

# Pesquisa de sequências em bases de dados

- Problema: encontrar sequências numa base de dados que apresentem uma semelhança **significativa** com uma determinada sequência

Factores a ter em consideração:

- Tamanho da base de dados
- Rapidez do algoritmo de alinhamento
- Sensibilidade do algoritmo de alinhamento
- DNA ou proteína ?

## As buscas com proteínas são mais sensíveis

- Sempre que possível, usar uma sequência de proteína (ou sequência de DNA traduzida) nas buscas
- Enquanto se podem detectar semelhanças entre sequências de proteína que divergiram há mais de 2 bilhões de anos, as comparações de DNA não permitem ir além dos 200-500 milhões de anos

# Exemplo de busca de DNA vs. proteína

Table 5: DNA vs. protein sequence comparison

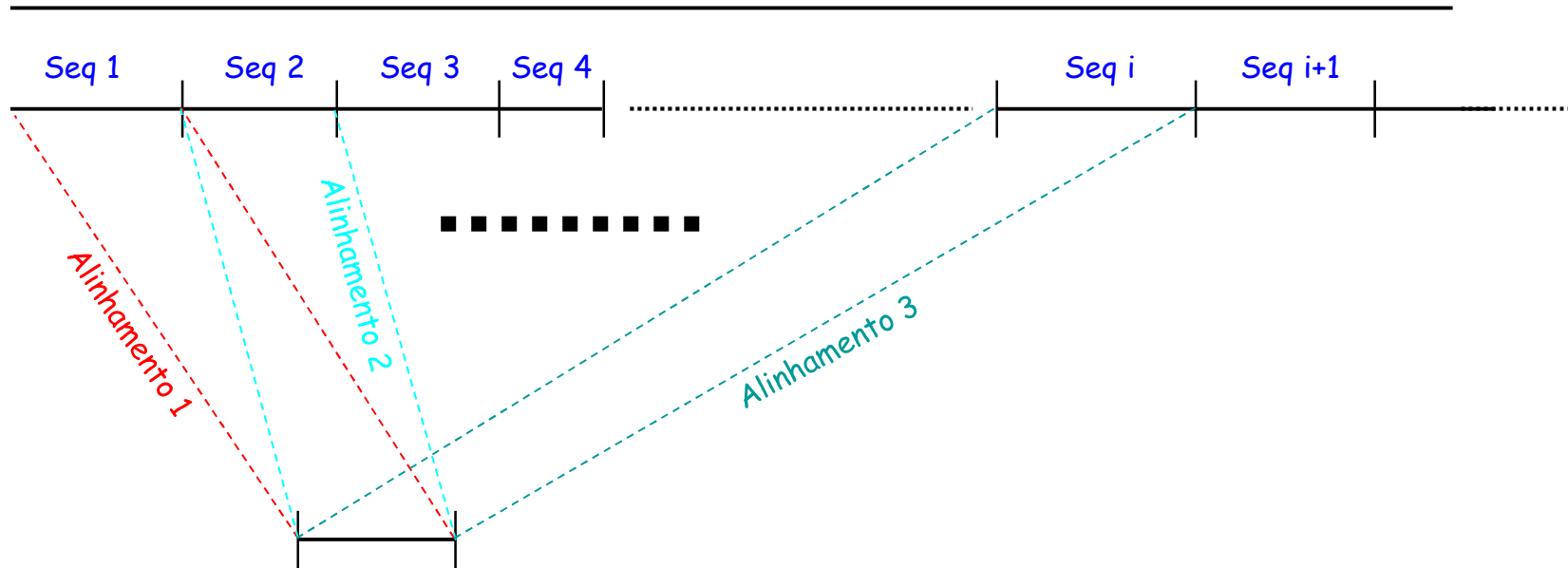
The best scores are:		DNA E(188,018)	tfastx3 E(187,524)	prot. E(331,956)
DMGST	D.melanogaster GST1-1	1.3e-164	4.1e-109	1.0e-109
MDGST1	M.domestica GST-1 gene	2e-77	3.0e-95	1.9e-76
LUCGLTR	Lucilia cuprina GST	1.5e-72	5.2e-91	3.3e-73
MDGST2A	M.domesticus GST-2 mRNA	9.3e-53	1.4e-77	1.6e-62
MDNF1	M.domestica nf1 gene. 10	4.6e-51	2.8e-77	2.2e-62
MDNF6	M.domestica nf6 gene. 10	2.8e-51	4.2e-77	3.1e-62
MDNF7	M.domestica nf7 gene. 10	6.1e-47	9.2e-77	6.7e-62
AGGST15	A.gambiae GST mRNA	3.1e-58	4.2e-76	4.3e-61
CVU87958	Culicoides GST	1.8e-41	4.0e-73	3.6e-58
AGG3GST11	A.gambiae GST1-1 mRNA	1.5e-46	2.8e-55	1.1e-43
EMO6502	Bombyx mori GST mRNA	1.1e-23	8.8e-50	5.7e-40
AGSUGST12	A.gambiae GST1-1 gene	2.3e-16	4.5e-46	5.1e-37
MOTGLUSTRA	Manduca sexta GST	5.7e-07	2.5e-30	8.0e-25
RLGSTARGN	R.leguminosarum <i>gstA</i> and <i>gstR</i>	0.0029	3.2e-13	1.4e-10
HUMGST2A	H. sapiens GSTT2	0.32	3.3e-10	2.0e-09
HSGSTT1	H.sapiens GSTT1 mRNA	7.2	8.4e-13	3.6e-10
ECAE000319	E. coli hypothet. prot.	—	4.7e-10	1.1e-09
MYMDCMA	Methylophilus dichlorometh. DH	—	1.1e-09	6.9e-07
BCU19883	Burkholderia maleylacetate red.	—	1.2e-09	1.1e-08
NFU43126	Naegleria fowleri GST	—	3.2e-07	0.0056
SP505GST	Sphingomonas paucim	—	1.8e-06	0.0002
EN1838	H. sapiens maleylacetoacetate iso.	—	2.1e-06	5.9e-06
HSU86529	Human GSTZ1	—	3.0e-06	8.0e-06
SYCCPNC	Synechocystis GST	—	1.2e-05	9.5e-06
HSEF1GMR	H.sapiens EF1g mRNA	—	9.0e-05	0.00065

Organismos  
mais distantes  
da Drosophila

Nesta zona a pesquisa  
com a sequência de de  
DNA não detecta  
similaridade

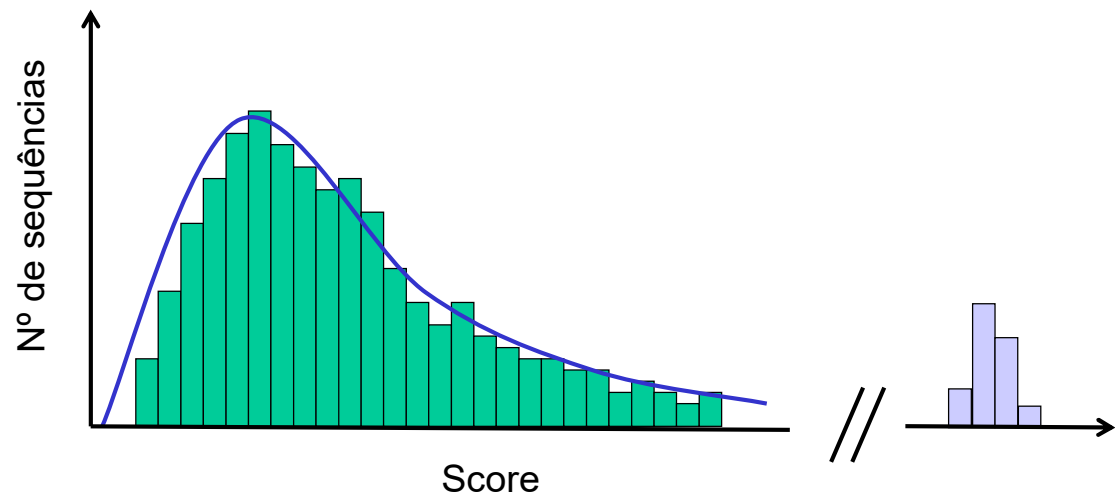
A pesquisa com proteína ou DNA traduzido permite identificar sequências da proteína GST em organismos mais distantes da Drosophila (mosca da fruta)

