UnADM
UNIVERSIDAD ABIERTA Y A DISTANCIA DE MÉXICO

Programa de la asignatura:

Bioinformática

U2 Análisis computacional de secuencias de ADN

DCSBA

BIOTECNOLOGÍA

Índice

Presentación de la unidad

En la actualidad existe un gran cúmulo de información genética producto de la secuenciación de genes particulares de un organismo o de genomas completos. Pero, ¿cómo se puede interpretar toda esta información para que tenga alguna utilidad práctica en nuestros proyectos de investigación? La bioinformática representa una herramienta muy útil para el análisis de toda la información generada hasta el momento, es por ello que uno de los objetivos de esta unidad es introducirte en el uso de las principales bases de datos disponibles en la actualidad para que seas capaz de buscar, obtener y analizar datos de secuencias de ácidos nucleicos que te permitan responder preguntas específicas que pudieran presentarse en tu actividad profesional.

Así, durante esta unidad se abordarán aspectos teóricos sobre la búsqueda y análisis de secuencias de ácidos nucleicos, particularmente del alineamiento de secuencias como una herramienta para buscar similitudes entre las secuencias y con ello, poder inferir su función. Posteriormente, nos centraremos en la utilización de distintas bases de datos y software para obtener información diversa de secuencias de ácidos nucleicos y sus aplicaciones prácticas. Se indicará el uso de estos programas paso por paso, esperando que refuerces y practiques estos conocimientos al realizar las actividades planteadas. En esta unidad, se espera un compromiso mayor de tu parte por practicar el uso del software y en el análisis de los datos obtenidos.

Propósitos de la unidad

- Identificar las bases teóricas del alineamiento de secuencias.
- Interpretar correctamente los resultados de un alineamiento.
- Determinar características intrínsecas de una secuencia en particular por medio del empleo de software.
- Distinguir herramientas bioinformáticas para el diseño de cebadores y construcción de plásmidos que son útiles en la ingeniería genética.

Competencia específica

Interpretar el contenido de secuencias de nucleótidos para su utilización en procesos moleculares específicos mediante el uso de software especializado en el diseño de herramientas moleculares.

2.1. Búsqueda de secuencias de ADN

En la unidad pasada buscamos una secuencia de ADN en la base de datos del NCBI, sin embargo, existen otras bases de datos en las que podemos hacer una búsqueda y dependiendo de la base datos que utilicemos, nuestra secuencia, así como los resultados mostrados, puede presentar diversos formatos.

2.1.1. Formatos de secuencias

Un formato de secuencia define el diseño y contenido de un texto, que incluye la secuencia, en un archivo. Contiene palabras de texto que definen los campos utilizados en las bases de datos. Los campos incluyen la secuencia, el nombre del identificador de secuencia y número de acceso, entre otros. La mayoría de los formatos pueden verse en la pantalla y son imprimibles. Aunque en esta sección utilizaremos el formato **FASTA**, es importante que sepas que existen otros formatos de secuencias como son: **ASN.1, EMBL UniProt, GenBank/GenPept, NEXUS, PHYLIP y NBRF-PIR**. Dichos formatos representan distintas formas de obtener y observar la información.

FASTA: es el formato más usado, consiste en una descripción en la línea de cabecera comenzando con un símbolo “*mayor que*” (>) seguido por un código de identificación de la secuencia y frecuentemente algunas palabras que describen la secuencia, como el

organismo al que pertenece, la proteína que codifica, entre otros. La línea de cabecera es seguida por una o más líneas conteniendo la secuencia, ya sea de nucleótidos o de aminoácidos. Ejemplo de secuencia de ADN en este formato:

Formato FASTA

```
>gi|71660716|ref|XM_816981.1|Trypanosoma cruzi strain CL Brener heat shock protein 20 partial mRNA

ATGTGGGATCCGTTTCGCGATGTGGAGCGCCTTCTCAATCGCATGCAGTCCGTCACCGGCACGAGTTTTCTCTCCACATCCGCT
CGTGGATCATGGGTGCCGGCGATGGACATTGTCGAGAGGGAGGACAGCTACAAGATTCTTGCTGACTTGCCCGGCATGAGCC
GCAACGACGTCTCTGTGGAGATTGAGGGCAGCCAACTGTGCATTGGAGGCAACCGCAAGTCCATGCTGAGTGAAGAGGAACAC
AAAAACGTTGTGATGGCAGAGCGCGGTTCCGGGAGATTTGAGCGCTGCGTGCGACTTCCCTCACCCCTTGAAGAGGGCAGCGT
GAAGGCCAGCCTGCGTGATTCGGTGCTGCTCGTGGAGGTGAAGAAAGTGACGGACGCCGTGCGGAAGCGCTCGGGGATCTCC
GTGAAGATCAATTAG
```

El formato FASTA es usado por las principales bases de datos:

EMBL	http://www.embl.de/
GenBank	http://www.ncbi.nlm.nih.gov/genbank/
SwissProt	http://www.expasy.org/
PIR	http://pir.georgetown.edu/

2.1.2. El caso específico de algunas bases de datos

Base de datos EMBL-EBI

Para la búsqueda de secuencias de nucleótidos, una de las bases de datos más utilizada es EMBL-EBI. Esta base de datos utiliza el Archivo Europeo de Nucleótidos o ENA, por sus siglas en inglés (***E****uropean* ***N****ucleotide* ***A****rchive*). ENA captura y presenta la información relacionada a trabajos experimentales que se basan en secuencias de nucleótidos. Un flujo de trabajo típico incluye el aislamiento y la preparación del material para secuenciación, un corrimiento en una máquina secuenciadora en la cual los datos de la secuencia son producidos y finalmente la realización de un análisis bioinformático. ENA registra esta información en un modelo de datos que expone la información de entrada (muestra, diseño experimental, configuración de la máquina), datos de la máquina de salida (trazo de la secuencia, lecturas y puntuaciones de calidad) y posteriormente la interpretación de la información (montaje, mapeo, anotaciones funcionales). En el siguiente enlace podrás entrar directamente a la página de ENA, realizar búsquedas y obtener más información: http://www.ebi.ac.uk/ena/about/about.

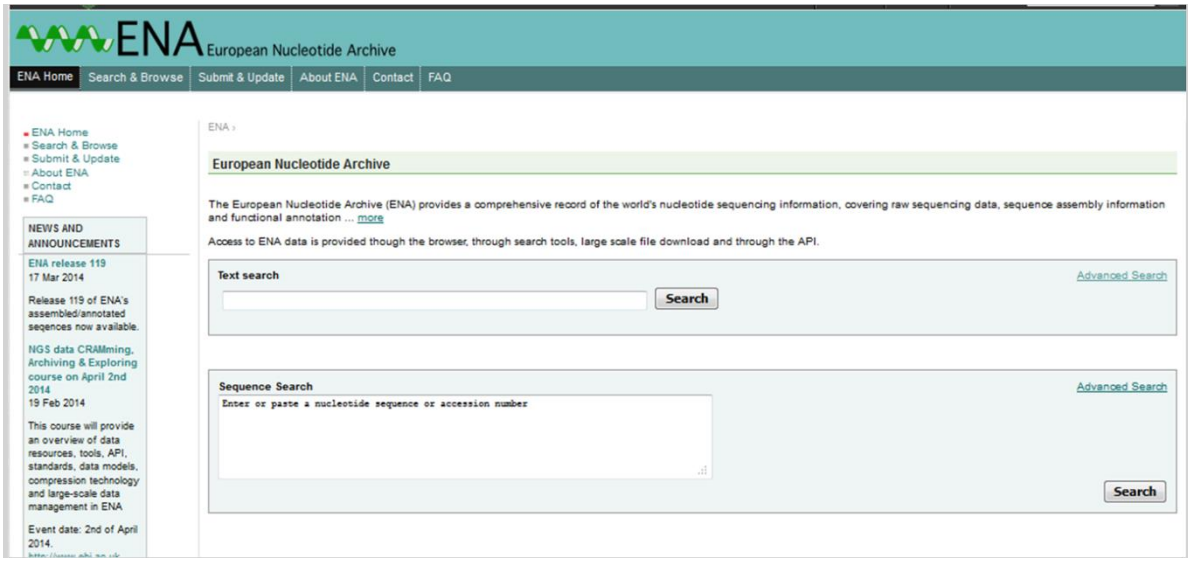
Los datos que llegan a ENA provienen de diversas fuentes. Estos datos incluyen envío de datos en bruto, secuencias ensambladas y notaciones a pequeña escala de secuencias.

Los datos alimentan a los principales centros europeos de secuenciación y permiten un intercambio global rutinario y comprensivo con los socios en la Colaboración Internacional de Bases de Datos de Secuencias de Nucleótidos o INSDC, por sus siglas en inglés (***I****nternational* ***N****ucleotide* ***S****equence* ***D****atabase* ***C****olaboration*). El suministro de los datos de secuencias de nucleótidos a ENA o sus socios INSDC se ha convertido en un paso central y obligatorio en la difusión de los resultados de la investigación a la comunidad científica. ENA trabaja con los editores de la literatura científica y los organismos de financiamiento para asegurar el cumplimiento de estos principios y para proporcionar sistemas de envío óptimos y herramientas de acceso a datos que funcionen a la perfección con la literatura publicada.

ENA se compone de un número de distintas clases de datos organizados en tres niveles. Cada clase tiene sus propios formatos y estándares. Aunque ENA tiene al menos 30 años de historia, los datos y servicios están en constante cambio con el fin de reflejar el crecimiento en volúmenes de datos. Asimismo, la tecnología de secuenciación que usan cada vez es mejor, ya que como parte del esfuerzo global para mejorar el acceso y uso de los datos de secuenciación de nucleótidos, colaboran ampliamente en el desarrollo de los servicios y tecnologías y también en las actividades de normalizar los datos .

A pesar de que en este curso no utilizaremos esta base de datos, es importante que al menos sepas reconocer sus características principales. Para acceder a ella, da click en la siguiente dirección de internet: http://www.ebi.ac.uk/ena/

Una ventana como la siguiente es la que aparecerá:



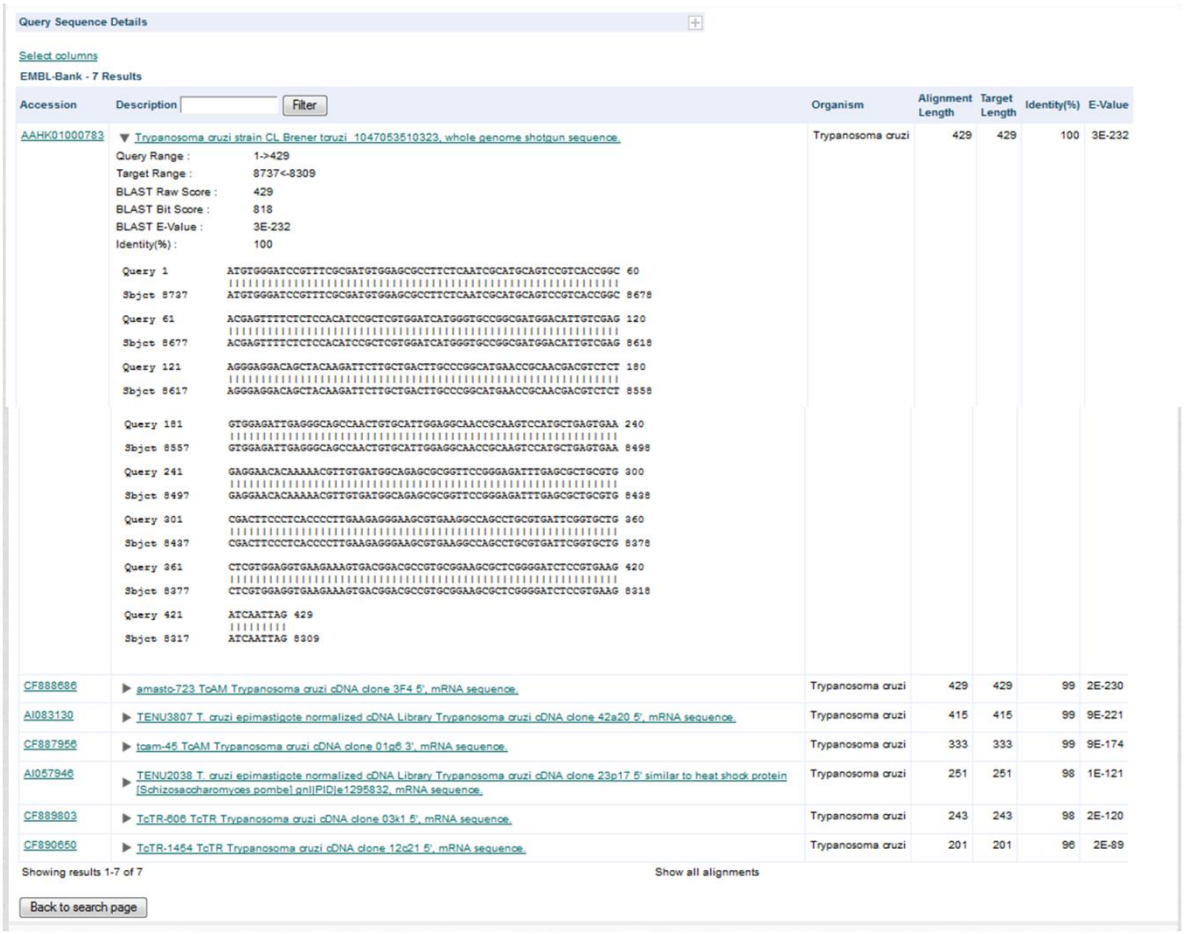
Los datos ENA pueden ser buscados y recuperados interactivamente y mediante programación, utilizando el navegador de ENA.

Clases de datos de ENA y los formatos

Texto libre: búsqueda de texto libre. Se proporciona desde la página principal de ENA a través de la búsqueda disponible en la parte superior de todas las páginas web EMBL-EBI. Las opciones de búsqueda avanzada están disponibles en la página de ENA (búsqueda avanzada).

Búsqueda de similitud de secuencias: se proporciona en la página principal ENA. Opciones de búsqueda avanzada están disponibles en la página de búsqueda de secuencias. La interfaz principal de programación para acceder a los datos es a través del navegador ENA. El navegador ENA está diseñado para acceder a través de URLs REST de acceso para recuperación de datos mediante programación y una variedad de formatos.

Una vez que introdujiste la secuencia de búsqueda, los resultados pueden verse como se muestra en el siguiente ejemplo:



Descarga de datos ENA: La principal herramienta para descargar datos ENA es su navegador. Este puede ser utilizado interactivamente y mediante programación a través de URLs REST. Todos los datos ENA incluyen un ensamblaje y anotación de secuencias que están disponibles para descargarlos a través de este navegador.

Envío de listas: Todos los usuarios son alentados a suscribirse a "*ena-announce*", para recibir los datos vía mail.

Base de datos NCBI

En la base de datos NCBI (con la cual ya te familiarizaste en la unidad pasada), se pueden hacer diversos alineamientos mediante el programa BLAST, este compara nuestra secuencia de consulta con todas las secuencias de la base de datos para encontrar regiones de similitud. En la sección siguiente nos adentraremos al uso de esta herramienta bioinformática.

2.1.3. Parámetros de búsqueda

Los parámetros son tipos de variables usadas para personalizar el análisis de secuencias, estas variables influyen en el análisis y de esta manera intervienen indirectamente en el resultado. En los alineamientos, la selección de parámetros ajusta los posibles resultados de la búsqueda, a lo que el investigador le interesa. Los parámetros por diseño que utiliza BLAST son óptimos, sin embargo en ocasiones se deben cambiar para obtener resultados más específicos o cuando se sospecha de homología entre especies muy divergentes, se deben “agrandar” o “achicar” los filtros para poder detectar adecuadamente.

En el caso de BLAST, los principales parámetros son: **descripción, formato de secuencia, expectativa, secuencias blanco máximas, alineación, nombre del organismo, clasificación taxonómica y filtro de baja complejidad, NCBI-gi, consulta de código genético** y **descripción gráfica**, entre otros (http://kinase.com/blast/docs/newoptions.html, http://viroblast.dbi.udel.edu/CHO/parameters.php).

Ejemplos de cómo se despliegan las ventanas donde uno puede cambiar los parámetros de este programa son mostrados en las imágenes siguientes, localiza cada uno de los parámetros. Cabe mencionar que si además colocas el cursor en el ícono del signo de interrogación, obtendrás una breve descripción de ese parámetro.

Descripción: Restringe el número de descripciones cortas reportadas que coinciden, el

límite por defecto es de 100.

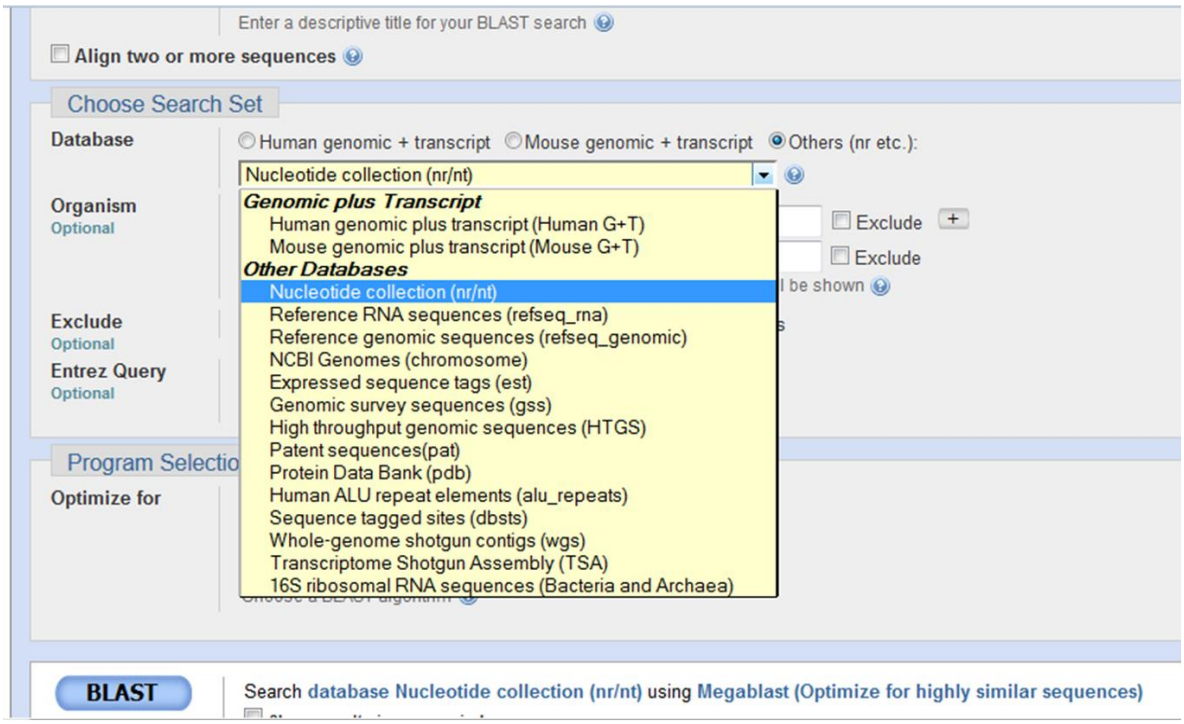
Formato de secuencia: Un formato de secuencia común para la mayoría de secuencias y de alineación y editores.

Expectativa: El umbral de significancia estadística coincide con la secuencia de la base de datos, el valor por defecto es 10, así que espera encontrar 10 coincidencias por casualidad.

Secuencias blanco máximas: Mantiene un número máximo de descripciones y alineaciones. El valor por defecto es 50.

Alineación: Restringe las secuencias de la base de datos con el número especificado para obtener pares de segmentos con la mayor puntuación, el límite por defecto es 100.

Nombre del organismo: Se introduce el nombre de la especie, por ejemplo: *Homo sapiens*. En el menú despegable se muestran las especies más comunes.



Clasificación taxonómica: Se introduce cualquiera de los nombres taxonómicos utilizado por NCBI:

Archaea
Bacteria

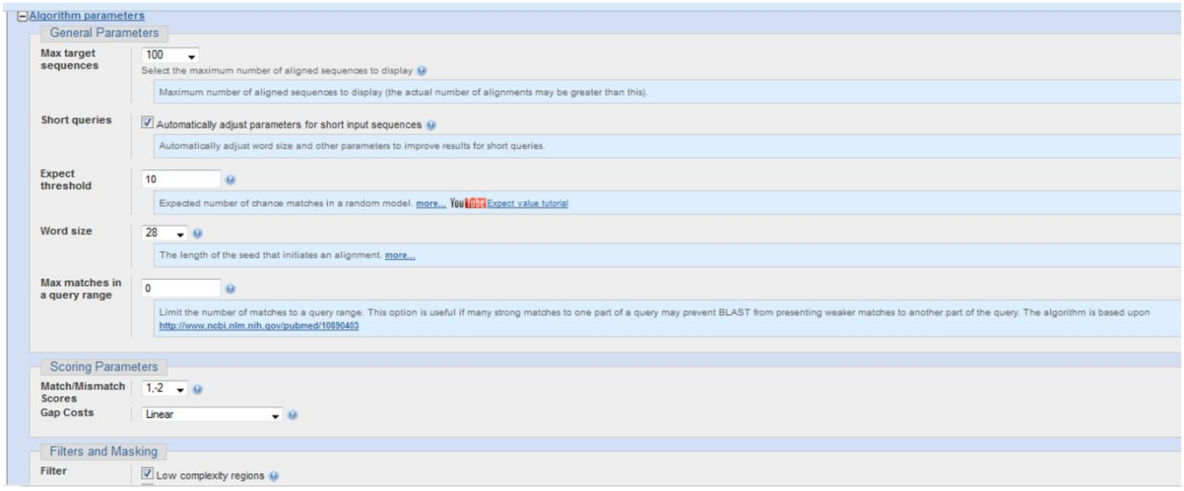
Eucarionte
Embryophyta
Hongos
Metazoos
Vertebrata
Mammalia
Rodentia
Primates

Filtro de baja complejidad: La filtración puede eliminar la significancia estadística, sin embargo no elimina informes biológicos, lo que permite tener regiones biológicas más interesantes en la secuencia de consulta para que esta coincida específicamente contra la base de datos.

