemos agora tentar relacioná-los de modo a obtermos um mapa funcional do organismo estudado. Nesse ponto deve-se identificar quais vias bioquímicas estão completas ou incompletas no organismo e quais vias alternativas ele possui. Aqui é fundamental a participação de biólogos especialistas em diversas áreas para que se possa descobrir como o metabolismo do organismo pode influenciar seu modo de vida e seu comportamento. Esse é o momento onde é possível levantar várias hipóteses que relacionem o funcionamento dos organismos com seus dados genômicos. Tais hipóteses devem ser testadas experimentalmente, por pesquisadores que trabalhem com o organismo estudado.

Como é realizada a anotação

Até aqui foi mostrado o que é normalmente feito em um processo de anotação gênica. Vejamos agora como tal processo é realizado. Lincoln Stein definiu muito bem como acontece a sociologia dos projetos de anotação gênica. Ele dividiu o processo de anotação de genomas em três etapas: a fábrica, o museu e a festa.

BOX6 – Principais softwares utilizados durante a anotação gênica

Repeat Masker http://repeatmasker.genome.washington.edu/
Utilizado para a identificação e o mascaramento de regiões repetitivas freqüentemente encontradas em genomas.
Genscan http://genes.mit.edu/GENSCAN.html
Utilizado para a predição de genes em genomas eucarióticos. Seu método de predição é baseado em cadeias escondidas de Markov.
tRNAscan-SE http://www.genetics.wustl.edu/eddy/tRNAscan-SE/
Utilizado para encontrar genes de tRNA em uma seqüência genômica.
BLAST http://www.ncbi.nlm.nih.gov/BLAST
Utilizado para encontrar similaridades entre seqüências de nucleotídeos e proteínas contra bancos de dados com grande número de seqüências dos mais diversos organismos. É um dos principais programas utilizados na identificação dos genes.
Interpro http://www.ebi.ac.uk/interpro
Utilizado para realizar buscas contra diferentes bancos de dados de domínios e famílias de proteínas. Integra os serviços do Pfam, PRINTS, ProDom, PROSITE, SMART, TIGRFAMs e SWISS-PROT.
GeneOntology http://www.geneontology.org
Consórcio destinado a produzir um vocabulário comum a ser aplicado para a classificação dos genes presentes em organismos eucarióticos. Cada gene é classificado em três níveis: função molecular, processos celulares e localização celular.

Na primeira etapa trabalham apenas as ferramentas de bioinformática, funcionando em larga escala, como uma fábrica. Assim, as seqüências obtidas passam por uma grande diversidade de programas, que devem ajudar os anotadores a identificá-las e agrupá-las para a próxima fase.

A segunda etapa necessita de especialistas que observem os dados obtidos na primeira etapa pelas ferramentas automáticas e que, como curadores de um museu, identifiquem as seqüências de acordo com critérios pré-definidos.

Após a identificação dos genes, é feita a anotação dos processos. Nesse momento deve-se promover a interação entre vários anotadores, bioinformatas e biólogos especialistas em diferentes áreas e no organismo estudado. Nessa festa deve-se discutir como as informações obtidas nas etapas anteriores podem estar relacionadas com a biologia do organismo em questão.

A era pós-genômica

Uma das características mais fascinantes da explosão, ocorrida nos últimos 10 anos, de projetos e consórcios destinados a compor o genoma completo dos mais diversos organismos, foi o estabelecimento de abordagens e tecnologias que permitiram um estilo "linha-de-montagem" na obtenção, em tempos cada vez mais curtos, de quantidades "industriais" de seqüências de ácidos nucleicos (DNA e RNA). Agora começamos a enfrentar o problema de interpretar e adicionar significado a essas seqüências. Temos agora que, a partir dos bancos de dados existentes, processar e correlacionar os dados brutos transformando-os em informação e a partir desta informação gerar conhecimento, que é a informação testada experimentalmente. No final, esta nova etapa promete ser uma jornada, provavelmente sem fim, através das proteínas, suas estruturas e funções, vias metabólicas e interações celulares. Esta mudança do foco de atenção, dos ácidos nucleicos para as proteínas, tem sido utilizada para batizar esta nova etapa da pesquisa biológica em larga escala como "Era Pós-Genômica". Contudo, trata-se apenas de mais uma etapa e, certamente, não a última para que os frutos dos programas de seqüenciamento de genomas possam ser colhidos. Etapas estas que foram previstas pelo Projeto do Genoma Humano. Das cinco metas a serem atingidas, o estudo da expressão de proteínas e a obtenção de mapas de interação proteína-proteína ocupam o segundo e terceiro estágios, dos quais se espera o maior impacto econômico, levando à descoberta de novas drogas e reduzindo o seu tempo de entrada no mercado.

