Introducción a la Bioinformática
Búsquedas de secuencias en bases de datos
Heurísticas, Hits, significancia de los resultados

Fernán Agüero
fernan@iib.unsam.edu.ar
Instituto de Investigaciones Biotecnológicas
Universidad Nacional de San Martín

Búsqueda de secuencias por similitud

- Tenemos un **método (algoritmo)** que nos garantiza un alineamiento óptimo entre dos secuencias
- Tenemos un **sistema de scoring** complejo que refleja mejor nuestras ideas biológicas acerca de lo que es un alineamiento
- Tenemos una **secuencia de interés** (*query*)
- Y una **base de datos** con secuencias
- Cómo implementaríamos una búsqueda **por similitud** contra una base de datos de secuencias?

Método de fuerza bruta = Búsqueda exhaustiva

- **Buscar secuencias similares al query en la base de datos**

- Tomar **una por una** las secuencias de la base de datos
- Calcular un **alineamiento y su score**
- **Ordenar los alineamientos** en base al score
- Finalmente usar **nuestro criterio** y evaluar si las secuencias al tope del ranking son lo suficientemente similares

https://en.wikipedia.org/wiki/Brute-force_search

Espacio de búsqueda

- Cuántos **espacios de búsqueda** reconocibles hay?
- Las búsquedas de secuencias por similitud son "exactas" si recorren completamente el espacio de búsqueda

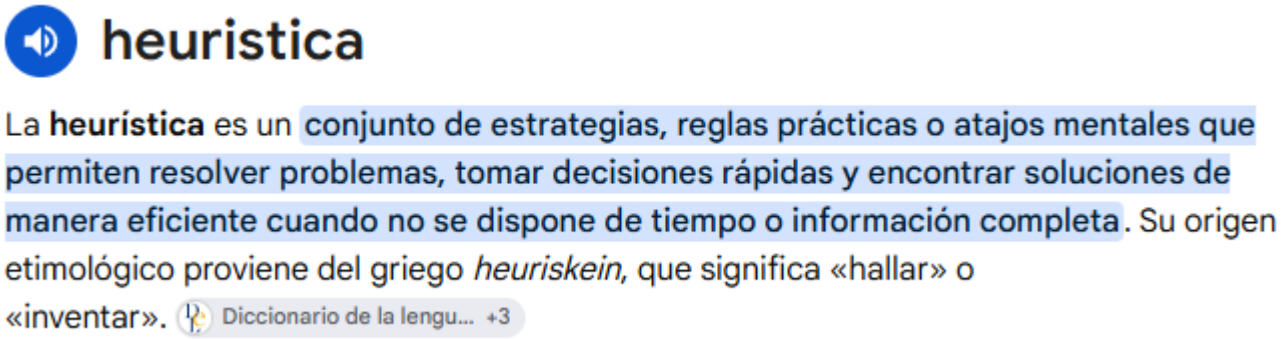
Espacios de búsqueda

- Hay dos espacios de búsqueda reconocibles:
  - El espacio de todas las secuencias de la base de datos
    - Database size (number of sequences)
  - El espacio de todos los alineamientos posibles entre dos secuencias
    - Dynamic programming matrix (M x N)
    - M = longitud de la secuencia query
    - N = longitud de la secuencia n de la base de datos

- Las búsquedas de secuencias por similitud son “exactas” si recorren en forma exhaustiva ambos espacios
  - Ej: Smith-Waterman sobre toda la base de datos

Heurísticas

- **A continuación vamos a introducir heurísticas para reducer estos espacios de búsqueda**
  - **Estrategias heurísticas para filtrar la base de datos**
  - **Heurísticas para reducir el espacio de alineamientos posibles que se explora efectivamente**



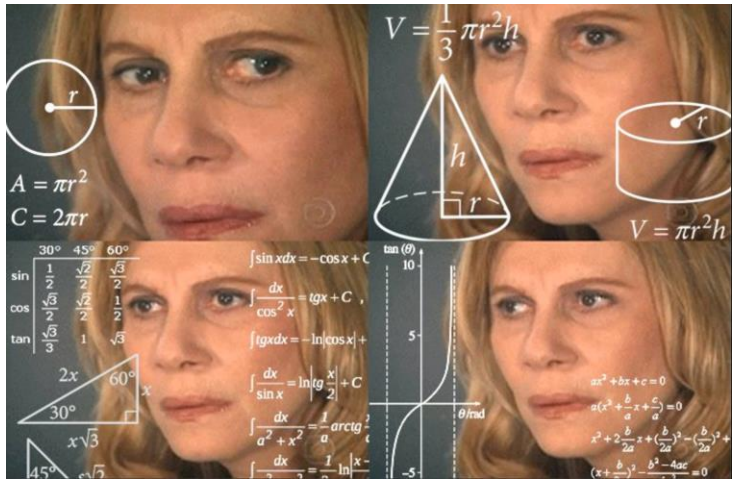
https://es.wikipedia.org/wiki/Heurística

Heurísticas para reducir el espacio de secuencias

MLIIKRDELVISWASHERE query sequence (ejemplo)

Vamos a intentar reducir el espacio de las secuencias de la base de datos
Database size (number of sequences)

Ideas? Heuriskein?

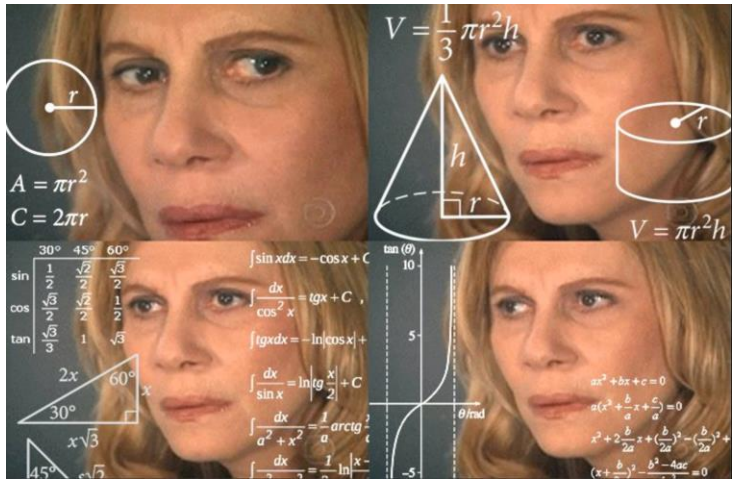


Heurísticas para reducir el espacio de secuencias

MLIIKRDELVISWASHERE query sequence (ejemplo)

Vamos a intentar reducir el espacio de las secuencias de la base de datos
Database size (number of sequences)

Ideas? Heuriskein?



Ideas:

- Descartar secuencias (saltearlas)
- Cuáles?
  - Las que no empiezan con Met?
  - Las que tienen una composición de aminoácidos diferente?
  - Las que son mucho mas largas?
  - Mucho mas cortas?

Indices - Hashing

Busquemos inspiración en los libros.

Los índices ayudan a encontrar cosas.

Que es un indice?

- Un mapa que conecta una palabra con una página
- Lookup tables, Hashes
  - Conectan datos

Hashes ~ Indices: ejemplo con anotaciones de genes

- Ejemplo de índices en bases de datos: buscar todos los registros que contengan la palabra **'kinase'** en la descripción de la secuencia

gi	acc	def
2147314077	X0	Xenopus laevis rhodopsin
1234567	AF435467	Mus musculus casein kinase

Ejemplo mínimo.

Tamaño real de la base de datos:
Cientos de millones de registros (10^8)

- Indexar la columna 'def' genera una nueva tabla

word	list of GIs
casein	1234,3245,43678,123456 ...
kinase	432,5678,32456,123456 ...
laevis	36314,214734, ...
mus	23467,98732,123456,312456,567
musculus	23467,98732,123456,567983 ...
rhodopsin	214734,223466,873212,23587,29
xenopus	8462,36314,98476,214734 ...

Tamaño del índice?
Cuántos registros tiene?

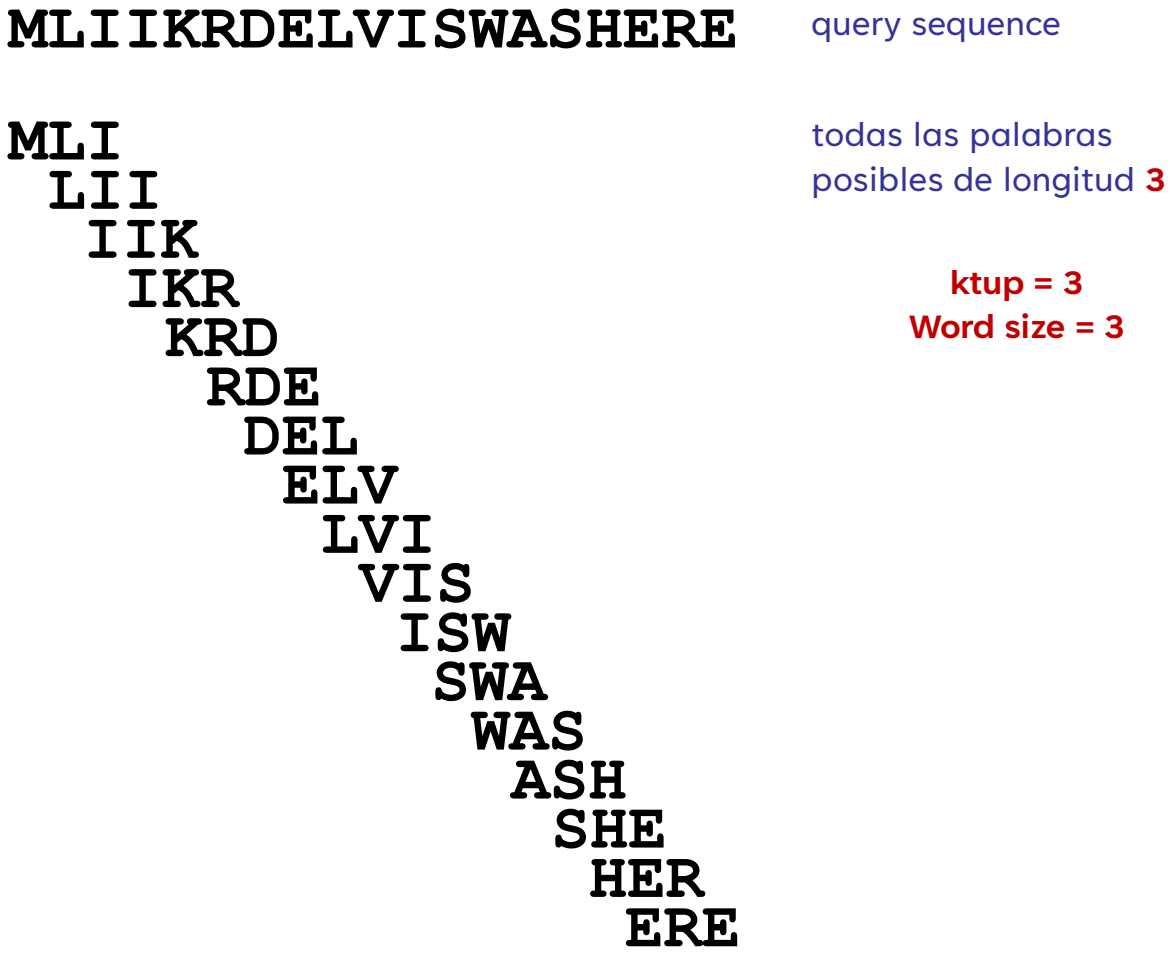
Hashing methods

Hashing es un método común para acelerar búsquedas en bases de datos.

Compilar un “diccionario” o ‘lookup table’ de palabras a partir de la secuencia ‘query’. Arma un índice con todas las posibles sub-palabras contenidas en la secuencia.

Longitud de la secuencia: 19

Cantidad de palabras: 17



Espacios de búsqueda

Secuencia Query

Seq: "MLIIKRDELVISWASHERE"

Longitud: 19

Ktup: 1; Cantidad de palabras: 19

Ktup: 2; Cantidad de palabras : 18

Ktup: 3; Cantidad de palabras : 17

Secuencias en la Base de Datos

Seq: "MLIIKRDELVISWASHERE"

Longitud: 19

Nro de Secuencias posibles: $20^{19} \sim 5.24\ 10^{24}$

Ktup 1; Cantidad de palabras?

Ktup: 2; Cantidad de palabras?

Ktup: 3; Cantidad de palabras?

