Lab Alineamiento de Secuencias usando el algoritmo BLAST

1. Introducción

La comparación de dos o más secuencias es un paso común en el flujo de trabajo de un bioinformático. Debido a que las secuencias de ADN definen la función de las proteínas en los seres vivos, comparar secuencias de interés contra base de datos de características conocidas es de gran utilidad. Estas bases de datos pueden incluir genes, elementos transponibles, repeticiones simples, motifs, proteínas, entre muchos otros. La comparación de secuencias se hace útil si tenemos en cuenta que entre más similares sean dos secuencias más similares tenderán a ser las funciones de las proteínas codificadas por ellas.

Además encontrar similitudes entre secuencias es útil para:

- Asegurarse de que dos secuencias son similares y cuantificar su similitud
- Encontrar dominios funcionales
- Comparar un gen y su producto
- Buscar posiciones homólogas en las secuencias

Para encontrar y cuantificar la similitud entre secuencias, debemos ejecutar un proceso llamado alineamiento de secuencias. Lo que se busca con un alineamiento óptimo es reducir al mínimo los "gaps" (o regiones que se dejan vacías en el alineamiento también llamados indels -insertions or deletions-) y los "mismatches" (residuos alineados que son diferentes entre ellos) y maximizar los "matches" (residuos alineados que son idénticos entre ellos).

Existen varios tipos de alineamientos, entre ellos el global y el local, si tenemos en cuenta las porciones de cada secuencia que usamos, los alineamientos pareados y múltiples si tenemos en cuenta la cantidad de secuencias y el gráfico, si tenemos en cuenta la presentación de los resultados. Cabe tener en cuenta que los alineamientos pueden ser entre ADN vs ADN, ADN vs aminoácidos, entre aminoácidos vs ADN o entre aminoácidos vs aminoácidos.

Alineamientos globales

Mediante un alineamiento global las secuencias se alinean en su totalidad, de principio a fin. El algoritmo básico para ello es el de Needleman-Wunsch. Este tipo de algoritmo es más útil cuando las secuencias tienen longitudes similares y se busca identificar si las secuencias en su totalidad son parecidas o no.

Alineamientos locales

Los alineamientos locales son más útiles para secuencias diferenciadas en las que se sospecha que existen regiones muy similares o motifs de secuencias similares dentro de un contexto mayor. El algoritmo Smith-Waterman es un método general de alineamiento local basado en programación dinámica.

Alineamientos pareados

Los métodos de alineamiento de pares, o emparejamientos, se utilizan para encontrar la mejor coincidencia en bloque (local) o alineamiento global de dos secuencias. Los alineamientos de pares sólo pueden utilizarse con dos secuencias a la vez, pero son eficientes de calcular, y son utilizados a menudo en métodos que no requieren precisión extrema, como la búsqueda en bases de datos de secuencias con alta homología de secuencia con respecto a una petición. Existen 3 métodos para calcular los alineamientos pareados, entre ellos la programación dinámica, la matriz de puntos y los algoritmos de búsqueda de palabras (para una mayor claridad en estos conceptos se recomienda leer el anexo 1).

Alineamientos múltiples

El alineamiento múltiple de secuencias es una extensión del alineamiento de pares que incorpora más de dos secuencias al mismo tiempo. Los métodos de alineamiento múltiple intentan alinear todas las secuencias de un conjunto dado. Los alineamientos múltiples son usados a menudos en la identificación de regiones conservadas en un grupo de secuencias que hipotéticamente están relacionadas evolutivamente, para la generación de árboles filogenéticos y para la identificación de SNPs (single nucleotide polymorphisms).

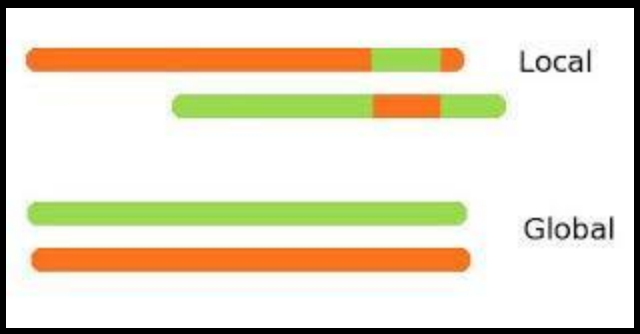


Figura 1. Representación de alineamientos locales y globales.

