Capítulo 1 Búsqueda de similitud usando BLAST

Kit J. Menlove, Mark Clement y Keith A. Crandall

31 de agosto de 2018

Resumen

Las búsquedas de similitud son un componente esencial de la mayoría de las aplicaciones bioinformáticas. Estas forman la base para la identificación de motivos estructurales, identificación de genes e ideas sobre asociaciones funcionales. Con el rápido aumento en los datos genéticos disponibles a través de una amplia variedad de bases de datos, las búsquedas de similitud son herramientas esenciales para acceder a estos datos de una manera informativa y productiva. En este capítulo, ofrecemos una descripción general de enfoques de búsqueda de similitud, bases de datos relacionadas y opciones de parámetros para lograr los mejores resultados para un variedad de aplicaciones. Al final, proporcionamos un ejemplo trabajado y algunas notas para su consideración.

Palabras clave: BLAST, alineamiento de secuencia, búsqueda de similitud

1. INTRODUCCIÓN

1.1. Una introducción a Bases de datos de Nucleótidos

Quizás el objetivo central de la genética es articular las asociaciones de fenotipos de interés con sus componentes genéticos subyacentes y luego comprender la relación entre variación genética y variación en el fenotipo. Este objetivo ha sido impulsado por el tremendo aumento en nuestra capacidad para obtener datos genéticos celulares, a través de poblaciones y especies. A medida que surgieron los métodos de recopilación de información sobre diversos aspectos de las macromoléculas biológicas, la información biológica se hizo abundante y la necesidad de consolidar y hacer accesible esta información se hizo cada vez más evidente. A principios de la década de 1960, Margaret Dayhoff y sus colegas en la Fundación Nacional de Investigación Biomédica (NBRF) comenzaron a recopilar información sobre secuencias de proteínas y estructuras en un volumen titulado *Atlas of Protein Sequence and Estructure* (1). Desde ese comienzo, las bases de datos han sido una parte importante y esencial de la investigación biológica y bioquímica.

En 1972, el tamaño del Atlas se había vuelto difícil de manejar, por lo que la Dra. Dayhoff–una pionera de la bioinformática–desarrolló una infraestructura de base de datos en la que se podía canalizar esta información. Aunque la información de nucleótidos se incluyó en el Atlas ya en 1966 (2), su volumen estaba compuesto por secuencias de aminoácidos con anotación estructural.

1.2. Colaboración internacional de bases de datos de secuencias de nucleótidos : DDBJ, EMBL y GenBank

No fue hasta 1982 que las bases de datos se desarrollaron con el expreso propósito de almacenar secuencias de nucleótidos por la *European Molecular Biology Laboratory* (EMBL: `http://www.embl.org/`) en Europa y los Institutos Nacionales de Salud (NIH - NCBI: `http://www.ncbi.nlm.nih.gov/`) en América del Norte. Japón siguió su ejemplo con la creación de su banco de datos de ADN (DDBJ: `http://www.ddbj.nig.ac.jp/`) en 1986.

Una considerable cantidad de intercambio ocurrió de forma natural entre estas tres bases de datos y la *Base de Datos de Secuencia del Genoma*, también de América del Norte, una condición que llevó a su coalición en 1988 bajo el título *International Nucleotide Sequence Database Collaboration* (INSDC). Siguen siendo entidades muy distintas, pero en la reunión de 1988 establecieron políticas para regular el formato y la administración de las secuencias que cada uno recibe. Sus políticas actuales incluyen el acceso sin restricciones y el uso de todos los registros de datos, la citación adecuada de los creadores de datos y las responsabilidades de los remitentes para verificar la validez de los datos y su derecho a enviarlos.

El INSDC actualmente contiene aproximadamente 80 mil millones de pares de bases (*bp*) (sin incluir las secuencias de *shotgun* de genoma completo) y casi 80 millones de entradas de secuencia. Ahora, incluyendo secuencias de *shotgun* (HTGS), el 22 de agosto de 2005 se superó la marca de 100 gigabases y para septiembre 2007 ya contenía aproximadamente 200 mil millones de *bp*. Durante más de 10 años, la cantidad de datos en estas bases de datos se duplicó aproximadamente cada 18 meses. Esta expansión ha comenzado a nivelare ya que nuestra capacidad de secuenciación de alto rendimiento está alcanzando gradualmente un máximo. Se espera que el próximo redoblamiento de los datos ocurra en aproximadamente 4 años (Fig. 1.1)

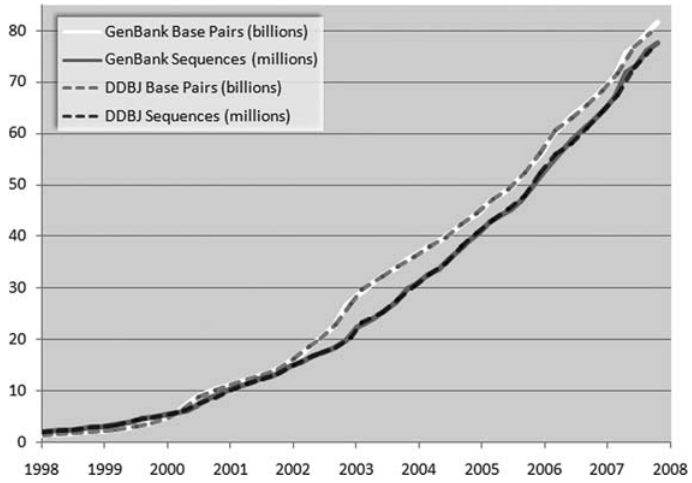


Fig. 1.1. Crecimiento de las bases de datos genéticas de GenBank y DDBJ en los últimos 10 años. Las bases de datos INSDC han crecido, en los últimos 10 años, aproximadamente 168 veces en total número de pares de bases. Mientras que en el pasado el número de entradas en las bases de datos INSDC duplicado aproximadamente cada 2 años, una regresión polinomial de segundo orden simple (R 2 = 0,9995) de los datos de los últimos 10 años indica que el próximo redoblamiento tomar un poco más de 4 años. Este gráfico no incluye datos de HTG

Desde que las primeras bases de datos de nucleótidos fueron iniciadas por EMBL y NIH (ahora en manos de NCBI), se han formado muchas bases de datos de ADN para atender las necesidades de grupos de investigación especializados. La edición 2007 de la base de datos de *Nucleic Acid Research* contenía 109 bases de datos de secuencias de nucleótidos que cumplen con los estándares requeridos para ser incluidos en su listado (3). Estas bases de datos suelen desarrollarse para incluir datos auxiliares asociados con los datos genéticos, como paciente o información de la muestra, incluida información clínica, imágenes, análisis posteriores. Muchos no cumplen con los estándares de calidad, cantidad y originalidad de los datos, así como la calidad de la interfaz web que deben considerarse para el problema (4). Incluso, muchos de ellos son de propiedad privada y solo permiten el acceso de datos costosos a unos pocos seleccionados.

Con todo, la cantidad de bases de datos de ADN es asombrosa y aumentan constantemente a medida que encontramos nuevas y poderosas formas de reunir, almacenar y utilizar las piezas que componen el rompecabezas de la vida.

2. USO DE PROGRAMAS

2.1. BASES DE DATOS Y FORMATOS

Una de las principales fuentes de diversidad entre las bases de datos de ADN reside en sus formatos de archivo. Si bien se han hecho grandes esfuerzos para estandarizar los formatos de archivo, los diversos tipos y propósitos de información de secuencia y anotación envían tipos de archivos personalizados.

2.1.1. Formato FASTA

Utilizado por primera vez con el programa FASTA de Pearson y Lipman para la comparación de secuencias (5), el formato de archivo FASTA es el más simple de los formatos ampliamente utilizados disponibles a través del INSDC. Se compone de una línea de definición o descripción seguida de la secuencia. La línea de definición comienza con un símbolo mayor que (y) y marca el comienzo de cada nueva entrada. La información que sigue al símbolo de mayor que varía varía según su fuente.

```
>gi|186681228|ref|YP_001864424.1| phycoerythrobilin:ferredoxin oxidoreductase
MNSERSDVTLYQPFLDYAIAYMRSRLDLEPYPIPTGFESNSAVVGKGKNQEEVVTTSYAFQTAKLRQIRA
AHVQGGNSLQVLNFVIFPHLNYDLPFFGADLVTLPGGHLIALDMQPLFRDDSAYQAKYTEPILPIFHAHQ
QHLSWGGDFPEEAQPFFSPAFLWTRPQETAVVETQVFAAFKDYLKAYLDFVEQAEAVTDSQNLVAIKQAQ
LRYLRYRAEKDPARGMFKRFYGAEWTEEYIHGFLFDLERKLTVVK
```

Figura 1: Ejemplo Archivo Secuencia en formato FASTA

En general, sigue un identificador (Tabla 1.1), después de lo cual se pueden incluir palabras de descripción opcional. Si la secuencia se recupera a través de las bases de datos del NCBI, un número GI precede al identificador. Aunque se recomienda que la línea de definición no tenga más de 80 caracteres, a menudo se incluyen varios tipos y niveles de información. La línea de definición es seguida por la secuencia de ADN en sí misma, en formato de línea única o múltiple. Los nucleótidos están representados por sus códigos estándar IUB / IUPAC, incluidos los códigos de ambigüedad (Tabla 1.2).