NCBI-gi: Causa identificadores gi NCBI para mostrarse en la salida, además del número de acceso y nombre del *locus*.

Consulta de código genético: Utiliza el código genético en la traducción BLASTX de consulta.

Descripción gráfica: Se muestran las secuencias de la base de datos alineadas con la secuencia de consulta. La puntuación de cada alineación se representa con cada uno de los 5 colores diferentes usados y cada alineamiento se conecta por una línea, al hacer click en una secuencia lleva al usuario a las alineaciones asociadas.



2.2. Alineamiento de secuencias

Una forma de representar la comparación entre dos o más secuencias, ya sea de ADN, ARN, o aminoácidos es el alineamiento de secuencias. En esta unidad revisaremos los fundamentos en los que se basan los programas computacionales para realizar estos análisis. Posteriormente nos enfocaremos en los alineamientos de secuencias de nucleótidos y en la próxima unidad retomaremos estos conceptos para el caso de secuencias de aminoácidos.

2.2.1. Concepto de alineamiento de secuencias

El alineamiento de secuencias se refiere al procedimiento de comparar dos o más secuencias, buscando caracteres individuales que se encuentren en el mismo orden en las secuencias comparadas. Los caracteres pueden ser ácidos nucleicos o aminoácidos. Así, un alineamiento involucra establecer correspondencias entre nucleótidos o codones de ADN o ARN o entre aminoácidos que forman una secuencia lineal, encontrando aquéllos que son idénticos (Baxevanis & Ouellette, 2001). Un ejemplo es mostrado en la Figura 1, donde se señalan, con una línea vertical, aquéllos nucleótidos que son idénticos en las dos secuencias.

Figura 1. Alineamiento de dos secuencias. El principio teórico del alineamiento de dos secuencias se basa en encontrar coincidencias entre éstas. En este caso, las coincidencias son señaladas por una línea vertical que comunica aquéllos nucleótidos idénticos entre una y otra secuencia en el orden en el que se encuentran dispuestos. Debido a que de 10 nucleótidos que conforman a la secuencia, 6 son idénticos, el porcentaje de similitud entre ellas es del 60%.

Una categorización de los análisis y búsqueda de secuencias, se realiza por el porcentaje de elementos de la que toman para el análisis, de acuerdo a esto, existen 2 tipos de alineamiento: global y local. El **alineamiento global** busca en toda la secuencia e introduce espacios o "*gaps*" donde no puede encontrar una coincidencia o apareamiento adecuado entre las secuencias (Figura 2A). En cambio, el **alineamiento local** busca sólo en regiones donde hay un apareamiento significativo (Figura 2B), éste se recomienda cuando se alinean secuencias con poca similitud, pues de esta forma sólo buscará en las regiones donde es más probable encuentre una identidad (Mount, 2001).

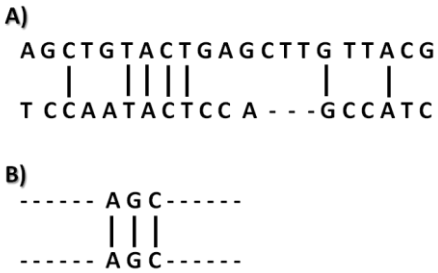


Figura 2. Tipos de alineamiento entre dos secuencias. De acuerdo a la región de comparación que abarque el alineamiento, éste puede ser clasificado en alineamiento global (A), donde se incluye toda la secuencia en su totalidad, o local (B), donde sólo se compara una parte de la secuencia.

2.2.2. Fundamentos teóricos del alineamiento de secuencias

La lógica que se sigue para llevar a cabo los análisis por alineamientos es comparar las diferencias entre dos o más secuencias, con el propósito de encontrar similitudes que representen relevancia a nivel biológico. Con la secuenciación de genes, proteínas y genomas completos, las bases de datos contienen una gran cantidad de información, con lo cual, hoy en día se puede contrastar una secuencia contra miles de secuencias conocidas ya descritas. El procedimiento está basado en distintos algoritmos con los cuales se buscan las posibles coincidencias entre éstas.

El análisis de secuencias se basa en fundamentos de estadística, que como veremos más adelante tiene ciertas particularidades, pero para entender las aproximaciones estadísticas es necesario comprender su fundamento científico, en este caso la probabilidad. La probabilidad en su definición intuitiva, acerca su significado al de frecuencia p(A) = #A/T donde p(A) es la probabilidad de A, #A es el número de eventos del tipo A y T son los eventos totales, así lo que se mide es la frecuencia de los eventos de tipo A. La ciencia de la probabilidad sirve para hacer inferencias a partir de la información que se tiene y aunque dar los detalles más técnicos sobre la probabilidad está fuera de los alcances de esta unidad, es necesario saber que las probabilidades ocurren en el rango de 0 a 1, por lo cual son siempre positivas. La probabilidad de un evento entre más cercano a 0 sea el evento es menos probable. Como mencionamos anteriormente, existe la probabilidad de que las similitudes entre secuencias se deban al azar y que necesitábamos elementos metodológicos y teóricos para determinar con argumentos científicos la existencia de homología entre secuencias.

En concreto las aproximaciones estadísticas para el análisis de secuencias se basan en crear un puntaje, calculado a partir de la distribución de puntajes generados al contrastar muchas secuencias no relacionadas, de esta forma entre mayor es la cantidad de secuencias con las que se construyen los puntajes, más confiable es el análisis.

Uno de los métodos usados actualmente, requiere del ajuste matemático de la cola de distribución de puntajes (es decir el área de la distribución donde se encuentran los valores extremos), realizados al contrastar secuencias no relacionadas elegidas al azar. Bajo la misma lógica de encontrar el valor máximo de un atributo, en el análisis de comparación de secuencias se busca el valor máximo de alineamiento. Ahora que ya hemos visto como se construyen los puntajes que dan el sustrato metodológico para las comparaciones estadísticas de secuencias, veamos el elemento teórico detrás de estas comparaciones, es decir la prueba de hipótesis.

Para poder tomar la decisión de homología se recurre a la estadística y el objetivo es probar una de dos hipótesis. La hipótesis nula, las similitudes entre 2 o más secuencias se deben al azar y la hipótesis alternativa, las similitudes entre 2 o más secuencias no se deben al azar. Para ello se determina la significancia estadística que se refiere a la probabilidad de que el resultado no sea debido al azar y por lo tanto depende de la probabilidad de distribución y de sus parámetros, que como más adelante veremos afectan la probabilidad de que el resultado sea un falso positivo o artefacto (Schuler, 2001). Los algoritmos que se han desarrollado para estos fines integran los análisis estadísticos y es necesario poner atención a los resultados para poder dar un juicio de evaluación de la similitud obtenida a partir de los alineamientos de secuencias (Schuler, 2001; Madden, 2003).

Usando el teorema central del límite que dice que la media de muchas muestras independientes tiende a la media poblacional, se puede establecer un criterio de significancia con fundamentos teórico-empíricos. Es recomendable contrastar el porcentaje de similitud contra muchas secuencias, lo cual realza la importancia de tener disponibles gran cantidad de secuencias para hacer comparaciones, si se obtiene un valor $p< 0.01$ y se hicieron 100 comparaciones, se puede afirmar que la probabilidad de que las similitudes se deban al azar son de 1 en 100 (Madden, 2003).

Dependiendo del software que se use para los análisis es como se presentarán los resultados, es por esto que es mejor saber identificar qué parámetros son importantes para la comparación de secuencias y la lógica detrás de cada forma de análisis.

2.2.3. Similitud, identidad y homología

Para entender los análisis de secuencias, se debe comprender los conceptos de **similitud**, **identidad** y **homología**.

Al encontrar coincidencias entre los caracteres de dos secuencias alineadas, se dice que las secuencias son similares. ¿Pero qué tan similares son las secuencias analizadas? El concepto de **similitud** se refiere a la cantidad medible de semejanzas entre una

secuencia y otra, que puede ser expresada bajo cualquier unidad cuantitativa, normalmente se expresa como porcentaje de **identidad**.

El **porcentaje de identidad** es la cantidad de nucleótidos o aminoácidos que son iguales entre las secuencias. Para determinarlo, simplemente se calcula el porcentaje que constituyen los caracteres idénticos con respecto a la longitud de la secuencia completa de acuerdo a la siguiente fórmula (Claverie & Notredame, 2007):

$$\frac{Número\ de\ nucleótidos\ o\ aminoácidos\ identicos}{Número\ de\ nucleótidos\ o\ aminoácidos\ totales\ que\ componen\ la\ secuencia} x\ 100$$

En el ejemplo de la Figura 1, la secuencia superior tiene 10 nucleótidos, de los cuales 6 son idénticos con la secuencia inferior, así, calculando el porcentaje de identidad de la secuencia superior con respecto a la inferior, tenemos que esos 6 nucleótidos representan un 60% de identidad, es decir, 60% de los nucleótidos son iguales.

En el caso de que quisiéramos conocer cuántos nucleótidos de la secuencia inferior son iguales con respecto a la secuencia superior, sería el mismo resultado, ya que también está constituida por 10 nucleótidos y 6 son iguales en comparación con la secuencias superior.

En un caso hipotético que la secuencia superior se conformara de 16 nucleótidos y la inferior de 10 y de los 16 nucleótidos de la superior, 8 coincidieran con la inferior, el porcentaje de identidad de la secuencia superior con la inferior sería del 50 %, en cambio, el porcentaje de identidad de la secuencia inferior con la superior sería del 80 %.

El objetivo principal del alineamiento de secuencias es determinar si el grado de similitud es tal, que pueda inferirse que dichas secuencias sean homólogas y con ello, encontrar información sobre la probable función de dicha secuencia. Dos secuencias son homólogas cuando provienen de un ancestro común, es decir, en términos evolutivos, se originaron de la misma secuencia ancestral.

La homología es un atributo discreto, es decir, dos secuencias son o no son homólogas, en consecuencia resulta erróneo expresar un porcentaje o nivel de homología entre las secuencias evaluadas (Baxevanis & Ouellette, 2001). Ahora bien, ¿qué tan similares deben ser dos secuencias para ser consideradas homólogas?
En términos generales, se acepta que, respecto a las secuencias de nucleótidos, un porcentaje de identidad del 70 % o mayor indica que los genes son homólogos (Claverie & Notredame, 2007). Para el caso de secuencias de aminoácidos, dos proteínas son homólogas si tienen un 25 % o más de identidad, lo cual puede considerarse una evidencia de que esas secuencias tienen un ancestro común (Brown, 2000).

2.2.4. Alineamiento de un par de secuencias

El alineamiento entre un par de secuencias, se refiere, como su nombre lo indica, al análisis por alineamiento entre sólo dos secuencias de ácidos nucleicos o de aminoácidos (Baxevanis & Ouellette, 2001).

En la Figura 3, se muestra un alineamiento de dos secuencias. En este alineamiento global puedes observar, entre otros datos, el porcentaje de identidad y el número de gaps generados, encerrados en un círculo rojo. La secuencia de búsqueda o consulta es aquélla de la cual nosotros queremos obtener información y se denomina **"Query"**, la **"Sbjct"** es la secuencia contra la cual se comparó. De aquí en adelante llamaremos secuencia de consulta a la secuencia Query. Así, en este alineamiento, la secuencia de consulta contenía 380 nucleótidos, de los cuales 372 coincidieron con la otra secuencia, produciendo un porcentaje de identidad del 98%. Por otro lado, en este alineamiento no fue necesario incluir gaps para que las secuencias quedaran alineadas (Figura 3).

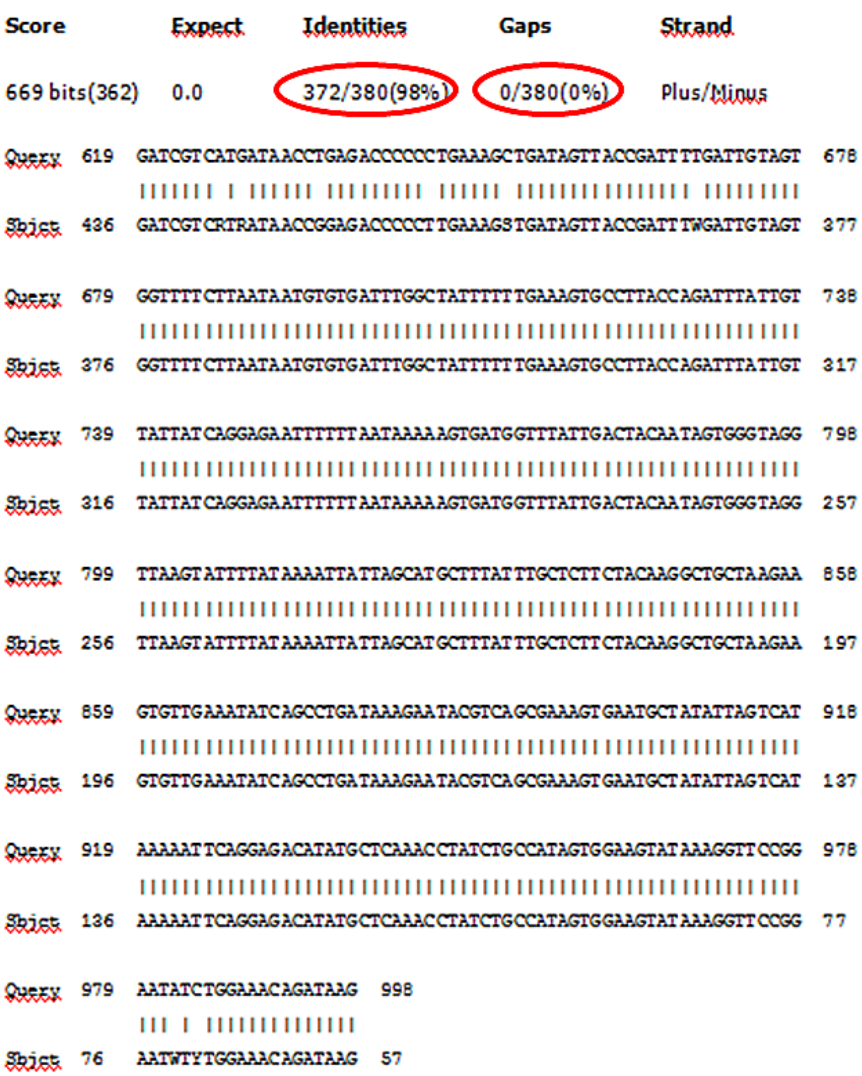


Figura 3. Alineamiento entre dos secuencias de nucleótidos. Cada coincidencia entre los nucleótidos se señala con una línea vertical. Se destaca el porcentaje de identidad y el número de gaps encontrados.

Los datos de secuencias genéticas de los que disponemos hoy en día son tan vastos, que la comparación de nuestra secuencia de consulta con las secuencias disponibles en las bases de datos involucra la comparación con miles de secuencias. Además, gracias al perfeccionamiento de los métodos de secuenciación, podemos tener secuencias con una longitud de varios miles de pares de bases. Por lo anterior, resultaría muy complicado alinear estas secuencias a mano, por lo que en la actualidad existe software especializado, del cual podemos hacer uso de manera gratuita en la red para poder realizar estos alineamientos de manera más rápida y eficiente.

Existen diversos programas disponibles para realizar alineamientos de secuencias, tanto de ácidos nucleicos como de aminoácidos. En esta sección aprenderás a realizar alineamientos de un par de secuencias de nucleótidos utilizando BLAST.

BLAST (por sus siglas en inglés, ***B****asic* ***L****ocal* ***A****lignment* ***S****earch* ***T****ool*) es una herramienta básica de búsqueda para alineamientos locales, mantenida por el NCBI, es la más usada para comparar regiones de similitud entre secuencias. Su programa compara secuencias de ácidos nucleicos o de aminoácidos de consulta con secuencias de las bases de datos y calcula la significancia estadística de los nucleótidos o aminoácidos que coinciden. Debido a los resultados que arroja, BLAST puede utilizarse para inferir las relaciones funcionales y evolutivas entre secuencias, así como ayudar a identificar a los miembros de las familias de genes (Baxevanis & Ouellette, 2001). La importancia de herramientas como BLAST se incrementa en la medida en que se secuencian más genomas o se describe la secuencia de más proteínas.

Existen cinco programas de la herramienta BLAST con los que pueden realizarse búsquedas o alineamientos, cuyas características principales son descritas en la Tabla 1. La diferencia entre estos radica en dos factores: el tipo de la secuencia que se quiere contrastar y el tipo de secuencias de la base de datos contra la que se contrastará nuestra secuencia de interés (Orengo, Jones, Thornton, 2003).

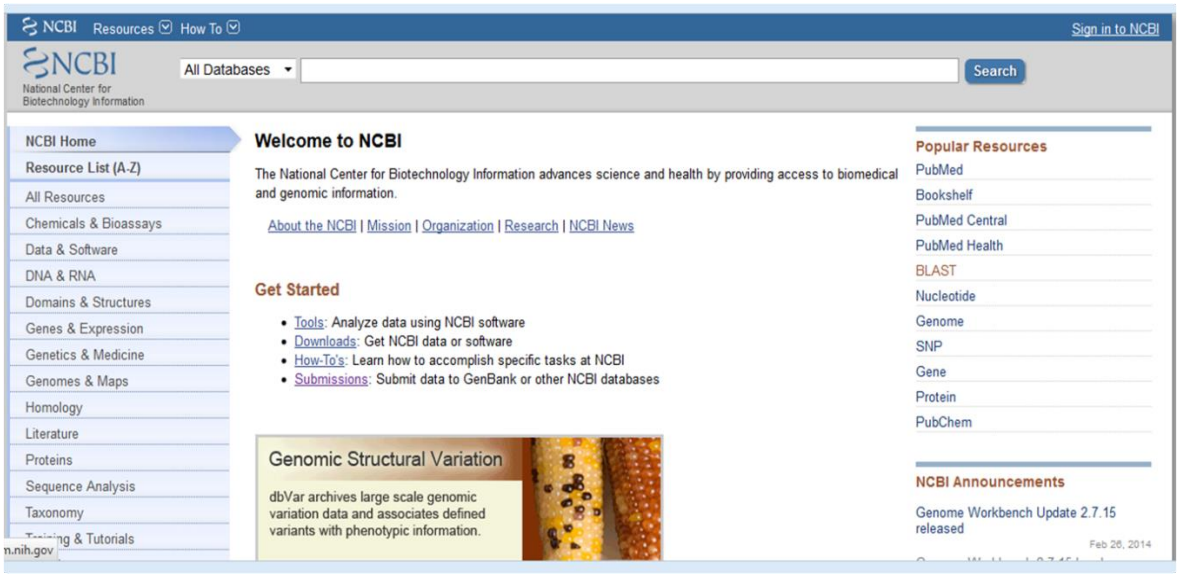
Tabla 1. PROGRAMAS DE BLAST PARA COMPARAR O ALINEAR SECUENCIAS

Programa	**Secuencia de consulta**	**Base de datos de búsqueda**	**Uso**
blastn	Nucleótidos	Nucleótidos	Compara secuencias de consulta de nucleótidos con secuencias de nucleótidos de la base de datos
blastp	Proteínas	Proteínas	Busca proteínas de la base de datos utilizando una secuencia de consulta de aminoácidos
blastx	Nucleótidos traducidos	Proteínas	Busca proteínas en la base de datos utilizando una secuencia de consulta de nucleótidos traducida
tblastn	Proteína	Nucleótidos traducidos	Busca nucleótidos traducidos de la base de datos usando una secuencia de consulta de aminoácidos

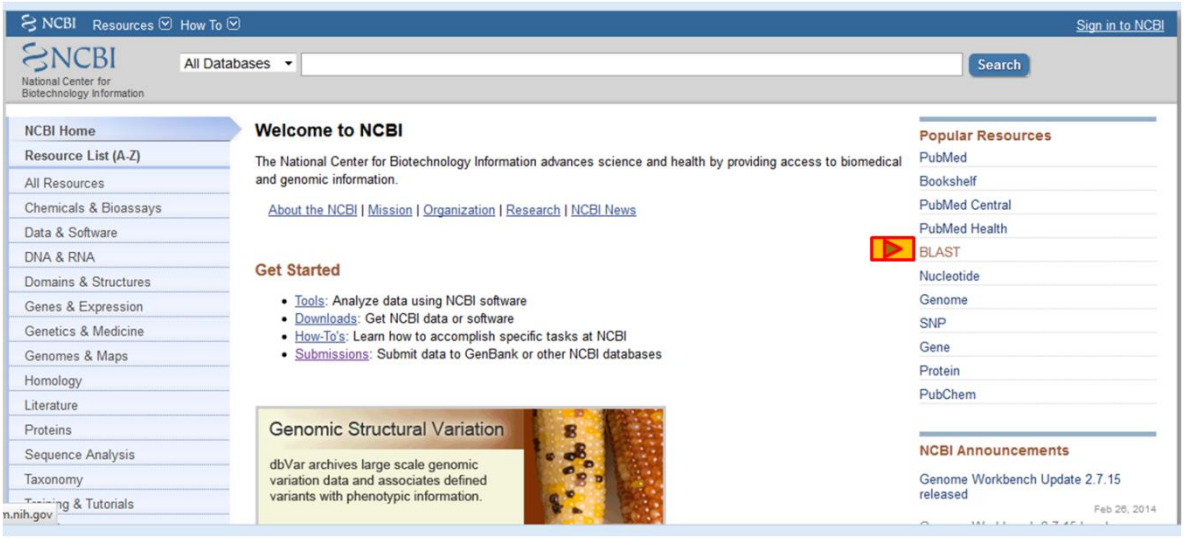
tblastx	Nucleótidos traducidos	Nucleótidos traducidos	Busca una secuencia traducida de nucleótidos usando un secuencia de consulta de nucleótidos

Para utilizar el programa BLAST, lo primero que debemos tener es una secuencia de consulta de la cual queremos obtener información. Una vez que le proporcionemos nuestra secuencia de consulta, el programa buscará las coincidencias que existan entre nuestra secuencia de consulta y las miles de secuencias depositas en esa base de datos. Es de suma importancia definir la naturaleza de nuestra secuencia, es decir, si se trata de nucleótidos o de aminoácidos, pues de lo contario, los resultados arrojados no serán correctos o simplemente el programa no correrá. De esta manera, si tenemos una secuencia de consulta conformada por nucleótidos, podemos pedirle al programa que la compare con una base de datos conformada por secuencias de nucleótidos, en cuyo caso, los resultados arrojados nos proporcionarán información sobre el gen al que pertenece nuestra secuencia y en qué otros organismos se encuentra. Cabe mencionar que en este caso, existe la posibilidad de que nuestra secuencia de consulta no codifique para ninguna proteína, por esta razón, la comparación con la base de datos de secuencias de nucleótidos es importante. Por otro lado, en el caso de que queramos conocer si nuestra secuencia de consulta contiene información genética expresada, podemos pedirle al programa que la secuencia de consulta conformada por nucleótidos la traduzca primero a aminoácidos, para posteriormente compararla con bases de datos de secuencias de aminoácidos o de nucleótidos traducidos. En este caso, podremos saber si nuestra secuencia de consulta, codifica alguna proteína. Las otras opciones son utilizadas cuando tenemos una secuencia de consulta conformada por aminoácidos, cuyas características se detallaran en la siguiente Unidad.