# Base de dados de seqüências:

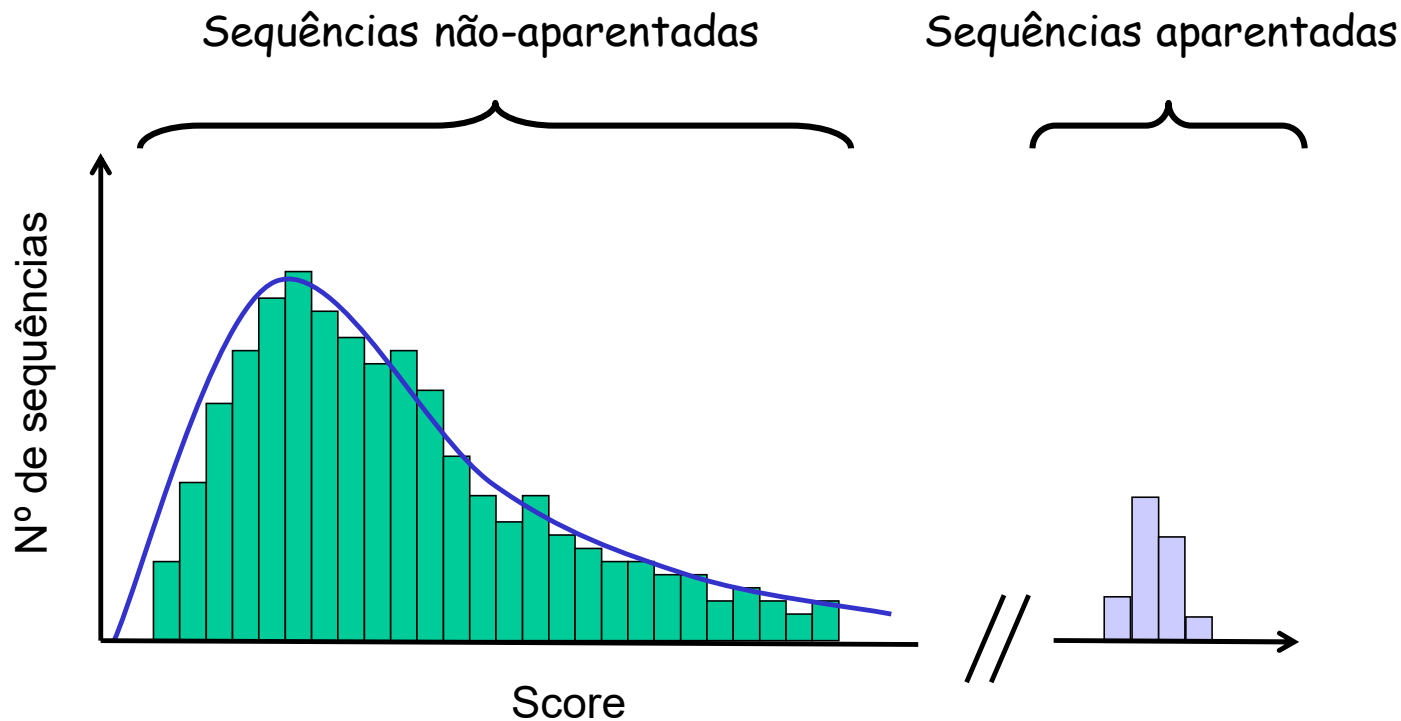


Sequência de pesquisa

## Distribuição de scores:



- Em geral a maioria das sequências na base de dados não apresenta semelhança detectável com a nossa sequência de pesquisa, produzindo alinhamentos *aleatórios*
- O pequeno grupo de sequências aparentadas com a sequência de busca irá produzir um score muito mais elevado
- Valores de  $K$  e  $\lambda$  podem ser estimados a partir da distribuição de scores



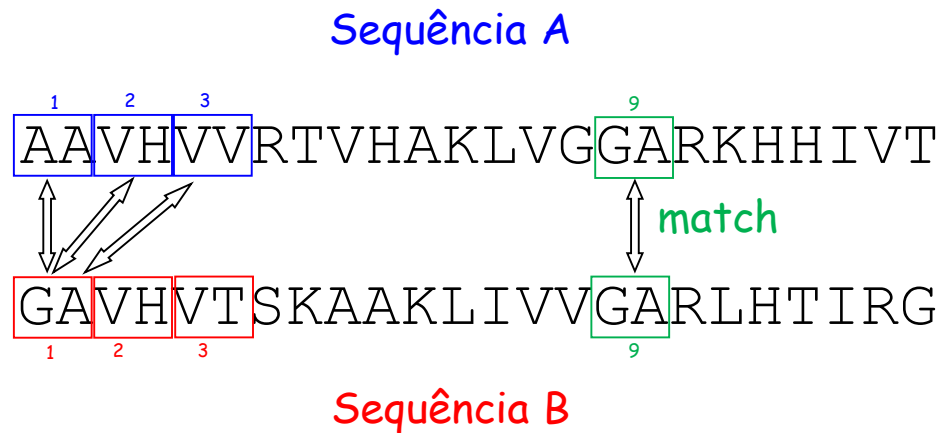
# Algoritmos de busca

- A grande dimensão das bases de dados de sequências e o elevado número de buscas suportado pelos servidores actuais torna impraticável o uso dos algoritmos de Needleman-Wunsch ou Smith-Waterman para a geração dos alinhamentos durante o processo de busca
- Recorre-se normalmente a uma classe de métodos, chamados **métodos heurísticos** de alinhamento, que sacrificam alguma sensibilidade na busca a troco de uma muito maior rapidez
- Os dois métodos heurísticos de busca mais usado são de longe os seguintes:
  - **FASTA** - <https://www.ebi.ac.uk/Tools/sss/fasta>
  - **BLAST** - <http://ncbi.nlm.nih.gov/BLAST>

# FASTA

- O primeiro programa a ser amplamente usado para pesquisa de bases de dados de seqüências, desenvolvida por Pearson e Lipman em 1985.
- Em vez de calcular sistematicamente alinhamentos óptimos, começa por analisar regiões de similaridade mais alta usando comparações entre grupos de resíduos, em vez de resíduos isolados. Em seguida produz um alinhamento óptimo dentro de uma zona restrita do dot plot. As seqüências que neste alinhamento retornam score mais alto são alinhadas rigorosamente com o método de Smith-Waterman.

<https://www.ebi.ac.uk/Tools/sss/fasta>



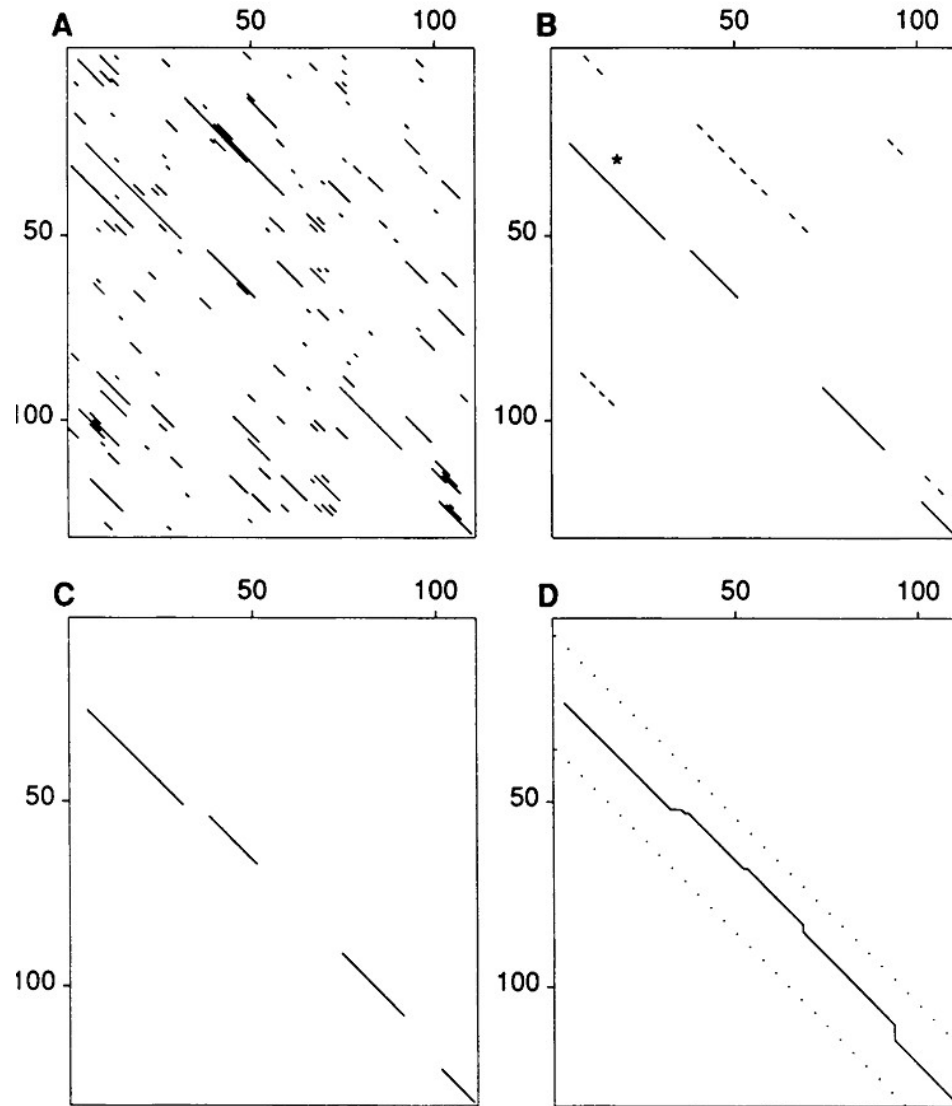
ktup=2

Tabela de "Hashing":

	A	B
AA	1	5
VH	2	2
...		
GA	9	9
...		
...		



