Basic Local Alignment Search Tool (BLAST)

Actualizado en: 24/01/2023

Objetivos

- Comprender el funcionamiento de BLAST.
- Conocer los diferentes tipos de búsquedas posibles.
- Capacidad de utilizar la interfaz web de NCBI BLAST.

BLAST, el algoritmo

Contexto histórico

- Lipman & Pearson. 1985. Rapid and sensitive protein similarity searches. *Science* 227(4693):1435-1441.

“One of the most rigorous programs for comparing amino acid sequences, SEQHP (5), requires more than 8 hours to compare a 200-residue protein to the 500,000-residue NBRF (National Biomedical Research Foundation) protein library on the VAX 11/750 computer.”

- Altschul et al. 1990. Basic Local Alignment Search Tool. *J. Mol. Biol.* 215:403-410.

Para qué sirve

- Comparar dos secuencias y encontrar similitudes locales (como Smith-Waterman).
- Buscar secuencias parecidas a una *query* en una base de datos (*target*).
- Múltiples aplicaciones:
  - Localizar dominios proteicos en una secuencia.
  - Recopilar secuencias homólogas para crear una filogenia.
  - Mapear secuencias cortas en un genoma de referencia.
  - Identificar una especie.

Qué meritos tiene

- Rapidez.
- Sensibilidad.
- Estadístico.

Cómo funciona

1. Detecta e ignora regiones repetitivas o de *baja complejidad* de la *query*.
2. Hace una lista de palabras de *k* letras de la *query* (*k* = 11 para DNA):

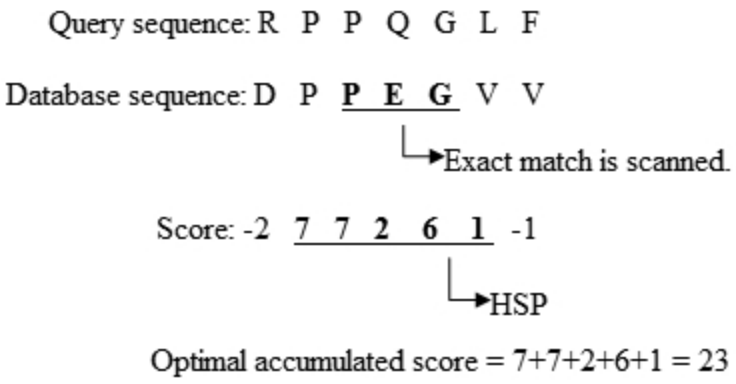
```
PQGEFG
PQG
 QGE
  GEF
   EFG
```

3. Añade a la lista palabras *vecinas* que alinearan con puntuación de al menos *T*.

Cómo funciona

4. Busca las palabras de la lista entre las secuencias de la base de datos (indexadas).

5. Alarga la *semilla* de los alineamientos encontrados (*High-scoring Segment Pair*, HSP).



Cómo funciona

6. Enumera HSPs con puntuación mayor de la que se produciría por azar.

7. Evalúa la significación de los HSPs.

8. Combina dos o más HSP en uno.

9. Muestra el alineamiento local Smith-Waterman de cada resultado.

10. Enumera los resultados con valor *E* menor o igual a un cierto umbral.

Evaluación estadística de los resultados

La distribución de puntuaciones de HSPs entre dos secuencias de longitudes *m* y *n* está descrita por los parámetros *K* y λ. El número esperado de HSPs con una puntuación de al menos *S* (**valor E**) es:

$$E = Kmne^{-\lambda S}$$

En una búsqueda en una base de datos, *n* es la longitud total de la base de datos entera. Los parámetros *K* y λ deben ser estimados mediante permutaciones. La probabilidad de observar al menos un HSP con una puntuación de al menos *S* por casualidad, es (distribución de Poisson):

$$P = 1 - e^{-E}$$

Este es el valor *p*.

Test

- El mismo HSP, en bases de datos de tamaños diferentes, ¿dónde tendrá un valor *E* mayor?
- ¿Cómo afectará el tamaño de palabra, *k*, a la sensibilidad?¿Y al tiempo de ejecución?
- ¿Para qué sirve conocer la distribución teórica de puntuaciones de HSPs?