Resumidamente, na Era Pós-Genômica procura-se estudar a expressão dos genes codificados pelo genoma dos organismos, tecidos, células ou compartimentos celulares em determinadas condições fisiológicas (por exemplo, uma doença, uma situação de estresse ou ainda a administração de uma droga). Tentando entender a resposta a essas condições, são alvos de estudos: a ativação ou repressão de determinados genes, a indução de mudanças no estado pós-traducional das proteínas e qualquer processo que resulte na modificação do número e/ou da composição das proteínas existentes.

Análise da Expressão Gênica

Lembrando do dogma central da biologia (DNA → mRNA → Proteína), é facil perceber que podemos avaliar a expressão gênica através da análise de transcritos (mRNA).

Em organismos eucariotos, a facilidade de isolamento dos mRNAs (usando oligonucleotídeos poli-T para capturar os mRNAs pela cauda poli-A), a possibilidade da transcrição reversa do mRNA para cDNA (usando a técnica de RT-PCR) e o domínio das técnicas de seqüenciamento em massa de cDNAs tornaram possível a análise qualitativa e quantitativa, em larga escala, dos genes transcritos em organismos, tecidos e células. Desta forma, nos projetos Transcriptoma, como já comentado, é feito o seqüenciamento parcial de cDNAs representativos da população de mRNA de maneira a permitir a identificação de diferentes transcritos (pela comparação das seqüências do cDNA) e sua abundância na população (pelo número de vezes em que cada transcrito é seqüenciado). As técnicas mais usadas são as de ESTs e SAGE (*Serial Analysis of Gene Expression*). Nesta última técnica, mais recente, são gerados e seqüenciados concatâmeros de fragmentos de cDNAs com apenas 10 ou 17 nucleotídeos de cada mensageiro, respectivamente denominados *SAGE tags* e *SAGE long tags*.

DNA chips e Microarrays

Uma outra forma de análise de transcritos, que permite a busca de transcritos de genes específicos na população dos mRNAs expressos, usa o já conhecido princípio da hibridação de DNA a sondas moleculares. As mais novas versões da técnica são os *DNA chips* e os *microarrays*, que permitem a análise simultânea da expressão de milhares de genes. Nestas duas técnicas, respectivamente, oligonucleotídeos ou fragmentos de cDNA conhecidos são ligados a uma lâmina de vidro e, em cada experimento de hibridação, os mRNAs de dois tipos celulares diferentes ou de células em duas condições patológicas ou tratamentos são analisados. As duas populações de mRNAs são amplificadas e marcadas com diferentes corantes fluorescentes (cianinas ou Cys), um verde e outro vermelho. Ao hibridarem com cada gene (oligo ou cDNA) aplicado sobre a lâmina de vidro, a cor verde ou vermelha de cada ponto (ou *spot*) indicará que esse gene está sendo mais transcrito em um tipo ou condição celular do que no outro. A cor amarela indicará que o gene é transcrito igualmente em ambos os tipos ou condições celulares. Além disso, a maior ou menor intensidade de cada cor indicará maior ou menor nível de expressão do gene.

A enorme quantidade de dados gerada nos experimentos de *DNA chips* e *microarrays* são analisados por softwares específicos que envolvem métodos de inferência estatística. Uma etapa bastante importante na fase de análise dos resultados é a que chamamos de normalização. Usando como referência os *spots* de genes controles (sabidamente expressos ou reprimidos nos tecidos ou células estudados), o que se busca é, basicamente, retirar dos valores de cada *spot* a influência de

manchas espúrias (*background*) e de variações do processo de hibridação. Desta forma, após a normalização, torna-se possível a comparação de *spots* de uma mesma lâmina ou de experimentos diferentes. Em uma etapa posterior, programas de *clustering* procuram identificar e agrupar os *spots* super-expressos, reprimidos ou que não tem expressão alterada nos tecidos ou células analisadas. Apesar dos métodos de análise empregados, a falta de reprodutibilidade dos resultados ainda é uma queixa bastante comum. O uso de maior número de réplicas de cada *spot* e/ou a busca de métodos de inferência estatística mais adequados parecem ser úteis para a validação destes resultados.