# Estratégia do programa FASTA



(A) Identificação de regiões de similaridade por comparação de "k-tuplas" entre a sequência-alvo e cada sequência da base de dados

(B) Cálculo de score para cada região de similaridade utilizando uma matriz de score e selecção das regiões de score mais alto

(C) Junção das regiões de score alto utilizando uma "junction penalty"

(D) Pesquisa do alinhamento óptimo numa banda (entre as linhas tracejadas) envolvendo a região definida no passo anterior

- [Help Index](#)
  - [General Help](#)
  - [Formats](#)
  - [Gaps](#)
  - [Matrix](#)
  - [References](#)
  - [FASTA Help](#)
  - [MView Help](#)
  - [VisualFASTA Help](#)
- 
- [View all FASTA's at EBI](#)
  - [FASTA Programmatic Access](#)
- 
- [Database Information](#)
    - [UniProt](#)
    - [UniParc](#)
- 
- [Similar Applications](#)
    - [FASTA](#)
    - [BLAST](#)
    - [MPsrch](#)
    - [ScanPS](#)

EBI > Tools > Similarity & Homology > FASTA

### FASTA - Protein Similarity Search

Provides sequence similarity searching against nucleotide and protein databases using the FASTA programs. FASTA can be very specific when identifying long regions of low similarity especially for highly diverged sequences. You can also conduct sequence similarity searching against complete [proteome](#) or [genome](#) databases using the [FASTA programs](#).

 [Download Software](#)

PROGRAM	DATABASES	RESULTS	SEARCH TITLE	YOUR EMAIL	
FASTA3 <input type="button" value="v"/>	Protein <input type="button" value="v"/> UniProt Knowledgebase <input type="button" value="▲"/> UniProtKB/Swiss-Prot <input type="button" value="▼"/> UniProt Clusters 100% <input type="button" value="v"/>	interactive <input type="button" value="v"/>	Sequence	<input type="text"/>	
MATRIX	GAP OPEN	GAP EXTEND	KTUP	EXPECTATION UPPER VALUE	EXPECTATION LOWER VALUE
BLOSUM5 <input type="button" value="v"/>	-10 <input type="button" value="v"/>	-2 <input type="button" value="v"/>	2 <input type="button" value="v"/>	10.0 <input type="button" value="v"/>	default <input type="button" value="v"/>
DNA STRAND	HISTOGRAM	MOLECULE TYPE			
none <input type="button" value="v"/>	no <input type="button" value="v"/>	Protein <input type="button" value="v"/>			
SCORES	ALIGNMENTS	SEQUENCE RANGE	DATABASE RANGE	FILTER	STATISTICAL ESTIMATES
50 <input type="button" value="v"/>	50 <input type="button" value="v"/>	START-END	START-END	none <input type="button" value="v"/>	Regress <input type="button" value="v"/>

Enter or Paste a  Sequence in any format

Upload a file:

**Please Note:** The way that the email submission results are sent back has changed, instead of returning the actual FASTA result, there is now a hyperlink to your result pages.

# FASTA no EBI

## FASTA

Protein | Nucleotide | Genomes | Proteomes | Whole Genome Shotgun | Web services | Help & Documentation | Also in this section ▾

Feedback | Share

Tools > Sequence Similarity Searching > FASTA

## Protein Similarity Search

This tool provides sequence similarity searching against protein databases using the FASTA suite of programs. FASTA provides a heuristic search with a protein query. FASTX and FASTY translate a DNA query. Optimal searches are available with SSEARCH (local), GGSEARCH (global) and GLSEARCH (global query, local database).

### STEP 1 - Select your databases

PROTEIN DATABASES

1 Database Selected X Clear Selection

- UniProt Knowledgebase (The UniProt Knowledgebase includes UniProtKB/Swiss-Prot and UniProtKB/TrEMBL)
- UniProtKB/Swiss-Prot (The manually annotated section of UniProtKB)
- UniProtKB/Swiss-Prot isoforms (The manually annotated isoforms of UniProtKB/Swiss-Prot)
- UniProtKB/TrEMBL (The automatically annotated section of UniProtKB)
- UniProtKB Reference Proteomes plus Swiss-Prot

▶ UniProtKB Taxonomic Subsets

▶ UniProt Clusters

▶ Patents

▶ Structures

▶ Other Protein Databases

### STEP 2 - Enter your input sequence

Enter or paste a  sequence in any supported format:

or Upload a file:  No file chosen [Use a example sequence](#) | [Clear sequence](#) | [See more example inputs](#)

### STEP 3 - Set your parameters

PROGRAM	GAP OPEN		GAP EXTEND	KTUP	EXPECTATION UPPER VALUE	EXPECTATION
<input type="text" value="FASTA"/>	<input type="text" value="-10"/>	<input type="text" value="-2"/>	<input type="text" value="2"/>	<input type="text" value="10"/>	<input type="text" value="0 (default)"/>	

DNA STRAND	HISTOGRAM	FILTER	STATISTICAL ESTIMATES
<input type="text" value="N/A"/>	<input type="text" value="no"/>	<input type="text" value="none"/>	<input type="text" value="None"/>

<https://www.ebi.ac.uk/Tools/sss/fast/>

# FASTA no EBI - parâmetros

Feedback

Share

or Upload a file:  No file chosen

[Use a example sequence](#) | [Clear sequence](#) | [See more example inputs](#)

## STEP 3 - Set your parameters

### PROGRAM

FASTA

### MATRIX

BLOSUM50

### GAP OPEN

-10

### GAP EXTEND

-2

### KTUP

2

### EXPECTATION UPPER VALUE EXPECTATION

10

0 (default)

### DNA STRAND

N/A

### HISTOGRAM

no

### FILTER

none

### STATISTICAL ESTIMATES

Regress

### SCORES

50

### ALIGNMENTS

50

### SEQUENCE RANGE

START-END

### DATABASE RANGE

START-END

### MULTI HSPs

no

### SCORE FORMAT

Default

### ANNOTATION FEATURES

no

## STEP 4 - Submit your job

Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

Submit

<https://www.ebi.ac.uk/Tools/sss/fast/>

# FASTA no EBI

Feedback

Share

FASTX and FASTY translate a DNA query. Optimal searches are available with SSEARCH (local), GGSEARCH (global) and GLSEARCH (global query, local database).

## STEP 1 - Select your databases

### PROTEIN DATABASES

1 Database Selected

X Clear Selection

- UniProt Knowledgebase (The UniProt Knowledgebase includes UniProtKB/Swiss-Prot and UniProtKB/TrEMBL)
- UniProtKB/Swiss-Prot (The manually annotated section of UniProtKB)
- UniProtKB/Swiss-Prot isoforms (The manually annotated isoforms of UniProtKB/Swiss-Prot)
- UniProtKB/TrEMBL (The automatically annotated section of UniProtKB)
- UniProtKB Reference Proteomes plus Swiss-Prot
- ▶ UniProtKB Taxonomic Subsets
- ▶ UniProt Clusters
- ▶ Patents
- ▶ Structures
- ▶ Other Protein Databases

## STEP 2 - Enter your input sequence

Enter or paste a **PROTEIN** sequence in any supported format:

```
>sp|P49789|FHIT_HUMAN Bis(5'-adenosyl)-triphosphatase OS=Homo sapiens OX=9606 GN=FHIT PE=1 SV=3
MSFRFGQHLLKPSVVFLKTELSFALVNRKPVVPGHVLVCPLRPVERFHDLRPDEVADLFQ
TTQRVGTVVEKHFHGTSLTFSMQDGPAGQTVKHVHVHVLPRKAGDFHRNDSIYEELQKH
DKEDFPASWRSEEEEMAAEAAALRVYFQ
```

or Upload a file:  No file chosen

[Use a example sequence](#) | [Clear sequence](#) | [See more example inputs](#)

## STEP 3 - Set your parameters

PROGRAM

<https://www.ebi.ac.uk/Tools/sss/fasta/>

# FASTA no EBI

## FASTA

[Protein](#)[Nucleotide](#)[Genomes](#)[Proteomes](#)[Whole Genome Shotgun](#)[Web services](#)[Also in this section ▾](#)[Feedback](#)[Share](#)

Tools > Sequence Similarity Searching > FASTA

### Results for job fasta-l20191021-112509-0711-52636113-p1m

[Summary Table](#)[Tool Output](#)[Visual Output](#)[Functional Predictions](#)[Submission Details](#)

#### Selection:

[Select All](#) [Invert](#) [Clear](#)

#### Apply to selection:

##### Annotations:

[Show](#) [Hide](#)

##### Alignments:

[Show](#) [Hide](#)

##### Entries:

[Download](#) in

fasta ▾

format

##### Tools:

[Launch](#)

