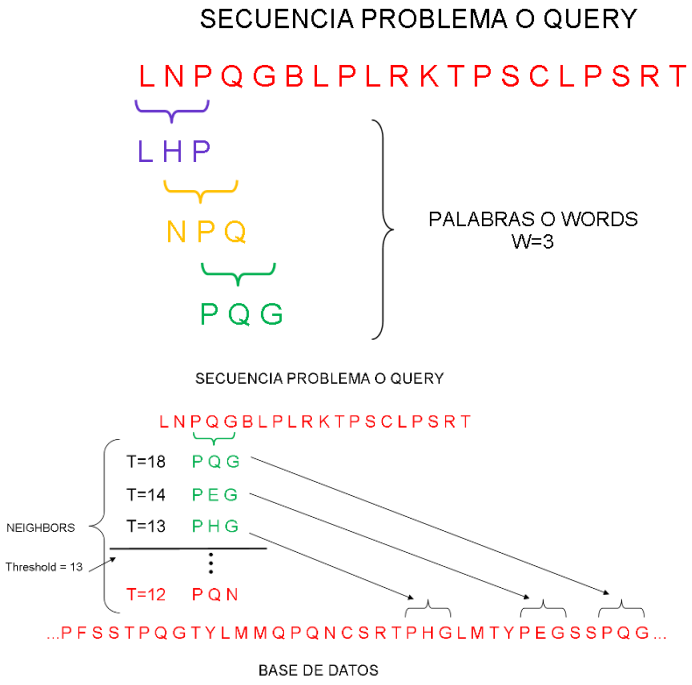
ALGORITMO BLAST

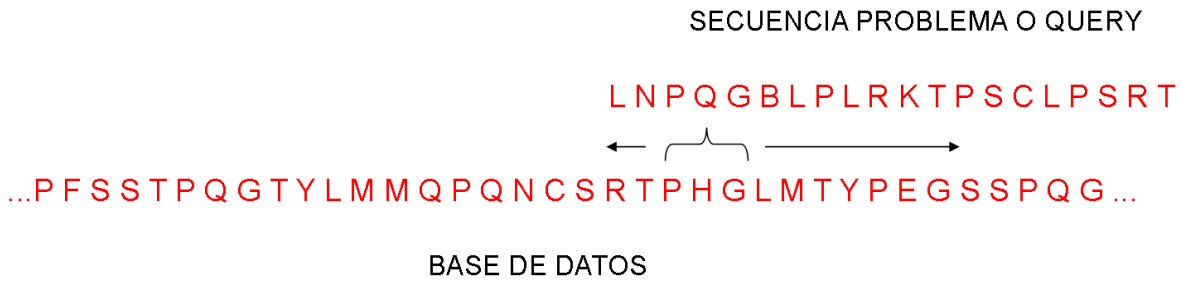
Realizar una búsqueda exacta necesitaría una cantidad de comparaciones excesiva desde el punto de vista computacional. Blast realiza alineamientos locales entre la secuencia problema y la base de datos, pero utiliza un algoritmo heurístico, o sea, que no nos garantiza un resultado óptimo pero a cambio nos permite realizar el alineamiento con buenos resultados. Blast se basa en la presunción de que buenos alineamientos contienen cortas regiones con perfectas (o muy buenas) coincidencias. Este algoritmo consta de tres fases:



1) ENSEMILLADO O SEEDING: Se divide la secuencia problema en pequeños fragmentos (palabras o words). En el caso de secuencias de ADN suelen ser de 11 nucleótidos (w=11) y el caso de proteínas suelen ser de 3 aminoácidos (w=3), aunque este parámetro puede ser ajustable. Posteriormente, a partir de cada una de estas palabras, se crean listas de palabras similares (Neighbors) hasta cierto valor humbral T (Threshold), y tomando como referencia una matriz de puntuación (por ejemplo match=2, mismatch=-3 gap=-5 para nucleótidos y Blosum62 para aminoácidos). Las palabras de la lista que estén por encima de dicho valor son marcadas en las secuencias de la base de datos (semillas o seeds) y las que estén por debajo de dicho umbral no son tenidas en cuenta por el programa.

