Ministério da Agricultura, Pecuária e Abastecimento

Instruções Técnicas 6

ISSN
Dezembro, 2001
Campinas, SP

Entendendo e Interpretando os Parâmetros Utilizados por BLAST

Roberto Hiroshi Higa[1]

O advento da tecnologia de obtenção rápida de seqüências de DNA, em meados dos anos 70, provocou uma explosão de informações sobre seqüências biológicas (Altschul et al., 1994). Desde então, o número de seqüências têm aumentado a uma velocidade cada vez maior, principalmente com o surgimento dos projetos Genoma, que visam obter as seqüências de todos os genes de um organismo completo. A maioria destas seqüências encontram-se organizadas em forma de bancos de dados, muitos de acesso público tais como Genebank (The National Center For Biotechnology Information, 2001b) e EMBL (European Molecular Biology Laboratory, 2001) para DNA, Swiss Prot (Swiss Institute of Bioinformatics, 2001) e PIR (Munich Information Center for Portein Sequences, 2001) para proteínas. Cabe observar que o termo banco de dados, aqui, refere-se apenas a um conjunto usualmente grande de seqüências catalogadas e as respectivas anotações, não existindo qualquer vínculo com Sistemas Gerenciadores de Bancos de Dados (SGBDs).

Atualmente, esses bancos de dados constituem-se em ferramenta de trabalho essencial para biologistas moleculares. Isto porque, baseado na observação de que genes ou proteínas com seqüências similares ou com regiões similares têm grande chance de possuírem funções similares, as primeiras informações para determinação da função de um gene, cuja seqüência foi recentemente obtida, quase sempre são obtidas pela busca de similaridades entre a nova seqüência e seqüências de proteínas ou famílias de proteínas conhecidas (Altschul et al., 1990).

Entretanto, para que essa tarefa de busca de seqüências similares possa ser efetivamente realizada é necessário que os biologistas moleculares tenham à sua disposição uma ferramenta computacional que os auxiliem. Neste sentido, diversos algoritmos para busca em banco de dados de seqüências foram criados. Abordagens baseadas em algoritmos de programação dinâmica, tais como o algoritmo de Smith-Waterman (Durbin et al., 1998; Setubal & Meidanis, 1997) são proibitivos devido ao custo computacional. Isto, então, levou ao desenvolvimento de métodos heurísticos para esta tarefa, tais como BLAST e FASTA (Altschul et al., 1990; Altschul et al., 1997).

[1] *Mestre em Engenharia Elétrica, Pesquisador da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo – 13083-970 – Campinas, SP. (roberto@cnptia.embrapa.br)*

Embrapa

BLAST – **B**asic **L**ocal **A**lignment **S**earch **T**ool é, hoje, o método mais utilizado para realizar buscas de seqüências similares em bancos de dados de seqüências, sendo que suas implementações mais conhecidas são a do NCBI – National Center for Biotechnology Information e o da University of Washington, conhecido como WU-BLAST.

O BLAST oferecido pelo NCBI é, na verdade, uma família de serviços, onde o usuário possui diversas opções, dependendo da seqüência de entrada, se ela é constituída de nucleotídeos ou aminoácidos, se o banco de dados alvo é de nucleotídeos, aminoácidos ou está restrito a um tipo de organismo, além dos parâmetros relacionados ao algoritmo de busca.

Uma vez que nem sempre estes parâmetros são de entendimento direto, nesta instrução alguns aspectos relacionados à teoria que suporta BLAST são apresentados, visando proporcionar uma melhor utilização deste método através do melhor entendimento dos parâmetros envolvidos.

O Método BLAST para Determinação de Similaridades entre Seqüências Biológicas

BLAST é uma heurística que tenta privilegiar a eficiência computacional ao mesmo tempo em que otimiza uma medida de similaridade específica (Altschul et al., 1997). Para que as heurísticas utilizadas no algoritmo de busca de BLAST possam ser entendidas, é preciso que alguns conceitos e resultados da teoria estatística que suporta BLAST sejam apresentados.

Score e matrizes de substituição

Para que seja possível estabelecer um alinhamento, buscando similaridades, entre duas seqüências é preciso que um esquema de score seja estabelecido. A discussão que se segue considera apenas esquemas de score utilizados para comparação de proteínas, uma vez que os esquemas para comparação de DNA são mais simples, e está baseada no exposto em (Ewens & Grant, 2001).

Ao alinhar duas seqüências de aminoácidos, deseja-se que um par alinhado (um aminoácido de cada seqüência) contribua para o score total do alinhamento com um score tanto maior quanto for a probabilidade de se encontrar essa substituição em seqüências biologicamente relacionadas. A abordagem utilizada em comparação seqüenciais de proteínas é a utilização de matrizes de substituição, sendo que as duas matrizes de substituição mais utilizadas são conhecidas como PAM e BLOSUM.

Determinação de Matrizes de Substituição BLOSUM

A construção de uma matriz BLOSUM (BLOcks SUbstitution Matrices) começa com a obtenção de um conjunto de seqüências protéicas oriundas de bases de dados públicas e que foram agrupadas em famílias. Daí, seguem-se os seguintes passos:

- a partir dessas seqüências, são extraídos blocos de seqüências alinhadas, onde blocos são alinhamentos sem gap de uma região altamente conservada da família protéica. Observe que para obter o alinhamento múltiplo é necessário utilizar um esquema de score. Como isto é exatamente o que se pretende, nesta fase é atribuído score 1 para um par alinhado constituído do mesmo aminoácido e 0 caso contrário (matriz de substituição unitária);
- para cada bloco, são determinados clusters de seqüências, tal que cada seqüência em um cluster possua identidade X% (ex.: 85%) para pelo menos uma seqüência naquele cluster naquele bloco. As freqüências calculadas nos passos seguintes são medidas com relação ao número de cluster e não de seqüências. A motivação para este passo é o fato desejável de que cada par de seqüências em um bloco tenha uma quantidade de "distância evolucionária" equivalente. O valor percentual X caracteriza a matriz, de forma que dependendo do seu valor têm-se matrizes BLOSUM62, BLOSUM85, etc.;
- é medida a freqüência de ocorrência no conjunto de blocos de cada aminoácido, denominada

$$\Rightarrow \mathbf{f}_i, \text{ p/ } 1 \leq i \leq 20.$$

- é medida a freqüência de ocorrência no conjunto de blocos de cada par de aminoácidos (x,y), denominada

$$\Rightarrow \mathbf{f}_{xy}, \text{ p/ } 1 \leq x,y \leq 20.$$

- para cada par de aminoácidos, determina-se a razão das probabilidades[2] de ocorrência do alinhamento dos aminoácidos ao acaso, dado a freqüência nos blocos, pela proporção observada nos blocos

$$\Rightarrow \mathbf{e_{xy}} = \begin{cases} 2\dfrac{f_x f_y}{f_{xy}}, \text{se } x = y \\ \dfrac{f_x f_y}{f_{xy}}, \text{se } x \neq y \end{cases} \quad p/1 \leq x, y \leq 20.$$

