# BLAST

## FERRAMENTA BÁSICA DE LOCALIZAÇÃO DE ALINHAMENTO LOCAL

O algoritmo BLAST e o programa computacional que o implementa foram desenvolvidos, na década de 1980, nos Estados Unidos ( ALTSCHUL et al. 1990)

www.ncbi.nlm.nih.gov

NCBI Resources How To Sign in to NCBI

NCBI
National Center for Biotechnology Information

All Databases Search

- NCBI Home
- Resource List (A-Z)
- All Resources
- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature
- Proteins
- Sequence Analysis
- Taxonomy
- Training & Tutorials
- Variation

## Welcome to NCBI

The National Center for Biotechnology Information advances sc
biomedical and genomic information.

About the NCBI | Mission | Organization | NCBI News | Blog

### Submit

Deposit data or manuscripts into NCBI databases

### Download

Transfer NCBI data to y computer

### Develop

Use NCBI APIs and code libraries to build applications

### Analyze

Identify an NCBI tool fo data analysis task

## Popular Resources

- PubMed
- Bookshelf
- PubMed Central
- PubMed Health
- BLAST
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem

s

refined

23 Aug 2016

+ provides a
lignment of

NLM: EDirect -
o NCBI's

22 Aug 2016

NCBI staff will

HIV-1 datasets in Gene updated

# Basic Local Alignment Search Tool

**BLAST** finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. Learn more

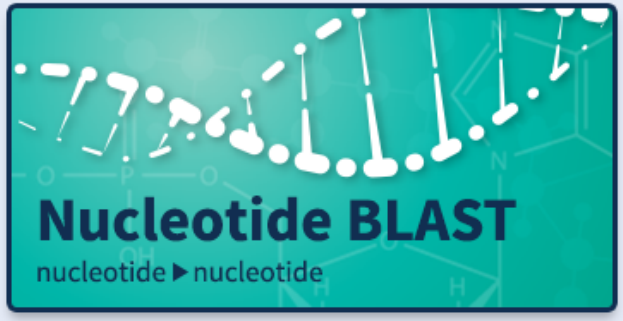
NEWS

**BLAST and HTTPS**

The BLAST URL API is moving to HTTPS.
Thu, 28 Jul 2016 12:00:00 EST

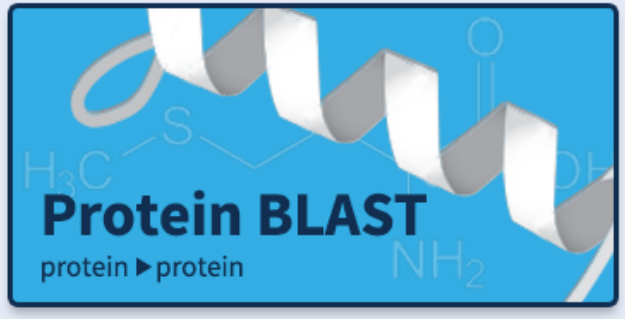
More BLAST news...

## Web BLAST




**blastx**
translated nucleotide ▶ protein

**tblastn**
protein ▶ translated nucleotide




## BLAST Genomes

*Enter organism common name, scientific name, or tax id* **Search**

Human Mouse Rat Microbes

**Tabela 2**. Simplificação das diferentes comparações possíveis com o uso do BLAST.

| Seqüência de entrada ("query") | Banco de dados ("subject") | |
|---|---|---|
| a.a. | a.a. | **Blastp** |
| nt | nt | **Blastn** |
| nt* | a.a. | **Blastx** |
| a.a. | nt* | **tBlastn** |
| nt* | nt* | **tBlastx** |

a.a. = aminoácido; nt = nucleotídeo; * = traduzidos para todas as seqüências (fases, "frames") de leitura possíveis

*Embrapa Recursos Genéticos e Biotecnologia*
Brasília, DF
2007

A utilização do BLAST é vasta e a ferramenta definida pelo objetivo do investigador:

- Comparar uma sequência gerada a outras sequências descritas em um banco de dados
- Avaliar a homologia de um primer
- Inferir relações funcionais e evolutivas entre sequências , bem como ajudar a identificar os membros de famílias de genes

# Standard Nucleotide BLAST

blastn | blastp | blastx | tblastn | tblastx

BLASTN programs search nucleotide databases using a nucleotide query. more...

Reset page | Bookmark

## Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

Clear

Query subrange

From

To

Or, upload file

Escolher arquivo | Nenhum arquivo selecionado

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

## Choose Search Set

Database

Human genomic + transcript | Mouse genomic + transcript | Others (nr etc.):

Nucleotide collection (nr/nt)

Organism
Optional

Enter organism name or id--completions will be suggested | Exclude | +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Exclude
Optional

Models (XM/XP) | Uncultured/environmental sample sequences

Limit to
Optional

Sequences from type material

Entrez Query
Optional

YouTube Create custom database

Enter an Entrez query to limit search

## Program Selection

Optimize for

- Highly similar sequences (megablast)
- More dissimilar sequences (discontiguous megablast)
- Somewhat similar sequences (blastn)

Choose a BLAST algorithm

BLAST

Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)

Show results in a new window

Algorithm parameters

# FASTA

- Formato baseado em texto para representar tanto peptídeos quanto nucleotídeos, que são representados usando códigos de uma única letra·
- Pode-se utilizar identificador e descrição, que devem vir após o símbolo ">"
- Recomenda-se que todas as linhas do texto sejam mais curtas do que 80 caracteres.
- ">“ pode ser utilizado para se inserir mais de uma sequência

```
>gi|37221963|gb|AAN78258.1| resistance protein [Arachis simpsonii]
LAKALYNSICDRFECACFLFNVRTISDQEEGLVRLQQTLLSKLLGEWEIKVRSVEEGISMIKEKLSKKRA
LIVLDDVNKIEQLKALAGECDWFSYGTRIVITRDKYLLTAHKVEKIYKMKLLSDPESLELFCWNAFKISR
PKENYEDLSNQAIHYAQ
```

# CXCL12

```
ORIGIN
        1 gccgcacttt cactctccgt cagccgcatt gcccgctcgg cgtccggccc ccgacccgcg
       61 ctcgtccgcc cgcccgcccg cccgcccgcg ccatgaacgc caaggtcgtg gtcgtgctgg
      121 tcctcgtgct gaccgcgctc tgcctcagcg acggtaagtg cgctcggcgg ggaggccctg
      181 gcgaggctct gggcttcggc tcccgcccgg cgcagagccg cggctcctct gcctgcgccc
      241 gcagtttgcg ccgcccgaca cagtgggggt ggaggcagct cgcttagctg ggcgctctgg
      301 tgcggtgtcg cgttcaaagc cctacttttg cgccgagggt ttgtgacctg cagagagcga
      361 ccgcctgacc tcaagctggc tgcagagcga ggtcagccgg gacaactggg caggaacgcc
      421 ccaagagagc gttcgggact ggcgcgggaa ggcgcgaggg tggggagccg cggaccccag
      481 acccctgcct gcccaccctc cacccactgc cttcccgctg cgggctcggt ttccacaggc
      541 gaatgggctc cgaggtcctg ctgcgcacag tgctggaggt cgggcggtgc atggcgagag
      601 cgcgctttgc aaagttcgcc tcatgccgcg acttgaggct gtgaggtccg atgcagccgc
      661 gcagcctccg ctccctgcga agccgatcct ctccccacac cctcgcgggg ctcttggcca
      721 ggcgcgggcg ggctgcgcgg cgcagctgtc gggcttgtgc cacccgccag gccgcgcttc
      781 tgcacagctc cgcggtccgc agcgacgcgg acctgggcgc gcgtccagcc gctgtcttcc
      841 tgtctcttct cgggttcacc tgagagaggc ccagggaccc tccgagcctc acacgccccc
      901 tccccagctc cccaaacaca cctgcagaca ggctcttttg tcaccagctg gccaagcgct
      961 ttgccccaga atgagagctg cagcaaagtt caatctccaa ggccctgggc cactgggaac
     1021 cgtggtgtcc cgtttcactg ggagagggct tttttctttt ctttcgctga agagaaactc
     1081 gctgcgctgt cacaggacac cgatgcaaac caaaataaac tgttgcctct gaacacacaa
     1141 aacaaaaccc tgtggctcag gccgctgagc ttccttggct gtgaagtttc ctgcaaagag
     1201 ttgaaggaca ctctgagtgc tcttgcagcc gcgggctgtg caggcccagc gcgcctcgag
     1261 cccttggctg ggttggtgca gacgtggccc cggcgtccta gctggcttgg actgtgcgct
     1321 ggtgtgagca ggaccagcgc agctgagtcg ggctggagaa gctgggcagc ctggggtcgt
     1381 ggcggactgc ctaagtgtta aggacacaga ctgaaggcaa ttgggaatgt tcgtcgttga
     1441 tgaaaagtcc ctgaggctgt gtctgcacac tgaccccccc cccccccgc aagtcatggg
```

Enter Query Sequence

BLASTN programs search nucleotide databases using a nucleotide query. more...

Reset page Bookmark

Enter accession number(s), gi(s), or FASTA sequence(s)

Clear

Query subrange

From

To

gccgcacttt cactctccgt cagccgcatt gcccgctcgg cgtccggccc ccgacccgcg

Or, upload file

Escolher arquivo Nenhum arquivo selecionado

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

## Choose Search Set

**Database** ○ Human genomic + transcript ○ Mouse genomic + transcript ◉ Others (nr etc.):

Nucleotide collection (nr/nt) ▾

**Organism** Optional

Enter organism name or id--completions will be suggested ☐ Exclude +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

**Exclude** Optional ☐ Models (XM/XP) ☐ Uncultured/environmental sample sequences

**Limit to** Optional ☐ Sequences from type material

**Entrez Query** Optional

YouTube Create custom database

Enter an Entrez query to limit search

## Program Selection

**Optimize for**

◉ Highly similar sequences (megablast)

○ More dissimilar sequences (discontiguous megablast)

○ Somewhat similar sequences (blastn)

Choose a BLAST algorithm

**BLAST** Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)

☐ Show results in a new window

⊕ Algorithm parameters

# Nucleotide Sequence (60 letters)

**RID** VU0AGG1701R (Expires on 08-25 10:07 am)

**Query ID** lcl|Query_116553
**Description** None
**Molecule type** nucleic acid
**Query Length** 60

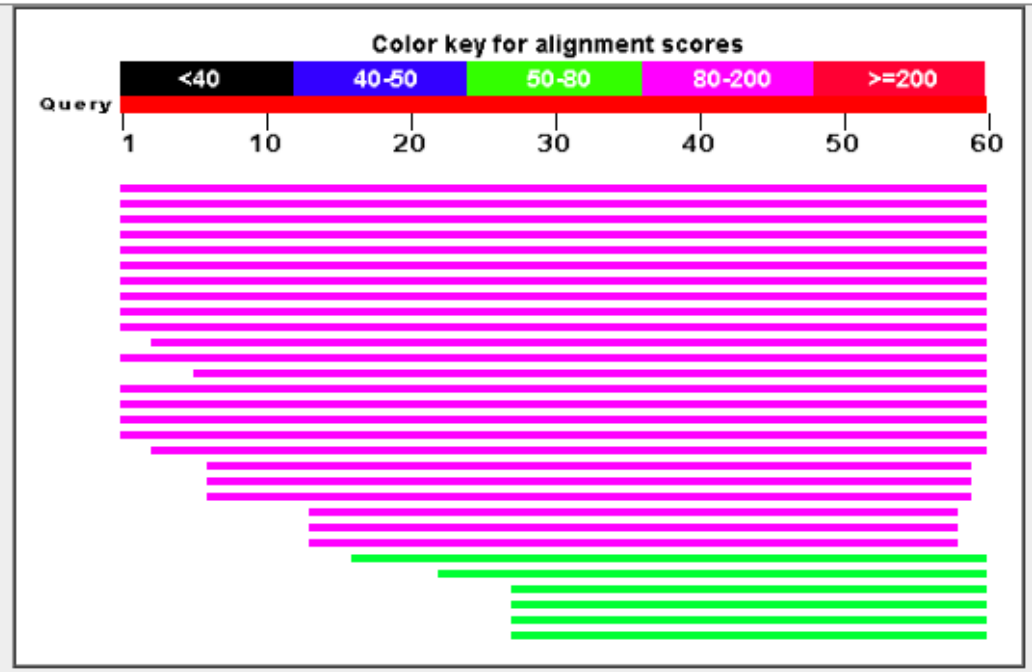
**Database Name** nr
**Description** Nucleotide collection (nt)
**Program** BLASTN 2.5.0+ ▸Citation