Clustal Omega ▾

Align.	DB:ID	Source	Length	Score (Bits)	Identities %	Positives %	E()
<input checked="" type="checkbox"/>	TR:A0A024R366_HUMAN	Fragile histidine triad gene, isoform CRA_a OS=Homo sapiens OX=9606 GN=FHIT PE=4 SV=1 <i>Cross-references and related information in:</i> <a href="#">▶ Gene expression</a> <a href="#">▶ Nucleotide sequences</a> <a href="#">▶ Genomes &amp; metagenomes</a> <a href="#">▶ Literature</a> <a href="#">▶ Samples &amp; ontologies</a> <a href="#">▶ Protein families</a> <a href="#">▶ Protein expression data</a> <a href="#">▶ Protein sequences</a>	147	252.0	100.0	100.0	5.5E-64
<input checked="" type="checkbox"/>	SP:FHIT_HUMAN	Bis(5'-adenosyl)-triphosphatase OS=Homo sapiens OX=9606 GN=FHIT PE=1 SV=3 <i>Cross-references and related information in:</i> <a href="#">▶ Gene expression</a> <a href="#">▶ Bioactive molecules</a> <a href="#">▶ Nucleotide sequences</a> <a href="#">▶ Genomes &amp; metagenomes</a> <a href="#">▶ Enzymes</a> <a href="#">▶ Literature</a> <a href="#">▶ Samples &amp; ontologies</a> <a href="#">▶ Molecular interactions</a> <a href="#">▶ Protein families</a> <a href="#">▶ Diseases</a> <a href="#">▶ Macromolecular structures</a> <a href="#">▶ Protein expression data</a> <a href="#">▶ Reactions &amp; pathways</a> <a href="#">▶ Protein sequences</a>	147	252.0	100.0	100.0	5.5E-64
<input checked="" type="checkbox"/>	TR:A0A2J8P7H3_PANTR	FHIT isoform 1 OS=Pan troglodytes OX=9598 GN=FHIT PE=4 SV=1 <i>Cross-references and related information in:</i> <a href="#">▶ Nucleotide sequences</a> <a href="#">▶ Genomes &amp; metagenomes</a> <a href="#">▶ Literature</a> <a href="#">▶ Samples &amp; ontologies</a> <a href="#">▶ Protein families</a> <a href="#">▶ Protein sequences</a>	147	248.0	98.6	99.3	8.4E-63
<input checked="" type="checkbox"/>	TR:A0A2I2YJR3_GORGO	Fragile histidine triad OS=Gorilla gorilla gorilla OX=9595 GN=FHIT PE=4 SV=1 <i>Cross-references and related information in:</i> <a href="#">▶ Genomes &amp; metagenomes</a> <a href="#">▶ Literature</a> <a href="#">▶ Samples &amp; ontologies</a> <a href="#">▶ Protein families</a> <a href="#">▶ Protein sequences</a>	147	245.8	97.3	99.3	3.9E-62

<https://www.ebi.ac.uk/Tools/sss/fasta/>

# FASTA no EBI

## FASTA

[Protein](#)[Nucleotide](#)[Genomes](#)[Proteomes](#)[Whole Genome Shotgun](#)[Web services](#)[Also in this section](#) ▾[Feedback](#)

Tools > Sequence Similarity Searching > FASTA

### Results for job fasta-l20191021-112509-0711-52636113-p1m

[Summary Table](#)[Tool Output](#)[Visual Output](#)[Functional Predictions](#)[Submission Details](#)[Download](#)[Download in XML format](#)

FASTA searches a protein or DNA sequence data bank  
version 36.3.8g Dec, 2017  
Please cite:  
W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

Query: @  
1>>>sp|P49789|FHIT\_HUMAN Bis(5'-adenosyl)-triphosphatase OS=Homo sapiens OX=9606 GN=FHIT PE=1 SV=3 - 147 aa  
Library: UniProt Knowledgebase  
60572120266 residues in 180781033 sequences

Statistics: Expectation\_n fit:  $\rho(\ln(x)) = 6.8884 \pm 0.000171$ ;  $\mu = 5.5199 \pm 0.009$   
 $\text{mean\_var} = 57.0663 \pm -11.317$ ,  $\theta$ 's: 203 Z-trim(117.5): 306 B-trim: 5238 in 2/62  
Lambda= 0.169779  
statistics sampled from 60000 (122918) to 30932256 sequences  
Algorithm: FASTA (3.8 Nov 2011) [optimized]  
Parameters: BL50 matrix (15:-5), open/ext: -10/-2  
ktp: 2, E-join: 1 (0.536), E-opt: 0.2 (0.172), width: 16  
Scan time: 2417.080

The best scores are: opt bits E(180781033)

[E]TR:A0A024R366_HUMAN	A0A024R366	Fragile histidine t ( 147)	996	252.0	5.5e-64
[E]SP:FHIT_HUMAN	P49789	Bis(5'-adenosyl)-triphosphata ( 147)	996	252.0	5.5e-64
[E]TR:A0A2J8P7H3_PANTR	A0A2J8P7H3	FHIT isoform 1 OS=P ( 147)	980	248.0	8.4e-63
[E]TR:A0A2I2YJR3_GORGO	A0A2I2YJR3	Fragile histidine t ( 147)	971	245.8	3.9e-62
[E]TR:H2PAD8_PONAB	H2PAD8	Fragile histidine triad OS= ( 147)	966	244.6	9e-62
[E]TR:A0A2J8TEL9_PONAB	A0A2J8TEL9	FHIT isoform 1 OS=P ( 147)	961	243.4	2.1e-61
[E]TR:A0A2K5S0M6_CEBCA	A0A2K5S0M6	HIT domain-containi ( 149)	954	241.7	7e-61
[E]TR:A0A2K6TUV1_SAIIB	A0A2K6TUV1	Fragile histidine t ( 149)	952	241.2	9.9e-61
[E]TR:A0A2K5DD36_AOTNA	A0A2K5DD36	HIT domain-containi ( 149)	949	240.4	1.6e-60
[E]TR:A0A2K6N901_RHIRO	A0A2K6N901	HIT domain-containi ( 149)	945	239.5	3.2e-60
[E]TR:A0A2K6K2R6_RHIBE	A0A2K6K2R6	HIT domain-containi ( 149)	945	239.5	3.2e-60
[E]TR:F6TCB7_MACMU	F6TCB7	Bis(5'-adenosyl)-triphospha ( 149)	940	238.2	7.6e-60
[E]TR:A0A2K5MIU2_CERAT	A0A2K5MIU2	HIT domain-containi ( 149)	940	238.2	7.6e-60
[E]TR:A0A2K6DGY6_MACNE	A0A2K6DGY6	HIT domain-containi ( 149)	940	238.2	7.6e-60
[E]TR:A0A096MQ24_PAPAN	A0A096MQ24	HIT domain-containi ( 149)	940	238.2	7.6e-60
[E]TR:A0A2K5TOF0_MARFA	A0A2K5TOF0	HIT domain-containi ( 149)	940	238.2	7.6e-60

<https://www.ebi.ac.uk/Tools/sss/fasta/>

# FASTA no EBI

```

>>TR:A0A2K6EP68_PROCO A0A2K6EP68 HIT domain-containi ( 148) 905 229.7 2.9e-57
>>TR:L5L555_MYODS L5L555 Bis(5'-adenosyl)-triphospha ( 149) 905 229.7 2.9e-57
>>TR:A0A3Q7Q453_CALUR A0A3Q7Q453 bis(5'-adenosyl)-tr ( 149) 903 229.2 4e-57
>>TR:G7MKZ1_MACMU G7MKZ1 HIT domain-containing prote ( 141) 900 228.5 6.3e-57
>>TR:G7NZB0_MACFA G7NZB0 HIT domain-containing prote ( 141) 900 228.5 6.3e-57
>>TR:A0A0N8ESW3_HETGA A0A0N8ESW3 Bis(5'-adenosyl)-tr ( 149) 899 228.2 8e-57
>>SP:FHIT_BOVIN Q1KZG4 Bis(5'-adenosyl)-triphosphata ( 149) 898 228.0 9.5e-57
>>TR:A0A4W2H904_BOBOX A0A4W2H904 HIT domain-containi ( 149) 898 228.0 9.5e-57
>>TR:A0A384AYQ4_BALAS A0A384AYQ4 bis(5'-adenosyl)-tr ( 149) 898 228.0 9.5e-57
>>TR:A0A4X1VGP0_PIG A0A4X1VGP0 Fragile histidine tri ( 149) 898 228.0 9.5e-57
>>TR:A0A384AZE5_BALAS A0A384AZE5 bis(5'-adenosyl)-tr ( 171) 898 227.9 1.1e-56

>>TR:A0A024R366_HUMAN A0A024R366 Fragile histidine triad gene,
isoform CRA_a OS=Homo sapiens OX=9606 GN=FHIT PE=4 SV=1 (147 aa)
initn: 996 init1: 996 opt: 996 Z-score: 1329.4 bits: 252.0 E(180781033): 5.5e-64
Smith-Waterman score: 996; 100.0% identity (100.0% similar) in 147 aa overlap (1-147:1-147)

      10      20      30      40      50      60
sp|P49 MSFRFGQHLIKPSVVFLKTELSFALVNRKPVVPGHVLVCLRPVERFHDLRPDEVADLFQ
      :
      :
      :
TR:A0A MSFRFGQHLIKPSVVFLKTELSFALVNRKPVVPGHVLVCLRPVERFHDLRPDEVADLFQ
      10      20      30      40      50      60

      70      80      90      100     110     120
sp|P49 TTQRVGTVEKHFHGTSLTFSMQDGPEAGQTVKHVHVHVLPRKAGDFHRNDSIYEELQKH
      :
      :
      :
TR:A0A TTQRVGTVEKHFHGTSLTFSMQDGPEAGQTVKHVHVHVLPRKAGDFHRNDSIYEELQKH
      70      80      90      100     110     120

      130     140
sp|P49 DKEDFPASWRSEEEEMAAEAAALRVYFQ
      :
      :
      :
TR:A0A DKEDFPASWRSEEEEMAAEAAALRVYFQ
      130     140

>>SP:FHIT_HUMAN P49789 Bis(5'-adenosyl)-triphosphatase OS=Homo
sapiens OX=9606 GN=FHIT PE=1 SV=3 (147 aa)
initn: 996 init1: 996 opt: 996 Z-score: 1329.4 bits: 252.0 E(180781033): 5.5e-64
Smith-Waterman score: 996; 100.0% identity (100.0% similar) in 147 aa overlap (1-147:1-147)

      10      20      30      40      50      60
sp|P49 MSFRFGQHLIKPSVVFLKTELSFALVNRKPVVPGHVLVCLRPVERFHDLRPDEVADLFQ
      :
      :
      :
SP:FHI MSFRFGQHLIKPSVVFLKTELSFALVNRKPVVPGHVLVCLRPVERFHDLRPDEVADLFQ
      10      20      30      40      50      60

      70      80      90      100     110     120
sp|P49 TTQRVGTVEKHFHGTSLTFSMQDGPEAGQTVKHVHVHVLPRKAGDFHRNDSIYEELQKH
      :
      :
      :
SP:FHI TTQRVGTVEKHFHGTSLTFSMQDGPEAGQTVKHVHVHVLPRKAGDFHRNDSIYEELQKH
      70      80      90      100     110     120

      130     140
sp|P49 DKEDFPASWRSEEEEMAAEAAALRVYFQ
      :
      :
      :
CD-FHI DKEDFPASWRSEEEEMAAEAAALRVYFQ