- Os scores da matriz de substituição são obtidos fazendo-se

$\Rightarrow \mathbf{S_{xy}} = \mathbf{ROUND}\ (-2 \log_2 (\mathbf{e_{xy}}))$, $p/1 \leq x, y \leq 20$.

Onde o operador ROUND faz o arredondamento do valor passado como parâmetro.

Para compensar a utilização da matriz de substituição unitária no primeiro passo do procedimento, o processo é repetido mais duas vezes utilizando a nova matriz de substituição.

Determinação de matrizes de substituição PAM

Um PAM ou "Point Accepted Mutations" é a substituição de um aminoácido em uma proteína e que é "aceito" pela evolução, no sentido de que na espécie em questão, a mutação não só apareceu, mas disseminou-se em praticamente toda a espécie.

A construção de uma matriz de substituição PAM começa, como no caso da matriz BLOSUM, pela obtenção de um conjunto de alinhamentos múltiplos de proteínas fortemente relacionadas arranjadas em blocos (alinhamento sem gap). Estes blocos são, então, utilizados para construir um modelo evolucionário, a partir do qual os parâmetros da matriz PAM são obtidos. Os passos a serem seguidos são:

- Para cada bloco de seqüências, é construída uma árvore filogenética utilizando o método de máxima parcimônia (Setubal & Meidanis, 1997; Durbin et al., 1998; Ewens & Grant, 2001). Este algoritmo constrói uma árvore com as seqüências originais nas folhas e as seqüências inferidas nos nós internos, tal que o número de substituições na árvore é mínimo. As arestas desta árvore representam a mutação em uma única posição que relaciona as duas seqüências nos nós que ele liga. Finalmente, observe que mais de uma árvore podem resultar do algoritmo.
- Para cada bloco, a árvore filogenética correspondente é utilizada para fazer uma contagem do número de mutações para cada par de aminoácidos j e k da seguinte forma:
  - $\Rightarrow$ Cada aresta contribui com 1 para a mutação j-k se j π k.
  - $\Rightarrow$ Cada aresta contribui com 2 para a mutação j-k se k = k.
  - $\Rightarrow$ Se um bloco possui mais uma árvore filogenética associada, as mutações são contabilizadas para todas as árvores filogenéticas e a contagem final é dividida por n, onde n é o número de árvores filogenéticas associadas ao bloco.
- Totalize a quantidade de mutações j-k, A_{jk}, para todos os blocos.
- Definindo a quantidade $\mathbf{a_{jk}} = \dfrac{A_{jk}}{\sum_m A_{jm}}$

  as seguintes quantidades podem ser definidas:

$$\mathbf{p_{jk}} = \begin{cases} \mathbf{ca_{jk}}, \text{se } j \neq k \\ 1 - \sum_{k \neq j} \mathbf{ca_{jk}}, \text{se } j = k \end{cases}, p/1 \leq j, k \leq 20.$$

  Note que esta equação implica em $\sum_k p_{jk} = 1$. Assim, se a constante c for suficientemente pequena para que cada p_{jj} seja não negativo, a matriz P = $\{P_{jk}\}$ tem as propriedades de uma matriz de transição de uma cadeia de Markov.
- Denotando p_j como a freqüência observada do aminoácido j, considerando as seqüências associadas aos blocos, a proporção de aminoácidos que sofreram mutação após uma interação da cadeia de Markov definida pela matriz de transição P é dada por:

$$\Rightarrow \sum_j p_j \sum_{k \neq j} p_{jk} = c \sum_j \sum_{k \neq j} p_j a_{kj}, p/1 \leq j, k \leq 20.$$

  Se a proporção de mutações esperada é fixada em 1%, o valor de c é dado por

$$\Rightarrow \mathbf{c} = \frac{0.01}{\sum_j \sum_k p_j a_{jk}}, p/1 \leq j, k \leq 20.$$

  A matriz de transição obtida, considerando-se essa proporção de mutações esperada, corresponde a uma distância evolucionária de 1 PAM e é denotada por M_1. De modo geral, a matriz correpondente a uma distância evolucionária de n PAM é obtida pela n-ésima potência da matriz M_1 e é denotada por M_n.
- Denotando-se o elemento da matriz M_n por $m_{jk}^{(n)}$, a matriz PAMn é dada por:

---

[2] O termo correto, em inglês é likelihood. Quando a massa de dados é suficientemente grande seu valor aproxima-se da probabilidade. Aqui, toma-se a liberdade de utilizar o termo probabilidade indistintamente.

$$\Rightarrow \mathbf{S_{jk}} = C\log\left(\frac{m_{jk}^{(n)}}{p_k}\right) p/1 \le j,k \le 20.$$

ou, denotando q(j,k) como a probabilidade conjunta de que o aminoácido j ocorra em uma dada posição no tempo 0 e o aminoácido k ocorra nesta mesma posição n passos depois, de acordo com a cadeia de Markov definida pela matriz M1, $q(j,k) = p_j m_{jk}^{(n)}$ e

$$\Rightarrow \mathbf{S_{jk}} = C\log\left(\frac{q(j,k)}{p_j p_k}\right), p/1 \le j,k \le 20.$$

Observe que a constante C não é importante e refere-se à escala do score.

