

Alineamiento local: búsqueda de homologías

Supongamos que el material de partida para realizar una búsqueda de homologías no es un gen o una proteína completos y bien caracterizados de los que podamos usar una clave de acceso o una palabra clave, sino que solo disponemos de un oligonucleótido:

TACAGCAGATAGCAGCCATAGCCGCATACGTCGCGACTAC...

O bien de un oligopéptido:

PTWRVPGRMEKWHALVKYLKYRTKDLEEV...

¿Cómo saber entonces si existe algún gen o proteína similar a ellos en la base de datos?

Para responder a esto, necesitamos hacer un 'rastreo' de la base de datos.

El alineamiento completo (**global**) de dos secuencias (Smith-Waterman) es muy preciso y garantiza obtener el alineamiento óptimo.

Pero ese algoritmo es muy lento. El tiempo de cálculo es proporcional al producto de las longitudes de las dos secuencias que se quieren alinear (o al producto de la longitud de nuestra secuencia problema y la de todas las secuencias de la base de datos).

Por el contrario, los algoritmos de alineamiento **local** son mucho más rápidos.

Alineamiento local

Se localizan todas las subsecuencias similares entre las dos secuencias:

```
Query: 181 acgatagcagatagcgcatagcgactagcgactgcagctacgcagcatagcagcagcaga 240
          ||| ||| |||
Sbjct: 189 tgagctagagatagctacgacgcatcagcgatagcagctagggcagctgcagcgactagca 247
```

El alineamiento se trata de extender en los dos sentidos mediante alineamiento global:

```
Query: 181 acgatagcagatagcgcatagcgactagcgactgcagctacgcagcatagcagcagcaga 240
          ||| ||| |||
          ← ( ) → ← ( ) → ← ( ) →
Sbjct: 189 tgagctagagatagctacgacgcatcagcgatagcagctagggcagctgcagcgactagca 247
```

Puntuación de un alineamiento

Range of Alignment

ATTGTCAAAGACTTGAGCTGATGCAT

GGCAGACATGA-CTGACAAGGGTATCG

Mismatch

Gap

$S = \sum(\text{identities, mismatches}) - \sum(\text{gap penalties})$

Score = **Max(S)**

Puntuación de un alineamiento (ejemplo)

```
AACGTTTCCAGTCCAAATAGCTAGGC
| | | * * | | |   | * | | | * | | * | | | | |
AACCGTTC---TACAATTACCTAGGC
```

