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:

PTWRVPGRMEKWHALVKYLKYRTKDLEEVR…

¿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).

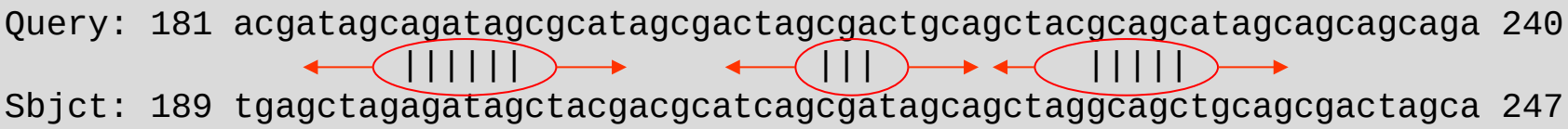
Por el contrario, los algoritmos de alineamiento **local** son mucho más rápidos.

Alineamiento local

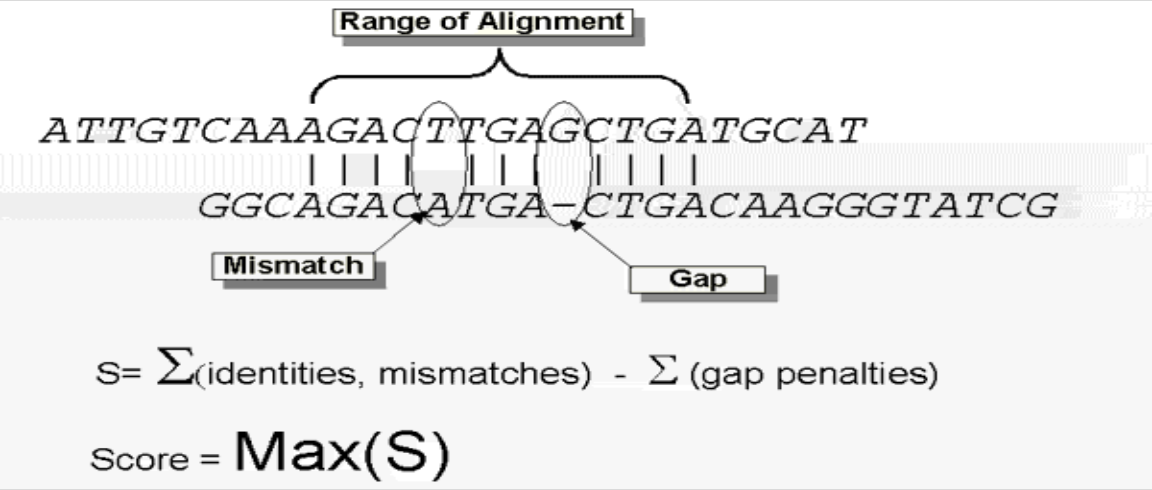
Se localizan todas las subsecuencias similares entre las dos secuencias:

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

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



Puntuación de un alineamiento



Puntuación de un alineamiento (ejemplo)