Resultados da Teoria Estatística de Comparação Local de Seqüências

Seja um alinhamento entre duas seqüências protéicas de comprimento N, obtidas ao acaso. Este evento pode ser modelado como um processo aleatório denominado caminhada aleatória, um caso particular de cadeia de Markov e fornece a teoria de probabilidade básica que suporta BLAST (Ewens & Grant, 2001). Da análise desse alinhamento, os seguintes parâmetros estatísticos são obtidos (Ewens & Grant, 2001):

- O número esperado de HSP (High-scoring Segment Pair) com score maior ou igual a um valor específico S é estimado por:

  $\Rightarrow E(S) = NKe^{-\lambda S}$

  Onde N é o comprimento das seqüências, K é uma constante obtida através de séries geométricas convergentes que dependem apenas do score s(j,k) e das probabilidades p_j e p'_k, as probabilidades de ocorrência dos aminoácidos j e k, e λ é obtida através da seguinte expressão, também dependente apenas da matriz de substituição utilizada s(j,k) e das probabilidades p_j e p'_k. A notação p_j e p'_k indicam que os dois aminoácidos j e k alinhados em posição específica foram gerados por mecanismos independentes.

  $$\Rightarrow \sum_J \sum_k p_j p'_k e^{\lambda S(j,k)} = 1, p/1 \le j,k \le 20.$$

- A probabilidade que, neste alinhamento, exista um HSP maior que S é dado por:

  $\Rightarrow P(Y > S) = 1 - e^{-E(S)}$

  Essa probabilidade é o ***P-value*** para a fdp associada a Y. Essa, por sua vez pode ser aproximada por uma fdp conhecida como Extreme Value Distribuição (EVD) dada por (Durbin et al., 1998):

  $$\Rightarrow P(Y > y) \cong e^{(-KNe^{\lambda(x-\mu)})}$$

Todos estes resultados foram obtidos considerando-se um alinhamento entre duas seqüências de tamanho fixo (N). Entretanto, BLAST manipula alinhamentos entre seqüências de tamanhos variáveis. Nenhuma teoria que estenda os resultados mencionados foi desenvolvida, mas diversos trabalhos de simulação têm mostrado sua validade, com pequenas adaptações, para o caso geral quando as seqüências não possuem o mesmo comprimento (Altschul et al., 1990; Altschul et al., 1994; Ewens & Grant, 2001). Assim, as equações apresentadas podem ser reescritas da seguinte forma:

$$\Rightarrow E(S) = MNKe^{-\lambda S}$$

$$\Rightarrow \sum_J \sum_k p_j p'_k e^{\lambda S(j,k)} = 1,\ 1 \le j,k \le 20$$

$$\Rightarrow P(Y > y) \cong e^{(-KMNe^{\lambda(x-\mu)})}$$

Onde M e N são os comprimentos das seqüências alinhadas.

Além disso, devido a melhoramentos nas heurísticas utilizadas por BLAST e a inclusão de tratamento para gaps nas seqüências (Altschul et al., 1997), os parâmetros K e λ não podem mais ser obtidos analiticamente. Eles, agora, são estimados através de processos de simulação executados previamente.

Com relação à probabilidade de que exista pelo menos um HSP com escore S ou mais, para comparações entre uma "seqüência query" e uma base de dados de seqüência, desde que a base de dados inteira é D/N vezes mais longa que a seqüência de interesse, a seguinte correção é aplicada:

$$\Rightarrow \textit{Expect} = (1 - e^{-E(S)})\frac{D}{N}$$

Este valor é conhecido como ***Expect*** ou ***E-value*** e é apresentado no relatório do BLAST para avaliação da significância do alinhamento obtido.

Finalmente, no relatório BLAST também é apresentado um score normalizado, denominado **Bit Score** e definido da seguinte forma:

$$\Rightarrow S' = \frac{\lambda S - \ln K}{\ln 2}$$

A relação entre o "bit score" S´e o E-value é dado por

$$\Rightarrow E = MN2^{-S'}$$

Dessa forma, ao contrário do que ocorre quando se avalia a significância de um alinhamento a partir do score S, em que é preciso conhecer os valores de M, N, λ e K, conhecendo-se o "bit score" S´, é necessário conhecer apenas os valores do espaço de busca M e N.

O Algoritmo Usado por BLAST

O algoritmo utilizado por BLAST pode ser resumido nos seguintes passos (Setubal & Meidanis, 1997) e baseia-se na idéia de que bons alinhamentos locais provavelmente contém pequenos segmentos de identidades (Durbin et al., 1998):

- Compilar uma lista de segmentos de alto score (*word* no jargão de BLAST). Para proteínas, essa lista é formada por todas as palavras com w caracteres (*w-mer*) com score no mínimo T com algum *w-mer* da seqüência query.
- Procurar por hits na base de dados (cada hit corresponde a uma semente). Um hit é um pequeno segmento alinhado onde cada posição do alinhamento corresponde a uma identidade (as duas seqüências possuem o mesmo aminoácido na posição correspondente).
- Estenda as sementes. Essa extensão é realizada nos dois sentidos; inicialmente era realizada sem considerar gaps (Altschul et al., 1990), mas atualmente, as extensões são feitas com gaps e o processo para estender uma semente só é disparado se o seu score for maior que um limiar T, ela possuir outra semente a uma certa distância máxima entre elas e se o score da extensão com gaps que elas geram excede a um dado limiar S_g. (Altschul et al., 1997).

Observe que, para o cálculo dos scores, é utilizado uma matriz de substituição tal como PAM ou BLOSUM. Na verdade, diferentes versões dessas matrizes estão disponíveis como parâmetro para o usuário.

Finalmente, a estratégia adotada para penalizar gaps é uma função linear do comprimento do gap.

$$\Rightarrow \text{pen}_{\text{gap}}(x) = a + bx$$

Onde a é o valor da penalidade pelo gap e b é a penalidade por cada unidade de gap.

NCBI-BLAST

A implementação de BLAST (The National Center For Biotechnology Information, 2001a) mais largamente utilizada atualmente é o serviço mantido pelo NCBI. Nesta seção, apresentaremos brevemente o funcionamento deste serviço e as várias opções disponíveis para o usuário.