Database name	Identifier syntax
GenBank	gb|*accession.version*
EMBL	emb|*accession.version*
DDBJ	dbj|*accession.version*
NCBI RefSeq	ref|*accession.version*
PDB	pdb|*entry*|*chain*
Patents	pat|*country*|*number*
NBRF PIR	pir||*entry*
SWISS-PROT	sp|*accession*|*entry*
Protein Research Foundation	prf|*name*
GenInfo Backbone Id	bbs|*number*
General database identifier	gnl|*database*|*identifier*
Local Sequence identifier	lcl|*identifier*

Table 1.1. FASTA Identificadores de secuencia de archivos. Información del NCBI Manual (25)

A	adenosine	M	A or C (amino)	V	A, C, or G
C	cytidine	K	G or T (keto)	H	A, C, or T
G	guanine	R	A or G (purine)	D	A, G, or T
T	thymidine	Y	C or T (pyrimidine)	B	C, G, or T
U	uridine	S	A or T (strong)	–	Gap of indeterminate length
		W	C or G (weak)	N	A, C, G, or T (any or unknown)

Tabla 1.2 Códigos de ambigüedad y nucleótidos IUB / IUPAC

2.1.2. Formato de archivo plano

GenBank, EMBL y DDBJ tienen cada uno su propio formato de archivo plano, pero contienen básicamente la misma información. Todos están basados en la Tabla de funciones, que se puede encontrar en `http://www.ncbi.nlm.nih.gov/collab/FT`. Para referencias a estos tipos de archivos, vea (6-9).

```
LOCUS       SCU49845     5028 bp    DNA             PLN       21-JUN-1999
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
            Saccharomycetales; Saccharomycetaceae; Saccharomyces.
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
  TITLE     Cloning and sequence of REV7, a gene whose function is required for
            DNA damage-induced mutagenesis in Saccharomyces cerevisiae
  JOURNAL   Yeast 10 (11), 1503-1509 (1994)
  PUBMED    7871890
REFERENCE   2  (bases 1 to 5028)
  AUTHORS   Roemer,T., Madden,K., Chang,J. and Snyder,M.
  TITLE     Selection of axial growth sites in yeast requires Axl2p, a novel
            plasma membrane glycoprotein
  JOURNAL   Genes Dev. 10 (7), 777-793 (1996)
  PUBMED    8846915
REFERENCE   3  (bases 1 to 5028)
  AUTHORS   Roemer,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (22-FEB-1996) Terry Roemer, Biology, Yale University, New
            Haven, CT, USA
FEATURES             Location/Qualifiers
     source          1..5028
                     /organism="Saccharomyces cerevisiae"
                     /db_xref="taxon:4932"
                     /chromosome="IX"
                     /map="9"
     CDS             <1..206
                     /codon_start=3
                     /product="TCP1-beta"
                     /protein_id="AAA98665.1"
                     /db_xref="GI:1293614"
                     /translation="SSIYNGISTSGLDLNNGTIADMRQLGIVESYKLKRAVVSSASEA
                     AEVLLRVDNIIRARPRTANRQHM"
     gene            687..3158
                     /gene="AXL2"
     CDS             687..3158
                     /gene="AXL2"
                     /note="plasma membrane glycoprotein"
                     /codon_start=1
                     /function="required for axial budding pattern of S.
                     cerevisiae"
                     /product="Axl2p"
                     /protein_id="AAA98666.1"
                     /db_xref="GI:1293615"
                     /translation="MTQLQISLLLTATISLLHLVVATPYEAYPIGKQYPPVARVNESF
                     TFQISNDTYKSSVDKTAQITYNCFDLPSWLSFDSSSRTFSGEPSSDLLSDANTTLYFN
                     VILEGTDSADSTSLNNTYQFVVTNRPSISLSSDFNLLALLKNYGYTNGKNALKLDPNE
                     VFNVTFDRSMFTNEESIVSYYGRSQLYNAPLPNWLFFDSGELKFTGTAPVINSAIAPE
                     TSYSFVIIATDIEGFSAVEVEFELVIGAHQLTTSIQNSLIINVTDTGNVSYDLPLNYV
                     YLDDDPISSDKLGSINLLDAPDWVALDNATISGSVPDELLGKNSNPANFSVSIYDTYG
```

Figura 2: Ejemplo Archivo Secuencia en formato GenBank

2.1.3. Números de acceso, números de versión, nombres de locus, identificadores de bases de datos, etc.

El estándar para identificar un registro de secuencia de nucleótidos es mediante un sistema *accession.version* donde el número de acceso es un identificador de dos letras seguido por seis dígitos y la versión es un número incremental que indica el número de cambios que se han realizado a la secuencia desde que fue presentada por primera vez.

Los nombres de los *locus* (ver Nota 1) son identificadores más antiguos, menos estandarizados, cuyo propósito original era agrupar las entradas con secuencias similares (10). El formato del *locus* original estaba destinado a contener información sobre el organismo y otras características comunes del grupo (como el producto del gen). Ese formato de diez caracteres ya no es capaz de contener esa información para la gran cantidad y variedad de secuencias ahora disponibles, por lo que el *locus* se ha convertido en otro identificador único a menudo configurado para tener el mismo valor que el número de acceso. Los identificadores de bases de datos son simplemente cadenas de dos o tres caracteres que sirven para indicar qué base de datos recibió y almacenó originalmente la información. El identificador de la base de datos es el primer valor enumerado en la sintaxis del identificador FASTA (Tabla 1.1).

Cuando una secuencia se envía por primera vez a GenBank, se envía con varias características definidas

asociadas con la secuencia. Algunos incluyen información de **CDS** (secuencia de codificación), **RBS** (sitio de unión a ribosoma), **rep_origin** (origen de replicación) y **ARNt** (ARN de transferencia madura). Se proporciona una traducción de secuencias de nucleótidos que codifican proteínas en aminoácidos como parte de la sección de características. Del mismo modo, el etiquetado de diferentes marcos de lectura abiertos, intrones, etc., es parte de la tabla de características. Puede encontrar una lista de características y sus descripciones, formatos y convenciones acordadas por INSDC en la Tabla de características (consulte la Sección 2.1.2).

2.2. SMITH-WATERMAN Y PROGRAMACIÓN DINÁMICA

En 1970, Needleman y Wunsch adaptaron la idea de programación dinámica al problema difícil de la alineación global de secuencias (11). En 1981, Smith y Waterman adaptaron este algoritmo a las alineaciones locales (12). Una alineación global intenta alinear dos secuencias en toda su longitud, mientras que una alineación local alinea regiones de dos secuencias donde se observa una gran similitud.

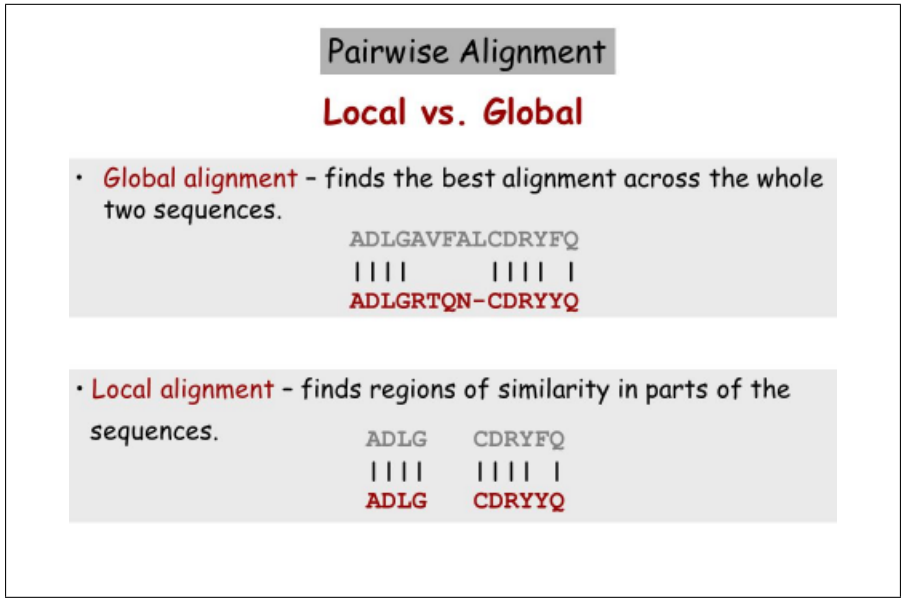


Figura 3: Ejemplo alineamiento local y global

Ambos métodos implican inicializar, puntuar y rastrear una matriz donde las filas y columnas corresponden a las bases o residuos de las dos secuencias alineadas (figura 1.2). En el caso de alineación local, la primera fila y la primera columna se rellenan con ceros. Las celdas restantes se rellenan con un valor de indicador recursivo derivado de valores vecinos:

$$max = \left\{ \begin{array}{c} 0 \\ \text{Vecino izquierdo + penalidad por gap} \\ \text{Vecino arriba + penalidad por gap} \\ \text{Vecino izquierdo-arriba +puntuación math/mistmatch} \end{array} \right\}$$

Si la celda actual corresponde a una coincidencia (bases idénticas o *match*), la puntuación del *match* se agrega al valor del vecino diagonal, de lo contrario se utiliza la puntuación de no-coincidencia o *mistmatch*. Los puntajes de penalización de hueco *(gap)* y *mistmatch* son generalmente cero o un número negativo pequeño, mientras que el puntaje de *match* es un número positivo, de mayor magnitud. Este método se usa recursivamente, comenzando desde la esquina superior izquierda de la matriz y avanzando hacia la esquina inferior derecha. La Figura 1.2b y c muestra matrices de dos conjuntos diferentes de puntajes de *gap* y *match*.

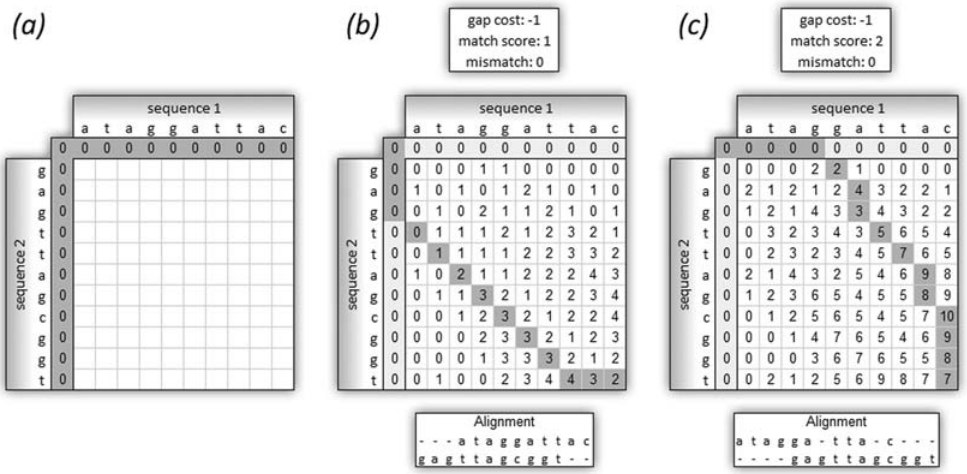


Fig. 1.2. Ejemplo de alineación local de Smith-Waterman. (A) muestra una matriz vacía, inicializada para una alineación de Smith-Waterman. (B) y (C) son alineamientos calculados utilizando los parámetros de puntuación especificados. La alineación producida en (B) es drásticamente diferente de la de (C), aunque solo difieren ligeramente en sus parámetros de puntuación, uno que usa un puntaje de coincidencia de 1 y el otro 2.