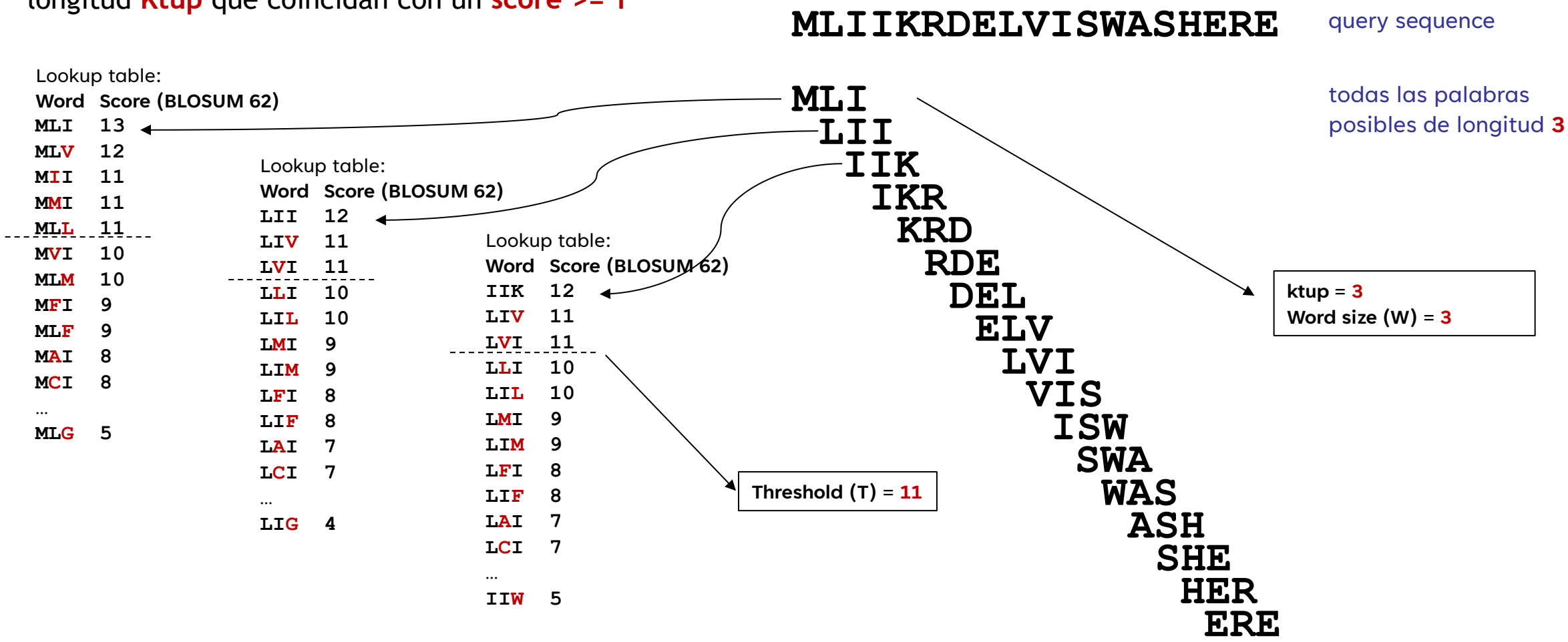
Ejemplo de lookup table de secuencias

Word	{seq,pos}
AAA	{L07770,232}; …
AAC	
AAD	
AAE	
AAF	
AAG	
AAH	
AAI	
AAK	
AAL	{L0771,221}; …
AAM	
AAN	
AAP	{L0770; 168}; …
AAQ	{BC002171,77}; …
AAR	
AAS	{BC002171,81}; …
AAT	{L0770; 336}
AAV	
AAW	
AAY	{L0770,41}; …
…	

Acc	Def	Prot Seq
L0770	Xenopus laevis rhodopsin mRNA, complete cds	MNGTEGPNFYVPMSNKTGVVRSPFDYPQY YLAEPWQYSALAAYMFLLILLGLPINFMTLF VTIQHKKLRTPLNYILLNLVFANHFMVLCG FTVTMYTSMHGYFIFGQTGCYIEGFFATLG GEVALWSLVVLAVERYMVVCKPMANFRFG ENHAIMGVAFTWIMALSCAAPPLFGWSRYI PEGMQCSCGVDYYTLKPEVNNESFVIYMFI VHFTIPLIVIFFCYGRLLCTVKEAAAQQQES ATTQKAEKEVTRMVVIMVVFFLICWVPYAY VAFYIFTHQGSNFGPVFMTVPAFFAKSSAIY NPVIYIVLNKQFRNCLITTLCCGKNPFGDED GSSAATSKTEASSVSSSQVSPA
L0771	Xenopus laevis (clone pXga1) transducin alpha subunit mRNA, complete cds	MGAGASAEEKHSRELEKKLKEDADKDART VKLLLLGAGESGKSTIVKQMKIIHQDGYSVE ECLEFIAIIYGNTLQSMIAIVKAMNTLNIQFG DPARQDDARKLMHLADTIDEGSMPKEMSDI IGRLWKDTGIQACFDRASEYQLNDSAGYYL NDLDRLVIPGYVPTEQDVLRSRVKTTGIIET QFGLKDLNFRMFDVGGQRSERKKWIHCFE GVTCIIFIAALSAYDMVLVEDDEVNRMHESL HLFNSICNHRYFATTSIVLFLNKKDVFTEKIK KAHLSICFPDYDGPNTYEDAGNYIKTQFLE LNMRRDVKEIYSHMTCATDTENVKFVFDAV TDIIIKENLKDCGLF
BC002171	Mus musculus casein kinase 1, alpha 1, mRNA (cDNA clone IMAGE:3486713), partial cds.	KKQKYEKISEKKMSTPVEVLCKGFPAEFAM YLNYCRGLRFEEAPDYMYLRQLFRILFRTL NHQYDYTFDWTMLKQKAAQQAASSSGQG QQAQTPTGKQTDKTKSNMKGF
…		

Busquedas usando lookup tables

Lookup table: todas las posibles sub-palabras de longitud **Ktup** que coincidan con un **score >= T**



Espacios de búsqueda

Secuencia Query

Seq: "**MLIIKRDELVISWASHERE**"

Longitud: **19**

Ktup: **3**; Cantidad de palabras : **17** ⟶

Secuencias en la Base de Datos

RefSeq (Julio 2026) = 482,864,455 (proteínas)

Parámetros:
Ktup = **3**
Threshold = **11**

Ktup: **3**; Cantidad de palabras = max 8000

Ktup: **3**; Secuencias con palabras con score >= **11**?

Cuántas secuencias con **17** palabras con score >= 11?
Cuántas secuencias con **16** palabras con score >= 11?
Cuántas secuencias con **15** palabras con score >= 11?
Cuántas secuencias con **14** palabras con score >= 11?
Cuántas secuencias con **13** palabras con score >= 11?
Cuántas secuencias con **12** palabras con score >= 11?
...

Recorte de espacios de búsqueda

Cuál es la frecuencia de coincidencias de k-meros en pares de secuencias biológicas conservadas?

Table 3. *Library search sensitivity*[a]

PROTEINS

Query	Superfamily	Family size	Related—missed/unrelated—found BLAST	*ktup* = 2	*ktup* = 1 opt	Smith-Waterman
HAHU	Hemoglobin	505	35	43 –	17 +	16 +
K1HUAG	Ig κ V region	280	54	100 –	17 +	21 +
OOHU	G-protein CR	165	27	42 –	26 +	22 +
CCHU	Cytochrome *c*	142	25	29 –	27 –	27 –
N2KF1U	Snake neurotoxin	109	3	5 –	4 –	4 –
XURT8C	Glutathione transferase	106	8	20 –	8	8
TPHUCS	Calcium binding	106	8	17 –	7 +	11 –
OKHU2C	Protein kinase	97	54	52 +	37 +	37 +
FEPE	Ferredoxin	93	53	54 –	53	53
RKMDS	Ribulose-bisphosphate carboxylase	77	0	1 –	0	0
K3HU	Ig κ C region	74	22	26 –	18 +	16 +
HMIVV	Hemagglutinin	73	1	4 –	0 +	0 +
HLHUB2	Histocompatibility Ag	71	2	15 –	1 +	0 +
IPHU	Insulin	69	3	3	3	3
CYBOA	α-Crystallin	67	7	8 –	4 +	4 +
PSHU	Phospholipase A2	58	2	2	2	2
DEHUGL	G3-PDH	46	0	1 –	1 –	1 –
TVHURA	Transforming protein (N-ras)	45	1	1	1	1

Pearson WR. Comparison of methods for searching protein sequence databases. Protein Sci. 1995 Jun;4(6):1145-60. doi: 10.1002/pro.5560040613. PMID: 7549879; PMCID: PMC2143149.

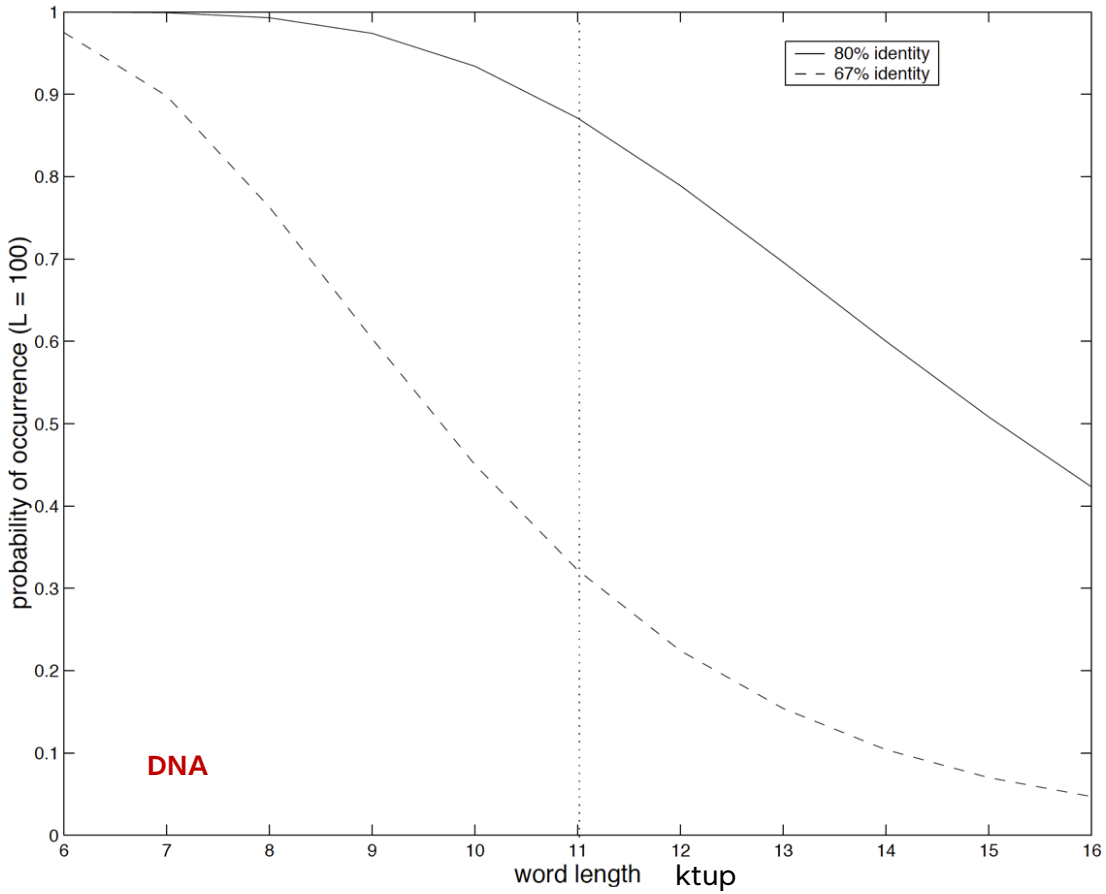
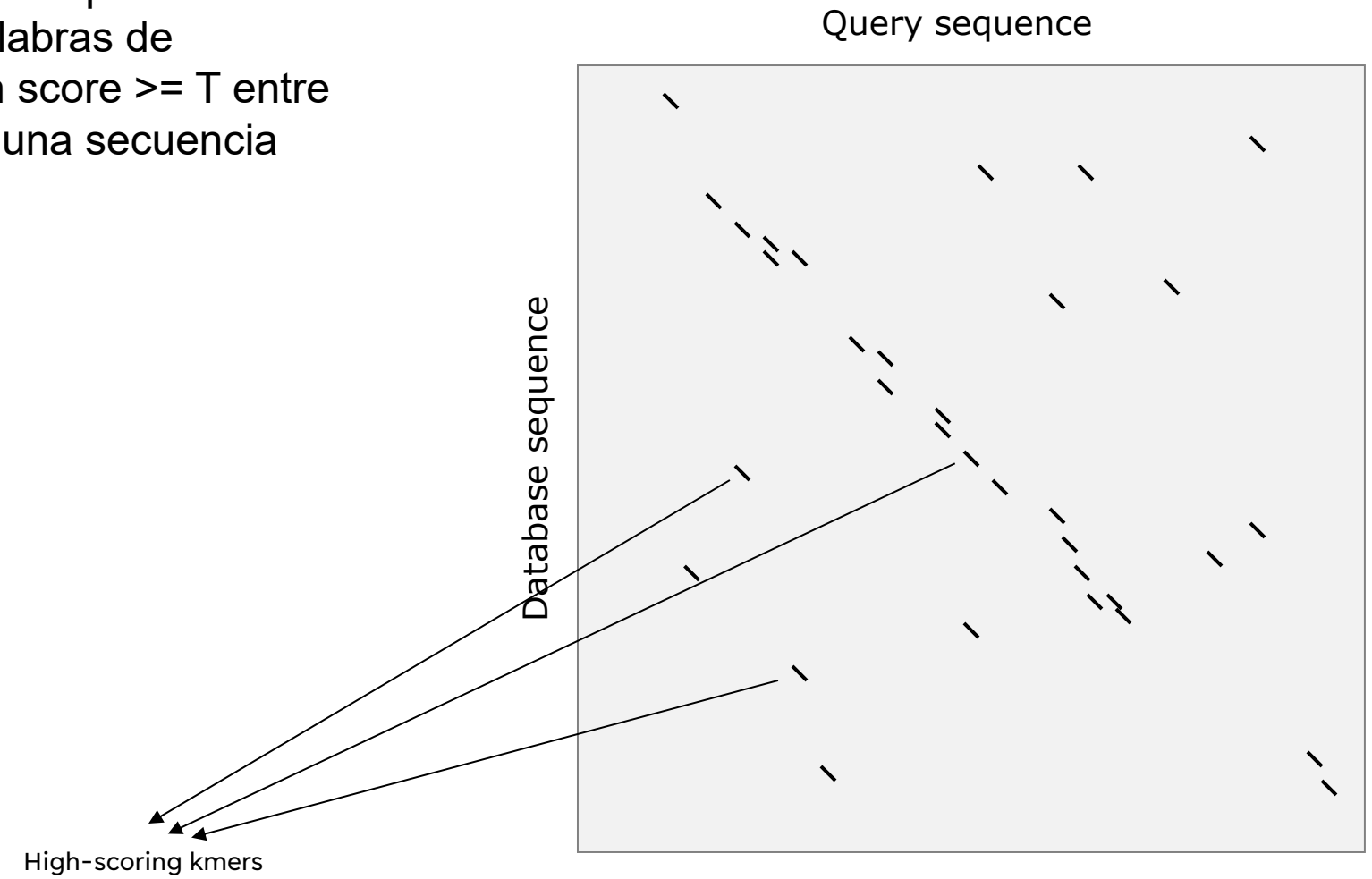