Mais recentemente, com novas técnicas para isolamento de mRNA de procariotos, projetos de ESTs e de microarray também têm sido desenvolvidos para estes organismos. Vários grupos de pesquisa em todo o Brasil estão iniciando projetos nesta área. Apenas como exemplo, entre os vários projetos brasileiros nesta área temos o projeto Cooperation for Analysis of Gene Expression (CAGE) (http://bioinfo.iq.usp.br/ e http://www.vision.ime.usp.br/~cage/) e o Projeto Genoma Raízes da Embrapa Soja (http://www.cnpab.embrapa.br/pesquisas/gp.html).

BOX7 – Exemplos de Projetos Transcriptoma:

Procuram avaliar quais são os genes expressos, e quanto deles é expresso, a partir do seqüenciamento parcial dos mRNAs transcritos.
Dados obtidos pela técnica de SAGE podem ser consultados na página http://www.ncbi.nlm.nih.gov/SAGE/. Já no banco dbEST estão depositadas ESTs de diversos Projetos Transcriptoma desenvolvidos em todo o mundo (http://www.ncbi.nlm.nih.gov/dbEST/).

Mais informações sobre *DNA Chips* e *Microarrays*

Nestas técnicas, a verificação da expressão de genes específicos é feita em experimentos de hibridação em lâminas de vidro contendo milhares de fragmentos de DNA.
Na página http://cmgm.stanford.edu/pbrown/, do pioneiro da técnica de *microarray*, Dr. Patrick Brown, há mais explicações, um forum de discussão e bancos de dados de *microarrays*. Na página http://ihome.cuhk.edu.hk/~b400559/array.html há informações sobre os equipamentes necessários, uma tabela de comparação dos programas de análise mais usados, noções de estatística aplicadas a *microarrays*, sugestões de bibliografia, etc.

Programa gratuíto para análise de *microarrays*

ScanAlyse: escrito por Michael Eisen, o programa pode ser obtido gratuitamente na página http://rana.lbl.gov/EisenSoftware.htm. Assinando um termo de compromisso, o autor permite, inclusive, o acesso ao código-fonte.

Projetos Proteoma

Um problema que surge com a abordagem descrita acima, de avaliação da expressão gênica a partir da análise dos mRNAs transcritos, é que nem sempre a quantidade de um mRNA reflete a quantidade da proteína correspondente expressa na célula e, assim, não podemos relacionar diretamente essa proteína a uma função nas células. Por isto, uma outra abordagem, embora muito mais trabalhosa, tem sido usada para avaliar a expressão gênica: a análise das proteínas expressas. Esta "contrapartida protéica" do genoma é conhecida como proteoma**.** Por permitir relacionar diretamente a uma proteína determinada função, esta abordagem constitui um instrumento particularmente poderoso para elucidar os mecanismos celulares relacionados ao desenvolvimento de doenças, ao mecanismo de funcionamento de compostos químicos (por exemplo, fármacos) e identificar novos alvos terapeuticos.

As bases experimentais da proteômica não são novas e pertencem ao arsenal "clássico" da bioquímica, mas houve, nos últimos anos, um salto qualitativo e quantitativo sem precedentes. Esse salto foi resultado de grandes investimentos privados na busca de abordagens mais agressivas e rápidas no isolamento, identificação e caracterização de proteínas, no mesmo estilo "industrial" que caracterizou a era genômica. O isolamento de proteínas em grande número, inicialmente repousava nas técnicas eletroforéticas, como a eletroforese mono e bi-dimensional em géis de poliacrilamida. Embora tais técnicas certamente sempre venham a ter um papel importante em qualquer laboratório de proteômica, nota-se hoje uma tendência cada vez maior no uso da cromatografia líquida de alta eficiência, com o uso de colunas capilares, no desempenho desta tarefa. A identificação e caracterização das proteínas depende de um conjunto de tecnologias (com certeza as que mais sofreram incremento no desempenho) envolvendo a espectrometria de massa, a ressonância magnética nuclear, além de recursos computacionais para a armazenagem, análise e compartilhamento dos diversos tipos de dados gerados por estas tecnologias (imagens de géis bidimensionais, sequências protéicas, estruturas protéicas, espectros de massa, etc.).

Nos últimos anos a espectrometria de massa, em conjunto com a cromatografia líquida de alta performance, vem se tornando a abordagem preferida para identificar e caracterizar proteínas, devido essencialmente a três motivos. O primeiro é o desenvolvimento de novos métodos para ionização de proteínas e peptídeos, especialmente o MALDI e o ESI (*Matrix-Assisted Laser Dessorption-Ionization* e *ElectroSpray Ionization*). O segundo é o desenvolvimento de recursos da bioinformática, permitindo a análise de dados obtidos por espectrometria de massas em bancos genômicos e de sequências protéicas. E o terceiro é que a espectrometria de massas fornece informação detalhada de modificações pós-traducionais, em particular as fosforilações e glicosilações.

