BLAST II. Búsquedas avanzadas

Actualizado en: 24/01/2023

Propósito y estrategia

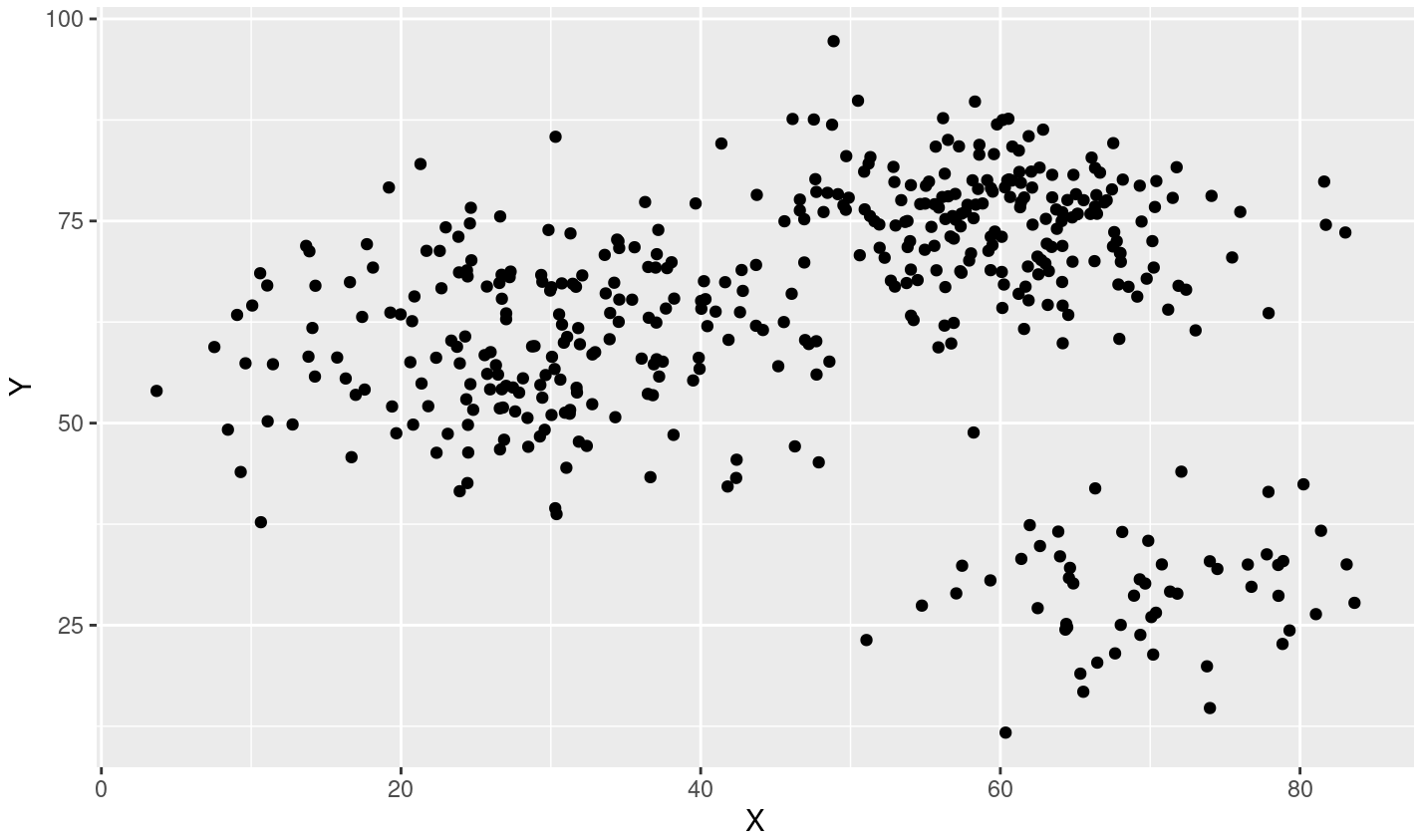
Finalidad	Estrategia
Identificar la especie de origen de una secuencia	Blastn o Megablast (k grande)
Saber si una secuencia nucleotídica codifica algo	Blastx. Raramente, tblastx
Inferir la función de una proteína	Pfam, Blastp
Construir la filogenia de una familia proteica	PSI-Blast. Raramente tblastn.
Buscar proteínas homólogas codificadas en un nuevo genoma	tblastn