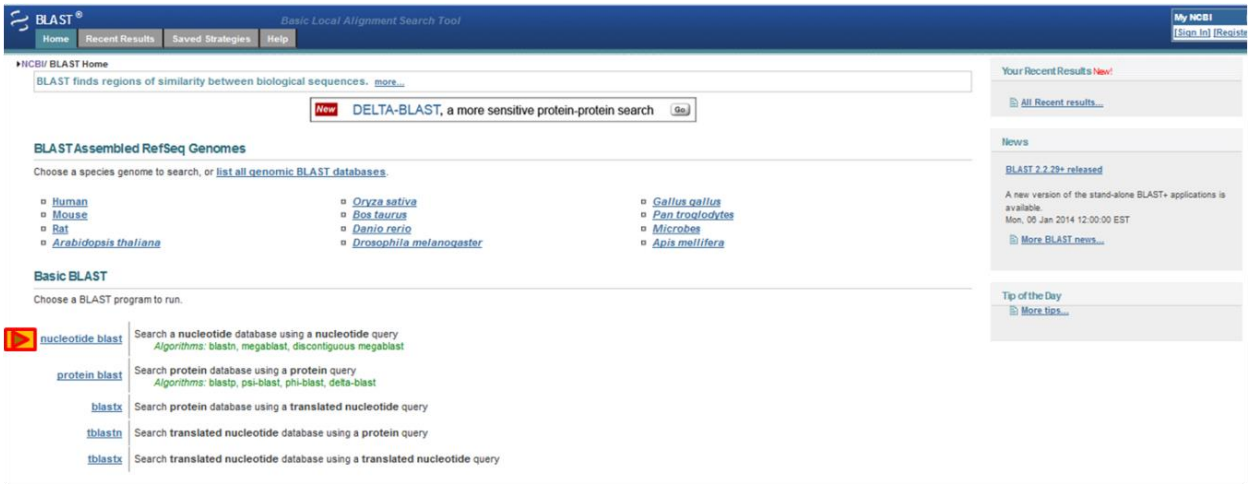
Ahora es el momento de empezar a utilizar la herramienta BLAST para realizar alineamientos de secuencias. Para ejecutar un alineamiento de un par de secuencias en BLAST, lo primero que tienes que hacer es acceder a la página principal del NCBI, como lo hiciste en la unidad pasada, en la siguiente dirección electrónica:
http://www.ncbi.nlm.nih.gov/



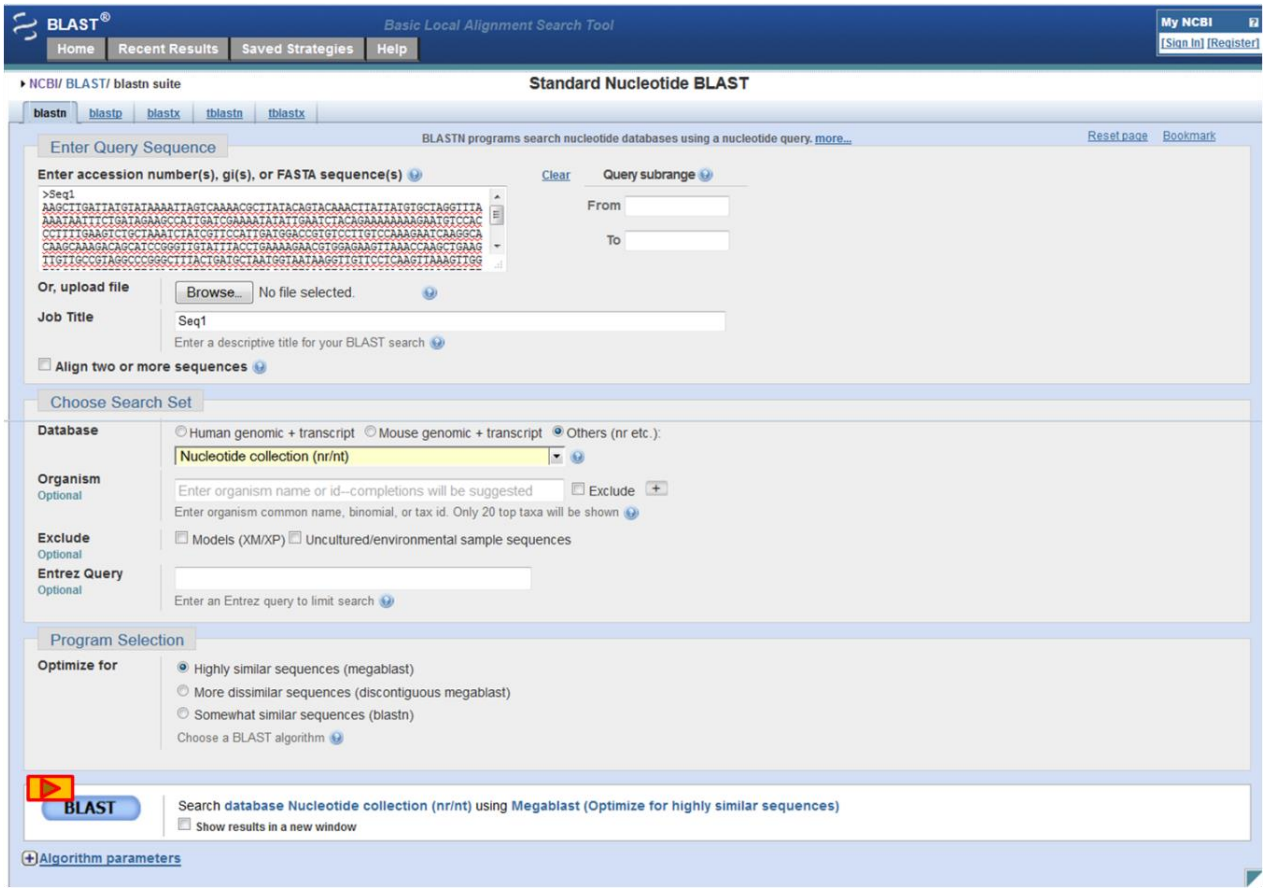
Una vez que accediste a la página del NCBI, del lado derecho de la pantalla selecciona la herramienta **"BLAST"**.



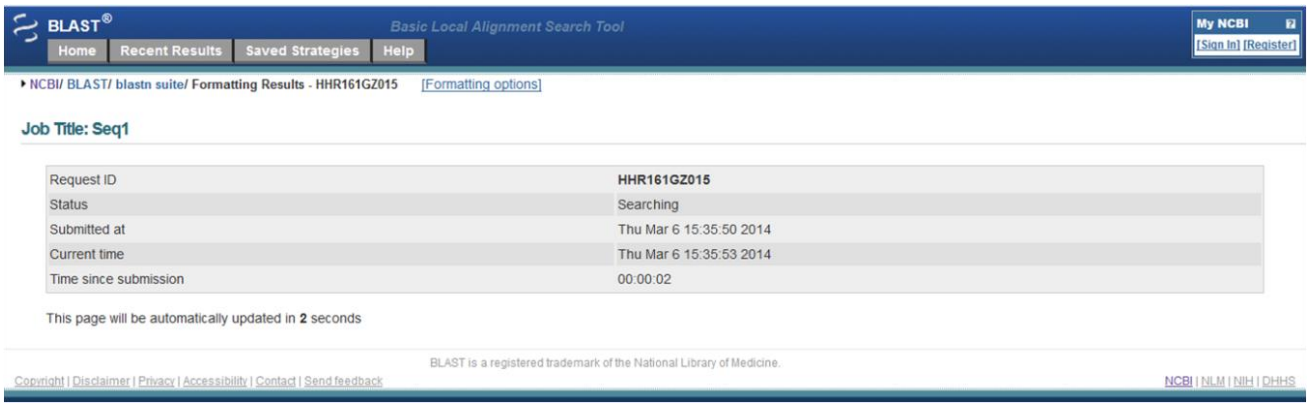
Ahora puedes ver en tu pantalla los diferentes programas de BLAST, cuyas características explicamos anteriormente y que se basan en el tipo de secuencia que introducirás y los resultados que esperas. Debido a que por el momento nuestro objetivo es introducir una secuencia de nucleótidos y alinearla contra una base de datos de nucleótidos, selecciona la herramienta "**nucleotide blast"**, que es lo mismo que **blastn**.



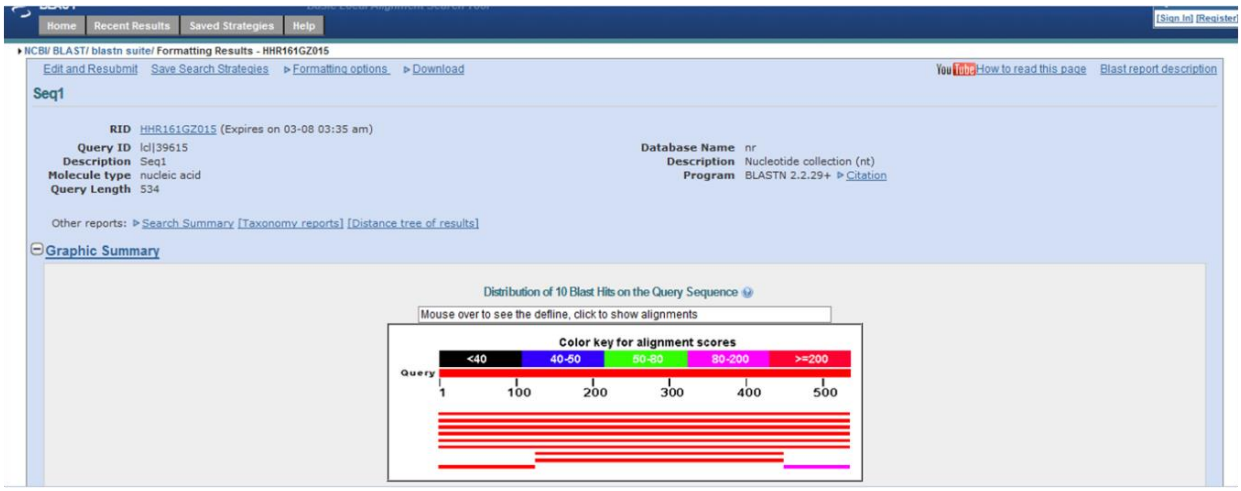
Acto seguido, se desplegará una ventana con el nombre "**Enter Query Sequence"**, donde puedes introducir tu secuencia o secuencias de nucleótidos, en caso de tener más de una, de las que quieres obtener información. Existen varias alternativas para introducir una secuencia en esta ventana. Si tu secuencia ya está depositada en una base de datos y conoces su **número de acceso** (*accession number*) o su **gI** (estos términos se definieron en la unidad pasada), puedes anotar directamente esos números para tener acceso a la secuencia deseada. Si no es el caso, y tu secuencia no está depositada en ninguna base de datos y la acabas de obtener, por ejemplo, como resultado de una clonación, o si no conoces ni el número de acceso ni el gI, entonces introducirás tu secuencia en formato FASTA, que ya se explicó anteriormente por qué se caracteriza. Puedes introducir tu secuencia pegándola directamente en la ventana o seleccionándola de un archivo en tu computadora. Debajo de la ventana donde pegaste tu secuencia podrás encontrar un apartado donde puedes poner el título a tu trabajo (*Job Tittle*). En la ventana "**Choose Search Set*",*** en el apartado **"Database",** escoge la opción **"Nucleotide collection"**, ya que queremos comparar nuestra secuencia contra una base de datos de nucleótidos. Finalmente, elige la opción "**BLAST"** para que empiece la búsqueda.



Una vez que inició la búsqueda, aparecerá una nueva ventana donde se indica que tu trabajo está siendo procesado, indicando el status, la fecha de búsqueda, así como el tiempo que ha transcurrido desde que solicitaste la búsqueda.



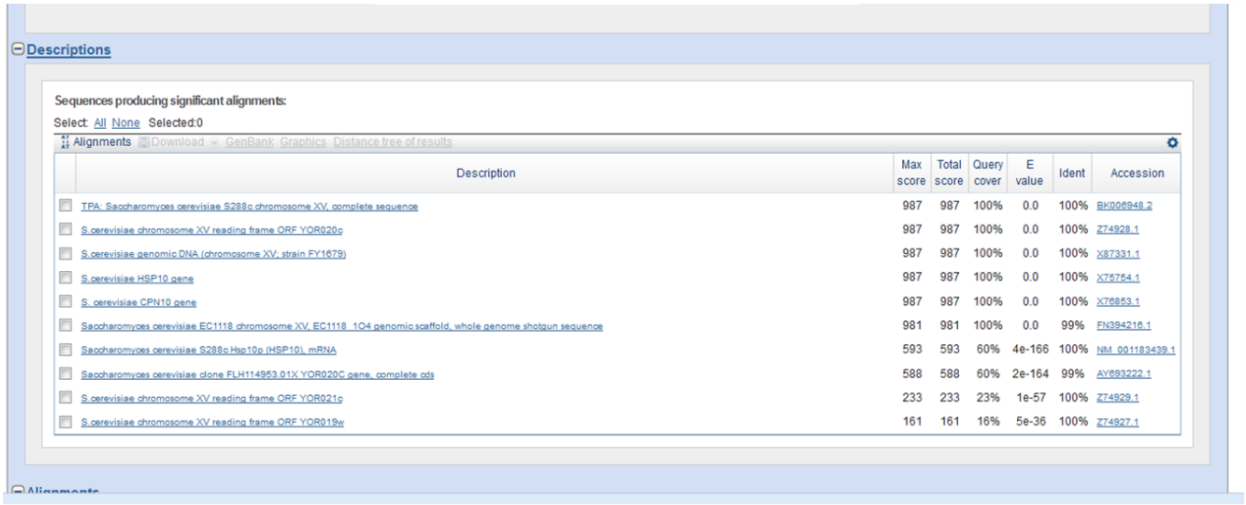
Una vez que termina la búsqueda, aparecerá una nueva ventana con los resultados obtenidos. En esta ventana se mostrarán datos como el nombre del trabajo solicitado, el tipo de molécula de que se trata, en este caso, ácidos nucleicos, la longitud de la secuencia, entre otros. Con respecto a los resultados, aparecerá un recuadro de colores donde se indica la longitud de tu secuencia de consulta (*Query*). Enseguida se muestra con líneas rojas, la longitud de coincidencia que abarcaron aquéllas secuencias que mostraron una identidad con tu secuencia de consulta y en orden descendente, es decir, aquéllas que mostraron la mayor identidad se colocan en la parte superior y van descendiendo conforme mostraron menor identidad.



Si te desplazas hacia abajo de la ventana, podrás encontrar un cúmulo de información referente a las secuencias con las que tu secuencia de consulta mostró una identidad. Así, en la primera columna se encuentra el nombre del organismo y el nombre del gen con el que alineó tu secuencia de búsqueda, además, posee un vínculo para acceder a dicha secuencia, donde podrás encontrar mayor cantidad de información sobre ésta. En las siguientes dos columnas se muestran los valores del "***Score***" (*Max score* y *Total score,* cuyos valores son iguales*),* el cual es una medida de la significancia estadística del alineamiento. Mientras más alto sea dicho valor, significa que más parecidas son las dos secuencias entre ellas. En general, un valor de score menor a 50 no es confiable. En la siguiente columna (*Query cover*) se encuentra el porcentaje de nucleótidos de tu secuencia de consulta que pudo ser alineada con la secuencia indicada, es decir, cuántos de los nucleótidos que conforman tu secuencia pudieron cotejarse con la secuencia comparada. Este valor depende de qué tan símiles sean las secuencias comparadas, pues podría darse el caso de que tuvieran tan poca similitud, que sólo fuera posible alinearlas con un alineamiento local, en cuyo caso no abarcaría toda la secuencia, sino sólo un fragmento.

Después se muestra el ***valor E*** (***E-value***), el cual proporciona la medida más importante de significancia estadística. Representa la probabilidad de que el alineamiento observado sea producto del azar. Entre menor es el valor de *E-value*, (valor cercano a 0) significa que más similares son tus secuencias y existe más confianza de que el alineamiento observado no es producto del azar y realmente representan secuencias homólogas. En la siguiente columna se anota el porcentaje de identidad que resultó de la comparación entre las secuencias (*Ident*). Finalmente, se muestra el número de acceso de la secuencia con la que se comparó tu secuencia de consulta, con un enlace para acceder a ella.

Observa detenidamente los resultados, ¿qué puedes concluir? Podemos observar que los primeros cinco resultados tienen valores de *Score*, *Query cover*, *E-value* e *Ident* idénticos, lo que nos lleva a señalar que esas cinco secuencias son las mismas, aunque tienen números de acceso y nombres de gen distintos. Sin embargo, todas coinciden en que pertenecen a *Saccharomyces cerevisiae*, por lo que concluimos que la secuencia pertenece a dicha levadura. Ahora bien, ya sabemos al organismo al que pertenece, pero ¿de qué gen se trata? Si te fijas en la cuarta secuencia, se lee que el gen es una HSP10, una proteína de choque térmico de 10 kDa. Las demás secuencias sólo indican en qué cromosoma se encuentra, pero no indican el nombre del gen, pero como ya concluimos que las primeras cinco corresponden a la misma secuencia, entonces podemos decir que nuestro gen pertenece a *S. cerevisiae* y codifica para una HSP10.


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	Ident	Accession
TPA: Saccharomyces cerevisiae S288c chromosome XV, complete sequence	987	987	100%	0.0	100%	BK006948.2
S.cerevisiae chromosome XV reading frame ORF YOR020c	987	987	100%	0.0	100%	Z74928.1
S.cerevisiae genomic DNA (chromosome XV, strain FY1679)	987	987	100%	0.0	100%	X87331.1
S.cerevisiae HSP10 gene	987	987	100%	0.0	100%	X75754.1
S. cerevisiae CPN10 gene	987	987	100%	0.0	100%	X76853.1
Saccharomyces cerevisiae EC1118 chromosome XV, EC1118_1O4 genomic scaffold, whole genome shotgun sequence	981	981	100%	0.0	99%	FN394216.1
Saccharomyces cerevisiae S288c Hsp10p (HSP10), mRNA	593	593	60%	4e-166	100%	NM_001183439.1
Saccharomyces cerevisiae clone FLH114953.01X YOR020C gene, complete cds	588	588	60%	2e-164	99%	AY693222.1
S.cerevisiae chromosome XV reading frame ORF YOR021c	233	233	23%	1e-57	100%	Z74929.1
S.cerevisiae chromosome XV reading frame ORF YOR019w	161	161	16%	5e-36	100%	Z74927.1

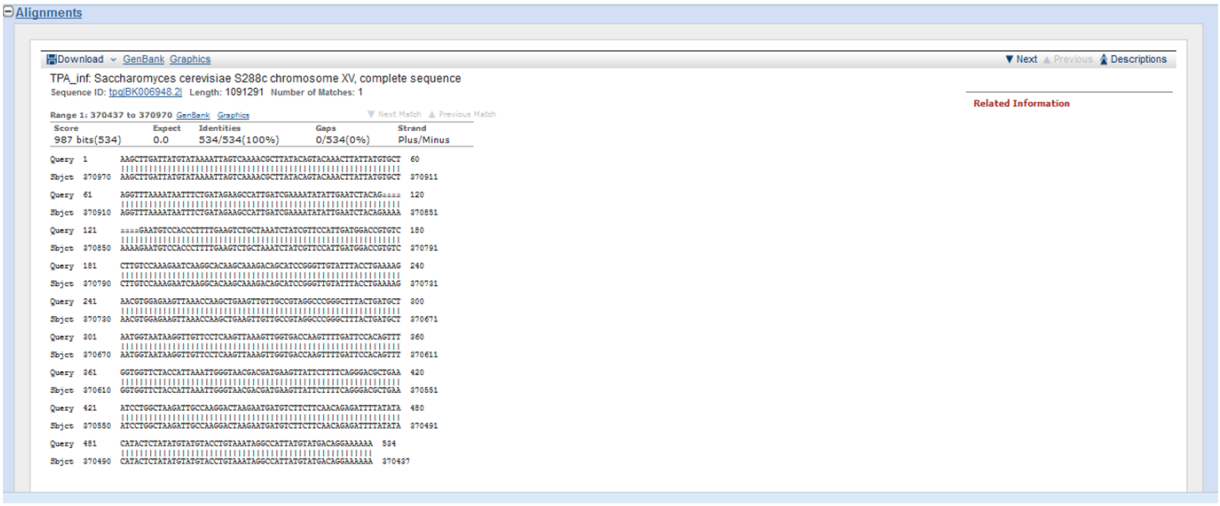
Alignments

En resumen, la interfaz que se observa en la imagen anterior corresponde a los resultados arrojados por BLAST, contiene la siguiente información para cada secuencia comparada.