Para encontrar una alineación local, uno simplemente encuentra el número más grande en la matriz y traza una ruta de regreso hasta que se alcanza un cero, cada paso se mueve a una celda que fue responsable del valor de la celda actual.

Si bien este método es robusto y garantiza la mejor alineación para un conjunto dado de puntajes y penalizaciones, es importante tener en cuenta que a menudo son posibles múltiples trayectorias y, por lo tanto, múltiples alineamientos para cualquier matriz dada cuando estos parámetros son usados. Como ejemplo, b y c de la figura 1.2 solo difieren ligeramente en sus puntuaciones de *gap* y *match*, pero producen alineaciones muy diferentes. Además, el conjunto de puntajes y penalizaciones utilizados afecta drásticamente la alineación, y encontrar el conjunto óptimo no es ni trivial ni determinista. Las matrices de peso para las secuencias de codificación de proteínas se desarrollaron a fines de la década de 1970 en un intento de superar estos desafíos.

2.3. MODELOS DE PONDERACIÓN

Para aumentar la especificidad de los algoritmos de alineación y proporcionar un medio para evaluar su significación estadística, fue necesario implementar un esquema de puntuación significativo para las sustituciones de nucleótidos y aminoácidos. Esto es especialmente cierto cuando se trata de secuencias de proteínas (o codificación de proteínas). En 1978, Dayhoff et al. desarrollaron las primeras matrices de puntuación o ponderación creadas a partir de sustituciones que se han observado durante la historia evolutiva (13).

2.3.1. Matrices PAM

Estas sustituciones han sido permitidas o aceptadas de forma natural y se llaman *mutaciones puntuales aceptadas* (PAM). Para las matrices de PAM de Dayhoff, se analizaron grupos de proteínas con un 85 % o más de similitud de secuencia y se catalogaron sus 1,571 sustituciones. Cada celda de una matriz PAM corresponde a la frecuencia en sustituciones por 100 residuos entre dos aminoácidos dados. Esta frecuencia se conoce como una unidad PAM. En la década de 1970, cuando se crearon, sin embargo, había un número limitado y poca variedad de secuencias de proteínas disponibles, por lo que están sesgadas hacia las proteínas globulares pequeñas. También es importante tener en cuenta que cada matriz PAM corresponde a una distancia evolutiva específica y que cada una es simplemente una extrapolación de la original. Por ejemplo, una matriz PAM250 (Fig. 1.3) se construye multiplicando la matriz PAM1 por sí misma 250 veces y se ve como una matriz de puntuación típica para las proteínas que se han separado por 250 millones de años de evolución.

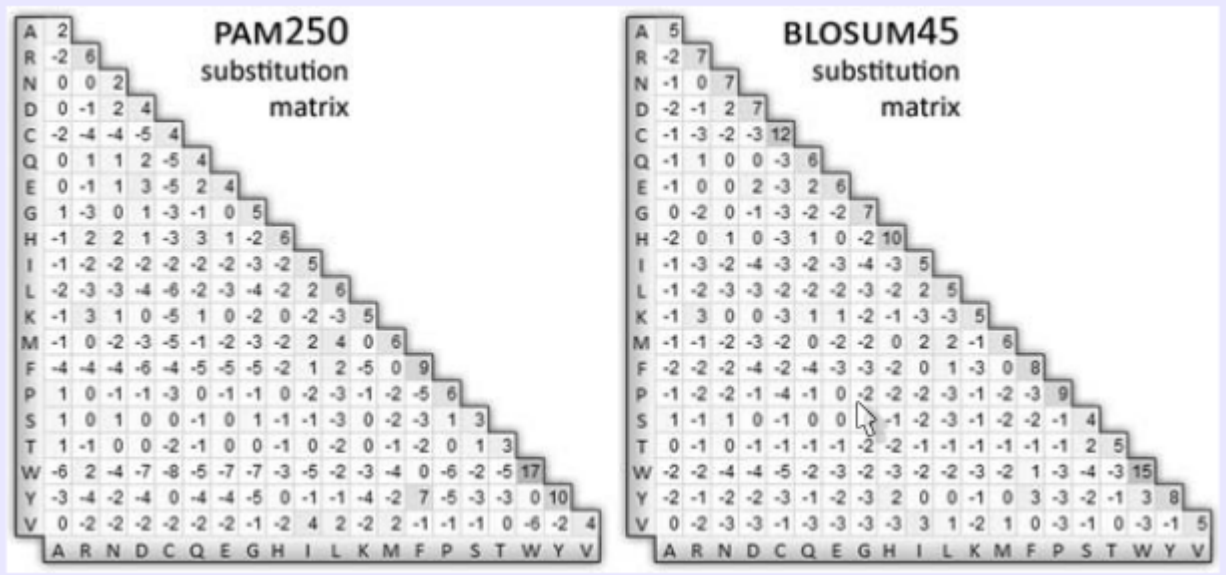


Fig. 1.3. Matrices de sustitución PAM250 y BLOSUM45.

2.3.2. Matrices Blosum

Para superar algunos de los inconvenientes de las matrices PAM, Henikoff y Henikoff desarrollaron las matrices BLOSUM en 1992 (14). Estas matrices se basaban en la base de datos BLOCKS, que organiza las proteínas en bloques, donde cada bloque, definido por una alineación de motivos, corresponde a una familia. Mientras que la matriz PAM original se calculó con proteínas con al menos un 85 % de identidad, las matrices BLOSUM se calculan cada una por separado utilizando motivos conservados a una distancia evolutiva específica o por debajo de ella. Esta diversidad de matrices, junto con el hecho de estar basado en conjuntos de datos más grandes, hace que las matrices BLOSUM sean más sólidas para detectar similitudes a mayores distancias evolutivas y más precisas, en muchos casos, en la realización de búsquedas de similitud local (15).

2.3.3. Selección de una matriz

Al elegir una matriz, es importante considerar las alternativas. No elija simplemente la configuración predeterminada sin una consideración inicial. En general, encontrar similitud al aumentar la divergencia corresponde al aumento de las matrices de PAM (PAM1, PAM40, PAM120, etc.) y la disminución de las matrices de BLOSUM (BLOSUM90, BLOSUM80, BLOSUM62, etc.) (16).

Las matrices PAM son fuertes para detectar similitudes altas debido a su uso de información evolutiva. Sin embargo, a medida que aumenta la distancia evolutiva, las matrices BLOSUM son más sensibles y precisas que sus contrapartes PAM. La Tabla 1.3 incluye una lista de usos sugeridos.

Alignment size	Best at detecting	Similarity (%)	PAM	BLOSUM
Short	Similarity within a **species**	75–90	**PAM30**	BLOSUM95
"	Similarity within a **genus**	60–75	**PAM70**	BLOSUM85
Medium	Similarity within a **family**	50–60	PAM120	**BLOSUM80**
"	The **largest range** of similarity	40–50	PAM160	**BLOSUM62**
Long	Similarity within a **class**	30–40	PAM250	**BLOSUM45**
"	Similarity within the **twilight zone**	20–30		BLOSUM30

Tabla 1.3 Usos sugeridos para matrices de sustitución comunes. Las matrices resaltadas en negrita están disponibles a través de la interfaz web BLAST de NCBI. Se ha demostrado que BLOSUM62 proporciona los mejores resultados en las búsquedas BLAST en general debido a su capacidad para detectar grandes rangos de similitud. Sin embargo, las otras matrices tienen sus puntos fuertes. Por ejemplo, si su objetivo es detectar secuencias de alta similitud para inferir homología dentro de una especie, las matrices PAM30, BLOSUM90 y PAM70 proporcionarían los mejores resultados. Esta tabla fue adaptada de los resultados obtenidos por David Wheeler (16)

2.4. PROGRAMAS BLAST

Las búsquedas de nucleótidos a nucleótidos son beneficiosas porque no se pierde información en la alineación. Cuando un codón se traduce de un nucleótido a un aminoácido, se pierde aproximadamente el 69 % de la complejidad (64 posibles combinaciones de nucleótidos asignadas a 20 aminoácidos). En contraste, sin embargo, la verdadera relación física entre dos secuencias de codificación se capta mejor en la vista traducida. Las matrices que tienen en cuenta las propiedades físicas, como PAM y BLOSUM, se pueden usar para agregar potencia a la búsqueda. Además, en una búsqueda de nucleótidos, solo hay cuatro posibles estados de caracteres en comparación con 20 en una búsqueda de aminoácidos. Por lo tanto, la probabilidad de una coincidencia debido a la casualidad frente a una coincidencia debido a ancestros comunes (identificar en estado versus idéntico por descendencia) es alta.

Las herramientas básicas de alineación y búsqueda local (BLAST) son las más utilizadas y entre las más precisas para detectar la similitud de secuencia (17) (ver Nota 2). Los programas BLAST estándar son Nucleotide BLAST (blastn), Protein BLAST (blastp), blastx, tblastn y tblastx. Otros también se han desarrollado para satisfacer necesidades específicas. Al elegir un programa BLAST, es importante elegir el correcto para su pregunta de interés. Algunos de los errores más comunes en la búsqueda de similitudes provienen de malentendidos de estas diferentes aplicaciones.