Búsquedas sensibles de proteínas

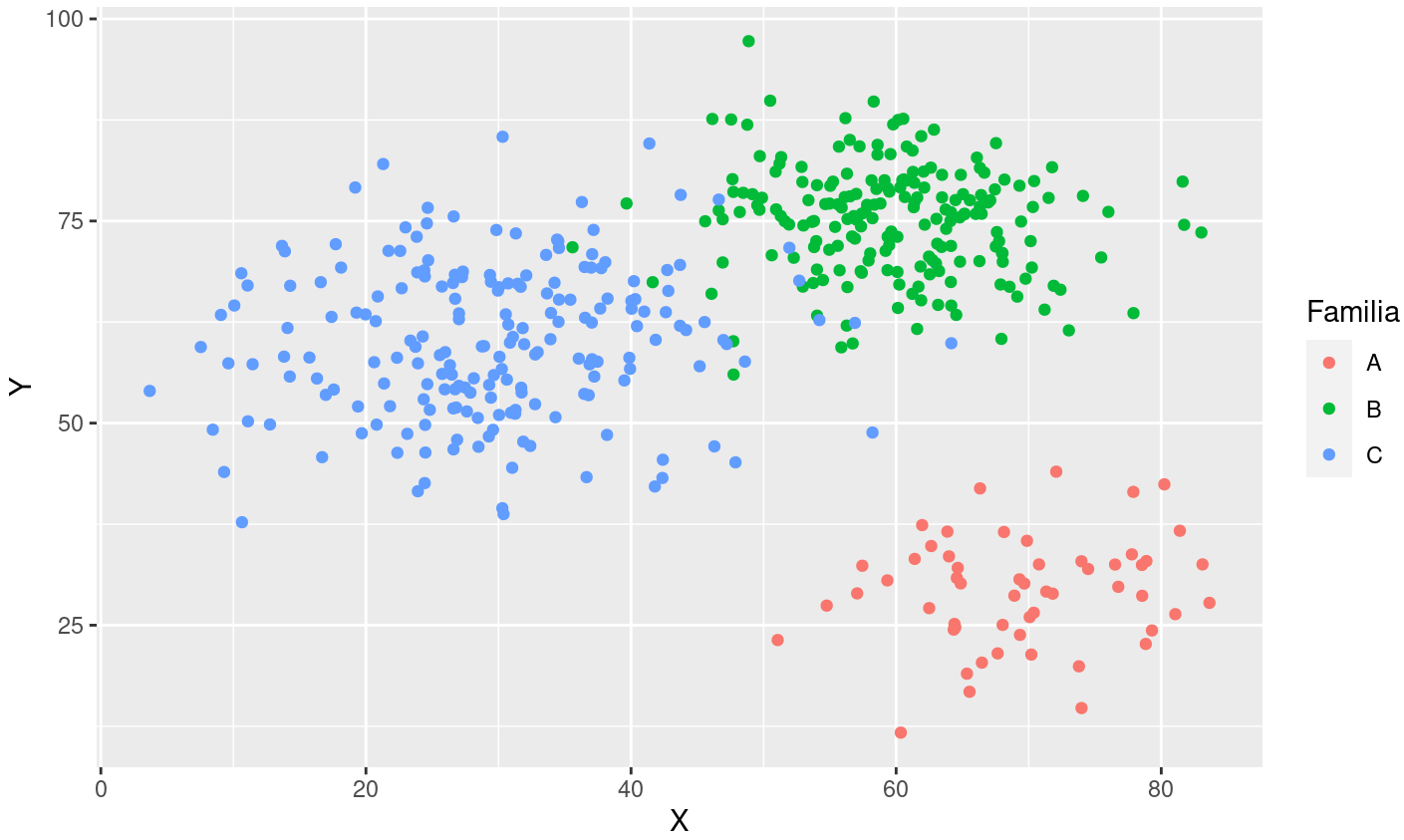
La analogía del *clustering*

- El alineamiento local permite medir similitudes entre porciones de proteínes.
- Similitudes/diferencias se pueden entender como **distancias** en un espacio.
- ¿Hasta qué distancia podemos extrapolar funciones?
- ¿Cómo definir dominios entre un conjunto de secuencias de función desconocida?