```

<https://www.ebi.ac.uk/Tools/sss/fasta/>





# Estatística de pesquisa com FASTA

- **E-value:** número de sequências aleatórias que é esperado possuírem um score superior ao observado. Este número aumenta com a dimensão da base de dados. Deve ser *peelo menos*  $< 0.01$  para podermos confiar no resultado do alinhamento.
- **z-score:** número de desvios padrões do score relativamente à média da distribuição. No programa FASTA os z-scores são normalizados para valores  $z'$ , dados por:

$$z' = 50 + 10 * z$$

- **Score otimizado:** o score produzido pelo passo final do programa.
- **ktup:** dimensão do conjunto de posições que é indexado por hashing, é também a dimensão do mínimo pormenor que o algoritmo é capaz de detectar na sequência. Para proteínas usa-se geralmente  $ktup=2$ , mas  $ktup=1$  produz buscas mais sensíveis e capazes de identificar similaridades mais distantes.

# BLAST

- O programa BLAST (Altschul et al, 1990) é actualmente o mais usado, devido a sua maior rapidez, conseguida a troco de uma sensibilidade um pouco inferior à de FASTA. O fundamento estatístico do método é bastante rigoroso e na sua versão iterativa PSI-BLAST (Position-Specific Iterated BLAST) é extremamente sensível, graças ao uso de uma matriz de score cujos valores dependem da localização dos aminoácidos e que é refinada em iterações sucessivas.
- O método baseia-se na localização de pequenas regiões de elevada similaridade que são prolongadas nas duas direcções de forma a maximizar o score
- O tratamento estatístico é baseada no modelo de distribuição de valor extremo de Karlin-Altschul, sendo rigoroso para alinhamentos locais sem gaps, e aproximado nas outras situações.

# NCBI BLAST

NIH U.S. National Library of Medicine NCBI Sign in to NCBI

**BLAST**® Home Recent Results Saved Strategies Help

## 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**

**End of updates for BLAST+ version 4 databases (dbV4)**  
Start moving to the new version 5 databases!  
Fri, 27 Sep 2019 16:00:00 EST [More BLAST news...](#)

## Web BLAST

**Nucleotide BLAST**  
nucleotide ► nucleotide

**blastx**  
translated nucleotide ► protein

**tblastn**  
protein ► translated nucleotide

**Protein BLAST**  
protein ► protein

## BLAST Genomes

**Search**

<https://blast.ncbi.nlm.nih.gov/Blast.cgi>

# NCBI BLAST

NIH U.S. National Library of Medicine NCBI National Center for Biotechnology Information Sign in to NCBI

**BLAST** ® » **blastp suite** Home Recent Results Saved Strategies Help

### Standard Protein BLAST

blastn **blastp** blastx tblastn tblastx

BLASTP programs search protein databases using a protein 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  No file chosen [Choose File](#)

Job Title   
Enter a descriptive title for your BLAST search [Help](#)

**Align two or more sequences** [Help](#)

**Choose Search Set**

Database  [Help](#)

Organism Optional   exclude   
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown. [Help](#)

Exclude Optional  Models (XM/XP)  Non-redundant RefSeq proteins (WP)  Uncultured/environmental sample sequences

**Program Selection**

Algorithm

- Quick BLASTP (Accelerated protein-protein BLAST)
- blastp** (protein-protein BLAST)
- PSI-BLAST (Position-Specific Iterated BLAST)
- PHI-BLAST (Pattern Hit Initiated BLAST)
- DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm [Help](#)


Search database nr using **Blastp** (protein-protein BLAST)

Show results in a new window

[+ Algorithm parameters](#)

**BLAST results will be displayed in a new format by default**

You can always switch back to the Traditional Results page.



<https://blast.ncbi.nlm.nih.gov/Blast.cgi>

# NCBI BLAST

**BLAST** Search **database nr** using **Blastp (protein-protein BLAST)**  
 Show results in a new window

**Algorithm parameters** [Restore default search parameters](#)

**General Parameters**

**Max target sequences**  Select the maximum number of aligned sequences to display

**Short queries**  Automatically adjust parameters for short input sequences

**Expect threshold**

**Word size**

**Max matches in a query range**

**Scoring Parameters**

**Matrix**

**Gap Costs**

**Compositional adjustments**

**Filters and Masking**

**Filter**  Low complexity regions

**Mask**  Mask for lookup table only  
 Mask lower case letters

**BLAST** Search **database nr** using **Blastp (protein-protein BLAST)**  
 Show results in a new window

<https://blast.ncbi.nlm.nih.gov/Blast.cgi>

# Versões de BLAST

- **BLASTP**

pesquisa uma *sequência de aminoácidos* em um banco de *sequências de proteína*.

- **BLASTN**

pesquisa uma *sequência de nucleótidos* em um banco de *sequências de nucleótidos* (DNA ou RNA)

- **BLASTX**

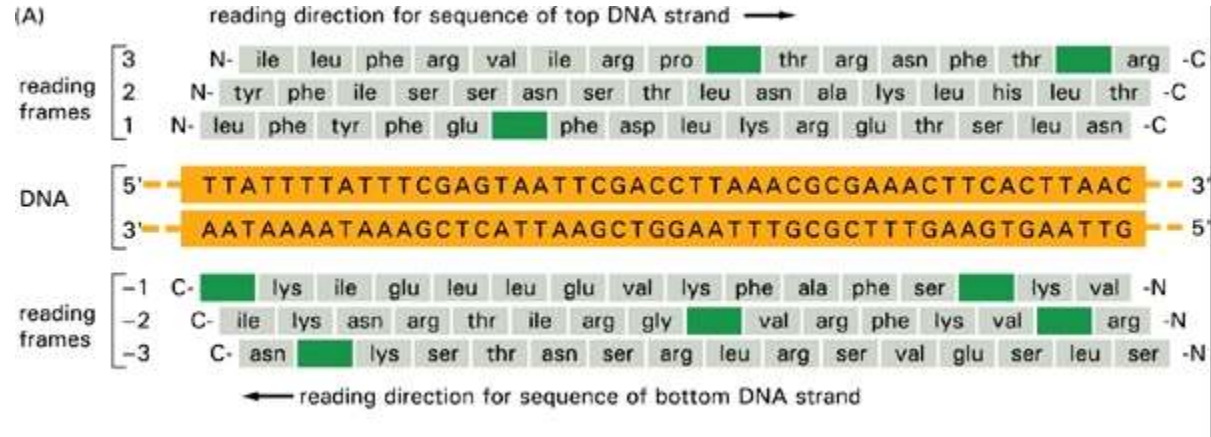
pesquisa uma *sequência nucleotídica* traduzida em todos os 6 quadros de leitura em um banco de *sequências de proteína*. Permite encontrar potenciais produtos da tradução de uma *sequência nucleotídica* desconhecida.