| Emparejamientos (+1): 18

* Desemparejamientos (-2): 5

- Huecos (existencia-2, extension -1): 1 de longitud 3

$$\text{Puntuación} = [18 * 1] + [5 * (-2)] + [(-2) + 2 * (-1)] = 4$$

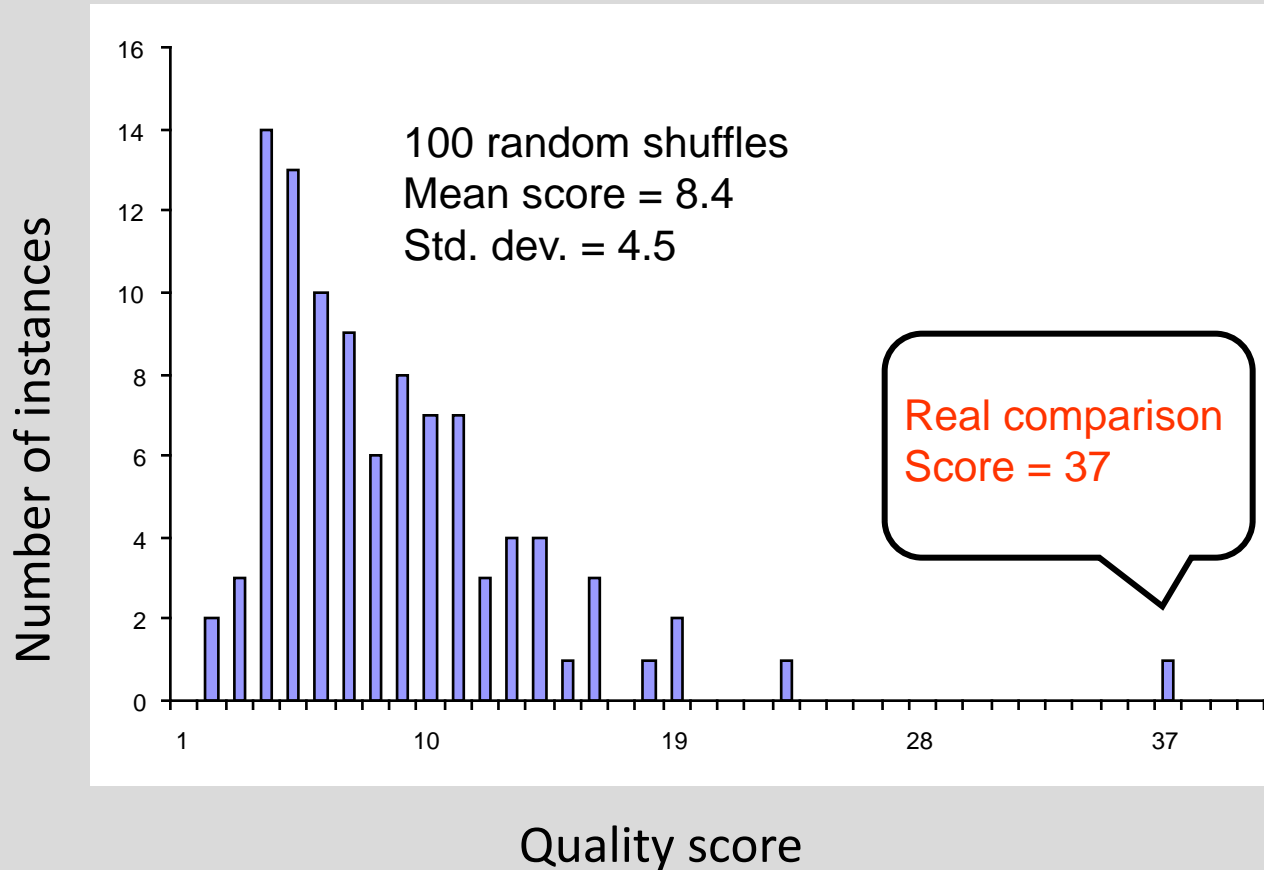
Significación estadística de un alineamiento: Test de randomización

- Se alinean las dos proteínas y se obtiene una puntuación real para el alineamiento obtenido:

```
RBP:                26  RVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFVDETGQMSATAKGRVRLLNWD- 84
                   + K++ + + +GTW++MA          + L   + A   V   T   +           +L+  W+
glycodelin:        23  QTKQDLELPKLAGTWHSMAA-TNNISLMATLKAPLRVHITSLLPTPEDNLEIVLHRWEN 81
```

- Se randomiza la segunda secuencia 100 veces, permutando al azar ('shuffling') las posiciones que ocupan los aminoácidos (manteniendo por tanto la longitud de la secuencia y la composición de aminoácidos)
- Se alinea cada secuencia randomizada con la primera secuencia y se obtienen 100, 1.000, 10.000... puntuaciones 'aleatorias'
- Cabe esperar que la puntuación real sea mucho mas grande que las puntuaciones 'aleatorias'

A randomization test shows that RBP is significantly related to b-lactoglobulin



Alineamiento local: FASTA

Fast Algorithm

Pearson & Lipman, 1988

The screenshot shows the EBI website interface for the FASTA/SSEARCH/FASTX/FASTY/FASTA tool. The top navigation bar includes links for Databases, Tools, Research, Training, Industry, About Us, and Help, along with Site Index, RSS, and Print icons. The left sidebar contains a navigation menu with categories like Help, FASTA website, Similar Applications, Programmatic Access, Download, Database Information (UniProt, UniParc), and FASTA related literature. The main content area is titled 'FASTA/SSEARCH/FASTX/FASTY/FASTA - Protein Similarity Search' and provides a description of the tool's capabilities. It features a 'Use this tool' section with two steps: 'STEP 1 - Select your databases' and 'STEP 2 - Enter your input sequence'. Step 1 includes a list of protein databases with checkboxes, a 'Clear Selection' button, and a sidebar for 'OTHER TYPES' (General and Specialised). Step 2 includes a dropdown menu for sequence type and a text input field.