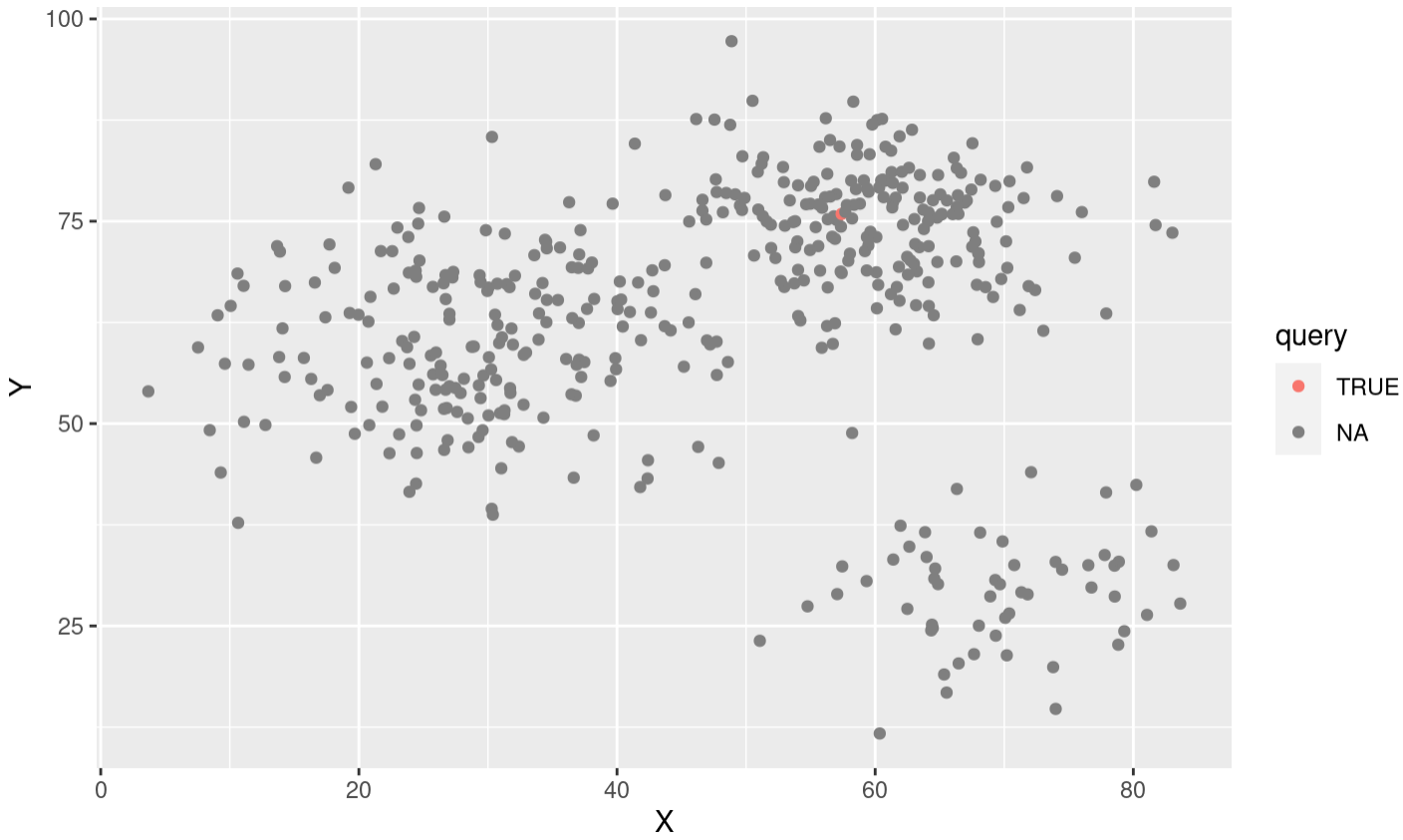
La analogía del *clustering*



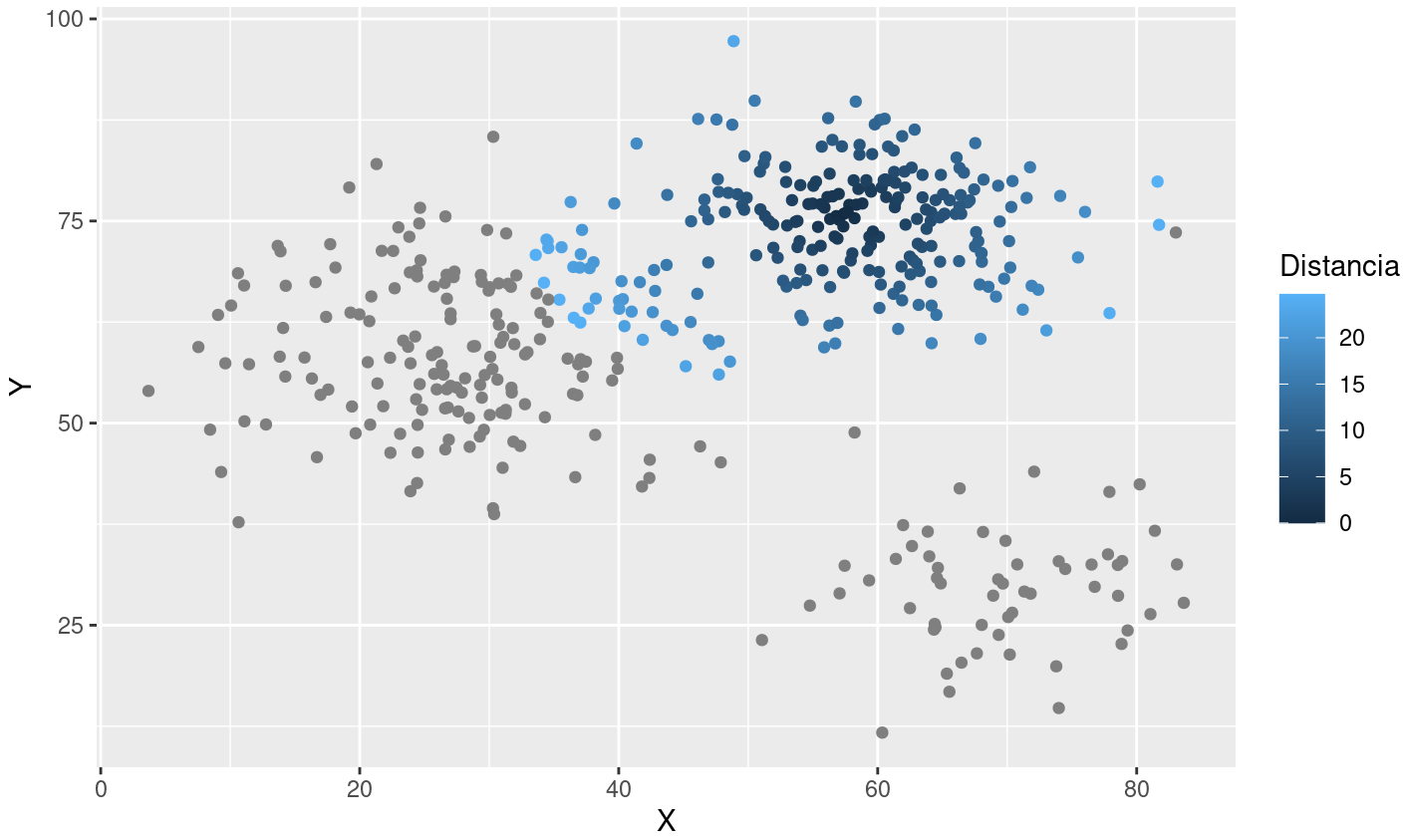
La analogía del *clustering*



La limitación del BLAST



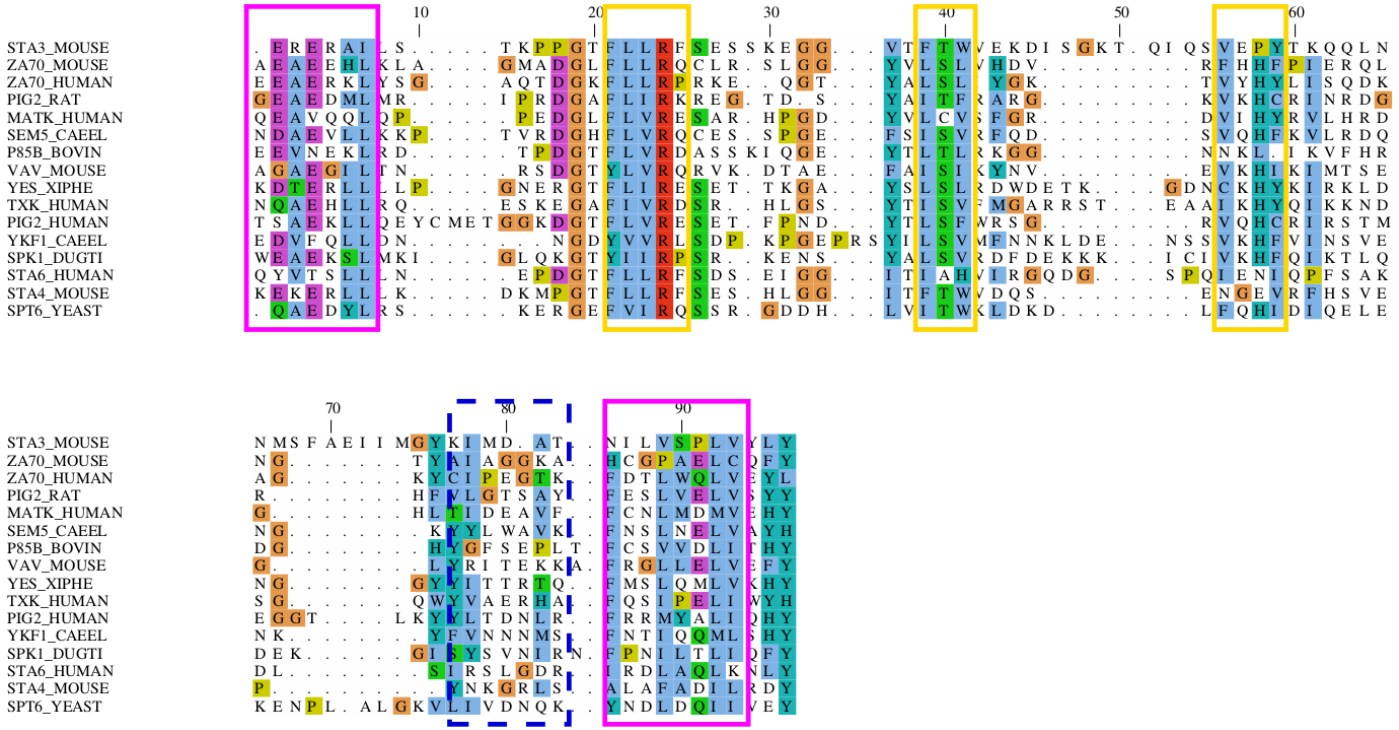
La limitación del BLAST



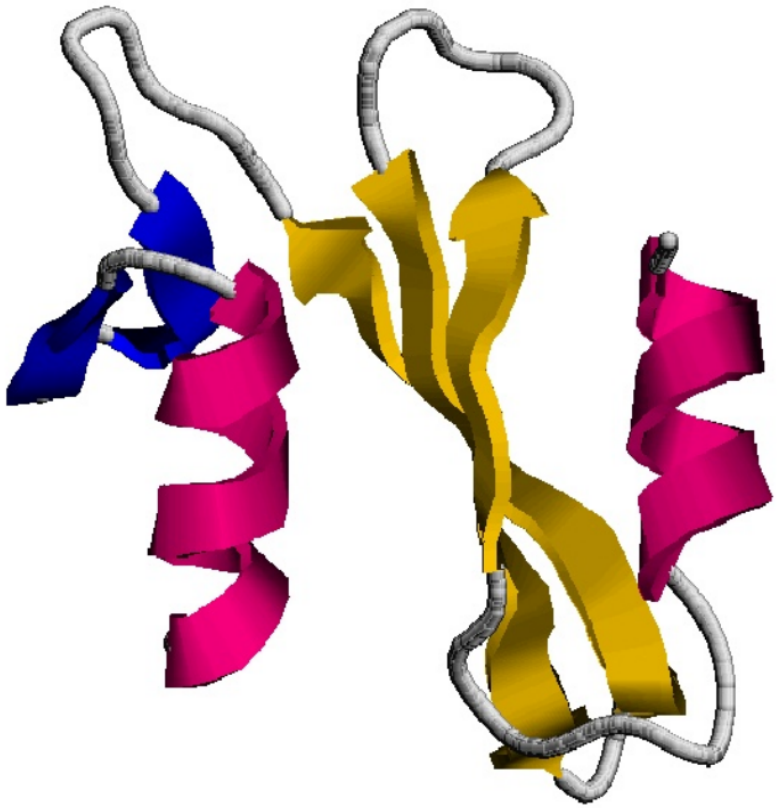
La limitación del BLAST

- Una única *query* sesga los resultados: falta visión de conjunto.
- Falta sensibilidad para detectar homologías distantes.
- La similitud de secuencia no es suficiente para definir dominios.
- ¿Qué tienen en común las proteínas de una familia o con una función?
    - Hace falta un **modelo** de lo que comparten unas u otras proteínas.

Ejemplo: alineamiento y estructura



Ejemplo: alineamiento y estructura



Modelos de regiones conservadas

- Patrones y expresiones regulares.
- *Position Specific Scoring Matrices* (PSSM).
- *Hiden Markov Models* (HMM): Pfam.
- PSI-BLAST
- Delta-BLAST

Patrones y expresiones regulares