BOX8 – MALDI e ESI

MALDI - Matrix-Assisted Laser Desorption-Ionization
Uma amostra de proteína ou peptídeo é misturada com um largo excesso de uma matriz, formada por uma substância que absorve no ultra-violeta, e posta para secar. Um *laser* com um comprimento de onda que seja absorvido pela matriz, em um compartimento sob vácuo, incide sobre a amostra seca e fragmentos ionizados da amostra são carreados pela vaporização da matriz e capturados por um campo elétrico do analisador de massas.

ESI - ElectroSpray Ionization
Um voltagem aplicada em uma fina agulha contendo uma solução protéica, gera uma névoa de pequenas gotículas da solução, contendo pequeno número de moléculas protéicas. A redução das gotículas por evaporação acaba colocando em fase gasosa as proteínas ionizadas. Elas são então capturadas pelo analisador de massas. A grande vantagem desta técnica é permitir o acoplamento direto de um sistema cromatográfico de alta eficiência ao espectrômetro de massas, possibilitando a análise em fluxo contínuo de misturas protéicas complexas.

No Brasil, apenas agora começamos a montar grupos de pesquisa nesta área. Merecem destaque as redes de proteômica em São Paulo, sediada no Laboratório Nacional de Luz Síncrotron (http://www.lnls.br/), e no Rio de Janeiro (http://www.faperj.br/interna. phtml?obj_id=219).

BOX9 - Links interessantes

Eletroforese bi-dimensional em géis de poliacrilamida (PAGE-2D)
http://us.expasy.org/ch2d/protocols/
http://www.aber.ac.uk/parasitology/Proteome/Tut_2D.html

Cromatografia líquida de alta eficiência, com o uso de colunas capilares (HPLC)
http://www.ionsource.com/tutorial/chromatography/rphplc.htm
http://www.ionsource.com/tutorial/capillary/introduction.htm

Espectrometria de Massas (MS)
http://ms.mc.vanderbilt.edu/tutorials/ms/ms.htm

Software gratuíto para análise de PAGE-2D - Melanie
Desenvolvido no Swiss Prot, está disponível diretamente na página do Swiss Prot, http://www.expasy.org/ ou num *link* na página http://www.science.gmu.edu/ ~ntongvic/Bioinformatics/software.html, que dá acesso a muitos outros programas de bioinformática.

As técnicas experimentais expostas acima, além de oferecerem respostas à curiosidade humana, constituem formas inovadoras na pesquisa para o combate de problemas globais como diabetes, câncer, hemofilia, etc... Na prática, independentemente do número de proteínas codificadas pelo genoma da espécie humana (o que ainda hoje é discutido), é previsível que em alguns anos possamos conhecer de 4000 a 10000 proteínas-alvo, sobre as quais medicamentos poderão agir. Para termos uma idéia da grandeza destes números, todo o arsenal terapêutico que conhecemos hoje atua sobre apenas 500 delas. O número de drogas disponíveis hoje nos EUA, derivadas destas novas tecnologias, chegou a 103 no ano passado (21 delas foram aprovadas em 2000).

Modelagem molecular

Ainda neste sentido, procurando associar proteínas a suas funções, a bioinformática pode e deverá trazer, nas próximas décadas, suas maiores contribuições à biologia. O conhecimento da estrutura terciária de uma proteína constitui uma informação valiosa para determinação de sua função, pois pode permitir a identificação de domínios conhecidos, como sítios catalíticos, sítios de modificação alostérica e outros.

Além disso, tendo as estruturas tridimensionais das proteínas determinadas, podemos então realizar pesquisas mais direcionadas no sentido de encontrar inibidores, ativadores enzimáticos e outros ligantes que permitam a produção de fármacos mais eficientes e específicos: o almejado Desenvolvimento Racional de Fármacos (*Rational Drug Design*).