Other reports: ▸Search Summary [Taxonomy reports] [Distance tree of results]

## Graphic Summary

Distribution of 30 Blast Hits on the Query Sequence

Mouse-over to show defline and scores, click to show alignments

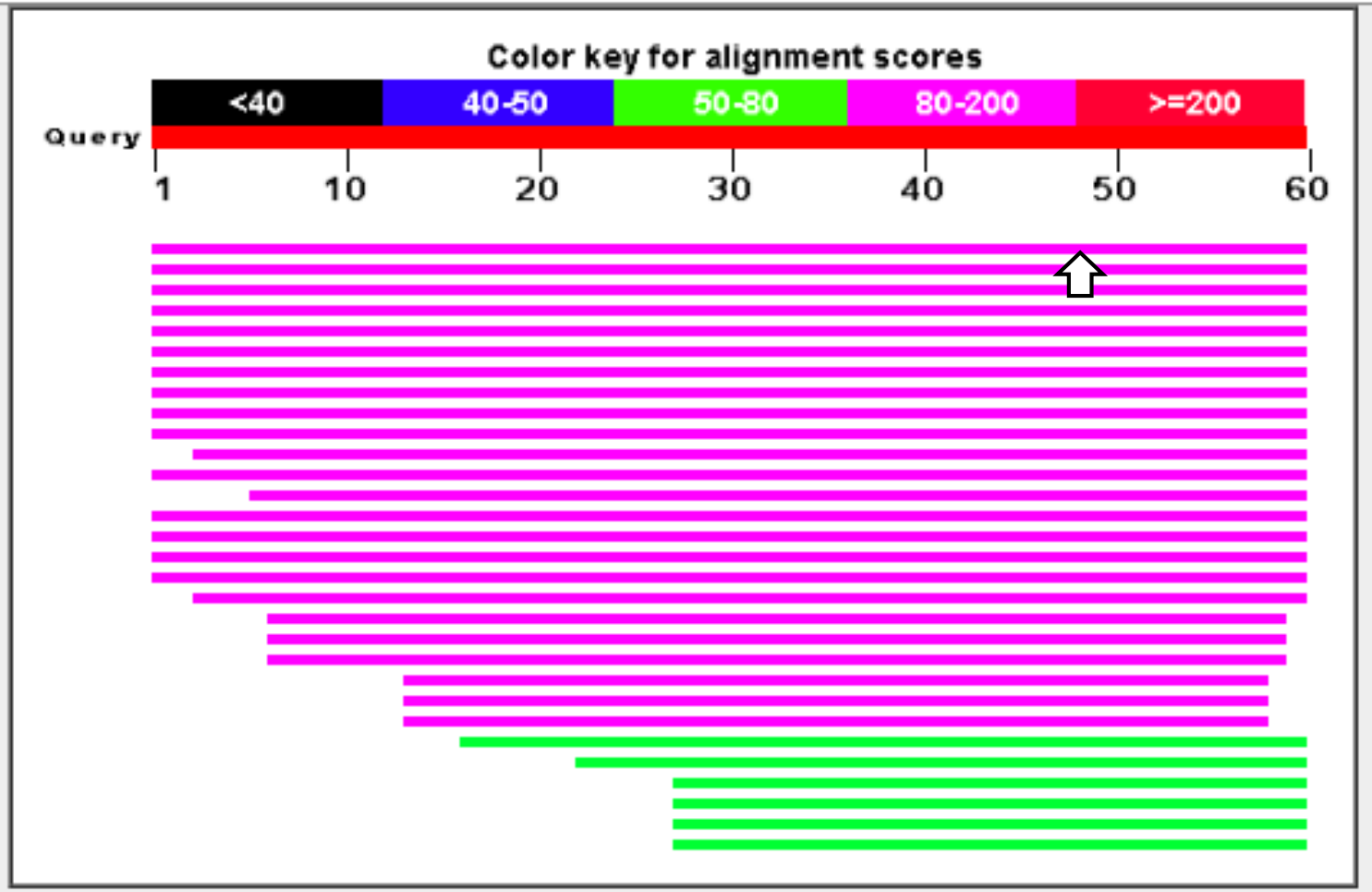


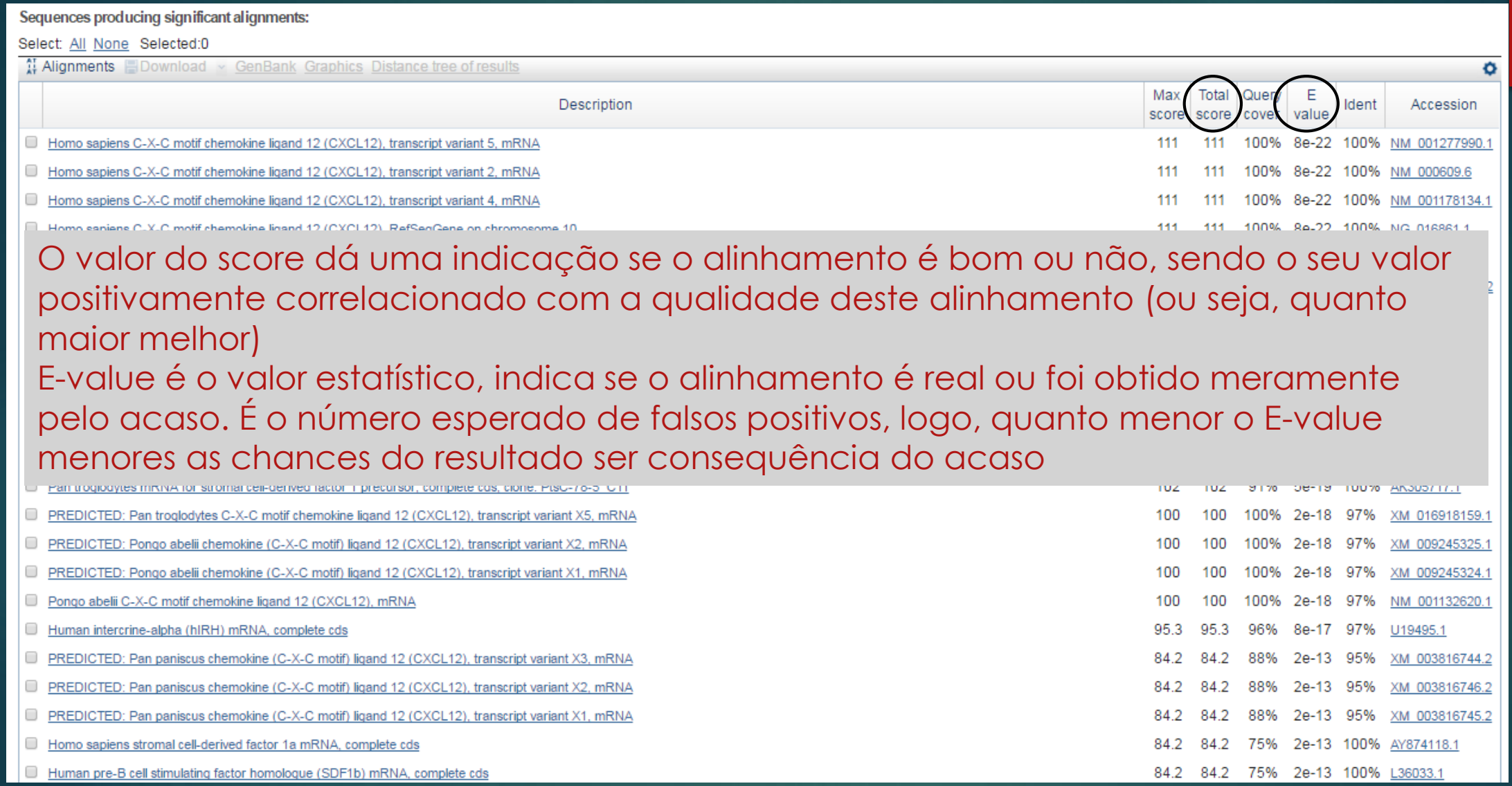
- Query – Sequência de entrada
- Subject – Banco de dados

- nr – Arquivos não redundantes
- Identificador -  as sequencias depositadas no NCBI recebem uma identificação que pode fornecer informações relevantes quanto sua origem e natureza

## Distribution of 30 Blast Hits on the Query Sequence

NM_001277990 Homo sapiens C-X-C motif chemokine ligand 12 (CXCL12), .. S=111 E=7.6e-22





Sequences producing significant alignments:

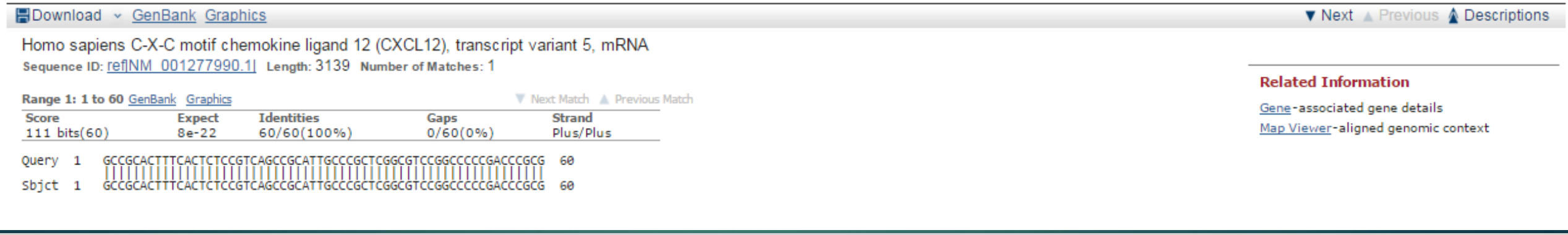
Select: All None Selected:0

Alignments Download GenBank Graphics Distance tree of results

| Description | Max score | Total score | Query cover | E value | Ident | Accession |
|---|---|---|---|---|---|---|
| Homo sapiens C-X-C motif chemokine ligand 12 (CXCL12), transcript variant 5, mRNA | 111 | 111 | 100% | 8e-22 | 100% | NM_001277990.1 |
| Homo sapiens C-X-C motif chemokine ligand 12 (CXCL12), transcript variant 2, mRNA | 111 | 111 | 100% | 8e-22 | 100% | NM_000609.6 |
| Homo sapiens C-X-C motif chemokine ligand 12 (CXCL12), transcript variant 4, mRNA | 111 | 111 | 100% | 8e-22 | 100% | NM_001178134.1 |
| Homo sapiens C-X-C motif chemokine ligand 12 (CXCL12), RefSeqGene on chromosome 10 | 111 | 111 | 100% | 8e-22 | 100% | NG_016861.1 |
| Pan troglodytes mRNA for stromal cell-derived factor 1 precursor, complete cds, clone: PtSC-78-5' CTT | 102 | 102 | 91% | 3e-19 | 100% | AK305717.1 |
| PREDICTED: Pan troglodytes C-X-C motif chemokine ligand 12 (CXCL12), transcript variant X5, mRNA | 100 | 100 | 100% | 2e-18 | 97% | XM_016918159.1 |
| PREDICTED: Pongo abelii chemokine (C-X-C motif) ligand 12 (CXCL12), transcript variant X2, mRNA | 100 | 100 | 100% | 2e-18 | 97% | XM_009245325.1 |
| PREDICTED: Pongo abelii chemokine (C-X-C motif) ligand 12 (CXCL12), transcript variant X1, mRNA | 100 | 100 | 100% | 2e-18 | 97% | XM_009245324.1 |
| Pongo abelii C-X-C motif chemokine ligand 12 (CXCL12), mRNA | 100 | 100 | 100% | 2e-18 | 97% | NM_001132620.1 |
| Human intercrine-alpha (hIRH) mRNA, complete cds | 95.3 | 95.3 | 96% | 8e-17 | 97% | U19495.1 |
| PREDICTED: Pan paniscus chemokine (C-X-C motif) ligand 12 (CXCL12), transcript variant X3, mRNA | 84.2 | 84.2 | 88% | 2e-13 | 95% | XM_003816744.2 |
| PREDICTED: Pan paniscus chemokine (C-X-C motif) ligand 12 (CXCL12), transcript variant X2, mRNA | 84.2 | 84.2 | 88% | 2e-13 | 95% | XM_003816746.2 |
| PREDICTED: Pan paniscus chemokine (C-X-C motif) ligand 12 (CXCL12), transcript variant X1, mRNA | 84.2 | 84.2 | 88% | 2e-13 | 95% | XM_003816745.2 |
| Homo sapiens stromal cell-derived factor 1a mRNA, complete cds | 84.2 | 84.2 | 75% | 2e-13 | 100% | AY874118.1 |
| Human pre-B cell stimulating factor homologue (SDF1b) mRNA, complete cds | 84.2 | 84.2 | 75% | 2e-13 | 100% | L36033.1 |

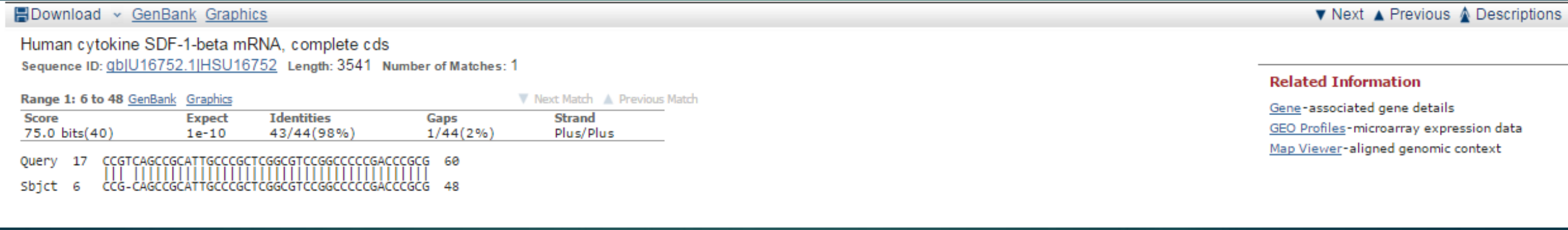
O valor do score dá uma indicação se o alinhamento é bom ou não, sendo o seu valor positivamente correlacionado com a qualidade deste alinhamento (ou seja, quanto maior melhor)

E-value é o valor estatístico, indica se o alinhamento é real ou foi obtido meramente pelo acaso. É o número esperado de falsos positivos, logo, quanto menor o E-value menores as chances do resultado ser consequência do acaso




Identidade é o número de matchs, ou seja de resíduos similares.

Gap é o espaço introduzido para compensação em um alinhamento. A medida em que os gaps aumentam para acomodar o pareamento entre as sequências o score diminui.



Faculdade de Odontologia da Universidade Federal de Minas Gerais

# Avaliação de expressão das quimiocinas CXCL10, CCL5 e CCL2 e dos receptores CCR5, CCR1 e CXCR2 em um modelo de infecção experimental por *F. nucleatum* e *E. faecalis* em camundongos isentos de germes

**Apresentação realizada como requisito parcial para obtenção de título de Mestre em Odontologia.**

Área de concentração: Endodontia

**Orientação:** Prof. Dr. Antônio Paulino Ribeiro Sobrinho

**Co-orientação:** Prof.Dra. Leda Quercia Vieira

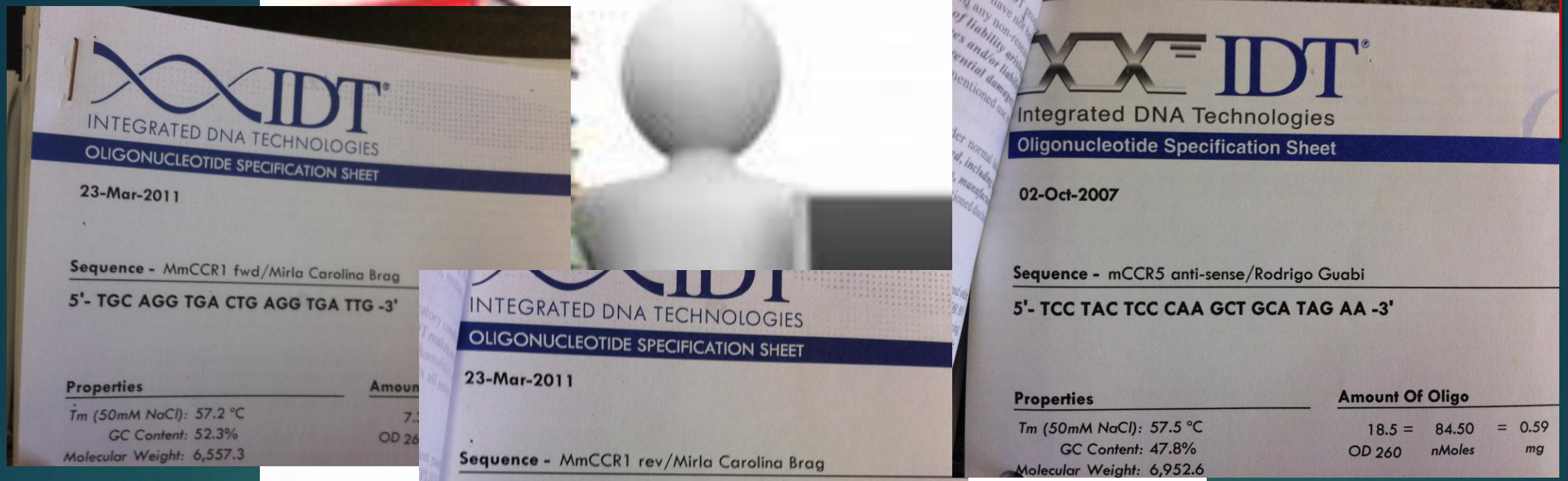
**Aluna :** Marcela Marçal Thebit

## Tabela

## Tabela I: Sequências dos primers utilizados

| | | |
|---|---|---|
| CXCL10 | 5’-CTC GCA AGG ACG GTC CGC TG-3’ | 5’-CTC GCA AGG ACG GTC CGC TG-3’ |
| CCL2 | 5’-AGG AAG ATC TCA GTG CAG AG-3’ | 5’-AGT CTT CGG AGT TTG CCT TTG-3’ |
| CCL5 | 5’-CGT GCC CAC ATC AAG GAG TA-3’ | 5’-CAC ACA CTT GGC GGT TCT TTC-3’ |
| CXCR2 | 5'-AGT GCC TGC CTC AAT GTC TCC A-3’ | 5’-CCA GGA GCA AGG ACA GAC CCC-3' |
| CCR5 | 5’-CAA GAC ATT CCT GAT CGT GCA A-3’ | 5’-TCC TAC CAA GCT GCA TAG AA-3’ |
| CCR1 | 5’-TGC AGG TGA CTG AGG TGA TTG-3’ | 5’-TGA AAC AGC TGC CGA AGG TAC-3’’’ |


LONGINES


IDT
INTEGRATED DNA TECHNOLOGIES
OLIGONUCLEOTIDE SPECIFICATION SHEET
23-Mar-2011
Sequence - MmCCR1 fwd/Mirla Carolina Brag
5'- TGC AGG TGA CTG AGG TGA TTG -3'
Properties
Tm (50mM NaCl): 57.2 °C
GC Content: 52.3%
Molecular Weight: 6,557.3
IDT
INTEGRATED DNA TECHNOLOGIES
OLIGONUCLEOTIDE SPECIFICATION SHEET
23-Mar-2011
Sequence - MmCCR1 rev/Mirla Carolina Brag
5'- TGA AAC AGC TGC CGA AGG TAC -3'
Properties
Tm (50mM NaCl): 57.6 °C
GC Content: 52.3%
Molecular Weight: 6,464.3
nmoles/OD260: 4.8
ug/OD260: 31.0
Amount Of Oligo
7.3 = 35.1
OD 260 nMoles
70
IDT
Integrated DNA Technologies
Oligonucleotide Specification Sheet
02-Oct-2007
Sequence - mCCR5 anti-sense/Rodrigo Guabi
5'- TCC TAC TCC CAA GCT GCA TAG AA -3'
Properties
Tm (50mM NaCl): 57.5 °C
GC Content: 47.8%
Molecular Weight: 6,952.6
Amount Of Oligo
18.5 = 84.50 = 0.59
OD 260 nMoles mg

# Obrigada

Documentos

ISSN 0102-0110
Dezembro, 2007

224

# O programa *BLAST*: guia prático de utilização

Embrapa

ISSN 0102 0110
Dezembro, 2007

**Empresa Brasileira de Pesquisa Agropecuária**
**Embrapa Recursos Genéticos e Biotecnologia**
**Ministério da Agricultura, Pecuária e Abastecimento**

*Documentos* 224

# O programa *BLAST*: guia prático de utilização

*Embrapa Recursos Genéticos e Biotecnologia*
**Brasília, DF**
**2007**

Exemplares desta edição podem ser adquiridos na

Embrapa Recursos Genéticos e Biotecnologia
Serviço de Atendimento ao Cidadão
Parque Estação Biológica, Av. W/5 Norte (Final) –
Brasília, DF CEP 70770-900 – Caixa Postal 02372 PABX: (61) 448-4600 Fax: (61) 340-3624
http://www.cenargen.embrapa.br
e.mail:sac@cenargen.embrapa.br



1ª edição
1ª impressão (2007):



**Dados Internacionais de Catalogação na Publicação (CIP)**
**Embrapa Recursos Genéticos e Biotecnologia**

P 965 O programa *BLAST*: guia prático de utilização Embrapa Recursos Genéticos e Biotecnologia / Alexandre Morais do Amaral, Marcelo da Silva Reis, Felipe Rodrigues da Silva (editores). -- Brasília, DF: Embrapa Recursos Genéticos e Biotecnologia, 2007.
24 p. -- (Documentos / Embrapa Recursos Genéticos e Biotecnologia, 1676 - 1340; 224).