Dependendo do tipo de busca que se queira realizar, pode-se utilizar o BLAST através de uma das seguintes formas:

- ***blastp*** para comparação de seqüências de aminoácidos em bancos de dados de proteínas;
- ***blastn*** para comparação de seqüências de nucleotídeos em bancos de dados de DNA;
- ***blastx*** para comparação de uma seqüência de nucleotídeos transladada em todos os ORFs (Open Reading Frames) com bancos de dados de proteínas;
- ***tblastn*** para comparação de seqüência de proteína com um banco de dados de seqüências de nucleotídeos dinamicamente transladados em todos os seus ORFs; e
- ***tblastx*** para comparar os ORFs de uma seqüência de nucleotídeos com os ORFs de todos os nucleotídeos em um banco de dados de nucleotídeos.

Além disso, ao formular um query para busca, pode-se delimitar o espaço de busca de várias formas:

- o banco de dados a ser utilizado na busca;
- o resultado de uma query *Entrez*[3] sobre um banco de dados ou
- um organismo específico.

Os formatos aceitos para especificação da seqüência query compreendem:

- o formato FASTA;
- identificadores, normalmente códigos para acesso aos bancos de dados mantidos pelo NCBI como o GenBank.

---

[3] *Entrez* é o Sistema de recuperação e busca sobre os diversos bancos de dados mantidos pelo NCBI (The National Center For Biotechnology Information, 2001b).

- seqüências puras, que podem ou não ser intercaladas por caracteres brancos ou numéricos.

BLAST oferece valores default para uma série de parâmetros utilizados pelo algoritmo de busca, entretanto todos são configuráveis pelo usuário:

- o valor de Expect a ser utilizado como valor de corte para a busca. Este é um valor importante, pois indica o nível de significância a partir do qual os resultados podem ser incluídos no relatório de respostas;
- o tamanho da palavra a ser utilizada nas duas primeiras etapas do algoritmo;
- a matriz de substituição a ser utilizada e a função de penalidades para gap. Matrizes do tipo PAMn e BLOSUMn estão disponíveis para o usuário, sendo a matriz BLOSUM62 é a default. Este parâmetro é importante, pois dependendo da distância evolucionária desejada uma matriz pode ser mais adequada que outra. Observe ainda que para a matriz PAM, maior n corresponde a maior distância evolucionária, enquanto que para matrizes BLOSUM mantém-se a relação inversa; e
- opções para filtragem de segmentos de baixa complexidade que poderiam levar à obtenção de seqüências que, apesar de apresentarem um alto score e estatisticamente significativo, não possuem significado biológico.

Existem ainda diversas opções para formatação do relatório da busca, incluindo formatos ASN.1, texto e html.

Finalmente, duas variantes importantes incorporadas ao algoritmo BLAST devem ser destacadas:

- **PHI-BLAST** (Pattern-Hit Initiated BLAST): esta variante do BLAST utiliza uma expressão regular para selecionar regiões para busca de HSP, ou seja, dada uma seqüência protéica S e um a expressão regular P, PHI-BLAST procura por seqüências que satisfaçam P e sejam homólogas a S na vizinhança das ocorrências dos padrões.
- **PSI-BLAST** (Position Specific Interactive BLAST): esta variante realiza uma busca interativa, onde as seqüências resultantes de uma interação são utilizadas para construção de um modelo de score específico por posição (profile). PSI-BLAST não utiliza uma matriz de substituição, mas constrói, a cada iteração, uma matriz QxA, onde Q é o tamanho da seqüência e A o tamanho do alfabeto (20). Além disso, a cada interação o usuário pode remover algumas das seqüências do conjunto de respostas da interação anterior, bem como salvar o profile corrente. Estudos comparativos têm mostrado que PSI-BLAST é mais sensível para detectar relações distantes que o BLAST tradicional (Altschul et al., 1997).

Exemplo de Busca

Neste exemplo, a base de dados utilizada é a base de seqüências não redundantes de aminoácidos do GenBank (nr) e a seqüência de aminoácidos utilizada como query é o ORF (Open Reading Frame) correspondente à proteína ainda não caracterizada MJ0577 do Methanococcus Jannaschii (The National Center For Biotechnology Information, 2001c):

>gi|2501594|sp|Q57997|Y577_METJA PROTEIN MJ0577
MSVMYKKILYPTDFSETAEIALKHVKAFKTLKAEEVILLHV
IDEREIKKRDIFXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXGIPHEEIVKIAEDEGVDIIIMGSHGK
TNLKEILLGSVTENVIKKSNKPVLVVKRKNS

As posições na seqüência indicadas com "X" informam ao BLAST que estas posições devem ser ignoradas durante o processo de busca. Esta região apresenta um tipo de composição de baixa complexidade que não é detectado pelos filtros de BLAST, um "Coiled-Coil". O algoritmo utilizado para realizar está análise é o COILS (Lupas, 1997). O valor de Expect utilizado para corte foi "1", a matriz de substituição utilizada foi o BLOSUM62 e os valores para penalidade por gap foram a=11 (para existência do gap) e b=1(para a extensão do gap).

As Fig. 1, 2 e 3 apresentam, a representação gráfica, as seqüências ordenada por score e os alinhamentos entre as seqüências que retornaram da busca e a seqüência query. A partir da Fig. 1, observa-se que existem diversas seqüências relacionadas à seqüência query por um score relativamente alto. A linha rosa, que se estende por toda a seqüência representa a própria seqüência query. Olhando o correspondente alinhamento, verifica-se tratar-se de seqüências associadas à determinação da estrutura da MJ0577. Os alinhamentos que se estendem por toda a seqüência representam possíveis homólogos da MJ0577 (parólogos em Pyrococcus horikoshii, Methanobacterum Thermoautotrophicum e Agrobacterium tumefaciens), sendo que à medida que o valor do score decai (e o correspondente Expect aumenta), mais distante é a seqüência homóloga.