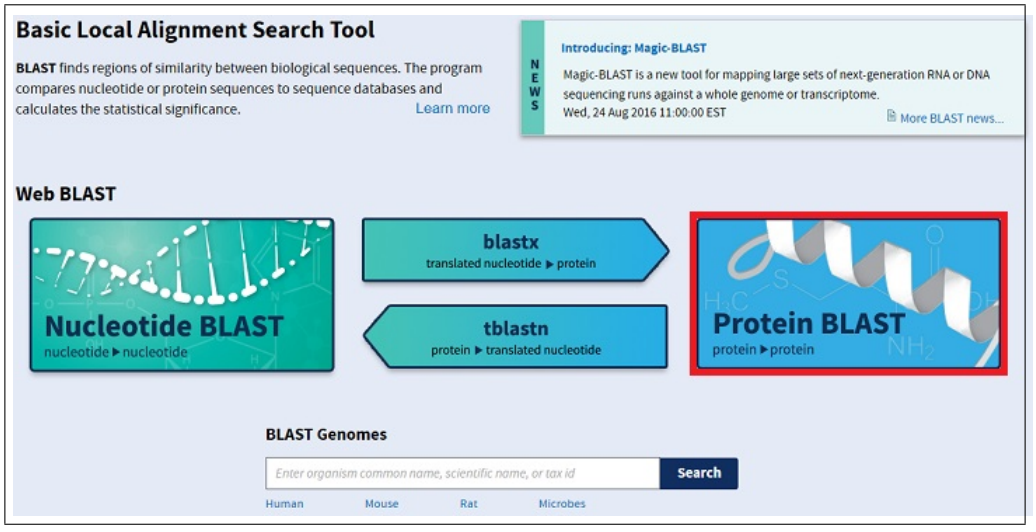


Figura 4: Entrada BLAST de NCBI para ejecución de los distintos tipos de BLAST

- **blastn: (Nucleotide BLAST):** compara una secuencia de consulta de nucleótidos *(query)* con una base de datos de secuencias de nucleótidos
- **blastp: (Protein BLAST):** compara una secuencia de consulta de proteínas *(query)* con una base de datos de secuencias de proteínas
- **blastx:** compara una secuencia de consulta de nucleótidos *(query)* traducida en los seis marcos de lectura con una base de datos de proteínas
- **tblastn**: compara una secuencia de consulta de proteínas *(query)* con una base de datos de secuencias de nucleótidos traducida dinámicamente en los seis marcos de lectura
- **tblastx**: compara una secuencia de consulta de nucleótidos *(query)* en los seis marcos de lectura con una base de datos de secuencias de nucleótidos en los seis marcos de lectura

El algoritmo BLAST es un programa heurístico, uno que no garantiza devolver el mejor resultado. Sin embargo, es bastante preciso. BLAST trabaja haciendo primero una tabla de búsqueda de todas las "palabras" y "palabras vecinas" de la secuencia de consulta (ver Figura 5 abajo). Las palabras son subsecuencias cortas de longitud W y las palabras vecinas son palabras que son altamente aceptadas en el sentido de la matriz de puntuación, determinado por un umbral T. La base de datos se escanea en busca de palabras y palabras vecinas. Una vez que se encuentra una coincidencia, las extensiones con y sin *gaps* se inician allí arriba y abajo. La extensión continúa, añadiendo la existencia de *gap* (iniciación) y las penalizaciones de extensión, y los puntajes de *match* y *mismatch*, según corresponda, como en el algoritmo de Smith-Waterman hasta que se alcanza un umbral de puntuación S. Alcanzar esta marca señala la secuencia de salida. La extensión continúa hasta que la puntuación caiga en un valor X desde el máximo, en cuyo punto la extensión se detiene y la alineación se recorta hasta el punto donde se alcanzó la puntuación máxima.

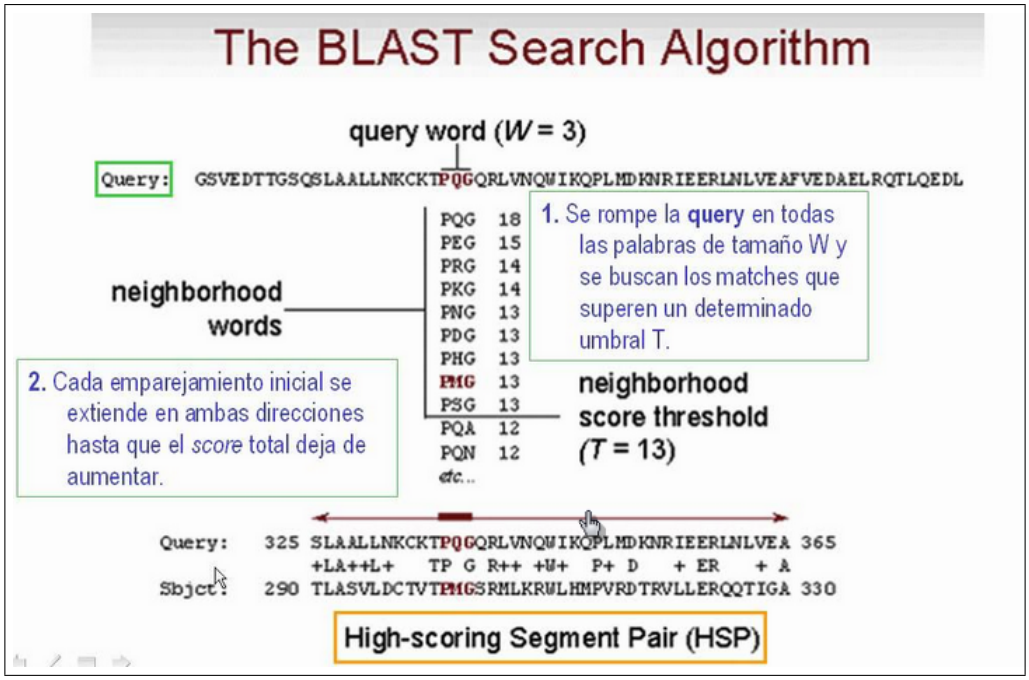


Figura 5: Pasos del Algortimo BLAST

Comprender este algoritmo es importante para los usuarios si desean seleccionar los parámetros óptimos para BLAST. La interacción entre los parámetros T, W, S, X y la matriz de puntuación permite al usuario encontrar un equilibrio entre sensibilidad y especificidad, alterar el tiempo de ejecución y modificar la precisión del algoritmo. Las interacciones entre estas variables se discutirán en la Sección 2.8.

2.5. SECUENCIA CONSULTA O SECUENCIA *Query*

Las secuencias de consulta se pueden ingresar cargando un archivo o ingresando uno manualmente en el cuadro de texto proporcionado (Fig. 1.4). La opción de carga acepta archivos que contienen una sola secuencia, secuencias múltiples en formato FASTA, o una lista de identificadores de secuencia válidos (números de acceso, números de GI, etc.). A diferencia de las versiones anteriores de BLAST en el sitio web de NCBI, la versión actual le permite al usuario especificar un título de trabajo descriptivo. Esto permite al usuario rastrear cualquier ajuste o versión de una búsqueda, así como su propósito e información de consulta. Esto es especialmente importante cuando los identificadores de secuencia no están incluidos en el archivo cargado.

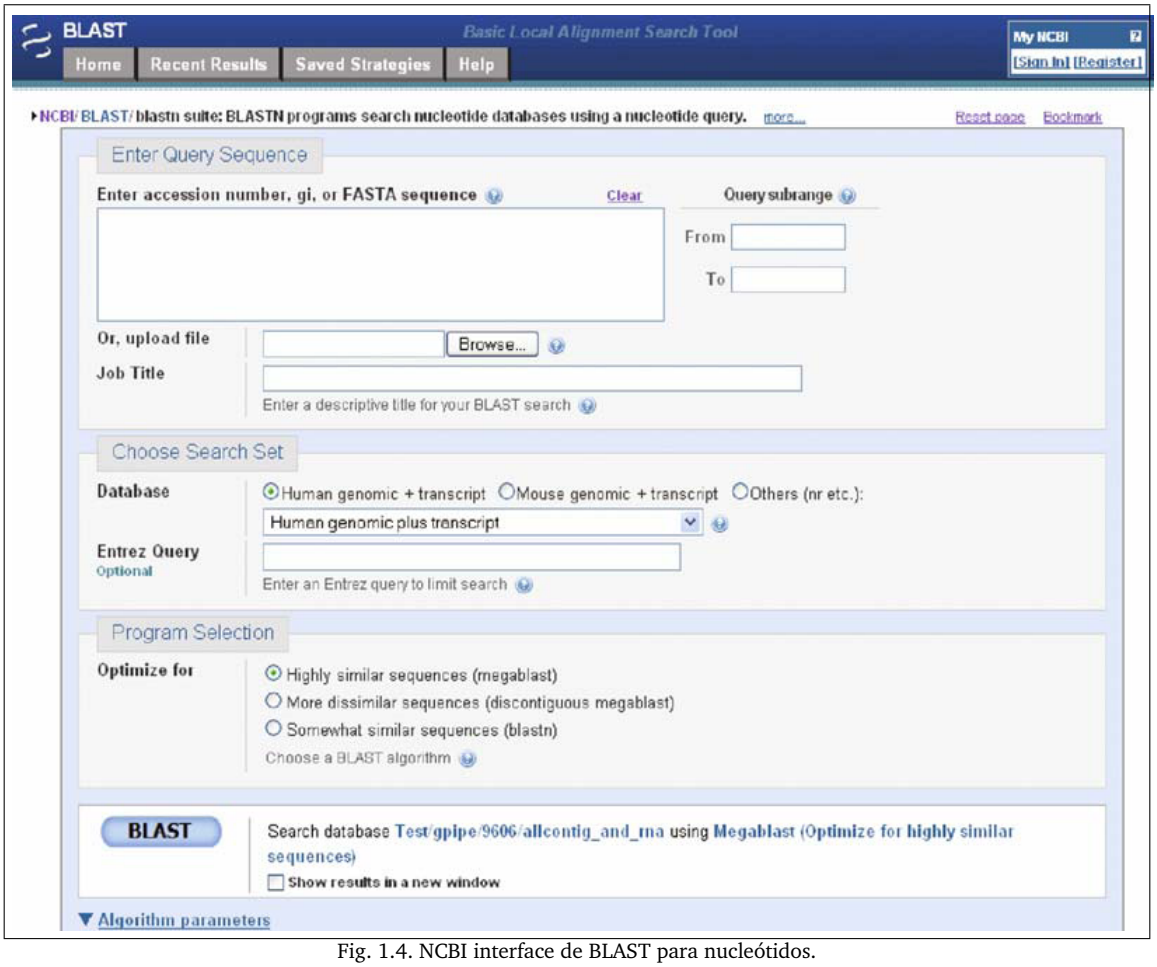


Fig. 1.4. NCBI interface de BLAST para nucleótidos.