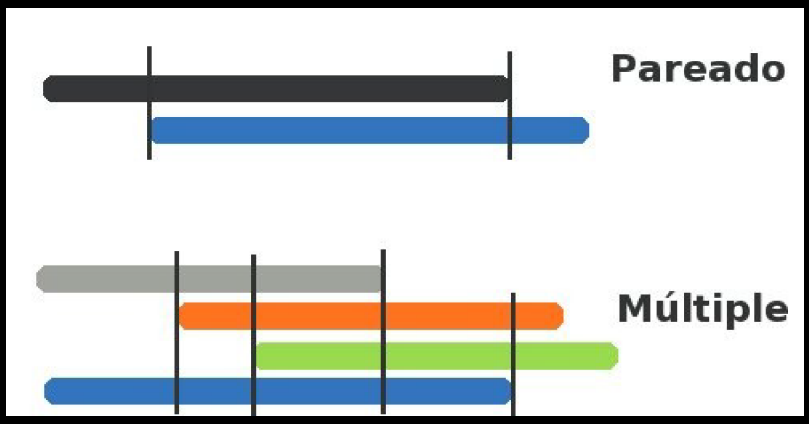


Figura 2. Representación de alineamientos pareados y múltiples.

En la mayoría de los algoritmos de alineamientos, es posible parametrizar los valores de penalización de los gaps y de los matches y mismatches. Para alinear aminoácidos, a menudo se usan matrices de sustitución que proveen al algoritmo de los valores necesarios para calcular el alineamiento óptimo y hallar la medida conocido como "score". Es importante mencionar que entre más alto sea este valor mayor significancia biológica tiene. Algunas de las matrices de sustitución más utilizadas son:

- PAM: Percent Accepted Mutation Matrix

  - Derivadas de alineamientos globales de secuencias cercanamente relacionadas.
  - PAM40 PAM250. A mayor N° mayor distancia evolutiva

- BLOSUM

  - Derivadas de alineamientos locales de secuencias distantes
  - BLOSUM90, BLOSUM45. El N° representa porcentaje de identidad

Además de esta medida, la mayoría de los algoritmos de alineamientos arrojan otros datos de interés como longitud del alineamiento (en el caso de los alineamientos locales), la similitud (que nos indica cual es el porcentaje de residuos iguales que contiene el alineamiento) y el e-value, que nos indica probabilísticamente la cantidad de hits que podría tener la secuencia buscada en una base de datos aleatoria (en el caso de blast, el cual es el más usado en la actualidad). Entre menor sea el e-value, mayor es la probabilidad de que las secuencias que estamos comparando compartan similitudes de forma no aleatoria.

2. Herramientas

Herramienta	Manual	Descripción

biolinux	http://nebc.nerc.ac.uk/courses/Bio-Linux/Bio-Linux_feb2009/IntroductionToBio-Linux5.0_Feb2009.pdf	Sistema operativo basado en Ubuntu (Linux) que cuenta con una gran cantidad de herramientas para análisis bioinformáticos.

Tabla 1. Herramientas necesarias para desarrollar esta guía.