1. BLAST - ferramenta de busca - similaridade - seqüências biológicas. 2. DNA. 3. Aminoácidos. 4. Embrapa Recursos Genéticos e Biotecnologia. I. Amaral, Alexandre Morais do. II. Reis, Marcelo da Silva. III. Silva, Felipe Rodrigues da. IV. Série.
631.5233 - CDD 21.

**Editores**

Alexandre Morais do Amaral

Pesquisador da EMBRAPA Recursos Genéticos e Biotecnologia, Brasília, DF

Marcelo da Silva Reis

Laboratório de Biotecnologia do Centro APTA Citros “Sylvio Moreira” - IAC, Cordeirópolis, SP.

Felipe Rodrigues da Silva

Pesquisador da EMBRAPA Recursos Genéticos e Biotecnologia, Brasília, DF

**Programa**

**e**

**Resumos Expandidos**

14-15 de Junho de 2007

BRASÍLIA – DF

**Sumário**

**Introdução**

Com a disponibilidade de grande volume de dados genéticos, gerados principalmente em projetos de seqüenciamento de genomas, e o envolvimento cada vez maior de novos pesquisadores a estes projetos, ferramentas da bioinformática que permitam a interpretação adequada de tais informações têm apresentado grande demanda por parte daqueles que se interessam pelo exame comparativo das seqüências.
Sem dúvida, o procedimento mais comum e difundido para a análise de seqüências é a utilização do programa BLAST para identificar rapidamente a existência de seqüências semelhantes entre si. Estas semelhanças podem indicar homologia e permitir a inferência de função. Ou seja, através da similaridade identificada pode-se atribuir - ainda que provisoriamente - uma função à seqüência sem a necessidade de realização imediata de experimento biológico.
O termo "BLAST" é uma sigla/trocadilho em inglês. A palavra *blast* significa "explosão". O algoritmo, entretanto, recebe este nome por ser uma *Basic Local Alignment Search Tool* (Ferramenta Básica de Busca de Alinhamentos Locais). Como o nome indica, trata-se de uma ferramenta de busca de similaridade entre seqüências biológicas (DNA ou aminoácidos). O algoritmo BLAST e o programa computacional que o implementa foram desenvolvidos, na década de 1980, nos Estados Unidos (ALTSCHUL et al., 1990).
Este algoritmo prioriza alinhamentos de locais específicos da seqüência, com respaldo estatístico, em lugar de realizar alinhamentos "globais", como os realizados por programas de alinhamento de múltiplas seqüências, como o Clustal (CHENNA et al., 2003) (apresentam uma boa revisão de toda a família de programas). Com isto, o programa é capaz de identificar relação entre seqüências que compartilham similaridade, mesmo que esta ocorra somente em algumas regiões isoladas (ALTSCHUL et al., 1990). Tal fato indica que se o recurso utilizado pelo programa fosse o alinhamentos global, ou seja, com o uso imediato do comprimento total da seqüência, menor número de similaridades seriam encontradas, sobretudo no caso de "domínios" e "motivos".
Devido ao seu uso constante, a sigla BLAST é hoje empregada como substantivo e até mesmo verbo ("blastar uma seqüência").

---

Saiba Mais (o algoritmo BLAST):

O BLAST é uma derivação do algoritmo Smith e Waterman (1981), que se caracteriza por apresentar a pontuação máxima do alinhamento local de duas seqüências. O algoritmo Smith-Waterman emprega um *método exaustivo* conhecido como *programação dinâmica*, e assim garante encontrar a pontuação máxima. Por sua vez, o algoritmo BLAST emprega uma *heurística* baseada em *palavras*, o que o torna 50 vezes mais rápido. Quando se considera o tamanho dos bancos de dados contra os quais são realizadas buscas atualmente (mais de 83 bilhões de bases de DNA no *GenBank* em dezembro de 2007), é simples entender porque velocidade é importante, mesmo que às custas da perda de *sensibilidade* (PERTSEMLIDIS e FONDON, 2001).

Uma busca utilizando BLAST compara a seqüência de interesse (*query* ou seqüência de entrada), contra um banco de dados (*subject*). A origem da seqüência de interesse assim como a pergunta que se tenta responder irão determinar o banco de dados a ser utilizado. Desta forma, um pesquisador tentando clonar um determinado gene (digamos, a Rubisco de uma planta recentemente descoberta) pode comparar a seqüência

que acabou de ser gerada em um seqüenciador automático contra todas as seqüências de proteínas já descritas em qualquer parte do mundo. Já um outro pesquisador, diante do resultado do seqüenciamento aleatório de uma biblioteca de cDNA, pode optar por "blastar" as seqüências recém-geradas em seu projeto contra um banco de dados formado pelas seqüências anteriormente geradas no projeto para avaliar o grau de redundância do mesmo.

Nos exemplos do parágrafo anterior, o primeiro pesquisador poderia acessar o BLAST em um *site* de acesso público, ou seja, um local que disponibiliza este recurso para qualquer pessoa com acesso à internet. Ainda que exista um número considerável de servidores públicos de BLAST, a imensa maioria dos usuários no mundo realiza suas buscas no NCBI (*National Center for Biotechnology Information*, ou Centro Nacional para Informação Biotecnológica (http://www.ncbi.nlm.nih.gov)), pertencente ao governo dos Estados Unidos. O segundo pesquisador, por outro lado, talvez esteja realizando suas buscas em um servidor com acesso restrito aos participantes do projeto (denominado "BLAST local").

Nas duas situações, o programa irá identificar um ou mais trechos da seqüência de entrada que são semelhantes a porções de seqüências presentes no banco de dados em questão e fornecer um relatório a respeito desta comparação, com várias informações, tais como o nome e descrição das seqüências do banco de referência e a demonstração numérica e gráfica dos pareamentos significativos.

O objetivo desta publicação é permitir para o usuário da ferramenta BLAST o melhor entendimento e compreensão das informações obtidas após a sua utilização, na interpretação dos resultados da pesquisa. Enfim, a finalidade é aumentar a exploração e o entendimento de dados gerados em estudos que utilizam informação genética a partir da análise de seqüências.

**Conceitos**

De forma bastante resumida, o algoritmo BLAST pode ser dividido em três estágios básicos (ALTSCHUL et al., 1990):

(1) Compilação de uma lista de "palavras" de alta pontuação;

(2) Procura destas palavras no banco de dados;

(3) Extensão de alinhamentos a partir da palavras encontradas.

**Saiba mais (os três estágios básicos do algoritmo):**

**1**. No primeiro estágio é gerada uma lista de todas as "palavras" encontradas na seqüência que está sendo buscada. "Palavra", no caso do BLAST, é um trecho de comprimento definido da seqüência. Os tamanhos-padrão de palavras são 3 aminoácidos para seqüências protéicas e 11 nucleotídeos para seqüências nucléicas. No caso de seqüências protéicas, a lista inicial inclui todas as combinações possíveis de 3 aminoácidos consecutivos, como esquematizado à esquerda na Figura 1. Esta lista é refinada através do cálculo da pontuação do alinhamento, sem *gaps* (*i.e.*, apenas comparando palavras inteiras, ou seja, contínuas, sem interrupção), destas palavras contra todas as 8.000 ($20^3$) palavras possíveis. Apenas "palavras" com pontuação acima de um determinado limiar são selecionadas. Pode-se notar que "palavras"

que não ocorrem na seqüência original, mas que pontuam acima do limiar com palavras presentes na seqüência buscada, são incluídas na lista de trechos selecionados. Em média, 50 palavras serão incluídas na lista para cada aminoácido da seqüência original. Em suma, no primeiro estágio, compila-se uma lista de "palavras" relevantes à seqüência buscada.

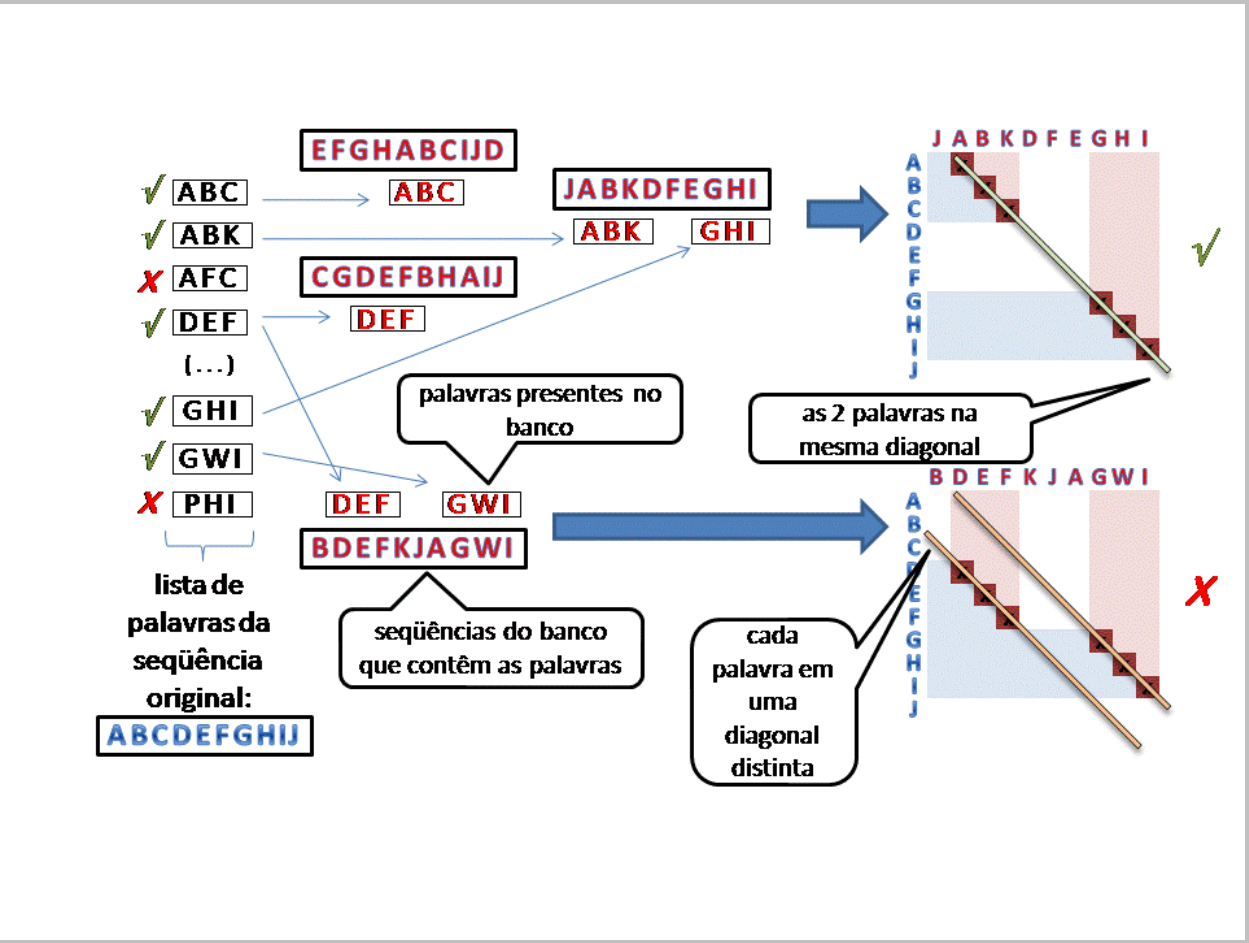



**Figura 1.** Exemplo de procedimento do programa BLAST para busca de seqüências protéicas, com tentativas iniciais de todas as combinações possíveis de 3 aminoácidos consecutivos.