- **TBLASTN**

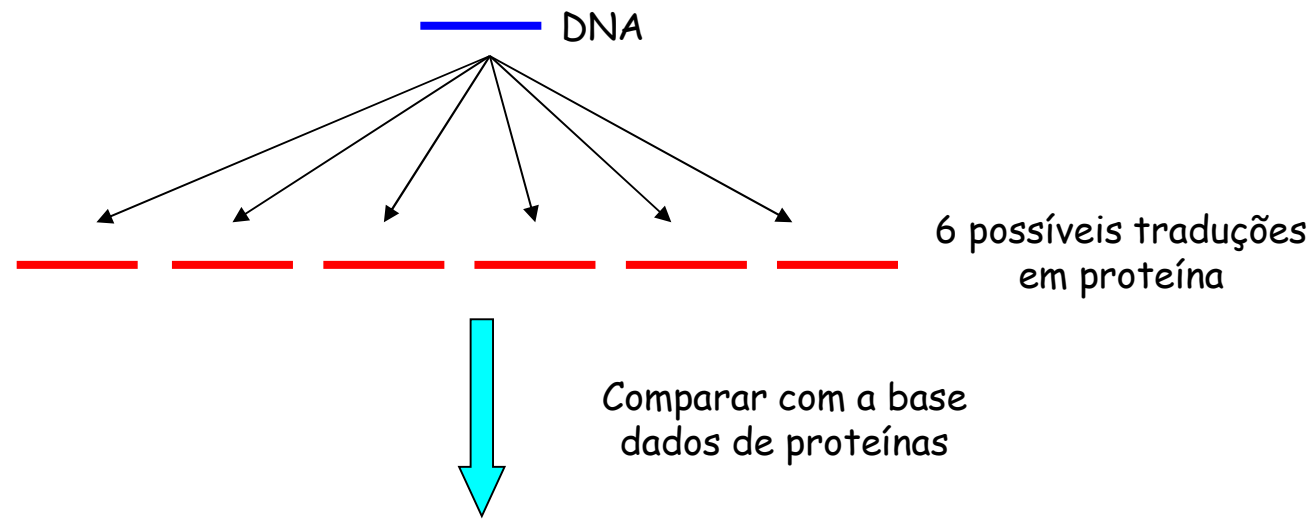
pesquisa uma *sequência de proteína* em um banco de *sequências nucleotídicas*, dinamicamente traduzidas nos 6 quadros de leitura

- **TBLASTX**

pesquisa os 6 possíveis quadros de leitura de uma *sequência nucleotídica* num banco de *sequências nucleotídicas*, dinamicamente traduzidas nos seus 6 quadros de leitura. Extremamente lento e pesado computacionalmente.



BLASTX





# Exemplo de pesquisa com o programa BLAST

[Distance tree of results](#) [NEW](#) [Related Structures](#)

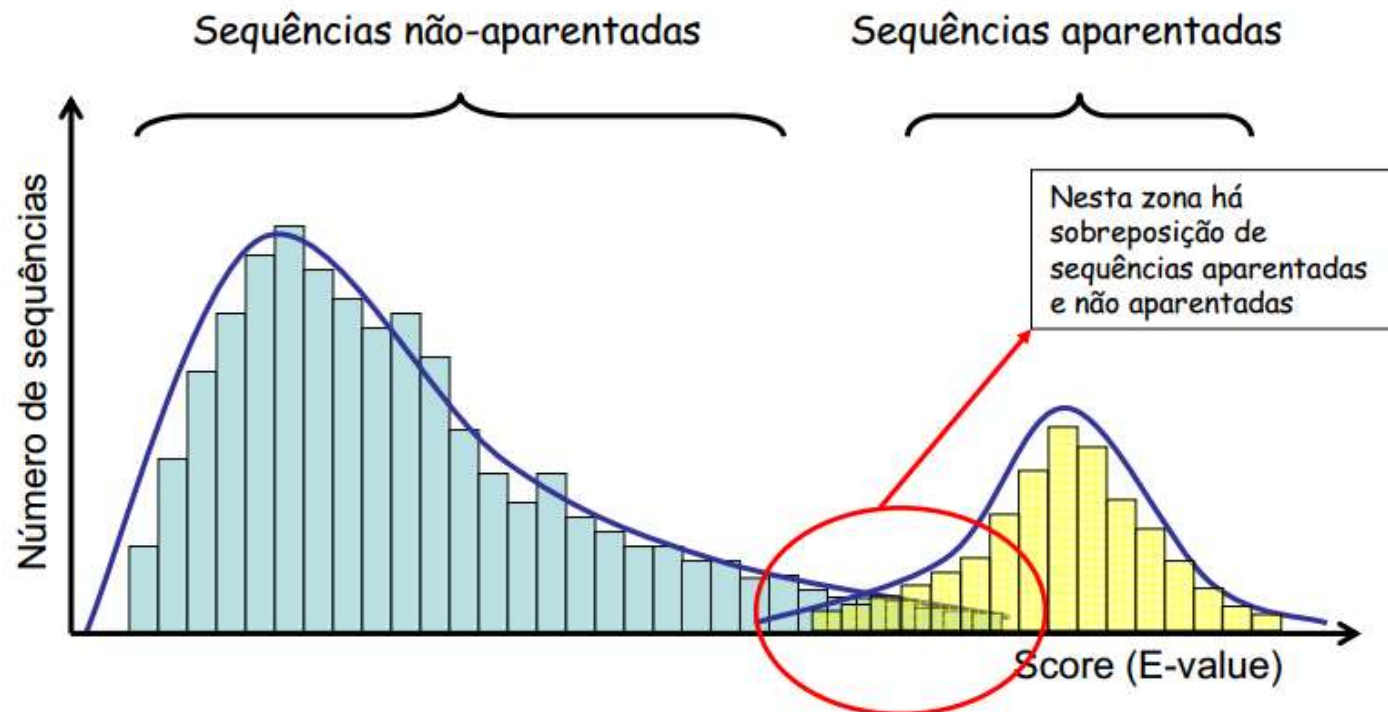
Sequences producing significant alignments:

	Score	E
	(Bits)	Value
<a href="#">sp P11540 BARS_BACAM</a> Barstar (Ribonuclease inhibitor) >gi 393...	186	5e-46
<a href="#">pdb 1B3S D</a> Chain D, Structural Response To Mutation At A Prot...	184	1e-45
<a href="#">pdb 1B2S D</a> Chain D, Structural Response To Mutation At A Prot...	184	2e-45
<a href="#">pdb 1AY7 B</a> Chain B, Ribonuclease Sa Complex With Barstar >pdb...	183	3e-45
<a href="#">pdb 1B2U D</a> Chain D, Structural Response To Mutation At A Prot...	183	3e-45
<a href="#">ref YP_001420464.1 </a> putative RNase inhibitor [Bacillus amylo...	183	3e-45
<a href="#">pdb 1A19 A</a> Chain A, Barstar (Free), C62a Mutant >pdb 1A19 B C...	183	3e-45
<a href="#">pdb 1B27 D</a> Chain D, Structural Response To Mutation At A Prot...	181	1e-44
<a href="#">pdb 1X1W D</a> Chain D, Water-Mediate Interaction At Aprotein-Pro...	179	5e-44
<a href="#">pdb 1BG5 E</a> Chain E, Recognition Between A Bacterial Ribonucle...	178	7e-44
<a href="#">pdb 1X1Y D</a> Chain D, Water-Mediate Interaction At Aprotein-Pro...	178	8e-44
<a href="#">pdb 1X1X D</a> Chain D, Water-Mediate Interaction At Aprotein-Pro...	176	3e-43
<a href="#">pdb 2HXX A</a> Chain A, Aminotryptophan Barstar >pdb 2HXX B Chain...	167	1e-40
<a href="#">ref YP_080807.1 </a> putative ribonuclease inhibitor YrdF [Bacill...	106	5e-22
<a href="#">ref YP_001422650.1 </a> YrdF [Bacillus amyloliquefaciens FZB42] >...	92.4	9e-18
<a href="#">ref YP_001307837.1 </a> Barstar (barnase inhibitor) [Clostridium ...	88.6	1e-16
<a href="#">ref NP_390550.1 </a> hypothetical protein BSU26730 [Bacillus subt...	88.2	1e-16
<a href="#">ref YP_001486227.1 </a> possible barstar-like protein ribonucleas...	83.2	5e-15
<a href="#">gb ABK00570.1 </a> putative ribonuclease inhibitor [Bacillus cereu...	80.5	3e-14
<a href="#">ref NP_347480.1 </a> Barstar-like protein ribonuclease (barnase) ...	80.1	4e-14
<a href="#">gb ABK00989.1 </a> putative ribonuclease inhibitor [Bacillus cere...	79.7	5e-14
<a href="#">ref YP_385561.1 </a> Barstar [Geobacter metallireducens GS-15] >g...	64.3	2e-09
<a href="#">ref YP_957328.1 </a> Barstar (barnase inhibitor) [Marinobacter aq...	62.0	1e-08
<a href="#">ref NP_273688.1 </a> ribonuclease inhibitor barstar [Neisseria me...	56.2	6e-07
<a href="#">ref NP_900871.1 </a> probable Barstar [Chromobacterium violaceum ...	54.7	2e-06
<a href="#">ref YP_001467473.1 </a> barstar (barnase inhibitor) superfamily [...	51.6	1e-05
<a href="#">ref ZP_00738603.1 </a> Ribonuclease inhibitor barstar [Bacillus t...	51.2	2e-05
<a href="#">ref YP_001346964.1 </a> barstar [Pseudomonas aeruginosa PA7] >gb ...	51.2	2e-05
<a href="#">ref YP_001007942.1 </a> putative ribonuclease inhibitor [Yersinia pe...	50.4	3e-05
<a href="#">ref NP_407142.1 </a> putative ribonuclease inhibitor [Yersinia pe...	50.1	5e-05
<a href="#">ref ZP_00828991.1 </a> COG2732: Barstar, RNase (barnase) inhibito...	49.7	5e-05
<a href="#">ref YP_624916.1 </a> Barstar (barnase inhibitor) [Burkholderia ce...	49.3	8e-05
<a href="#">ref ZP_00834161.1 </a> COG2732: Barstar, RNase (barnase) inhibito...	49.3	8e-05
<a href="#">ref ZP_00823899.1 </a> COG2732: Barstar, RNase (barnase) inhibito...	48.9	9e-05
<a href="#">ref YP_259489.1 </a> possible ribonuclease inhibitor [Pseudomonas...	48.5	1e-04
<a href="#">ref ZP_00822864.1 </a> COG2732: Barstar, RNase (barnase) inhibito...	48.5	1e-04
<a href="#">gb EAZ59705.1 </a> hypothetical protein PA2G_03002 [Pseudomonas aeru...	48.1	2e-04
<a href="#">ref ZP_00967937.1 </a> COG2732: Barstar, RNase (barnase) inhibito...	48.1	2e-04
<a href="#">ref NP_628172.1 </a> ribonuclease inhibitor [Streptomyces coelic...	47.8	2e-04
<a href="#">gb AAN16065.1 </a> possible ribonuclease inhibitor [Pseudomonas stut...	47.4	3e-04
<a href="#">ref YP_447728.1 </a> hypothetical protein Msp_0687 [Methanospaer...	47.4	3e-04
<a href="#">ref NP_931254.1 </a> hypothetical protein plu4062 [Photorhabdus 1...	47.0	4e-04
<a href="#">ref YP_560425.1 </a> hypothetical protein Bxe_A0560 [Burkholderia...	46.6	5e-04
<a href="#">ref ZP_01512999.1 </a> Barstar (barnase inhibitor) [Burkholderia ...	46.2	5e-04
<a href="#">gb AAC38289.1 </a> ribonuclease inhibitor [Streptomyces aureofaciens	45.4	0.001
<a href="#">ref YP_001580933.1 </a> hypothetical protein Emul_2752 [Burkholde...	44.7	0.002
<a href="#">ref ZP_00985480.1 </a> hypothetical protein Bdo1A_01002754 [Burkh...	44.3	0.002