3. Procedimiento

a. Obteniendo los datos

Al igual que en otros procesos bioinformáticos, el primer paso consiste en obtener los datos desde alguna base de datos libre o privada. En nuestro caso usaremos el cromosoma 1 de la planta modelo *Arabidopsis thaliana*, el cual podremos descargar desde el siguiente enlace:

ftp://ftp.arabidopsis.org/home/tair/Sequences/whole_chromosomes/TAIR10_chr1.fas

El Segundo paso es obtener algunas secuencias de ADN complementario (cDNA) obtenidos de la misma planta, que se puede descargar desde el siguiente enlace: ftp://ftp.arabidopsis.org/home/tair/Sequences/ATH_cDNA_EST_sequences_FASTA/ATH_cDNA_sequences_20101108.fas. Nuestro objetivo entonces es comparar las secuencias de cDNA contra la secuencia del cromosoma 1 de la planta, para identificar posibles secuencias de interés en este cromosoma.

b. Usando el algoritmo blast localmente

Para optimizar el tiempo de cómputo, el algoritmo blast formatea la secuencia que usaremos para comparar (la cual llamaremos base de datos), creando varios archivos binarios. Este es el primer paso que debemos hacer al usar el algoritmo blast. Para crear la base de datos formateada de blast usamos el siguiente comando en la consola de Biolinux (debemos navegar en las carpetas hasta llegar a donde se encuentre la secuencia del cromosoma que descargamos en el paso anterior):

```
makeblastdb -in TAIR10_chr1.fas -dbtype nucl
```

El parámetro -in indica la secuencia de entrada que usaremos como base de datos, mientras que el parámetro -dbtype indica la naturaleza de los datos contenidos en el archivo de entrada. En nuestro caso son nucleótidos, por lo tanto usamos la palabra nucl. En el último parámetro también podemos usar la palabra prot, cuando nuestra base de datos contenga proteínas.

A continuación usaremos el comando para generar el alineamiento de secuencias. Cabe mencionar que existen varios comandos que usa blast dependiendo de los datos contenidos en la base de datos y en nuestras secuencias de interés que llamaremos query. Estos comandos son (el formato que usamos es "query vs base de datos"):

- blastn: para ADN vs ADN
- blastp: para proteína vs proteínas
- blastx: para ADN vs proteínas
- tblastn: para proteínas vs ADN

En el caso de blastx, el algoritmo traduce las secuencias query a proteínas, teniendo en cuenta las 6 posiciones diferentes de inicio (3 para *forward* Y 3 para *reverse*). En nuestro ejemplo usaremos el comando blastn, debido a que haremos comparaciones entre nucleótidos.

Otro punto a considerar, es que el algoritmo blast usa el método de alineamiento local y hace las comparaciones a través de alineamientos pareados, en donde cada una de las secuencias del archivo multifasta (que puede contener muchas secuencias en formato fasta) que usaremos como query, se alineará contra la secuencia que usaremos como base de datos. En el caso de que el archivo de base de datos también sea multifasta, el algoritmo ejecutará alineamientos

pareados entre cada una de las secuencias del query contra cada una de las secuencias de la base de datos.

Para ejecutar el alineamiento usaremos entonces el siguiente comando en el terminal de Biolinux:

```
blastn -query ATH_EST_sequences_20101108.fas -db TAIR10_chr1.fas -num_threads 4 -outfmt 6 -out cDNA_vs_cromosoma.blast
```

En donde los parámetros -query y -db indican las secuencias que vamos a comparar, -num_threads indica la cantidad de procesos en el cual podemos ejecutar el algoritmo (depende en gran medida de la cantidad de CPUs de la máquina en donde será ejecutado), -outfmt nos indica el formato de salida (6 para formato tabular; consulte la tabla 2 para los otros tipos de formatos permitidos) y -out indica el archivo donde quedarán guardados los resultados. Si este parámetro no es indicado, la salida del blast se imprimirá en la consola.