Databases Tools Research Training Industry About Us Help Site Index

EBI > Tools > Sequence Similarity Searching > FASTA

FASTA/SSEARCH/FASTX/FASTY/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).

Use this tool

STEP 1 - Select your databases

1 Databank Selected X Clear Selection

PROTEIN DATABASES

- UniProt Knowledgebase
- UniProtKB/Swiss-Prot
- UniProtKB/Swiss-Prot isoforms
- UniProtKB/TrEMBL
- UniProtKB Taxonomic Subsets
- UniProt Clusters

OTHER TYPES

General

- Nucleotide Databases

Specialised

- Proteomes Databases
- Genomes Databases
- WGS Databases

STEP 2 - Enter your input sequence

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

FASTA Results

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

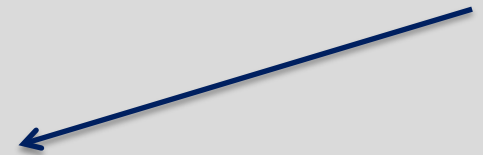
Alignments

Selection: [Show Annotations](#) [Hide Annotations](#) | [Show Alignments](#) [Hide Alignments](#)

[Download](#) in [fasta](#) format

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

Align.	DB:ID	Source	Length	Score	Identities	Positives	E()
<input checked="" type="checkbox"/> 1	SP:VIF_HV2BE	Virion infectivity factor OS=Human immunodeficiency virus type 2 subtype A (isolate BEN) GN=vif PE=2 SV=1 <i>Cross-references and related information in:</i> ▶ Nucleotide Sequences ▶ Ontologies ▶ Protein Families ▶ Literature	215	1177	100.0	100.0	1.5E-82
<input checked="" type="checkbox"/> 2	SP:VIF_HV2D1	Virion infectivity factor OS=Human immunodeficiency virus type 2 subtype A (isolate D194) GN=vif PE=2 SV=1 <i>Cross-references and related information in:</i> ▶ Nucleotide Sequences ▶ Ontologies ▶ Protein Families ▶ Literature	215	1120	93.9	98.2	4.1E-78
<input checked="" type="checkbox"/> 3	TR:P88143_9HIV2	Virion infectivity factor OS=Human immunodeficiency virus 2 GN=vif PE=4 SV=1 <i>Cross-references and related information in:</i> ▶ Nucleotide Sequences ▶ Ontologies ▶ Protein Families	215	1117	93.3	98.2	7.0E-78
<input checked="" type="checkbox"/> 4	TR:Q6R792_9HIV2	Virion infectivity factor OS=Human immunodeficiency virus 2 GN=vif PE=4 SV=1 <i>Cross-references and related information in:</i> ▶ Nucleotide Sequences ▶ Ontologies ▶ Protein Families ▶ Literature	215	1102	91.4	98.2	1.0E-76
<input checked="" type="checkbox"/> 5	TR:Q6R783_9HIV2	Virion infectivity factor OS=Human immunodeficiency virus 2 GN=vif PE=4 SV=1 <i>Cross-references and related information in:</i> ▶ Nucleotide Sequences ▶ Ontologies ▶ Protein Families ▶ Literature	215	1102	91.4	98.2	1.0E-76



Valor E: probabilidad de que la similitud encontrada se deba al azar

Valor P y valor E

Valor P: Probabilidad de que un suceso ocurra por azar.