Atualmente a abordagem mais eficaz na determinarção da estrutura terciária de proteínas é aquela que se utiliza de técnicas experimentais como NMR (Ressonância Magnética Nuclear) e cristalografia por difração de raios-X. Dezenas de milhares de protéinas tiveram suas estruturas terciárias conhecidas através destes métodos e têm fornecido dados para o desenvolvimento de programas de modelagem e para a modelagem por homologia. Entretanto os métodos experimentais são, frequentemente, procedimentos dispendiosos e de difícil execução. Além disso, existem limitações técnicas que dificultam a determinação de várias proteínas. A obtenção de cada proteína pura é um desses fatores limitantes. Outro fator é a dificuldade de cristalização das proteínas, etapa necessária para a determinação de estrutura por difração de raios-X. Este é um problema comum em proteínas de membrana ou glicosiladas. Mesmo usando robôs para acelerar o processo experimental, estas e outras dificuldades fazem com que a determinação de novas estruturas protéicas não consiga acompanhar a velocidade de obtenção de dados dos projetos genoma.

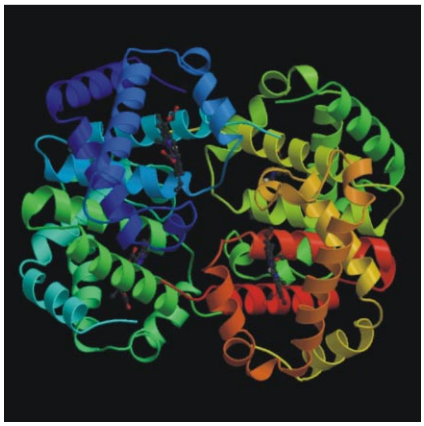

Figura 8: Estrutura terciária e quaternária da Deoxi-hemoglobina humana obtida por Difração de Raios X e depositada no PDB. A molécula é um tetrâmero, composta por 4 cadeias, e ligada a 4 átomos de ferro

A modelagem molecular é um método alternativo, não experimental, que permite, com base nos conhecimentos da estereoquímica dos aminoácidos e nas informaçoes adquiridas das estruturas terciárias já resolvidas, prever a conformação de proteínas a partir da seqüência primária dos aminoácidos.

Uma das formas de se realizar a modelagem de proteínas é utilizar como referência uma ou mais protéinas homólogas e de estrutura terciária já conhecida. Este tipo de modelagem é conhecido como *modelagem por homologia* ou *modelagem comparativa*, e, por enquanto, é a abordagem que obtém melhores resultados. O primeiro passo do processo é a pequisa de proteínas homólogas em bancos de dados de estruturas terciárias de proteínas. O PDB (*Protein Database Bank*) é o mais utilizado para este fim. A seguir, deve ser realizado o alinhamento das seqüências de aminoácidos das protéinas homólogas e a proteína-alvo (o programa Clustal, citado anteriormente no artigo, pode ser usado). A modelagem, propriamente dita, é realizada através de softwares como o Modeller, SWISS-MODEL, 3D-PSSM, dentre outros. Esses programas normalmente procuram encontrar a estrutura terciária que melhor se aproxime da disposição dos átomos das proteínas utilizadas como modelo, e ao mesmo tempo atenda às restrições estereoquímicas. Após a definição de uma estrutura candidada, esta pode ser avaliada através de outros softwares de verificação de restrições estereoquímicas, como o programa Procheck.

A modelagem por homologia é um processo iterativo de ajuste de parâmetros e verificação dos resultados. Normalmente é necessário que o processo seja repetido várias vezes até que uma estrutura terciária adequada seja obtida. Além disso, a modelagem de proteínas, como um todo, é uma técnica heurística: mesmo que a estrutura obtida concorde perfeitamente com todas as restrições impostas, não há garantias de que esteja correta. Deve-se lembrar que uma estrutura bastante semelhante à real pode ser o suficiente para formulação de novas hipóteses e atingir as expectativas do usuário desta técnica.

Uma abordagem recente, que possui um crescente números de adeptos e acumula bons resultados, é a modelagem através de *threading* de proteína. Esta técnica é baseada na comparação da proteína em questão com modelos descritivos dos enovelamentos de proteínas homólogas. Nesses modelos são descritas: a distância entre os resíduos de aminoácidos, a estrutura secundária de cada fragmento e as características fisico-químicas de cada resíduo.

Entretanto, um grande desejo dos que trabalham com proteínas é o desenvolvimento de programas realmente eficientes para a modelagem *ab initio*, ou seja, que sejam capazes de predizer a estrutura terciária de uma proteína, tendo como informação apenas a seqüência dos resíduos de aminoácidos e suas interações fisico-químicas, entre si e com o meio. Programas assim existem hoje mas têm muito a melhorar para que possamos confiar unicamente no seu resultado.