1. ***Description***: Descripción de la secuencia
2. ***Max score*** (***Máximo puntaje***): Puntaje del mejor alineamiento local

3. ***Total score*** (***Puntaje total):*** La suma de puntajes de todos los alineamientos
4. ***Query cover:*** Porcentaje de alineamiento entre secuencias
5. ***E value*** (***Valor E***): Significancia estadística de la comparación
6. ***Ident***: El porcentaje de identidad entre las secuencias
7. ***Accession***: El identificador de secuencia

Si ahora sigues el acceso de alguna secuencia en particular, aparecerá el alineamiento completo, mostrando las secuencias y algunos de los valores que se incluyeron anteriormente como el score y el porcentaje de identidad. Otro dato calculado es el número de *gaps*. Ahora puedes guardar este alineamiento en tu computadora para tener evidencia de tus resultados, copiándolo y pegándolo en un procesador de textos.



¡Felicidades!, acabas de realizar tu primer alineamiento entre dos secuencias.

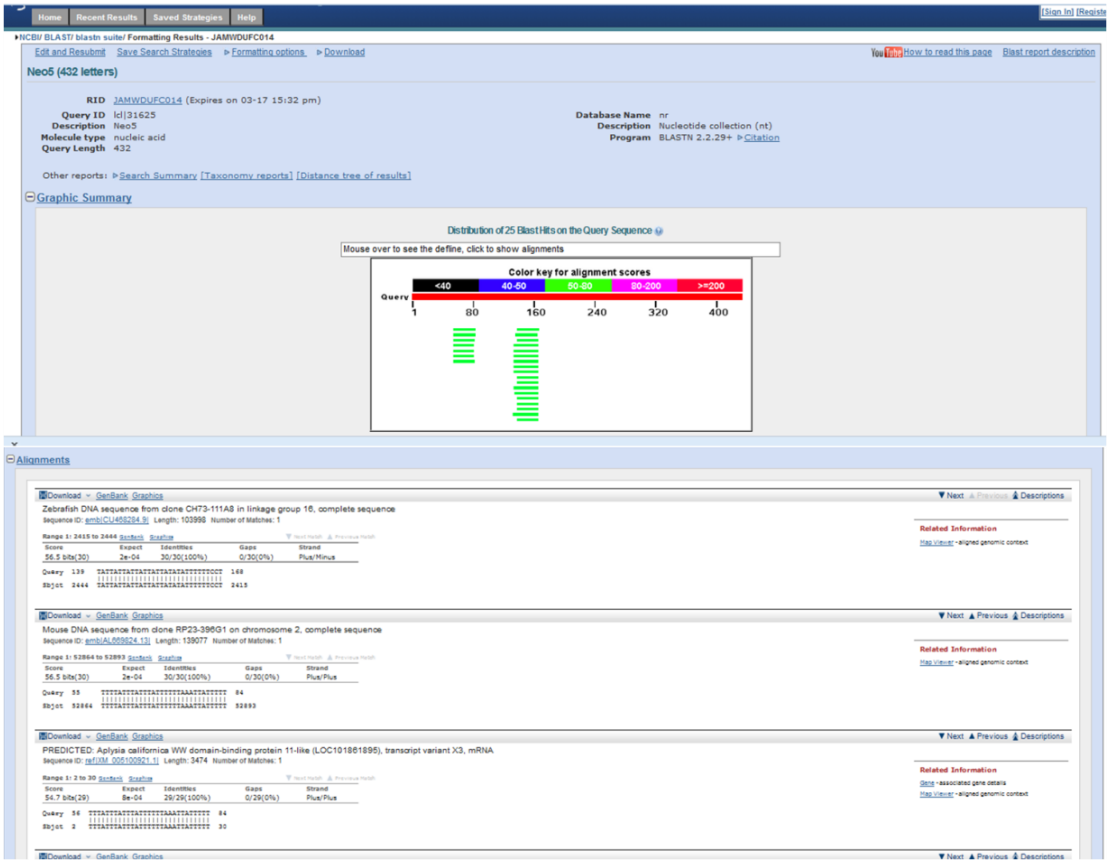
¿Qué crees que pase si usando la misma secuencia de consulta ahora utilizas las opciones blastx o tblastx?, ¿encontrarías los mismos resultados?, ¡inténtalo! No olvides poner tu secuencia en formato FASTA.

Secuencia de consulta:
GCGGATAGTTTGTACACATAGTGTCCCTAAAATTCCTATTGATGAATAGATCAATTTTA
TTAGCAGACAATTGGGGGCAGCAACTGAATAGCAGAAGAAATTTGAGTTCAATTATTT
TTTTTTCCTGTCATACATAATGGCCTATTTACAGGTACATACATATAGAGTATGTATATA
AAATCTCTGTTGAAGAAGACATCATTCTTAGTCCTTGGCAATCTTAGCCAGGATTTCAG
CGTCCCTGAAAAGAATAACTTCATCGTCGTTACCCAATTTAATGGTAGAACCACCAAA
CTGTGGAATCAAAACTTGGTCACCAACTTTAACTTGAGGAACAACCTTATTACCATTAG
CATCAGTAAAGCCCGGGCCTACGGCAACAACTTCAGCTTGGTTTAACTTCTCCACGTT
CTTTTCAGGTAAATACAACCCGGATGCTGTCTTTGCTTGTGCCTTGATTCTTTGGACAA

GGACACGGTCCATCAATGGAACGATAGATTTAGCAGACTTCAAAAGGGTGGACATTC
TTTTTTTTCTGTAGATTCAATATATTTTCGATCAATGGCTTCTATCAGAAATTATTTTAAA
CCTAGCACATAATAAGTTTGTACTGTATAAGCGTTTTGACTAATTTTATACATAATCAAG
CTTCTTTTTCCCATTCCTTCAAGATTCTAGAAATTTCTATCATTGATGACGGGCATTAC
CCCGTTAATGACCTTCACACGAATGAGAATTGGGCGGCTAATGAGAGAACTTCGAGA
GGTGCAATAAGTGAGAAATAACAACTTTAGAACTCATTATGATTGCTTCCAATACCTAA
TCCTACGTATGTACTAAATTAAAAAGACAGACATGCATTATTGAATATTGACATTTTGA
GAGTAACTTTTTATTATGAGTGGCATAATAAGATAATCGACGCAAGCCACAATTTATAC
AATAAAAAATGCTACCATCGCTGCTACATATGAACGAAAATAATACAACTATCGTTACG
GCCTTTGCTGAACCGTAATAAAATAAAATTCCTTGTTACATTTTTTTAGCCAGCTGCCT
CAGAAAGAGGCGTTTACTATTTAATGGAGAAAGAAAGCAAAGAAAAATAAAAGGTATT
TTCTTTACGGAAAGCCTTCGAGCAATCCAGGAGAAAGTGGACCTTTTTTTCCCAATGA
AGAGATCATAGGAGTATGGATTGAAAATATAATAGAACTTCGGGTAACGAGGTGTAAT
TTCACAGTCCATAATACAGAGCTAACGGTTTAAGGGTAAATAGTTATCTAAGTCAAGTT
TTGAAGGAACAAGTAAGAAAGGTCGCTACTGTTTCTAAACATAAGATATACAAAAATAA
ATATAGCTATCTCAATGGGTGCTGCATACAAAGTATTTGGGAAGACGGTTCAACCTCA
CGTATTGGCTATATCTACGTTTATCGCTACTGCTGCAGTGGCATCTTACTTTACCACG
AAACCAAAAACCAAAAATGAAGGCAAGAATAGTTCTGCCTTGAGCCAACAAAAAAGCG
GTGAAAGTTCAAACTCAGATGCTATGGGAAAGGACGATGATGTCGTAAAGAGTATTGA
AGGATTTTTAAATGATTTAGAGAAAGATACGAGGCAGGATACGAAAGCCAACTGATTA
TGTATAAAAATTTCTGAAATGGTGGTGTTCTTCATCGTTCAGTGAAGGGATGCACTGA
TTTCTATAAACTTGAAGCACTTTTTGAAACTACTGTTCTATAACGAAAATTAGCGTCCTT
CTTTCTATTAAGTATGCATTATACATATAATTCAATATATTCTGAATAGCAAAACGGCAA
TGAAAAAAAAAAAACACTGAAAATACTTGCCTTAGGCCATTGTGCATGATACGAATATG
CACAAAACTTGCCTCTTTTTACTTTACGGATCAATGACAACACTCAGGTGTAAGTGATA
GTTGATGGCCTTTCAATATTTGAAAGGCTGGAAGATAATTATAAAAAATCGAACTAATT
GCCTATGATTGTTTCATTACTGAGACTATTTTTTCACCTCAAGGGGGTCGTCTGAATTA
GCAAAGCCATGGCAACTAGTGCAGTGCTTGAAGTCACCAGCTCGTTTGGTTTTAGAT
GCAAGTAATGTAAAGAAATTAATTTAAATAAAATAATAAAAGTTTCTACTTTTTTTTTCAA
TTAAAAAGCATAATACAACCAATCAATTTTATCCTATTTGGCCTGACAATGATGATATC
ATAAAAGTAAACGGTTCCTTGTTTTTATTTTTCTTGCATGCACTTTTCAGAAGTCTTGGT
AGCGCTACTAACGCAAAATACGAAATATTCATTGGCTAATAAACTTGATTTTTTCATTG
AATGTCGTTTTTGAACTATATACAATATAATTAATGCTACGACCCTAAACTTTTCAACTA
ACTCTTTGACAAGGAAGCATCATCACTTATTACAACCATAGAATGTTACTTAAAGGACT
GTTCTCAT

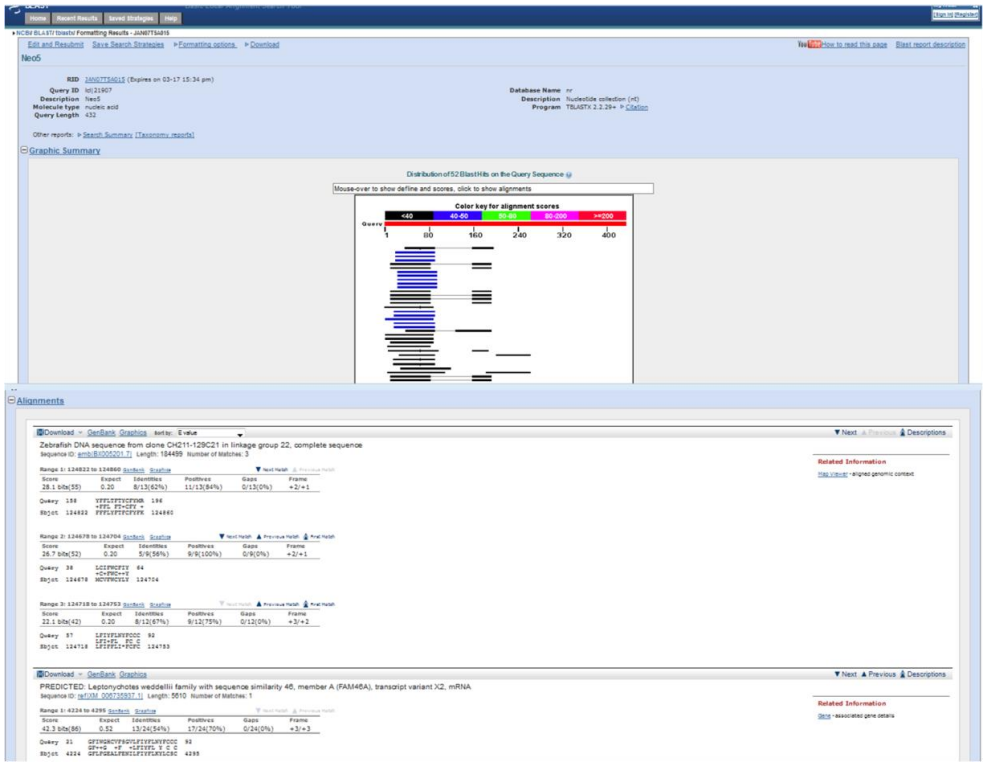
Un aspecto importante a considerar es que en las bases de datos podemos encontrar secuencias idénticas depositadas más de una vez por distintos investigadores, a las cuales se les asignó un nombre distinto, por lo que es estos casos, habrá que recurrir a la secuencia que tiene un nombre asignado. En nuestro ejemplo anterior, la secuencia de consulta codifica para alguna proteína, que es la HSP10, basándonos en los valores de

Score, *Query cover*, *E-value* e *Ident* el alineamiento generado es confiable. Sin embargo, puede haber casos en los que las secuencias de consulta no codifiquen para ninguna proteína y se trate por ejemplo, de una secuencia intergénica (por tanto, no codificante). Los resultados entonces no arrojaran alineamientos confiables y serán más bien alineamientos locales. Ya que en las bases de datos se depositan sólo secuencias codificadoras de genes, si nuestra secuencia de consulta no codifica para ninguna proteína, el programa no encontrará coincidencias entre las secuencias, pues nuestra secuencia, aunque sea real, no estará depositada en las bases de datos, por ser no codificante. Sin embargo, por probabilidad, encontrará alguna coincidencia con alguna secuencia depositada, pero si observamos los valores de *Score*, *Query cover*, *E-value* e *Ident*, estos nos dirán que dicho alineamiento no es confiable. Enseguida se muestra el ejemplo de resultados obtenidos al introducir como secuencia de consulta una secuencia intergénica que pertenece al protozoario *Trypanosoma cruzi*. En el primer caso, el alineamiento se realizó con blastn. Como puedes observar, los alineamientos generados son sólo locales, por las razones que ya se explicaron anteriormente.

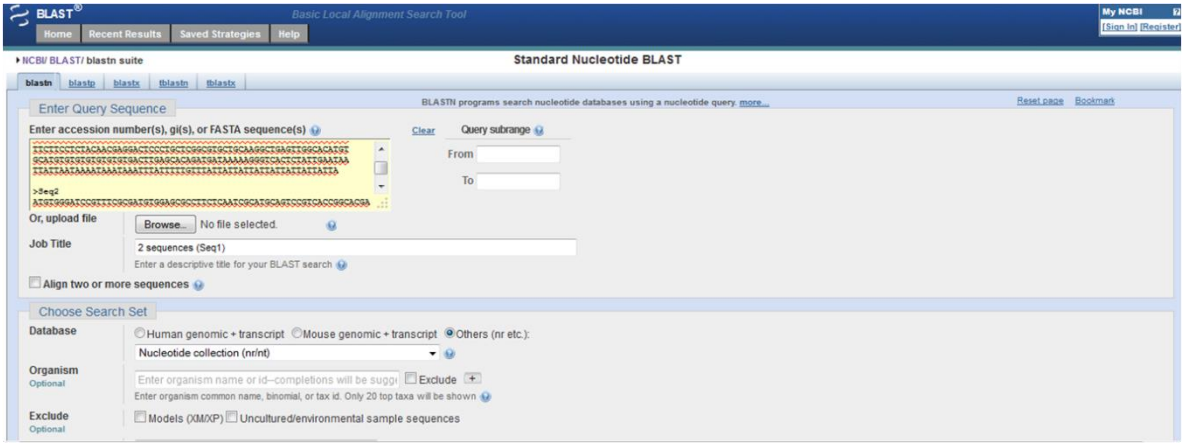


Usando ahora el programa tblastx, se encontraron los mismos resultados, sólo alineamientos locales, que fueron generados por azar, pues la secuencia de consulta o

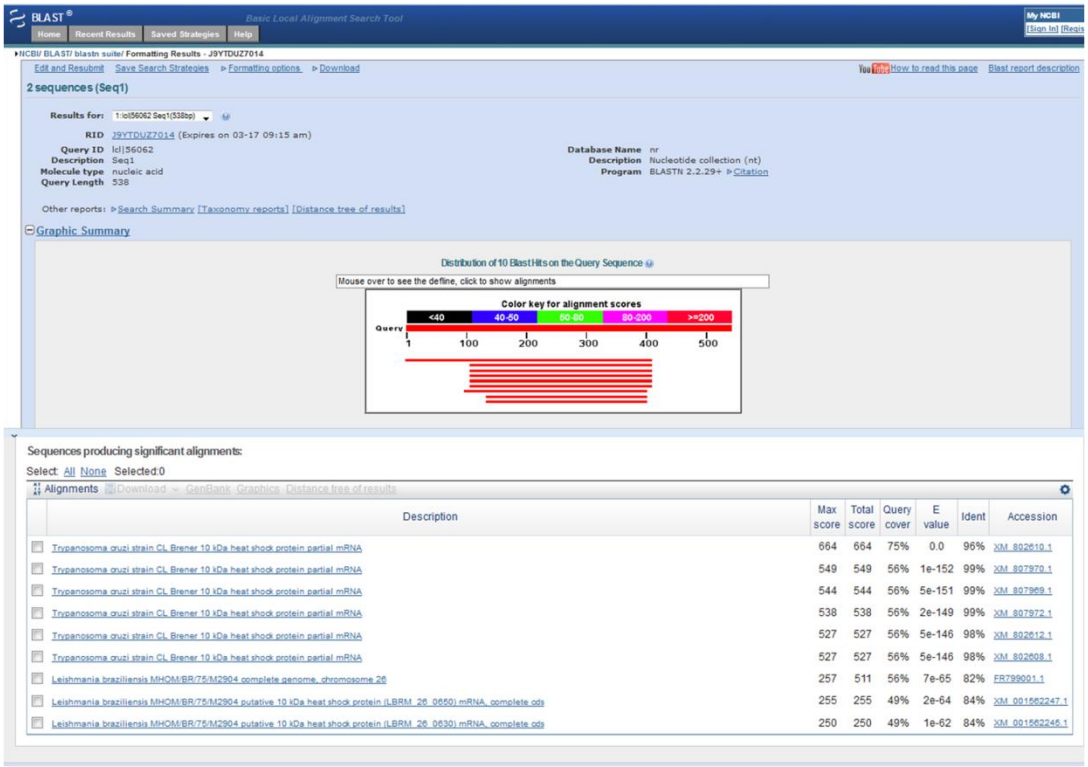
secuencias similares no se encuentran depositadas en la base de datos de nucleótidos ni de nucleótidos traducidos a aminoácidos. Observa los valores del *Score*, *Query cover*, *E-value* e *Identy* y analízalos respecto a lo que ya sabes respecto a sus valores esperados.



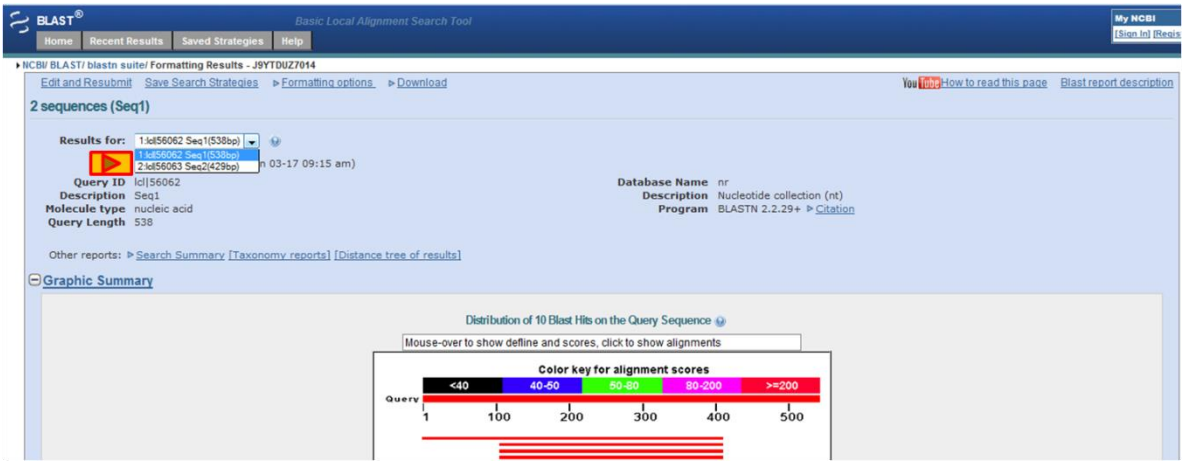
¿Qué pasaría si tuvieras más de una secuencia de consulta? En este caso puedes introducir cada una de tus secuencias de manera individual como lo hicimos en el ejemplo anterior o puedes ser más práctico y en la misma ventana introducir las varias secuencias de consulta que quieras analizar. Lo anterior siempre y cuando el programa de análisis a utilizar sea el mismo (blastn, tblastx, etc.), así como la naturaleza de tus secuencias, es decir, todas deben ser nucleótidos o todas secuencias de aminoácidos. En el ejemplo de abajo se muestra que en la misma ventana se copiaron dos secuencias de nucleótidos, cada una en formato FASTA.