En el contexto del alineamiento de secuencias, el valor P asociado a una determinada puntuación S de un alineamiento es la probabilidad de obtener por azar una puntuación al menos tan alta como S .

Valor E (expectation value): Corrección del valor P para ensayos múltiples.

En el contexto del alineamiento de secuencias, el valor E asociado a una puntuación S es la proporción de alineamientos obtenidos por azar en un rastreo de la base de datos con puntuaciones al menos tan buenas como S .

Cuanto más bajo el valor E, más significativa es la puntuación obtenida para un alineamiento.

Alineamiento local: BLAST

Basic Local Alignment Search Tool

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990)

BLAST® Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

NCBI/ BLAST/ **blastn suite** Standard Nucleotide BLAST

blastn blastp blastx tblastn tblastx

Enter Query Sequence BLASTN programs search nucleotide databases using a nucleotide query. [more...](#)

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#) Query subrange [?](#)

GGAAACTAA	GACTCCTTCG	CCTTCTGCAC	CAGACAAGTG	AGTATGGAGC	CTGGTAGGAA
TCAGCGITTT	GTIGTCATTT	TACTAACAAG	TGCTTGCCTA	GTATATTGTA	GCCAGTATGT
GACTGTTTTT	TATGGCATAAC	CCGCGTGGAA	AAATGCATCT	ATCCCTTAT	TTTGTCAC
TAAAAATAGA	GACACTTGGG	GGACCATAACA	GTGCTTGCCA	GACAATGATG	ATTATCAGGA
AATAATTTTA	AATGTGACAG	AGGCTTTTGA	TGCAATGGAAT	AATACAGTGA	CAGAACAAGC
AGTAGAAGAT	GTCIGGCATC	TATTTGAGAC	ATCAATAAAA	CCATGTGTCA	AGCTAACACC

Or, upload file [?](#)

Job Title

Enter a descriptive title for your BLAST search [?](#)

Align two or more sequences [?](#)