2.6. CONJUNTOS DE BÚSQUEDA

Al elegir una base de datos, es importante comprender su propósito, contenido y limitaciones. La lista de bases de datos de nucleótidos está dividida en secciones de *Genomic plus Transcript y Other Databases*.

2.6.1. Bases de Datos

Algunas de las bases de datos, compuestas de secuencias de referencia, provienen de la base de datos *RefSeq*, un conjunto altamente curado, todo incluido, no redundante de INSDC (EMBL + GenBank + DDBJ), entradas de ADN, ARNm y proteínas. Las secuencias de RefSeq tienen números de acceso de la forma AA_######, donde AA es una de las siguientes combinaciones de letras (Tabla 1.4) y ###### es un número único que representa la secuencia.

Experimentally determined and curated		Genome annotation (computational predictions from DNA)	
NC	Complete genomic molecules		
NG	Incomplete genomic region		
NM	mRNA	XM	Model mRNA
NR	RNA (non-coding)		
NP	Protein	XP	Model protein

Table 1.4 Categorías RefSeq

Una descripción de las bases de datos de nucleótidos se incluye a continuación. Se puede encontrar una lista de bases de datos de proteínas accesibles a través de la interfaz web de BLAST en `http://www.ncbi.nlm.nih.gov/BLAST/blastcgihelp.shtml`.

- **Human genomic plus transcript:** contiene todas las secuencias genómicas y de ARN humanas.
- **Mouse genomic plus transcript:** contiene todas las secuencias genómicas y de ARN de ratón.
- **Colección de nucleótidos (nr / nt):** contiene secuencias de nucleótidos INSDC + RefSeq + secuencias PDB, sin incluir EST, STS, GSS o secuencias de HGT inacabadas. La colección de nucleótidos es el conjunto más completo de secuencias de nucleótidos disponibles a través de BLAST.
- **Secuencias de ARNm de referencia (refseq_rna):** contiene las secuencias de ARNm de RefSeq no redundantes.
- **Secuencias genómicas de referencia (refseq_genomic):** contiene las secuencias genómicas RefSeq no redundantes.
- **Secuencias de etiquetas expresadas (Expressed sequence tags -est):** contiene lecturas cortas y únicas de secuenciación de ARNm (a través de ADNc). Estas secuencias de ADNc representan el ARNm en una célula en un momento particular en un tejido particular.
- **EST no humanos, no ratones (est_others):** se eliminó la base de datos previa con secuencias humanas y de ratón.
- **Secuencias de estudio genómico (gss):** contiene secuencias genómicas aleatorias obtenidas de estudios de genoma de paso único, cósmidos, BAC, YAC y otros métodos de estudio. Su calidad varía.
- **Secuencias genómicas de alto rendimiento (HTGS):** contiene secuencias obtenidas de centros de genoma de alto rendimiento. Las secuencias en esta base de datos contienen un número de fase, 0 es la fase inicial y 3 es la fase final. Una vez terminado, las secuencias se mueven a la división apropiada en su respectiva base de datos.
- **Secuencias de patente (pat):** contiene secuencias de las oficinas de patentes de cada una de las organizaciones del INSDC.
- **Banco de datos de proteínas (pdb):** las secuencias de nucleótidos del Brookhaven Protein Data Bank gestionado por el Research Collaboratory for Structural Bioinformatics (http: // www.rcsb.org/pdb).
- **Elementos de repetición de ALU humanos (alu_repeats):** contiene un conjunto de elementos de repetición de ALU que se pueden usar para enmascarar elementos repetidos de secuencias de consulta. Las secuencias ALU son regiones sujetas a división por endonucleasas de restricción Alu, de aproximadamente 300 pb de longitud, y se estima que constituyen aproximadamente el 10 % del genoma humano (18).
- **Secuencias de sitios etiquetados (dbsts):** una colección de secuencias únicas utilizadas en la PCR y el mapeo del genoma que identifican una región particular de un genoma.

- **Lecturas de *shotgun* de genoma completo (wgs):** contiene secuencias de *shotgun* a gran escala, en su mayoría sin montar y sin anotaciones.
- **Muestras medioambientales (env_nt):** contiene conjuntos completos de lecturas de *shotgun* de muchos organismos muestreados, cada conjunto desde una ubicación particular de interés. Estos conjuntos permiten a los investigadores observar la diversidad genética existente en un lugar y entorno particular.

2.6.2. Organismo

El cuadro de organismo permite al usuario especificar un organismo particular para buscar. Sugiere automáticamente organismos cuando comienza a escribir. Esta opción no está disponible cuando se seleccionan las bases de datos Genomic plus Transcript (figura 1.5).

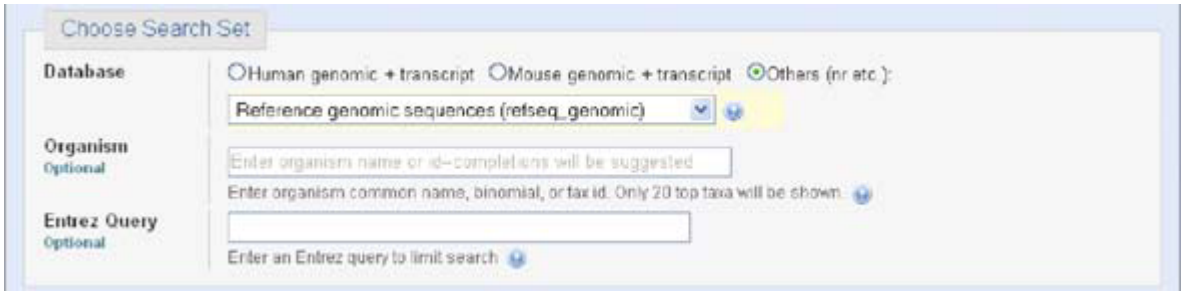


Fig. 1.5. Parámetros del algoritmo BLAST de nucleótidos NCBI.

2.6.3. Consultas Entrez

Las consultas de Entrez proporcionan una forma de limitar su búsqueda a un tipo específico de organismo o molécula. Es una forma eficiente de filtrar los resultados no deseados mediante la exclusión de organismos o la definición de criterios de longitud de secuencia. Además, las consultas de Entrez le permiten al usuario encontrar secuencias enviadas por un autor en particular, de una revista en particular, con una propiedad o clave de función particular, o enviadas o modificadas dentro de un rango de fechas específico. Para obtener ayuda con las consultas de Entrez, consulte el documento de ayuda de Entrez en `http://www.ncbi.nlm.nih.gov/entrez/query/static/help/helpdoc.html`.

2.7. PARÁMETROS DE BÚSQUEDA DE BLAST

Además de ingresar una secuencia de consulta, elegir un conjunto de búsqueda y seleccionar un programa, hay varios parámetros adicionales disponibles que le permiten ajustar su búsqueda a sus necesidades. Estos parámetros están disponibles haciendo clic en el enlace "Parámetros del algoritmo"en la parte inferior de la página BLAST (Fig. 1.6) (ver Notas 3 y 4).

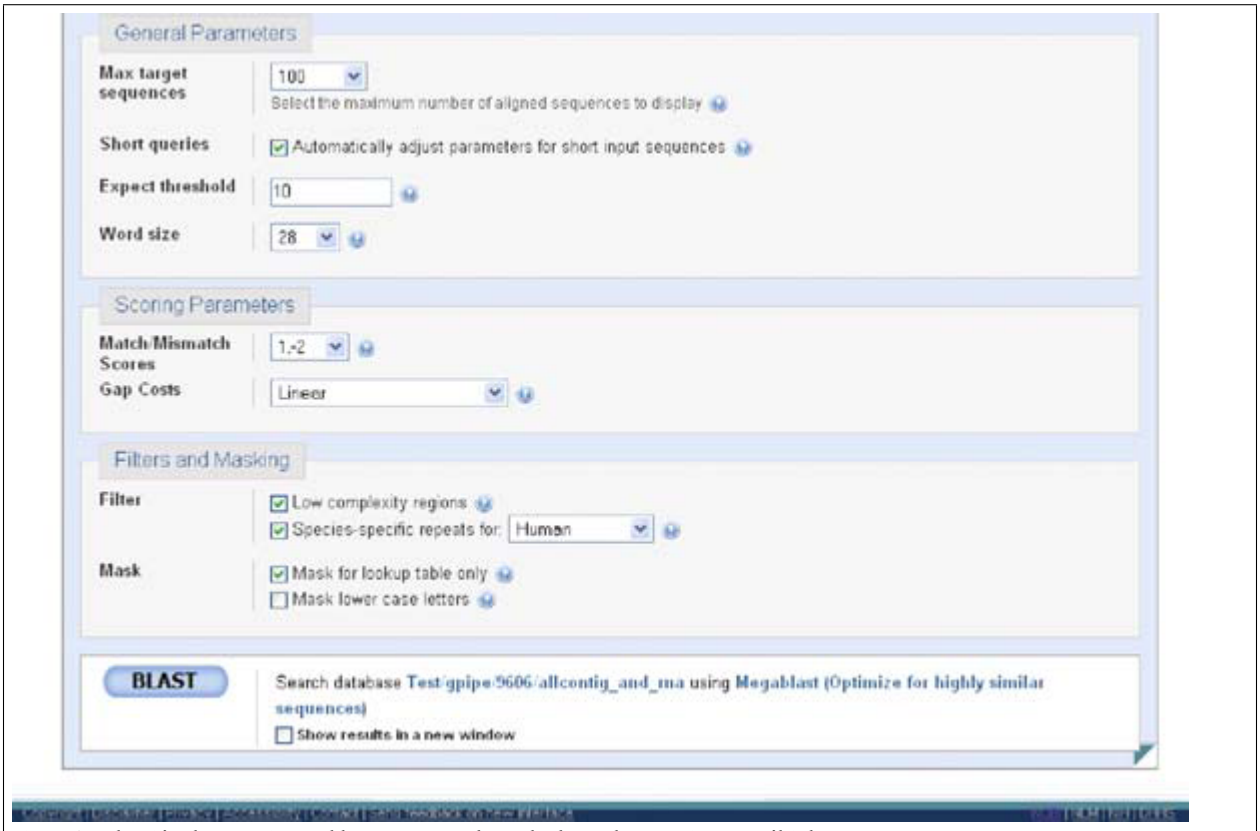


Fig. 1.6. Selección de organismo al buscar en una base de datos de organismos múltiples.