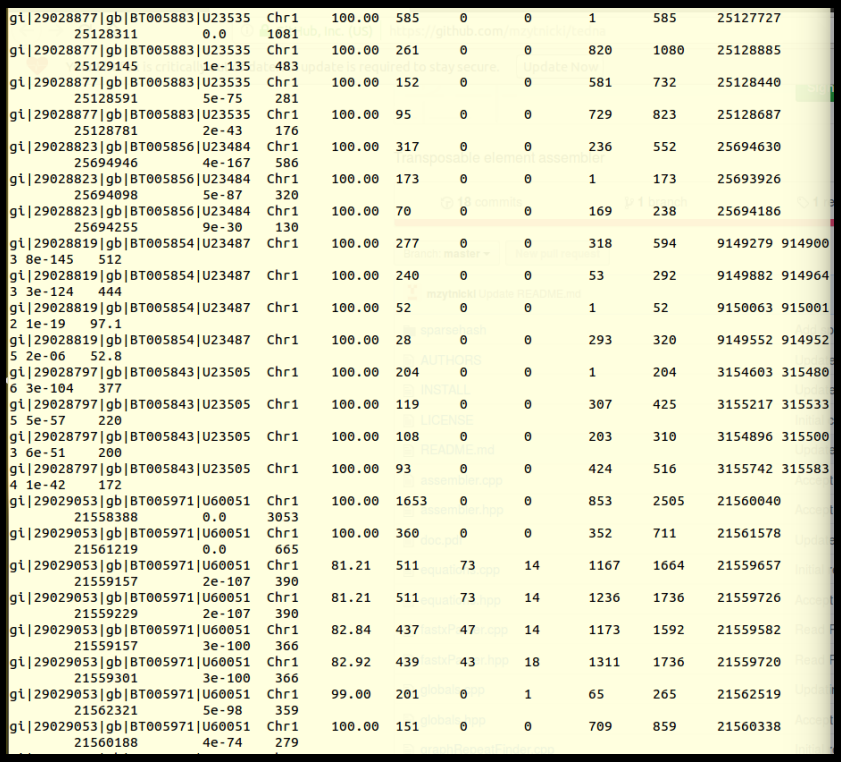
Número	Formato
0	Pairwise
1	Query-anchored showing identities
2	Query-anchored no identities
3	Flat query-anchored, show identities
4	Flat query-anchored, no identities
5	XML Blast output
6	Tabular
7	Tabular with comment lines
8	Text ASN.1
9	Binary ASN.1
10	Comma-separated values
11	BLAST archive format (ASN.1)

Tabla 2. Formatos de salida permitidos por el algoritmo blast.

Una vez ejecutado el blast, podremos ver un nuevo archivo en nuestro directorio a través del comando 'ls -l' y podemos visualizarlo si escribimos en la consola:

```
less cDNA_vs_cromosoma.blast
```

Este comando nos mostrará en pantalla lo siguiente (recuerde que para salir de la visualización del archivo debe usar la tecla q):



```
gi|29028877|gb|BT005883|U23535  Chr1    100.00  585     0       0       1       585     25127727
        25128311        0.0     1081
gi|29028877|gb|BT005883|U23535  Chr1    100.00  261     0       0       820     1080    25128885
        25129145        1e-135  483
gi|29028877|gb|BT005883|U23535  Chr1    100.00  152     0       0       581     732     25128440
        25128591        5e-75   281
gi|29028877|gb|BT005883|U23535  Chr1    100.00  95      0       0       729     823     25128687
        25128781        2e-43   176
gi|29028823|gb|BT005856|U23484  Chr1    100.00  317     0       0       236     552     25694630
        25694946        4e-167  586
gi|29028823|gb|BT005856|U23484  Chr1    100.00  173     0       0       1       173     25693926
        25694098        5e-87   320
gi|29028823|gb|BT005856|U23484  Chr1    100.00  70      0       0       169     238     25694186
        25694255        9e-30   130
gi|29028819|gb|BT005854|U23487  Chr1    100.00  277     0       0       318     594     9149279 914900
3 8e-145   512
gi|29028819|gb|BT005854|U23487  Chr1    100.00  240     0       0       53      292     9149882 914964
3 3e-124   444
gi|29028819|gb|BT005854|U23487  Chr1    100.00  52      0       0       1       52      9150063 915001
2 1e-19    97.1
gi|29028819|gb|BT005854|U23487  Chr1    100.00  28      0       0       293     320     9149552 914952
5 2e-06    52.8
gi|29028797|gb|BT005843|U23505  Chr1    100.00  204     0       0       1       204     3154603 315480
6 3e-104   377
gi|29028797|gb|BT005843|U23505  Chr1    100.00  119     0       0       307     425     3155217 315533
5 5e-57    220
gi|29028797|gb|BT005843|U23505  Chr1    100.00  108     0       0       203     310     3154896 315500
3 6e-51    200
gi|29028797|gb|BT005843|U23505  Chr1    100.00  93      0       0       424     516     3155742 315583
4 1e-42    172
gi|29029053|gb|BT005971|U60051  Chr1    100.00  1653    0       0       853     2505    21560040
        21558388        0.0     3053
gi|29029053|gb|BT005971|U60051  Chr1    100.00  360     0       0       352     711     21561578
        21561219        0.0     665
gi|29029053|gb|BT005971|U60051  Chr1    81.21   511     73      14      1167    1664    21559657
        21559157        2e-107  390
gi|29029053|gb|BT005971|U60051  Chr1    81.21   511     73      14      1236    1736    21559726
        21559229        2e-107  390
gi|29029053|gb|BT005971|U60051  Chr1    82.84   437     47      14      1173    1592    21559582
        21559157        3e-100  366
gi|29029053|gb|BT005971|U60051  Chr1    82.92   439     43      18      1311    1736    21559720
        21559301        3e-100  366
gi|29029053|gb|BT005971|U60051  Chr1    99.00   201     0       1       65      265     21562519
        21562321        5e-98   359
gi|29029053|gb|BT005971|U60051  Chr1    100.00  151     0       0       709     859     21560338
        21560188        4e-74   279
```

