Glossário de Termos em Bioinformática

Uma grande dificuldade que se apresenta ao iniciante em Bioinformática é a imensa quantidade de jargões, cujo conhecimento se torna necessário para execução dos principais recursos que se tenciona usar. Aqui apresentaremos algumas definições para termos úteis nesse processo de familiarização com o mundo da Bioinformática e da Biologia Computacional.

Ab initio – é um termo em latim que significa “desde o começo” ou a partir do zero, assim como os cálculos orbitais moleculares, o quais usam todos os orbitais moleculares em um cálculo, não só a valência dos orbitais dos elétrons.

Accession number (Número de acesso/identificador) – um identificador ou registro único dado a uma sequência submetida a um dos repositórios mundiais de DNA (GenBank, EMBL, DDBJ). O depósito inicial de um registro de sequência é referido como *version 1*. Se a sequência for atualizada, o número da versão é incrementada, mas o número de acesso permanece constante.

Alelo – uma das formas variáveis de um gene em um lócus particular de um cromossomo. Alelos diferentes produzem variação em características herdadas, tais como cor do cabelo ou tipo sanguíneo. Num dado indivíduo, uma forma do alelo (o dominante) pode ser expresso mais que a outra forma (a recessiva).

Alinhamento – o processo de alinhar/comparar duas ou mais sequências para alcançar os níveis máximos de identidade (e de conservação, no caso de sequências de aminoácidos) objetivando atribuir o grau de similaridade e a possibilidade de homologia.

Alinhamento Global – o alinhamento de duas sequências em toda a sua inteira extensão.

Alinhamento Local – o alinhamento em alguma porção de duas sequências (de ácidos nucléicos ou aminoácidos).

Alinhamento Múltiplo de Sequências (*Multiple Sequence Alignment*, MSA) – Um alinhamento de três ou mais sequências com *gaps* inseridos de forma que os resíduos contendo posições estruturais comuns e/ou resíduos

ancestrais sejam alinhados na mesma coluna. Clustal W é o programas de MSA mais usado e conhecido, embora outros programas mais recentes (ate mais rápidos e eficientes) já estejam disponíveis como o MUSCLE, o T-Coffee, o MAFFT e o ProbCons.

Algoritmo – um procedimento fixo, predefinido, incorporado num programa computacional. Um conjunto finito de instruções passo-a-passo para um procedimento computacional ou solução de problemas, especial-, mas não necessariamente, um procedimento que possa ser implementado por um computador.

Anotação – uma combinação de comentários, notações, referências, e citações, ou em formato livre ou utilizando um vocabulário controlado, que juntos descrevem todas as informações experimentais e inferidas sobre um gene ou proteína. Anotações também podem ser aplicadas para descrição de outros sistemas biológicos. O *batch* (lote) é igual a anotações automáticas sobre um pacote ou *bulk* de sequências biológicas.

Anticódon – É um triplet de bases contíguas no tRNA que se liga a sequência de códon de nucleotídeos no RNAm. Exemplo: código GGG para Glicina.

API – *Application Programming Interface.* Uma API é um conjunto de passos/atividades que uma aplicação usa para requerer, carregar e executar tarefas de baixo nível (mais próximo da linguagem de máquina) realizadas pelo sistema operacional (SO). Para computadores executando uma interface gráfica de usuário (*graphical user interface* ou GUI), uma API gerência as janelas da aplicação, ícones, menus, e caixas de diálogo.

Arquivo flat – *flat file* – Um arquivo de dados que contém registros (cada um correspondendo a uma linha em uma tabela); esses registros, entretanto, não possuem relações estruturadas e, para interpretar esses arquivos, as propriedades do formato do arquivo precisam ser conhecidas.

Árvore filogenética – uma variedade de dendrograma (diagrama) no qual os organismos são mostrados em ramos que os ligam de acordo com sua descendência e relação evolutiva.

ASN.1 – *Abstract Syntax Notation* 1 é um formato padrão internacional de representação de dados usado para alcançar interoperabilidade entre plataformas computacionais. Ele permite o intercâmbio confiável de dados (em termos de estrutura e conteúdo) por sistemas computacionais e softwares de todos os tipos.

BAC – *Bacterial Artificial Chromosome*. Um BAC é um grande segmento de DNA (100,000 -200,000 pares de bases, pb ou *bp*) de uma outra espécie clonada em bactérias. Uma vez que o DNA estranho (*foreign DNA*) tenha sido clonado em uma bactéria hospedeira, muitas cópias dele poderão ser feitas.

Bacteriófago – É um vírus que infecta bactérias. O DNA bacteriófago tem servido como uma base para a clonagem de vetores, e é também utilizado para criar bibliotecas de fagos contendo genes humanos ou outros genes.

Bancos de Dados (BDs) – *databases* – coleção lógica e coerente de dados relacionados e com um significado inerente; um local de armazenamento de dados relacionados representantes de um ou vários domínios do conhecimento; os dados são fatos que podem ser gravados e que possuem um significado implícito.

BankIt - é uma ferramenta para submissão on-line de uma ou algumas sequências no GenBank; ela se destina a tornar o processo de submissão rápido e fácil (*BankIt* automaticamente usa o VecScreen para identificar segmentos de ácidos nucléicos que sejam de origem do vetor, ou adaptador, para evitar o problema de contaminação por vetores de clonagem no GenBank).

Biblioteca de cDNA – coleção de clones de cDNA recombinante.

Biblioteca Genômica – coleção de fragmentos de DNA de uma determinada espécie de organismo, obtidos a partir da ação de endonucleases de restrição.

Biblioteca de Subtração – uma biblioteca de cDNA que somente contém cDNAs exclusivamente expressos em uma dada célula ou tecido, p.ex., células T e células B expressarão muitos RNAs comuns, assim como uma pequena porcentagem, a qual será exclusiva para células T e B, respectivamente. Para fazer uma biblioteca de subtração de células T, o cDNA de uma biblioteca de células T é hibridizado com bastante RNA de células B. Os genes expressos comumente resultarão em híbridos RNA-cDNA, os quais podem ser removidos (ou subtraídos) para deixar somente cDNAs específicos de células T.

Biblioteca Virtual – a criação e estocagem de uma vasta coleção de estruturas moleculares em um BDs eletrônico. Esses BDs podem ser consultados para pesquisa por subconjuntos de dados que exibam características físico-químicas específicas, ou possam ser “virtualmente rastreados” para encontrar potencialidades de se ligar a drogas-alvo.

Bioinformata X Bioinformaticista – uma distinção tem sido feita entre bioinformaticistas e bioinformatas. Um bioinformaticista é um expert que não somente sabe como usar as ferramentas de bioinformática, mas também sabe como desenvolver e escrever interfaces para uso efetivo das ferramentas. Um bioinformata, por outro lado, é um indivíduo treinado que sabe usar ferramentas de bioinformática sem entendimento muito profundo. Assim, o bioinformaticista está para as ciências genômicas como um engenheiro mecânico estaria para um automóvel, enquanto o bioinformata está para as ciências genômicas como um mecânico estaria para um automóvel.

Bioinformática – a fusão da biotecnologia e da tecnologia da informação (TI) com o objetivo de revelar novas hipóteses e princípios em biologia.

Biologia de Sistemas – é o estudo coordenado dos sistemas biológicos pela investigação de componentes de redes celulares e suas interações, por aplicação de técnicas experimentais pan-genômicas e em larga escala (*high-throughput*), e integrando métodos de modelagem e automatização computacionais.

BLAST - *Basic Local Alignment Search Tool* (Altschul et al., J Mol Biol 215:403-410; 1990) é um algoritmo de comparação de sequências otimizado para rapidez; ele é usado para busca em bancos de dados de sequências a fim de gerar alinhamentos locais a partir de uma seqüência-consulta (uma *query-sequence*). É a ferramenta mais conhecida, usada e difundida da Bioinformática.

BLASTn – tipo de BLAST que pega sequências de nucleotídeos em FASTA e as compara contra BDs de nucleotídeos.

BLASTp - Tipo de BLAST que pega sequências de proteínas (aminoácidos) em FASTA e as compara contra BDs de proteínas.

BLAT – um programa de análise de sequências (de DNA/Proteína) para identificação rápida de sequências com mais de 95% de similaridade em mais de 25 bases de comprimento. O BLAT (http://genome.ucsc.edu/cgi-bin/hgBlat?command=start) é diferente do BLAST, embora sejam ambos ferramentas de alinhamento.

BLOSUM 62 - *Blocks Substitution Matrix*. Uma matriz de substituição na qual os escores para cada posição são derivados de observações de frequências de substituições em blocos de alinhamentos locais em proteínas relacionadas. Cada matriz é adaptada/modificada para uma dada distância evolutiva. Na matriz BLOSUM 62, por exemplo, o alinhamento do qual os escores derivaram foi criado usando sequências que compartilham não mais que 62% de identidade. Sequências com identidades maiores de 62% são representadas por uma única sequência no alinhamento para evitar superestimar membros muito próximos de famílias de proteínas (Henikoff & Henikoff, Proc Natl Acad Sci U S A 89:10915-10919; 1992).