No geral, a modelagem de proteínas através de programas de computador é um campo de pesquisa recente e ainda não gerou softwares de eficiência comprovada. Para estimular o desenvolvimento de programas de molelagem molecular de proteínas, foi criado um evento para a avaliação desses *softwares* denominado CASP (*Critical Assesment of Structural Prediction*). A cada dois anos este evento reúne os mais conhecidos pesquisadores desta área, que são desafiados e suas diferentes metodologias avalia-

BOX10 – Programas e sites relacionados com modelagem e estruturas de proteínas

PDB http://www.rcsb.org/pdb/
Mais famoso e completo banco de dados de estrutura de proteínas.
Protein explorer http://molvis.sdsc.edu/protexpl/
Programa derivado do RasMol para a visualização de estruturas de proteínas.
SWISS-PDBviewer http://www.expasy.org/spdbv/
Programa para a visualização e análise da estrutura de várias proteínas ao mesmo tempo. Permite a realização de mutações de aminoácidos, alterações em pontes de hidrogênio, ângulos de torção e distâncias entre átomos.
Modeller http://guitar.rockefeller.edu/modeller
Um dos programas mais utilizados para a modelagem de proteínas por homologia.
SWISS-MODEL http://www.expasy.org/swissmod
Programa via web para a modelagem de proteínas por homologia.
PROCHECK http://www.biochem.ucl.ac.uk/~roman/procheck/procheck.html
Programa que checa a qualidade estereoquímica de uma estrutura de proteína, gerando análises gráficas sobre a geometria espacial da proteína, resíduo por resíduo.
Libra http://www.ddbj.nig.ac.jp/E-mail/libra/LIBRA_I.html
Programa on-line que utiliza *threading* para encontrar uma seqüência de resíduos de aminoácidos que melhor se adequem a uma estrutura terciária conhecida e vice-versa.
CASP http://predictioncenter.llnl.gov/Center.html
Critical Assesment of Structural Prediction. Competição que avalia os softwares de predição de estrutura de proteínas.

das. Nesta competição cada grupo recebe seqüências de proteínas tiveram sua estrutura resolvida experimentalmente por NMR e/ou cristalografia por difração de raios X, mas que ainda não foram publicadas. Vence o grupo que conseguir prever *ab initio,* com maior exatidão, a estrutura do maior número proteínas. Apesar dos esforços, até hoje não houve 100% de acerto.

Métodos em filogenética molecular

Uma das aplicações mais antigas da bioinformática é a de desenvolvimento de programas que, a partir das seqüências de DNA ou de proteínas de diferentes organismos, sejam capazes de reconstruir a relação de parentesco entre as espécies, o que chamamos de sistemática molecular, ou de reconstruir o parentesco entre as espécies associando essas informações a uma escala temporal, o que chamamos de filogenia molecular. A representação gráfica desses resultados é feita na forma de árvores filogenéticas.

Atualmente, árvores filogenéticas são extremamente comuns em artigos que abordam assuntos de biologia molecular, refletindo o reconhecimento de que estas árvores representam uma maneira legítima de entender os processos biológicos e a evolução dos mais diversos caracteres. Estes estudos e as ferramentas criadas para este fim têm aplicações tão diversas como procurar entender a origem do homem ou reconstituir a história epidemiológica da AIDS a partir de dados do genoma do vírus HIV.

Para realizar inferências a respeito das relações de parentesco entre organismos, tomando como base seqüências de DNA ou proteínas, o primeiro passo é identificar seqüências de interesse que apresentem ancestralidade comum, ou seja, que sejam homólogas. Para isto, muitas vezes estas seqüências são escolhidas por similaridade nos grandes bancos de dados disponíveis na rede, sem que tenhamos, sobre elas, dados das funções bioquímicas e biológicas que possam confirmar sua homologia. Por isso, é importante ressaltar que, ao fazermos uma reconstrução filogenética, a escolha de seqüências homólogas é fundamental para gerar uma árvore confiável, pois só assim teremos certeza de que estaremos comparando um mesmo marcador que apresenta similaridades entre vários organismos a partir de uma origem comum, garantindo que eles compartilham um mesmo ancestral. Quando não se comparam caracteres homólogos, pode-se incidir no erro de considerar similaridades sem origem comum e, portanto, com histórias evolutivas diferentes. Uma das formas de avaliar esta escolha é incluir nas análises, seqüências de grupos externos (organismos com historia evolutiva conhecida em relação ao grupo em estudo), que funcionam como controles no processo de reconstrução de parentescos.