Figura 3. Salida tabular del alineamiento de secuencias obtenido a través del algoritmo blast.

El archivo está organizado en filas, donde cada línea es un hit (o un alineamiento local entre las secuencias query y la base de datos) y en donde cada columna corresponde a un campo específico, que nos brinda información relevante del alineamiento (ver Tabla 3).

Posición del campo	Nombre del campo
1	Query (e.g., gene) sequence id
2	Database (e.g., reference genome) sequence id
3	Percentage of identical matches
4	Alignment length
5	Number of mismatches
6	Number of gap openings
7	Start of alignment in query

8	End of alignment in query
9	Start of alignment in database
10	End of alignment in subject
11	Expect value
12	Bit score

Tabla 3. Nombres de los campos y su posición en el archivo de salida en formato tabular.

Por lo tanto podremos ver en la primera línea del archivo de salida, que la secuencia con id "gi|29028877|gb|BT005883|U23535" tiene una porción similar a la secuencia de referencia (o que usamos como base de datos) con id "Chr1" de una longitud de 585 nucleótidos, con un porcentaje de similitud del 100%, con 0 mismatches, con 0 gaps y que se encuentra en el cromosoma 1 entre los nucleótidos 25.127.727 y 25.128.311.

Adicionalmente podemos obtener algunos datos de interés usando comandos de la terminal de Linux, tales como:

- Cantidad de hits entre las secuencias usadas a través del comando `wc -l cDNA_vs_cromosoma.blast`
- Cantidad de hits entre una secuencia específica y la base de datos, a través del comando `grep -c "gi|29028877|gb|BT005883|U23535" cDNA_vs_cromosoma.blast`
- Obtener una lista de las secuencias query que tuvieron hits con la base de datos, a través del comando `cut -f1 cDNA_vs_cromosoma.blast | sort | uniq`
- Obtener la máxima longitud de alineamiento a través del comando `cut -f4 cDNA_vs_cromosoma.blast | sort -n | tail -n 1`

```
manager@bl8vbox[taller_Alineamientos] wc -l cDNA_vs_cromosoma.blast                [ 7:30PM]
127194 cDNA_vs_cromosoma.blast
manager@bl8vbox[taller_Alineamientos] cut -f4 cDNA_vs_cromosoma.blast | sort -n | tail -n 1 [ 7:30PM]
6067
manager@bl8vbox[taller_Alineamientos] grep -c "gi|29028877|gb|BT005883|U23535" cDNA_vs_cromosoma.blast

4
manager@bl8vbox[taller_Alineamientos] cut -f1 cDNA_vs_cromosoma.blast | sort | uniq  [ 7:31PM]
```

Figura 4. Análisis de los resultados del alineamiento.

c. Ejecución del algoritmo blast a través del portal web

El Centro Nacional de información de Biotecnología (NCBI por sus siglas en inglés) ofrece un servicio en línea para la ejecución

de blast entre secuencias queries de nuestro interés y algunas bases de datos que se han recopilado.