<a href="#">ref YP_001118454.1 </a> hypothetical protein Bcep1808_0607 [Burkh...	44.3	0.002
<a href="#">ref YP_441738.1 </a> barstar family protein [Burkholderia thailan...	43.9	0.003
<a href="#">ref YP_334843.1 </a> barstar family protein [Burkholderia pseudom...	43.9	0.003
<a href="#">ref YP_109551.1 </a> hypothetical protein BPSL2957 [Burkholderia ...	43.5	0.004
<a href="#">ref YP_104020.1 </a> barstar family protein [Burkholderia mallei ...	43.5	0.004
<a href="#">ref YP_347771.1 </a> ribonuclease inhibitor barstar [Pseudomonas ...	42.4	0.009
<a href="#">ref ZP_01165878.1 </a> hypothetical protein MED92_10709 [Oceanosp...	42.0	0.012
<a href="#">ref ZP_01075124.1 </a> possible ribonuclease inhibitor [Marinomon...	40.8	0.023
<a href="#">ref NP_879850.1 </a> hypothetical protein BP1066 [Bordetella pert...	40.8	0.025
<a href="#">ref YP_001480610.1 </a> Barstar (barnase inhibitor) [Serratia pro...	40.8	0.025
<a href="#">ref YP_001632237.1 </a> hypothetical protein Bpet3626 [Bordetella...	40.8	0.026
<a href="#">ref YP_297050.1 </a> hypothetical protein Reut_A2845 [Ralstonia e...	40.8	0.028
<a href="#">ref YP_727594.1 </a> hypothetical protein H16_A3151 [Ralstonia eu...	40.4	0.032
<a href="#">ref YP_551634.1 </a> hypothetical protein Bpro_4860 [Polaromonas ...	40.0	0.039
<a href="#">pdb 1L1K A</a> Chain A, Nmr Identification And Characterization O...	40.0	0.040
<a href="#">ref YP_001019402.1 </a> hypothetical protein Mpe_A0205 [Methylibi...	40.0	0.048
<a href="#">gb ABA55898.1 </a> putative ribonuclease inhibitor [Vibrio sp. DAT72	40.0	0.049
<a href="#">ref YP_367960.1 </a> hypothetical protein Bcep18194_A3715 [Burkho...	39.3	0.068
<a href="#">emb CAJ48472.1 </a> putative ribonuclease inhibitor [Bordetella avi...	39.3	0.072
<a href="#">ref YP_988286.1 </a> hypothetical protein Ajs_4108 [Acidovorax sp...	39.3	0.075
<a href="#">ref YP_973066.1 </a> hypothetical protein Aave_4761 [Acidovorax a...	39.3	0.076
<a href="#">ref YP_001354261.1 </a> hypothetical protein mma_2571 [Janthinoba...	38.9	0.088
<a href="#">ref YP_133861.1 </a> hypothetical protein PRSB101_28 [uncultured ...	38.9	0.093
<a href="#">ref YP_001567045.1 </a> hypothetical protein Daci_6032 [Delftia a...	38.9	0.094
<a href="#">ref YP_001100725.1 </a> hypothetical protein HEAR2476 [Herminiimo...	38.9	0.10
<a href="#">ref YP_001100903.1 </a> hypothetical protein HEAR2660 [Herminiimo...	38.9	0.10
<a href="#">ref NP_520886.1 </a> hypothetical protein RSc2765 [Ralstonia sola...	38.1	0.17
<a href="#">ref ZP_00943824.1 </a> Hypothetical Protein RRSL_02692 [Ralstonia...	38.1	0.17
<a href="#">ref YP_984284.1 </a> hypothetical protein Pnap_4072 [Polaromonas ...	38.1	0.17
<a href="#">ref YP_585127.1 </a> hypothetical protein Rmet_2985 [Ralstonia me...	38.1	0.18
<a href="#">ref YP_999519.1 </a> hypothetical protein Veis_4809 [Verminephrob...	37.7	0.19
<a href="#">ref YP_001354585.1 </a> hypothetical protein mma_2895 [Janthinoba...	37.7	0.20
<a href="#">ref YP_703323.1 </a> possible barnase inhibitor [Rhodococcus sp. ...	37.7	0.24
<a href="#">emb CAJ42314.1 </a> hypothetical protein [Streptomyces steffiasburgen	37.4	0.27
<a href="#">ref ZP_01520139.1 </a> conserved hypothetical protein [Comamonas ...	37.4	0.29
<a href="#">ref ZP_01694365.1 </a> hypothetical protein M23134_00594 [Microsc...	37.0	0.34
<a href="#">ref ZP_02007004.1 </a> conserved hypothetical protein [Ralstonia ...	37.0	0.36
<a href="#">ref ZP_01660530.1 </a> conserved hypothetical protein [Ralstonia ...	37.0	0.36
<a href="#">ref ZP_01504924.1 </a> conserved hypothetical protein [Burkholder...	37.0	0.37
<a href="#">ref ZP_00749220.1 </a> Sti [Beggiatoa sp. PS] >gb EDN65480.1  Sti [B	37.0	0.39
<a href="#">ref YP_001103709.1 </a> hypothetical protein SACT_462 [Saccharop...	36.6	0.46
<a href="#">ref YP_860993.1 </a> hypothetical protein GFC_952 [Gramella fors...	35.8	0.79
<a href="#">ref YP_971688.1 </a> hypothetical protein Aave_3344 [Acidovorax a...	35.4	1.1
<a href="#">ref YP_884624.1 </a> ribonuclease inhibitor [Mycobacterium smegma...	35.0	1.4
<a href="#">ref YP_001623583.1 </a> hypothetical protein RSal33209_0415 [Reni...	34.7	1.9
<a href="#">ref XP_001511604.1 </a> PREDICTED: hypothetical protein [Ornithorhyn	33.9	2.9
<a href="#">ref XP_001604032.1 </a> PREDICTED: similar to jumonji domain cont...	33.9	3.6
<a href="#">emb CAM13843.1 </a> heterogeneous nuclear ribonucleoprotein R [Mus m	33.5	4.0
<a href="#">ref YP_297129.1 </a> hypothetical protein Reut_A2925 [Ralstonia e...	33.5	4.1
<a href="#">ref YP_579588.1 </a> hypothetical protein Pcryo_0320 [Psychrobact...	33.5	4.4
<a href="#">ref ZP_02170649.1 </a> phenylalanyl-tRNA synthetase, beta subunit...	33.5	4.6

