BLAST en el servidor del NCBI

BLAST es la **herramienta bioinformática más utilizada** en todo el mundo. **Compara una secuencia problema** (*query sequence*) de nucleótidos o de proteínas c**on todas las secuencias de una BD** para encontrar regiones de similitud local. Además, calcula la significación estadística de los resultados. BLAST **permite inferir relaciones** funcionales, estructurales o evolutivas entre dos secuencias de modo que se pueden **identificar nuevos miembros de una familia** de genes o de proteínas.

Por regla general, las búsquedas con BLAST obedecen a uno de estos dos objetivos:

- **Localizar una secuencia dentro de otra**, como cuando se quiere encontrar la ubicación de un gen, oligo, cDNA o EST en un genoma o cuando se quiere determinar la estructura de un gen (localizar los intrones, los exones y las regiones reguladoras)
- **Explorar las BD en busca de secuencias relacionadas** funcional o evolutivamente

Hay **dos formas de utilizar BLAST**:

- **conectados a Internet** (*on-line*). Además de la página del NCBI (**http://blast.ncbi.nlm.nih.gov/Blast.cgi**) hay muchas otras páginas que ofrecen esta herramienta. En estas condiciones, corremos el peligro de que alguien pueda acceder a los resultados de nuestra búsqueda.
- **sin conexión a Internet**. Cuando interesa mantener la confidencialidad de los resultados, lo mejor es descargar el programa e instalarlo en un ordenador personal. El programa se puede descargar gratuitamente desde la dirección: **ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/**, las BD se pueden descargar desde: **ftp://ftp.ncbi.nlm.nih.gov/blast/db/** y las instrucciones de instalación y los comandos de uso se pueden encontrar en la dirección **http://www.ncbi.nlm.nih.gov/books/NBK1762/**.

Para utilizar BLAST en el servidor del NCBI hay que:

- 1.- Ir a la dirección **http://blast.ncbi.nlm.nih.gov/Blast.cgi**
- 2.- Seleccionar la variante del programa que se va a utilizar
- 3.- Introducir la secuencia problema. Se puede poner directamente un código de acceso, un código *gi* (*gene identifier*), o una secuencia en formato FASTA
- 4.- Seleccionar la base de datos
- 5.- Ajustar diversos parámetros de la búsqueda (*E-value*, *word size*, *scoring parameters: substitution matrix*, *gap penalties*)
- 6.- ¡BLAST!

Versiones del programa BLAST (NCBI)

Existen diversas versiones del programa BLAST. Es importante **saber cuál es la que mejor se adapta a los objetivos de la búsqueda**. Para ello, hay que tener en cuenta 3 factores: (1) la naturaleza de la secuencia problema, (2) el objetivo de la búsqueda y (3) la BD donde se va a llevar a cabo la búsqueda.

En la siguiente Tabla se muestran las distintas versiones del programa BLAST:

Table 1 BLAST programs

Program	Query sequence type	Target sequence type	
BLASTN	Nucleotide	Nucleotide	Compares a nucleotide query sequence against a nucleotide sequence database
BLASTP	Protein	Protein	Compares an amino acid query sequence against a protein sequence database
BLASTX	Nucleotide (translated)	Protein	Compares a nucleotide query sequence translated in all reading frames against a protein sequence database
TBLASTN	Protein	Nucleotide (translated)	Compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames
TBLASTX	Nucleotide (translated)	Nucleotide (translated)	Compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database

Blastn

Compara una secuencia de nucleótidos con una BD que también contiene secuencias de nucleótidos. Se utiliza para:

- **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 (genes de RNA o de proteínas, regiones reguladoras, etc.)
- **Generación de *contigs*** a partir de las lecturas más cortas obtenidas durante el proceso de secuenciación
- Eliminar subsecuencias pertenecientes a **vectores**
- Detección de **contaminaciones**

Este tipo de búsqueda **no es el más apropiado para encontrar regiones que codifican proteínas homólogas en otros organismos**. En este caso es mejor hacer búsquedas a nivel de proteína directamente con blastp o traducir la secuencia problema, la BD, o ambas, según las seis pautas de lectura posibles. Son tres las razones que explican esta circunstancia: (1) la degeneración del código genético, (2) las secuencias proteicas albergan más información que las secuencias de nucleótidos y (3) las matrices de sustitución utilizadas para el alineamiento de secuencias de proteínas son más sofisticadas que las utilizadas para alinear secuencias de nucleótidos.