Para acceder a esta plataforma, debemos ingresar al siguiente enlace: https://blast.ncbi.nlm.nih.gov/Blast.cgi. Allí podremos usar cualquiera de los 4 tipos de blast que hablamos anteriormente (blastn, blastp, blastx o tblastn).

En nuestro caso, nos interesa saber si ya ha sido reportada la secuencia que tuvo una similitud del 100% con una región del cromosoma 1 de la planta *Arabidopsis thaliana*, para buscar su función. Para lograr este objetivo, se extrae la secuencia con id "gi|29028877|gb|BT005883|U23535", del archivo ATH_EST_sequences_20101108.fas. Lo podemos visualizar con el comando less y seleccionamos el encabezado y todos los nucleótidos, luego seleccionamos clic derecho y copiar.

Figura 5. Extracción de la secuencia de interés del archivo query

El siguiente paso consiste en ir al sitio web de blast del NCBI, seleccionar blastn y en el campo titulado "Enter sequence query" pegamos la secuencia que copiamos del archivo query. A continuación en el campo "Choose Search Set" escribimos *arabidopsis thaliana* (taxid:3702). Seleccionamos la especie de la planta para que la búsqueda sea más limitada y así poder obtener resultados más rápidamente. Si dejamos este campo vacío, el algoritmo buscará en las bases de datos de todas las

especies disponibles, esto puede ser interesante para ver secuencias homólogas en otras especies, como por ejemplo genes o elementos transponibles. Por último damos clic en el botón blast.