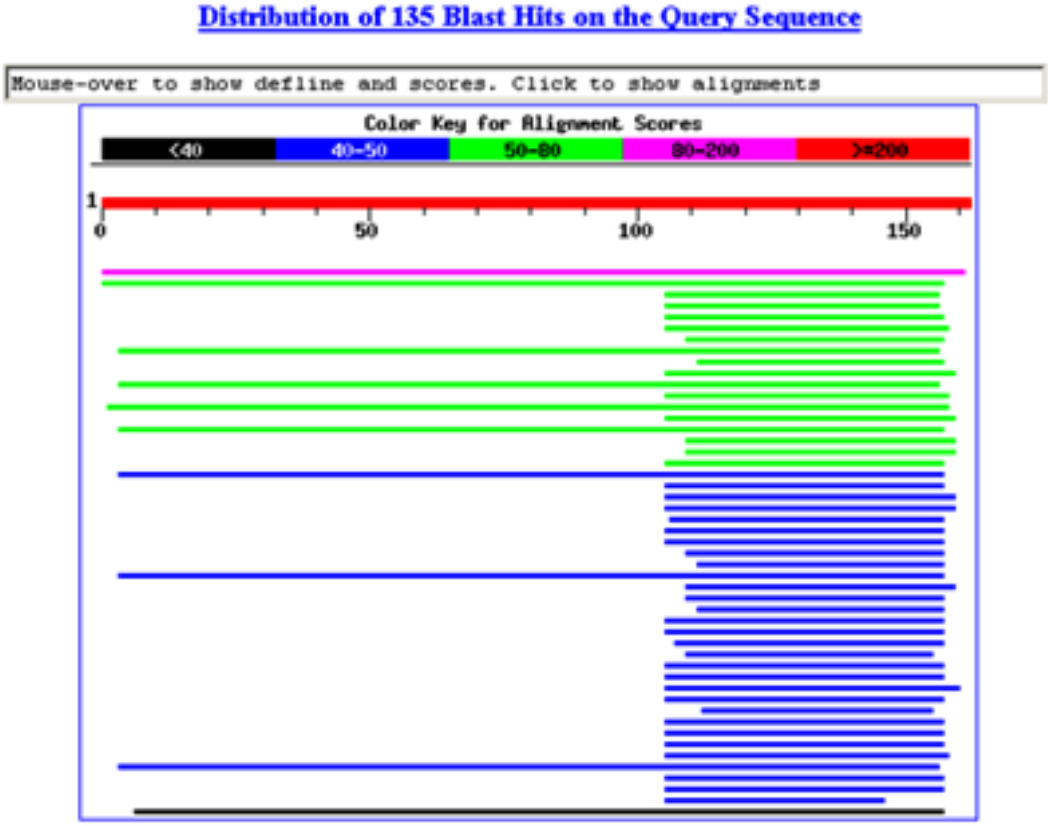


Fig. 1. Visualização gráfica do resultado de busca.

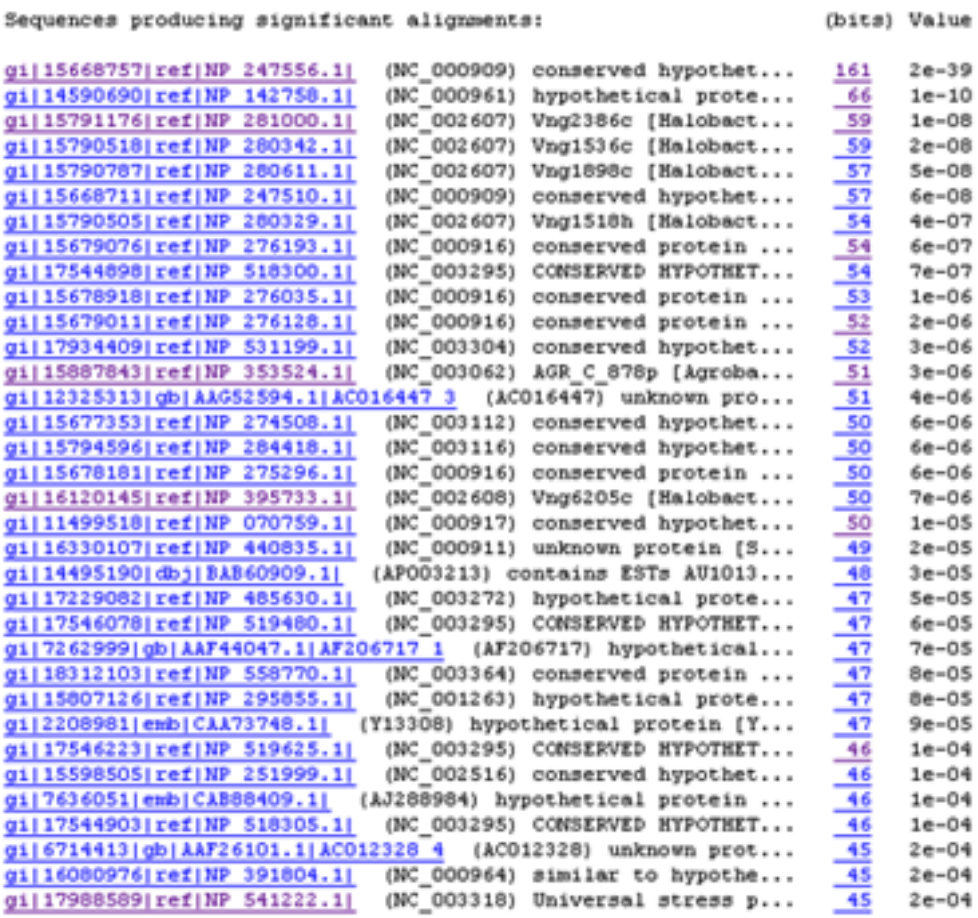
Observando as Fig. 2 e 3, procura-se por seqüências anotadas próximas à MJ057. As primeiras seqüências com anotação encontradas são aquelas com entrada no GeneBank gi|15887843 e gi|15155425, com score 51.2 e Expect 3 e-06. Trata-se, portanto, de um alinhamento bastante significativo, o que indica que a proteína MJ057 está provavelmente associada à família de proteínas USP (Universal stress protein).