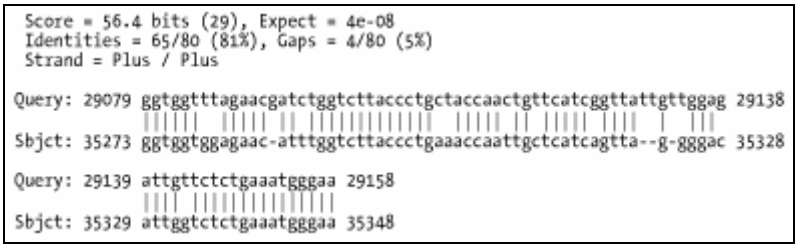
Dentro del programa blastn se pueden seleccionar **varios algoritmos**:

- **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 problema

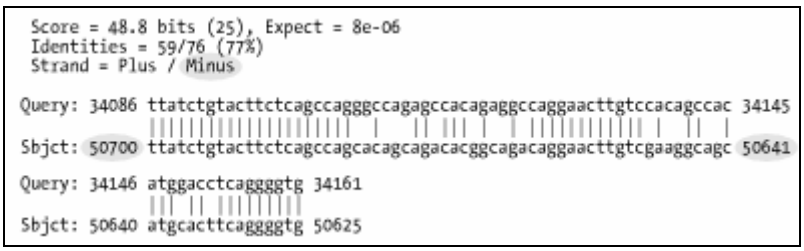
en las BD no es necesario que ambas sean idénticas, sino que permite la presencia de discontinuidades.

Cuando se introduce una secuencia problema para hacer búsquedas en una BD, **BLASTN utiliza las dos hebras de la molécula de DNA**. La **hebra plus** es la que se ha introducido en formato FASTA y la complementaria inversa es la **hebra minus**.

Cuando la región de similitud local está situada en la misma hebra, al representar el alineamiento las dos secuencias (la secuencia problema y la secuencia de la BD) están señaladas como "plus" y las coordenadas de las dos secuencias avanzan en sentido creciente.



Cuando la hebra minus de la secuencia problema es similar a una secuencia de la BD, el alineamiento representa la hebra plus de la secuencia problema y la hebra minus de la BD. Por tanto, las coordenadas de la secuencia de la BD están colocadas en sentido decreciente.



Blastp

Este programa compara una secuencia de aminoácidos con una BD que también contiene secuencias proteicas. Se utiliza para:

- **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 a**lineamientos 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**.

Dentro del programa blastp se pueden seleccionar **varios algoritmos**:

- **Blastp** compara una secuencia proteica con una BD de proteínas
- **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
- **PHI-BLAST** busca proteínas que contienen un patrón especificado por el usuario y que, además del patrón, presentan otras regiones de similitud con la secuencia problema
- **DELTA-BLAST** construye una PSSM basándose en una búsqueda en la BD de dominios conservados y, a continuación, hace una búsqueda en una BD de proteínas

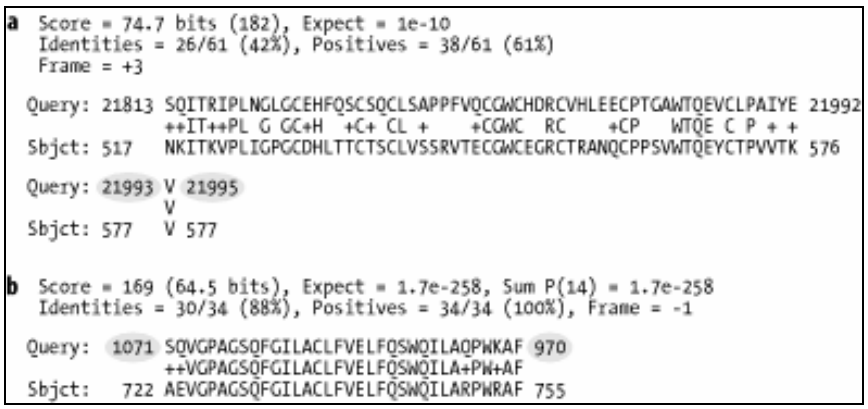
Blastx

La secuencia problema es una secuencia de nucleótidos. El programa traduce esta secuencia en sus seis posibles marcos de lectura (tres marcos de lecturas por hebra) y compara estas secuencias traducidas con una BD de proteínas. Es un programa lento que se usa cuando **se tiene sospecha de que la secuencia problema codifica una proteína** pero no se sabe exactamente cuál. Si la secuencia problema corresponde a una región no codificante del DNA, blastx no encontrará nada.

Se utiliza para:

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

A la hora de interpretar los alineamientos generados por blastx hay que tener en cuenta **la hebra, la pauta de lectura y las coordenadas**. En la hebra plus, las pautas de lectura (*frame*) se denominan +1, +2 y +3. En la hebra minus, las pautas de lectura se denominan –1, –2 y –3. Las coordenadas de la secuencia problema aumentan de tres en tres (parte **a** de la figura inferior) porque cada aminoácido corresponde a tres nucleótidos. Si el alineamiento se produce en la hebra minus de la secuencia problema (parte **b** de la figura inferior, *frame* = –1), las coordenadas de la secuencia problema aparecen en orden descendente.