Secuencia consenso

```
 GHEGVGKVVKLGAGA   |
 GHEKKGYFEDRGPSA   |
 GHEGYGGRSRGGGYS   |----- alineamiento
 GHEFEGPKGCGALYI   |
 GHELRGTTFMPALEC   |
-----------------
 GHEGVGKVVKLGAGA   |
    KK YFEDRAPSS   |      residuos
    FY GRSRG GYI   |----- observados
    LE PKGCP LEC   |      por columna
     R TTFM        |
-----------------
 GHE**G*****G***   <----- secuencia consenso
```

Secuencia consenso

- Ventajas
  - Rápido y fácil de implementar
- Limitaciones
  - Sin información sobre variancián en columnas.
  - Muy dependiente del alineamiento inicial.
  - Sin puntuación: resultado binario (sí/no).
- Utilidad
  - Patrones muy conservados, como lugares de restricción en DNA.

Expresiones regulares

```
<A-x-[ST](2)-x(0,1)-{V}
```

- Alanina en N-terminal.
- Seguida de cualquier aminoácido.
- Seguido de serina o treonina dos veces.
- Seguidas (o no) por otro residuo cualquiera.
- Seguido por cualquier aminoácido excepto valina.

Expresiones regulares

- Señales post-traduccionales
  - *Protein splicing signature*: [DNEG]-x-[LIVFA]-[LIVMY]-[LVAST]-H-N-[STC]
  - Sitio de fosforilación por tirosina kinasa: [RK]-x(2)-[DE]-x(3)-Y or [RK]-x(3)-[DE]-x(2)-Y
- Señales de ineracción entre DNA y RNA
  - Marca de histona H4: G-A-K-R-H
  - Marca de p53: M-C-N-S-S-C-[MV]-G-G-M-N-R-R
- Enzimas
  - Sitio activo de L-lactato deshidrogenasa: [LIVMA]-G-[EQ]-H-G-[DN]-[ST]
  - Marca de enzima activador de ubiqüitina: P-[LIVM]-C-T-[LIVM]-[KRH]-x-[FT]-P

Expresiones regulares

- Ventajas
  - Rapidez de implementación.
  - Sencillez de diseño e interpretación.
- Desventajas
  - No modela bien las inserciones y deleciones.
  - Patrones cortos detectan muchos falsos positivos.
  - Poca capacidad de predicción.
  - Sin puntuación: solo resultados binarios.
- Utilidad
  - Detección de marcas pequeñas y sitios activos.
  - Comunicación escrita.