Gráfico de longitud de kmeros (eje x) vs probabilidad de ocurrencia para secuencias de ADN de longitud = 100 (eje y) y alineamientos de 80% o 67% de identidad.

https://gepstaging.wustl.edu/lessons/Smith_Waterman_to_BLAST/Smith_Waterman_to_BLAST.html

High-Scoring Words Map

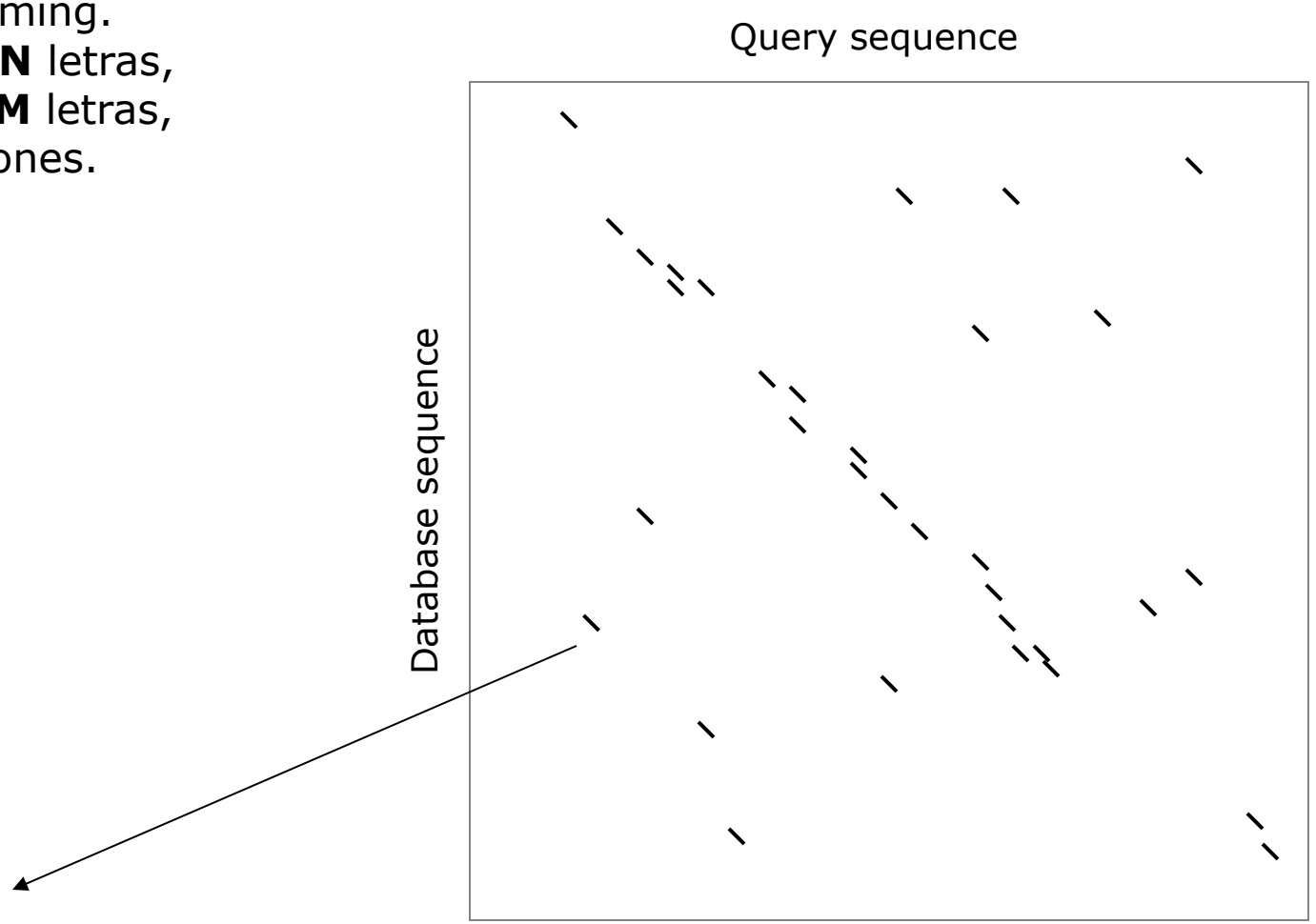
El mapa de palabras compartidas muestra todas las palabras de longitud = ktup, y con score >= T entre la secuencia query y una secuencia de la base de datos.



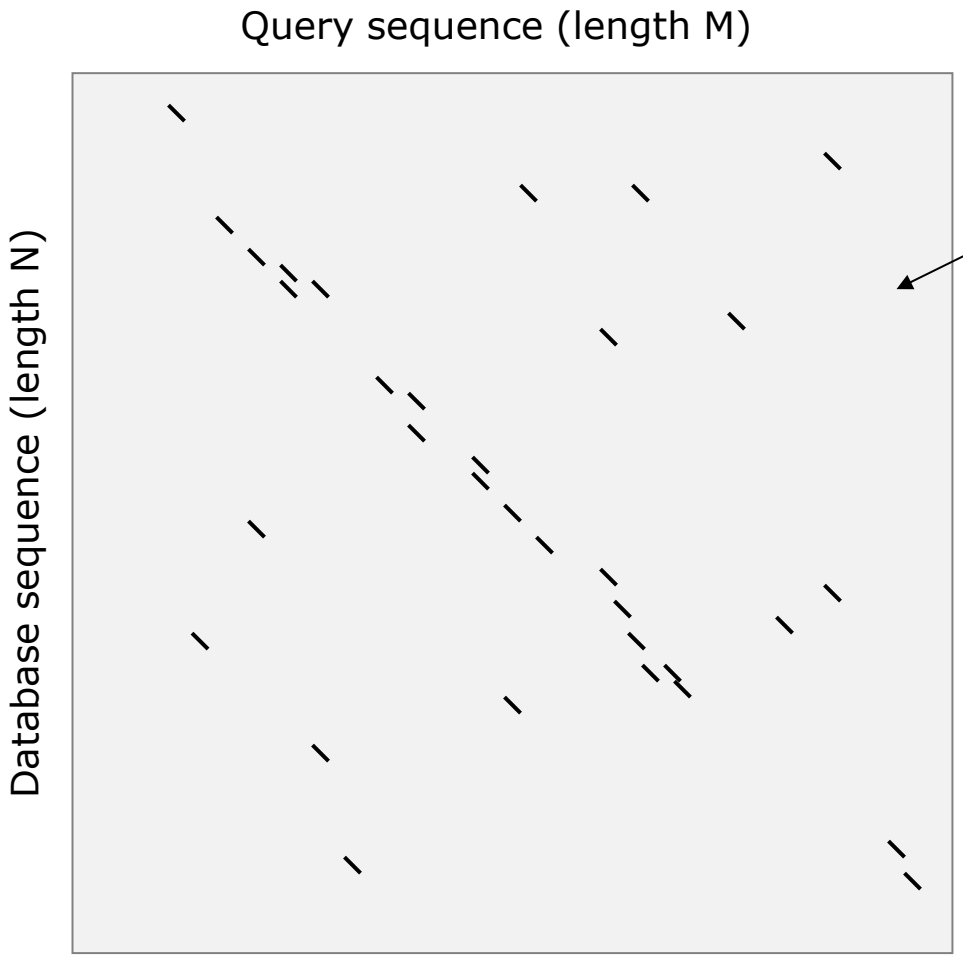
High-Scoring Words Map

La búsqueda más simple es un gran ejemplo de dynamic programming. Para una secuencia query de **N** letras, contra una base de datos de **M** letras, se requieren **MxN** comparaciones.

Cómo reducir este espacio de búsqueda?



Heurísticas para reducir espacio de búsqueda

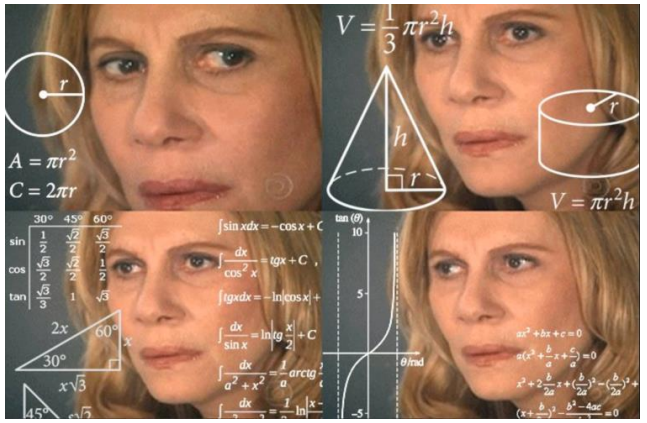


Ahora vamos a intentar reducir

El espacio de todos los alineamientos posibles entre dos secuencias

- Dynamic programming matrix (M x N)
  - M = longitud de la secuencia query
  - N = longitud de la secuencia de la base de datos

Cómo? Ideas?



Heurísticas: FASTA

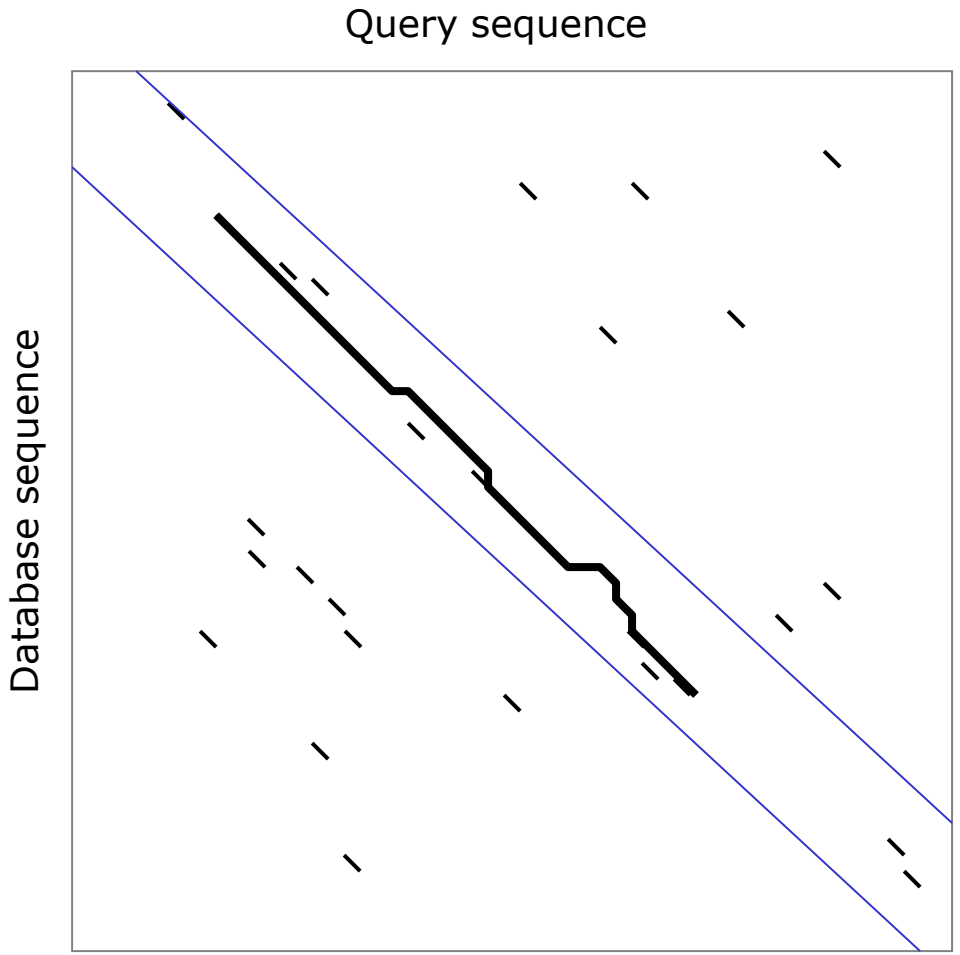
Las palabras de alto score de longitud ktup son las semillas del alineamiento.

1) FASTA identifica regiones de alta densidad de semillas

2) Elige la region de mayor score

3) Traza una banda diagonal que representa los límites para calcular el alineamiento

4) Dentro de los límites se calcula el mejor alineamiento usando el algoritmo Smith-Waterman



FASTA searches in a band

Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A. 1988 Apr;85(8):2444-8. doi: 10.1073/pnas.85.8.2444. PMID: 3162770; PMCID: PMC280013.

Heurísticas: FASTA

Las palabras de alto score de longitud ktup son las **semillas** del alineamiento.