Boleano (*Boolean)* – termo que se refere a álgebra binária que usa os operadores lógicos AND, OR, XOR, e NOT; as saídas consistem de valores lógicos (ou TRUE ou FALSE). A palavra-chave *boolean* indica que a expressão associada com o identificador assume o valor TRUE ou FALSE. O operador lógico AND (&&) produz o valor 1 se ambos operandos possuem valores diferentes de zero; caso contrário, ele produz o valor 0. O operador lógico OR produz o valor 1 se qualquer de seus operandos possui um valor diferente de zero. O operador lógico NOT (!)produz o valor 0 se seu operando for verdadeiro (TRUE, *nonzero*) e o valor 1 se seu operando for falso (FALSE, 0). O operador OU exclusivo (XOR) será TRUE somente se um de seus

operandos for TRUE e o outro for FALSE. Se ambos operandos forem iguais (seja TRUE ou FALSE), a operação será FALSE.

Cadeia de Markov – qualquer densidade de probabilidade multivariada cujo diagrama de independência seja uma cadeia. Essas variáveis são ordenadas, e cada variável "depende" somente de seus vizinhos no sentido de serem condicionalmente independentes dos outros. As cadeias de Markov são um componente integral dos modelos "escondidos" (*hidden Markov models)*.

Domínio Conservado (CD - *Conserved Domain*) - CD refere-se a um domínio (uma unidade distinta estrutural e/ou funcional de uma proteína) que seja conservado durante a evolução. Evolutivamente, alterações em posições específicas de uma sequência de aminoácidos na proteína ocorrem de maneira que propriedades físico-químicas dos resíduos originais sejam preservadas, preservando também as propriedades estruturais e/ou funcionais daquela região da proteína.

CDART - *Conserved Domain Architecture Retrieval Tool*. Uma ferramenta do NCBI. Considerando uma sequência de proteína sendo consultada (*query sequence*), o CDART apresenta os domínios funcionais que formam a proteína e lista todas as proteínas com arquiteturas similares de domínios. Os domínios funcionais para uma sequência são encontrados pela comparação da sequência da proteína com um BDs de alinhamentos de domínios conservados, o CDD do NCBI.

CDD - *Conserved Domain Database*. Um BDs de alinhamentos de domínios conservados, um recurso muito válido do NCBI, que consiste numa coleção de perfis de alinhamentos representativos de domínios de proteína conservados durante a evolução.

cDNA – DNA complementar. Uma sequência de DNA obtida por transcrição reversa de uma sequência de RNA mensageiro (mRNA).

CDS - *coding region*, *coding sequence*. CDS se refere a porção de uma sequência de DNA genômico que é traduzida, desde o códon iniciador até o códon finalizador, inclusive, se for a CDS completa. Uma CDS parcial carece de parte da CDS completa (ela pode não ter ou o iniciador ou o finalizador, ou ambos). A tradução bem sucedida de uma CDS resulta na síntese de uma proteína.

CGI - *Common Gateway Interface*. Um mecanismo que permite que um servidor Web rode um programa ou script num servidor e envie a saída (o output) para um *browser*.

Camundongo *Knockout* (gene alvo) – camundongo produzido com um determinado gene ausente.

Ciência da Computação – o estudo sistemático dos sistemas computacionais e da computação. O corpo do conhecimento resultante desta disciplina/área contém teorias para o entendimento dos métodos e sistemas da computação;

o design metodológico, algoritmos e ferramentas; métodos para testes de conceitos; métodos para análises e verificação; e representação do conhecimento e implementação. Ou resumidamente, o estudo da implementação, organização, e aplicação de *softwares* para computadores e recursos de *hardware*.

Ciência da Informação – ciência pura e aplicada envolvendo a coleção, organização e manutenção da informação.

Clone – uma população de células geneticamente idênticas ou moléculas de DNA.

Clonagem – a formação/produção de clones ou réplicas genéticas exatas.

Cluster – um grupo criado com base em certos critérios. P. ex., o cluster de um gene pode incluir um conjunto de genes cujos perfis de expressão similar sejam similares de acordo com certos critérios, ou um cluster pode se referir a um grupo de clones que são relacionados uns com os outros por homologia. O termo "*clusterização*" se refere ao procedimento de agrupar elementos similares e, em genômica, é muito utilizado nas fases de fechamento de *gaps* e finalização de genomas recém-sequenciados.

Cn3D – uma ferramenta de visualização de alinhamento de sequências e de estruturas 3-D para os bancos de dados do NCBI. O Cn3D pode funcionar como uma aplicação auxiliar ao browser ou como um cliente (www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml).

Código Genético – as instruções contidas em um gene que determinam ou orientam, num contexto celular, a formação de uma proteína em particular. A, T, G, e C são as letras do código do DNA, respectivamente correspondem as bases nitrogenadas adenina, timina, guanina, e citosina, que formam os nucleotídeos ou resíduos de uma sequência de DNA. O código de cada gene combina os quatro nucleotídeos de várias maneiras (especificamente 64 maneiras) para representar trincas de nucleotídeos (os códons) que especifiquem qual aminoácido deve ocorrer em cada posição para montar uma dada proteína.

Códon – uma sequência de três nucleotídeos no DNA ou mRNA que específica um aminoácido particular durante a síntese protéica; também chamado de trinca (*triplet*). Dos 64 códons possíveis, 3 são finalizadores (ou *stop codons*) e não codificam aminoácidos.

Códon Iniciador (*translation start site*) – a posição dentro de um mRNA no qual a síntese de uma proteína se inicia. O sitio de inicio da tradução geralmente é um códon AUG, mas, ocasionalmente, os códons GUG ou CUG são usados para iniciar a síntese protéica.

Códon Finalizador (*stop códon*) – um dos três códons no RNAm que não especificam nenhum aminoácido e, assim, provocam o fim da tradução do

mRNA em proteína. Eles marcam o fim da sequência codificadora de proteína.

COGs - *Clusters of Orthologous Groups* (de genes/proteínas) - foram designados para comparar sequências de proteínas de genomas completos/finalizados. Cada COG consiste de proteínas individuais ou grupos de parálogos de pelo menos três linhagens e, assim, correspondem a um domínio ancestral conservado.

Configuração – (em *software*) - a ordenação e descrição completa de todas as partes de um software ou sistema de BDs. A manutenção da configuração é o uso de softwares para identificar, fazer levantamentos (inventórios) e manter os componentes dos módulos que juntos compreendem um ou mais sistemas ou produtos.

Conformação – é o arranjo tri-dimensional preciso de átomos e de ligações em uma molécula, descrevendo sua geometria e, por consequência, sua função molecular.

Consenso (*consensus sequence*) – Os nucleotídeos ou aminoácidos mais comumente observados/encontrados em cada posição nas sequências de DNAs, RNAs, ou proteínas homólogas.

Consenso possível (*tentative consensus* ou TC) – a identificação de uma sequência de um *cluster* de EST que representa o gene completo ou parte dele. Os TCs são geralmente determinados por clusterização de ESTs permitindo, por erros sequenciais, artefatos tais como clones quiméricos, e fenômenos biológicos naturalmente ocorrentes tais como o splicing alternativo. A criação de um cluster permite a geração de uma sequência consenso e, então, a identificação de uma ORF longa (uma matriz aberta de leitura), a qual sugeriria a possibilidade de que o consenso representa realmente um gene (um gene *bona fide*).

Constitutivo(a) – Síntese ou Expressão Constitutiva - síntese de RNAm e de proteína a uma taxa constante (ou pouco variável) independente das exigências da célula (veja genes estruturais ou *housekeeping*).

Contig – Um segmento contíguo do genoma obtido pela junção de sequências (previamente clonadas) que se sobrepõem (através de porções comuns). Um clone contig consiste de um grupo de pedaços copiados (clonados) de DNA representando regiões sobrepostas de um cromossomo. Uma sequência *contig* é uma sequência estendida, criada pela fusão de sequências primárias que se sobrepõem. Um mapa de contigs mostra as regiões de um cromossomo onde os segmentos contíguos de DNA se sobrepõem. Os mapas de contigs possibilitam o estudo de genomas pelo avaliação de uma série de clones sobrepostos, os quais demonstram a sucessão ininterrupta de resíduos naquela região clusterizada.

Cosmídio – vetores que permitem a inserção de fragmentos longos de DNA (acima de 50 kb).

CPU - *Central Processing Unit* - A unidade de controle e processamento do computador; o dispositivo que interpreta e executa as instruções do sistema operacional.

Data mining (mineração de dados) – a habilidade de consultar/pesquisar grandes BDs no intuito de satisfazer uma hipótese (*"top-down" data mining* = mineração de cima-para--baixo); ou de examinar ("interrogar") BDs no intiuito de gerar novas hipóteses baseadas em correlações estatísticas rigorosas (*"bottom-up" data mining* = mineração de baixo-para-cima).