**2**. Palavras idênticas às presentes na lista são identificadas nas seqüências do banco de dados. Como o banco de dados teve sua lista de palavras pré-compilada e indexada no momento em que foi criado, não no instante em que se executa a busca, este passo é extremamente rápido.

**3.a.** Caso sejam encontradas duas palavras na mesma diagonal, o algoritmo inicia tentativas de estender o alinhamento, em ambas as direções e sem *gaps*, a partir da região idêntica representada pela 2ª palavra. Este mesmo algoritmo atribui nota (*score*) para o alinhamento durante a sua formação: enquanto esta nota permanecer alta a extensão é mantida, mas é interrompida quando são detectadas variações que levem à diminuição consistente deste valor. Caso o *score* deste alinhamento sem *gaps* atinja um determinado patamar, um alinhamento que permite *gaps* é disparado a partir da região já alinhada.

**3.b.** Se o programa identifica alinhamento com alto grau de continuidade (ou seja, com poucas regiões de inserção ou deleção), é produzido um alinhamento da comparação, realizada uma extensão deste mesmo alinhamento, a partir de adaptação do algoritmo de Smith-Waterman, e os resultados estatisticamente significativos são demonstrados graficamente para o usuário.

A utilização do programa com suas várias possibilidades também leva ao uso de uma série de recursos com expressões bastante peculiares e que são de uso muito comum durante a sua manipulação, tanto na preparação e submissão dos dados assim como na análise dos resultados. Dentre tais, estão as expressões FASTA, banco de dados, identidade, alinhamento, *e*-value etc..., como demonstrado abaixo.

Formato FASTA

A submissão de dados para uma análise através do algoritmo BLAST requer a formatação desta informação de tal maneira que possa ser reconhecida e interpretada pelo programa. O formato FASTA é a forma de apresentação da seqüência, representada pelo código que utiliza uma letra para cada nucleotídeo ou aminoácido, segundo as normas IUB/IUPAC, antecedida por uma linha que pode conter qualquer registro dado pelo usuário, mas que normalmente é utilizada para descrever o que representa, a origem da seqüência ou mesmo qualquer comentário de interesse de quem está manuseando a seqüência (Fig. 2). Esta "identificação" não é obrigatória para a utilização do programa BLAST se o objetivo é submeter apenas uma seqüência (*query*), ou seja, individualmente, para comparação com um banco de dados em que os seus registros apresentem tal identificação. Preferencialmente, cada linha da seqüência deve conter no máximo 80 caracteres.

Estes dados (seqüências) são opcionalmente identificados, em seu cabeçalho, ou seja, no texto que antecede os nucleotídeos ou aminoácidos, pelo símbolo ">" ("maior que") seguido de curta descrição da seqüência em uma linha (também idealmente, com menos de 80 caracteres), onde são informados pelo usuário a denominação, base de dados, comentário etc...

O símbolo ">" deve vir na primeira coluna e sem espaço entre ele e a primeira letra da identificação. Este formato permite também que o programa identifique, ao reconhecer novamente o símbolo, o início de uma próxima seqüência que ocorra em seguida, se o usuário submeter simultaneamente, para análise, mais de uma seqüência (formato "multi-FASTA").

**>gi|37221963|gb|AAN78258.1| resistance protein [Arachis simpsonii]**
LAKALYNSICDRFECACFLFNVRTISDQEEGLVRLQQTLLSKLLGEWEIKVRSVEEGISMIKEKLSKKRA
LIVLDDVNKIEQLKALAGECDWFSYGTRIVITRDKYLLTAHKVEKIYKMKLLSDPESLELFCWNAFKISR
PKENYEDLSNQAIHYAQ

**Figura 2.** Exemplo de seqüência de aminoácidos em formato FASTA, com identificação (cabeçalho) destacado em negrito.

Programas do BLAST

De forma geral, o programa permite a comparação de seqüências de cinco formas ("sabores") diferentes: Blastp, Blastn, Blastx, tBlastn e tBlastx. Cada comparação a ser solicitada ao algoritmo tem um objetivo e um programa do BLAST a ser utilizado, que é baseado na necessidade do usuário e natureza dos dados e do banco de referência.

Basicamente, serão descritos aqui aqueles programas mais freqüentemente utilizados e que permitem a comparação da seqüência de entrada contra um banco de dados, ambos podendo ser apresentados, simultaneamente (MULTI-FASTA) ou não, como conjunto de nucleotídeos ou aminoácidos (Tabelas 1 e 2).

Embora menos usuais, análises mais específicas são oferecidas por outros programas também derivados do BLAST, tais como o PSI-BLAST (*Position specific iterative* BLAST), que é um refinamento do BLAST e que gera resultados considerando-se uma tabela (matriz) construída automaticamente a partir da freqüência de cada aminoácido em cada posição da seqüência da proteína.

**Tabela 1**. Descrição dos programas presentes no algoritmo BLAST para comparação de seqüências.

| Programa | Descrição |
|---|---|
| Blastp | Compara a seqüência de aminoácido de entrada (*query*) contra um banco de dados de seqüências de proteínas (*subject*). |
| Blastn | Compara a seqüência de nucleotídeos de entrada contra um banco de dados de seqüências de nucleotídeos. |
| Blastx | Compara a seqüência de nucleotídeos de entrada traduzida para todas as seqüências de leitura possíveis contra um banco de dados de seqüências de proteínas. É o programa mais utilizado em grandes projetos de seqüenciamento, pois permite identificar possíveis proteínas a partir de uma seqüência de nucleotídeos desconhecida. |
| tBlastn | Compara a seqüência de aminoácido de entrada contra um banco de dados de seqüências de nucleotídeos traduzidas para todas as seqüências de leitura possíveis. |
| tBlastx | Compara as seis seqüências de leitura possíveis de um nucleotídeo contra um banco de dados de nucleotídeos traduzidos para todas as seqüências de leitura possíveis. |

Fonte: www.ncbi.nlm.nih.org

**Tabela 2**. Simplificação das diferentes comparações possíveis com o uso do BLAST.

| Seqüência de entrada ("query") | Banco de dados ("subject") | |
|---|---|---|
| a.a. | a.a. | **Blastp** |
| nt | nt | **Blastn** |
| nt* | a.a. | **Blastx** |
| a.a. | nt* | **tBlastn** |
| nt* | nt* | **tBlastx** |

a.a. = aminoácido; nt = nucleotídeo; * = traduzidos para todas as seqüências (fases, "frames") de leitura possíveis

*"Blast local"*

O chamado "Blast local", como o próprio nome indica, trata-se de uma versão do programa BLAST disponível para *download* no NCBI, tanto em versão *web* quanto versão terminal (*blastall*), que o pesquisador pode utilizar para analisar, em um computador local de seu laboratório, seqüências confidenciais (*i.e.,* seqüências sigilosas, que ainda não tornaram-se públicas). O BLAST local, disponível em diversas plataformas (Windows, Solaris, Linux, etc), pode ser obtido através no endereço ftp://ftp.ncbi.nih.gov/blast/ .

Para a versão *web* funcionar, é necessário algum conhecimento de informática, pois é preciso configurar algum servidor *web* (*i.e., software* que exibe páginas *web* para outras máquinas, p.e. Apache) na máquina hospedeira.

banco de dados para a utilização do programa BLAST (ver seção 2.3). Já o blastall é o programa BLAST propriamente dito. Embora o blastall funcione em plataforma Windows, por questões de compatibilidade e de eficiência é altamente recomendável que seja utilizado um sistema operacional do tipo UNIX (Linux, Solaris, etc). Uma distribuição Linux bem fácil de instalar e de utilizar é a Ubuntu (http://www.ubuntu.com).

**Saiba mais (Blast local em terminal):**

Para disparar o blastall (BLAST) em terminal, digite o seguinte comando:
*nohup blastall -p blastx -e 1e-5 -F F -d nr -i meu_multifasta.fasta -o meu_multifasta.blastx 1> arq.out 2> arq.err &*
Onde:
*nohup :* "no hang up"; evita que o fechamento do terminal encerre a execução do programa (o comando *nohup* serve para qualquer comando terminal UNIX, não somente para o blastall);
*blastall* : arquivo binário (executável) da ferramenta BLAST;
*-p blastx* : indica que o blast é de uma seqüência nucléica contra um banco de proteínas;
*-e 1e-5* : *cutoff* (valor de corte) do *e-value* (significância estatística), geralmente assumido como o valor $10^{-5}$;
*-F F* : desabilita o filtro de baixa complexidade (*i.e.*, não esconde regiões de baixa complexidade da seqüência em questão);
*-d nr* : o banco de dados (database) utilizado; neste exemplo, o nr;
*-i meu_multifasta.fasta* : especifica arquivo multifasta de entrada;
*-o meu_multifasta.blastx* : especifica arquivo do relatório (*report*) do blast de saída;
*1> arq.out* : redireciona mensagens de avisos (*warnings*) para esse arquivo especificado (isto também serve pra qualquer comando UNIX);
*2> arq.err* : redireciona mensagens de erros para esse arquivo especificado (isto também serve pra qualquer comando UNIX);
*&* : força o programa chamado à rodar em *background* (isto também serve pra qualquer comando UNIX).

Banco de Dados

O banco de dados nada mais é que o local onde são armazenadas as seqüências (nucleotídeos ou aminoácidos) que são utilizadas como referência para a comparação. Como dito anteriormente, este banco pode ser de acesso restrito ou irrestrito. O banco de dados de acesso restrito se caracteriza por não compartilhar publicamente seu conteúdo e o programa BLAST é instalado para funcionamento em sigilo ("Blast local"). Os bancos públicos permitem o acesso irrestrito de dados por qualquer usuário, sem prévia autorização. Eventualmente, estes bancos públicos podem restringir (ou limitar), temporariamente, o acesso de algum usuário se detectarem excesso de demanda, o que também leva muitos usuários a preferirem o "Blast local". A descrição do mecanismo de limitação por excesso de uso do Blast do NCBI pode ser encontrada em http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?CMD=Web&PAGE_TYPE=BlastFAQs#Queuetime .
Há 3 grandes bancos públicos mundiais de seqüências e que trocam diariamente dados entre si: EMBL (*European Molecular Biology Laboratory*, http://www.embl-heidelberg.de/), *GenBank* (*National Center for Biotechnology Information*, http://www.ncbi.nlm.nih.gov/Genbank/index.html) e DDBJ (*DNA Databank of Japan*, http://www.ddbj.nig.ac.jp/).