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

| Emparejamientos (+1): 18

* Desemparejamientos (-2): 5

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

Puntuación = [18 * 1] + [5 * (-2)] + [(– 2) + 2*(-1)] = 4

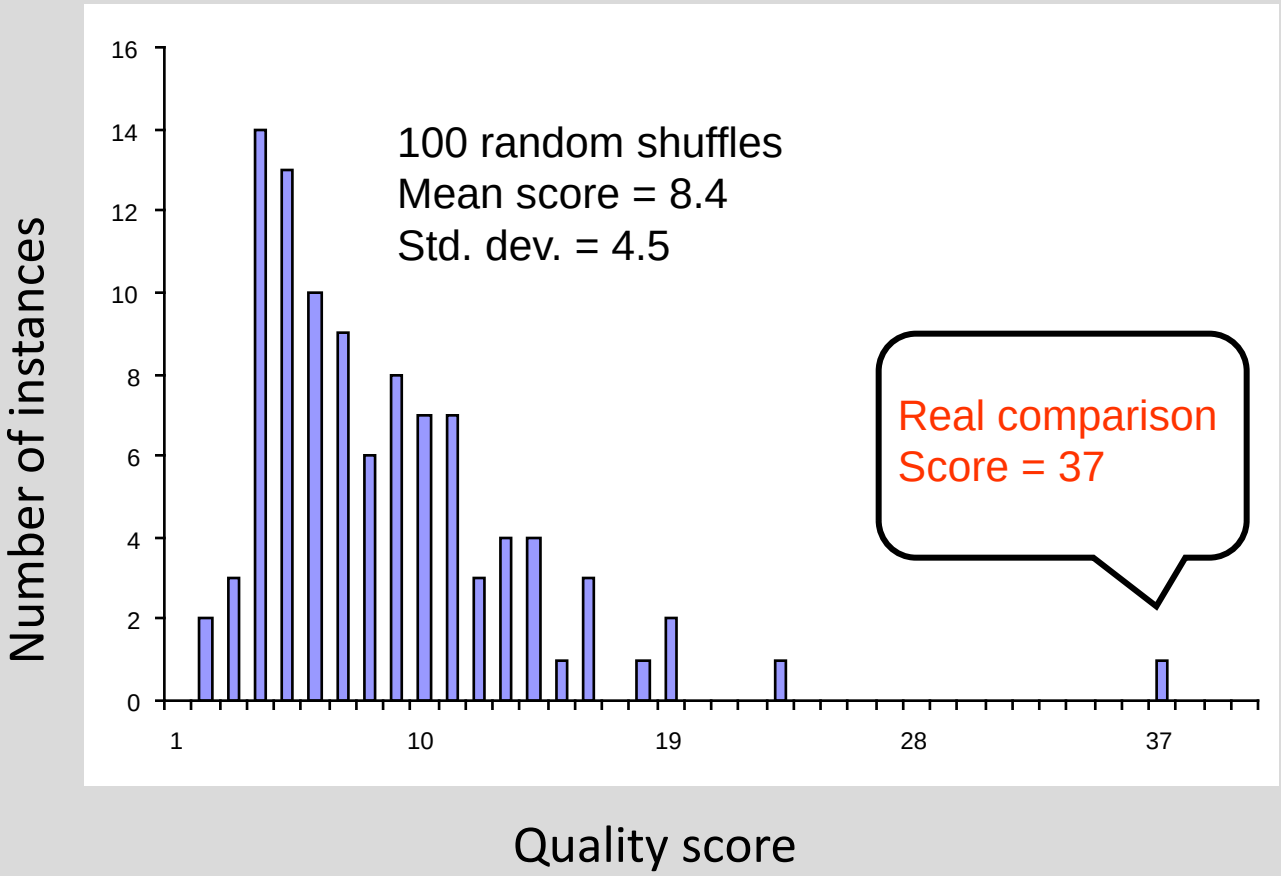
Significación estadística de un alineamiento: Test de randomización