Position Specific Scoring Matrices (PSSM)

Frecuencias en cada posición

```
GHEGVGKVVKLGAGA
GHEKKGYFEDRGPSA
GHEGYGGRSRGGGYS
GHEFEGPKGCGALYI
GHELRGTTFMPALEC
```

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15
A	0	0	0	0	0	0	0	0	0	0	0	2	1	0	2
C	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
D	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
E	0	0	5	0	1	0	0	0	1	0	0	0	0	1	0
F	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0
G	5	0	0	2	0	5	1	0	1	0	2	3	1	1	0
H	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
K	0	0	0	1	1	0	1	1	0	1	0	0	0	0	0
L	0	0	0	1	0	0	0	0	0	0	1	0	2	0	0
M	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
P	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0
R	0	0	0	0	1	0	0	1	0	1	1	0	0	0	0
S	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1
T	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
V	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0
Y	0	0	0	0	1	0	1	0	0	0	0	0	0	2	0

Cambiamos 0 por un valor pequeño

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15
A	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.12	0.08	0.04	0.12
C	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.08	0.04	0.04	0.04	0.04	0.08
D	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.08	0.04	0.04	0.04	0.04	0.04
E	0.04	0.04	0.24	0.04	0.08	0.04	0.04	0.04	0.08	0.04	0.04	0.04	0.04	0.08	0.04
F	0.04	0.04	0.04	0.08	0.04	0.04	0.04	0.08	0.08	0.04	0.04	0.04	0.04	0.04	0.04
G	0.24	0.04	0.04	0.12	0.04	0.24	0.08	0.04	0.08	0.04	0.12	0.16	0.08	0.08	0.04
H	0.04	0.24	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04
I	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.08
K	0.04	0.04	0.04	0.08	0.08	0.04	0.08	0.08	0.04	0.08	0.04	0.04	0.04	0.04	0.04
L	0.04	0.04	0.04	0.08	0.04	0.04	0.04	0.04	0.04	0.04	0.08	0.04	0.12	0.04	0.04
M	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.08	0.04	0.04	0.04	0.04	0.04
P	0.04	0.04	0.04	0.04	0.04	0.04	0.08	0.04	0.04	0.04	0.08	0.04	0.08	0.04	0.04
R	0.04	0.04	0.04	0.04	0.08	0.04	0.04	0.08	0.04	0.08	0.08	0.04	0.04	0.04	0.04
S	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.08	0.04	0.04	0.04	0.04	0.08	0.08
T	0.04	0.04	0.04	0.04	0.04	0.04	0.08	0.08	0.04	0.04	0.04	0.04	0.04	0.04	0.04
V	0.04	0.04	0.04	0.04	0.08	0.04	0.04	0.08	0.08	0.04	0.04	0.04	0.04	0.04	0.04
Y	0.04	0.04	0.04	0.04	0.08	0.04	0.08	0.04	0.04	0.04	0.04	0.04	0.04	0.12	0.04