Conforme o banco a ser utilizado e o objetivo do usuário, há também a possibilidade de restringir a busca dentro do conjunto de seqüências. Ou seja, definir qual coleção de dados será utilizada dentro deste banco. O *GenBank* - o banco público de seqüências mais utilizado no mundo, com dados de mais de 100.000 organismos e que compartilha vários recursos de utilização com o EMBL e DDBJ - possui várias alternativas de busca, seja pela natureza dos dados [DNA, proteína, EST (sítio com seqüência expressa), STS (sítio com seqüência única marcada), GSS (basicamente, EST e DNA genômico) etc...], seja pelo tipo de organismo (humano, plantas, camundongo, fungos, procariotos etc...). De fato, a grande maioria dos usuários do programa BLAST utiliza o *GenBank* com acesso público para a comparação de suas seqüências. Nas Tabelas 3 e 4, são descritas informações de bancos de dados contidos no *GenBank*.

**Tabela 3**. Descrição dos bancos de dados disponíveis no algoritmo BLAST para comparação de seqüências de nucleotídeos.

| Banco de dados | Descrição |
|---|---|
| nr | Todas as seqüências do *GenBank* + EMBL + DDBJ + PDB (sem seqüências de EST, STS, GSS ou HTGS). Atualmente, não exclui redundância. |
| refseq_rna | Seqüências de mRNA do NCBI. |
| refseq_genomic | Seqüências genômicas do NCBI. |
| est | Todos os ESTs dos bancos de dados de *GenBank* + EMBL + DDBJ. |
| est_human | Sub-divisão do banco "est" com coleção apenas de seqüências do ser humano. |
| est_mouse | Sub-divisão do banco "est" com coleção apenas de seqüências do camundongo. |
| est_others | Coleção de todas as seqüências originadas de ESTs, com exceção de seres humanos e camundongo. |
| gss (g*enome survey sequence*) | Seqüências originadas de genomas, como EST, mas também inclui dados de PCR originadas do banco "Alu" e pontas de cosmídeos, BAC e YAC, por exemplo. |
| htgs (h*igh throughput genomic sequences*) | Seqüências depositadas em massa no banco, através de centros de seqüenciamento. Ou seja, sem "refinamento" dos dados. |
| Pat | Seqüências de nucleotídeos derivadas do banco de patentes do *GenBank*. |
| pdb | Seqüências originadas de banco de dados para estrutura tridimensional de proteínas. |
| month | Todas as seqüências recentes, revisadas ou não, dos bancos *GenBank* + EMBL + DDBJ + PDB depositadas nos últimos 30 dias. |
| alu_repeats | Banco com a tradução de todas as repetições de Alu selecionadas do REPBASE, ideal para eliminar efeitos destas repetições presentes nas seqüências de entrada (ftp://ncbi.nlm.nih.gov/pub/jmc/alu). Refere-se a dados de genoma de humanos e outros primatas. |
| dbsts | Dados de STS (*Sequence Tag Site*) dos bancos *GenBank* + EMBL + DDBJ. |
| chromosome | Genomas e cromossomos completos do NCBI, presentes também no banco Refseq_genomic |
| wgs | Montagem de seqüências de projetos de genoma completo por shotgun (*whole genome shotgun*) |
| env_nt | Seqüências geradas de amostras com DNA de vários organismos simultaneamente (metagenoma). Não estão presentes no banco nr. |

**Tabela 4**. Descrição dos bancos de dados disponíveis no algoritmo BLAST para comparação de seqüências de proteínas.

| Banco de dados | Descrição |
|---|---|
| nr | Todos os arquivos não redundantes (*i.e.*, ausência de arquivos "repetidos") dos bancos de dados: seqüências codantes ("CDS") traduzidas do *GenBank*+ PDB+SwissProt+PIR+PRF. |
| refseq | Coleção (não redundante) de todas as seqüências de proteínas do NCBI. |
| swissprot | Banco que armazena, a partir do banco EMBL, seqüências de proteínas com alto grau de anotação (atribuição de função, estrutura de domínios, modificações pós-tradução, entre outros), baixo nível de redundância e alta integração com outros bancos de dados. |
| pat | Seqüências de proteínas derivadas da divisão de patentes do *GenBank*. |
| pdb | Seqüências derivadas do banco de dados "Brookhaven", para estrutura tridimensional de proteínas. |
| env_nr | Tradução dos CDs não redundantes dos depósitos env_nt. |
| month | Todos os CDS novos ou revisados, traduzidos do *GenBank*+ PDB+SwissProt+PIR+PRF, e liberados nos últimos 30 dias. |

Fonte: www.ncbi.nlm.nih.org

CDS

Se parte da seqüência do nucleotídeo codifica uma proteína, uma tradução conceitual, chamada CDS (*coding sequence*), é anotada. Em outras palavras, os CDS são as seqüências codantes ou a porção de uma seqüência que permite a formação dos códons (unidade de três nucleotídeos) responsáveis pela codificação dos aminoácidos. No arquivo da seqüência, o código traz a localização, se houver, do códon de iniciação e de terminação.

## Filtragem (*Filtering*)

É o processo de retirada (omissão) de regiões da seqüência de entrada que apresentam baixa "complexidade", ou seja, que trazem pouca informação do ponto de vista biológico, tais como regiões poliA, e que, presentes, podem levar a *scores* que não correspondem necessariamente à realidade do alinhamento. Embora o programa apresente esta filtragem como já automaticamente acionada, eventualmente ela pode ser desconsiderada pelo usuário se o seu objetivo é obter informações adicionais sobre a seqüência.

## Matriz de substituição

A avaliação da qualidade do alinhamento de proteínas, um procedimento muito importante durante a análise, é realizada com o uso da matriz de substituição, que atribui uma pontuação (*score*) para cada resíduo substituído dentro de um alinhamento: alterações drásticas são penalizadas com valores baixos, e até negativos, pois a presença de aminoácidos não idênticos no alinhamento pode representar repercussões biológicas relevantes, enquanto alterações "tênues" (também sob o ponto de vista biológico) são premiadas com valores altos. Além disso, conservações significativas são melhor pontuadas que conservações irrelevantes.

Se o alinhamento dos resíduos resulta em similaridade das propriedades físico-químicas e a substituição é freqüente entre proteínas da mesma família, a substituição é considerada conservadora e indica possível fator evolutivo envolvido. Do ponto de vista biológico, a substituição tem diferentes efeitos, as quais podem ser ponderados de acordo com as propriedades do aminoácido. Como se pode perceber, para aminoácidos, a identidade entre as seqüências pondera a repercussão estrutural da proteína a partir da presença de diferentes

aminoácidos na cadeia. As matrizes de substituição, que fornecem valores referentes à probabilidade que um determinado aminoácido x tem de sofrer uma mutação e originar o aminoácido y, consideram todos os pareamentos possíveis de aminoácidos. Percebe-se que esta abordagem também permite gerar a probabilidade de uma determinada mutação ocorrer ao longo da evolução.
Há matrizes de substituição que podem ser escolhidas no programa BLAST conforme a necessidade do usuário, tais como PAM ("Point Accepted Mutation") e BLOSUM ("Blocks Substitution Matrix"). Atualmente verifica-se a preferência pela BLOSUM, que é uma matriz mais recente e baseada nos dados genômicos gerados a partir dos anos 90, ao contrário da matriz PAM, estabelecida na década de 70.

BLOSUM
A denominação BLOSUM significa matriz de substituição em blocos. Esta matriz dá a pontuação para o alinhamento a partir da avaliação da freqüência de substituições em blocos de alinhamentos locais em proteínas relacionadas. A matriz mais comumente utilizada é a BLOSUM62.
A construção da matriz específica é baseada em determinada distância evolucionária, que é indicada na sua denominação. Como exemplo, na matriz BLOSUM62 (a mais utilizada atualmente) o alinhamento cujos *scores* foram derivados foi construída a partir de blocos com, no máximo, 62% de identidade (*i.e.* se duas seqüências são mais de 62% idênticas o alinhamento entre elas não será empregado no cálculo da matriz). Ou seja, quanto maior o número da matriz BLOSUM, menos divergentes devem ser as seqüências relacionadas.

**Relatório do BLAST (*report*)**
Toda a submissão correta de dados de seqüências ao programa que aplica o algoritmo BLAST irá gerar um relatório da análise realizada (Figs. 3.a. e 3.b.). Este relatório, o "*report* do BLAST", traz uma série de informações referentes à análise e, dentre elas, a demonstração gráfica dos alinhamentos estatisticamente significativos, a lista das seqüências (com respectivas identificações) que apresentaram similaridade no banco de dados e o grau desta similaridade. Cabe salientar que as versões gráficas do relatório sofrem modificações freqüentes, ao contrário do texto em si.

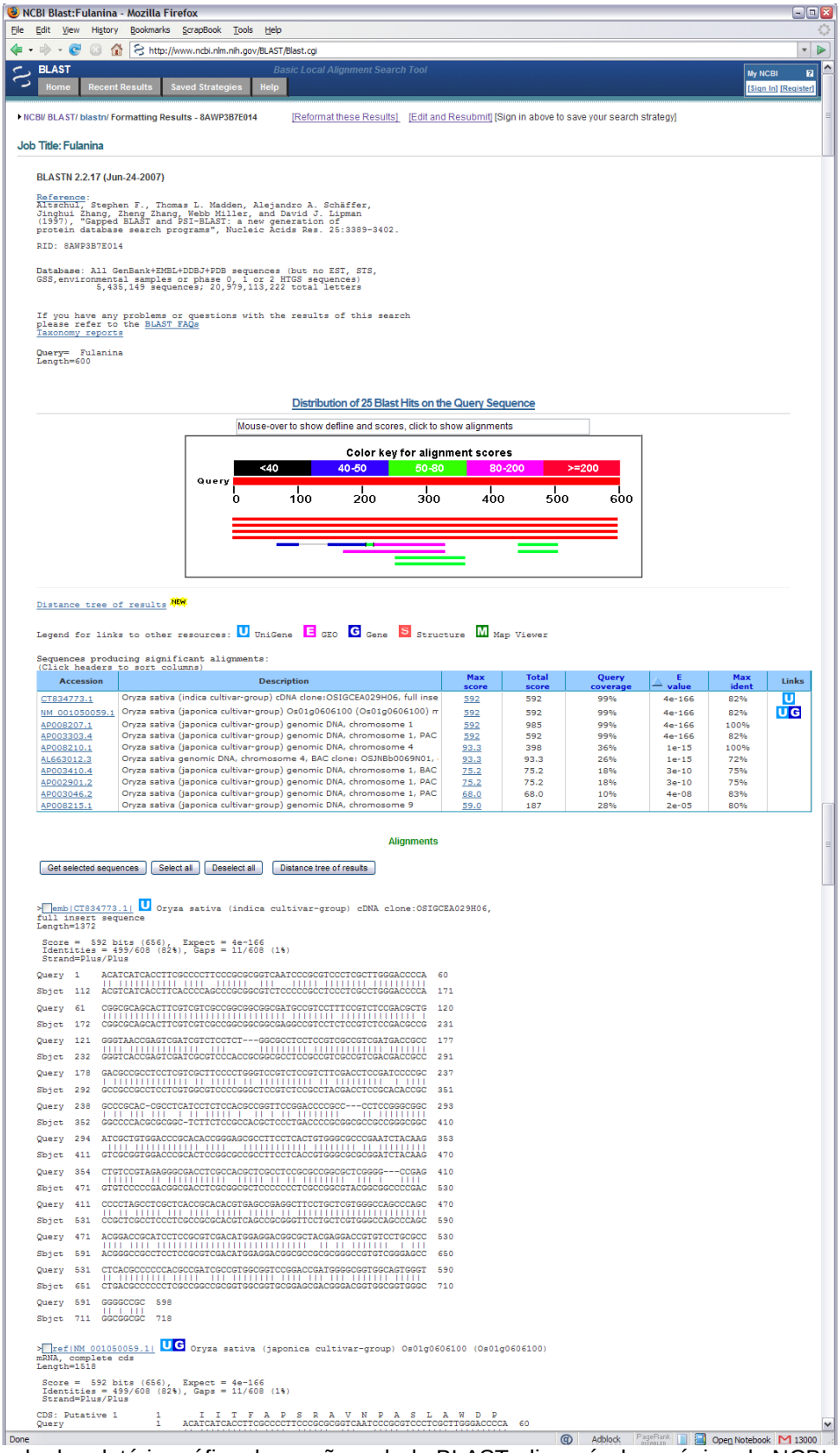



**Figura 3.a.** Exemplo de relatório gráfico da versão web de BLAST, disponível na página do NCBI.

```
BLASTN 2.2.16 [Mar-25-2007]
```

Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

```
Query= Contig2
         (613 letters)

Database: All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS,
GSS,environmental samples or phase 0, 1 or 2 HTGS sequences)
           5,259,275 sequences; 20,596,195,900 total letters

Searching.................................................done


                                                                  Score    E
Sequences producing significant alignments:                       (bits) Value

ref|NM_001065446.1| Oryza sativa (japonica cultivar-group) Os07g...    92   3e-15
dbj|AP008213.1| Oryza sativa (japonica cultivar-group) genomic D...    92   3e-15
dbj|AP005632.3| Oryza sativa (japonica cultivar-group) genomic D...    92   3e-15
dbj|AK066146.1| Oryza sativa (japonica cultivar-group) cDNA clon...    92   3e-15
gb|AC117689.11| Mus musculus chromosome 12, clone RP23-446I19, c...    46   0.16
emb|BX649258.11| Zebrafish DNA sequence from clone CH211-203F4 i...    44   0.64

>ref|NM_001065446.1| Oryza sativa (japonica cultivar-group) Os07g0153300 (Os07g0153300)
           mRNA, complete cds
          Length = 2881

 Score = 91.7 bits (46), Expect = 3e-15
 Identities = 70/78 (89%)
 Strand = Plus / Plus


Query: 25  agacaagtaattctctgaggctggcaaccttcttctggctgatgccgcacaaagttgctg 84
           |||| |||| |||||| ||||| ||||| |||||||||||||||| || ||||||||||
Sbjct: 513 agacgagtagttctcttaggctagcaacattcttctggctgatgccccataaagttgctg 572


Query: 85  catggatatactgggagg 102
           || ||||||||||||||
Sbjct: 573 cagggatatactgggagg 590



 Score = 75.8 bits (38), Expect = 2e-10
 Identities = 122/150 (81%)
 Strand = Plus / Plus


Query: 100  aggccctcagactttggctgaagaacgtcaagttcctagaccatcccaggtatctgaacc 159
            ||||| |||||||||||||||||||||||| ||||| |||||||| || ||  |||||
Sbjct: 2267 aggccatcagactttggctgaagaacgtcaaattcctggaccatccaagatacttgaact 2326
(…)
```

**Figura 3.b.** Exemplo de relatório texto da versão local de BLAST, obtido rodando o programa blastall em um terminal.

Identificadores

Todas as seqüências depositadas nos bancos de dados do NCBI recebem automaticamente uma identificação e que, dependendo do código empregado, fornece uma série de informações relevantes quanto à origem e natureza destes dados.

Identificadores de seqüências (gi)

O *gi* é um código único, representado por números e atribuído a toda seqüência de nucleotídeos ou proteína traduzida depositada no banco de dados do *GenBank*, não importando a origem. É uma espécie de "RG" da seqüência: individual, intransferível e não modificável. Mesmo que haja dois registros de seqüências dentro do mesmo arquivo, a versão em nucleotídeo apresentará um código e a proteína, outro. Com alterações realizadas na seqüência, será registrado um novo *gi* que pode não guardar qualquer relação numérica com o arquivo original – que não será extinto nem modificado.

Identificador de proteína (Protein ID)

O identificador *Protein ID* é atribuído a uma seqüência de aminoácidos que pertence a algum arquivo. Como padrão, adota-se um formato composto de três letras seguidas de cinco dígitos, um ponto e o número da versão. Por exemplo, o código CAC40990.1 se refere a uma proteína com *gi* 14331118 e número de acesso AJ404328.1.
Além do *gi*, cada seqüência apresenta, concomitantemente, uma identificação que representa o código do local onde foi realizado o primeiro depósito daquela seqüência presente no *GenBank*. As origens podem ser representadas pelos seguintes códigos: gb (*GenBank*), emb (EMBL), dbj (DDBJ), ref (RefSeq), sp (Swiss Prot), pdb (*Protein Databank*), pir (PIR), prf (PRF), tpg (TPA *GenBank*), tpe (TPA EMBL) e tpj (TPA DDBJ). A expressão TPA (*Third party annotation*) significa que são seqüências que, mesmo já presentes no banco de dados, sofreram alteração de anotação posterior por terceiros, ou seja, indivíduos que não os autores ou a própria equipe do banco de dados.

Número de acesso

O número de acesso ou *accession number* é o identificador do registro da seqüência depositada no *GenBank*, que combina letras e números, e que pertence então à coleção de seqüências do banco de dados. Normalmente, este identificador compreende a combinação de uma letra seguida de cinco dígitos ou duas letras e seis dígitos. Ele representa o relatório completo da seqüência e não somente a seqüência em si, ao contrário do "Sequence Identifier", que é representado pelo código "gi" ou "ProteinID".

Exemplos:
gi|14331118|emb|CAC40990.1|
gi|41052472|dbj|BAD07483.1|
gi|15218218|ref|NP_173005.1|
gi|92894031|gb|ABE92044.1|

Número de acesso RefSeq

Identificador de um registro de seqüência e natureza dos dados, sendo as duas primeiras indicativas do tipo da seqüência. Não confundir com os códigos de acesso do *GenBank*. O formato segue o uso de duas letras, uma "sub-barra" ("_") e números. Os símbolos mais comumente utilizados são aqueles que qualificam que tipo de seqüência (mRNA, proteína, genoma completo etc.) ou se houve qualquer avaliação/correção ("cura") por parte do corpo técnico do NCBI ou colaboradores.

**Exemplos:**
NT_12345 (*contig* construído a partir de genoma, não "curado"), NM_12345 [mRNA (na verdade, seqüências de cDNA formados a partir de mRNA), curado], NP_12345 [proteína (completa ou parcial), curado], NC_12345 [seqüência genômica completa (genomas, organela, plasmídeo ou cromossomo), curado], NR_12345 (transcritos não codificantes, como RNA estrutural e pseudogenes, curado), XM_12345 [mRNA (seqüências de cDNA formados a partir de mRNA), não curado], XP_12345 [proteína (completa ou parcial), não curado], entre outros.

Alinhamento

O alinhamento é a disposição de duas seqüências, com a demonstração gráfica da comparação através do pareamento da seqüência de entrada (ou seja, que o usuário está investigando) (*query*) na linha superior, com a seqüência do banco de dados (*subject*) na linha inferior. No alinhamento, as similaridades são destacadas. Cabe lembrar que a simbologia é modificada se a seqüência se refere a nucleotídeo ou aminoácido. As combinações imperfeitas (*mismatchs*) são consideradas mutações, enquanto que os intervalos vazios (*gaps*) são considerados como deleções ou inserções.

Nucleotídeos
Para a representação de regiões onde há identidade de nucleotídeos no pareamento, é utilizada uma barra vertical ("|") (Fig. 4). Por outro lado, as combinações imperfeitas não apresentam qualquer símbolo e os intervalos vazios (regiões onde não há seqüência para a comparação) são preenchidos por um traço ("-"). O número que aparece ao lado das seqüências significa a localização dentro da mesma onde se iniciou, significativamente, a similaridade.

```
Query  302    ATAA---CAAGGGATAAATATCTTCTAACAGCTCAT  334
              ||||   ||||||||||||||||||||| |||||
Sbjct  328    ATAACACCAAGGGATAAATATCTTCTAACACCTCAT  363
```

**Figura 4**. Exemplo com representação de pareamento de nucleotídeos, com indicação de regiões similares e com "combinações imperfeitas**".**

*Aminoácido*

A representação da similaridade em alinhamentos entre aminoácidos é ligeiramente mais complexa do que para nucleotídeos. Além da identidade entre aminoácidos, representado por pelo código de uma letra, surge o conceito de *similaridade*. Quando os aminoácidos alinhados entre as duas seqüências, embora não sejam

idênticos, recebem pontuação positiva na Matriz de Substituição escolhida para a busca dizemos que eles são similares. A similaridade, representado por sinal positivo (“+”) depende, portanto, da Matriz de Substituição (Fig. 5).

```
Query  13   SASENK---KKGMVLPFDPHSITFDEVVYSVD  41
            S +ENK   K+GM+L F+PH ITFDEV YSVD
Sbjct  532  SVTENKHYGKRGMILSFEPHCITFDEVTYSVD  563
```

**Figura 5.** Exemplo de representação da similaridade em alinhamentos entre aminoácidos. O sinal positivo (+) indica que a par de aminoácidos alinhados pontua positivament na Matriz de Substituição empregada no alinhamento.

*Score*

Nota atribuída pelo algoritmo e baseada no número de pareamentos perfeitos (*match*) e imperfeitos (*mismatch*) entre a seqüência de entrada e alguma seqüência do banco de dados. É conseqüência do número de inserções, deleções e substituições neste pareamento. Contudo, mesmo quando o *score* é alto e, em princípio, melhor o pareamento, este valor não pode ser analisado individualmente, mas acompanhado do respaldo estatístico (*e*-value). Este *score* de um alinhamento (S) é dado pela soma de *scores* de *matchs*, *gaps* (inserções e deleções) e, no caso de aminoácidos, substituições, sendo estas últimas obtidas através de uma tabela (normalmente a BLOSUM) que reflete o grau de relevância desta “troca” de aminoácidos.
Enfim, o valor do score dá uma indicação se o alinhamento é bom ou não, sendo o seu valor positivamente correlacionado com a qualidade deste alinhamento (ou seja, quanto maior, melhor).