Data warehouse – uma coleção de dados adquiridos e organizados, de modo que possam ser facilmente analisados, extraídos, sintetizados e mesmo usados para o propósito de entender posteriormente os dados. Isto deve ser contrastado com dados adquiridos para suprir objetivos imediatos tais como, p.ex., transações de ordem de pagamento numa empresa. Grandes coleções de dados heterogêneos (biológicos), armazenados dentro de um único repositório de dados lógico, que são acessíveis para diferentes consultas e métodos de manipulação.

DDBJ – DNA Banco de Dados do Japão - DNA Data Bank of Japan http://www.ddbj.nig.ac.jpl

Deconvolução – procedimento matemático que separa os efeitos de sobreposição de moléculas, tais como mistura de compostos numa solução rica ou misturas de cDNAs em altas densidades.

Dendrograma – é um tipo específico de diagrama ou representação icônica que organiza determinados fatores e variáveis. Um dendrograma é um diagrama tipo-árvore que sumariza o processo de clustering. Casos similares são agrupados por associações cuja posição no diagrama é determinada pelo nível de similaridade entre os casos (eventos ou elementos). Dendro = árvore.

Dímero – uma molécula composta pela ligação de duas moléculas (veja homo e heterodímeros).

Dinâmica Molecular – o estudo das conformações intramoleculares e movimentos moleculares, usando simulações computacionais. Cálculos simulando o movimento de cada átomo num sistema molecular em uma energia e temperatura constante, ou com mudanças controladas de temperatura.

DNA Genômico (sequência de) – inclui sequências de íntrons e éxons (sequências codificantes), assim como sequências regulatórias não-codificantes tais como promotores. Diz-se DNA genômico para diferenciar, p.ex., do DNA mitocondrial.

Domínio – Um domínio se refere a exata porção/trecho de uma proteína que se supõe enovelar/dobrar independentemente do resto da proteína e que deve possuir função própria.

Duplo Híbrido (*yeast two-hybrid*) – um método baseado em levedura que é usado para simultaneamente identificar e clonar o gene codificador de proteínas que interagem com uma proteína conhecida. A base deste método é um "ensaio repórter (ou de registro) transcricional" no qual a expressão do gene repórter é dependente de dois domínios. O primeiro domínio se liga aa proteína conhecida, enquanto o segundo domínio se liga geneticamente a uma biblioteca. Se a biblioteca é rastreada ou vasculhada para a proteína conhecida, os dois domínios irão interagir somente se a proteína da biblioteca se ligar à proteína conhecida, resultando na ativação da transcrição do gene repórter e numa cor azul. O "clone de levedura azul" conterá o gene que codifica a proteína recém identificada.

Elementos repetitivos - fornecem importantes indícios sobre a dinâmica de cromossomos, forças evolucionárias, e mecanismos para mudança de informação genética entre organismos.

Enzima – uma classe de proteínas que são capazes de catalizar reações químicas (a formação ou dissolução de ligações químicas). Eles fazem isso pela condução de seus substratos para uma geometria adequada em uma localização particular (o sitio ativo) onde resíduos de aminoácidos eletrofílicos ou nucleofílicos podem participar da reação. Enzimas são proteínas catalíticas que acelaram reações químicas que de outro modo seriam mais lentas sobre condições fisiológicas.

Enzimas de restrição – endonucleases de restrição – um tipo de enzima que reconhece de forma específica as sequências de DNA (usualmente sequências palindrômicas de 4, 6, 8 ou 16 pares de bases de comprimento) e realiza cortes em ambas as fitas de DNA que contendo aquelas sequências. São chamadas de "tesouras moleculares" da tecnologia do DNA recombinante.

Epigenômica – o estudo de redes complexas de expressão ou de ligações físicas e temporais (nas diferentes fases ou locais de desenvolvimento).

E-value - *Expect value* – Valor esperado – Um parâmetro que descreve o numero de *hits* que se espera encontrar ao acaso quando buscando um BDs de determinado tamanho. Ele diminui exponencialmente com o escore (S) que é atribuído a uma combinação ou acasalamento (um *match*) entre duas sequências. Em suma, o *E-value* descreve a aleatoriedade implícita (ou o *random background noise*) que existe para os *matches* entre as sequências. Então, um *E-value* de 1 atribuído a um *hit* significaria que em um BDs daquele tamanho exato se esperaria apenas um *match* com um escore similar ao acaso. Isso significa que quanto menor o *E-value*, ou quão mais próximo ele é de 0, maior a significância do *match*. Entretanto, vale ressaltar que buscas com sequências muito curtas podem ser virtualmente idênticas e possuírem *E-values* relativamente altos. Isso se deve ao fato de que o calculo

do *E-value* também considera a extensão (ou o comprimento) da sequência de consulta (*query sequence*), já que sequências mais curtas tem uma alta probabilidade de ocorrência meramente ao acaso no banco de dados. Para detalhes, veja o tutorial: www.ncbi.nih.gov/BLAST/tutorial/Altschul-1.html.

Entrez – O sistema de recuperação (*retrieval*) de buscas que integra as consultas em vários BDs ligados no NCBI, tais como PubMed, GenBank, Protein, Structure, Genome, PopSet, OMIM, Taxonomy, Books, ProbeSet, 3D Domains, UniSTS, SNP, e CDD. Também chamado de *All Databases* que significa "Todos os Bancos de Dados". Acesse em http://www.ncbi.nlm.nih.gov/Entrez/.

Especiação – o processo pelo qual uma ou mais populações de um espécie se torna geneticamente diferente o suficiente para formar uma nova espécie. O processo geralmente requer que as populações estejam isoladas por um longo período de tempo.

EST - *Expressed Sequence Tag*. As etiquetas de sequências expressas (ESTs) são sequências curtas (usualmente de cerca de 300-500 pares de bases), oriundas de leitura em sequenciadores de clones de bibliotecas de cDNA. Em geral, as ESTs representam os genes expressos em um dado tecido e/ou estágio de desenvolvimento. Elas são como rótulos ou etiquetas de expressão (algumas codificantes, outras não) para uma determinada biblioteca de cDNA.

Estrutura (de proteínas) – Estrutura Primária (de um sequência de aminoácidos) numa proteína – a sequência linear de um polipeptídeo ou proteína. Estrutura Secundária – a organização do arcabouço peptídico de uma proteína que ocorre como resultado das pontes ou ligações de hidrogênio, por exemplo, uma alfa hélice e uma folha beta pregueada. Estrutura Terciária – o enovelamento (*folding*) de uma cadeia de proteína via interações de suas moléculas de cadeias laterais, incluindo a formação de pontes ou ligações dissulfeto entre resíduos de cisteína.

Éxon – o trecho de um gene que codifica uma parte do RNAm daquele gene. Um gene pode conter muitos éxons, alguns dos quais podem incluir somente sequências codificantes de proteínas; no entanto, um éxon pode também incluir sequências não-traduzidas 5' or 3'. Cada éxon codifica uma porção específica da proteína completa. Em algumas espécies de eucariotos (os humanos, inclusive), os éxons de um gene podem estar separados por extensão regiões de DNA (os íntrons), que aparentemente não possuem função conhecida, embora se saiba que alguns íntrons podem codificar pequenos RNAs não-traduzidos.

Farmacogenômica – capítulo da ciência genômica que estuda a correlação entre o perfil genético dos pacientes (seus genótipos) e as respostas individuais ao tratamento com os diferentes fármacos.

FASTA – O primeiro algoritmo amplamente utilizado para busca de similaridade de sequências em bancos de dados de DNA e de proteínas. O programa executa alinhamentos locais ideais (ou ótimos) através da procura de pequenos *matches*, chamados de *words*. O termo FASTA também se refere ao formato de arquivo padrão de sequências de DNA ou aminoácidos (http://www.ncbi.nlm.nih.gov/BLAST/fasta.html).

Fator(es) de Transcrição – um grupo de proteínas reguladoras que são requeridas para a transcrição em eucariotos. Os fatores de transcrição se ligam a região promotora de um gene e facilitam a transcrição pela RNA polimerase.

Fenótipo – o conjunto de traços ou características observáveis (manifestadas) de um organismo, tais como cor do cabelo, peso, ou a presença ou ausência de uma doença. Traços fenotipicos não são necessariamente genéticos (ou seja, herdados).

Filogenia – relações evolutivas dentro e entre níveis taxonômicos, particularmente os padrões de linhas da descendência. Filogenética – a classificação taxonômica de organismos baseada no seu grau de relação (ou parentesco) evolutivo; árvore filogenética – uma variedade de dendrograma (diagrama) no qual os organismos são mostrados em ramos que os ligam de acordo com sua descendência e relação evolutiva.

Frameshift – um deleção, substituição, ou duplicação de uma ou mais bases que causam mudança na matriz (ou fase) de leitura de um gene estrutural alterando a série normal de trincas (*triplets*).

FTP - *File Transfer Protocol*. – Protocolo de Transferência de Arquivo - um método de recuperar/receber arquivos numa rede (interna ou externa) de computadores diretamente da fonte para o computador do usuário usando um conjunto de protocolos que regulam como os dados serão transportados. Geralmente é usado para transferência de grande volume de dados que não seriam facilmente enviados por outros meios como e-mail ou download por internet.