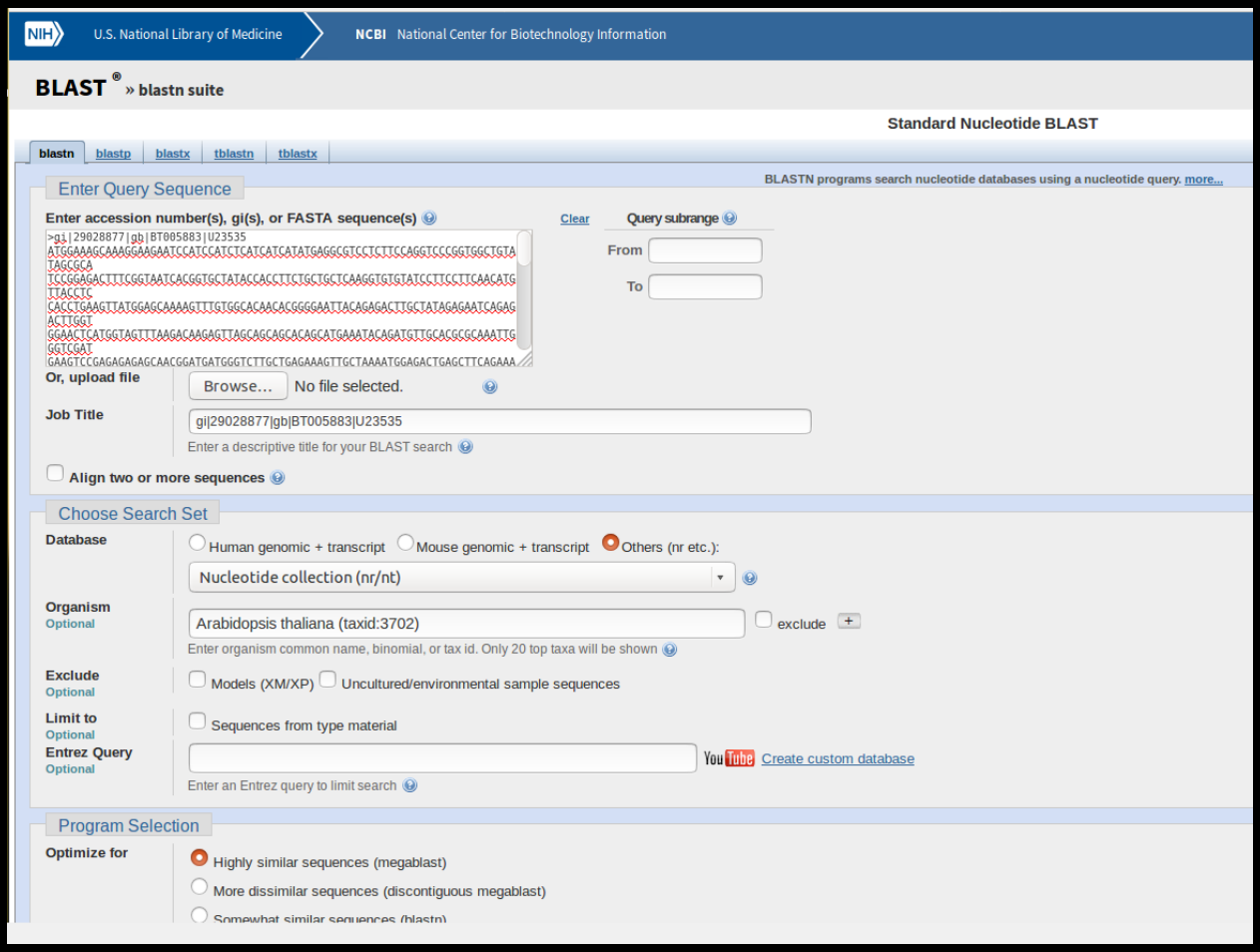
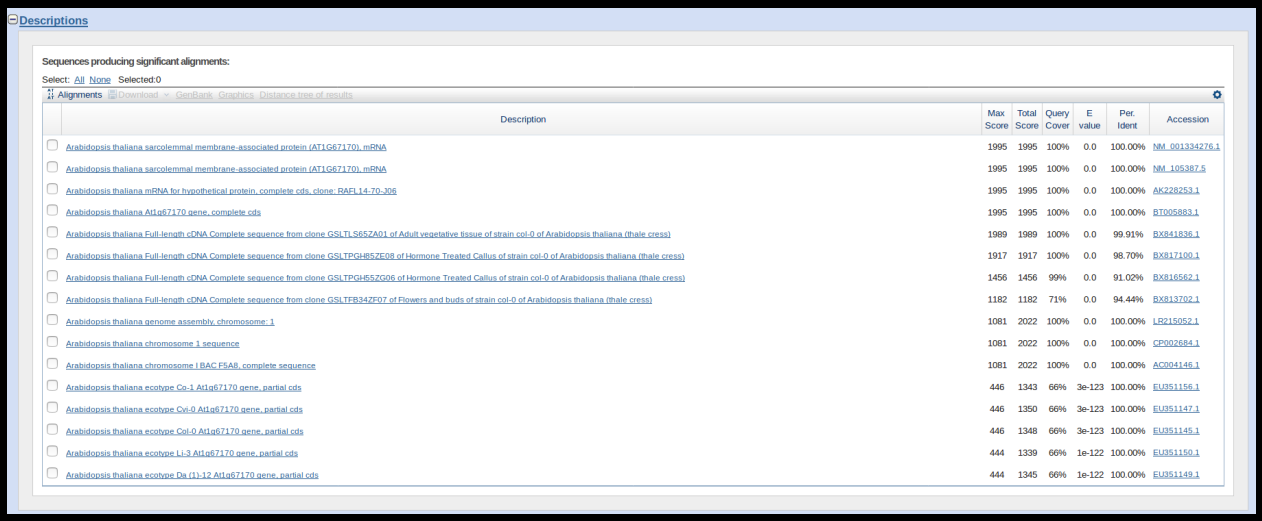


Figura 6. Parámetros necesarios para la ejecución del algoritmo blast en el sitio web del NCBI.