1) FASTA identifica regiones de alta densidad de semillas

2) Elige la region de mayor score

3) Traza una **banda diagonal** que representa los límites para calcular el alineamiento

4) Dentro de los límites se calcula el mejor alineamiento usando el algoritmo Smith-Waterman

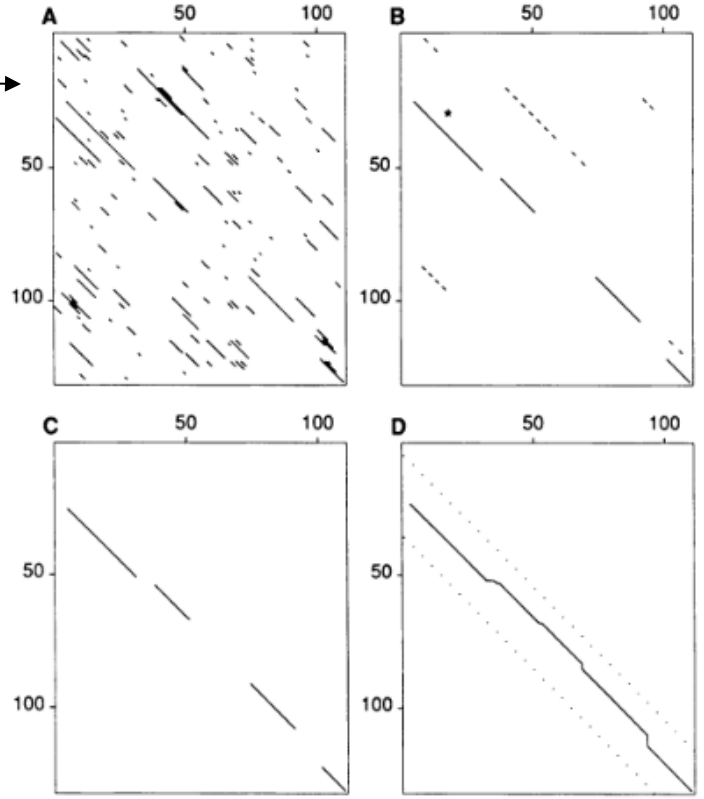


FIG. 1. Identification of sequence similarities by FASTA. The four steps used by the FASTA program to calculate the initial and optimal similarity scores between two sequences are shown. (*A*) Identify regions of identity. (*B*) Scan the regions using a scoring matrix and save the best initial regions. Initial regions with scores less than the joining threshold (27) are dashed. The asterisk denotes the highest scoring region reported by FASTP. (*C*) Optimally join initial regions with scores greater than a threshold. The solid lines denote regions that are joined to make up the optimized initial score. (*D*) Recalculate an optimized alignment centered around the highest scoring initial region. The dotted lines denote the bounds of the optimized alignment. The result of this alignment is reported as the optimized score.

Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A. 1988 Apr;85(8):2444-8. doi: 10.1073/pnas.85.8.2444. PMID: 3162770; PMCID: PMC280013.

Heurísticas: BLAST

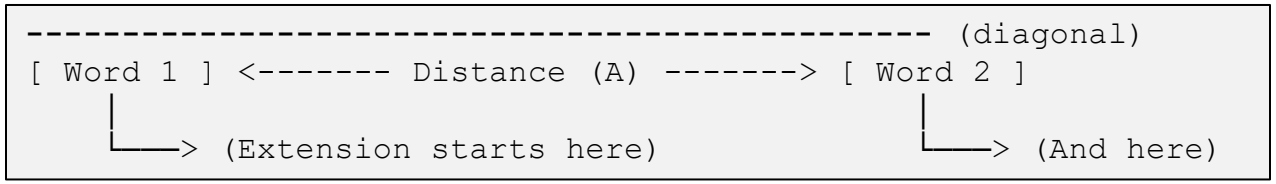
Las palabras de alto score de longitud ktup son las semillas del alineamiento.

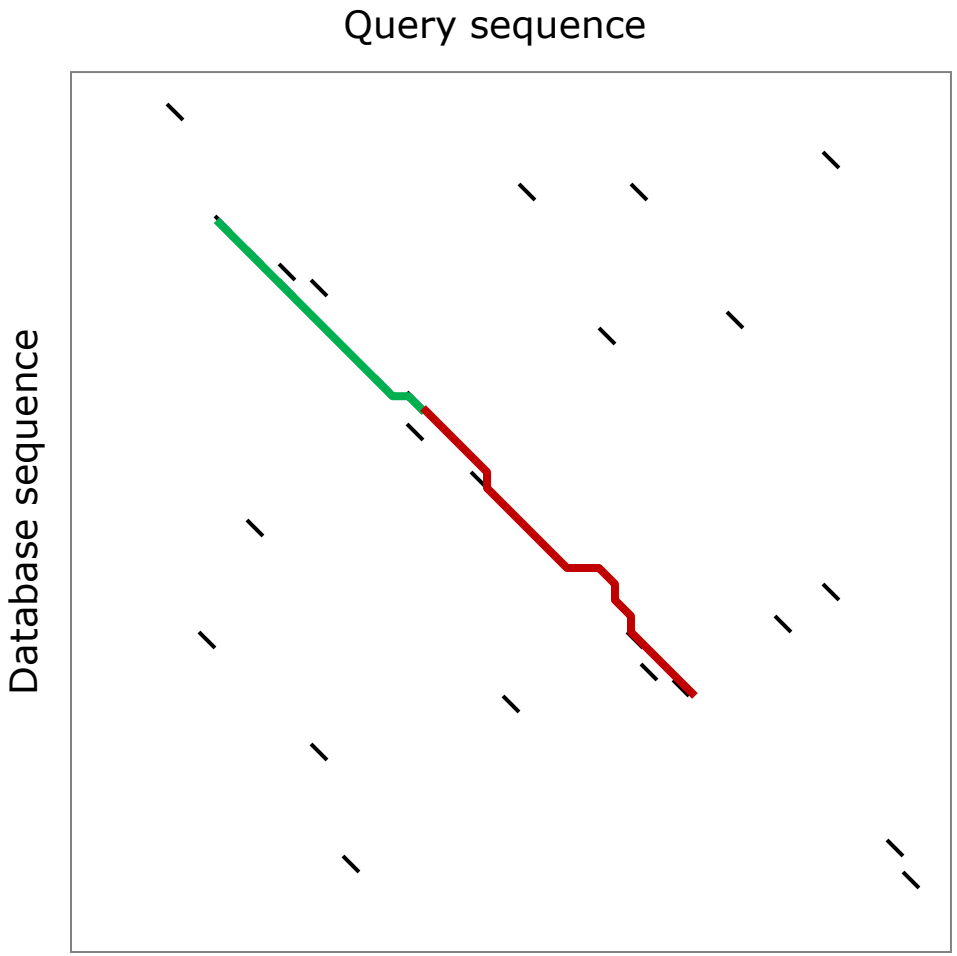
1) BLAST identifica regiones de alta densidad de semillas

2) Regla de proximidad: elige palabras de alto score a menos de una distancia A

3) Extiende las palabras en ambas direcciones para calcular el alineamiento

4) El alineamiento se corta cuando el score cae más que un valor X predeterminado (drop-off)





BLAST extends from word hits

Heurísticas: BLAST

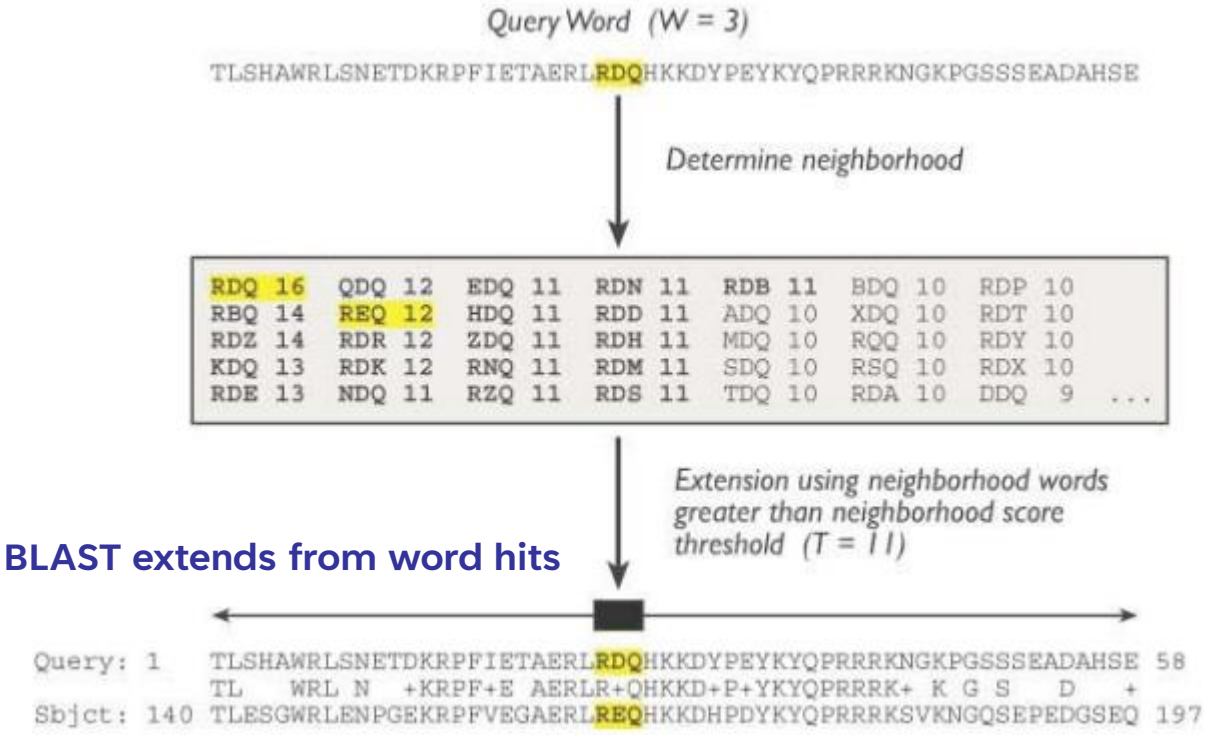
Las palabras de alto score de longitud ktup son las semillas del alineamiento.

1) BLAST identifica regiones de alta densidad de semillas

2) Regla de proximidad: elige palabras de alto score a menos de una distancia A

3) BLAST extiende las palabras en ambas direcciones para calcular el alineamiento

4) El alineamiento termina cuando el score cae más que un valor X predeterminado (drop-off)

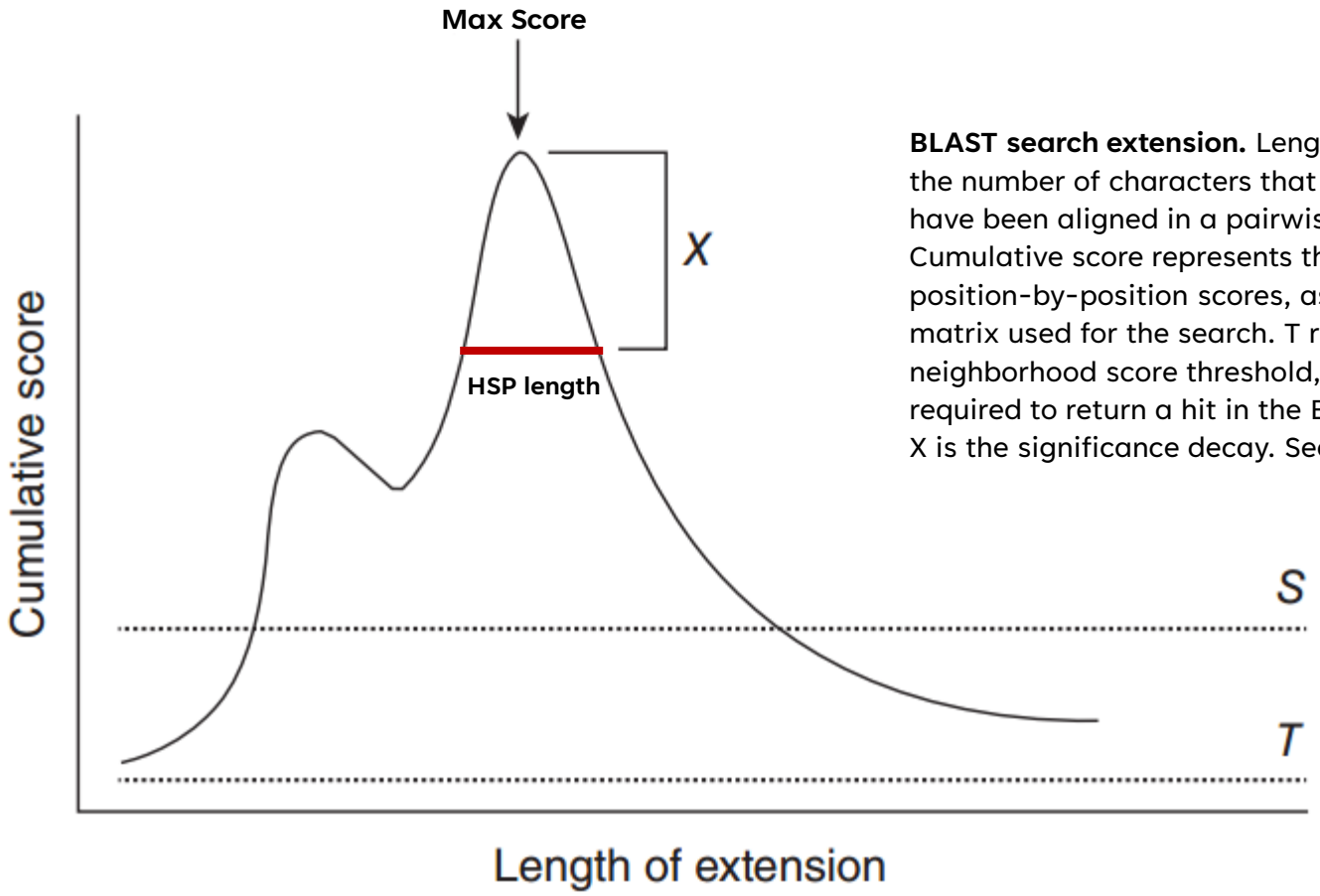


Bioinformatics. Andreas D. Baxevanis, Gary D. Bader, David S. Wishart. John Wiley & Sons, 4th Edition (2020)

```
------------------------------------------------ (diagonal)
[ Word 1 ] <------- Distance (A) -------> [ Word 2 ]
    |                                         |
    └──> (Extension starts here)              └──> (And here)
```

BLAST puede producir varios alineamientos (HSPs)

HSP = "*high scoring pair*" (o sea un alineamiento corto, local, de alto score, entre la secuencia "*query*" y una secuencia "*subject*" de la base de datos.)