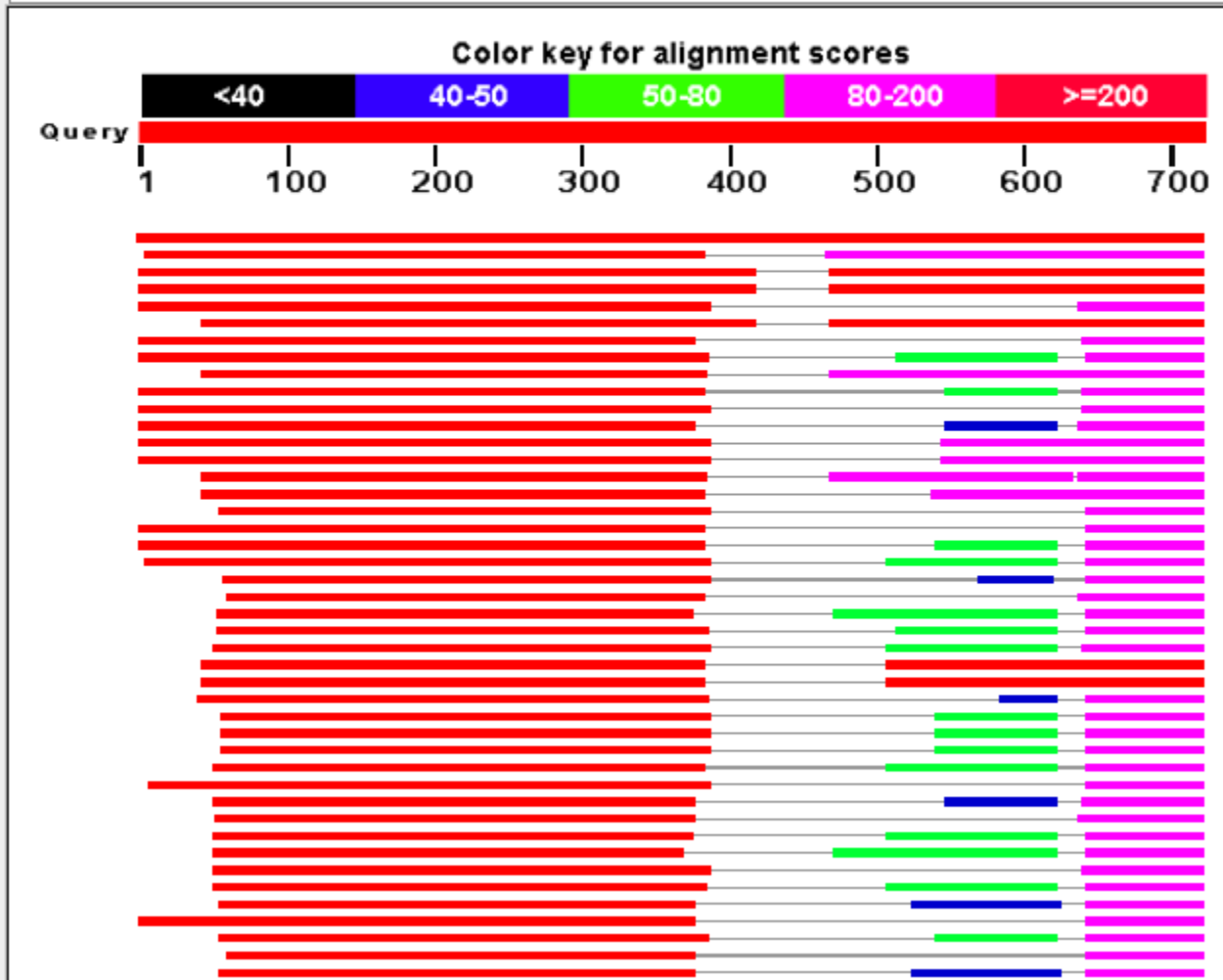
Choose Search Set

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

◆ Nucleotide collection (nr/nt) [?](#)

Distribution of 241 Blast Hits on the Query Sequence

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



Sequences producing significant alignments:


Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
gi 1332355 M30502.1	Human immunodeficiency virus 2 isolate BEN, complete genome	1385	1385	100%	0.0	100%	G
gi 1845204 U38293.1	Human immunodeficiency virus type 2, complete proviral genome	550	703	88%	5e-153	92%	
gi 60155 X52223.1	Human immunodeficiency virus type 2 strain D194 proviral genome	533	805	92%	8e-148	89%	
gi 325654 J04542.1	Human immunodeficiency virus type 2 (HIV-2), complete proviral genome	533	805	92%	8e-148	89%	
gi 747644 U22047.1	Human immunodeficiency virus type 2, complete genome	477	612	65%	5e-131	94%	
gi 77379272 DQ213026.1	HIV-2 isolate P1-1986 from Sweden envelope glycoprotein (env) gene,	475	680	87%	2e-130	89%	
gi 41056775 AY509259.1	HIV-2 isolate MCN13, complete genome	475	601	63%	2e-130	93%	
gi 4007991 AF082339.1	HIV-2 isolate ALI from Guinea-Bissau, complete genome	475	651	79%	2e-130	94%	
gi 77379274 DQ213027.1	HIV-2 isolate P1-1991 from Sweden envelope glycoprotein (env) gene,	471	658	82%	3e-129	91%	
gi 3153166 D00835.1	Human immunodeficiency virus 2 proviral DNA, complete genome	465	650	75%	1e-127	94%	
gi 1339798 M31113.1	Human immunodeficiency virus type 2 (HIV-2), complete proviral genome	465	585	65%	1e-127	92%	
gi 41056785 AY509260.1	HIV-2 isolate MCR35, complete genome	464	632	74%	6e-127	92%	



Valor E: probabilidad de que la similitud encontrada se deba al azar

Alignments

Select All [Get selected sequences](#) [Distance tree of results](#)

> [gi|1332355|gb|M30502.1|HIV2BEN](#)  Human immunodeficiency virus 2 isolate BEN, complete genome
Length=10359