Después de algunos minutos, se mostrará en pantalla los resultados del alineamiento. El primer campo nos muestra de forma gráfica los mejores puntajes obtenidos, pero el segundo campo es el que nos brinda la información que necesitamos para cumplir con nuestro objetivo.



Descriptions

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenBank Graphics Distance tree of results

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
Arabidopsis thaliana sarcolemmal membrane-associated protein (AT1G67170), mRNA	1995	1995	100%	0.0	100.00%	NM_001334276.1
Arabidopsis thaliana sarcolemmal membrane-associated protein (AT1G67170), mRNA	1995	1995	100%	0.0	100.00%	NM_105387.5
Arabidopsis thaliana mRNA for hypothetical protein, complete cds, clone: RAFL14-70-J06	1995	1995	100%	0.0	100.00%	AK228253.1
Arabidopsis thaliana At1g67170 gene, complete cds	1995	1995	100%	0.0	100.00%	BT005883.1
Arabidopsis thaliana Full-length cDNA Complete sequence from clone GSLTLS65ZA01 of Adult vegetative tissue of strain col-0 of Arabidopsis thaliana (thale cress)	1989	1989	100%	0.0	99.91%	BX841836.1
Arabidopsis thaliana Full-length cDNA Complete sequence from clone GSLTPGH85ZE08 of Hormone Treated Callus of strain col-0 of Arabidopsis thaliana (thale cress)	1917	1917	100%	0.0	98.70%	BX817100.1
Arabidopsis thaliana Full-length cDNA Complete sequence from clone GSLTPGH55ZG06 of Hormone Treated Callus of strain col-0 of Arabidopsis thaliana (thale cress)	1456	1456	99%	0.0	91.02%	BX816562.1
Arabidopsis thaliana Full-length cDNA Complete sequence from clone GSLTFB34ZF07 of Flowers and buds of strain col-0 of Arabidopsis thaliana (thale cress)	1182	1182	71%	0.0	94.44%	BX813702.1
Arabidopsis thaliana genome assembly, chromosome: 1	1081	2022	100%	0.0	100.00%	LR215052.1
Arabidopsis thaliana chromosome 1 sequence	1081	2022	100%	0.0	100.00%	CP002684.1
Arabidopsis thaliana chromosome I BAC F5A8, complete sequence	1081	2022	100%	0.0	100.00%	AC004146.1
Arabidopsis thaliana ecotype Co-1 At1g67170 gene, partial cds	446	1343	66%	3e-123	100.00%	EU351156.1
Arabidopsis thaliana ecotype Cvi-0 At1g67170 gene, partial cds	446	1350	66%	3e-123	100.00%	EU351147.1
Arabidopsis thaliana ecotype Col-0 At1g67170 gene, partial cds	446	1348	66%	3e-123	100.00%	EU351145.1
Arabidopsis thaliana ecotype Li-3 At1g67170 gene, partial cds	444	1339	66%	1e-122	100.00%	EU351150.1
Arabidopsis thaliana ecotype Da (1)-12 At1g67170 gene, partial cds	444	1345	66%	1e-122	100.00%	EU351149.1

Figura 7. Resultados del algoritmo blast en línea.

Si observamos con detenimiento los dos primeros resultados, podemos ver que los alineamientos tienen el 100% de identidad con una proteína asociada a una membrana sarcolemal. Además podemos ver que el e-value es de 0, esto nos indica que la probabilidad de que las dos secuencias sean homologas es muy alta.

De esta forma podemos concluir que la secuencia que encontramos en el cromosoma 1 de la planta modelo *Arabidopsis thaliana*, con id "gi|29028877|gb|BT005883|U23535", tiene como función codificar una proteína asociada a las membranas celulares.