>gi|15887843|ref|NP_353524.1| (NC_003062)
AGR_C_878p [Agrobacterium tumefaciens]
[Agrobacterium tumefaciens str. C58 (Cereon)]
gi|15155425|gb|AAK86309.1| (AE007985)
AGR_C_878p [Agrobacterium tumefaciens str. C58 (Cereon)]
Length = 160
Score = 51.2 bits (121), Expect = 3e-06
Identities = 42/158 (26%), Positives = 62/158 (38%), Gaps = 11/158 (6%)

Query: 2 SVMYKKILYPTDFSETAEIALKHVK
AFKTLKAEEVILLHVIDER
EIKKRDIFXXXXXXXX 61
+VM+K IL PTD S A+IA+ A +V ++ V +
+ D+
Sbjct: 14
NVMFKHILIPTDGSPLAQIAIDQGFALAREAGA
KVTVVTVSEPFHVIASDV---------- 64

Query: 62
XXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXGIPHEEIVK
IAEDEGV 121
G P E I++IA+ G
Sbjct: 65 —EDIAAIAEEEFHRCEEAEHLL
RDTQAHAAAMGLDCEALLARAGRPDEA
IIEIADRTGC 122

Query: 122
DIIIMGSHGKTNLKEILLGSVTENVIKKSNK
PVLVVKR 159
D+I M SH + + E+LLGSVT V+K S PVLV ++
Sbjct: 123
DLIAMASHRRRSFIEMLLGSVTAKVLKNSKI
PVLVYRQ 160



```
Sequences producing significant alignments:                          (bits) Value

gi|15668757|ref|NP_247556.1|  (NC_000909) conserved hypothet...   161   2e-39
gi|14590690|ref|NP_142758.1|  (NC_000961) hypothetical prote...    66   1e-10
gi|15791176|ref|NP_281000.1|  (NC_002607) Vng2386c [Halobact...    59   1e-08
gi|15790518|ref|NP_280342.1|  (NC_002607) Vng1536c [Halobact...    59   2e-08
gi|15790787|ref|NP_280611.1|  (NC_002607) Vng1898c [Halobact...    57   5e-08
gi|15668711|ref|NP_247510.1|  (NC_000909) conserved hypothet...    57   6e-08
gi|15790505|ref|NP_280329.1|  (NC_002607) Vng1518h [Halobact...    54   4e-07
gi|15679076|ref|NP_276193.1|  (NC_000916) conserved protein ...    54   6e-07
gi|17544898|ref|NP_518300.1|  (NC_003295) CONSERVED HYPOTHET...    54   7e-07
gi|15678918|ref|NP_276035.1|  (NC_000916) conserved protein ...    53   1e-06
gi|15679011|ref|NP_276128.1|  (NC_000916) conserved protein ...    52   2e-06
gi|17934409|ref|NP_531199.1|  (NC_003304) conserved hypothet...    52   3e-06
gi|15887843|ref|NP_353524.1|  (NC_003062) AGR_C_878p [Agroba...    51   3e-06
gi|12325313|gb|AAG52594.1|AC016447_3  (AC016447) unknown pro...    51   4e-06
gi|15677353|ref|NP_274508.1|  (NC_003112) conserved hypothet...    50   6e-06
gi|15794596|ref|NP_284418.1|  (NC_003116) conserved hypothet...    50   6e-06
gi|15678181|ref|NP_275296.1|  (NC_000916) conserved protein ...    50   6e-06
gi|16120145|ref|NP_395733.1|  (NC_002608) Vng6205c [Halobact...    50   7e-06
gi|11499518|ref|NP_070759.1|  (NC_000917) conserved hypothet...    50   1e-05
gi|16330107|ref|NP_440835.1|  (NC_000911) unknown protein [S...    49   2e-05
gi|14495190|dbj|BAB60909.1|  (AP003213) contains ESTs AU1013...    48   3e-05
gi|17229082|ref|NP_485630.1|  (NC_003272) hypothetical prote...    47   5e-05
gi|17546078|ref|NP_519480.1|  (NC_003295) CONSERVED HYPOTHET...    47   6e-05
gi|7262999|gb|AAF44047.1|AF206717_1  (AF206717) hypothetical...    47   7e-05
gi|18312103|ref|NP_558770.1|  (NC_003364) conserved protein ...    47   8e-05
gi|15807126|ref|NP_295855.1|  (NC_001263) hypothetical prote...    47   8e-05
gi|2208981|emb|CAA73748.1|  (Y13308) hypothetical protein [Y...    47   9e-05
gi|17546223|ref|NP_519625.1|  (NC_003295) CONSERVED HYPOTHET...    46   1e-04
gi|15598505|ref|NP_251999.1|  (NC_002516) conserved hypothet...    46   1e-04
gi|7636051|emb|CAB88409.1|  (AJ288984) hypothetical protein ...    46   1e-04
gi|17544903|ref|NP_518305.1|  (NC_003295) CONSERVED HYPOTHET...    46   1e-04
gi|6714413|gb|AAF26101.1|AC012328_4  (AC012328) unknown prot...    45   2e-04
gi|16080976|ref|NP_391804.1|  (NC_000964) similar to hypothe...    45   2e-04
gi|17988589|ref|NP_541222.1|  (NC_003318) Universal stress p...    45   2e-04
```

Fig. 2. Resultado de busca: seqüências com expect menor que o limite estipulado.

```
                                                Alignments

>gi|15668757|ref|NP_247556.1| (NC_000909) conserved hypothetical protein [Methanococcus
          jannaschii]
 gi|2501594|sp|Q57997|Y577_METJA Protein MJ0577
 gi|2128018|pir||A64372 hypothetical protein homolog MJ0577 - Methanococcus jannaschii
 gi|5107802|pdb|1MJH|B Chain B, Structure-Based Assignment Of The Biochemical Function Of
          Hypothetical Protein MjO577: A Test Case Of Structural
          Genomics
 gi|5107801|pdb|1MJH|A Chain A, Structure-Based Assignment Of The Biochemical Function Of
          Hypothetical Protein MjO577: A Test Case Of Structural
          Genomics
 gi|1591284|gb|AAB98568.1| (U67506) conserved hypothetical protein [Methanococcus jannaschii]
          Length = 162

 Score =  161 bits (407), Expect = 2e-39
 Identities = 110/162 (67%), Positives = 110/162 (67%)

Query: 1   MSVMYKKILYPTDFSETAEIALKHVKAFKTLKAEEVILLHVIDEREIKKRDIFXXXXXXX 60
           MSVMYKKILYPTDFSETAEIALKHVKAFKTLKAEEVILLHVIDEREIKKRDIF
Sbjct: 1   MSVMYKKILYPTDFSETAEIALKHVKAFKTLKAEEVILLHVIDEREIKKRDIFSLLLGVA 60

Query: 61  XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXGIPHEEIVKIAEDEG 120
                                                        GIPHEEIVKIAEDEG
Sbjct: 61  GLNKSVEEFENELKNKLTEEAKNKMENIKKELEDVGFKVKDIIVVGIPHEEIVKIAEDEG 120

Query: 121 VDIIIMGSHGKTNLKEILLGSVTENVIKKSNKPVLVVKRKNS 162
           VDIIIMGSHGKTNLKEILLGSVTENVIKKSNKPVLVVKRKNS
Sbjct: 121 VDIIIMGSHGKTNLKEILLGSVTENVIKKSNKPVLVVKRKNS 162

>gi|14590690|ref|NP_142758.1| (NC_000961) hypothetical protein [Pyrococcus horikoshii]
 gi|7444946|pir||B71132 hypothetical protein PH0823 - Pyrococcus horikoshii
 gi|3257233|dbj|BAA29916.1| (AP000003) 170aa long hypothetical protein [Pyrococcus horikoshii]
          Length = 170
```

Fig. 3. Resultado de busca: alinhamentos.