2) **EXTENSIÓN:** A partir de cada una de las "semillas" el programa va extendiendo el alineamiento entre la base de datos y la secuencia problema en ambas direcciones. La extensión se realiza utilizando el algoritmo de Smith-Waterman y siguiendo de nuevo la matriz de puntuación. El alineamiento se detiene cuando se alcanza un valor X por debajo del valor máximo anterior. En ese momento el programa regresa al punto del valor máximo, que sera el "score" de dicho alineamiento. Esta es la clave para que al algoritmo sea viable,

ya que no estamos seguros de que se creen todos los alineamientos posibles, debido a que el programa podría detenerse en un máximo local. Los alineamientos que obtengan un score igual o superior al previamente seleccionado serán seleccionados y reciben el nombre de HSP (High-scoring Segment Pair) y el que obtiene la máxima puntuación es el MSP (Maximal Segment Pair).



3) EVALUACIÓN: Una vez terminado el alineamiento y calculada la puntuación final de cada resultado obtenido, el programa realiza una significación estadística, o sea, un estudio de la probabilidad de que cada uno de estos resultados haya sido obtenido al azar. Como resultado de este estudio el programa nos muestra un valor E (e-value) para cada uno de los alineamientos obtenidos.

El valor Esperado (E) es un parámetro que indica el número de coincidencias que se pueden "esperar" por casualidad, con un score igual o mejor al obtenido, al buscar en una base de datos de tamaño similar. Por ejemplo, un valor de E = 1 significa que en una base de datos del tamaño actual uno esperaría ver 1 coincidencia con un puntaje igual o mejor simplemente por azar. Cuanto menor sea el valor E, o cuanto más cerca esté de cero, más "significativa" será la coincidencia.

0 = E-value < 10e-100	SECUENCIAS IDÉNTICAS
10e-100 < E-value < 10e-50	SECUENCIAS CASI IDÉNTICAS
10e-50 < E-value < 10e-10	SECUENCIAS ESTRECHAMENTE RELACIONADAS
10e-10 < E-value < 1	PODRÍA EXISTIR HOMOLOGÍA
E-value => 1	EL ALINEAMIENTO MUY PROBABLEMENTE SEA CASUAL

TIPOS DE BLAST

En función del tipo de secuencia problema (ADN o proteína) y del tipo de base datos con la que hagamos la búsqueda, podemos encontrar que existen diferentes tipos de BLAST:

- **BLASTN**

A partir de una secuencia de NUCLEÓTIDOS realiza una búsqueda en una base de datos de NUCLEÓTIDOS.

APLICACIONES:

- Localizar oligonucleótidos, cDNA, EST, productos de PCR o elementos repetitivos en un genoma.
- Identificación de secuencias de DNA y anotación del DNA genómico.
- Localizar secuencias homólogas en especies distintas.
- Generación de contigs a partir de las lecturas más cortas obtenidas durante el proceso de secuenciación.
- Eliminar subsecuencias pertenecientes a vectores.
- Detección de contaminaciones.

VARIANTES:

- **MEGABLAST:** diseñado para identificar una secuencia problema (el parecido es del 100%) o para encontrar secuencias muy parecidas (> 95% de residuos idénticos). Es muy rápido porque utiliza un tamaño de palabra (el parámetro w) de 28 residuos.
- **Blastn**: Es más sensible que el anterior porque utiliza por defecto un parámetro w = 11, pero es más lento. Está diseñado para encontrar secuencias similares en organismos distintos. Si es preciso, también puede buscar con w = 7, aumentando la sensibilidad pero reduciendo notablemente la velocidad.
- **MEGABLAST discontinuo:** también está diseñado para encontrar secuencias similares en organismos distintos. Utiliza w = 11 y, en estas mismas condiciones, es más sensible y eficaz que blastn porque ignora algunas bases (la tercera de cada codón) y porque al buscar las palabras de la secuencia problemaen las BD no es necesario que ambas sean idénticas, sino que permite la presencia de discontinuidades.

- **BLASTP**

A partir de una secuencia de PROTEÍNA realiza un búsqueda en una base de datos de PROTEÍNAS.

APLICACIONES:

- Identificar una secuencia problema: en este caso, el parecido es del 100% y el programa genera un alineamiento global. Para que la identificación sea inequívoca puede ser una buena idea desactivar el filtro de las regiones de poca complejidad (low complexity filter)
- Encontrar secuencias parecidas en una BD de secuencias proteicas. Si el parecido es grande, puede tratarse de proteínas homólogas y es bastante probable que las anotaciones de las secuencias homólogas también sean válidas para la secuencia problema. BLAST permite reunir una colección de secuencias homólogas procedentes de distintos organismos para hacer alineamientos múltiples de secuencias o análisis filogenéticos.
- Localizar regiones de similitud: en este caso el parecido se limita a una región de las secuencias y el programa genera alineamientos locales que pueden corresponder a dominios conservados.