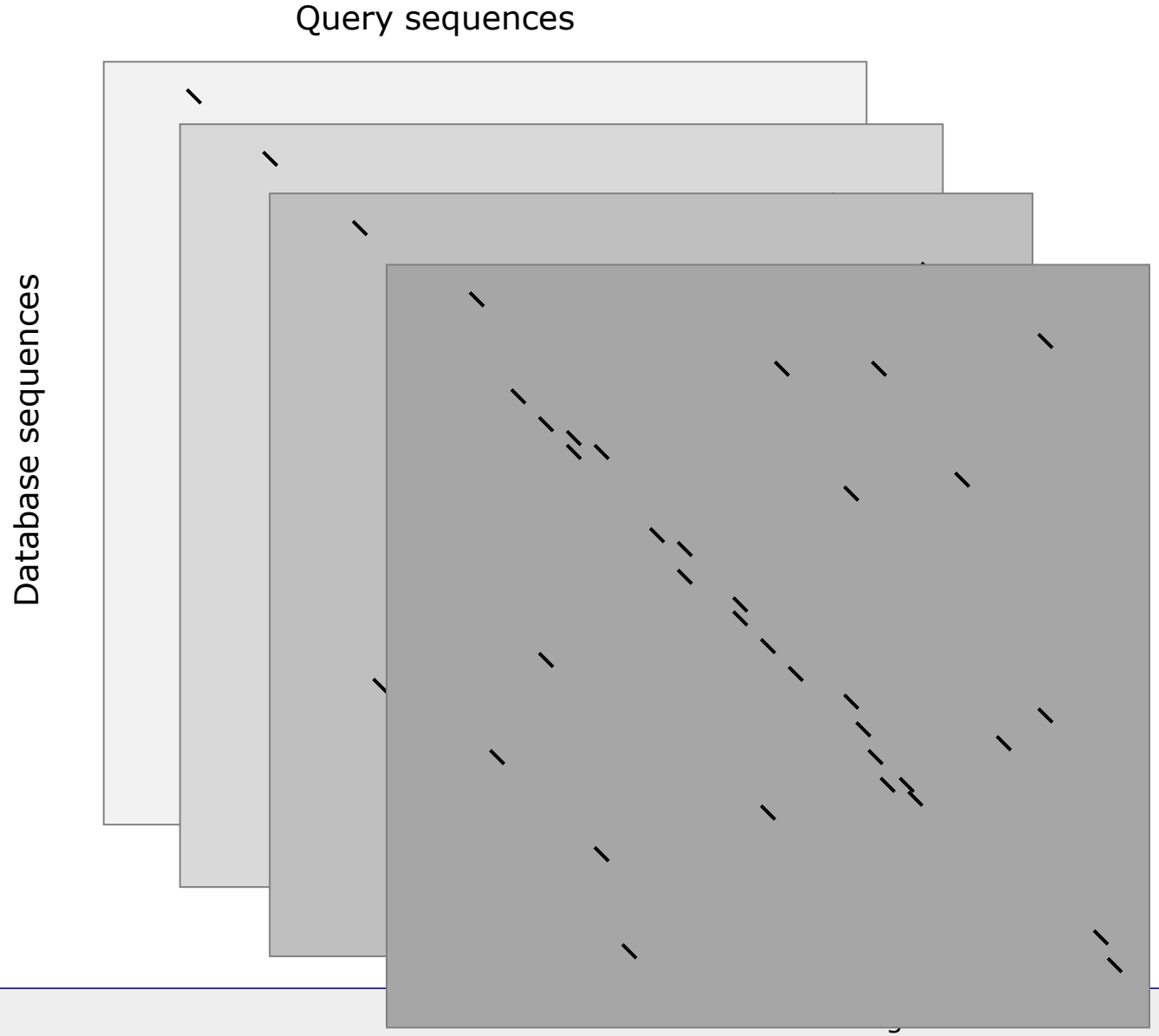


BLAST search extension. Length of extension represents the number of characters that have been aligned in a pairwise sequence comparison. Cumulative score represents the sum of the position-by-position scores, as determined by the scoring matrix used for the search. T represents the neighborhood score threshold, S is the minimum score required to return a hit in the BLAST output, and X is the significance decay. See text for details.

Bioinformatics. Andreas D. Baxevanis, Gary D. Bader, David S. Wishart. John Wiley & Sons, 4th Edition (2020)

Heurísticas para búsquedas

Esto se repite por cada secuencia de la base de datos (BLAST, FASTA).

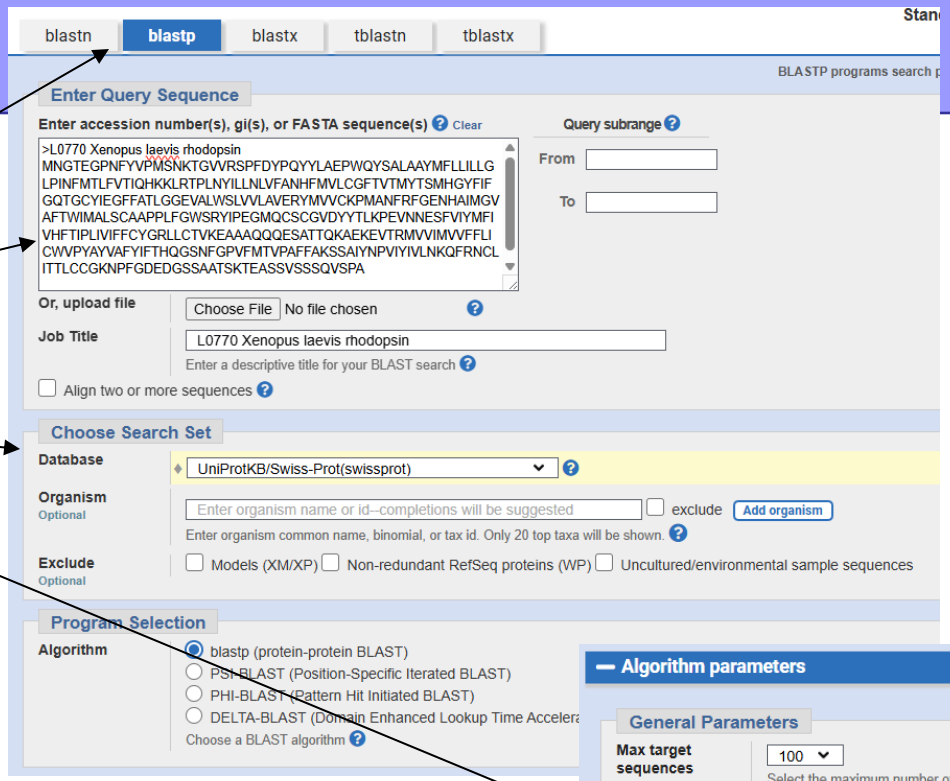


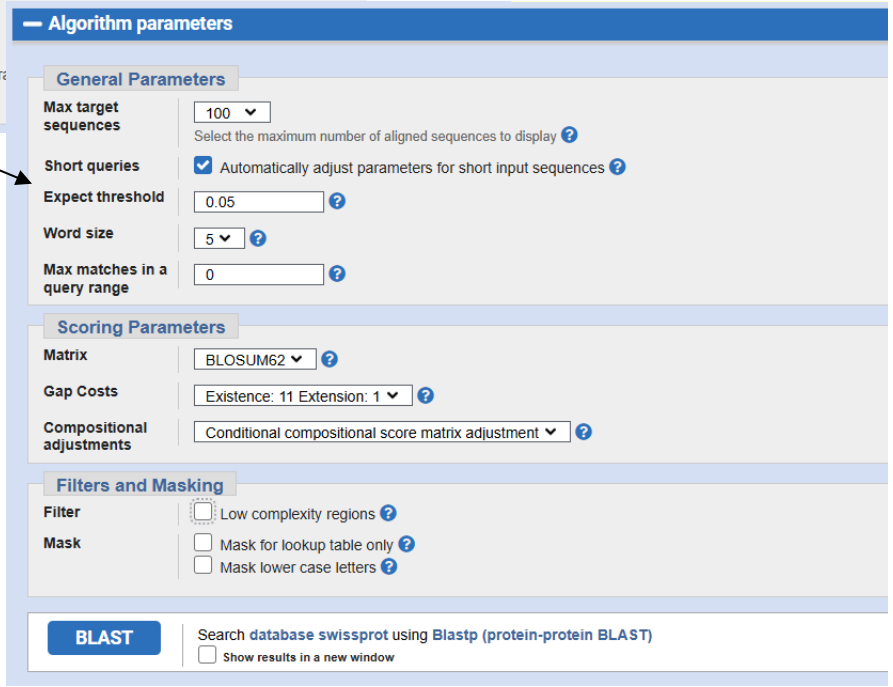
INTERVALO

Fin del Slot 1

Búsquedas en bases de datos

Una búsqueda típica tiene
4 elementos básicos:
1) Programa / Algoritmo
2) Secuencia query
3) Base de Datos
4) Parámetros opcionales



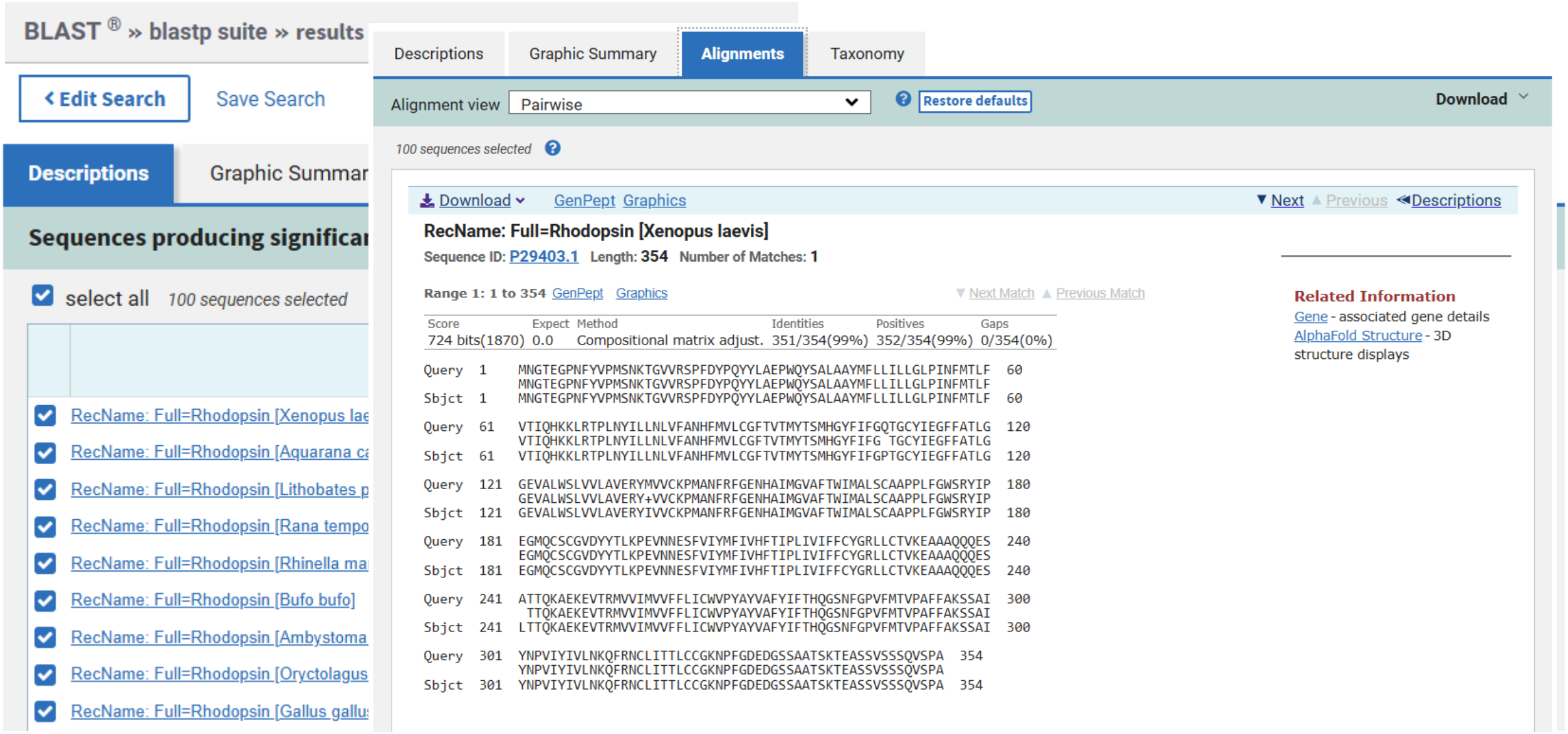



```
> blastp myquery  swissprot  -word_size 5
```

Programa — Secuencia query — Base de datos — Parámetros opcionales

https://www.ncbi.nlm.nih.gov/books/NBK279684/table/appendices.T.blastp_application_options/
https://www.ncbi.nlm.nih.gov/books/NBK279684/table/appendices.T.blastn_application_options/