Una vez que se ejecutó el BLAST, los resultados aparecieron de la siguiente forma. En esta ventana se muestra el alineamiento de una de las secuencias (Seq1), que corresponden a una proteína de choque térmico de 10 kDa del parásito *T. cruzi*.



Ahora coloca el cursor en la pestaña de "*Results for*" y selecciona la Seq2, de 429 nucleótidos.

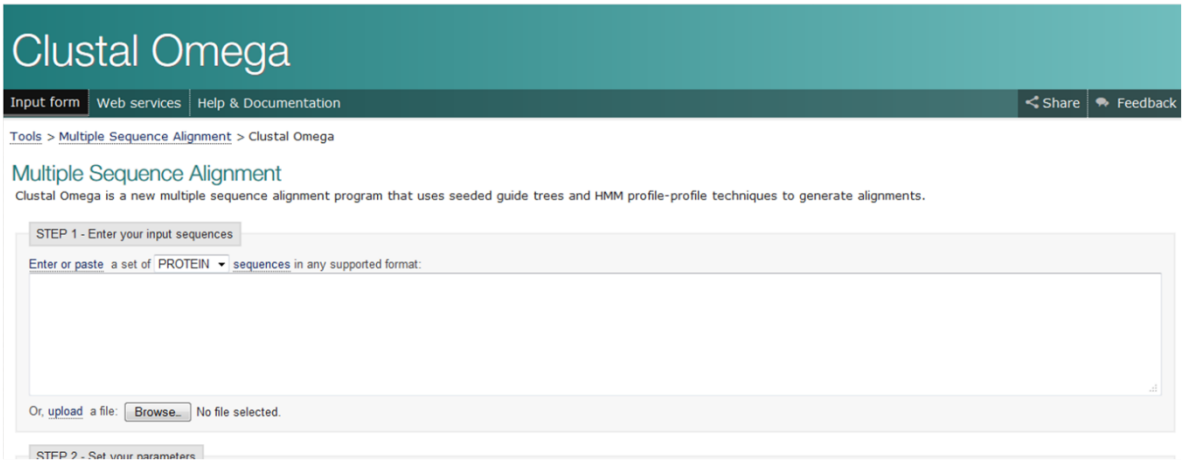


La nueva ventana entonces te mostrará los resultados para la otra secuencia de consulta, en este caso, el alineamiento muestra que codifica para una proteína de la familia *alpha-crystallin small heat shock protein*, también de *T. cruzi*. De esta forma, en un mismo paso de ejecución puedes hacer el alineamiento de varias secuencias, accediendo a ellas después, en la misma ventana de resultados.

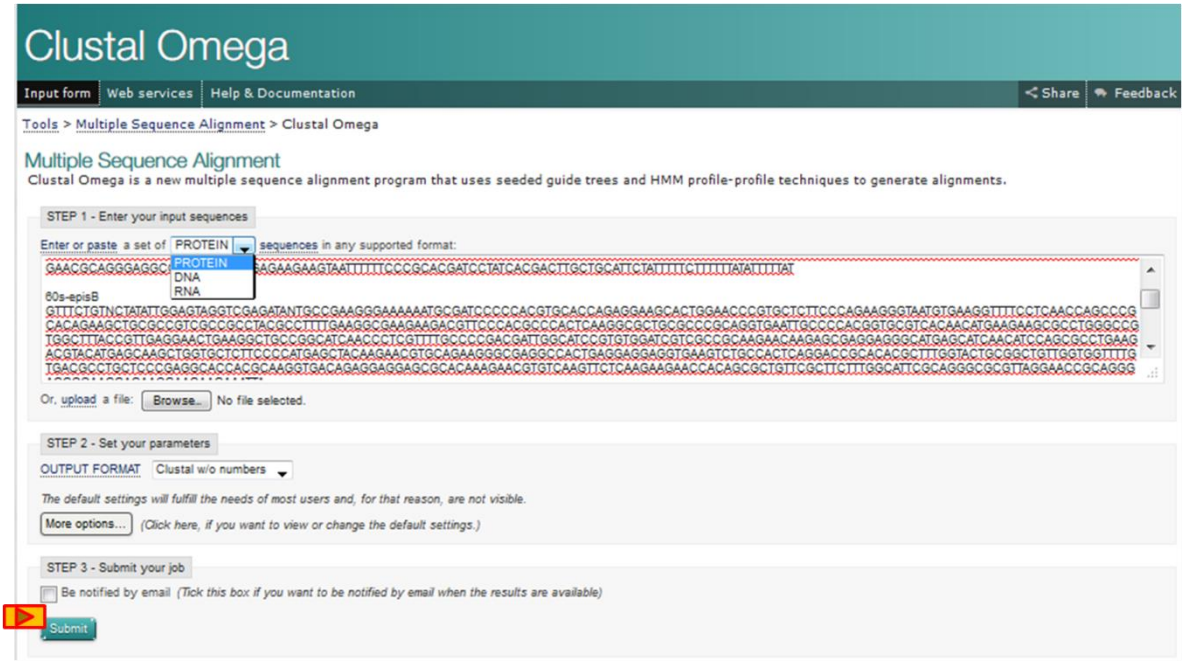
2.2.5. Alineamiento múltiple de secuencias

El alineamiento múltiple implica alinear más de dos secuencias. Para alinear más de dos secuencias utilizaremos el programa especializado Clustal Omega, disponible en la red de manera gratuita.

Así, accede al sitio de internet de este software: https://www.ebi.ac.uk/Tools/msa/clustalo/. Una ventana como la siguiente es mostrada.



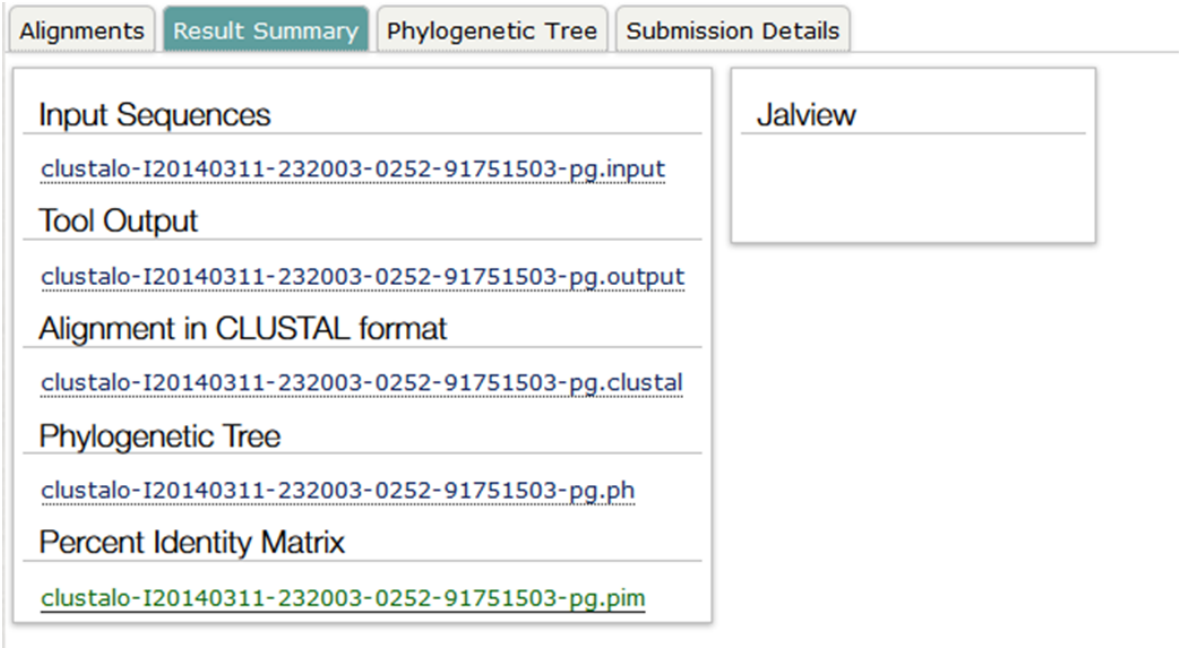
Ahora pega tus secuencias a alinear. Este software admite distintos formatos de secuencia, entre ellos FASTA. Introduce tus secuencias en este formato. Selecciona en la pestaña donde se lee **"*Enter or paste a set of….*"** la opción que corresponda de acuerdo a la naturaleza de tus secuencias, es decir, si se trata de una secuencia de aminoácidos (PROTEIN), de ADN (DNA) o de ARN (RNA). Finalmente, ejecuta el comando **"*Submit*"**.



Enseguida se muestran los resultados del alineamiento.

```
CLUSTAL O(1.2.1) multiple sequence alignment


60s-episB       ---------------------------GTAGGTCGAGATAATGC-CGAAGGAAAAAACGC
60s-epis        --------GTTTCTGTNCTATATTGGAGTAGGTCGAGATANTGCCGAAGGGAAAAAATGC
60s-amastis     GTGATACAGTTTCTGTACTATATTGGAGTAGGTCGAGATAATGCCGAAGGGAAAAAACGC
60s-amastisB    GTGATACAGTTTCTGTACTATATTGGAGTAGGTCGAGATAATGCCGAAGGGAAAAAACGC
60s-amastisC    GTGATACAGTTTCTGTACTATATTGGAGTAGGTCGAGATAATGCCGAAGGGAAAAAACGC
                                          ************** *** * ******** **

60s-episB       GATCCCCCACGTGCACCAGAGGAAGCACTGGAACCCGTGCTCTTCCCAGAAGGGTAATGT
60s-epis        GATCCCCCACGTGCACCAGAGGAAGCACTGGAACCCGTGCTCTTCCCAGAAGGGTAATGT
60s-amastis     GATCCCCCACGTGCACCAGAGGAAGCACTGGAACCCGTGCTCTTCCCAGAAGGGTAATGT
60s-amastisB    GATCCCCCACGTGCACCAGAGGAAGCACTGGAACCCGTGCTCTTCCCAGAAGGGTAATGT
60s-amastisC    GATCCCCCACGTGCACCAGAGGAAGCACTGGAACCCGTGCTCTTCCCAGAAGGGTAATGT
                ************************************************************

60s-episB       GAAGGTTTTCCTCAACCAGCCCGCACAGAAGCTGCGCCGTCGCCGCCTACGTCTTTTGAA
60s-epis        GAAGGTTTTCCTCAACCAGCCCGCACAGAAGCTGCGCCGTCGCCGCCTACGCCTTTTGAA
60s-amastis     GAAGGTTTTCCTCAACCAGCCCGCACAGAAGCTGCGCCGTCGCCGCCTACGTCTTTTGAA
60s-amastisB    GAAGGTTTTCCTCAACCAGCCCGCACAGAAGCTGCGCCGTCGCCGCCTACGTCTTTTGAA
60s-amastisC    GAAGGTTTTCCTCAACCAGCCCGCACAGAAGCTGCGCCGTCGCCGCCTACGTCTTTTGAA
                ************************************************** ********

60s-episB       GGCGAAGAAGACGTTCCCACGCCCACTCAAGGCGCTGCGCCCGCAGGTGAATTGCCCCAC
60s-epis        GGCGAAGAAGACGTTCCCACGCCCACTCAAGGCGCTGCGCCCGCAGGTGAATTGCCCCAC
60s-amastis     GGCGAAGAAGACGTTCCCACGCCCACTAAAGGCGCTGCGCCCGCAGGTGAATTGCCCCAC
60s-amastisB    GGCGAAGAAGACGTTCCCACGCCCACTAAAGGCGCTGCGCCCGCAGGTGAATTGCCCCAC
60s-amastisC    GGCGAAGAAGACGTTCCCACGCCCACTAAAGGCGCTGCGCCCGCAGGTGAATTGCCCCAC
                ****************************** **********************************

60s-episB       GGTGCGTCACAACATGAAGAAGCGCCTGGGCCGTGGCTTTACCGTTGAGGAGCTGAAGGC
60s-epis        GGTGCGTCACAACATGAAGAAGCGCCTGGGCCGTGGCTTTACCGTTGAGGAACTGAAGGC
60s-amastis     GGTGCGTCACAACATGAAGAAGCGCCTGGGCCGTGGCTTTACCGTTGAGGAGCTGAAGGC
60s-amastisB    GGTGCGTCACAACATGAAGAAGCGCCTGGGCCGTGGCTTTACCGTTGAGGAGCTGAAGGC
60s-amastisC    GGTGCGTCACAACATGAAGAAGCGCCTGGGCCGTGGCTTTACCGTTGAGGAGCTGAAGGC
                ************************************************************

60s-episB       TGCCGGCATCAACCCTCGTTTTGCCCCGACGATTGGCATCCGTGTGGATCGTCGCCGCAA
60s-epis        TGCCGGCATCAACCCTCGTTTTGCCCCGACGATTGGCATCCGTGTGGATCGTCGCCGCAA
60s-amastis     TGCCGGCATCAACCCTCGTTTTGCCCCGACGATTGGCATCCGTGTGGATCGTCGCCGCAA
60s-amastisB    TGCCGGCATCAACCCTCGTTTTGCCCCGACGATTGGCATCCGTGTGGATCGTCGCCGCAA
60s-amastisC    TGCCGGCATCAACCCTCGTTTTGCCCCGACGATTGGCATCCGTGTGGATCGTCGCCGCAA
                ************************************************************
```

Para conocer el porcentaje de identidad que existe entre cada par de secuencias, selecciona la opción **"*Result Summary*"** y enseguida **"*Percent Identity Matrix*"**. En este punto, cabe mencionar que en los alineamientos múltiples, los porcentajes de identidad tienen que ser calculados necesariamente asumiendo alineamientos entre un par de secuencias.



Ahora se muestra una matriz con el porcentaje de identidad de la secuencia 1 con la secuencia 1, de la secuencia 1 con la secuencia 2 y así sucesivamente, obteniendo todas las combinaciones de comparación entre las 5 secuencias.

```
#
#
#  Percent Identity  Matrix - created by Clustal2.1
#
#

     1: 60s-episB     100.00   96.35   96.73   96.92   96.92
     2: 60s-epis       96.35  100.00   91.42   91.72   98.40
     3: 60s-amastis    96.73   91.42  100.00   99.47   99.83
     4: 60s-amastisB   96.92   91.72   99.47  100.00  100.00
     5: 60s-amastisC   96.92   98.40   99.83  100.00  100.00
```

A pesar de que los alineamientos múltiples tienen mayor significancia en secuencias de aminoácidos, para el caso de ácidos nucleicos estos pueden servir, por ejemplo, para comparar la secuencia de un gen antes de ser mutado *in vitro* contra las versiones del gen después de diferentes mutaciones generadas. En este caso no podemos establecer relaciones de homología, si no sólo de comparación entre secuencias para identificar cambios en las secuencias. Por otro lado, cabe mencionar que en este programa también se pueden realizar alineamientos de una par de secuencias.

2.3. Aplicaciones prácticas del análisis de las secuencias de ADN

Continuaremos con los alineamientos de secuencias para el estudio de proteínas en la siguiente Unidad. Mientras tanto, en esta sección abordaremos algunas herramientas bioinformáticas que se pueden aplicar para identificar características intrínsecas de las secuencias de nucleótidos y también analizaremos algunas herramientas con aplicaciones en las técnicas de ingeniería genética.

2.3.1. Contenido GC

Los genomas, genes o segmentos de ADN o ARN de los organismos, tienen características intrínsecas que los identifican. Una de estas características es el contenido GC, también conocido como porcentaje GC, el cual representa la cantidad de las bases nitrogenadas Guanina (G) y Citocina (C) que están presentes en el genoma o un segmento de ADN y es expresado en porcentaje. La fracción restante de la molécula de ADN contendrá las bases Adenina (A) y Timina (T), de tal forma que a partir del contenido GC, podemos saber el contenido AT. Así por ejemplo, un contenido GC de 51%, tendrá un contenido AT de 49%. Si se conoce la secuencia de un genoma, gen o fragmento de ADN del que quieres determinar el contenido GC, éste se calcula como sigue (Krebs, 2010):

$$\frac{Número\ de\ bases\ G + C}{Número\ de\ bases\ A + T + G + C} x\ 100$$

El contenido GC también puede ser determinado experimentalmente, en caso de que no se conozca la secuencia. Importantemente, el contenido GC tiene un significado biológico. Las bases nitrogenadas G y C se encuentran unidas por tres enlaces de hidrógeno, a diferencia de A y T, que sólo están unidas por dos enlaces. Como resultado de esta interacción, se produce una unión más fuerte entre G y C, lo que hace más resistentes a estos enlaces a sufrir una desnaturalización como producto de altas temperaturas. Por esta razón, un contenido GC alto es característico de organismos que viven en ambientes con altas temperaturas o termófilos (Krebs, 2010).

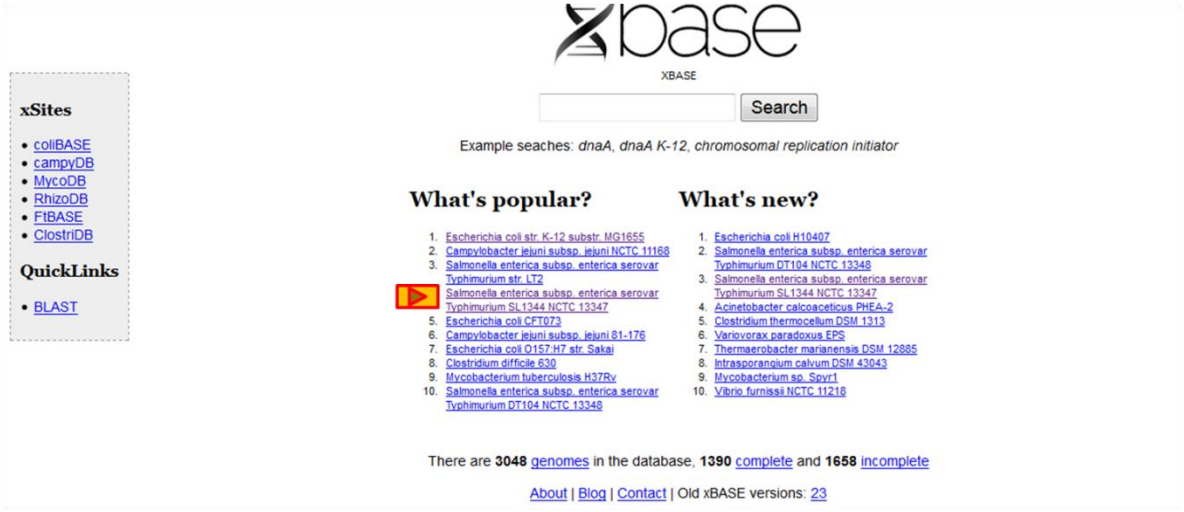
Otro fenómeno biológico asociado al contenido GC es la transferencia horizontal de genes. Entre los organismos, un evento común es que los genes sean transferidos de una especie a otra, representando este hecho, por tanto, una transferencia horizontal y no vertical como sería de padres a hijos entre la misma especie o entre ancestros evolutivos. Los genes adquiridos de esta forma, que no se encontraban en la especie ancestral, pueden conferir ventajas a los organismos que los adquirieron, en cuyo caso los

conservan como parte de su genoma. Los genes adquiridos de esta forma son diferenciables de los genes que conformaban al organismo ancestral debido al contenido GC. Ya que el contenido GC es característico de cada organismo, en general, los genes que se adquirieron de otros organismos, tienen un contenido distinto al del organismo que los recibe. Ejemplos de genes que han sido adquiridos de esta manera en bacterias con los que confieren resistencia a antibióticos. Otro ejemplo son los genes que están involucrados en producir virulencia, es decir, aquéllos genes que en bacterias les permiten infectar a un organismo. Por ejemplo, en la bacteria *Salmonella enterica*, cuyo contenido GC es de aproximadamente 52%, ciertos genes asociados a virulencia presentan un contenido GC menor, entre 37-44% (Main-Hester *et al.*, 2008). En la bacteria emparentada *Escherichia coli*, el contenido GC es del 51%.