Gap - Um espaço introduzido em um alinhamento para compensar inserções e deleções feitas em uma sequência em relação a outra(s). Para evitar o acumulo excessivo de gaps em um alinhamento, a introdução de um gap provoca a dedução de uma quantidade fixa (o escore do gap) no escore do alinhamento. A extensão do gap (seu prolongamento) para comportar resíduos (nucleotídeos ou aminoácidos) adicionais também é penalizado no escore de um alinhamento. Maiores detalhes podem ser vistos em:
<http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/Alignment_Scores2.html>

GenBank – um BDs de sequências de nucleotídeos de mais de 100,000 organismos. Os registros que são anotados com características de regiões codificantes também incluem traduções de aminoácidos. O GenBank faz parte de uma colaboração internacional de BDs de sequências que também

inclui o EMBL e o DDBJ, sendo um dos vários BDs que compõem o NCBI (http://www.ncbi.nlm.nih.gov/Genbank/submit.html)

Genoma – o conteúdo genético completo de um organismo ou o conjunto total de genes e material extragênico.

Genômica – a análise do genoma inteiro de um organismo de escolha; uma tentativa de analisar ou comparar o repertório genético total de uma espécie ou espécies. Também é possível analisar genomas pela comparação de subconjuntos mais ou menos representativos de genes dentro dos genomas.

Genômica Estrutural ou Bioinformática Estrutural – referente as análises de estrutura macromolecular, particularmente de proteínas, usando ferramentas computacionais e plataformas teóricas. Um dos objetivos da genômica estrutural é obter modelos estruturais tridimensionais acurados para todas as famílias de proteínas conhecidas, domínios de proteínas e enovelamento de proteínas. Alinhamento estrutural é uma ferramenta da genômica estrutural.

Genótipo – A identidade genética de um indivíduo que não se manifesta como características externas (fenotípicas). O genótipo se refere ao par de alelos para uma dada região do genoma (um gene ou conjunto de genes) que um indivíduo porta.

Glicosilação – a adição de grupamentos de carboidratos (açúcares) a, p.ex., cadeias de polipeptídeos.

Grupo Carboxila – grupamento funcional -COOH, naturalmente ácido, encontrado em todos os aminoácidos.

GSS - *Genome Survey Sequences* – são sequências análogas as ESTs, com exceção de que sua origem é genômica (DNA) e não de transcritos (cDNA oriundo de mRNA).

Hairpin (Grampo) – uma região de dupla hélice em uma fita simples de DNA ou de RNA formado pela ligação de hidrogênio entre as sequências complementares adjacentes que formam uma estrutura enroscada / grampeada ou um grampo.

Haplóide – uma célula ou organismo contendo somente um conjunto de cromossomos sem os pares homólogos (diplóides).

Hibridização – a interação de fitas complementares de ácidos nucléicos. Isto pode ocorrer entre duas fitas de DNA ou entre fitas de DNA e RNA, sendo a base de muitas técnicas tais como a *Southern blot*.

Hidden Markov *model* (HMM) – um modelo estatístico de uma sequência ordenada de variáveis (ver também Cadeia de Markov). O resultado de

estocasticamente perturbar as variáveis em uma cadeia de Markov (as variáveis originais estão assim escondidas ou “hidden”), onde a cadeia de Markov possui variáveis discretas as quais selecionam o “estado” do HMM a cada etapa. Os valores perturbados podem ser contínuos e são as saídas (os “outputs”) do HMM. Um HMM é equivalente a um modelo de mistura acoplada em que a distribuição conjunta sobre os estados é uma cadeia de Markov. Os HMMs são muito valiosos na bioinformática porque eles possibilitam que um algoritmo de busca ou de alinhamento seja treinado usando sequências de entrada não-alinhadas ou não-ponderadas; também permitem parâmetros de escore dependentes de posição tais como penalidades por gap, o que pode modelar muito mais precisamente as consequências de eventos evolutivos em famílias de sequências.

Homeobox – uma região altamente conservada em um gene homeótico composto de 180 bases (60 aminoácidos) que especifica um domínio protéico (o homeodomínio) que serve como elemento mestre de regulação gênico na diferenciação celular durante o desenvolvimento em espécies tão divergentes quanto vermes, moscas e humanos.

Homologia – (conceito estrito) - duas ou mais espécies, sistemas ou moléculas biológicas que dividem um ancestral evolutivo comum. (conceito geral) – duas ou mais sequências de gene ou proteína que compartilham um significativo grau de similaridade, geralmente medido pela quantidade de identidade (no caso do DNA) ou de substituições conservativas (no caso da proteína), que elas registram ao longo de suas extensões. Buscas por homologia em sequências são realizadas com uma sequência de DNA ou proteína alvo de consulta (query) para identificar genes ou produtos gênicos conhecidos que compartilhem similaridade significativa e daí possam informar algo sobre a ancestralidade, herança e possível função do gene procurado (consultado).

Homólogo – O termo se refere a similaridade atribuível a um descendente a partir de um ancestral/ascendente comum. Cromossomos homólogos são membros de um par de cromossomos essencialmente idênticos, cada um originário de um dos dois ascendentes (genitores); eles possuem os mesmos (ou alélicos) genes com *loci* genéticos dispostos na mesma ordem.

HTML - *Hypertext Markup Language*. A linguagem HTML (derivada da SGML) é baseada em texto e destinada a sua marcação; é usada primordialmente para disponibilizar informações usando um web browser e para associar partes informação via hiperlinks. As *tags* ou etiquetas usadas num documento HTML fornecem informação somente sobre como o conteúdo deve ser apresentado, mas não fornece informação sobre o conteúdo em si.

Identidade – O grau ou porcentagem o quanto duas sequências (de nucleotídeos ou aminoácidos) são invariáveis em seus resíduos.

In silico (biologia) – é o uso de computadores para simular, processar, ou analisar de um experimento biológico.

Íntron - trecho de uma sequência de DNA que está presente no transcrito primário; será removido por *splicing* durante o processamento do RNA e não estará incluído na sequência final, madura e funcional do mRNA, rRNA, ou tRNA. Também chamada de sequência interveniente.

ISSN - *International Standard Serial Number*, um número de oito dígitos que identifica publicações periódicas, revistas e jornais, incluindo as versões impressas e eletrônicas.

Iteração – uma série de passos em um algoritmo por meio do qual o processamento de dados é executado repetitivamente até o resultado exceder um limite particular (limiar ou *threshold*). A iteração geralmente é usada nos alinhamentos múltiplos de sequências em que cada conjunto de alinhamentos pareados é comparado um com todos os outros, começando com os pares mais similares e progredindo até os menos similares, até que não restem mais quaisquer pares de sequências remanescentes a serem alinhadas.

Knockout (camundongo) – um animal desenvolvido para ter um dado gene ausente; o gene é inativado nas chamadas células tronco embrionárias usando a técnica de recombinação homóloga. Tais células são então reintroduzidas num embrião de estágio inicial (um blastocisto) e este é então transplantado num animal recipiente. A progênie subsequente não vai possuir o gene alvo em algumas células. Esta técnica é usada para determinar a função de um gene alvo.

"Lab on a chip" – termo descrito para microdispositivos que permitem análises rápidas e microanalíticas de DNA ou de proteínas em um único sistema, miniaturizado e completamente integrado.

Léxico – (estrito) na bioinformática, um léxico se refere a uma lista pré-definida de termos que juntos definem completamente o conteúdo de um BDs particular. (geral) O componente na gramática que é uma lista de palavras ou entradas léxicas.

Ligante – qualquer molécula pequena que se liga a uma proteína ou receptor; o complemento cognato de muitas proteínas, enzimas e receptores celulares.

Linkage (Ligação) – a associação/ligação de genes (ou locos genéticos) no mesmo cromossomo. Genes que se ligam tendem a ser transmitidos juntos.

Locus (ou Loco) – num contexto genômico, refere-se a posição em um cromossomo. Pode, contudo, referir-se a um marcador, um gene, ou qualquer outro elemento molecular que seja descrito assim.

Mapa de Restrição – um mapa físico ou descrição de um gene (ou genoma) derivado do ordenamento de fragmentos de restrição sobrepostos produzidos pela digestão de DNA com um número de enzimas de restrição.

Matriz de Peso – (*Weight matrix*) – a densidade dos sítios de ligação em um gene ou sequência que pode ser usada para obter uma taxa de densidade para cada elemento em um padrão de interesse. As taxas combinadas de densidade individuais de todos os elementos são então coletivamente usadas para construir um perfil de escore conhecido como matriz de peso. Esse perfil pode ser usado para testar a predição da identificação do padrão selecionado e da habilidade do algoritmo em discriminar as sequências-padrão das não-padronizadas.

Matriz de Substituição – uma matriz de substituição contendo valores proporcionais a probabilidade de que o aminoácido *a* modifique-se em aminoácido *b* para todos os pares de aminoácidos. Tais matrizes são construídas pela montagem de uma amostragem ampla e diversa de alinhamentos pareados verificados de aminoácidos. Se a amostragem for grande o suficiente para ser estatisticamente significativa, as matrizes resultantes devem refletir as probabilidades verdadeiras de mutações que ocorrem durante a evolução.