BLAST results



BLASTP results

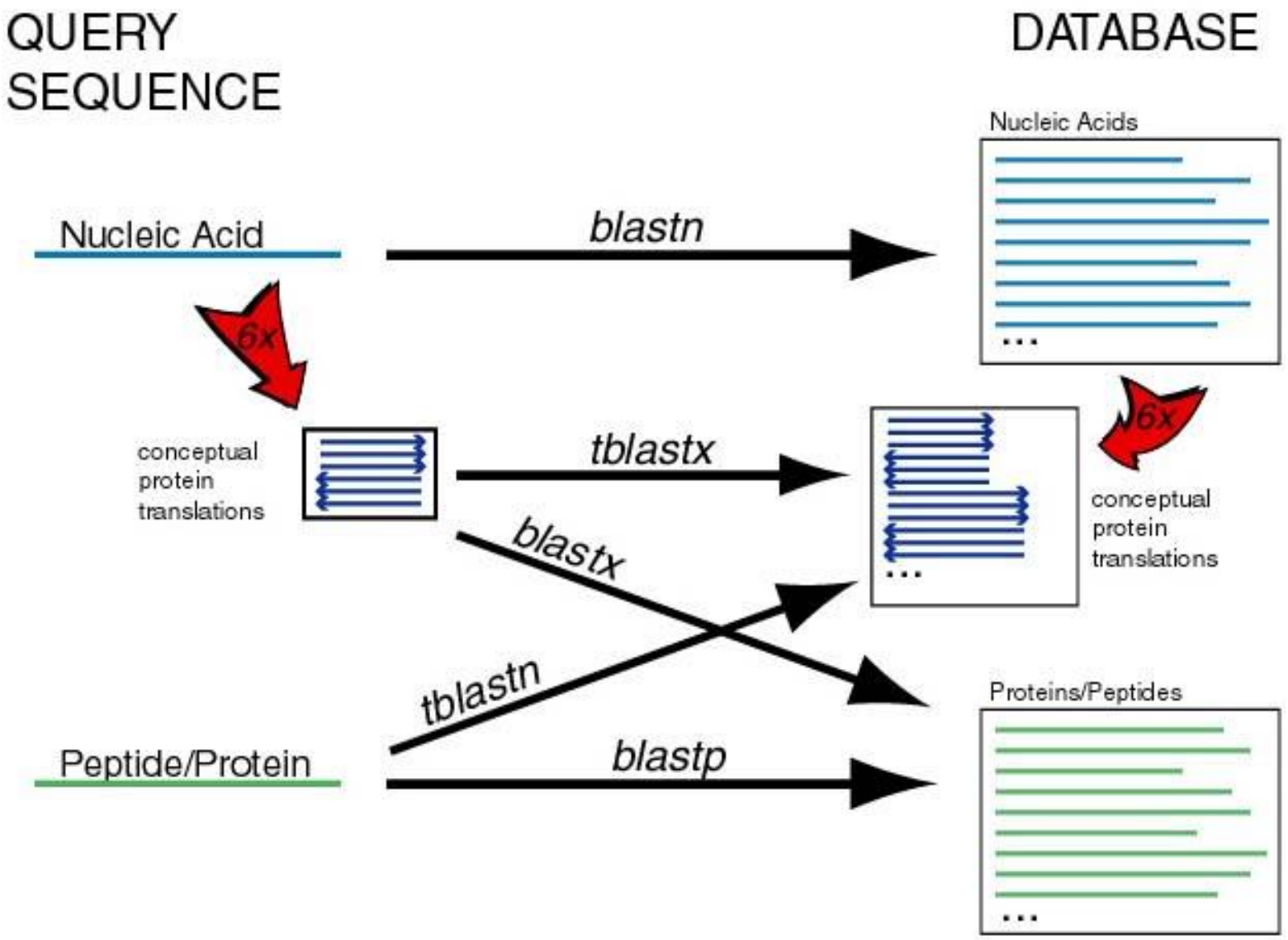
```
> blastp myquery  swissprot  -word_size 5
```

Programa — Secuencia query — Base de datos — Parámetros opcionales

```
>O62791.1 RecName: Full=Rhodopsin [Delphinus delphis]
Length=348

 Score = 616 bits (1589),  Expect = 0.0, Method: Compositional matrix adjust.
 Identities = 287/354 (81%), Positives = 326/354 (92%), Gaps = 6/354 (2%)

Query  1    MNGTEGPNFYVPMSNKTGVVRSPFDYPQYYLAEPWQYSALAAYMFLLILLGLPINFMTLF  60
            MNGTEG NFYVP SNKTGVVRSPF+YPQYYLAEPWQ+S LAAYMFLLI+LG PINF+TL+
Sbjct  1    MNGTEGLNFYVPFSNKTGVVRSPFEYPQYYLAEPWQFSVLAAYMFLLIVLGFPINFLTLY  60

Query  61   VTIQHKKLRTPLNYILLNLVFANHFMVLCGFTVTMYTSMHGYFIFGQTGCYIEGFFATLG  120
            VT+QHKKLRTPLNYILLNL  AN FMV  GFT T+YTS+H YF+FG TGC +EGFFATLG
Sbjct  61   VTVQHKKLRTPLNYILLNLAVANLFMVFGGFTTTLYTSLHAYFVFGPTGCNLEGFFATLG  120

Query  121  GEVALWSLVVLAVERYMVVCKPMANFRFGENHAIMGVAFTWIMALSCAAPPLFGWSRYIP  180
            GE+ALWSLVVLA+ERY+VVCKPM+NFRFGENHAIMG+A TWIMA++CAAPPL GWSRYIP
Sbjct  121  GEIALWSLVVLAIERYVVVCKPMSNFRFGENHAIMGLALTWIMAMACAAPPLVGWSRYIP  180

Query  181  EGMQCSCGVDYYTLKPEVNNESFVIYMFIVHFTIPLIVIFFCYGRLLCTVKEAAAQQQES  240
            EGMQCSCG+DYYTL PEVNNESFVIYMF+VHFTIPL++IFFCYG+L+ TVKEAAAQQQES
Sbjct  181  EGMQCSCGIDYYTLSPEVNNESFVIYMFVVHFTIPLVIIFFCYGQLVFTVKEAAAQQQES  240

Query  241  ATTQKAEKEVTRMVVIMVVFFLICWVPYAYVAFYIFTHQGSNFGPVFMTVPAFFAKSSAI  300
            ATTQKAEKEVTRMV+IMVV FLICWVPYA VAFYIFTHQGS+FGP+FMT+P+FFAKSS+I
Sbjct  241  ATTQKAEKEVTRMVIIMVVAFLICWVPYASVAFYIFTHQGSDFGPIFMTIPSFFAKSSSI  300

Query  301  YNPVIYIVLNKQFRNCLITTLCCGKNPFGDEDGSSAATSKTEASSVSSSQVSPA  354
            YNPVIYI++NKQFRNC++TTLCCG+NP GD++ S+ A SKTE     +SQV+PA
Sbjct  301  YNPVIYIMMNKQFRNCMLTTLCCGRNPLGDDEASTTA-SKTE-----TSQVAPA  348
```

```
BLASTP 2.12.0+


Reference: Stephen F. Altschul, 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.


Reference for composition-based statistics: Alejandro A. Schaffer,
L. Aravind, Thomas L. Madden, Sergei Shavirin, John L. Spouge, Yuri
I. Wolf, Eugene V. Koonin, and Stephen F. Altschul (2001),
"Improving the accuracy of PSI-BLAST protein database searches with
composition-based statistics and other refinements", Nucleic Acids
Res. 29:2994-3005.



Database: Non-redundant UniProtKB/SwissProt sequences
           487,492 sequences; 185,956,103 total letters



Query= L07770

Length=354
                                                                    Score     E
Sequences producing significant alignments:                        (Bits)  Value

P29403.1 RecName: Full=Rhodopsin [Xenopus laevis]                   724     0.0
P51470.1 RecName: Full=Rhodopsin [Aquarana catesbeiana]             694     0.0
P31355.1 RecName: Full=Rhodopsin [Lithobates pipiens]               688     0.0
P56516.1 RecName: Full=Rhodopsin [Rana temporaria]                  681     0.0
P56515.1 RecName: Full=Rhodopsin [Rhinella marina]                  672     0.0
P56514.1 RecName: Full=Rhodopsin [Bufo bufo]                        660     0.0
Q90245.1 RecName: Full=Rhodopsin [Ambystoma tigrinum]               654     0.0
P49912.1 RecName: Full=Rhodopsin [Oryctolagus cuniculus]            630     0.0
P22328.1 RecName: Full=Rhodopsin [Gallus gallus]                    628     0.0
Q68J47.1 RecName: Full=Rhodopsin [Loxodonta africana]               628     0.0
Q95KU1.1 RecName: Full=Rhodopsin [Felis catus]                      626     0.0
P15409.2 RecName: Full=Rhodopsin [Mus musculus]                     625     0.0
P52202.1 RecName: Full=Rhodopsin [Alligator mississippiensis]       622     0.0
Q28886.1 RecName: Full=Rhodopsin [Macaca fascicularis]              621     0.0
```

BLAST: algoritmos



Evaluando alineamientos

- Qué hacemos cuando estamos comparando dos secuencias que no son similares, pero que muestran un alineamiento prometedor?
- Necesitamos un **test de significancia**
- Tenemos que responder a la pregunta:
  - **Cuál es la probabilidad de que un alineamiento similar (con un score similar) ocurra entre proteínas no relacionadas?**

Solamente con el score basta?

Dos secuencias **no relacionadas** (que no son homólogas), pueden tener alineamientos con **el mismo puntaje** o muy cercano?

Estadística de los alineamientos

Dos estrategias alternativas:

- Generar **secuencias al azar** de la misma longitud y composición que la secuencia query y alinearlas
  - Karlin & Altschul (1990); Altschul et al (1994); Altschul & Gish (1996)
- Suponer que la base de datos **es suficientemente grande** y está compuesta mayormente por secuencias **no relacionadas**
  - Pearson (1998)

- Analizar **la distribución de scores** que se obtiene

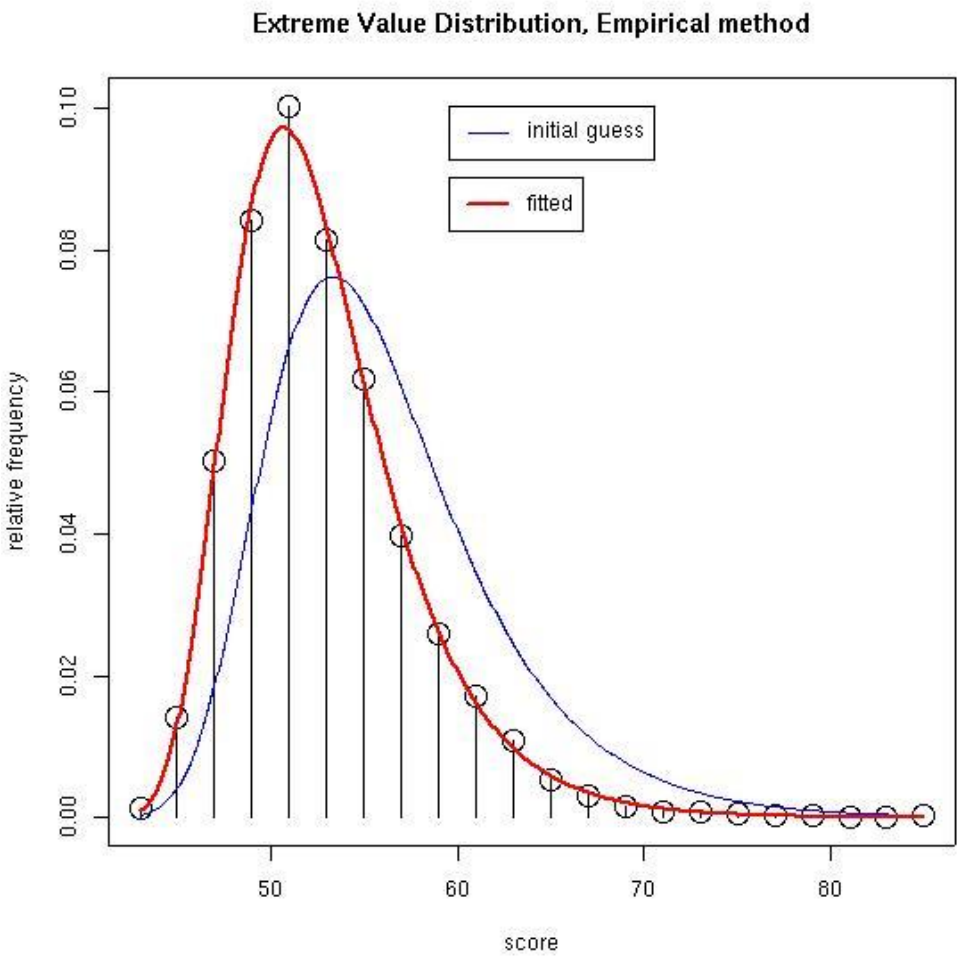
https://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html

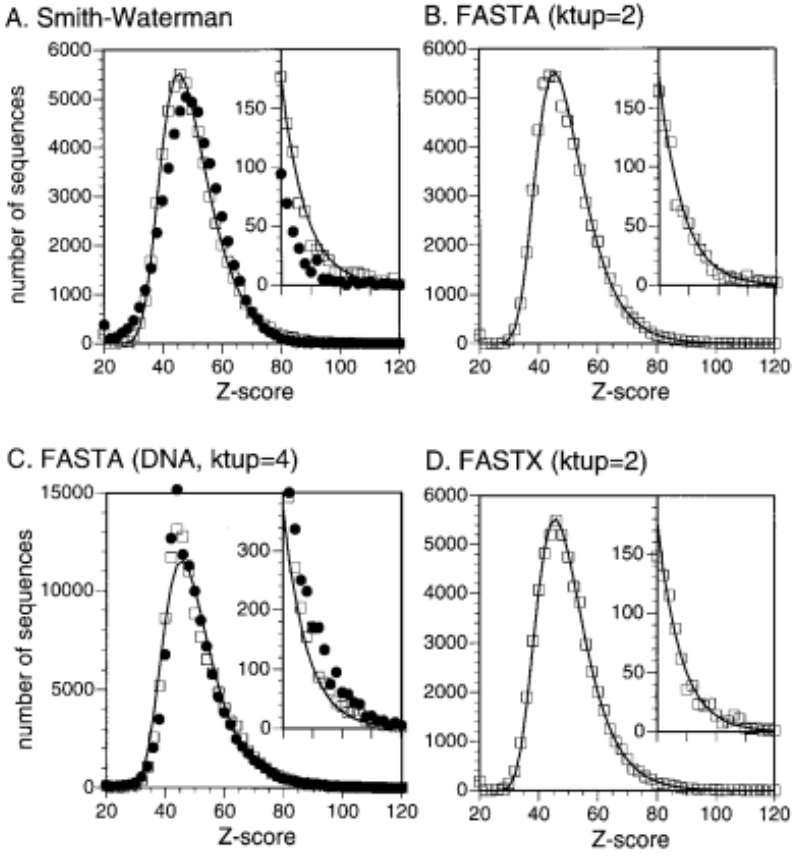
Pearson WR. Empirical statistical estimates for sequence similarity searches. J Mol Biol. 1998 Feb 13;276(1):71-84. doi: 10.1006/jmbi.1997.1525. PMID: 9514730.