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

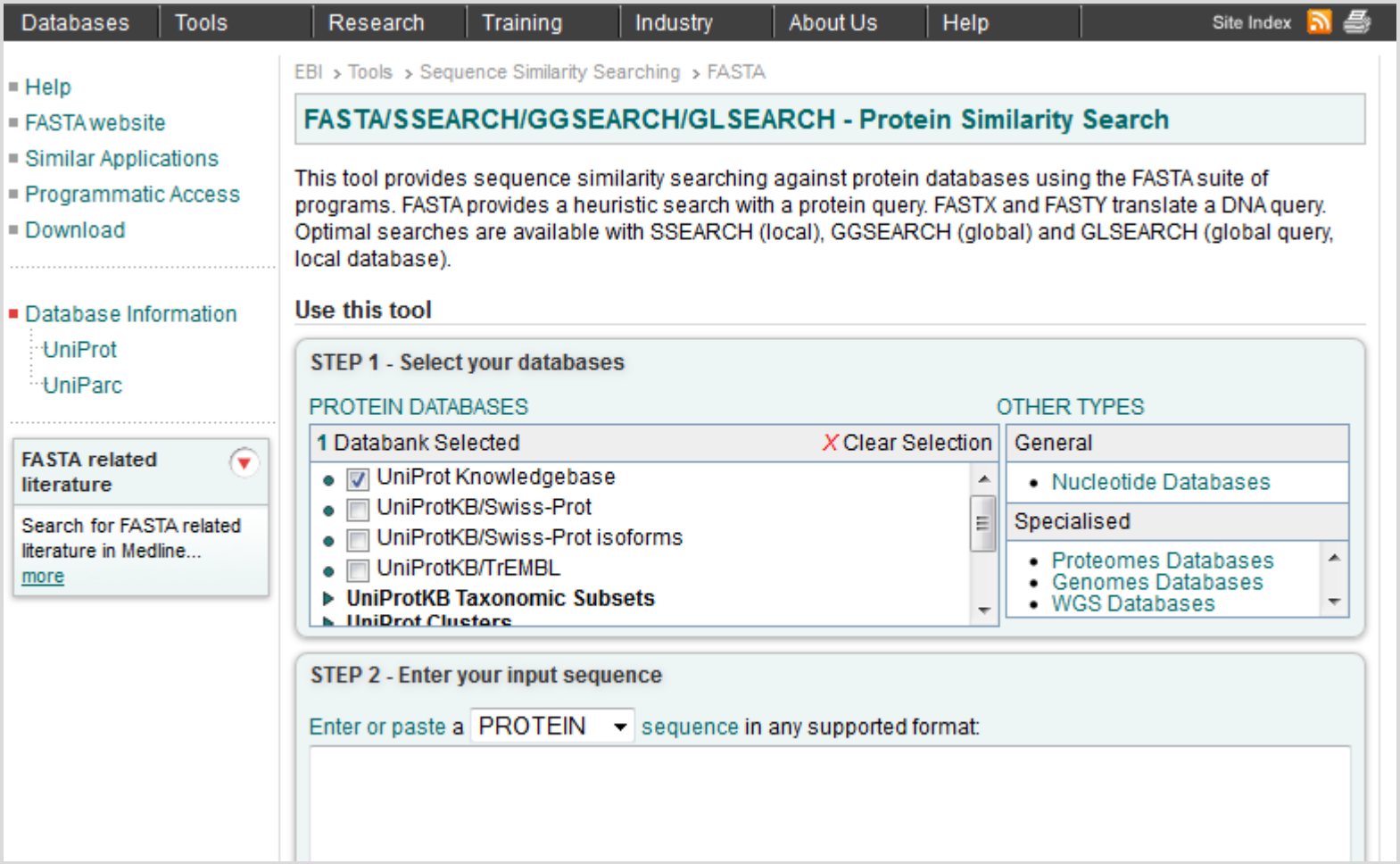
```
RBP:          26  RVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWD- 84
                  + K++ +  + +GTW++MA      + L   + A   V  T  +        +L+  W+
glycodelin:   23  QTKQDLELPKLAGTWHSMAMA-TNNISLMATLKAPLRVHITSLLPTPEDNLEIVLHRWEN 81
```

- 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



FASTA Results

Summary Table | Tool Output | Visual Output | Functional Predictions | Submission Details | Submit Another Job

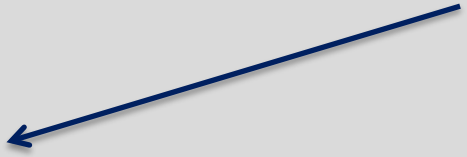
Alignments

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()
1	SP:VIF_HV2BE	Virion infectivity factor OS=Human immunodeficiency virus type 2 subtype A (isolate BEN) GN=vif PE=2 SV=1 *Cross-references and related information in:* ▶ Nucleotide Sequences ▶ Ontologies ▶ Protein Families ▶ Literature	215	1177	100.0	100.0	1.5E-82
2	SP:VIF_HV2D1	Virion infectivity factor OS=Human immunodeficiency virus type 2 subtype A (isolate D194) GN=vif PE=2 SV=1 *Cross-references and related information in:* ▶ Nucleotide Sequences ▶ Ontologies ▶ Protein Families ▶ Literature	215	1120	93.9	98.2	4.1E-78
3	TR:P88143_9HIV2	Virion infectivity factor OS=Human immunodeficiency virus 2 GN=vif PE=4 SV=1 *Cross-references and related information in:* ▶ Nucleotide Sequences ▶ Ontologies ▶ Protein Families	215	1117	93.3	98.2	7.0E-78
4	TR:Q6R792_9HIV2	Virion infectivity factor OS=Human immunodeficiency virus 2 GN=vif PE=4 SV=1 *Cross-references and related information in:* ▶ Nucleotide Sequences ▶ Ontologies ▶ Protein Families ▶ Literature	215	1102	91.4	98.2	1.0E-76
5	TR:Q6R783_9HIV2	Virion infectivity factor OS=Human immunodeficiency virus 2 GN=vif PE=4 SV=1 *Cross-references and related information in:* ▶ 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.