Caso esta entrada não estivesse na base de dados, ou para reforçar a homologia detectada, o exame dos alinhamentos deveria prosseguir até o nível de similaridade (score e Expect) que se julgasse adequado. Neste caso, encontraríamos as seguintes seqüências anotadas (até o limiar onde seguramente homologias podem ser detectadas – Expect ≈ 0.1), que reforçam a suposição de que a seqüência sob análise pertence à família USP:

>gi|17988589|ref|NP_541222.1| (NC_003318)
Universal stress protein family [Brucella melitensis]
gi|17984389|gb|AAL53486.1| (AE009663) Universal stress protein family [Brucella melitensis]
Length = 148
Score = 45.4 bits (106), Expect = 2e-04
Identities = 20/51 (39%), Positives = 34/51 (66%)

>gi|13541477|ref|NP_111165.1| (NC_002689)
Nucleotide-binding protein (UspA-related) [Thermoplasma
volcanium]
gi|14324861|dbj|BAB59787.1| (AP000993)
hypothetical protein [Thermoplasma volcanium]
Length = 150
Score = 44.3 bits (103), Expect = 4e-04
Identities = 24/44 (54%), Positives = 33/44 (74%)

>gi|17229756|ref|NP_486304.1| (NC_003272) Na+/H+-exchanging protein [Nostoc sp. PCC 7120]
gi|17131355|dbj|BAB73963.1| (AP003588) Na+/H+-exchanging protein [Nostoc sp. PCC 7120]
Length = 543
Score = 41.2 bits (95), Expect = 0.004
Identities = 19/52 (36%), Positives = 34/52 (64%)

>gi|18313304|ref|NP_559971.1| (NC_003364)
universal stress protein family [Pyrobaculum
aerophilum]
gi|18160828|gb|AAL64153.1| (AE009873) universal stress protein family [Pyrobaculum aerophilum]
Length = 137
Score = 39.3 bits (90), Expect = 0.013
Identities = 36/152 (23%), Positives = 59/152 (38%), Gaps = 18/152 (11%)

>gi|17986386|ref|NP_539020.1| (NC_003317)
Universal stress protein family [Brucella melitensis]
gi|17981977|gb|AAL51284.1| (AE009453) Universal stress protein family [Brucella melitensis]
Length = 149
Score = 39.3 bits (90), Expect = 0.014
Identities = 19/47 (40%), Positives = 31/47 (65%)

>gi|13541547|ref|NP_111235.1| (NC_002689)
Nucleotide-binding protein (UspA-related) [Thermoplasma
volcanium]
gi|14324932|dbj|BAB59858.1| (AP000993)
hypothetical protein [Thermoplasma volcanium]
Length = 142
Score = 38.9 bits (89), Expect = 0.017
Identities = 18/46 (39%), Positives = 31/46 (67%)

>gi|15608774|ref|NP_216152.1| (NC_000962)
hypothetical protein Rv1636 [Mycobacterium tuberculosis
H37Rv]
gi|15841091|ref|NP_336128.1| (NC_002755)
universal stress protein family [Mycobacterium
tuberculosis CDC1551]
gi|7444951|pir||B70560 hypothetical protein Rv1636 - Mycobacterium tuberculosis (strain
H37RV)
gi|2113920|emb|CAB08889.1| (Z95554) hypothetical protein Rv1636 [Mycobacterium tuberculosis
H37Rv]
gi|13881306|gb|AAK45942.1| (AE007031) universal stress protein family [Mycobacterium
tuberculosis CDC1551]
Length = 146
Score = 37.0 bits (84), Expect = 0.062
Identities = 18/52 (34%), Positives = 34/52 (64%)

>gi|16122525|ref|NP_405838.1| (NC_003143)
putative stress protein [Yersinia pestis]
gi|15980297|emb|CAC91106.1| (AJ414151) putative stress protein [Yersinia pestis]
Length = 318
Score = 36.6 bits (83), Expect = 0.078
Identities = 17/53 (32%), Positives = 30/53 (56%)

>gi|2507515|sp|P44195|YDAA_HAEIN Protein HI1426
Length = 309
Score = 36.2 bits (82), Expect = 0.11
Identities = 19/53 (35%), Positives = 34/53 (63%)

Para fins de ilustração, a Fig. 4 apresenta o resultado da busca utilizando a mesma seqüência query, mas utilizando a matriz PAM70. O Resultado é essencialmente o mesmo e a hipótese final criada seria a mesma. Entretanto, observa-se que as seqüências apresentam um score menor e as similaridades encontradas uma extensão menor, o que indica que a utilização desta matriz apresenta uma sensibilidade menor, o que pode prejudicar a detecção de similaridades mais fracas. As matrizes PAM foram utilizadas como default por BLAST por muito tempo, mas atualmente a matriz recomendada como default é a matriz BLOSUM62. Caso nenhum resultado significativo ser obtido, recomenda-se a utilização da matriz BLOSUM30.