Matriz PAM - *Percent Accepted Mutation* – Uma unidade introduzida por Dayhoff et al. para quantificar o montante de alteração evolutiva em uma sequência de proteína. Uma unidade PAM 1.0 é a quantidade de evolução que irá modificar, em média, 1% dos aminoácidos em uma sequência protéica. Uma matriz de substituição PAM(x) é uma tabela de referência na qual os escores para cada substituição de aminoácido foram calculados com base na frequência daquela substituição em proteínas intimamente relacionadas (próximas) que sofreram uma certa quantidade (x) de divergência evolutiva.

MEDLINE – BDs (da National Library of Medicine os EUA) de citações de periódicos indexados e resumos na área de biomedicina e saúde; compreende aproximadamente 5,000 periódicos publicados nos EUA e em outros 70 países.

MegaBLAST – programa para alinhar sequências que diferem pouco entre si ou para manipular eficientemente sequências muito longas de DNA (como os genomas inteiros, p.ex.); mais adequado do que o blastn tradicional nesses dois casos. Dependendo da situação, o MegaBLAST pode ser ate 10 vezes mais rápido do que os programas comuns de busca de similaridade, ao usar um algoritmo GREEDY.

Metabolômica – é a computação de sistemas biológicos emergentes de interesse tais como o desenvolvimento, relógios biológicos e modelos de inferência cinética de DNA, RNA e proteínas.

Metilação – a adição de grupos –CH3 (metil) a um sitio alvo. Geralmente tal adição ocorre nas bases citosina do DNA.

Microarranjos (*Microarrays*) – um arranjo bidimensional (2D), tipicamente feito sob uma superfície fina de vidro, filtro ou silicone, sobre os quais os genes ou fragmentos de genes são depositados ou sintetizados em uma ordem espacial predeterminada, permitindo que sejam feitas sondas em larga-escala e de forma paralela maciça.

Microfluidics – miniaturização de reações químicas ou ensaios farmacológicas dentro de tubos ou recipientes microscópicos no sentido de aumentar sua performance pela disposição de muitas amostras lado-a-lado num arranjo ou ensaio.

Microssatélites – trechos repetitivos de sequências curtas de DNA usadas como marcadores genéticos moleculares para rastrear a herança de traços em famílias de indivíduos (p.ex, CC(TATATATA)CCCT); também conhecidos como *short tandem repeats* (STRs) ou repetições curtas em *tandem*.

Mimética/Mimetizar – compostos que mimetizam a função de outras moléculas através de seu alto grau de similaridade estrutural, e, consequentemente, de propriedades físicoquimicas similares.

Mineração de dados (data mining) – ver *data mining*. Uma atividade de extração de informação cujo objetivo é descobrir fatos implícitos contidos nos BDs. Usando uma combinação de aprendizado de máquina, análise estatística, técnicas de modelagem e tecnologia de BDs, a mineração de dados encontra padrões e relações sutis nos dados, infere regras que permitem a predição de futuros resultados. Aplicações usuais em bioinformática incluem a descoberta de novas relações entre elementos biológicos aparentemente independentes.

Modelagem / Modelar – (*Modeling*) geralmente se refere a modelagem molecular, um processo pelo qual a arquitetura tridimensional das moléculas biológicas é interpretada (ou predita), visualmente representada e manipulada com o intuito de determinar suas propriedade moleculares (sentido na bioinformática). Uma serie de equações ou procedimentos matemáticos os quais simulam um processo da vida real, dados um conjunto de assertivas, parâmetros limites e condições iniciais (sentido geral).

Modificações – Modificação Pós-transcricional: alterações feitas ao nível do RNA pré-mensageiro, ou antes de deixar o núcleo e tornar-se um mRNA maduro. Modificação Pós-translacional – alterações feitas na proteína após sua síntese no ribossomo. Tais modificações, como adição de carboidrato ou cadeias de ácidos graxos, podem ser criticas para a função da proteína.

Monômero – uma unidade básica de qualquer molécula ou macromolécula biológica, tais como um aminoácido, ácido nucléico, domínio polipeptídico ou proteína.

Motivo (*Motif*) - Uma pequena, curta, região conservada em uma sequência de proteína. Geralmente os motivos protéicos são partes altamente

conservadas dos domínios.

Mutagênico – qualquer agente que pode causar um aumento na taxa de mutações em um organismo.

Mutação – uma alteração herdável no genoma que inclui mudanças genéticas pontuais (ou de um só nucleotídeo) ou grandes alterações, tais como deleções ou rearranjos cromossômicos.

NMR - *Nuclear Magnetic Resonance* – uma técnica de espectroscopia física usada para determinação de estrutura tridimensional de proteínas.

nr-PDB - *non-redundant* Protein Data Bank – BDs não redundante do PDB, (http://www.rcsb.org/pdb/), especializado em estruturas tridimensionais resolvidas de moléculas (em geral, proteínas).

Nucleosídeo – um açúcar com cinco carbonos covalentemente ligados a uma base nitrogenada.

Nucleotídeo – uma unidade de ácido nucléico composta de um açúcar com cinco carbonos unida a um grupo fostato e a uma base nitrogenada.

Objeto-Relacional (OR) – um tipo de BDs que une elementos de linguagens de programação orientada a objetos com a capacidade de BDs, possibilitando armazenamento mais persistente de objetos na linguagem de programação. Os BDs OR ampliam a funcionalidade de C++, Smalltalk, ou Java, p.ex., resultando num alto nível de congruência entre o modelo de dados para a aplicação e para o BDs. Os BDs Objeto-relacional são usados na Bioinformática para mapear objetos moleculares (tais como sequências, estruturas, mapas e *pathways*) até suas representações subjacentes (tipicamente dentro de linhas e colunas das tabelas relacionais). Isso permite ao usuário lidar com objetos biológicos de maneira muito mais intuitiva, como se eles estivessem num laboratório, sem preocupações quanto ao modelo de dados subjacente de sua representação. Como resultado se tem uma melhor interatividade entre os modelos de dados e as aplicações.

Oligonucleotídeo – uma molécula curta consistindo de diversos nucleotídeos (tipicamente entre 10 e 60) ligados covalentemente e unidos por ligações fosfodiéster.

Ontologia – em computação é uma definição explicita e formal para o compartilhamento de um vocabulário comum. Em filosofia é o estudo do ser, e compreende tudo envolvido com os seres enquanto humanos, o processo de se tornar plenamente humano, e as relações entre os níveis de ser e as palavras ontológicas criadas por elas.

Open reading frame (ORF) ou Matriz aberta de leitura – qualquer trecho de DNA que potencialmente codifica uma proteína e começa com o códon de iniciação e termina com o códon de terminação.Nenhum códon de terminação pode estar presente no interior de uma ORF. A identificação de

uma ORF é a primeira indicação de que um segmento de DNA é parte de um gene funcional.

Operador – um segmento de DNA que interage com os produtos de genes regulatórios e facilita a transcrição de um ou mais genes estruturais.

Operon – uma unidade de transcrição consistindo de um ou mais genes estruturais, um operador e um promotor.

Ortólogo – termo que descreve a relação de genes em diferentes espécies que se originam de um mesmo ancestral no último ancestral comum da respectiva espécie, ou seja, os genes ortólogos são evolutivamente contra-partes diretos.

Padrão – padrões moleculares biológicos geralmente ocorrem a um nível de características que formam as sequências de genes ou proteínas. Uma linguagem padrão deve ser definida no sentido de empregar diferentes critérios a diferentes posições de uma sequência. Para se ter uma comparação posição-especifica feita por computador, um algoritmo de *matching* de padrão precisa permitir resíduos alternativos em uma dada posição, repetições de um resíduo, exclusão de resíduos alternativos, o peso e preferencialmente a representação combinatória.

Padrão filético – padrão de presença ou ausência de um cluster de genes ortólogos (COG) em espécies diferentes.

Palíndromo – uma região do DNA com um arranjo simétrico de bases ocorrendo a partir de uma dado ponto em que as bases de ambas as fitas são idênticas (se as fitas forem lidas na mesma direção), p.ex., 5' GAATTC 3' cuja sequência complementar é 3' CTTAAG 5'.

Parálogo – um de um conjunto de genes homólogos que divergiram em consequência de uma duplicação gênica. Os ortólogos retém a mesma função no curso evolutivo, enquanto os parálogos evoluem para novas funções, mesmo se estas estiverem relacionadas com a função original.

Parâmetros – valores estabelecidos pelo usuário, tipicamente determinados experimentalmente, e que regulam os limites de um algoritmo ou programa. Por ex., a seleção dos parâmetros de entrada adequados determina o sucesso de um algoritmo de busca. Alguns dos parâmetros de busca mais comuns em bioinformática incluem a estringência de uma ferramenta de busca de alinhamento, e os pesos (penalidades) fornecidos para os mismatches e gaps.