TBlastn

Compara una secuencia proteica con una BD de nucléotidos. Para ello, primero traduce todas las secuencias de nucleótidos de la BD en sus seis marcos de lectura y luego realiza la comparación. TBLASTN es un programa lento que **se usa cuando el análisis con Blastp no ha tenido éxito porque la proteína no aparece en las BD**. Sin embargo, es posible que las BD de EST o de proyectos genómicos en curso (que carecen de anotaciones) incluyan alguna secuencia que pueda corresponder al transcrito que codifica esa proteína o una similar. Se utiliza para:

- **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

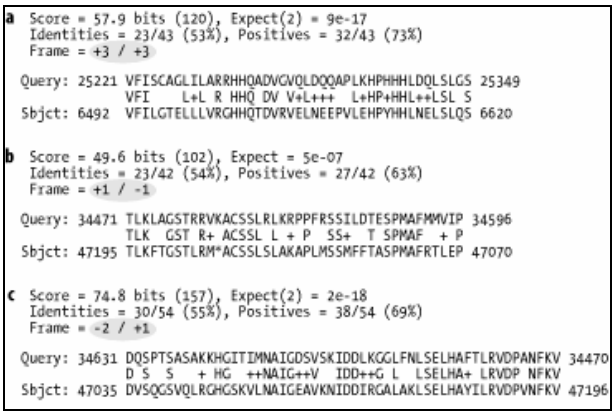
Hay que tener cuidado con los resultados obtenidos con esta variante de Blast, porque una buena cantidad de las secuencias traducidas no son proteínas que existan en la naturaleza.

Tblastx

Compara una secuencia de nucleótidos con una BD de nucleótidos, pero primero traduce la secuencia problema y las secuencias de las BD en los seis marcos de lectura posibles. Se aprovecha del hecho de que las secuencias codificantes evolucionan más lentamente que el DNA adyacente. Se trata de una búsqueda más sensible que BLASTP, pero requiere mucho esfuerzo computacional y sólo debería utilizarse como último recurso y, preferentemente, sin conexión a Internet. Se utiliza para:

- **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

Los alineamientos generados por tblastx son difíciles de interpretar porque hay que tener en cuenta **la hebra, la pauta de lectura y las coordenadas**, tanto en la secuencia problema como en la secuencia de la BD.

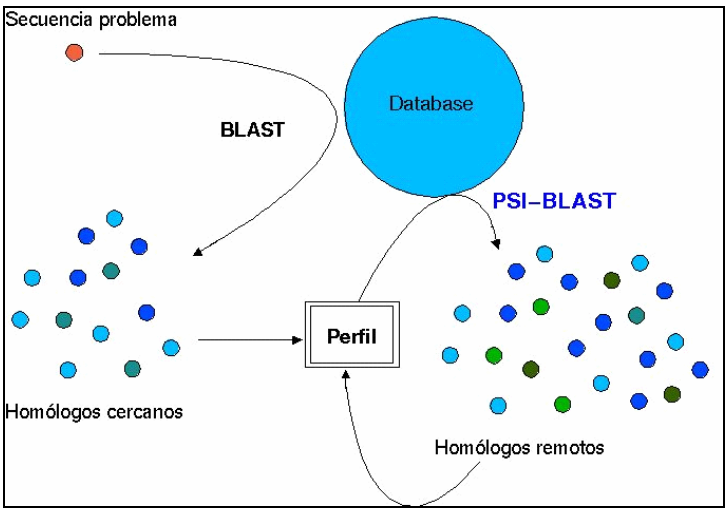


PSI-BLAST (PSI = *Position Specific Iterative*)

Si una búsqueda con BLASTP no ha conseguido encontrar proteínas similares o si muchos de los resultados son dudosos ("*hypothetical protein*", "*predicted*" o "*similar to...*"), podemos utilizar PSI-BLAST. Este programa es **el más sensible** de todos y es muy útil a la hora de (1) **encontrar proteínas con parentesco remoto**, (2) **identificar nuevos miembros de una familia** de proteínas, o (3) **descubrir proteínas con secuencias muy divergentes pero con una estructura tridimensional parecida**.

PSI-BLAST se ejecuta en varias etapas:

- La primera etapa consiste en una búsqueda **BLASTP normal** utilizando una matriz de sustitución como BLOSUM62
- En la segunda etapa, **se seleccionan las secuencias** con un valor E menor que cierto umbral (por defecto E = 0,005, pero se puede cambiar) y se hace un **alineamiento múltiple** (AMS) con el que se construye un **perfil**, también denominado una matriz de puntuación específica de la posición (**PSSM**, *position-specific scoring matrix*). Esta PSSM **asigna una puntuación distinta a cada posición del AMS**: a los residuos conservados en una determinada posición se les asigna una puntuación muy alta, mientras que, en esa misma posición, a los demás residuos se les asigna una puntuación muy negativa. En las regiones no conservadas se asigna una puntuación cercana a cero a todos los residuos.
- En la tercera etapa se lleva a cabo una **nueva búsqueda** en la BD utilizando la PSSM en lugar de la matriz de sustitución. Las **nuevas secuencias** con el valor E apropiado se añaden a las seleccionadas en la segunda etapa para hacer una **nueva PSSM** con la que iniciar una **nueva búsqueda**. Esta etapa es **iterativa**, es decir, se puede repetir las veces que uno quiera o hasta llegar a la **convergencia**, que es el momento en el que la búsqueda no consigue encontrar nuevas secuencias.



El resultado de la búsqueda depende en gran medida de las secuencias seleccionadas en la segunda etapa para construir la PSSM. **El usuario puede modificar la selección de secuencias** utilizadas para construir la PSSM, eliminando algunas que el programa haya incluido de forma automática o introduciendo otras descartadas por el programa.