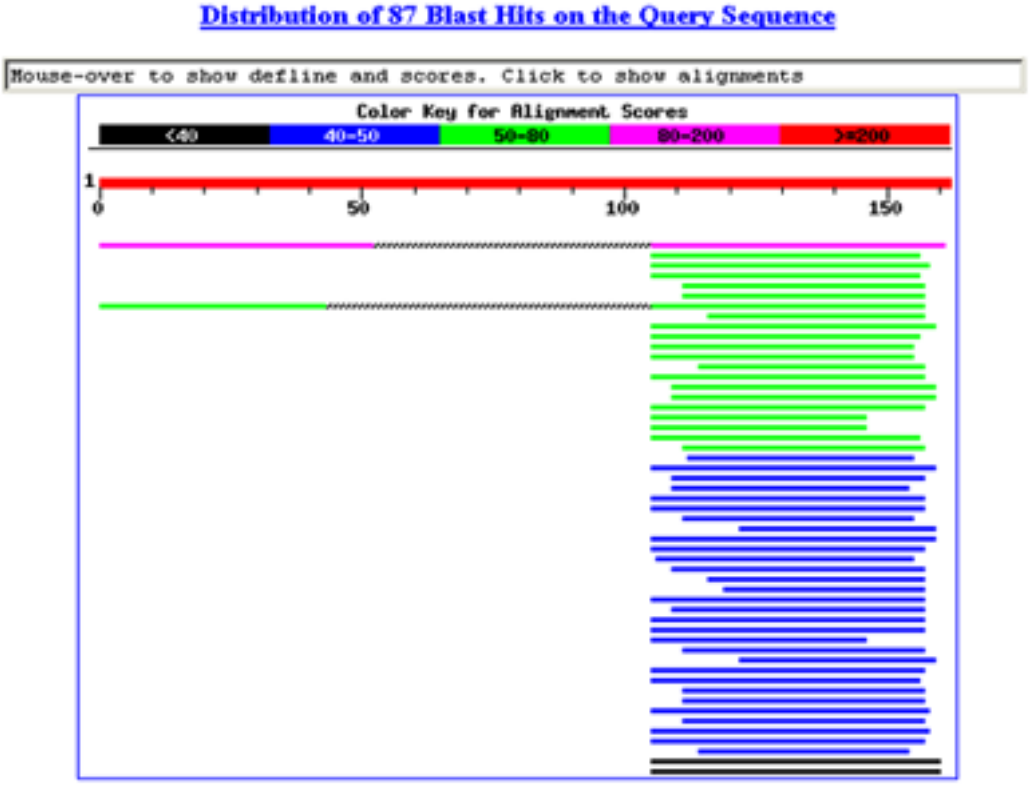


Fig. 4. Visualização gráfica do resultado de busca utilizando matriz PAM70.

Comentários Finais

Nesta instrução foram apresentados conceitos e resultados teóricos que suportam BLAST e sua utilização foi exemplificada através do serviço mantido pelo NCBI.

BLAST constitui-se, hoje, numa ferramenta de fundamental importância para biologistas moleculares, pois permite que, no estudo de uma seqüência, seqüências potencialmente homólogas sejam encontradas, fornecendo ainda medidas estatísticas para a avaliação da significância da similaridade detectada.

Entretanto, a decisão sobre se uma similaridade detectada representa uma homologia passa obrigatoriamente pela interpretação biológica do alinhamento obtido por BLAST. Assim, BLAST por si só não é suficiente para revelar uma homologia, mas constitui-se num primeiro passo fundamental neste sentido. Porisso a importância da correta interpretação de seus parâmetros e resultados.

Referências Bibliográficas

ALTSCHUL, S. F.; BOGUSKI, M. S.; GISH, W.; WOOTTON, J. C. Issues in searching molecular sequence databases. **Nature Genetics**, v. 6, p. 119-129, 1994.

ALTSCHUL, S. F.; GISH, W.; MILLER, W.; MYERS, E. W.; LIPMAN, D. J. Basic Local Alignment Search Tool. **Journal of Molecular Biology**, v. 215, p. 403-410, 1990.

ALTSCHUL, S. F.; MADDEN, T. L.; SCHAFFER, A. A.; ZHANG, J.; ZHANG, Z.; MILLER, W.; LIPMAN, D. L. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. **Nucleic Acid Research**, v. 25, n. 17, p. 3389-3402, 1997.

DURBIN, R.; EDDY, S.; KROGH, A.; MITCHISON, G. **Biological sequence analysis**: probabilistic models of proteins and nucleic acids. Cambridge, UK: CambridgeUniversity Press, 1998. 356 p.

EUROPEAN MOLECULAR BIOLOGY LABORATORY. **EMBL – European Molecular Biology Laboratory**. Disponível em: <http://www.embl-heidelberg.de>. Acesso em: 20 dez. 2001.

EWENS, W. J.; GRANT, G. R. **Statistical methods in bioinformatics**: an introduction. New York: Springer, 2001. 476 p. (Statistics for Biology and Health).

LUPAS, A. Current opinion on structural biology. **Current Opinion on Structural Biology**, v. 7, n. 3, p. 388-393, 1997.

MUNICH INFORMATION CENTER FOR PROTEIN SEQUENCES. **The PIR – International Protein Sequence Database**. Disponível em: <http://www.mips.biochem.mpg.de/proj/protseqdb>. Acesso em: 20 dez. 2001.

THE NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION. **NCBI BLAST home page**. Disponível em: <http://www.ncbi.nlm.nih.gov/BLAST>. Acesso em: 20 dez. 2001a.

THE NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION. **NCBI homepage**. Disponível em: <http://www.ncbi.nlm.nih.gov>. Acesso em: 20 dez. 2001b.

THE NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION. **The statistics of sequence similarity scores**. Disponível em: <http://www.ncbi.nlm.nih.gov/BLAST/tutorial/ Altschul-1.html>. Acesso em: 20 dez. 2001c.

SETUBAL, J.C.; MEIDANIS, J. **Introduction to computational molecular biology**. Boston: PWS Publishing, 1997. 296 p.

SWISS INSTITUTE OF BIOINFORMATICS. **ExPASy - SWISS-PROT and TrEMBL**. Disponível em: <http://ca.expasy.org/sprot>. Acesso em: 20 dez. 2001.

Instruções Técnicas, 6

MINISTÉRIO DA AGRICULTURA, PECUÁRIA E ABASTECIMENTO

GOVERNO FEDERAL
Trabalhando em todo o Brasil

Embrapa Informática Agropecuária
Área de Comunicação e Negócios
Av. Dr. André Tosello s/nº
Cidade Universitária - "Zeferino Vaz"
Barão Geraldo - Caixa Postal 6041
13083-970 - Campinas, SP
Telefone/Fax: (19) 3789-5743
E-mail: sac@cnptia.embrapa.br

1ª edição