Pathways (Vias Metabólicas) – a bioinformática se esforça para definir as representações dos tipos de dados biológicos chaves, algoritmos e procedimentos de inferência, incluindo sequências, estruturas, reações e vias bioquímicas. A representação e computação das vias biológicas exige ontologias para representar o conhecimento.

PCR - *Polymerase Chain Reaction* - A técnica de reação em cadeia da polimerase para amplificação de segmentos específicos de DNA em uma mistura complexa, na qual estão presentes também curtos iniciadores (primers) oligonucleotídeos para o segmento de interesse, além de reagentes para síntese de DNA. O PCR se baseia na habilidade do DNA de separar suas duas fitas complementares a altas temperaturas (um processo chamado de desnaturação) e na fusão das duas fitas a uma temperatura baixa ideal (anelamento). A fase de anelamento é seguida de uma etapa de síntese de DNA à temperatura ideal com uma polimerase termoestável. Apos múltiplos ciclos de desnaturação, anelamento e síntese de DNA, a sequência alvo de DNA especificada pelos *primers* é amplificada em milhares de cópias.

PCR *Nested* – é a segunda etapa da amplificação de uma sequência já amplificada por PCR, usando um novo par de *primers* os quais são internos aos *primers* originais. Geralmente feita quando uma reação simples de PCR gera quantidades insuficientes de produto.

Peptídeo – um trecho curto de aminoácidos ligados covalentemente um ao outro por uma ligação peptídica (ou amida) que consiste numa ligação covalente formada entre 2 aminoácidos quando o grupo amino de um está ligado ao grupo carboxílico de outro (resultando na eliminação de uma molécula de água)

Perfil (*profile*) - uma tabela que lista as frequências de cada aminoácido em cada posição de uma sequência de proteína. As frequências são calculadas a partir de alinhamentos múltiplos de sequências (MSAs) contendo um domínio de interesse. Veja também PSSM. Perfil de sequências são usualmente derivados de múltiplos alinhamentos de sequências com relações conhecidas, e consiste de tabelas de posições específicas e penalidades por gaps. Cada posição no perfil contém escores para todos os possíveis aminoácidos, assim como um escore de penalidade para o inicio e um para a continuidade do gap na posição especificada. Tentativas foram feitas para incrementar a sensibilidade do perfil pelo refinamento dos procedimentos para construir um perfil começando de um MSA. Outras representações para domínios de sequências ou motivos não exigem necessariamente a presença de um alinhamento múltiplo correto e completo, tal como nos HMMs.

Pfam – *Protein family* (http://pfam.wustl.edu/index.html) é um BDs que alberga uma grande coleção de alinhamentos múltiplos de sequências e cadeias de Markov (os *hidden Markov models* ou HMM) contemplando muitos domínios comuns de proteínas.

PHRAP – Um programa de computador que faz a montagem (*assembly*) de sequências (ou dados brutos) em grupos ou contigs e atribui a cada posição na sequência um escore associado, com base nos escores PHRED das leituras (reads) de sequências puras. Um escore de qualidade PHRAP de /X/ corresponde a uma probabilidade de erro de aproximadamente 10^-/X//10. Assim, um escore de qualidade PHRAP de 30 corresponde a 99.9% de acurácia para uma base na sequência montada (*assembled*).

PHRED - Um programa de computador que analisa sequências puras para produzir um *base call* ou chamada de bases com um escore de qualidade associado para cada posição na sequência. Um escore de qualidade PHRED /X/ corresponde a uma probabilidade de erro de aproximadamente 10^-/X//10. Assim, um escore de qualidade PHRED 30 corresponde a uma acurácia de 99.9% para o base call na leitura bruta (*raw read*).

PHYLIP - *PHYLogeny Inference Package*, um pacote multiplataforma de programas para inferência filogenética e geração de arvores evolutivas, disponível gratuitamente: http://evolution.genetics.washington.edu/phylip.html.

Pirimidina – um composto nitrogenado semelhante ao benzeno, mas com um anel heterocíclico: dois átomos de nitrogênio substituem o carbono nas posições 1 e 3.

Plasmídio – Elementos de replicação de DNA que pode existir na célula independente dos cromossomos.

Pleiotropia – os efeitos múltiplos num fenótipo devido a um único gene ou alelo, p.ex., as citocinas que podem se ligar a vários receptores celulares e afetar o crescimento e reações imunes.

PNG - *Portable Network Graphics*, um formato de arquivo extensível para armazenamento seguro, estável e bem comprimido de imagens *raster* (imagens compostas de linhas horizontais de pixels, tais como aquelas geradas em computador. Compressão de arquivos de imagem e aplicativos eh necessário para reduzir o tempo de transmissão na internet. O formato PNG ultrapassa as patentes do GIF (*Graphic Interchange Format*) e pode substituir muitos usos comuns do TIFF (*Tagged Image File Format*). Vários aspectos como cores indexadas, escala de cinza, e cores reais são suportados, assim como um canal-alfa opcional. O PNG eh valido para trabalhos online de visualização de aplicações como um padrão de imagem bem suportado pela internet.

Poli(A) – cauda poli(A) e sítio de poliadenilação – o trecho de resíduos Adenina (A) na ponta 3' do RNAm eucariótico que é adicionado ao pre-mRNA enquanto é processado, antes de ser transportado do núcleo ao citoplasma e da tradução subsequente no ribossomo. A cauda aumenta a estabilidade do mRNA e permite o isolamento do mRNA por PCR usando iniciadores poly(T).

Polimorfismo – uma variação numa sequência comum de DNA entre indivíduos. Variações genéticas ocorrendo em mais de 1% da população seriam consideradas polimorfismos validos para analise de ligação genética (*genetic linkage*).

Polipeptídeo – um polímero linear (uma cadeia) de aminoácidos conectados por ligações peptídicas. Proteínas são grandes polipeptídeos, e os dois termos são comumente usados.

Predição estrutural – algoritmos que predizem a estrutura secundária, terciária, e às vezes, quaternária de proteínas a partir de suas sequências. A determinação da estrutura protéica a partir da sequência tem sido denominada de “segundo código genético” já que a estrutura terciária enovelada da proteína determina como ela funciona enquanto produto gênico.

Primer ou Iniciador – um oligonucleotídeo curto o qual fornece uma hidroxila livre na ponta 3’ para a síntese de DNA ou RNA por uma polimerase apropriada.

Procarioto – um organismo ou célula que não possui membrana nuclear.

Promotor – um trecho de DNA onde a RNA polimerase se ligará e iniciará a transcrição. Uma sequência de DNA que se localiza antes de um gene e controla a sua expressão.

Proteoma – o conjunto inteiro de proteínas de um dado organismo numa dada condição. O mesmo organismo vai apresentar diversos proteomas, de acordo com as diferentes condições em que estiver mantido, enquanto seu genoma será sempre basicamente o mesmo.

Proteômica – o estudo dos proteomas. Geralmente, a catalogação de todas as proteínas expressas por um determinado tipo celular ou tecido, obtido pela identificação de proteínas a partir de extratos celulares usando uma combinação de eletroforese bidimensional e espectrometria de massa. A análise em larga-escala da composição e função protéica.

Pseudogene - uma sequência de DNA que é muito parecida com um gene normal, mas que foi levemente alterada de forma que não é expressa. Tais genes provavelmente foram funcionais em algum momento, mas, ao longo da evolução, adquiriram uma ou mais mutações que os tornaram incapazes de codificar um produto protéico.

PSI-BLAST - *Position-Specific Iterated* BLAST – usado para buscas iterativas de similaridade de sequências de proteínas usando uma matriz de escore posição-especifica (PSSM). Variação do BLAST para verificar BDs de proteínas usando consultas (*queries*) que identificam outros membros da mesma família protéica. Todos os alinhamentos estatisticamente significativos encontrados pelo BLAST são combinados em um alinhamento múltiplo, do qual uma PSSM é construída. Essa matriz também é usada para buscar no BDs outros alinhamentos significativos, e o processo pode ser iterado (repetido) até que nenhum novo alinhamento seja encontrado.

PSSM - *Position-Specific Score Matrix*. A PSSM fornece o escore log-odds de busca para um *matching* particular de aminoácido em uma sequência alvo.

PubMed – Um sistema de coleção/recuperação/resgate contendo citações, resumos, e termos de indexação para artigos de periódicos nas ciências biomédicas. Ele inclui citações da literatura fornecidas diretamente ao NCBI

pelos editores/*publishers*, bem como a URL para artigos na integra (full text) pelos web sites das editoras. O PubMed possui o conteúdo completo dos BDs MEDLINE e PREMEDLINE, tendo também artigos e periódicos considerados fora do escopo do MEDLINE, comportando um total de mais de 1,100,000 artigos de mais de 340 *journals.*

Purina – base nitrogenada (denominada então base púrica), compostos orgânicos heterocíclicos. Todas são compostas por um anel aromático duplo (anel purina).