Score = 1385 bits (720), Expect = 0.0
Identities = 720/720 (100%), Gaps = 0/720 (0%)
Strand=Plus/Plus

```
Query 1      GGAAAATAAGACTCCTTCGCCTTCTGCACCAGACAAGTGAGTATGGAGCCTGGTAGGAA 60
            |||
Sbjct 6661   GGAAAATAAGACTCCTTCGCCTTCTGCACCAGACAAGTGAGTATGGAGCCTGGTAGGAA 6720

Query 61     TCAGCTGTTTGTGTCATTTTACTAACAAGTGCTTGCTTAGTATATTGTAGCCAGTATGT 120
            |||
Sbjct 6721   TCAGCTGTTTGTGTCATTTTACTAACAAGTGCTTGCTTAGTATATTGTAGCCAGTATGT 6780

Query 121    GACTGTTTTCTATGGCATAACCCGCGTGGAAAAATGCATCTATTCCCTTATTTTGTGCAAC 180
            |||
Sbjct 6781   GACTGTTTTCTATGGCATAACCCGCGTGGAAAAATGCATCTATTCCCTTATTTTGTGCAAC 6840
```

```
Query 467    ATAATAAATGAAACTTCTAACTGCATAGAAAACAACACATGCGCAGGATTAGGGTATGAG 526
            |||
Sbjct 6532    ATAATAGATGAAAATTCTACCTGTATAGGCGACAACAACACTGCACAGGATTAGGGAAAGAA 6591

Query 527    GAGATGATGCAATGTGAGTTCAATATGAAGGGGTTAGAACAAGATAAGAAAAGGAGGTAT 586
            |||
Sbjct 6592    GAGGTGGTTGAGTGTGAGTTCAATATGACGGGGCTAGAACAAGATAAGAAAAGGAAAGTAT 6651

Query 587    AAGGACACATGGTATTTAGAAGATGTGGTTTGTGACAACACAA---CAGCTGGCACATGT 643
            |||
Sbjct 6652    AATGACGCATGGTACTCAAGAGATGTGGTTTGTGACAAGACAAACGGAACAGGCACATGT 6711

Query 644    TACATGAGACATTGCAACACATCAATCATCAAAGAGTCATGTGATAAGCACTATTGGGAT 703
            |||
Sbjct 6712    TACATGAGACATTGCAACACATCAGTCATCAAAGAGTCATGTGACAAGCACTATTGGGAT 6771

Query 704    GCTATGAGGTTTAGATA 720
            |||
Sbjct 6772    GCTATGAAGTTTAGATA 6788
```


Table 5.1 Five ways to perform BLAST searches.

Program	Description
blastp	Compares an amino acid query sequence against a protein sequence database.
blastn	Compares a nucleotide query sequence against a nucleotide sequence database.
blastx	Compares a nucleotide query sequence translated in all reading frames against a protein sequence database. Use this option to find potential translation products of an unknown nucleotide sequence.
tblastn	Compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames.
tblastx	Compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database. Note that the tblastx program cannot be used with the nr database on the BLAST Web page because it is computationally intensive.

Source: National Institutes of Health

Table 5.1 Phylogenomics: A Primer (© Garland Science 2013)

Program Name	Description	Abbreviation
FASTA	Scan a protein or DNA sequence library for similar sequences.	fasta
FASTX	Compare a DNA sequence to a protein sequence database, comparing the translated DNA sequence in forward and reverse frames.	fastx
FASTY	Compare a DNA sequence to a protein sequence database, comparing the translated DNA sequence in forward and reverse frames.	fasty
SSEARCH	Compare a protein or DNA sequence to a sequence database using the Smith-Waterman algorithm.	ssearch
GGSEARCH	Compare a protein or DNA sequence to a sequence database using a global alignment (Needleman-Wunsch)	ggsearch
GLSEARCH	Compare a protein or DNA sequence to a sequence database with alignments that are global in the query and local in the database sequence (global-local).	glsearch