BLAST, los programas

Principales programas

Programa	Query	Base.de.datos
blastn	DNA	DNA
blastp	proteína	proteína
blastx	DNA	proteína
tblastn	proteína	DNA
tblastx	DNA	DNA

PSI-BLAST

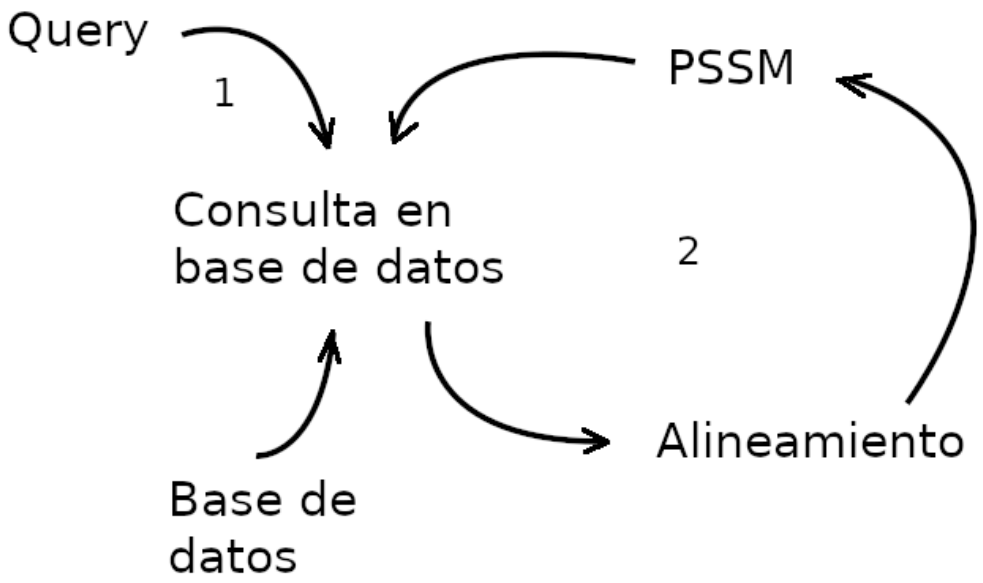
GAGGTAAAC
TCCGTAAGT
CAGGTTGGA
ACAGTCAGT
TAGGTCATT
TAGGTACTG
ATGGTAACT
CAGGTATAC
TGTGTGAGT
AAGGTAAGT

PSSM

A	0.25	1.23	-1.24	-4.70	-4.70	1.23	1.45	-0.31	-1.24
C	-0.31	-0.31	-1.24	-4.70	-4.70	-0.31	-1.24	-1.24	-0.31
G	-1.24	-1.24	1.45	1.96	-4.70	-1.24	-1.24	0.97	-1.24
T	0.66	-1.24	-1.24	-4.70	1.96	-1.24	-1.24	-0.31	1.23

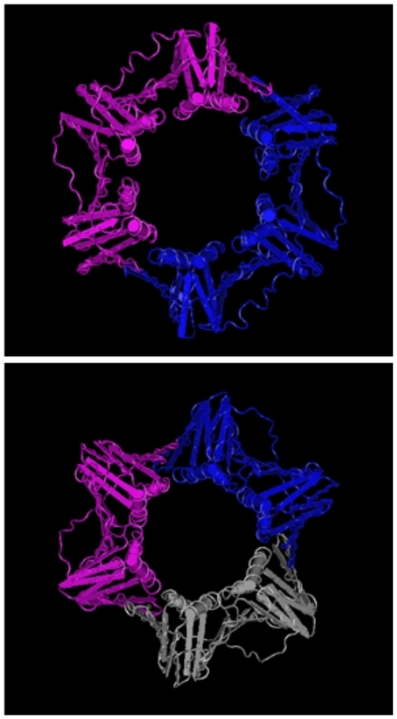
Ejemplo de *Position Specific Scoring Matrix*

PSI-BLAST



El PSI-BLAST empieza como un BLASTP (1). A partir del alineamiento de las secuencias homólogas que superan el umbral de valor E, genera una PSSM (2). Y utiliza la PSSM como nueva consulta (query) para añadir secuencias homólogas al alineamiento y repetir el ciclo.

PSI-BLAST



(Humana Press, 2007)

El PSI-BLAST es capaz de detectar la homología entre las secuencias de la subunidad β de la DNA polimerasa III de *E. coli* (arriba, número de acceso NP_002583) y la proteína humana PCNA (abajo, NP_002583), de estructura y función similares, pero muy divergente en la secuencia aminoacídica.

BLAST en la web

NCBI BLAST

EMBL-EBI BLAST

Existe tamién un paquete de programas `blast` para línea de comandos que utilizaremos en prácticas.