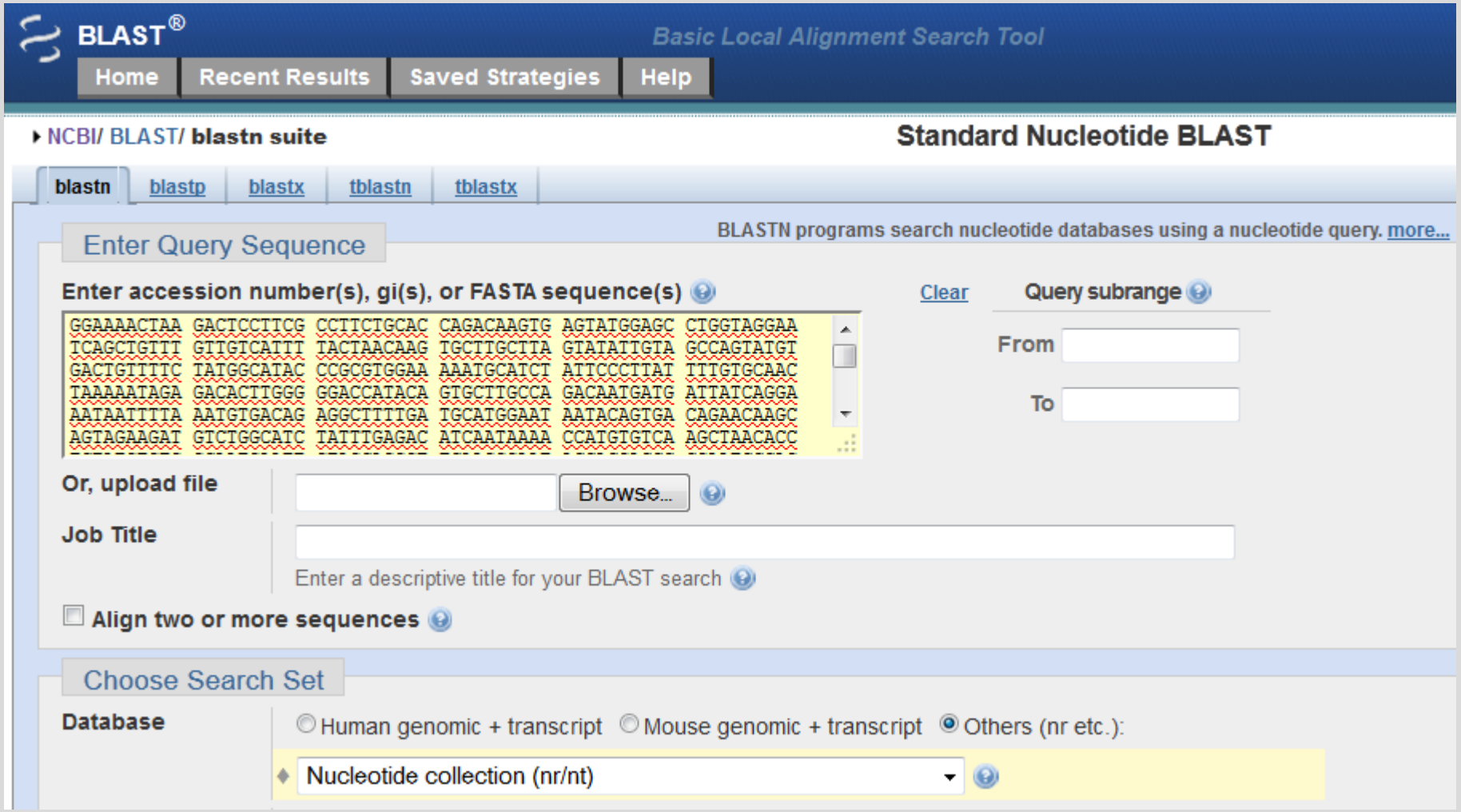
Valor E (expectation value): Corrección del valor P para ensayos múltiples.