QTL - *Quantitative Trait Locus*. Um QTL é uma possibilidade de que uma certa região do cromossomo contenha genes que contribuam significativamente para a expressão de um traço/característica complexa. QTLs são geralmente identificados pela comparação de ligação de marcadores moleculares polimórficos e medidas de traços fenotípicos. A densidade de um mapa de ligação é importante na localização precisa e acurada dos QTLs; quanto maior a densidade do mapa, mais precisa a localização dos QTLs putativos, embora exista uma crescente probabilidade que falsos positivos sejam detectados.

Query (sequência) – uma sequência de DNA, RNA ou proteína usada para busca em um BDs de sequências no intuito de identificar membros de famílias próximas ou remotas (homólogos) de função conhecida, ou sequências com sítios ou regiões ativas similares (análogas), de onde a função da *query* pode ser deduzida.

Recessivo - qualquer característica que é expressa fenotìpicamente somente quando presente em ambos os alelos de um gene (antônimo de dominante).

Reciprocal best hits – Os melhores hits recíprocos são proteinas de diferentes organisms que são os melhores hits uma da outra (*each other's top* BLAST *hit*), quando os proteomas daqueles organisms são comparados um com o outro.

Recombinante (rDNA) – moléculas de DNA que resultam da fusão de DNAs de diferentes fontes. A tecnologia empregada para recortar DNA e amplificar os DNAs heterogêneos resultantes.

Recombinação – uma nova combinação de alelos resultando de um rearranjo ocorrendo por crossing-over ou por variedade independente.

Recursão – um procedimento algorítmico pelo qual um algoritmo se aciona para executar um calculo até que o resultado exceda um limiar (*threshold*), situação na qual o algoritmo cessa. É um procedimento poderoso para processar dados, sendo computacionalmente bastante eficiente.

Rede Neural – é um conjunto interconectado de elementos, unidades ou nós, de processamento simples, cuja funcionalidade é levemente inspirada no cérebro animal. A habilidade de processamento da rede é armazenada nas

conexões interunidades, ou pesos, obtidos por um processo de adaptação a um conjunto de padrões de treinamento. Em bioinformática, as redes neurais são usadas para mapear dados e fazer predições, tais como tomar um MSA de uma família de proteínas como um training set a fim de identificar novos membros da família a partir de seus dados de sequências.

RefSeq – um BDs do NCBI de sequências-referências; um conjunto curado, não-redundante que inclui contigs de DNA genômico, mRNAs e proteínas de genes conhecidos, e cromossomos inteiros.

Relacional (BDs) – um BDs que segue as onze regras de E.F. Codd, uma série de passos matemáticos e lógicos para a organização e sistematização de dados dentro de um sistema de software que permite fácil recuperação, atualização e expansão.

Relational Database Management Systems (RDBMS) ou Sistema de Gerenciamento de BDs Relacional (SGBDR) – um sistema de software que inclui uma arquitetura de BDs, uma linguagem de consulta (*query language*), carregamento dos dados, ferramentas de atualização e outros softwares subordinados que, juntos, permitem a criação de uma aplicação de BDs relacional.

Repetições (*repeat sequences*) - sequências repetitivas e repetições aproximadas ocorrem ao longo do DNA de organismos superiores (mamíferos). Por exemplo, as sequências Alu (cerca de 300 caracteres de comprimento), aparecem centenas de milhares de vezes no DNA humano com cerca de 87% de homologia com um consenso de trecho Alu. Alguns trechos curtos (substrings) como os TATA-boxes, poly-A e (TG)* também aparecem mais frequentemente do que se esperaria aleatoriamente. As sequências reptitivas também podem aparecer dentro dos genes, na forma de mutações ou alterações nestes genes. As sequências repetitivas, especialmente os elementos móveis, possuem muitas aplicações na pesquisa genética. DNA transposons e retroposons são rotineiramente utilizados em mutagênese de inserção, mapeamento e rotulação gênica, e em transferência gênica em vários sistemas modelo.

Replicação – a síntese de uma macromolécula informativamente idêntica (ex. DNA) a uma molécula molde.

Repressor – o produto da proteína de um gene regulatório que combina com um operador específico (sequência de DNA regulatória) e depois bloqueia a transcrição de genes em um operon.

RFLP (*Restriction Fragment Length Polymorphism*) - Variações genéticas no sitio onde uma enzima de restrição “cliva” ou “corta” um fragmento de DNA. Tais variações afetam o tamanho dos fragmentos resultantes; e essas sequências podem ser usadas como marcadores em mapas físicos e de ligação.

RNA - Ácido ribonucléico – uma categoria de ácidos nucléicos, nos quais, o componente açúcar é uma ribose e consiste de 4 nucleotídeos, timina, uracila, guanina, e adenina. Existem vários tipos de RNA: RNA mensageiro, RNA transportador, RNA ribossômico, pequenos RNAs, RNA nucleolar, etc.

RNA Mensageiro (mRNA) – cópia do RNA complementar de um DNA formado de uma fita simples de DNA alvo durante a transcrição que migra do núcleo para o citoplasma, onde este é processado em uma sequência carregando a informação que codifica um domínio polipeptídico.

RNA transportador (tRNA) – uma pequena molécula de RNA que reconhece um aminoácido específico, transporta-o para um códon específico no RNAm, e posiciona-o corretamente na cadeia polipeptídica nascente.

RPS-BLAST - *Reverse Position-Specific* BLAST - Um programa usado para identificar domínios conservados em uma seqüência sendo consultada (*protein query sequence*). Ele faz isso pela comparação de uma *query sequence* com matrizes de escore posição especifica (PSSMs) que foram obtidas a partir de alinhamentos de domínios conservados. O RPS-BLAST é uma versão reversa do PSI-BLAST; entretanto, o RPS-BLAST compara uma *query sequence* contra um BDs de perfis preparados com alinhamentos pré-ordenados, enquanto o PSI-BLAST constrói os alinhamentos partindo de uma única sequência protéica.

Reverse transcriptase-PCR (RT-PCR) – procedimento no qual a amplificação de PCR é realizada no DNA que é primeiramente gerado pela conversão de RNAm para cDNA usando a transcriptase reversa.

SAGE - *Serial Analysis of Gene Expression* – Uma técnica experimental destinada a medir quantitativamente a expressão gênica.

Seletividade – de algoritimos de busca de similaridades é definida como um limiar de significância para relatar matches de sequências em BDs.

Sensibilidade - de algoritimos de busca de similaridade giram em torno de 2 áreas: primeiro, quão bem pode o método detectar relações significativas entre 2 sequências relacionadas na presença de mutações e erros de sequenciamento; segundo, como a natureza heurística do algoritimo afeta a probabilidade com que um match poderá ou não ser detectado.

Sequência de Consulta (*query sequence*) – em Bioinformática pode se referir a sequência (de DNA, RNA ou proteína) que está sendo “jogada contra” um BDs para procurar similaridade, como, p.ex., num BLAST.

Sequência Rascunho (*draft sequence*) - refere-se a sequência (geralmente de DNA) que não está finalizada, mas geralmente é de alta qualidade (ou seja, tem acurácia maior que 90%).

Sequência Sinal (sequência líder) – uma sequência curta localizada na região amino- terminal.

Sequenciamento *single-pass* – sequenciamento rápido de grandes segmentos de um genoma de um organismo por isolamento do maior número possível de sequências expressas e execução de corridas de sequências finais únicas 5' ou 3'.

Sequenciamento Multiplex – abordagem para sequenciamento em larga-escala que usa diversas amostras mistas de DNA, através de métodos de separação e análise simultânea de géis.

Sequin – Um software desenvolvido pelo NCBI para submissão e atualização de entradas de sequências nos BDs do GenBank, EMBL, ou DDBJ. Ele é capaz de manipular submissões simples (contendo uma sequência única e curta de mRNA) e complexas (contendo longas sequências, anotações múltiplas, conjuntos segmentados de DNA, ou estudos amplos filogenéticos e de populações.

SGML - *Standard Generalized Markup Language*- O padrão internacional de especificação da estrutura e conteúdo dos documentos eletrônicos. A SGML é usada para marcação de dados de maneira auto-explicativa. A SGML não é propriamente uma linguagem, mas sim uma forma de definir linguagens desenvolvidas sob seus princípios. Uma subdivisão da SGML chamada XML é muito mais usada para marcação de dados, enquanto outra, a HTML, usa alguns de seus conceitos na oferta de uma linguagem universal de marcação para disponibilizar informações e associar diferentes partes daquelas informações.

Sintenia – na mesma fita – O termo sintenia conservada se refere a ordem gênica conservada em cromossomos de espécies diferentes, mas relacionadas.

Sitio(s) – sítios em sequências podem ser localizados no DNA (ex. sítios de ligação, sítios de clivagem) ou em proteínas.

SMART - *Simple Modular Architecture Research Tool* – Uma ferramenta (online ou local) para identificação e anotação automáticas de domínios em sequências fornecidas pelo usuário. O BDs do SWISS-PROT, por exemplo, é uma coleção extensivamente anotada e não-redundante de sequências de proteínas. As anotações do SWISS-PROT são mineradas para a geração dos alinhamentos que originam as anotações derivadas do SMART.