2.7.1. Max Target Sequences:

El parámetro de máximo de secuencias objetivo le permite seleccionar la cantidad de secuencias que le gustaría que se muestren en sus resultados. Los números más bajos no reducen el tiempo de búsqueda, pero reducen el tiempo para devolver los resultados. Esto generalmente es solo un problema en una conexión lenta.

2.7.2. Short Queries:

Cuando usa secuencias de consulta pequeñas (de tamaño 30 o menor), los parámetros deben ser ajustados o usted no obtendrá resultados significantes estadísticamente. Seleccionando la caja "*short queries*" automáticamente ajusta los parámetros para retornar respuestas validas para secuencias de consulta cortas.

2.7.3. Expected Threshold:

El umbral esperado limita los resultados mostrados a estos que tienen un *E-valor* menor que este. Este valor corresponde al número de secuencias coincidentes que son esperadas encontrarse meramente por azar.

2.7.4. Tamaño de la Palabra:

El tamaño de la palabras W, como se discutió antes, determina la longitud de las palabra y palabras vecinas usadas como consultas de búsqueda inicial. Incrementado el tamaño de palabra generalmente resulta en menos inicializaciones de extensión, lo que incremente la velocidad de las búsquedas de BLAST pero decrementa sensibilidad.

2.7.5. Parámetros de Puntaje

Los parámetros de puntaje de una búsqueda de nucleótidos son los puntajes de *match* (coincidencia), *mismatch (no-coincidencia) y gaps* (huecos). En búsquedas de proteínas, los puntajes de match y mismatch

son indicados por una ***matriz de puntajes*** (vea Sección 2.3). Un conjunto limitado de puntajes sugeridos de match y mismatch están disponibles desde el menú desplegable sobre la forma de búsqueda de BLAST de NCBI. Incrementado el porcentaje de la siguiente forma (match, mismatch): (1,–1) → (4,–5) → (2,–3) → (1,–2) → (1,–3) → (1,–4) previene nucleótidos no coincidentes en el alineamiento, incrementando el número de gaps, pero decrementando mismatches. Entre más grande es la divergencia que usted espera en secuencias que está buscando, mayor es el porcentaje que usted debería seleccionar.

NCBI ha provisto las guías mostradas en la Tabla 1.5. Adicionalmente, decrementando la existencia de *gaps* y las penalidades por extensión incrementará la incidencia de *gaps*

Match/mismatch ratio	Similarity (%)
0.33 (1/–3)	99
–0.5 (1/–2)	95
–1 (1/–1)	75

Tabla 1.5. Parámetros de puntuación sugeridos para búsquedas BLAST nucleótido-nucleótido. Cuando realice una búsqueda nucleótido-nucleótido-nucleótido, estas guías generales pueden ser usadas para escoger el puntaje *match/mismatch* basado en el grado de conservación que usted espera ver en sus resultados. Si usted está buscando secuencias con un alto grad de similaridad (i.e., dentro de una especie), los parámetros por defecto de (match +1, mismatch -2) serían los apropiados. Si, no obstante, usted está buscando secuencias entre organismo muy distintos (ejemplo: un gusano y un ratón), un porcentaje más pequeño sería el apropiado (por ejemplo, -1). Información provista por en NCBI (26).

2.7.6. Filtres

El filtro de regiones de baja complejidad remueve las regiones de la secuencia con baja complejidad, previniendo que estos segmentos produzcan resultados estadísticamente significativos pero no-informativos. El programa *DUST* de Tatusov and Lipmant (no publicados) se usa para búsquedas de nucleótidos en BLAST. A menudo, cuando una búsqueda tomo mucho más tiempo del esperado, la secuencia consulta (*query)* contiene una región de baja complejidad que está siendo igualada (*matched*) con muchas secuencias similares pero no relacionadas.

Sin embargo, es importante tener en cuenta que activar este filtro puede eliminar algunas coincidencias interesantes e informativas de los resultados. En las búsquedas de nucleótidos, también es posible eliminar las repeticiones específicas de especie al marcar la casilla "*Especies específicas para*:" y seleccionar las especies apropiadas. Esto evita que las repeticiones que son comunes en una especie en particular produzcan falsos positivos con otras partes de su propio genoma o genomas estrechamente relacionados.

2.7.7. Máscaras

La opción "*Máscara para tabla de búsqueda solamente*" permite al usuario enmascarar las regiones de baja complejidad (regiones de composición sesgada que incluyen ejecuciones homopoliméricas, repeticiones de períodos cortos, etc.) durante la etapa de siembra, donde las palabras y palabras vecinas se escanean, pero desenmascararlos durante las fases de extensión. Esto evita que los E-valores se vean afectados en resultados biológicamente interesantes al tiempo que evita que las regiones de baja complejidad ralenticen la búsqueda y la introducción de resultados poco interesantes.

La opción "*Máscara de letras minúsculas*" le da al usuario la opción de anotar su secuencia usando letras minúsculas donde se desea enmascarar.

2.8. INTERPRETACIÓN DE LOS RESULTADOS

Por defecto, los resultados de BLAST contienen cinco secciones básicas:

- Un resumen de su entrada (consulta y parámetros),
- Una descripción gráfica de los resultados principales,
- Una tabla de secuencias que producen alineaciones significativas,

- Las mejores 100 alineaciones, y
- Estadísticas de resultados.

El número de aciertos o *hits* que se muestran en la vista general gráfica, así como el número de alineaciones, entre otras opciones, se pueden cambiar haciendo clic en "*Actualizar estos resultados*" en la parte superior de la página de resultados o haciendo clic en "Formatear opciones" en la página de Resultados de formato (la página que aparece después de hacer clic en BLAST y antes de que aparezcan los resultados).

En la tercera sección, la tabla de resultados contiene ocho columnas: número de acceso, descripción, puntuación máxima, la puntuación total, cobertura consulta, E-valor, ident máximo, y enlaces.

El número de acceso proporciona un enlace a información detallada sobre la secuencia. La descripción proporciona información sobre la especie y el tipo de muestra del que se generó el *hit*. El puntaje máximo proporciona una métrica de cuán buena es la mejor alineación local. La puntuación total indica qué tan similar es la secuencia a la consulta, teniendo en cuenta todas las alineaciones locales entre las dos secuencias. Si el puntaje máximo es mayor que el puntaje total, entonces se encontró más de un alineamiento local entre las dos secuencias.

Los puntajes más altos se correlacionan con secuencias más similares. Ambas puntuaciones, informadas en bits, se calculan a partir de una fórmula que tiene en cuenta las coincidencias (o residuos similares, si se realiza una búsqueda de proteínas) y las penalizaciones de falta de coincidencia junto con las penalizaciones por inserción de huecos. Los puntajes de bit se normalizan para que puedan compararse directamente aunque las alineaciones entre diferentes secuencias puedan ser de diferentes longitudes.

El valor de expectativa o ***E-valor*** proporciona una estimación de la probabilidad de que esta alineación ocurra por azar. Un valor E de 2e-02 indica que la similitud encontrada en la alineación tiene una posibilidad de 2 en 100 de ocurrir por casualidad. Cuanto menor es el E-valor, más significativa es la puntuación. Un E-valor de corte apropiado depende de los objetivos de los usuarios.

El campo de identidad máxima muestra el porcentaje de la secuencia de consulta que fue idéntica al *hit* de la base de datos.

El campo de enlaces proporciona enlaces a UniGene, Gene Expression Omnibus, Entrez Gene, Entrez's Related Structures (para secuencias de proteínas) y Map Viewer (para secuencias genómicas).

2.9. Futuro de las Búsquedas de Similaridad

Como las matrices PAM y BLOSUM se derivan experimentalmente de un conjunto limitado de secuencias en una base de datos que estaba disponible en el momento en que se crearon, es casi seguro que no proporcionan valores óptimos para las búsquedas con nuevas familias de secuencias. La investigación actual se está realizando para determinar qué propiedades químicas están cambiando en una secuencia para proporcionar una magnitud de cambio que es independiente de las matrices de puntuación.

Las técnicas actuales para encontrar regiones promotoras carecen de precisión (19). En el futuro, surgirán técnicas que pueden mejorar los métodos actuales mediante el uso de algoritmos de tipo BLAST para evaluar la similitud de una secuencia con elementos promotores conocidos, lo que ayuda a identificarlo como un promotor.

3. Ejemplos

Esta sección proporcionará tres ejemplos de usos comunes de BLAST: un nucleótido-nucleótido BLAST, un BLAST iterado de posición específica y un blastx.

3.1. BLAST nucleótido-nucleótido para búsqueda de alelos

Aquí presentamos un ejemplo de uso de BLAST para buscar los alelos conocidos de una secuencia de nucleótidos dada. Este enfoque se puede usar para responder a la pregunta: ¿cuáles son las variantes conocidas de mi gen de interés (dentro de su especie)? Nuestro ejemplo será encontrar todas las variantes conocidas de una secuencia de nucleótidos de Tp53 (número de acceso AF151353) de un ratón. Si bien esta secuencia codifica una proteína, las secuencias no codificantes funcionarían igual de bien con este enfoque.