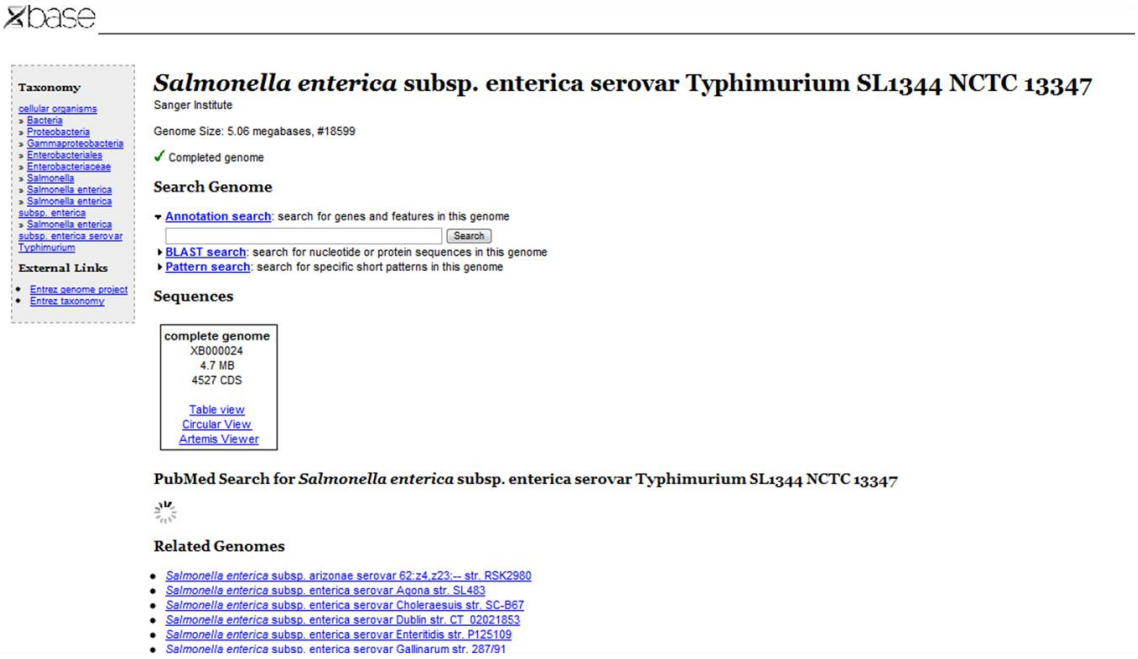
Existen distintas bases de datos donde entre otros datos, se indica el contenido GC del genoma de los organismos. Generalmente dichas bases de datos se encuentran disponibles para aquéllos organismos cuyos genomas han sido completamente secuenciados. En esta sección utilizaremos la base de datos XBase, que contiene información relacionada a distintas especies de bacterias.

Puedes acceder a la página de la base de datos XBase, cuya dirección web es la siguiente: http://www.xbase.ac.uk/colibase/.

Lo primero es escoger, de entre las distintas especies de bacterias, de la que queramos obtener información. En nuestro ejemplo, seleccionaremos la especie *Salmonella enterica* subsp. enterica serovar Typhimurium SL1344 NCTC 13347. Accede a dicha especie bacteriana.

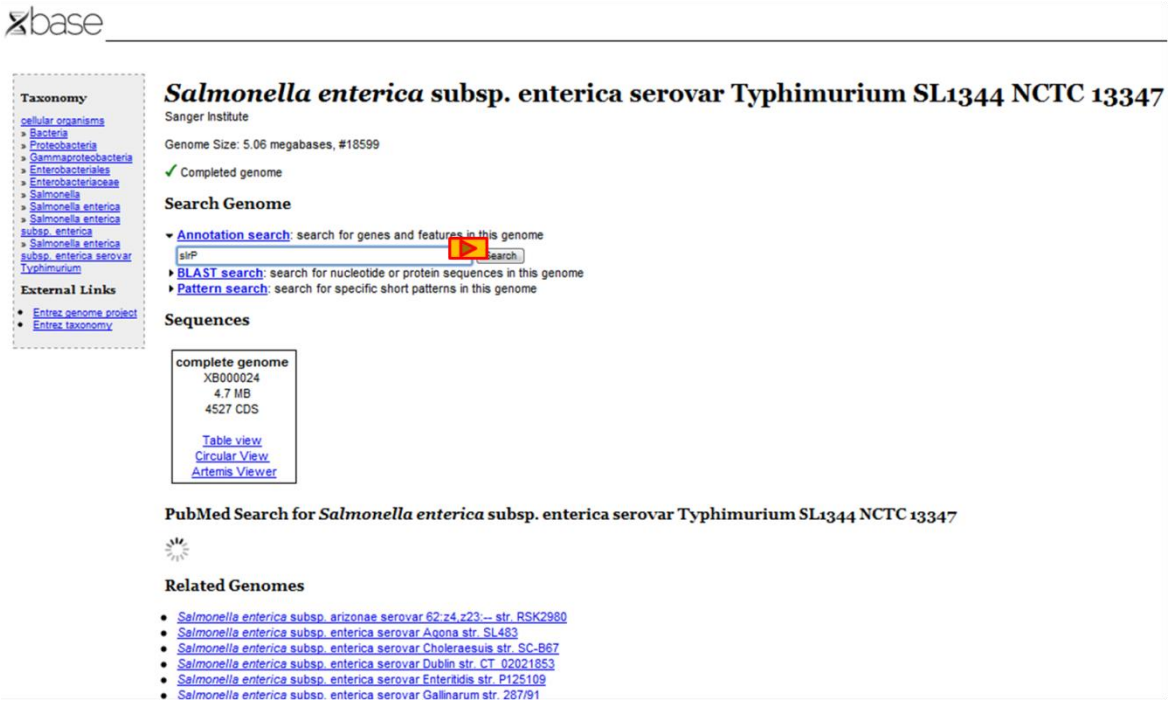


Una ventana como la siguiente es desplegada.

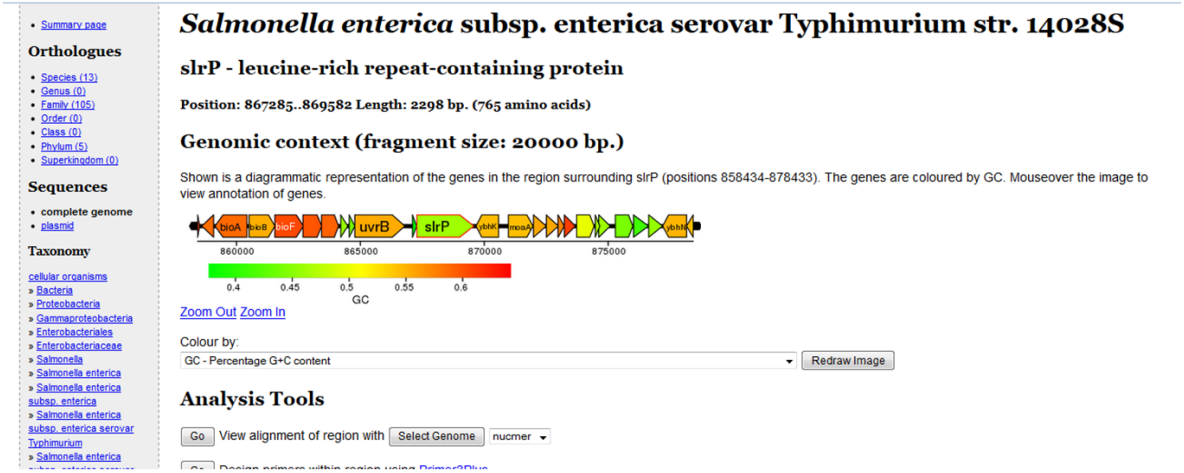


Se muestran de arriba a abajo, los siguientes datos: el organismo que seleccionaste para obtener información (*Salmonella enterica*), la institución que alimenta la base datos (Instituto Sanger), el tamaño del genoma (5.06 megabases), el estado del proyecto de secuenciación (completado). Posteriormente, se muestra una ventana (*Annotation search*) donde se introduce el nombre del gen de consulta. Después hay una opción de BLAST (*BLAST search*) para realizar una búsqueda utilizando únicamente la información correspondiente al genoma de esta bacteria. Más abajo hay diversos enlaces para acceder al genoma completo y a genomas de bacterias emparentadas.

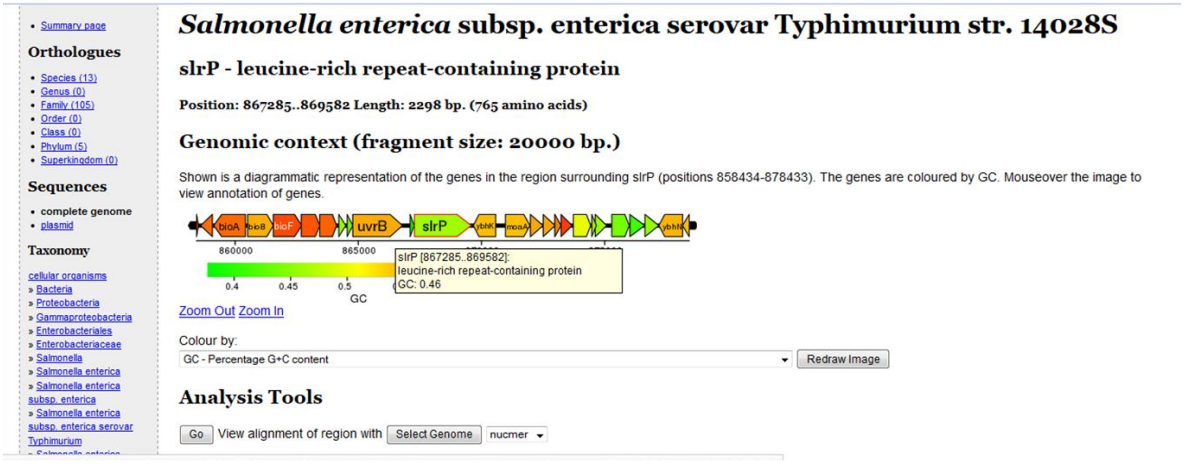
Ahora coloca el nombre del gen *slrP* en la ventana **"*Annotation search*"** y oprime search.



Acto seguido, aparecerá una ventana como la siguiente. Debajo del nombre del organismo, aparece el nombre del gen y más abajo su contexto genómico. El **contexto genómico** se refiere al contenido genético de esa zona particular del genoma, es decir, cuáles genes se encuentran alrededor de nuestro gen de búsqueda. Como puedes observar, los distintos genes están coloreados de manera diferencial. El color depende del contenido GC, en la escala encontrada debajo del diagrama genético, se indica el porcentaje GC. Así por ejemplo, un porcentaje GC de aproximadamente 45% corresponde al color verde. De esta manera, se vuelve fácil y rápido de visualizar el contenido GC de tus genes de consulta. Con las opciones *Zoom Out* y *Zoom In* puedes abarcar una región mayor del contexto genético o una región menor, respectivamente.

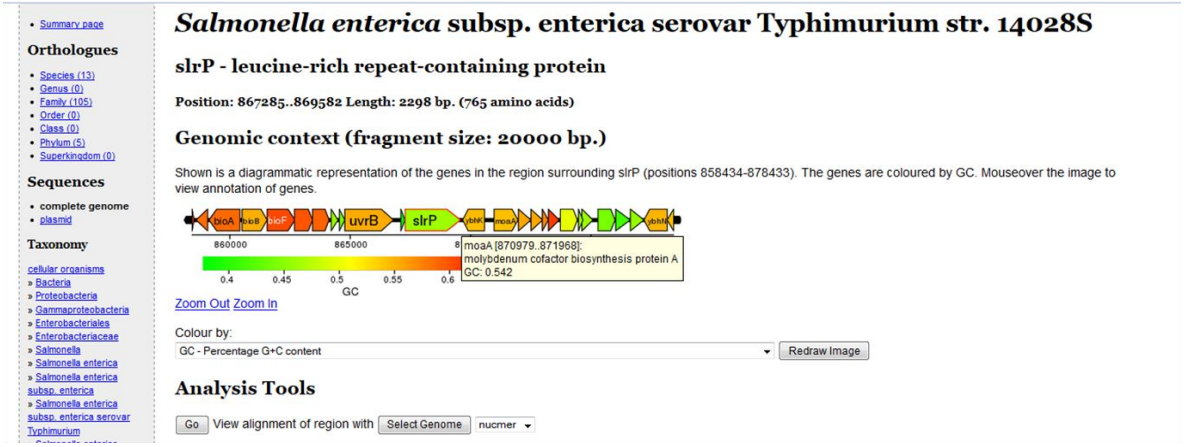


Para conocer el porcentaje GC exacto de tu gen, coloca el cursor sobre el gen del que quieras obtener la información y el porcentaje GC se mostrará en una pequeña ventana, además del nombre del gen y de dónde a dónde abarca la secuencia de dicho gen, es decir, del nucleótido 23 al 45, por ejemplo. De acuerdo a su porcentaje GC, el gen *slrP* parece haber sido adquirido por transferencia horizontal. Recordemos que *S. enterica* tiene un porcentaje GC de aproximadamente 52%, mientras que *slrP* tienen un contenido GC de 46%, más bajo que el del resto de genoma. Lo anterior sugiere que hoy en día, dicho gen está presente en el genoma de esta bacteria como consecuencia de un evento de transferencia horizontal entre bacterias que ocurrió durante su evolución. Cabe mencionar que la proteína para la que codifica el gen *slrP* ha sido implicada en la virulencia de la bacteria.



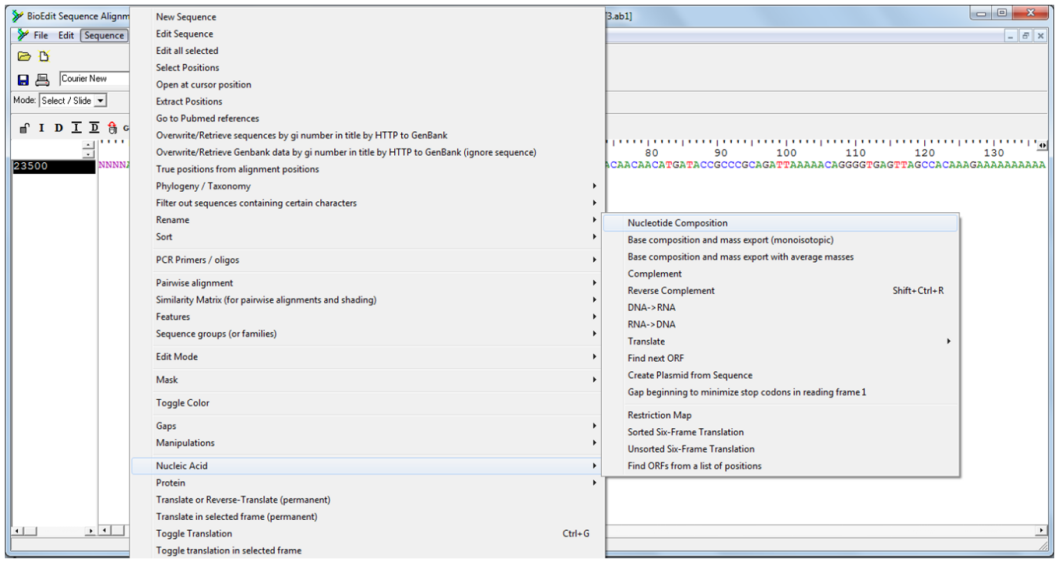
Miremos ahora dos genes más allá de *slrP*, el gen *moaA*, que codifica para un cofactor de molibdeno para biosíntesis de proteína A, implicado en el metabolismo bacteriano. El gen tiene un porcentaje GC de 54%, muy parecido y no menor al del resto de genoma, que es

de 52%. Lo anterior sugiere que este gen no es resultado de un evento de transferencia horizontal. Si un gen está implicado en el metabolismo central de la bacteria, es muy probable que estuviera presente desde los ancestros de esta bacteria, es decir, no fue adquirido posteriormente.

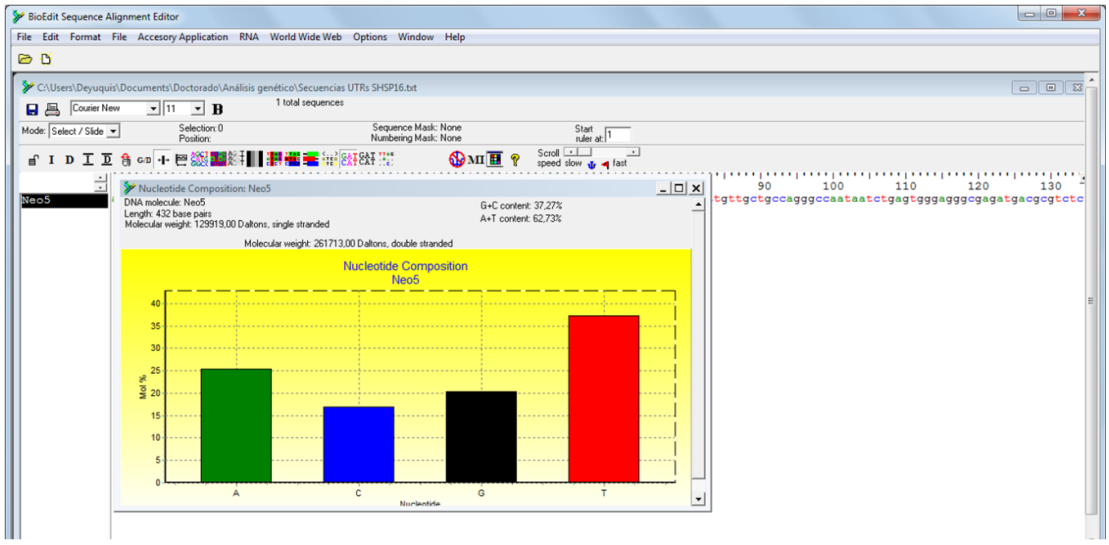


En esta sección exploramos el uso de sólo algunas funciones de esta base de datos, sin embargo, te invitamos a que explores por tu cuenta todo el tipo de información que puedes obtener.

Con BioEdit también es posible conocer la composición de nucleótidos de tu secuencia, tanto de las bases individuales, como el porcentaje GC. Así, si quisieras determinar el contenido GC de alguna secuencia en particular, puedes utilizar esta herramienta. Usando la misma secuencia con la que hemos venido trabajando, en la pestaña **"*Sequence*"**, selecciona la opción **"*Nucleic Acid*"** y enseguida **"*Nucleotide Composition*"**.



Enseguida se mostrará un esquema como el siguiente, donde puedes encontrar los porcentajes de ocurrencia de las bases en tu secuencia, así como el contenido GC y el contenido AT.



2.3.2. Diseño de cebadores

Como recordaras**,** un **cebador** también llamado **primer** u **oligonucléotido**, es una secuencia de nucleótidos de cadena simple que sirve como punto de anclaje para la enzima ADN polimerasa y que ésta replique una secuencia determinada de ADN en la reacción en cadena de la polimerasa (PCR). Como se muestra en la figura 4, se utilizan

dos cebadores que se hibridan con la secuencia de ADN complementaria de manera específica para poder amplificar una secuencia definida. Debido a la importancia y a las múltiples aplicaciones en la biología molecular de la amplificación de secuencias de ADN por PCR, se han generado herramientas bioinformáticas para facilitar el diseño de los cebadores.

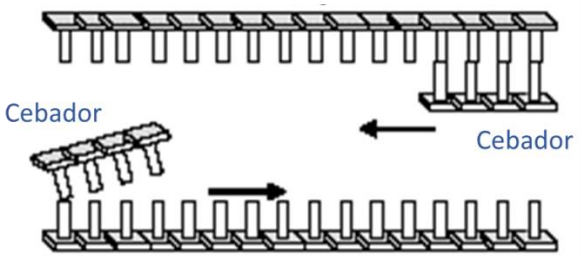


Figura 4. Se muestra la doble cadena de ADN desnaturalizada y la unión de los cebadores para dar inicio a la reacción en cadena de la polimerasa (PCR) (Imagen modificada de Konietzny y Greiner, 2003).

Entender las características de un cebador es importante para su diseño y correcta implementación en las técnicas de biología molecular, por ello a continuación se describen brevemente estas características.

Al diseñar cebadores se busca una asociación específica y complementaria de dos cadenas sencillas de ácidos nucleicos, en este caso la **especificidad** debe ser entre el cebador y la secuencia que se quiera amplificar. La **longitud del cebador** influye en la especificidad, en la **temperatura de alineamiento** y en el tiempo necesario para la hibridación con su secuencia complementaria. El tamaño del oligonucleótido es proporcional a la especificidad de la hibridación con la secuencia complementaria, en cuanto más largo el cebador más específico; sin embargo, con cebadores más largos es más ineficiente el proceso de hibridación. Por estas razones se recomienda utilizar cebadores de 18 a 24 bases (Dieffenbach *et al.,* 1995).