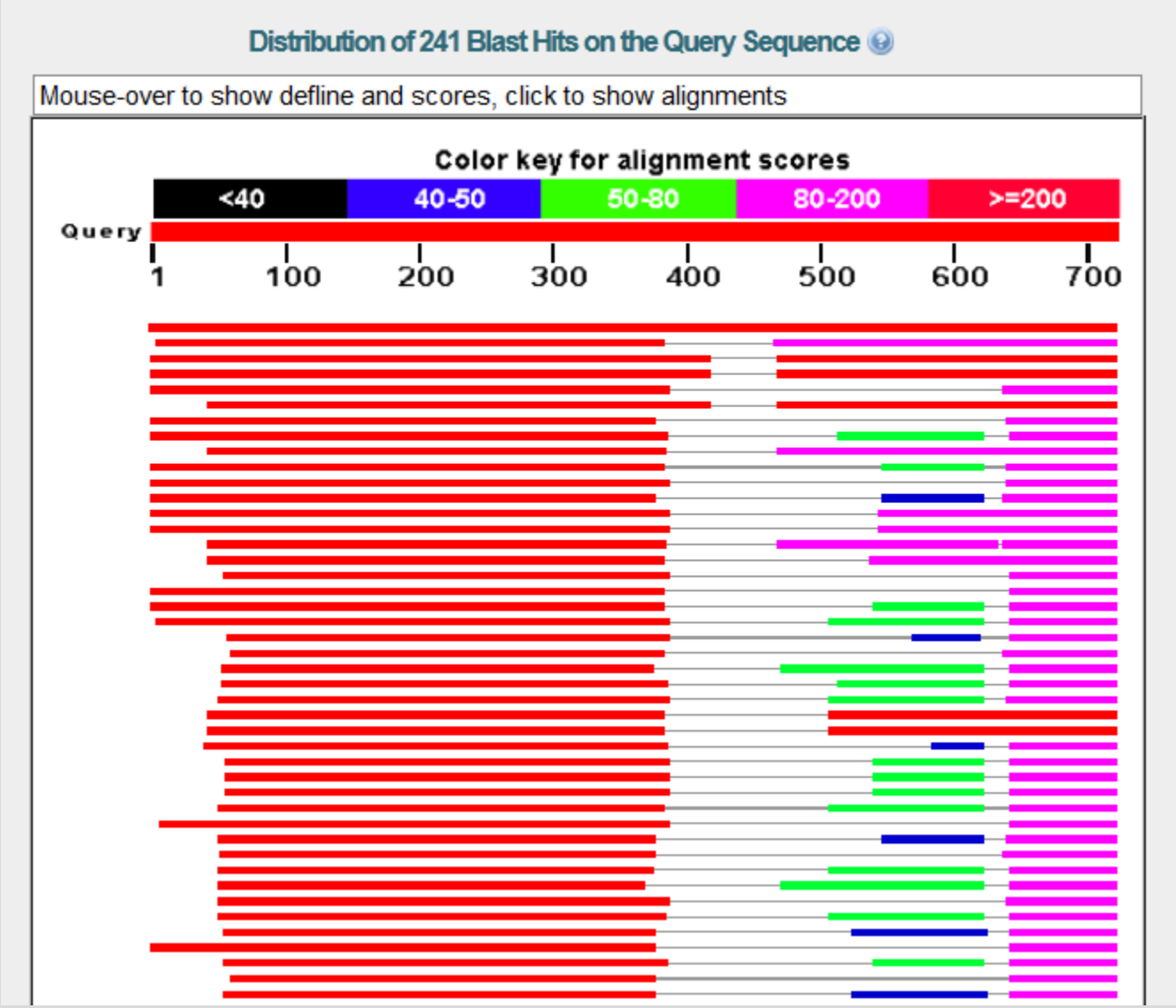
En el contexto del alineamiento de secuencias, el valor E asociado a una puntuación S es la proporción de alineamientos obtenidos por azar en un rastreo de la base de datos con puntuaciones al menos tan buenas como S.

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)




Distribution of 241 Blast Hits on the Query Sequence
Mouse-over to show defline and scores, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
1
100
200
300
400
500
600
700

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%	
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 genon	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 genon	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     GGAAAACTAAGACTCCTTCGCCTTCTGCACCAGACAAGTGAGTATGGAGCCTGGTAGGAA  60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  6661  GGAAAACTAAGACTCCTTCGCCTTCTGCACCAGACAAGTGAGTATGGAGCCTGGTAGGAA  6720

Query  61    TCAGCTGTTTGTTGTCATTTTACTAACAAGTGCTTGCTTAGTATATTGTAGCCAGTATGT  120
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  6721  TCAGCTGTTTGTTGTCATTTTACTAACAAGTGCTTGCTTAGTATATTGTAGCCAGTATGT  6780

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

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

Query  527   GAGATGATGCAATGTGAGTTCAATATGAAGGGGTTAGAACAAGATAAGAAAAGGAGGTAT  586
             |||  ||  |   | ||||||||||||||||  ||||  ||||||||||||||||||||  ||||
Sbjct  6592  GAGGTGGTTGAGTGTGAGTTCAATATGACGGGGCTAGAACAAGATAAGAAAAGGAAGTAT  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

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