Comenzaremos visitando la página principal de BLAST en `http://www.ncbi.nlm.nih.gov/BLAST/` y seleccionando *Blast de nucleótidos*. En el cuadro "*Introducir secuencia de consulta*", escribimos el número de

acceso: AF151353. Notará que el recuadro "Título del trabajo" completa automáticamente un título para usted "AF151353: Supresor tumoral Mus musculus p53 ... ". Si tuviéramos que pegar una secuencia en lugar de un número de acceso o GI, nos gustaría ingresar un título de trabajo para ayudarnos a realizar un seguimiento de nuestros resultados. En "*Elegir conjunto de búsqueda*", seleccionamos la base de datos "*Colección de nucleótidos (nr / nt)*", ya que es la base de datos más completa (recuerde que nr ya no es redundante). Para una búsqueda completa, también debería realizar una búsqueda en la base de datos "Expressed sequence tags (est)". En el cuadro Organismo, seleccionamos el tipo "mouse" y seleccionamos "mouse (taxid: 10090)," que corresponde a *Mus musculus*, el mouse de la casa. Como estamos buscando alelos, seleccionamos "Secuencias muy similares (megablast)" en el cuadro "*Selección de programa*".

Luego, cambiamos los parámetros del algoritmo. Haga clic en "Parámetros de algoritmo" para visualizarlos. Dado que la secuencia tiene una longitud de 1.409 pb, anulamos la selección de la casilla "Ajustar parámetros automáticamente para secuencias de entrada cortas". Dado que esperamos que la proteína p53 sea una proteína bien conservada (debido a su función crítica), establecemos el umbral esperado a un valor bajo Vamos a elegir 1e-8.

Para un tamaño de palabra, no nos preocupa la velocidad en este caso, por lo que la cantidad de extensiones realizadas no es una preocupación. Seleccionemos un tamaño de palabra de 20 para asegurarnos de no perder ninguna coincidencia (aunque en este caso un tamaño de palabra más grande no debería hacer mucha diferencia). En cuanto a los parámetros de puntuación, elegimos la proporción más grande, que corresponde a la identidad más grande: "1, -4". Dado que esta es una secuencia de codificación de proteínas, no esperamos que las repeticiones sean un factor, entonces dejamos los *Filtros y sección de enmascaramiento* en la configuración predeterminada.

Los resultados indican que se encontraron 108 *hits (coincidencias)* en la secuencia de consulta. Al observar la alineación gráfica (figura 1.7), observamos que solo alrededor de 2/3 de ellos abarcan una buena parte de la consulta. Cuando nos desplazamos hacia abajo a las descripciones de los genes, la mayor parte del último cuarto son pseudogenes (secuencia parcial) (figura 1.8), que pueden ofrecer una idea de los diferentes alelos y sus correspondientes fenotipos, pero que no fueron secuenciados experimentalmente. Al realizar una búsqueda en la base de datos EST con los mismos parámetros, se obtienen 101 visitas adicionales.

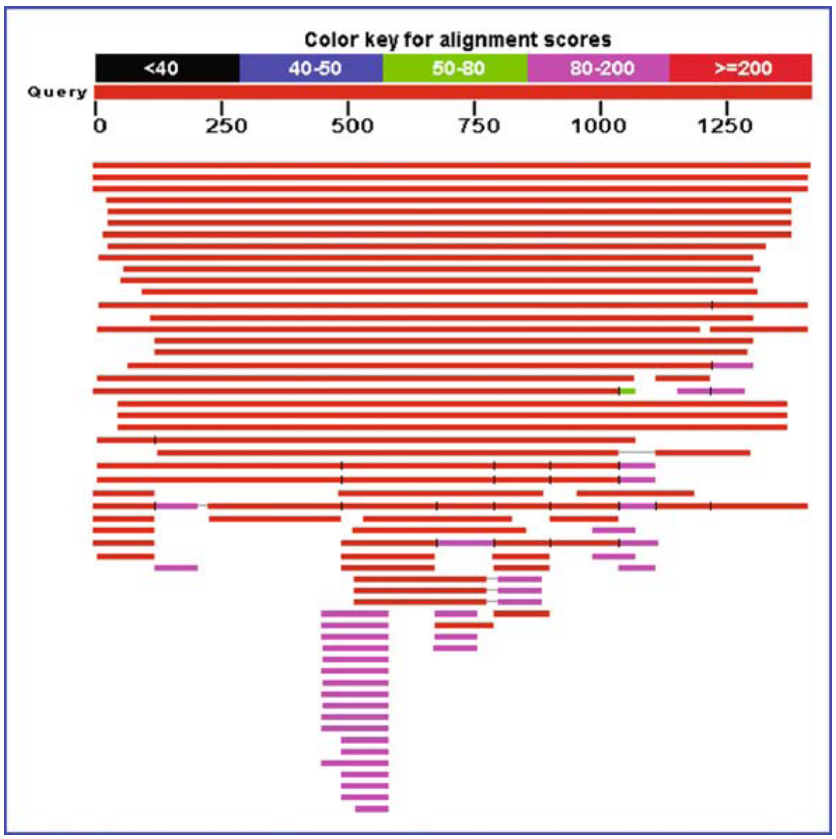
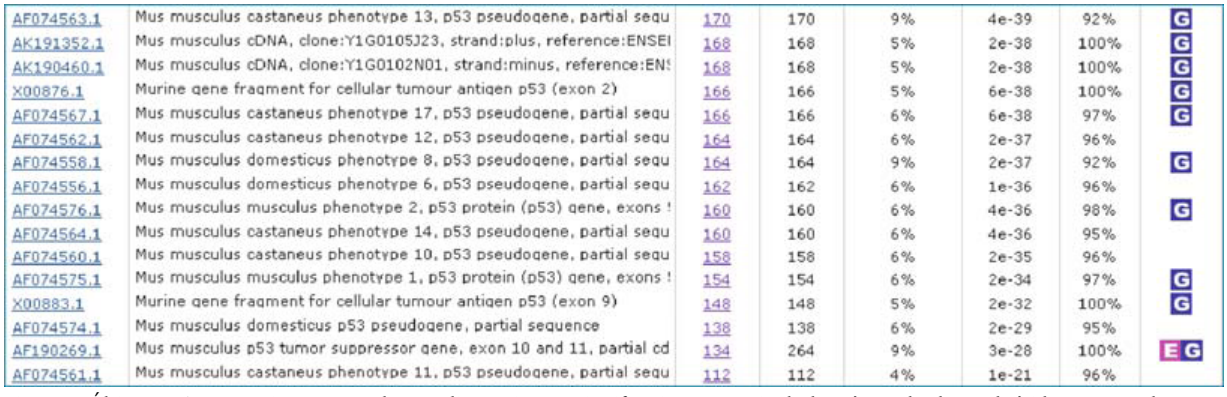


Fig. 1.7. Graphical distribution of top 100 BLAST hits.



AF074563.1	Mus musculus castaneus phenotype 13, p53 pseudogene, partial sequ	170	170	9%	4e-39	92%	G
AK191352.1	Mus musculus cDNA, clone:Y1G0105J23, strand:plus, reference:ENSEI	168	168	5%	2e-38	100%	G
AK190460.1	Mus musculus cDNA, clone:Y1G0102N01, strand:minus, reference:EN	168	168	5%	2e-38	100%	G
X00876.1	Murine gene fragment for cellular tumour antigen p53 (exon 2)	166	166	5%	6e-38	100%	G
AF074567.1	Mus musculus castaneus phenotype 17, p53 pseudogene, partial sequ	166	166	6%	6e-38	97%	G
AF074562.1	Mus musculus castaneus phenotype 12, p53 pseudogene, partial sequ	164	164	6%	2e-37	96%	
AF074558.1	Mus musculus domesticus phenotype 8, p53 pseudogene, partial sequ	164	164	9%	2e-37	92%	G
AF074556.1	Mus musculus domesticus phenotype 6, p53 pseudogene, partial sequ	162	162	6%	1e-36	96%	
AF074576.1	Mus musculus musculus phenotype 2, p53 protein (p53) gene, exons !	160	160	6%	4e-36	98%	G
AF074564.1	Mus musculus castaneus phenotype 14, p53 pseudogene, partial sequ	160	160	6%	4e-36	95%	
AF074560.1	Mus musculus castaneus phenotype 10, p53 pseudogene, partial sequ	158	158	6%	2e-35	96%	
AF074575.1	Mus musculus musculus phenotype 1, p53 protein (p53) gene, exons !	154	154	6%	2e-34	97%	G
X00883.1	Murine gene fragment for cellular tumour antigen p53 (exon 9)	148	148	5%	2e-32	100%	G
AF074574.1	Mus musculus domesticus p53 pseudogene, partial sequence	138	138	6%	2e-29	95%	
AF190269.1	Mus musculus p53 tumor suppressor gene, exon 10 and 11, partial cd	134	264	9%	3e-28	100%	EG
AF074561.1	Mus musculus castaneus phenotype 11, p53 pseudogene, partial sequ	112	112	4%	1e-21	96%	

Fig. 1.8. Últimas 16 secuencias que producen alineamientos significativos a partir de la Búsqueda de nucleótidos BLAST de un gen p53 de ratón. Diecinueve de las últimas 26 secuencias informadas son pseudogenes.

Al buscar secuencias lejanas relacionadas, hay dos opciones de BLAST disponibles. Uno es el BLAST estándar nucleótido-nucleótido con BLAST discontiguo, un método muy similar al trabajo de Ma et al. (20), seleccionado como el programa. El otro es usar un enfoque más sensible, PSI-BLAST, que realiza una búsqueda iterativa en una consulta de secuencia de proteínas. Aunque el segundo enfoque solo funcionará si se trata de secuencias de codificación de proteínas, es más sensible y preciso que el primero.

En este ejemplo, buscaremos parientes del gen del citocromo *b* del lagarto nocturno de Durango (*Xantusia extorris*). Comenzamos seleccionando explosión de proteína desde la página de inicio de BLAST e ingresando el número de acceso ABY48155 en el cuadro de consulta. Si su secuencia no está disponible como una secuencia de proteínas, tendrá que traducirla. Esto se puede hacer fácilmente usando un programa como MEGA (21), disponible en `http://www.megasoftware.net`, o una herramienta en línea como *JustBio Translator* (`http://www.justbio.com/translator/`) o la herramienta de traducción ExPASy (`http://www.expasy.org/herramientas/dna.html`).