Karlin S, Altschul SF. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci U S A. 1990 Mar;87(6):2264-8. doi: 10.1073/pnas.87.6.2264. PMID: 2315319; PMCID: PMC53667.

Análisis de las distribuciones de scores

In a database search (BLAST/FASTA) the alignment scores **do not** follow a normal/Gaussian distribution!





The Gumbel/Extreme value distribution

In a database search (BLAST/FASTA) the alignment scores follow a Gumbel or Extreme Value distribution!

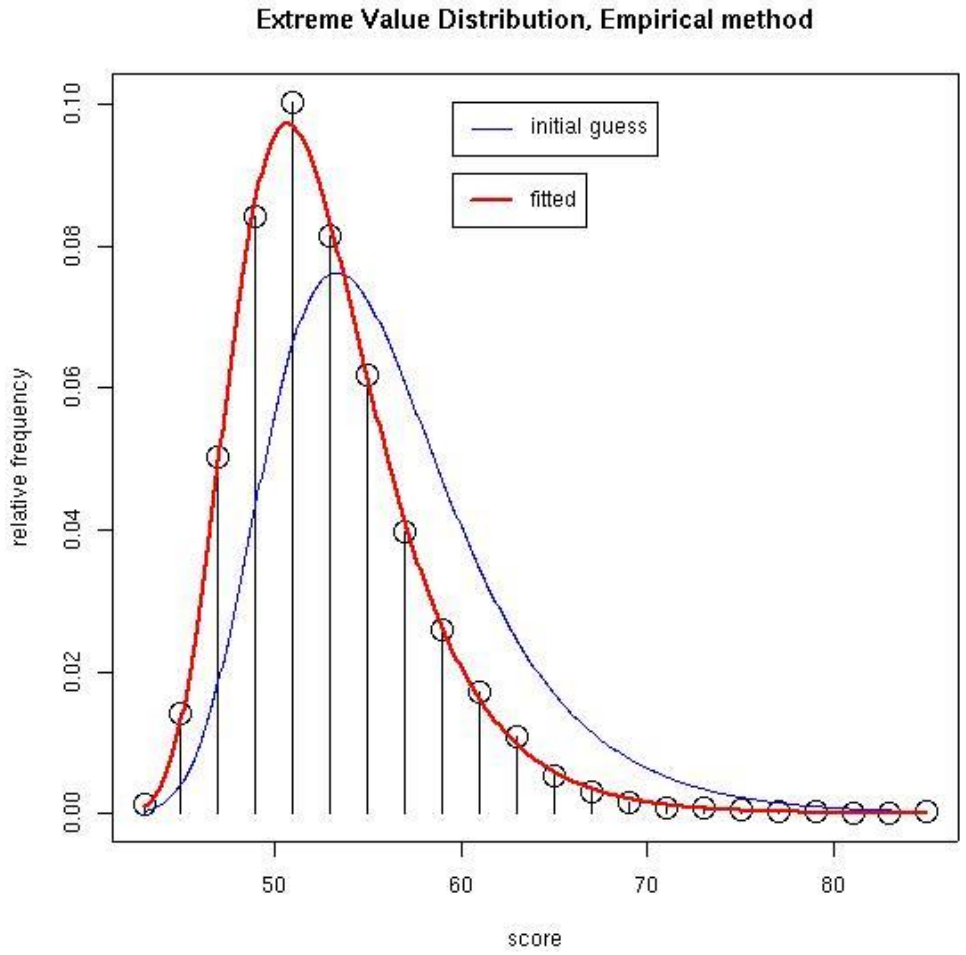
$$E = K\,mn\,e^{-\lambda S}$$

E is the **number** of alignments with a score = S

m,n: length of the sequences

K,λ, : estimated parameters estimated (depend on the scoring matrix and the size of the database)

e: Euler's number (mathematical constant)

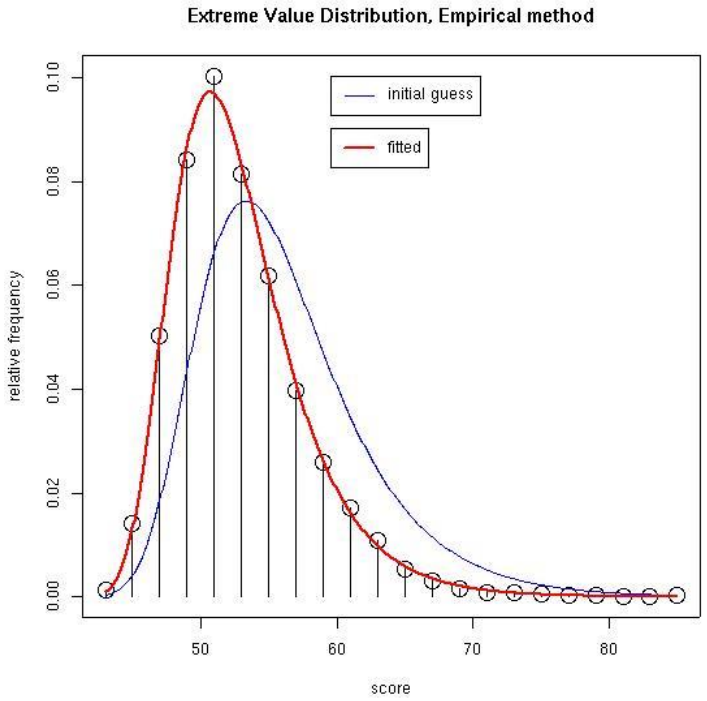


E-value

Los hits pueden ser ordenados de acuerdo a su E-value o a su Score.
El E-value – más conocido como **EXPECT** value – es una función del score, el tamaño de la base de datos y de la longitud de la secuencia 'query'.
E-value: Número de alineamientos con un score = S que se espera encontrar si la base de datos es una colección de letras al azar.
Ejemplo: En el caso de un score=1 (un match o identidad) debería haber un número enorme de alineamientos. Uno espera encontrar menos alineamientos con un score de 5, 10, etc. Eventualmente, cuando el score es lo suficientemente alto, uno espera encontrar un número insignificante de alineamientos que sean debidos al azar.
Valores de E-value menores que 1e-6 ($1* 10^{-6}$) son generalmente muy buenos para proteínas, mientras que E<1e-2 puede considerarse significativo. Es posible que un hit cuyo E > 1 sea biológicamente importante, aunque es necesario analizarlo más detalladamente para confirmarlo.

Observed vs expected

Si la base de datos es suficientemente grande y contiene mayoritariamente secuencias no relacionadas la distribución de scores **observados** debería coincidir bastante con la distribución de scores **esperados por azar** (Pearson 1998)



Pearson WR. Empirical statistical estimates for sequence similarity searches. J Mol Biol. 1998 Feb 13;276(1):71-84. doi: 10.1006/jmbi.1997.1525. PMID: 9514730.

Karlin S, Altschul SF. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci U S A. 1990 Mar;87(6):2264-8. doi: 10.1073/pnas.87.6.2264. PMID: 2315319; PMCID: PMC53667.

Influencia del tamaño de la base de datos

$$E(S > x) = p(S > x)\, D$$

El número de alineamientos con un score > S se incrementa linealmente con el tamaño de la base de datos

Tamaño de la base de datos: un ejemplo

Objetivo: encontrar en *E. coli* el gen homólogo al gen AroA de *Bacillus* subtilis (DAHP synthase)

- ***Database* = *E. coli* proteome (4,283 proteínas)**
  - **kdsA, E(4283) < 0.00015**
- **Database = Swissprot (74,417 proteínas)**
  - **kdsA, E(74417) < 0.0017**
- **Database = OWL (260,784 proteínas)**
  - **kdsA, E(260784) < 0.0085**

- **El mismo alineamiento, con el mismo score es 50 veces más significativo en la base de datos más chica.**

Disclaimer: los tamaños de las bases de datos son ilustrativos y corresponden a 1998-1999 cuando se hizo este ejercicio.

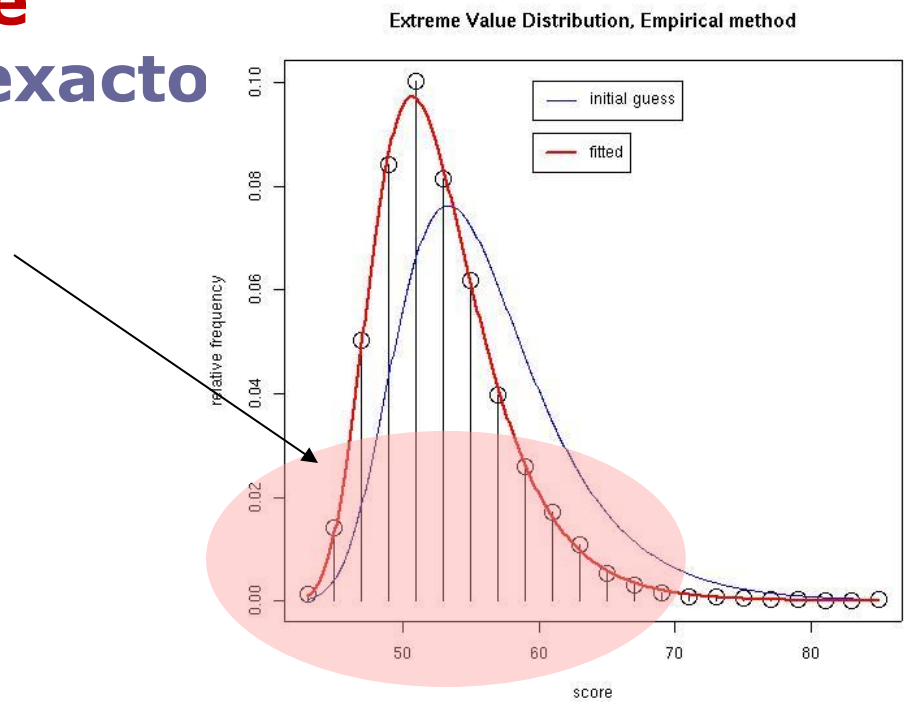
Bleasby AJ, Akrigg D, Attwood TK. OWL--a non-redundant composite protein sequence database. Nucleic Acids Res. 1994 Sep;22(17):3574-7. PMID: 7937061; PMCID: PMC308323.

Identificar homólogos con eficiencia

Problema: estamos buscando homólogos lejanos

Estrategia:

- Buscar en bases de datos **pequeñas primero**
- Repetir la búsqueda en una base de datos pequeña con un **algoritmo más sensible** (word_size / ktup = 1 o un algoritmo exacto sin heurísticas, ej ssearch)
- Dudar de alineamientos de bajo score



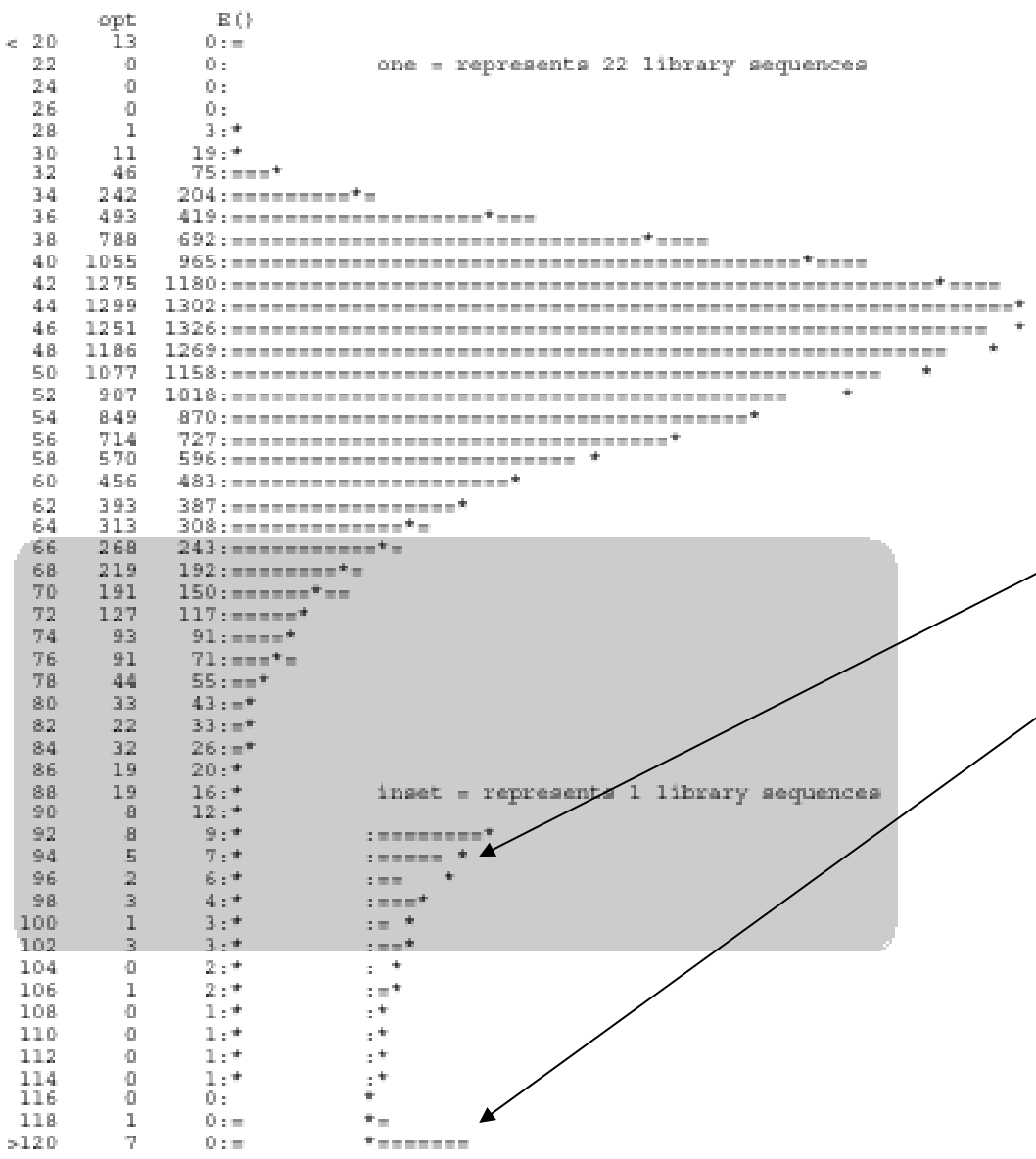
Límites de la estadística

- **En ciertos casos, la estadística de los alineamientos falla**
  - **Lo que falla son las suposiciones que hicimos para llegar al modelo estadístico**
- **Se obtienen estimaciones incorrectas del E-value cuando**
  - **Se usan penalidades de gap incorrectas o extremas**
  - **Existen regiones de baja complejidad en la secuencia query**
  - **La secuencia query tiene una composición sesgada**