SNP – *Single Nucleotide Polymorphism* – Uma variação comum e numerosa que ocorre no DNA com freqüências de 1 a cada 1,000 bases. Um SNP é um sitio de uma única base dentro do genoma na qual mais de uma das 4 bases possíveis é comumente encontrada em populações regulares/naturais. Varias centenas de milhares de SNPs estão sendo identificados e mapeados no genoma humano, propiciando um mapa dos mais densos possíveis sobre diferenças genéticas individuais.

Sonda (Probe) = Qualquer produto bioquímico que possa ser marcado ou rotulado de forma tal que possa ser usado para identificar ou isolar um gene, RNA, ou proteína.

Sequence Tagged Site (STS) – uma sequência única de uma localização cromossômica conhecida que pode ser amplificada por PCR. STSs atuam como marcadores físicos para mapeamento genômico e clonagem.

Clonagem Shotgun – clonagem de um segmento de gene ou genoma inteiro pela geração de um conjunto randômico de fragmentos usando endonucleases de restrição para criar uma biblioteca gênica que pode ser subsequentemente mapeada e sequenciada para reconstruir um genoma inteiro.

Busca de similaridade (homologia) – dada uma nova sequência gênica, existem 2 abordagens principais para a predição de estrutura e função da sequência de aminoácido. Métodos de homologia são mais poderosos e são baseados na detecção de similaridade significativa em sequências a uma proteína de estrutura conhecida, ou de uma sequência padrão característica de uma família de proteína.

Splicing – a junção de partes componentes de DNA e RNA separados. O splicing de RNA em eucariotos, p.ex., envolve a remoção de íntrons e junção dos éxons do transcrito pré-RNAm antes da maturação. O splicing significa a remoção dos íntrons do transcrito de RNA, e a união de todos os éxons em ordem correta para liberar um RNAm maduro functional. O splicing alternativo é o splicing de um único RNA transcrito primário que ocorre em 2 ou mais diferentes padrões, bem definidos. Em cada splicing padrão, um conjunto definido de éxons são aglomerados para liberar uma molécula de RNAm. O efeito final do splicing alternativo é a geração de um grande número de diferentes proteínas a partir de um número relativamente pequeno de genes.

SWISS-PROT - <http://www.ebi.ac.uk/swissprot/> - um BDs (curado) de sequências de proteínas com alto nível de anotação (descrição da função protéica, de estruturas de domínios, modificações pos-traducionais, variantes, etc.), com mínima redundância, alto grau de integração com outros BDs e mediana interoperabilidade de funções permitidas.

Tecnologia da Informação – Aquisição, processamento, armazenagem e disseminação de todos os tipos de informação usando tecnologia de computação e sistemas de telecomunicação.

TIGR - The Institute for Genomic Research <http://www.tigr.org>

Timina – base pirimídica encontrada no DNA, mas não no RNA.

Tradução – o processo de conversão de RNA em proteína através da montagem de uma cadeia polipeptídica mediante uma molécula de RNAm nos ribossomos.

Transcrição – a montagem de uma fita única complementar de RNA através de um DNA molde.

Transcriptoma – o conjunto completo de moléculas de RNA produzidas pelo genoma é usualmente referido como um transcriptoma. No caso dos eucariotos, um único gene pode produzir mais que um tipo de RNAm maduro por um fenômeno chamado de *splicing* alternativo.

Transcriptase Reversa – uma DNA polymerase que pode sintetizar uma fita de DNA complementar (cDNA) usando RNA como molde, também chamados de RNA dependentes de DNA polimerase.

Transcrito – a fita simples de uma cadeia de RNAm que é montada a partir de um gene molde.

Transgene / Transgênico – um gene estranho que é introduzido em uma célula ou organismo inteiro (ex.camundongos transgênicos) para propósitos terapéuticos ou experimentais.

3-D ou 3D - Tridimensional.

UNIX – Um sistema operacional (SO) desenvolvido por Dennis Ritchie e Kenneth Thompson nos Laboratórios Bell há mais de 30 anos com funções multitarefa e multiusuarios e portabilidade para outros SOs; funcionalmente organizado em três níveis: o kernel, que programa as tarefas e gerencia o armazenamento; o shell, o qual conecta e interpreta os comandos do usuário, chama os programas da memória e os executa; e as ferramentas e aplicações, as quais oferecem funcionalidades suplementar ao SO, tais como editores de texto. O UNIX foi registrado pela Bell Laboratories <http://www.lucent.com/> como uma trademark para SOs de computadores e atualmente a marca esta em poder do The Open Group <http://www.opengroup.org/>.

URF - *Unidentified reading frame* (URF) – Uma ORF (matriz aberta de leitura) codificando uma proteína de função indefinida.

Uracila – base nitrogenada pirimídica encontrada no RNA, mas não no DNA.

URL - *Uniform Resource Locator* – o endereço de um recurso na Internet. A sintaxe URL vem na forma de *protocolo://host/localinfo*, where *protocol* especifica o sentido de atrair o objeto (tal como HTTP, usado pelos servidores e browsers web para trocar informações, ou FTP), *host* especifica a localização remota onde o objeto reside, e *localinfo* é uma string (geralmente um nome de arquivo) passado ao manipulador do protocolo na localização remoto. Também chamdo de *Uniform Resource Identifier* (URI).

UTR - *Untranslated Region*- A região não traduzida 3’ é aquela porção de um mRNA desde a posição do ultimo códon que é usado na tradução até a terminação 3’. A UTR 5’ é aquela porção de um mRNA desde a terminação 5’ até posição do primeiro códon usado na tradução.

Valor **H** – a entropia relativa das frequências de resíduos (explícitos e implícitos) (Karlin and Altschul, 1990). O valor H pode ser entendido como uma medida da informação media (em bits) disponível por posição que distingue um alinhamento ao acaso. No caso de altos valores de H, alinhamentos curtos podem ser diferenciados ao caso; ao passo que em baixos valores de H, um alinhamento mais longo pode ser necessário. (Altschul, 1991)

Valor **K** – um parâmetro estatístico usado no calculo de escores do BLAST que pode ser entendido como uma escala natural para a busca de tamanho de espaço. O valor K é usado na conversão de um escore bruto (*raw score*, (S)) em um escore bit (*bit score* (S')).

Valor **P** - a probabilidade de um alinhamento ocorrer com o escore em questão ou melhor. O valor p é calculado pela relação do escore de alinhamento observado, S, com a distribuição esperada dos escores HSP de comparações de sequências aleatórias de mesmo comprimento e composição da consulta (*query*) ao BDs. Os valores de P mais significantes serão aqueles próximos de 0. Os valores P e E values são maneiras diferentes de representar a significância de um alinhamento.

Variable numbers of tandem repeats (VNTRs) – Blocos de sequências de DNA que possuem entre 2-60 pares de bases, os quais se repetem desde 2 até mais de 20 vezes em diferentes indivíduos. Estes marcadores polimórficos (VNTRs) são muito usados em mapeamento genômico, análise de ligações e também DNA fingerprinting.

Vetor – Qualquer agente que transfira material (geralmente DNA) de um hospedeiro a outro. Em geral, vetores de DNA são elementos de DNA autônomos (tais como os plasmídios) que podem ser manipulados e integrados dentro de um DNA hospedeiro ou vírus recombinante.

Visualização – visualização é o processo de representação de dados científicos resumidos como imagens que podem ajudar no entendimento do significado dos dados

Sequência WGS - *Whole Genome Shotgun* – Uma estratégia de sequenciamento automático de DNA, onde o DNA de alto peso molecular é picotado/clivado em fragmentos aleatórios, selecionado por tamanho (geralmente em fragmentos de 2, 10, 50, e 150 kb), e clonado em um vetor apropriado ao tamanho. Os clones são, então, sequenciados em ambas as terminações (3' e 5'). As duas terminações de um mesmo clone são chamadas de *mate pairs* (casais). A distancia entre os pares pode ser inferida se o tamanho da biblioteca for conhecido e tiver uma faixa limitada de desvio. As sequencias são alinhadas com um software de montagem (tipo o pacote PHRED, PHRAP, CONSED, o PPC).

XML - *Extensible Markup Language* – descreve uma classe de objetos de dados chamados documentos XML (uma subdivisão da SGML) que se conformam com os documentos SGML, sendo feitos de unidades de armazenamento chamadas entidades, as quais contem dados parseados ou não (*unparsed*). Dados parseados são feitos de caracteres (uma unidade de texto), alguns dos quais formam dados de caracteres e outros formam marcação. A marcação inclui rótulos (*tags*) de informação sobre os dados, i.e., a descrição de uma estrutura e conteúdo do documento. Os dados de caracteres compreendem todo o texto que não for marcação.

YAC - *Yeast Artificial Chromosome* – Um vetor de clonagem para segmentos extremamente grandes de DNA de uma outra espécie “spliceado” em DNA de levedura. YACs são usados para clonar ate um milhão de bases DNA estranho em uma célula hospedeira, onde o DNA será propagado como outros cromossomos da célula de levedura.