VARIANTES:

- **PSI-BLAST** utiliza los resultados de BLASTP para construir una matriz de puntuación específica de la posición (PSSM) y, a continuación, localizar secuencias con un parentesco remoto. Si una búsqueda con BLASTP no ha conseguido encontrar proteínas similares o si muchos de los resultados son dudosos, podemos utilizar PSI-BLAST. Este programa es el más sensible de todos y es muy útil a la hora de encontrar proteínas con parentesco remoto, identificar nuevos miembros de una familia de proteínas, o descubrir proteínas con secuencias muy divergentes pero con una estructura tridimensional parecida.

- **BLASTX**

A partir de una secuencia de NUCLEÓTIDOS realiza un búsqueda en una base de datos de PROTEÍNAS. El programa traduce la secuencia de nucleótidos en sus seis posibles marcos de lectura (tres marcos de lecturas por hebra) y compara estas secuencias traducidas contra una base de datos de proteínas. BLASTX tiene utilidad cuando sospechamos que nuestra secuencia ADN puede codificar para alguna proteína.

APLICACIONES:

- Localizar genes que codifican proteínas en el DNA genómico.
- Determinar si un transcrito (convertido en cDNA o en EST) codifica alguna proteína conocida.
- Definir las regiones codificantes y no codificantes de un mRNA.

- **TBLASTN**

A partir de una secuencia de PROTEÍNA se realiza una búaqueda con una base de datos de NUCLEÓTIDOS. Para realizar esto traduce todas las secuencias de nucleótidos en sus seis marcos de lectura. Se suele utilizar cuando el análisis con Blastp no ha dado resultados positivos.

APLICACIONES:

- Localizar una proteína en el DNA genómico, lo que permite ver si existen elementos reguladores cerca de la región codificante del gen y localizar exones.
- Buscar en BD de EST los transcritos que correspondan a la secuencia problema o a una secuencia parecida.

- **TBLASTX**

A partir de una secuencia de NUCLEÓTIDOS realiza un búsqueda en una base de datos de NUCLEÓTIDOS, pero a diferencia de Blastn, TBLASTX compara las traducciones de seis marcos de lectura de la secuencia de consulta de nucleótidos con las traducciones de seis marcos de lectura de la base de datos de secuencias de nucleótidos.

APLICACIONES:

- Detectar nuevos genes en secuencias genómicas (de la misma especie o de especies distintas), especialmente los que resultan difíciles de encontrar por los métodos tradicionales (genes dentro de otros genes, procesamientos alternativos o genes con bajos niveles de expresión).
- Descubrir transcritos (en forma de cDNA o EST) cuyos productos aún no están incluidos en las BD.

OTRAS CONSIDERACIONES

Podemos realizar un blast para encontrar secuencias de nucleótidos similares a una secuencia problema desconocida, pero lo más común es trabajar con secuencias de aminoácidos. De este modo podemos identificar proteínas, encontrar proteínas homólogas, seleccionar proteínas para realizar MSA o identificar regiones o dominios conservados en proteínas de diversas especies.

Cuando dos secuencias se parecen es muy probable que sean homólogas, esto quiere decir que evolutivamente hablando tienen un ancestro común y por tanto su estructura puede ser muy parecida, y también su función. Esto nos permite conocer muchas cosas de nuestra secuencia recién descubierta con solo compararla con el resto de secuencias de las que ya se tiene mucha información. En el caso de las proteínas, se considera que por encima del 25% de similitud es muy probable que exista homología (en el caso de los acidos nucléicos debe superar el 70%), siempre que comparemos secuencias de al menos 100 residuos. Pero en realidad la existencia o no de homología no la podemos saber con certeza, es posible que proteínas con un 15% de identidad en sus residuos de aminoácidos posean una misma estructura y función.

A esa zona alrededor del 25% se le llama “twilight zone” o zona de penumbra. Los valores alrededor de la zona de penumbra deben de ser tomados con cautela y observar otros parámetros para confirmar o rechazar la existencia de homología. Por ejemplo secuencias problema cortas en bases de datos extensas es más probable que nos den coincidencias elevadas. Un parámetro muy interesante y que nos permite valorar nuestra búsqueda teniendo en cuenta estos factores es el E-Value (Expected value) que veremos con más detalle en el siguiente apartado del tema.