Una vez más, la casilla "Título del trabajo "se completa con "ABY48155: citocromo b [Xantusia extorris]". Seleccionaremos la base de datos "Proteínas de referencia (refseq_protein)", que es más altamente curada y no redundante (por gen) que la base de datos predeterminada nr. No especificamos un organismo porque queremos resultados de todos los organismos relacionados. Para el algoritmo, seleccionamos PSI-BLAST debido a su capacidad para detectar secuencias relacionadas más distantemente. Esperamos incluir tantas secuencias como sea posible en nuestras iteraciones, por lo que elegimos 1,000 como las secuencias de destino máximo. Podemos, una vez más, eliminar la comprobación "Ajustar automáticamente los parámetros para secuencias de entrada cortas", ya que nuestra secuencia es suficientemente larga (380 aminoácidos). Como deseamos detectar todas las secuencias relacionadas, mantenemos el umbral de espera en su valor predeterminado de 10. Si bien la disminución puede eliminar los falsos positivos, también puede evitar que se devuelvan algunos resultados significativos. Como no tenemos un alcance particular en mente (dentro del género o familia, por ejemplo), usaremos la matriz BLOSUM62 debido a su capacidad de detectar homología en grandes rangos de similitud.

La primera iteración da como resultado 1000 *hits* en la secuencia de consulta, todas las cuales cubren al menos el 93 % de la secuencia de consulta y tienen un E-valor de $10^{1}26$ o menos. Dejamos todas las secuencias seleccionadas y presionamos el botón "Ejecutar iteración PSI-Blast 2". La segunda iteración también devuelve 1000 hits, pero esta vez tienen E-valores inferiores a $10^{-}99$ y cubren al menos un 65 % de la secuencia de consulta (todos excepto seis cubren el 90 % o más). Desmarcamos el último *hit*, Bi4p [*Saccharomyces cerevisiae*], ya que no estamos seguros de su homología, y repetimos una última vez.

En este punto, sería útil ver el informe de taxonomía de los resultados. Puede hacerlo haciendo clic en "Informes de taxonomía" cerca de la parte inferior de la primera sección del informe BLAST. Notarás que tenemos una buena selección de organismos, desde peces óseos hasta Proteobacteria. Si bien esta lista debería reducirse para producir una buena taxonomía, sería un buen punto de partida si desea realizar una amplia reconstrucción filogenética. Para realizar una búsqueda de secuencias más estrechamente relacionadas, es probable que realice una blastp estándar (proteína-proteína BLAST) en lugar de una PSI-BLAST y use la matriz PAM 70 o PAM 30.

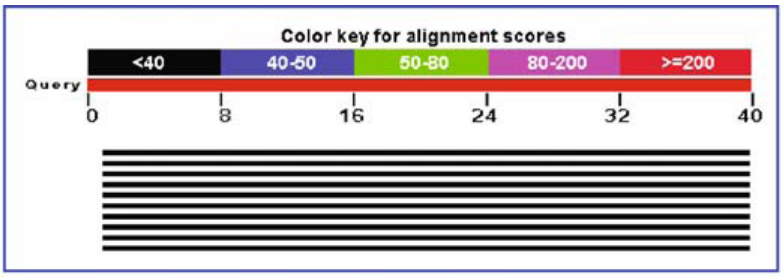
¿Qué pasa si tienes una secuencia de nucleótidos tal como una etiqueta de secuencia expresada (EST) y deseas saber si codifica una proteína conocida? Puede buscar en la base de datos de nucleótidos o tomar el enfoque más directo de Blastx. Blastx le permite buscar en la base de datos de proteínas mediante una consulta de nucleótidos, que primero se traduce en los seis marcos de lectura. En este ejemplo, realizaremos

un blastx en la siguiente secuencia:

TCTCTATAGTTATGGTGTTCTGAATCAGCCTTCCCTCATA

Como la secuencia tiene solo 40 pb de longitud, debemos tener cuidado con nuestros parámetros. Comenzamos seleccionando blastx desde la página principal de BLAST. Luego ingresamos la secuencia en el cuadro de consulta e ingresamos un título de trabajo relevante, como "EST blastx Search 1". Buscaremos en la base de datos "Secuencias de proteínas no redundantes (nr)", ya que tiene el mayor número de secuencias de nucleótidos anotadas. En "Parámetros de algoritmo", debemos elegir un umbral y una matriz de espera adecuados. Si elegimos un umbral de espera demasiado bajo, es posible que no encontremos nada. Del mismo modo, si elegimos la matriz incorrecta, es posible que no obtengamos resultados significativos debido a la corta longitud de nuestra secuencia. Elegiremos 10 (el valor predeterminado) como nuestro umbral esperado y PAM70 como nuestra matriz, ya que corresponde a la búsqueda de similitud en el nivel de familia/género o por debajo de este. Como no sabemos cuál es nuestra secuencia, queremos filtrar las regiones de baja complejidad para garantizar que si nuestra secuencia contiene dichas regiones, no obtendrán resultados engañosamente significativos.

Nuestra búsqueda produce un gran número (más de 1,000) de resultados con un E-valor de 0.079 (figura 1.9). Si tuviéramos que usar la matriz PAM70, esencialmente se obtendrían los mismos resultados, pero cada uno con un E-valor de 3.0. Dado que los 2,117 resultados son entradas diferentes de la proteína nucleocápside del virus de la Influenza A, podemos estar seguros de que nuestra proteína está relacionada, especialmente si tenemos algún conocimiento previo que respalde nuestros hallazgos.



Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer

Sequences producing significant alignments:
(Click headers to sort columns)

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
ABY81430.1	nucleocapsid protein [Influenza A virus (A/swine/Iowa/1/1986(H1N1))	38.3	38.3	97%	0.079	92%	
ABQ43788.1	nucleocapsid protein [Influenza A virus (A/mallard/ON/499/2005(H5N	38.3	38.3	97%	0.079	92%	
CAN89845.1	nucleoprotein [Influenza A virus (A/wild boar/Germany/R169/2006(H:	38.3	38.3	97%	0.079	92%	
ABQ12377.1	nucleocapsid protein [Influenza A virus (A/mallard/Manitoba/458/200	38.3	38.3	97%	0.079	92%	
ABV53582.1	nucleoprotein [Influenza A virus (A/chicken/Nigeria/1071-30/2007(H5	38.3	38.3	97%	0.079	92%	
ABV53572.1	nucleoprotein [Influenza A virus (A/chicken/Nigeria/1071-29/2007(H5	38.3	38.3	97%	0.079	92%	
ABV53532.1	nucleoprotein [Influenza A virus (A/chicken/Nigeria/1071-10/2007(H5	38.3	38.3	97%	0.079	92%	
ABV53512.1	nucleoprotein [Influenza A virus (A/chicken/Nigeria/1071-7/2007(H5N	38.3	38.3	97%	0.079	92%	
ABV53492.1	nucleoprotein [Influenza A virus (A/chicken/Nigeria/1071-4/2007(H5N	38.3	38.3	97%	0.079	92%	
ABV53482.1	nucleoprotein [Influenza A virus (A/chicken/Nigeria/1071-3/2007(H5N	38.3	38.3	97%	0.079	92%	

Fig. 1.9. BLASTX resultados muestran E-valores de 0,079 para los diez mejores <10> hits, todos los cuales son proteínas o nucleoproteínas nucleocápside.

4. Notas

1. Una de las opciones que ofrece el NCBI desde su página de inicio es buscar entre sus bases de datos utilizando un identificador (número de acceso, número de identificación de secuencia, ID de Locus, etc.). Esta opción puede ser bastante simple si está usando un identificador único para una secuencia en particular; Sin embargo, si está buscando un *locus* a través de organismos o personas, es posible que deba prestar mucha atención a los términos de búsqueda que está utilizando. Por ejemplo, dado que la subunidad Cytochrome b/b6 se conoce con los términos "Cytochrome b," "Cytochrome b6," "cyt-b", "cytb", "cyb", "COB", "COB1", "cyb6", "petB", "mtcyb", y "mt-cyb", en una búsqueda de todos los homólogos posibles de esta subunidad, es necesario buscar todos sus nombres y abreviaciones usadas en los organismos de interés. Dado que los grupos de investigación que estudian diferentes organismos crean

Fig. 1.10. Guardado de Estrategias de búsquedas.

sus propios nombres de *locus* únicos para el mismo gen, es importante usarlos todos en su búsqueda. IHOP (`www.ihop-net.org`) es un excelente recurso para nombres de proteínas (22). Además, querrás realizar una búsqueda BLAST para asegurarte de tener todo.

2. Además del programa BLAST proporcionado por NCBI, existen otros programas BLAST que han mejorado el algoritmo BLAST de varias maneras. El Dr. Warren Gish en la Universidad de Washington en St. Louis ha desarrollado WU-BLAST, el primer algoritmo BLAST que permitió alineamientos abiertos con estadísticas (23). Cuenta con velocidad, precisión y flexibilidad, asumiendo incluso los trabajos más grandes. Otro programa, FSA-BLAST (algoritmo de búsqueda más rápido), se desarrolló para implementar las mejoras recientemente publicadas al algoritmo original BLAST (24). Promete ser el doble de rápido que NCBI e igualmente preciso. WU-BLAST es gratuito para uso académico y sin fines de lucro y FSA-BLAST es una fuente abierta bajo el acuerdo de licencia de BSD.

3. Mi NCBI es una herramienta que le permite personalizar sus preferencias, guardar búsquedas y configurar búsquedas automáticas que envíen resultados por correo electrónico. Si se encuentra realizando las mismas búsquedas (o incluso búsquedas similares) repetidamente, ¡puede aprovechar esta opción! Para registrarse, vaya a la página de inicio NCBI y haga clic en el enlace "Mi NCBI" en "Hot Spots". Una vez que se haya registrado e iniciado sesión, encontrará una nueva opción en todas las búsquedas BLAST y Entrez (Fig. 1.10).

4. Para guardar una estrategia de búsqueda de BLAST, simplemente haga clic en el enlace "Guardar estrategias búsqueda" en la página de resultados. Esto agregará la búsqueda en su página "Estrategias guardadas", que está disponible a través de una pestaña en la parte superior de cada página en el sitio web BLAST cuando hayas iniciado sesión en Mi NCBI. Hacer eso no salvará sus resultados, pero guardará su consulta y todos los parámetros que especificó para su búsqueda para que pueda ejecutarlo más tarde para recuperar resultados actualizados.