Otro factor importante es la composición del extremo 3’ del cebador, ya que errores de hibridación en el extremo 3’ resultan en reducción significativa en la eficiencia de una PCR (Yang*, et al.*, 2006). Se recomienda que el cebador tenga, como máximo, dos residuos G o C en la últimas cinco bases del extremo 3’, lo cual ayuda a asegurar que las secuencias hibriden correctamente debido a los enlaces de hidrógeno que se dan entre los residuos de GC. Sin embargo, se debe evitar que haya secuencias de poliG (4 o más guaninas seguidas) o poliC (4 o más citosinas seguidas) para evitar la formación de dímeros (Yang, *et al.,* 2006; Brown, 2000). Así mismo debe evitarse que haya secuencias complementarias dentro de la secuencias del cebador sentido y anti sentido, más de 3 pares de bases complementarias pueden causar la formación de estructuras secundarias y que los cebadores hibriden entre ellos, resultando en menor cantidad de cebador

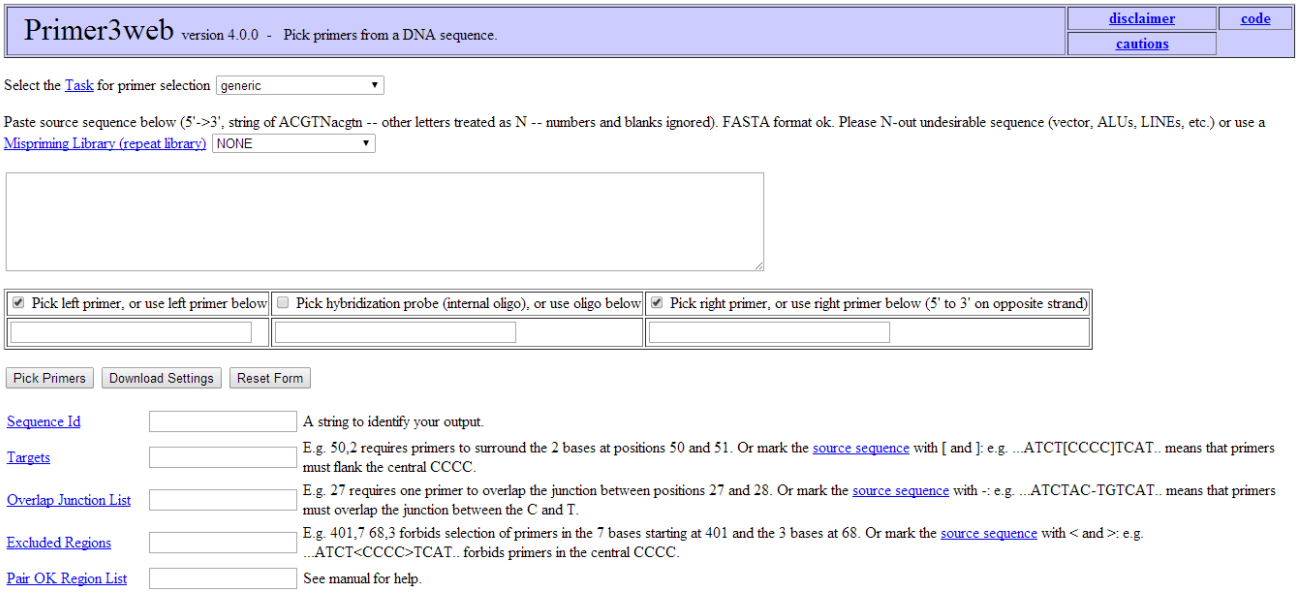
disponible para amplificar, dando como consecuencia menor eficiencia de la PCR (Yang, et al., 2006). Para una eficiente hibridación se recomienda que la composición de bases de los cebadores sea del 45% al 55% en GC y que cebador sentido y anti-sentido hibriden a la misma temperatura, la cual deberá estar dentro del rango de los 50 °C a los 65 °C (Yang, *et al.,* 2006).

Existen disponibles varios programas informáticos para el diseño de cebadores, **CODEHOP**, **Primer-Blast** y **Primer3Plus** son algunos ejemplos. En estos programas se introduce la secuencia que se quiere amplificar y el sistema realiza el diseño de los cebadores de acuerdo a las características antes mencionadas.

Si bien existe una variedad de programas que ayudan en el diseño y selección de cebadores se recomienda probarlos experimentalmente para una secuencia y contrastar su capacidad de amplificación a la misma temperatura (Brown, 2000).
Así mismo es muy importante verificar en las bases de datos de ADN que el cebador diseñado tenga como blanco la secuencia deseada y no vectores, genes, secuencias repetidas, que podrían interferir en la PCR (Sambrook y Russell, 2001).

A continuación se muestra el diseño de cebadores a partir de la herramienta Primer3, la cual es de acceso público (Untergrasser *et al*., 2012). Puedes entrar a partir del siguiente enlace http://bioinfo.ut.ee/primer3/, en donde podrás abrir la página web similar a la imagen que se muestra a continuación.

Este programa permite especificar un gran número de parámetros para el diseño de cebadores de acuerdo a las necesidades como Tm, porcentaje de GC, máxima auto-complementariedad. También permite discriminar las regiones de la secuencia que se quieren amplificar, las que se deben excluir y el tamaño del producto de PCR.

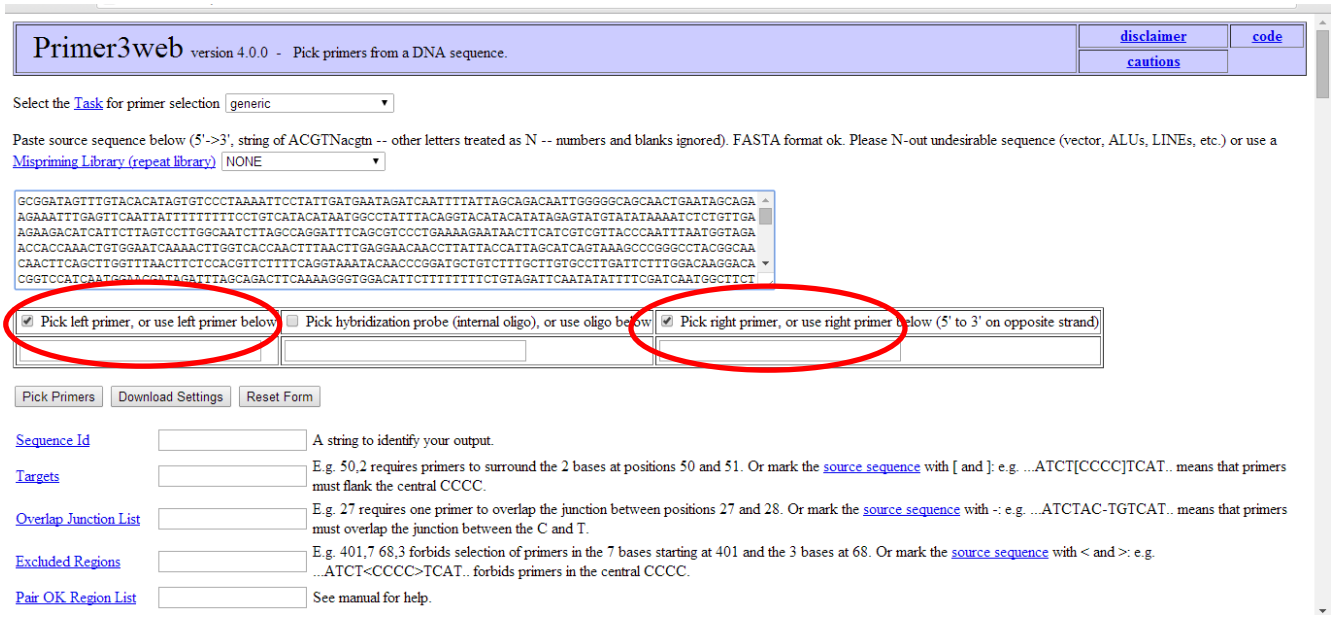


Utilizaremos la secuencia que corresponde al gen de una quitinasa de *Arthrobacter sp.* para ejemplificar el procedimiento de diseño de cebadores.

>gi|50871751|emb|AJ504427.1| Arthrobacter sp. TAD20 chiC gene for chitinase
ATGTTCAGCCGTTACGTGGATGTTACGGCAACCCCGTCATACAGTTTTGAAAAACCGG
TCTCCGAGTCTGCTGATTCCGTTGTTTTGTCCTTTATTGTCTCGTCCAAGGACAAAGC
GTGCGAGCCAAGTTGGGGAACTTTCTACTCCATGGATGCCGCCGGGCAAGAATCGG
ACGTGGATCGGCGTATTGCACGCCTACTCCAGCAAGGCGGTTCCGTATCCGTCTCGT
TTGGCGGCCAAGCTAATGATGAGCTTGCCGTGCGCTGTACGGATGTTGCGGAACTGC
AGGCCGCCTACGCCTCAGTGGTAGAACGCTACGATCTTTCCGCGATTGACCTTGACT
TAGAAGGCCCAGGTCTATCCGATACCGCTGCTCTCAAGCGACGAGCCACAGCAATTG
CTGCGCTGCAGTCGCAACGACTTGCTGCAAAGCATCCGCTGTCAGTGTGGCTGACAC
TCCCAGTTGCTCCCACCGGGCTGACCGCTGAGGGGACCTCGGCAGTTGCTGCCATG
CTGGATGCAAAGGTTGACCTTGCGGGAGTCAACATCATGACTATGGATTACGGTGGC
AGCAGGCGCAGGAAGCACTATGTTAGAAGCTTCTACGGCGGCTGCACGGCGACACA
TGCGCAGCTCGGGGCACTCTATAAGGCCGATGGACAAGATTTTGGCGCCGATTGGTT
GTGGCGGAAAATTGGGCTCACCCCAATCATTGGCCAAAACAGTGTTGCTGGTGAGAT
TTTCACTTTGCAGGATGCCGTAGTCTCCATGATTTCGCTGTGGGAAAGGCGTTGGCC
GCGTCTCCATGTGGTCCTTGAACCGGGATGCCACCTGTGTCCCAACTACCCTGACCT
GACTCGGGTTTCCGACGGGTGCAGCGGCATTGACCAGAAGGGCAAATTGTTCTCAA
CTGTGCTGGGTGAAGGGTTAAGCGTTCTGCCTACGCAATCTGCCACCAGTGCACCGG
CGCAACCCACCGTTCAGTCCACCTTTCCCACGGATAATCCCGCAACGAGTCCTTATC
CGATCTGGAGCGATCTGGCAGTCTATGTAGAAGGAGACAGGATTGTTCTCAACGGCA
ACGTCTACATGGCTAAATGGTGGACTCAGGGAGACGTGCCTGATAACCCAGTGGCAA
CGGACGGGCTGACTCCGTGGCAGCTTATTGGTCCTGTACTTCCCGGGGACAAGCCG

GCCCCGCAAGTGACTGCGCCGGCAGGAACCTATCCGCTGTGGAGTGCTGCGAAGGT
ATATGACCAAGGGGACCGAGTGATGTTTGATGGCCGTATATTCGAGGCTAAATGGTG
GAATCGCGAAGAAAGCCCGGTAGCTTCCCTGCAGGGATCGCCCTCGGCAGCGTGGA
AGTTGTTCTCCAACGCTCAGGTGGCGCAGATCCTGGCCACCCCGGATGGTAAATAA

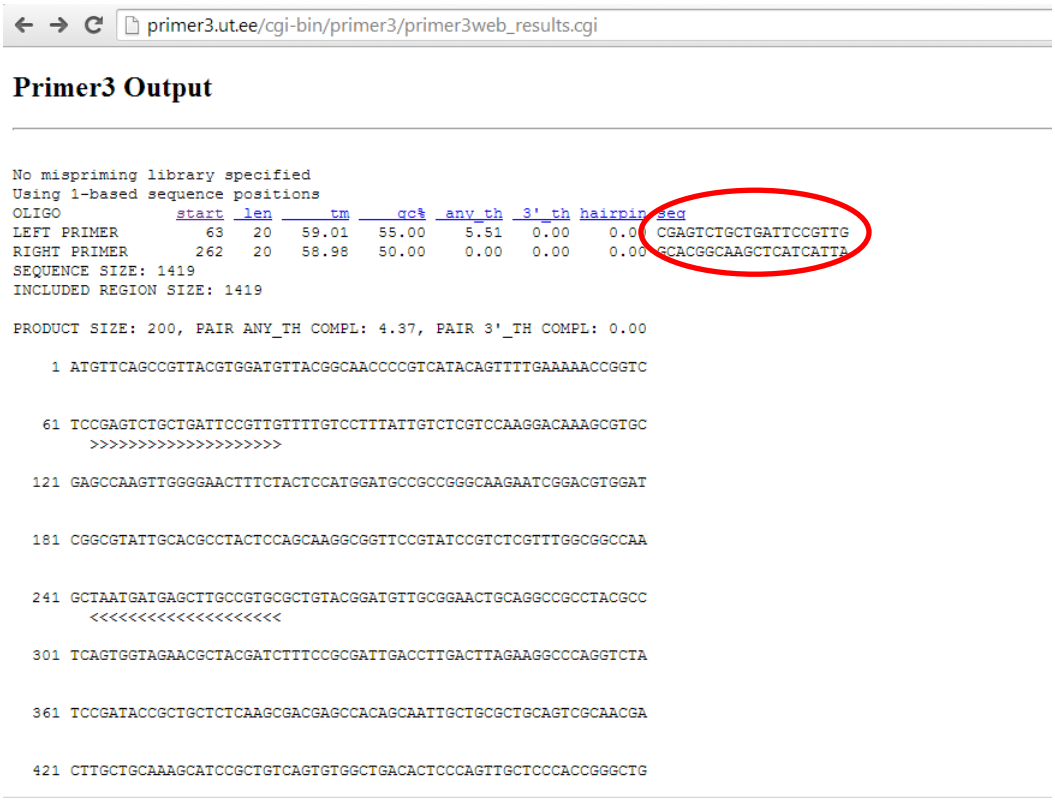
Copia y pega esta secuencia en la ventana del programa Primer3, como se muestra en la siguiente imagen. En la parte inferior de la ventana del programa encontrarás diversas secciones que le permiten controlar una amplia variedad de parámetros relacionados con el diseño de cebadores.



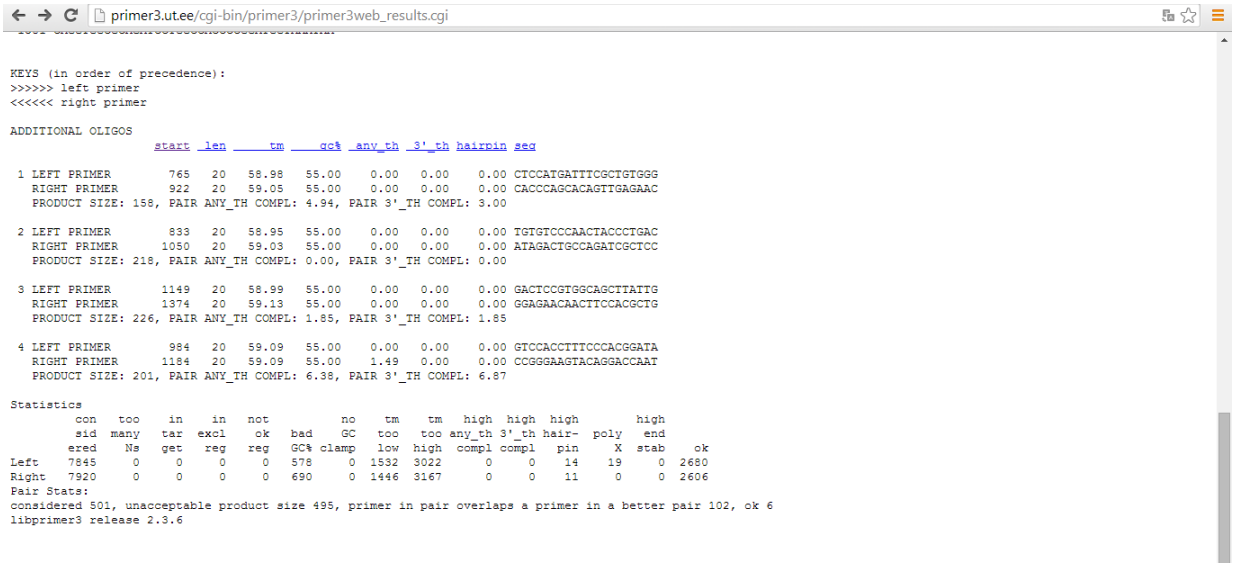
Como puedes observar en los círculos rojos, Primer3 tiene habilitadas las casillas de las opciones para seleccionar el cebador corriente arriba y corriente debajo de nuestra secuencia de interés (*Pick left primer, Pick right primer*). A continuación escoge la opción "*Pick primers*" para que el programa te muestre las opciones de cebadores para la secuencia de consulta.

En la siguiente interfaz aparecen los resultados del diseño de cebadores en donde se muestra, en un círculo rojo, las secuencias que corresponden al cebador sentido y antisentido que fueron generadas por el programa. Así mismo, aparece indicada la región complementaria de los cebadores dentro de la secuencia de consulta y el tamaño del producto de la amplificación (en el caso de nuestro ejemplo es de 200). Además para cada secuencia se indican las siguientes características:

Start (***Inicio***): posición de la base del extremo 5´, con respecto a la secuencia de consulta.
Len: Longitud del cebador en cantidad de nucleótidos.
Tm: temperatura de alineamiento.
GC %: Contenido de GC en porcentaje.
Any th: Auto-complementariedad local la cual refleja la tendencia de los cebadores a hibridar entre su misma secuencia y formar estructuras secundarias.
"*3*'": Alineamiento Global Anclado en 3', que refleja la complementariedad entre los cebadores sentido y antisentido, y se utiliza para predecir la formación de dímeros.



Además de esta opción de cebadores, el programa te muestra en la parte inferior otras cuatro opciones de pares de cebadores con sus respectivas características.



A partir de estos resultados analiza si las secuencias de los oligonucleótidos presentan las características ideales para poder ser utilizados en la técnica de PCR.

Como te pudiste dar cuenta, a partir de estas herramientas bioinformáticas es fácil y rápido poder diseñar cebadores, además de que tienes la oportunidad de seleccionar en el programa características específicas para generarlos de acuerdo a tus necesidades.

2.3.3. Diseño de plásmidos y patrón de restricción

Como recordarás de las asignaturas de Biología Molecular II y Genética Molecular Bacteriana los **vectores** o **vehículos moleculares** son herramientas muy importantes en la ingeniería genética. Un ejemplo de vectores moleculares son los plásmidos, éstos son moléculas constituidas por una doble cadena de ADN circular extracromosomal y tienen la capacidad de replicarse de forma autónoma a la del ADN cromosómico, ya que poseen una secuencia de origen de replicación (Figura 5) (Bolívar Zapata, 2007).

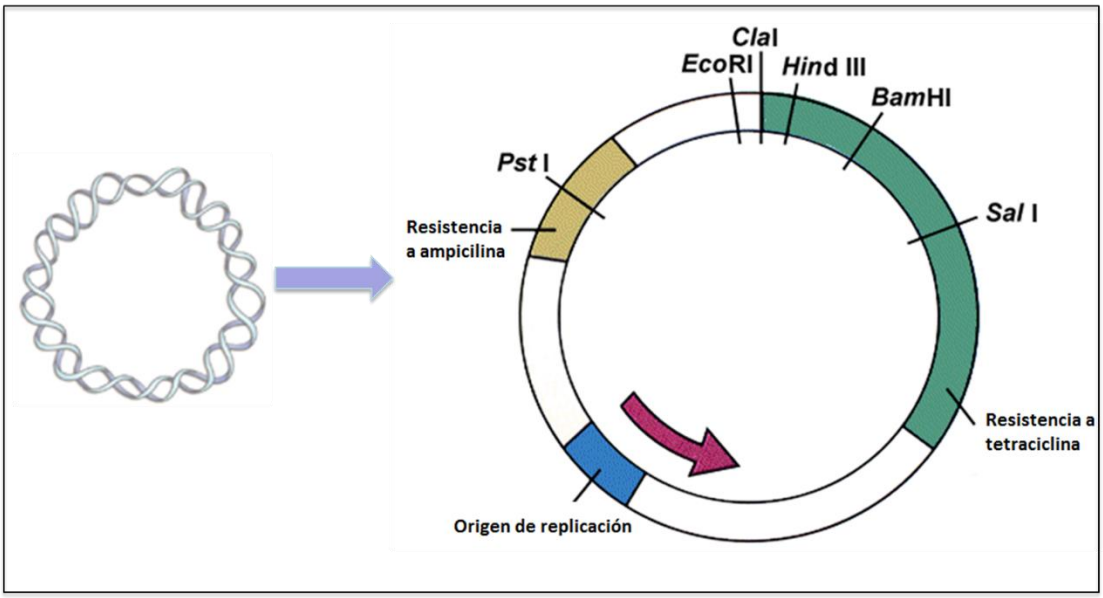


Figura 5. Se representa una cadena de ADN covalentemente cerrada que tiene secuencias importantes para poder fungir como un vector de clonación (Origen de replicación y secuencias de reconocimiento para enzimas de restricción). Además algunas de las secuencias de ADN del plásmido codifican para genes cuyos productos confieren resistencia a antibióticos (ampicilina y tetraciclina).

En general los vectores moleculares deben cumplir con las siguientes características:

a) **Tamaño pequeño**. Para tener una mayor eficiencia de entrada del ADN plasmídico a la célula huésped por el proceso de transformación.
b) **Estabilidad dentro de la célula huésped**.
c) **Una secuencia de origen de replicación (ori)**.
d) **Presencia de genes "marcadores de selección**". Que sean útiles para poder diferenciar entre las células que adquirieron el vector de clonación de aquellas que no lo hicieron, por lo general, los marcadores de selección más comunes son los que confieren resistencia a antibióticos.
e) **Presencia de un "sitio múltiple de clonación" MCS**, por sus siglas en inglés (***M****ultiple* ***C****loning* ***S****ite*) o *poli-linker*, que es la región donde se encuentran los sitios de corte únicos que serán reconocidos por diversas enzimas de restricción (Krebs, 2010).

El diseño de los plásmidos como vectores es muy importante y está en función de objetivos que se quieran alcanzar. Por ejemplo, los **vectores de clonación** se diferencian de los **vectores de expresión** de proteínas, ya que los segundos expresarán al gen clonado, por lo que tienen en su secuencia un sitio de unión a ribosoma o RBS, por sus siglas en inglés (***R****ibosome* ***B****inding* ***S****ite*) que induzca la traducción de proteínas, una

secuencia de inicio de la traducción incluyendo el codón de inicio (ATG) y una secuencia de término de la traducción, incluyendo el codón de paro, mientras que el vector de clonación no tiene estos elementos.

En la figura 6 se muestra el mapa de un vector de clonación, llamado pUC118, donde se señalan los orígenes de replicación: pMB1 ori y f1 ori, el MCS, así como el gen que confiere la resistencia a ampicilina (ampR) además de otros sitios de corte (Bgl I y Pvu I).