*E-value*

É uma das expressões mais utilizadas na análise de seqüências e representa o valor estatístico (probabilidade) que indica se o alinhamento é real ou foi obtido meramente pelo acaso naquele banco de dados (“falso positivo”). Em outras palavras, é o número esperado de falsos positivos que obteriam *score* igual ou maior que o reportado em um determinado alinhamento entre a seqüência de entrada e uma do banco.
Fundamentalmente, quanto menor o *e*-value, menores as chances daquele resultado ser conseqüência do acaso.
Como exemplo, um *e*-value = 0.02, dentre outras interpretações, pode significar que em um determinado alinhamento há 2 chances em 100 (ou 1 em 50) de que as similaridades sejam resultantes meramente do acaso. Do ponto de vista estatístico, tal número pode representar um valor significativo, mas do ponto de vista biológico representa, em princípio, valor bastante aquém do ideal – embora alguns pesquisadores considerem que ainda assim, tal resultado pode representar alguma informação biológica (Pertsemlidis e Fondon, 2001).
Pode-se perceber então que quanto menor o *e*-value (quanto mais próximo de zero), maior a confiabilidade da predição (mais significativo é o alinhamento). Adicionalmente, em geral estabelece-se uma “nota de corte”, que determina que sejam considerados para uma primeira análise apenas *e*-values menores que e-5 ou e-10.
Geralmente, o valor do *e*-value é apresentado de três formas possíveis, por exemplo: 0.0, 3e-25 e 0.12. O valor 3e-25 é uma abreviação de $3 \times 10^{-25}$, o que indica que este número está mais próximo de 0.0 (máximo de confiabilidade) do que de 0.12.
Outro fator importante a ser considerado é que buscas dentro de um banco de dados utilizando-se seqüência de entrada muito curta podem resultar em homologia total, porém com valor de *e*-value pior (alto), pois o

cálculo também pondera o comprimento da seqüência, uma vez que seqüências menores tendem naturalmente a ter maiores chances (probabilidade) de encontrarem trechos semelhantes em um banco de dados e, por isso, podem representar apenas um evento do acaso.

$$E = K\ m\ n\ e^{-\lambda S}$$

Onde E é o *e-value*, m e n são os comprimentos das seqüências, K e λ (lambda) são parâmetros e S é o *score*. Através da fórmula, percebe-se então que o *e*-value decresce (i.e., melhora) à medida que aumenta o *score* encontrado para o alinhamento das duas seqüências.

Identidade

É o número de resíduos (letras) similares (*matchs*) identificados no alinhamento e expresso, em porcentagem, a partir da comparação com o comprimento deste alinhamento. Nos resultados do BLAST, os "positivos" (que já não representam a identidade, mas a similaridade) indicam a conservação evolutiva, ou seja, são a soma do número de aminoácidos idênticos e aqueles que são diferentes na comparação mas que apresentam *score* positivo na tabela empregada.

*Gap*

Espaço introduzido em um alinhamento para compensar regiões de inserção ou deleção em uma das duas seqüências (entrada ou banco). À medida que aumenta o número de *gaps* no alinhamento, para poder acomodar o pareamento entre as seqüências, há o aumento da penalização aplicada ao *score* deste mesmo alinhamento. Ou seja, embora a introdução de *gaps* resulte em um melhor pareamento, há como conseqüência uma diminuição ponderada do *score*, o que, em algum grau, irá afetar desfavoravelmente o *e*-value.

Similaridade

Similaridade é o grau da semelhança entre as seqüências. Este valor é baseado na identidade e/ou conservação da seqüência. No programa BLAST, este termo é baseado nos valores das tabelas de matrizes. Logicamente, o julgamento da similaridade vai ser limitado à disponibilidade de seqüências do banco de dados para a comparação.
Uma prática comum na sua utilização e interpretação é o uso do *hit* (seqüência do banco) mais similar nos relatórios que analisam vários genes ou proteínas.

Homologia

A expressão se refere à similaridade atribuída a descendentes originados de ancestral comum. Ou seja, seqüências homólogas são seqüências que, supostamente, têm a mesma origem. A homologia pode ser classificada como paralogia ou ortologia. Se a homologia é resultante de especiação, ou seja, a seqüência ocorre em uma espécie que originou outra espécie que, por sua vez, apresenta uma cópia desta mesma seqüência, elas são ortólogas, embora possam não ser responsáveis por função semelhante. Por outro lado,

se há duas cópias da mesma seqüência no mesmo organismo, ou seja, uma duplicação da seqüência, estas são parálogas (FITCH, 2000).

Freqüentemente, na comparação entre seqüências, de forma errônea utiliza-se a expressão, por exemplo, “93% de homologia” enquanto o correto é dizer-se que há “93% de identidade” entre elas, uma vez que a terminologia “homólogo” se refere basicamente à origem e não ao grau de similaridade. Ou, colocado de outra forma, a porcentagem de identidade é uma medida concreta, enquanto homologia é uma hipótese sustentada por esta evidência. Em suma, a homologia se descreve apenas binominalmente (homólogo ou não-homólogo).

**Exemplos de alinhamentos**

Proteína

**gi|41052472|dbj|BAD07483.1| PDR-type ABC transporter 1 [Nicotiana tabacum]**

**gi|75326590|sp|Q76CU2|PDR1_TOBAC Pleiotropic drug resistance protein 1 (NtPDR1)**

```
Length=1434

 Score = 64.7 bits (156),  Expect = 1e-09
 Identities = 33/42 (78%), Positives = 35/42 (83%), Gaps = 2/42 (4%)

Query  2    SPQITSTQEGDSASE--NKKKGMVLPFDPHSITFDEVVYSVD  41
            S QITST  GDS SE  N KKGMVLPF+PHSITFD+VVYSVD
Sbjct  806  SSQITSTDGGDSISESQNNKKGMVLPFEPHSITFDDVVYSVD  847
```

Algumas informações que podem ser obtidas do alinhamento acima:

1. A seqüência de referência (banco de dados), embora presente no *GenBank* (como verificado pelo código *gi*), foi originalmente depositada no banco de dados DDBJ (dbj) e também no Swiss Prot (sp) e apresenta 1434 aminoácidos;
2. O *score* indica a pontuação atribuída pelo programa, que é baseada no número de pareamentos perfeitos (*match*), inserções e deleções (*gaps*) e substituições, entretanto tem pouca utilidade analisado isoladamente;
3. Considerando o *e-value* resultante (*Expect*), há uma probabilidade de 1 x $10^{-9}$ do alinhamento ter sido apenas casual, o que não se pode considerar como um ótimo valor, sobretudo por se tratar de seqüência curta (*i.e.*, poucos resíduos);
4. O alinhamento se iniciou no 2º aminoácido da seqüência de entrada e foi interrompido em seu 41º aminoácido, enquanto que na seqüência de referência o alinhamento se iniciou no 806º aminoácido e foi finalizado no 847º aminoácido;
5. Neste trecho da seqüência de referência, de 42 aminoácidos da seqüência de entrada, 33 são idênticos aos da seqüência de referência , enquanto 35 apresentaram “similaridade”, ou seja, além dos 33 idênticos entre as duas seqüências, dois aminoácidos não são idênticos (foram substituídos) mas pontuam positivamente na Matriz de Substituição empregada (*i.e.,* são conservados – representados pelo sinal “+” no alinhamento).
6. Há duas deleções presentes na seqüência de entrada de 42 aminoácidos [*Gaps* = 2/42 (4%)].

Nucleotídeo

**gi|102139737|gb|AC186746.1| Musa acuminata clone MA4_25J11, complete sequence**

```
Length=105019

Score = 46.1 bits (23),  Expect = 0.24
Identities = 41/47 (87%), Gaps = 0/47 (0%)
Strand=Plus/Plus

Query  113    TTCTCAATGGGAAACCATTTTACGCTAATGGCTTCAATGCCTACTGG  159
              ||||||||||| || ||||| |  || ||||||||||||||||||||
Sbjct  27822  TTCTCAATGGCAATCCATTCTTTGCCAATGGCTTCAATGCCTACTGG  27868
```

Algumas informações que podem ser obtidas do alinhamento acima:

1. A seqüência de referência foi primeiramente depositada no *GenBank* e possui comprimento total de 105.019 resíduos;
2. O total de pontos (*score*) do alinhamento foi de 23. O alto valor de "*e*-value" (0.24) indica que há uma grande possibilidade deste *score* ser obtido apenas por chance, não indicando homologia.
3. A seqüência de entrada apresentou alinhamento entre os seus nucleotídeos da posição 113 e 159, enquanto na seqüência de referência, o alinhamento se deu nas posições 27.822 e 27.868 do arquivo;
4. Dos 47 nucleotídeos da seqüência de entrada, 41 foram iguais àquele trecho que estava no banco de dados. Como é uma seqüência de nucleotídeos, ao contrário do que ocorre com aminoácidos, não há julgamento de semelhança ou "positivos", por outro lado, também há pontuação de acordo com o pareamento inexato ou presença de *gaps*;
5. Não foram detectadas deleções ou inserções (*gap*) dentro do alinhamento formado pelo BLAST;
6. O DNA utilizado para o alinhamento é originado da fita *sense* de ambas as seqüências ("plus/plus").

## Glossário de termos computacionais

*Algoritmo* – Lista finita de instruções (código) para a execução de uma determinada tarefa que, dado um estado inicial definido (entrada do programa), irá gerar um estado final também definido (saída do programa).

*Heurística* – Tipo de algoritmo que, dado um estado inicial, busca pelo melhor resultado possível durante um certo período de tempo ou de iterações, porém não garantindo que o resultado ótimo seja atingido. O algoritmo do BLAST é na verdade uma heurística, por conter métodos heurísticos em seu código.

*Método Exaustivo* – Tipo de busca ou análise onde são testadas todas as possibilidades possíveis para determinar o resultado.

*Palavras* – Aqui entende-se por cadeias de caracteres (*strings*), com comprimento definido, que representam seqüências de nucleotídeos (palavras de 11 caracteres são o padrão do Blast) ou de aminoácidos (3 caracteres).

*Programação Dinâmica* – Tipo de programação (construção de algoritmos) onde a solução ótima de um estado inicial é atingida a partir do aproveitamento da solução ótima de um sub-problema do mesmo estado inicial. Este tipo de programação visa principalmente evitar o recálculo, tornando o algoritmo mais rápido.

## REFERÊNCIAS


ALTSCHUL, S. F.; GISH, W.; MILLER, W.; MYERS, E. W.; LIPMAN, D. J. Basic local alignment search tool. **Journal of Molecular Biology**, Amsterdam, v. 215, n. 3, p. 403-410, 1990.

ALTSCHUL, S. F.; MADDEN, T. L.; SCHAFFER, A. A.; ZHANG, J.; ZHANG, Z.; MILLER, W.; LIPMAN, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. **Nucleic Acids Research**, London, v. 25, n. 17, p. 3389-33402, 1997.

CHENNA, R.; SUGAWARA, H.; KOIKE, T.; LOPEZ, R.; GIBSON, T. J.; HIGGINS, D. G.; THOMPSON, J. D. Multiple sequence alignment with the Clustal series of programs. **Nucleic Acids Research**, London, v. 31, p. 3497–3500, 2003.

FITCH, W. M. Homology a personal view on some of the problems. **Trends in Genetics: DNA Differentiation & Development**, Amsterdam, v. 16, n. 5, p. 227-231, 2000.

HENIKOFF, S.; HENIKOFF, J. G. Amino acids substitution matrices from protein blocks. **Proceedings of the National Academy of Sciences of the United States of America**, Washington, v. 89, n. 22, p. 10915-10919, 1992.

KOSKI, L. B.; GOLDING, G. B. The closest BLAST hit is often not the nearest neighbor. **Journal of Molecular Evolution**, New York, v. 52, n. 6, p. 540-542, 2001.

KRAUTHAMMER, M.; RZHETSKY, A.; MOROZOV, P.; FRIEDMAN, C. Using BLAST for identifying gene and protein names in journal articles. **Gene**, Amsterdam, v. 259, n. 1-2, p. 245-252, 2000.

PERTSEMLIDIS, A.; FONDON, J. W. 3rd. Having a BLAST with bioinformatics (and avoiding BLASTphemy). **Genome Biology**, v. 2, n. 10, p. reviews2002.1-2002.10, 2001.

SMITH, T. F.; WATERMAN, M. S. Identification of common molecular subsequences. **Journal of Molecular Biology**, Amsterdam, v. 147, p. 195-197, 1981.