La puntuación

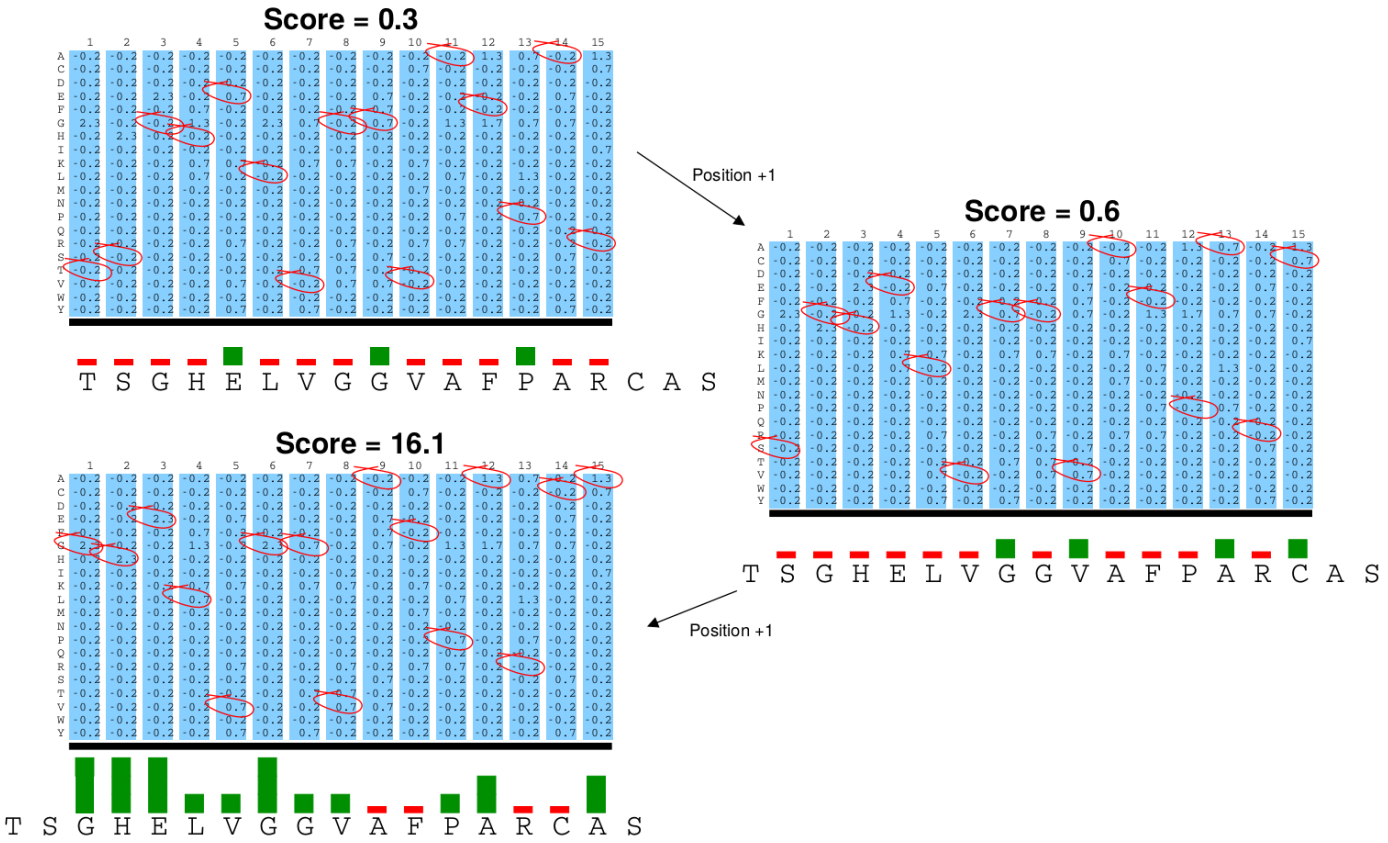
$$S = \log \frac{f'_{ij}}{q_i}$$

La puntuación de un aminoácido *i* en cada posición *j* depende de cuanto más frecuente es ese aminoácido en esa posición del alineamiento (f'_{ij}) de lo que lo seria en una secuencia al azar (q_i).

PSSM

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15
A	-0.32	-0.32	-0.32	-0.32	-0.32	-0.32	-0.32	-0.32	-0.32	-0.32	-0.32	1.26	0.68	-0.32	1.26
C	-0.32	-0.32	-0.32	-0.32	-0.32	-0.32	-0.32	-0.32	-0.32	0.68	-0.32	-0.32	-0.32	-0.32	0.68
D	-0.32	-0.32	-0.32	-0.32	-0.32	-0.32	-0.32	-0.32	-0.32	0.68	-0.32	-0.32	-0.32	-0.32	-0.32
E	-0.32	-0.32	2.26	-0.32	0.68	-0.32	-0.32	-0.32	0.68	-0.32	-0.32	-0.32	-0.32	0.68	-0.32
F	-0.32	-0.32	-0.32	0.68	-0.32	-0.32	-0.32	0.68	0.68	-0.32	-0.32	-0.32	-0.32	-0.32	-0.32
G	2.26	-0.32	-0.32	1.26	-0.32	2.26	0.68	-0.32	0.68	-0.32	1.26	1.68	0.68	0.68	-0.32
H	-0.32	2.26	-0.32	-0.32	-0.32	-0.32	-0.32	-0.32	-0.32	-0.32	-0.32	-0.32	-0.32	-0.32	-0.32
I	-0.32	-0.32	-0.32	-0.32	-0.32	-0.32	-0.32	-0.32	-0.32	-0.32	-0.32	-0.32	-0.32	-0.32	0.68
K	-0.32	-0.32	-0.32	0.68	0.68	-0.32	0.68	0.68	-0.32	0.68	-0.32	-0.32	-0.32	-0.32	-0.32
L	-0.32	-0.32	-0.32	0.68	-0.32	-0.32	-0.32	-0.32	-0.32	-0.32	0.68	-0.32	1.26	-0.32	-0.32
M	-0.32	-0.32	-0.32	-0.32	-0.32	-0.32	-0.32	-0.32	-0.32	0.68	-0.32	-0.32	-0.32	-0.32	-0.32
P	-0.32	-0.32	-0.32	-0.32	-0.32	-0.32	0.68	-0.32	-0.32	-0.32	0.68	-0.32	0.68	-0.32	-0.32
R	-0.32	-0.32	-0.32	-0.32	0.68	-0.32	-0.32	0.68	-0.32	0.68	0.68	-0.32	-0.32	-0.32	-0.32
S	-0.32	-0.32	-0.32	-0.32	-0.32	-0.32	-0.32	-0.32	0.68	-0.32	-0.32	-0.32	-0.32	0.68	0.68
T	-0.32	-0.32	-0.32	-0.32	-0.32	-0.32	0.68	0.68	-0.32	-0.32	-0.32	-0.32	-0.32	-0.32	-0.32
V	-0.32	-0.32	-0.32	-0.32	0.68	-0.32	-0.32	0.68	0.68	-0.32	-0.32	-0.32	-0.32	-0.32	-0.32
Y	-0.32	-0.32	-0.32	-0.32	0.68	-0.32	0.68	-0.32	-0.32	-0.32	-0.32	-0.32	-0.32	1.26	-0.32