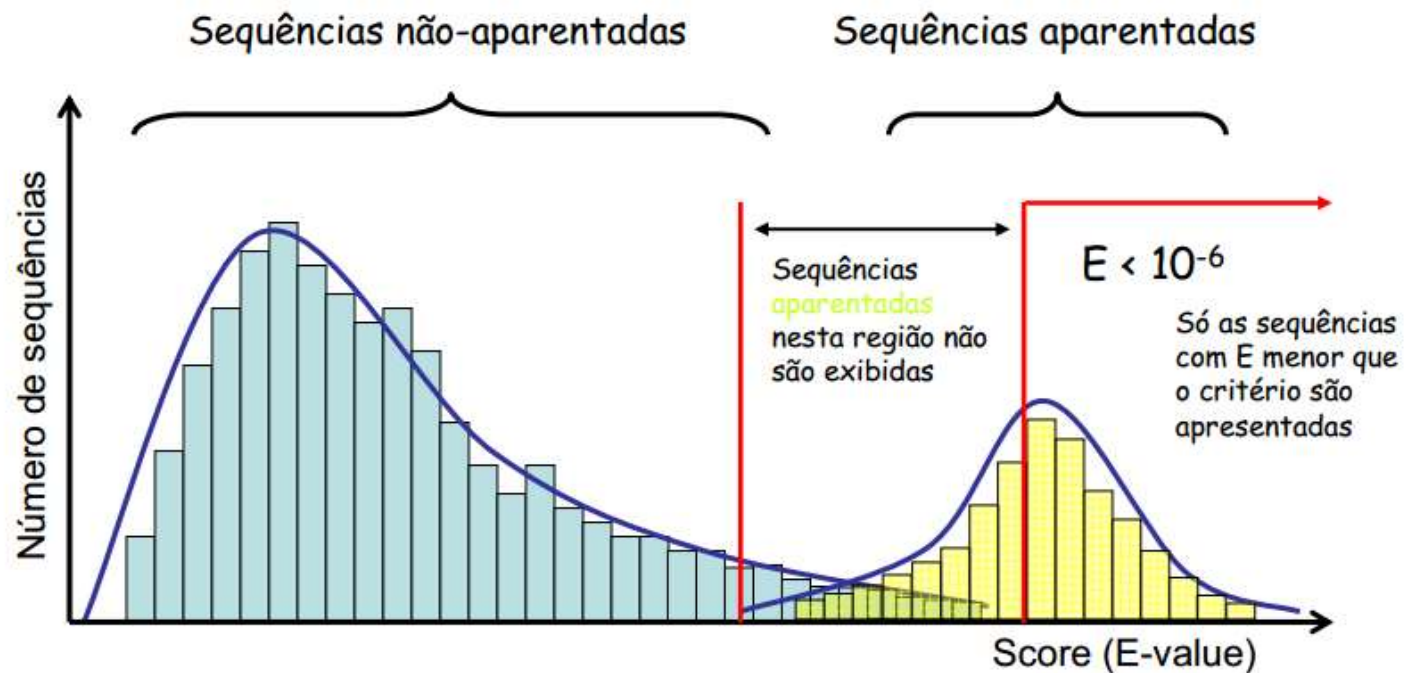
## Sensibilidade vs. Selectividade: a escolha do "E-value"

- Em geral a maioria das sequências numa base de dados não são aparentadas com a nossa sequência de busca e produzem uma distribuição de scores semelhante à de um conjunto de alinhamentos aleatórios
- As sequências aparentadas deverão produzir scores bem superiores à distribuição. No entanto existe quase sempre uma zona de sobreposição das duas regiões



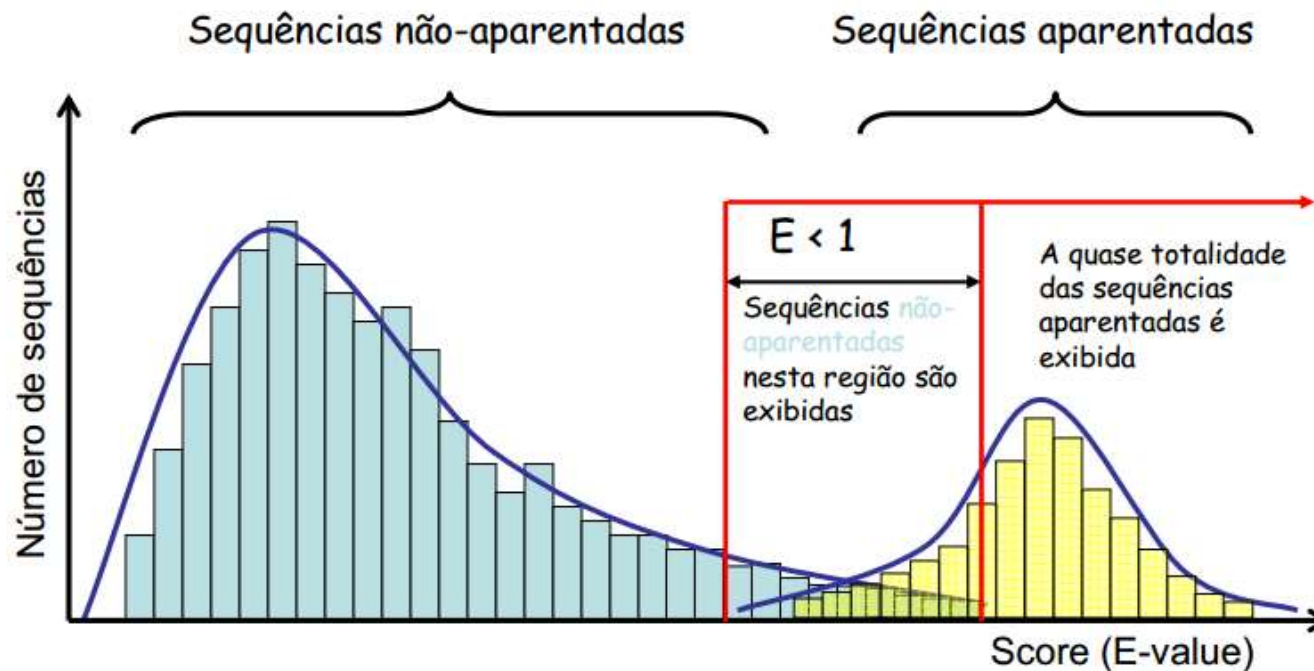
## Sensibilidade vs. Selectividade: pesquisa muito selectiva mas pouco sensível

- Se impusermos um valor de E máximo muito pequeno para a apresentação dos resultados, iremos garantir que TODAS as seqüências apresentadas estão relacionadas com a nossa seqüência de busca (alta selectividade)
- Seqüências aparentadas com a nossa seqüência de busca e possuindo um E value superior ao critério não serão apresentadas (baixa sensibilidade)



## Sensibilidade vs. Selectividade: pesquisa muito sensível mas pouco selectiva

- “Relaxando” o critério de busca, ou seja permitindo a exibição de seqüências com um E value mais alto iremos conseguir identificar a quase totalidade das seqüências aparentadas com a nossa seqüência de busca (alta sensibilidade)
- Como consequência da relaxação do critério, irão ser exibidas muitas seqüências **não-aparentadas** com a nossa seqüência de busca (baixa selectividade)



# Pesquisas mais sensíveis

O uso de métodos heurísticos acelera as pesquisas mas sacrifica sensibilidade e pode impedir a identificação de relações distantes entre sequências. Para as situações em que é necessária sensibilidade adicional, dispomos de algumas aplicações que realizam alinhamentos de Smith-Waterman contra toda a base de sequências, sem filtragem prévia por métodos heurísticos:

- ~~MPSearch:~~ <http://www.ebi.ac.uk/MPsrch/>
- Ssearch (parte do conjunto de aplicações de FASTA):
- <https://www.ebi.ac.uk/Tools/sss/fast/>
- <http://molsim.sci.univr.it/bioinfo/tools/OlfactionDB/ssearch.html>
- [https://fasta.bioch.virginia.edu/fasta\\_www2/fasta\\_www.cgi?rm=select&pgm=sw](https://fasta.bioch.virginia.edu/fasta_www2/fasta_www.cgi?rm=select&pgm=sw)

Note-se que esta abordagem é muito mais pesada computacionalmente, pelo que a resposta é mais demorada e o uso intensivo dos servidores é geralmente desencorajado pelos seus responsáveis. Este tipo de serviços deve pois ser usado de forma moderada.

# Regras gerais para pesquisa

- Usar uma sequência de proteína, ou converter DNA para proteína sempre que possível
- Utilizar a base de dados mais pequena que possa conter as sequências que pretendemos encontrar, no caso de não termos resultados então passar para uma base de dados maiores
- Escolha do esquema da *gap penalty* compatível com o modelo estatístico usado, nomeadamente com a matriz de score usada.
- Sempre que possível verificar a qualidade do ajuste da distribuição de scores aleatórios à distribuição de valor extrema produzida pelo software de alinhamento
- Procura valores de  $E < 0.01$ , para valores superiores o número de falsos positivos torna-se importante
- No caso de realizar um grande número de buscas devemos usar valores de  $E$  mais pequenos, pois o número de falsos positivos cresce linearmente com o número de pesquisas