>sp|P42858|HD_HUMAN Huntingtin OS=Homo sapiens OX=9606 GN=HTT
MATLEKLMKAFESLKSFQQQQQQQQQQQQQQQQQQQQQPPPPPPPPPPQLPQPPPQAQP
LLPQPQPPPPPPPPPGPAVAEEPLHRPKKELSATKKDRVNHCLTICENIVAQSVRNSPE
FQKLLGIAMELFLLCSDDAESDVRMVADECLNKVIKALMDSNLPRLQLELYKEIKKNGAP
RSLRAALWRFAELAHLVRPQKCRPYLVNLLPCLTRTSKRPEESVQETLAAAVPKIMASFG
NFANDNEIKVLLKAFIANLKSSSPTIRRTAAGSAVSICQHSRRTQYFYSWLLNVLLGLLV
PVEDEHSTLLILGVLLTLRYLVPLLQQQVKDTSLKGSFGVTRKEMEVSPSAEQLVQVYEL
TLHHTQHQDHNVVTGALELLQQLFRTPPPELLQTLTAVGGIGQLTAAKEESGGRSRSGSI
[…]

>sp|P04156|PRIO_HUMAN Major prion protein OS=Homo sapiens OX=9606 GN=PRNP
MANLGCWMLVLFVATWSDLGLCKKRPKPGGWNTGGSRYPGQGSPGGNRYPPQGGGGWGQP
HGGGWGQPHGGGWGQPHGGGWGQPHGGGWGQGGGTHSQWNKPSKPKTNMKHMAGAAAAGA
VVGGLGGYMLGSAMSRPIIHFGSDYEDRYYRENMHRYPNQVYYRPMDEYSNQNNFVHDCV
NITIKQHTVTTTTKGENFTETDVKMMERVVEQMCITQYERESQAYYQRGSSMVLFSSPPV
ILLISFLIFLIVG

Evaluando la estadística en casos extremos



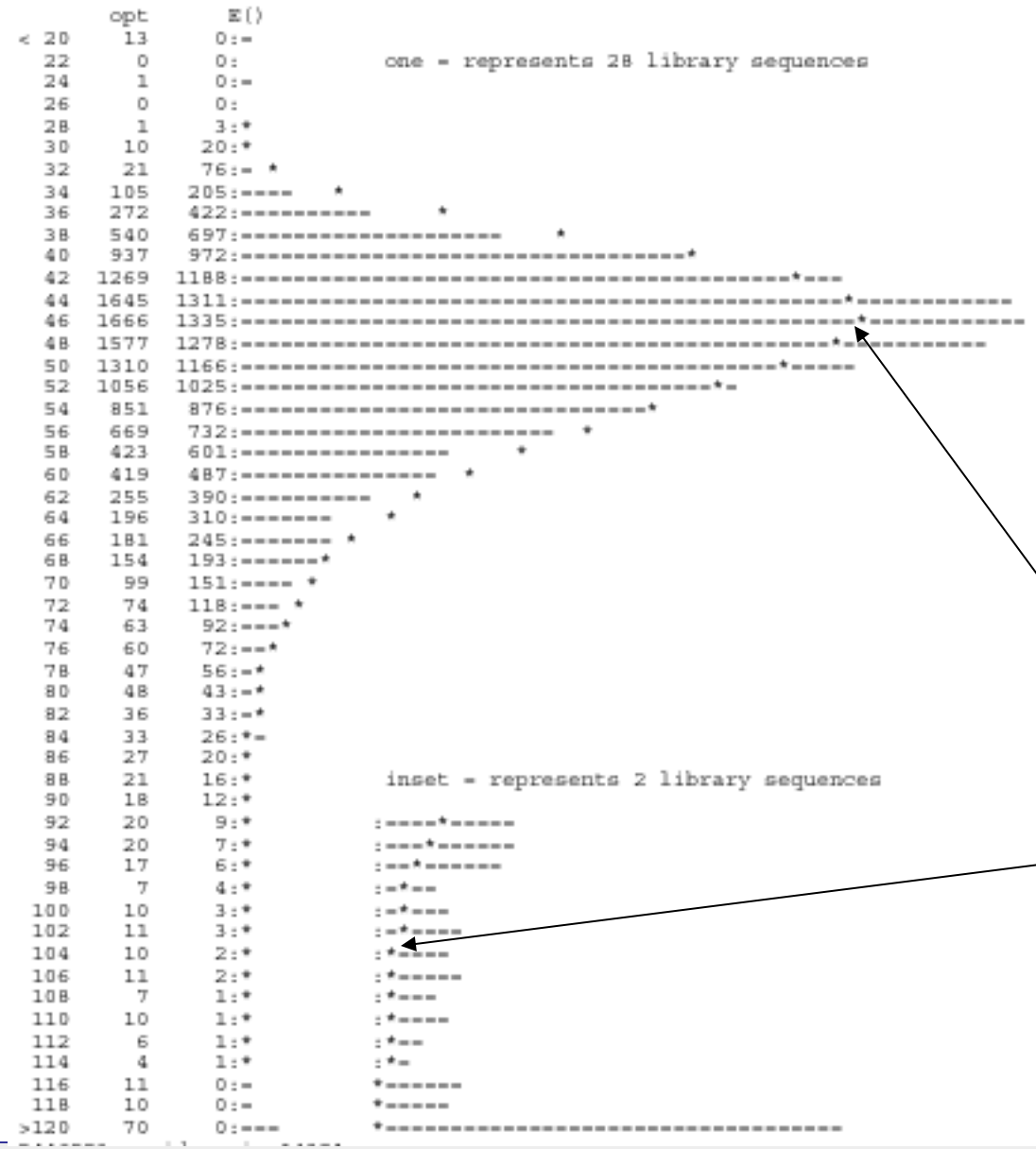
Distribución (histograma) de scores esperados y observados

* → observados

= → esperados

Mirar el E-value de la secuencia no relacionada con mayor score

Evaluando la estadística (cont)



Si los histogramas de Observados vs Esperados coinciden

Y si el E del mejor alineamiento no relacionado es <<< 1

La estimaciones estadísticas están funcionando bien.

Y en este caso?

* → observados

= → esperados

Buscando homólogos en los límites

- **Secuencias homólogas distantes a menudo no tienen similitud estadísticamente significativa**
- **Secuencias con regiones de baja complejidad pueden tener similitud estadísticamente significativas, aunque no sean homólogas**

- **Secuencias homólogas generalmente son similares sobre toda la longitud de la secuencia o de un dominio**

- **Secuencias homólogas comparten un ancestro común**
  - **Si hay homología entre A y B; entre B y C; y entre C y D, A y D deben ser homólogos, aun cuando no muestren similitud estadísticamente significativa**

Low complexity sequences

- **Secuencias (o sub-secuencias) con bajo contenido de información**
  - AAAAAAAAAAAAAAAAAAAA
  - CAACAACAACAACAACAACAA
- **Secuencias con sesgo en la composición de bases (nucleótidos) o residuos (aminoacidos)**
- **Por el bajo contenido de información y el sesgo en la composición, suelen dar falsos positivos en las búsquedas por similitud**
  - PolyQ: proteínas con tractos de polyglutamina no están relacionadas por ancestría
    - Ej OTX2 (Transcription factor); CREB-binding protein (connects proteins with different functions); MED15/GAL11 (Subunit of the RNA polymerase II mediator complex)

BLAST compositional adjustment of scoring matrices

Las matrices de scoring (ej BLOSUM62) son derivadas a partir de alineamientos de proteinas globulares, con una composición de aminoacidos definida

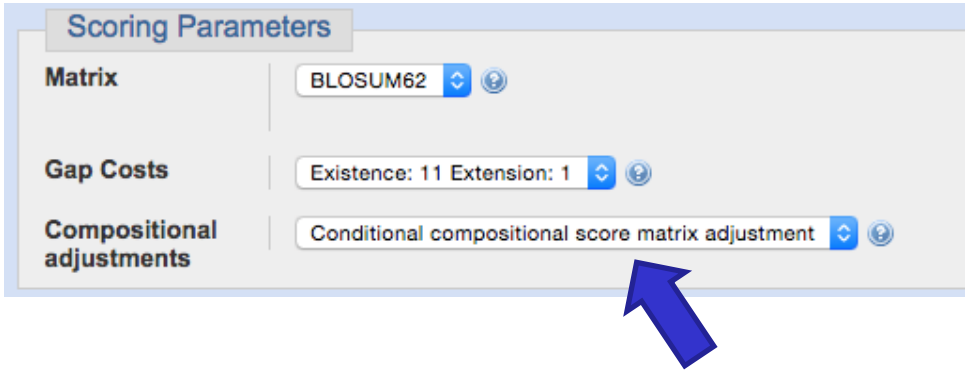
Pero muchas veces las secuencias “query” tienen composiciones de aminoácidos con sesgos muy marcados y diferentes a las de las secuencias utilizadas para derivar las matrices

- Ej proteínas hidrofóbicas, Cys-rich, proteínas codificadas por genomas con alto sesgo (AT rich, GC rich), que afectan los codones más utilizados

En esos casos hay maneras de ajustar las matrices (ej recalcular BLOSUM62 on-the-fly) para que funcionen mejor frente a estos casos

BLAST compositional adjustment of scoring matrices

Query = human insulin NP_000198
Program = blastp
Database = *C. elegans* RefSeq



Option = NO compositional adjustment

```
>ref|NP_501926.1| INSulin related family member (ins-1) [Caenorhabditis elegans]
Length=109

 Score = 34.7 bits (78),  Expect = 0.009
 Identities = 30/100 (30%), Positives = 41/100 (41%), Gaps = 14/100 (14%)

Query  11   LALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELGG  70
            LA+L L  P P+ A +   LCGS L   L  VC  +        +R A+
Sbjct  17   LAILLLSSPTPSDASIR--LCGSRLTTTLLAVCRNQLCTGLTAFKRSADQSY--------  66

Query  71   GPGAGSLQPLALEGSLQKRG-IVEQCCTSICSLYQLENYC  109
             A + + L      QKRG I  +CC   CS   L+ +C
Sbjct  67   ---APTTRDLFHIHHQQKRGGIATECCEKLCSFAYLKTFC  103
```

Option = conditional compositional score matrix adjustment

```
>ref|NP_501926.1| INSulin related family member (ins-1) [Caenorhabditis elegans]
Length=109

 Score = 33.5 bits (75),  Expect = 0.020, Method: Compositional matrix adjust.
 Identities = 27/100 (27%), Positives = 39/100 (39%), Gaps = 12/100 (12%)

Query  10   LLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELG  69
             LA+L L  P P+ A +   LCGS L   L  VC  +        +R A+
Sbjct  16   FLAILLLSSPTPSDASIR--LCGSRLTTTLLAVCRNQLCTGLTAFKRSADQ--------S  65

Query  70   GGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYC  109
              P    L    +    ++ GI  +CC   CS   L+ +C
Sbjct  66   YAPTTRDL--FHIHHQQKRGGIATECCEKLCSFAYLKTFC  103
```

https://blast.ncbi.nlm.nih.gov/doc/blast-topics/blastsearchparams.html#compositional-adjustments

BLAST compositional adjustment of scoring matrices

Fig. 1 Examples of alignment extensions yielded by compositional adjustment of the scoring system.

a) standard BLOSUM-62 substitution matrix

b) composition-adjusted matrix derived from BLOSUM-62 with unconstrained relative entropy

c) composition-adjusted matrix derived from BLOSUM-62 with relative entropy constrained to 0.566 bits

Yu YK, Altschul SF. The construction of amino acid substitution matrices for the comparison of proteins with non-standard compositions. Bioinformatics. 2005 Apr 1;21(7):902-11. doi: 10.1093/bioinformatics/bti070. Epub 2004 Oct 27. PMID: 15509610.

Fin