Uma vez selecionadas as seqüências homólogas dos organismos de interesse e de grupos externos, será necessário realizar o alinhamento múltiplo entre elas e então gerar árvores filogenéticas a partir de métodos de distância ou de caracteres discretos (máxima parcimônia ou máxima verossimilhança) para podermos realizar a inferência filogenética desejada. Para tanto, os seguintes métodos são freqüentemente utilizados pelos softwares:

Métodos de Distância

Funcionam basicamente em dois passos, sendo que o primeiro deles é a redução das variações entre seqüências alinhadas a valores de distância dispostos em uma matriz. No segundo passo, estes valores são utilizados na reconstrução filogenética. Um dos métodos de distância mais comuns é a chamada *distância p*, que expressa o número de sítios variáveis entre duas seqüências com relação ao total de sítios comparados. Além deste, existem também muitos outros modelos evolutivos utilizados para o cálculo de distâncias genéticas, como o Jukes-Cantor, Kimura 2 parâmetros, Tajima e Nei e Tamura 3 parâmetros. Na reconstrução filogenética, os algoritmos mais utilizados são o UPGMA (*Unweighted Pair Group Method with Arithmetic means*) e o Neighbor-joining, que realizam uma série de cálculos com a matriz de distância gerada a partir do alinhamento para estimar a árvore filogenética.

Máxima Parsimônia (MP)

Este método baseia-se na teoria de que a melhor hipótese para explicar um processo é aquela que requer o menor número de passos. Para a análise filogenética, isto significa que a árvore que possuir um menor número de mudanças (substituições) para explicar os dados do alinhamento é a mais próxima da real. Na MP não há a fase de cálculo de distância, sendo que as árvores são calculadas diretamente dos dados do alinhamento. Entretanto, esta metodologia requer muito mais tempo quando se usa a busca exaustiva de árvores, uma vez que o computador precisa reconstruir todas as árvores possíveis para "escolher" aquelas com um número mínimo de mudanças, que são chamadas de árvores mais parcimoniosas. Para contornar este problema do tempo, existem também algoritmos heurísticos de reconstrução filogenética, mas é preciso lembrar que, nestes casos, a árvore final pode ser subótima.

Máxima Verossimilhança (MV)

Este método baseia-se na reconstrução filogenética através da busca por uma árvore que maximize a probabilidade dos dados observados. Neste sentido, o método de MV calcula as probabilidades associadas a diferentes topologias e cada uma delas com as variações nos tamanhos dos ramos, considerando o modelo evolutivo escolhido. Portanto, encontrar a árvore mais verossímil envolve não somente a análise das topologias possíveis, mas também das variações de comprimento de ramos para cada topologia. Deste modo, o emprego de algoritmos heurísticos pode auxiliar enormemente na busca pela árvore ideal, já que o tempo computacional aumenta de acordo com o número de espécies e de parâmetros considerados na análise.

A cada vez que um programa de filogenia molecular é rodado para gerar uma árvore sobre o conjunto de dados escolhidos, o resultado pode ser diferente. Por isso, para validar uma árvore filogenética, o que se faz é rodar repetidas vezes o programa escolhido e, estatisticamente, testar cada ramo para escolher um a um aqueles com maior probabilidade de ocorrência para a composição final da árvore. O método estatístico mais usado nessas análises é o chamado *bootstrap*.

O *bootstrap* funciona gerando conjuntos modificados de dados, obtidos aleatoriamente a partir dos dados do alinhamento. Para cada conjunto aleató-

BOX11 - Programas mais utilizados na análise filogenética
Clustal

Programa para o alinhamento múltiplo de seqüências
Acesso on line - http://www.ebi.ac.uk/clustalw/
Download do clustal X para diversas plataformas - http://inn-prot.weizmann.ac.il/software/ClustalX.html
PAUP 4.0 (Phylogenetic Analysis Using Parsimony and other methods) - http://paup.csit.fsu.edu/
Análises filogenéticas utilizando métodos de distância, máxima parcimônia e máxima verossimilhança
PHYLIP (Phylogeny Inference Package) – inferências filogenéticas
http://evolution.genetics.washington.edu/phylip.html
MEGA (Molecular Evolutionary Genome Analysis) - http://www.megasoftware.net/
Inferências filogenéticas com métodos de distância e parcimônia. Download gratuito.
Treeview http://taxonomy.zoology.gla.ac.uk/rod/treeview
Software gratuito para edição gráfica e impressão de árvores filogenéticas

rio de dados obtidos é estimada uma árvore. As novas árvores, geradas a partir dos conjuntos modificados dos dados de entrada, são comparadas. Cada um dos ramos da árvore final recebe então um valor de probabilidade, que é obtido do número de novas árvores onde esse ramo ocorreu dividido pelo número total de novas árvores estimadas. Probabilidades altas indicam que, mesmo com algumas alterações, os dados suportam o ramo ao qual essa probabilidade se refere e probabilidades baixas significam que, com a amostra analisada, não se pode ter certeza de que determinado ramo seja correto.