Funcionamiento de la PSSM

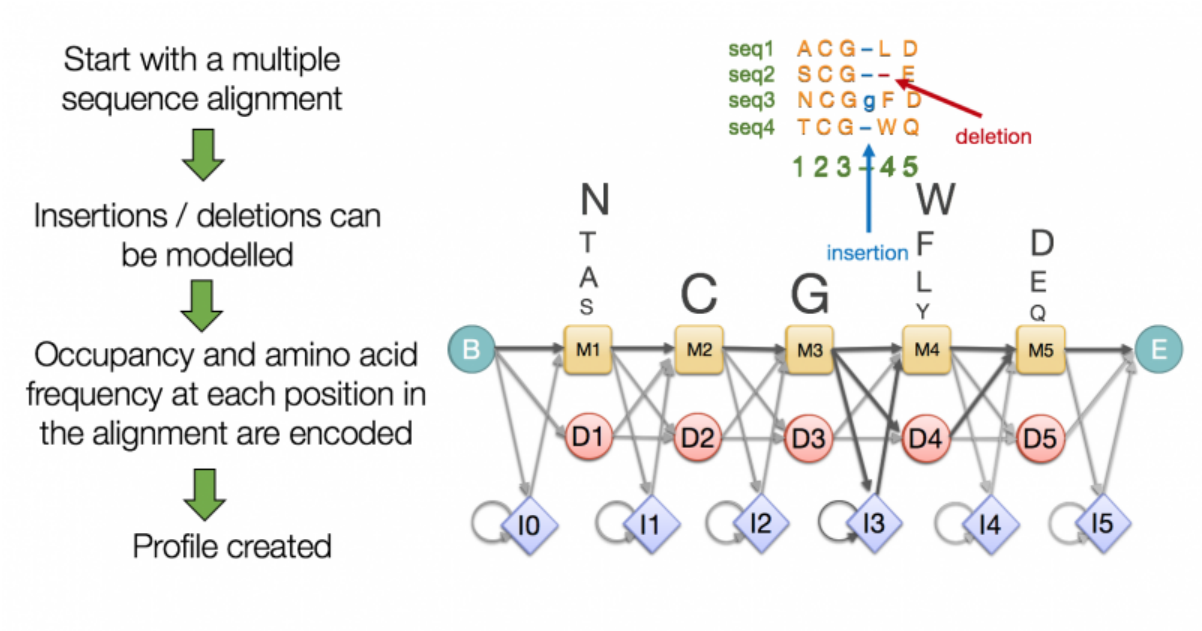


Balance de las PSSM

- Ventajas
  - Adecuadas para regiones cortas y conservadas.
  - Rapidez y facilidad de implementación.
  - Puntuaciones interpretables y valor E.
- Limitaciones
  - Inserciones y deleciones no permitidas.
  - Eso las hace inadecuadas para regiones largas.

Hiden Markov Models (HMM)

Ejemplo de HMM sencillo



PSI-BLAST

Principios del PSI-BLAST

1. BLAST estándard inicial, con BLOSUM62, e.g.
2. Construcción automática de una PSSM que permite indels.
3. La PSSM reemplaza la BLOSUM62 en un segundo BLAST.
4. Se repiten los pasos 2 y 3 incorporando nuevas secuencias al PSSM cada vez.
5. Cuando ya no se encuentran nuevas secuencias, el PSI-BLAST ha convergido.

Dos valores *E*

1. El umbral de valor *E* para el BLAST inicial (opción -*e*, por defecto, 10).
2. El valor *E* de inclusión, para aceptar nuevas secuencias en PSSM (opción -*h*, por defecto 0.001).

Ventajas del PSI-BLAST

- Rápido y eficiente.
- Mucho más sensible que el BLAST.
- Uso sencillo.

Peligros del PSI-BLAST

- Si sólo acepta secuencias muy parecidas, puede quedarse corto.
- Si acepta demasiada divergencia, puede incluir secuencias no homólogas (falsos positivos) e incluso no convergir.
- Requiere revisión manual: difícil de automatizar.

Otros BLASTs

Delta BLAST

- *Domain Enhanced Lookup Time Accelerated BLAST*
- Busca la proteína en una base de datos de PSSMs primero.
- Las PSSMs provienen de un subconjunto de la base de datos de dominios del NCBI (CDD).
- Utiliza la(s) PSSM(s) significativas en la búsqueda de más proteínas.
- Es más rápido que PSI-BLAST y se puede automatizar.
- No ayudaría mucho a identificar dominios nuevos. ¿Para qué puede servir?

Delta BLAST. Ejemplos

> "To identify potential *Ostreococcus* transporter proteins, mammalian sequences for all classes of SLC and all known magnesium transporters were blasted onto the Ostreococcus proteome using DELTA-BLAST (National Center for Biotechnology Information), and gene models were then taken from the latest version of the Ostreococcus genome using the Orcae service (Ghent University)."

Feeney et al. 2016. Daily magnesium fluxes regulate cellular timekeeping and energy balance. *Nature* 532: 375-379.

> "ORFs were predicted […]. To identify more distant homologues, ORFs were queried by DeltaBLAST against the Caudivirales subset of Genbank (NCBI taxonomy identifiers 28883 and 102294) with an e-value cutoff of 0.01."

Dutilh et al. 2014. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nature Communications* 5: 4498.