CONSIDERAÇÕES FINAIS

Tentamos abordar nesse artigo os principais tópicos desenvolvidos em bioinformática. Este artigo não pretende esgotar cada um dos assuntos abordados, mas imaginamos que os leitores interessados poderão encontrar mais informações e trilhar seu próprio caminho visitando os links e observando as referências sugeridas.

Agradecimentos

Sendo este trabalho fruto do aprendizado obtido no II Curso de Especialização em Bioinformática, realizado de agosto a novembro de 2002 em Petrópolis - RJ, os autores gostariam de agradecer principalmente ao CNPq pelo suporte financeiro concedido para a realização do curso e ao LNCC (Laboratório Nacional de Computação Científica) por sediar este evento, em especial à coordenadora do curso, Ana Tereza Vasconcelos. Agradecemos também a todos os nossos professores: Darcy de Almeida, Richard Garratt, Glaucius Oliva, Patricia Palagi, Marie Anne Van Sluys, Cláudia Russo, Anamaria Camargo, Helena Brentani, Sandro de Souza, Jorge de Souza, Luiz Gonzaga, Frank Alarcon, Fernanda Raupp, Daniele Quintella, Helio Barbosa, Alexandre Plastino, Dorival Leão, Marcos Grivet, Simone Martins e a todo o pessoal do Laboratório de Bioinformática do LNCC.

Agradecemos também a nossos orientadores e às instituições e órgãos de financiamento nacionais e estaduais pelo apoio dado a cada um de nós para a participação no Curso de Especialização em Bioinformática do LNCC.

REFERÊNCIAS BIBLIOGRÁFICAS

1. Altschul SF *et al.* **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 25: 3389-3402. 1997.
2. Baxevanis AD, Ouellette BFF. **Bioinformatics: A practical guide to the analysis of genes and proteins.** Ed. Wiley-interscience. 2nd ed. 2001. 470p.
3. Clote P, Backofen R. **Computational Molecular Biology: An introduction.** John Wiley & Sons, LTD. 2000. 286p.
4. Ewing B, Green P. **Base-calling of automated sequencer traces using phred. II. Error probabilities.** *Genome Res* 8:186-94. 1998.
5. Frishman D *et al.* **Comprehensive, comprehensible, distributed and intelligent databases: current status.** *Bioinformatics Review*, **14**, 551-561. 1998.
6. Huang X, Madan A. **CAP3: A DNA Sequence Assembly Program.** *Genome Biol* 9: 868-877. 1999.
7. Hunt SP, Livesey FJ. **Functional genomics.** Oxford University Press. 2000. 253p.
8. Matioli RM. **Biologia Molecular e Evolução.** Ed. Ribeirão Preto: Holos, 2001. 202 p.
9. Nei M, Kumar S. **Molecular evolution and phylogenetics.** 1 Ed. New York: Oxford, 2000. 333 p.
10. Lander ES *et al.* **Initial sequencing and analysis of the human genome.** *Nature* 409:860-921. 2001.
11. Li WH, Graur D. **Fundamentals of molecular evolution.** 2. Ed. Sunderland: Sinauer Associates, 2000.480p.
12. Prosdocimi F *et al.* **Clustering of *Schistosoma mansoni* mRNA sequences and analysis of the most transcribed genes: implications in metabolism and biology of different developmental stages.** *Mem Inst Oswaldo Cruz* 97: 61-69. 2002.
13. Schena M. **Microarray Analysis.** Ed. John Wiley & Sons. 2002.
14. Setubal JC, Meidanis J. **Introduction to Computational Molecular Biology.** Brooks Cole Publishing Company. 1997. 296p.
15. Stein L. **Genome annotation: from sequence to biology.** *Nat Reviews* 2: 493-505. 2001.
16. Strohman R. Five stages of the Human Genome Project. *Nat. Biotechnol* 17, 112. 1999.
17. Schwartz RL. **Learning Perl.** Ed. O'Reilly & Associates, Inc. 1993. 247p.
18. Tisdall JD. **Beginning Perl for Bioinformatics.** Ed. O'Reilly & Associates, Inc. 2001. 368p.
19. Venter JC *et al.* **The sequence of the human genome.** *Science* 29:1304-51. 2001.