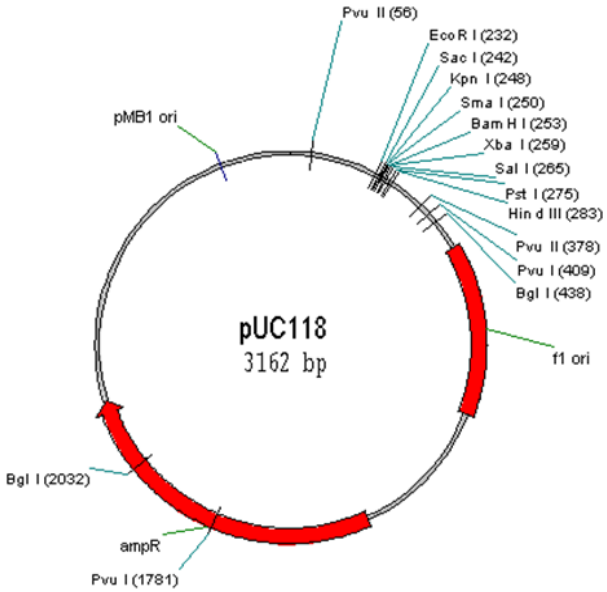
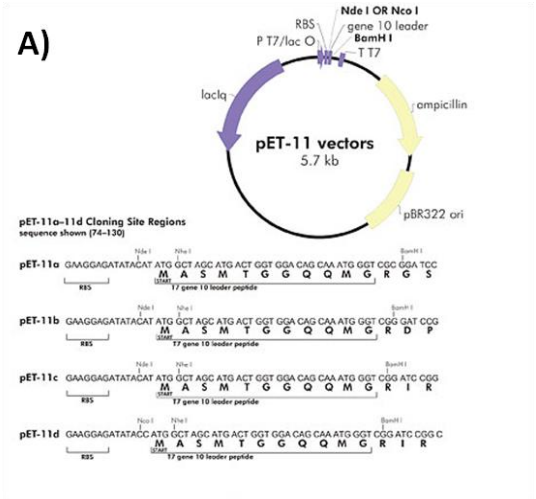


Figura 6. Vector de clonación pUC118. Se indican los orígenes de replicación, el MCS, así como el gen de resistencia a ampicilina.

En la figura 7 se esquematizan dos vectores de expresión. En el pET-11 (Figura 7A) se señalan, además del origen de replicación, el gen de resistencia a ampicilina y el sitio de clonación, el RBS, la región promotora (*T7 promoter*) y la región terminadora (*T7 terminator*), regiones necesarias para la expresión de proteínas. En el vector pIX 4.0 (Figura 7B) se muestran los elementos comunes como el MCS, el origen de replicación y el gen de resistencia a ampicilina, pero también el codón de inicio de la traducción (ATG) y la región terminadora (*T7 Term*).

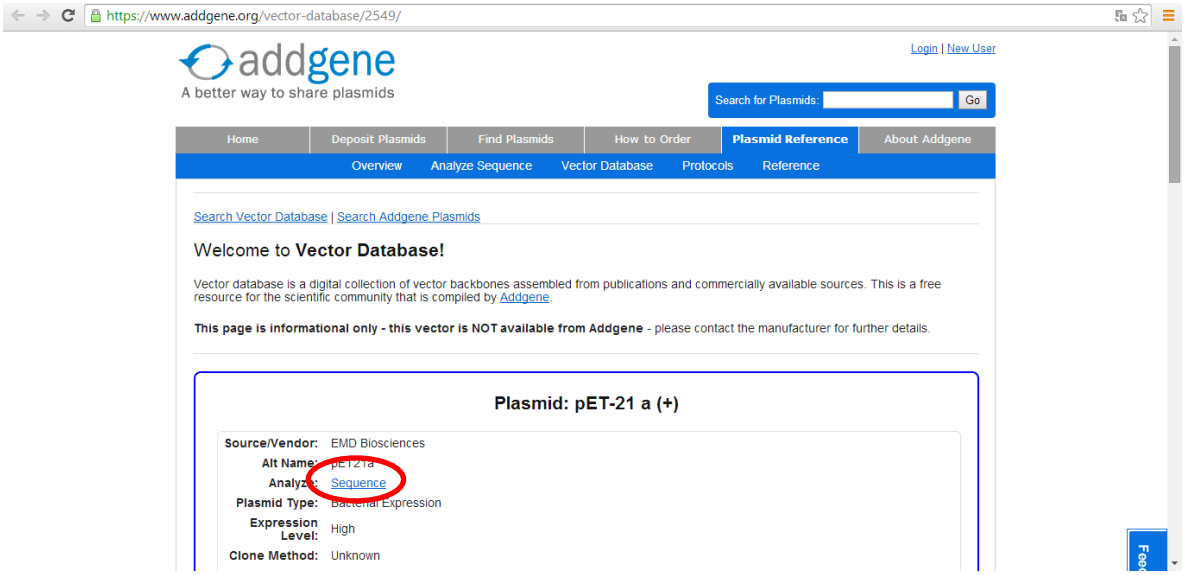


Feature	Nucleotide Position			
	pET-11a	pET-11b	pET-11c	pET-11d
T7 promoter with lac operator	1–43	1–43	1–43	1–43
ribosome binding site (RBS)	74–80	74–80	74–80	74–80
Nde I (pET-11a–c) or Nco I (pET-11d) cloning site	86–91	86–91	86–91	86–91
T7 gene 10 translated leader	89–121	89–121	89–121	88–120
BamH I cloning site	125–130	124–129	123–128	122–127
T7 terminator	199–245	198–244	197–243	196–242
ampicillin resistance (bla) ORF	657–1514	656–1513	655–1512	654–1511
pBR322 origin of replication	1665–2332	1664–2331	1663–2330	1662–2329
lacI^q repressor ORF	4212–5291	4211–5290	4210–5289	4209–5288

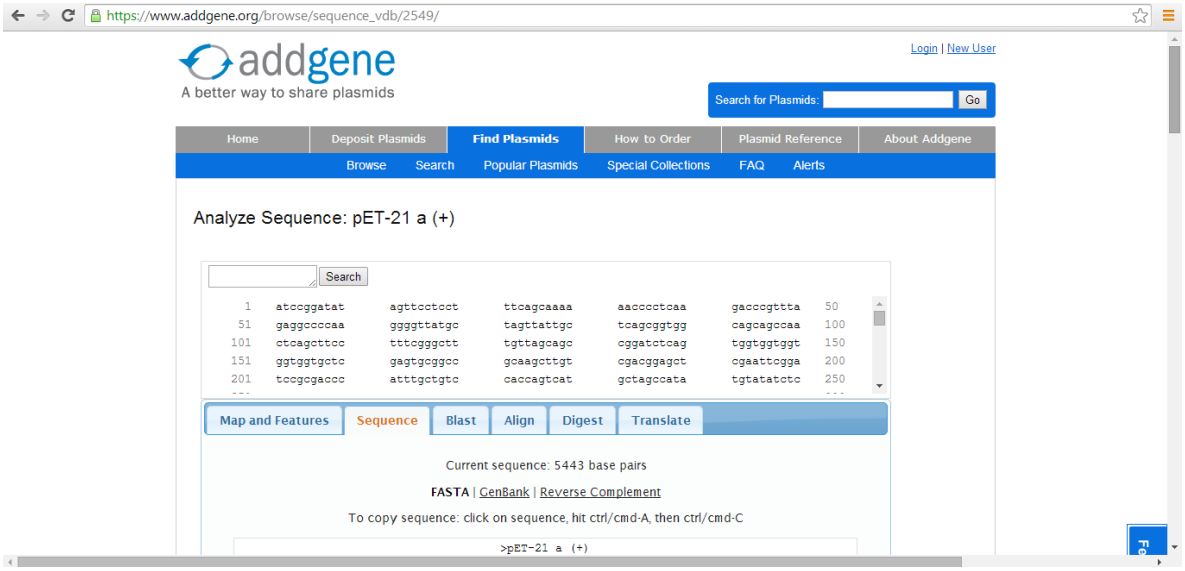
Figura 7. Vectores de expresión. En ambos vectores, A) pET-11 y B) pIX 4.0, se señalan las regiones necesarias para la expresión de genes como por ejemplo el RBS, la región promotora, la región terminadora o el ATG de inicio de la traducción.

Debido a que los plásmidos son muy utilizados como vectores de clonación y de expresión, se cuenta con bases de datos en donde se localizan las secuencias y características de estas moléculas. En la siguiente dirección de internet encontrarás el sitio llamado Addgene en donde puedes buscar secuencias de plásmidos http://www.addgene.org/.

A continuación realizaremos la búsqueda de la secuencia de la secuencia del plásmido pET-21a (+), que corresponde a un vector de expresión. Entra al sitio de Addgene y localiza la secuencia ingresando el nombre del plásmido en la sección de búsqueda.



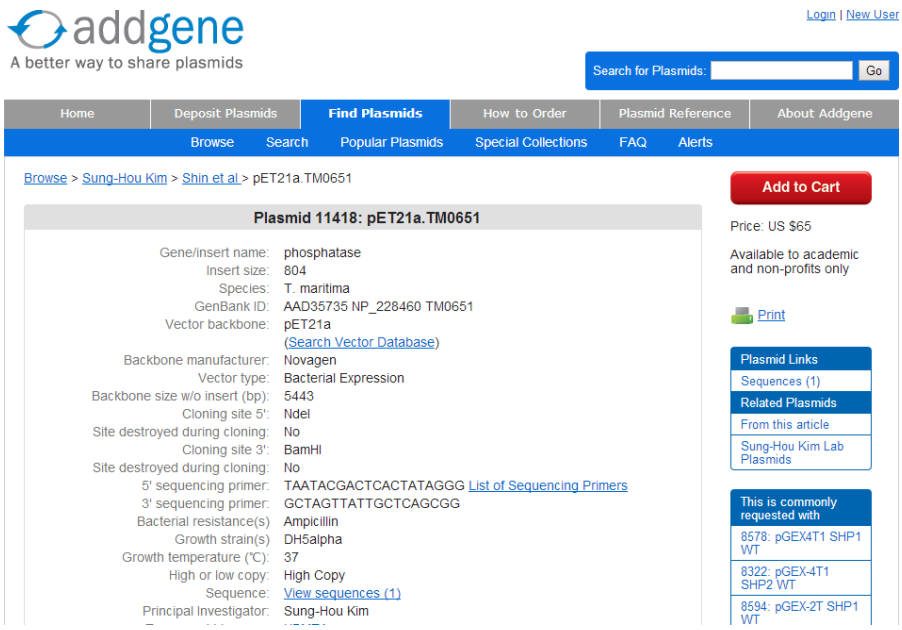
La secuencia del plásmido pET-21 a (+) aparece en una interfaz como la que se muestra a continuación. Tienes la opción de visualizar el mapa lineal y circular del plásmido si seleccionas la pestaña “Map and Features” (mapa y características) o puedes obtener la secuencia como se muestra si seleccionas la pestaña que dice “Sequence” (Secuencia).



Un ejemplo de un plásmido que ya tiene el inserto (gen de la proteína) es el plásmido pET21aTM0651 el cual tiene el gen de una enzima fosfatasa del organismo *T. marítima* (Shin *et al.,* 2003). En el sitio de Addgene puedes encontrar este plásmido con el número

de registro 11418 que corresponde a la secuencia del vector pET21a más la secuencia del gen que codifica para una fosfatasa.

Como puedes observar en la siguiente interfaz se presenta el nombre del plásmido (pET21aTM0651) y su clasificación (11418). Así como el inserto que corresponde al gen de la fosfatasa y que tiene un tamaño de 804 pb. También te presentan el número de acceso en el GenBank. De acuerdo a la información el gen de la fosfatasa se insertó en el sitio múltiple de clonación haciendo cortes con las enzimas de restricción NdeI y BamHI.



En este caso el mapa del plásmido se observa con el inserto de color rojo que corresponde a la secuencia de la fosfatasa que fue clonada en el vector pET-21a.

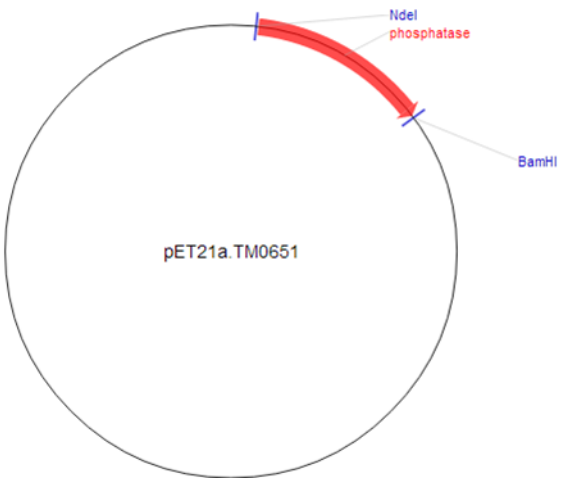
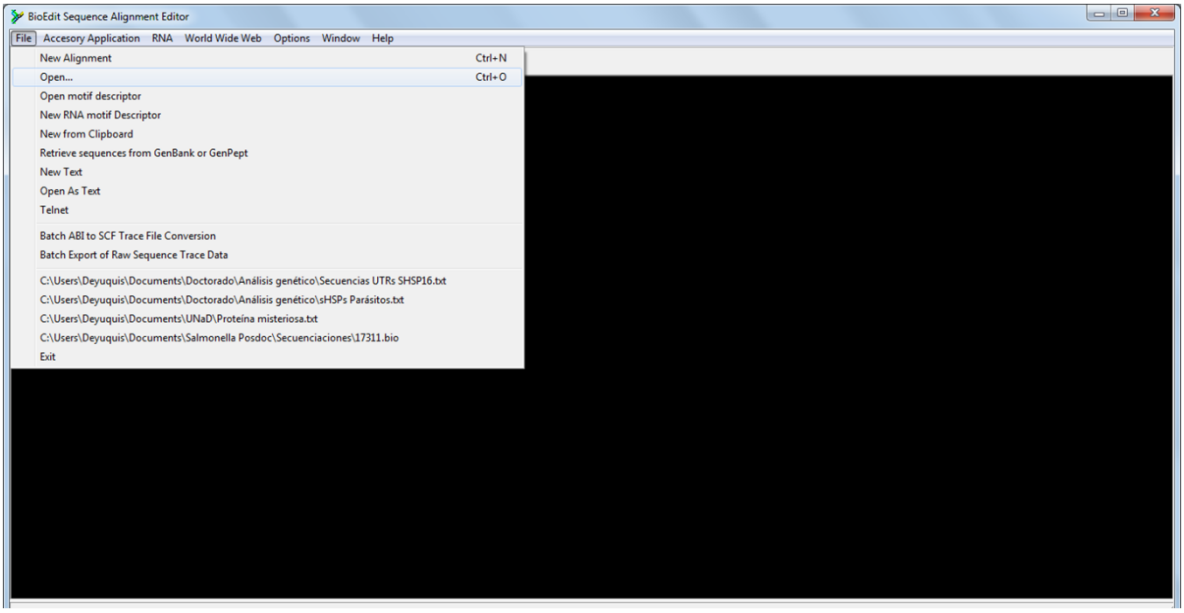


Figura 8. Plásmido pET-21 a con el inserto del gen de la fosfatasa de Thermotoga maritima.

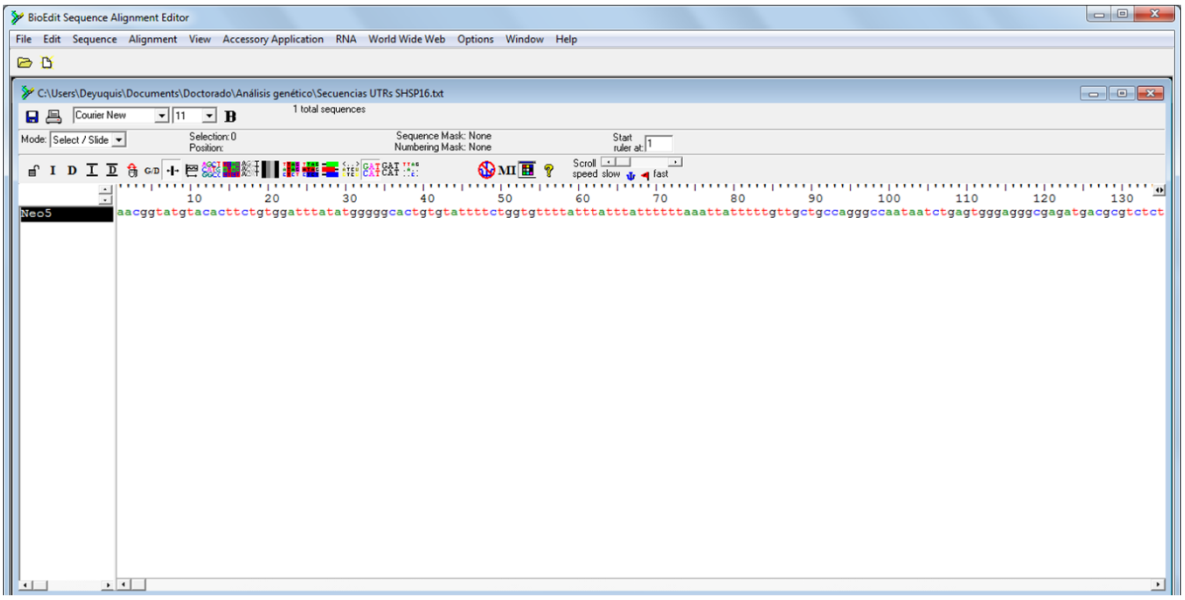
Ahora que ya conoces como buscar secuencias de vectores utilizaremos programas informáticos con el propósito de hacer análisis de plásmidos, en este caso realizaremos la determinación de un **patrón de restricción** y un **mapa circular**. Un **patrón de restricción** se refiere a los sitios de corte que reconocen enzimas de restricción particulares en una secuencia, este patrón puede ser representado en una secuencia linear, denominado mapa de restricción (Krebs, 2010). Determinar el patrón de restricción resulta crucial al momento de clonar genes en algún vector, debido a que conociendo este patrón, podremos escoger las enzimas con las que clonaremos nuestro gen de manera experimental.

En esta sección aprenderemos a usar el editor de secuencias que no requiere intercambiar datos de internet, denominado BioEdit (por sus siglas en inglés, ***Bio****logical Sequence Alignment* ***Edit****or*). La dirección del sitio de internet donde puedes descargarlo e instalarlo en tu computadora de manera gratuita es la siguiente: http://www.mbio.ncsu.edu/bioedit/bioedit.html.

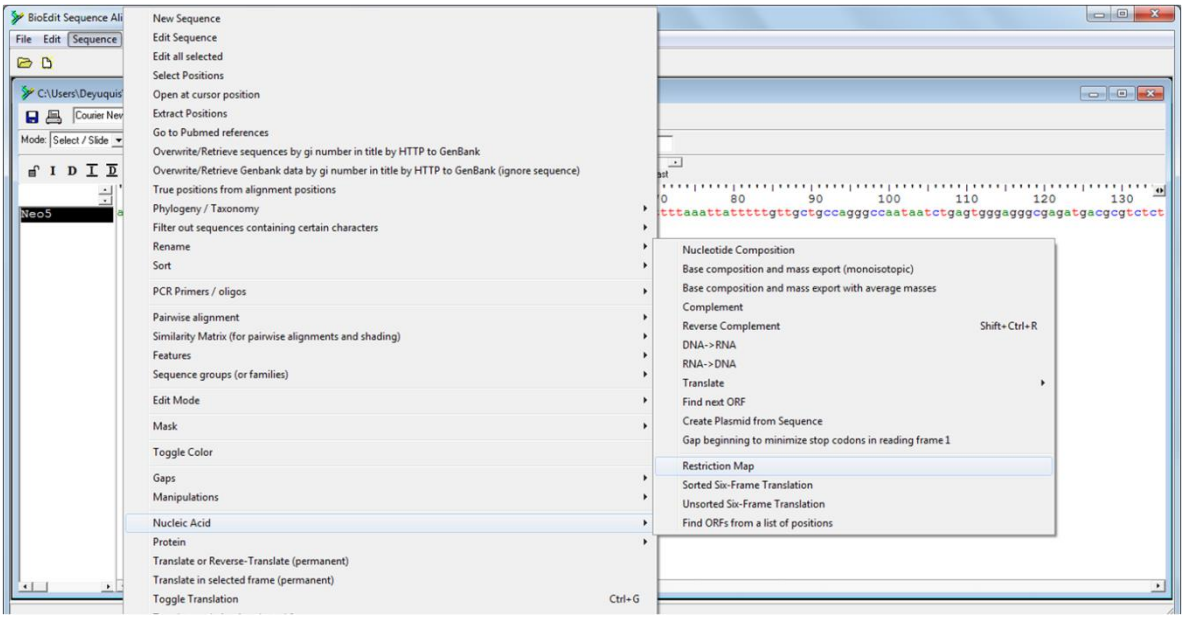
Una vez que has instalado el programa en tu computadora, empecemos a trabajar. Para poder abrir una secuencia de estudio en el editor, es necesario que primero la guardes en un procesador de textos, como el bloc de notas, en formato FASTA. Una vez que tienes tus secuencias guardadas en este formato, puedes acceder a ellas en el Bioedit. Abre el programa y en la pestaña de abrir documentos, selecciona tu archivo.



Una vez abierto tu archivo, tu secuencia se visualizará de la siguiente forma. El nombre de tu secuencia aparece del lado izquierdo, en una columna separada.

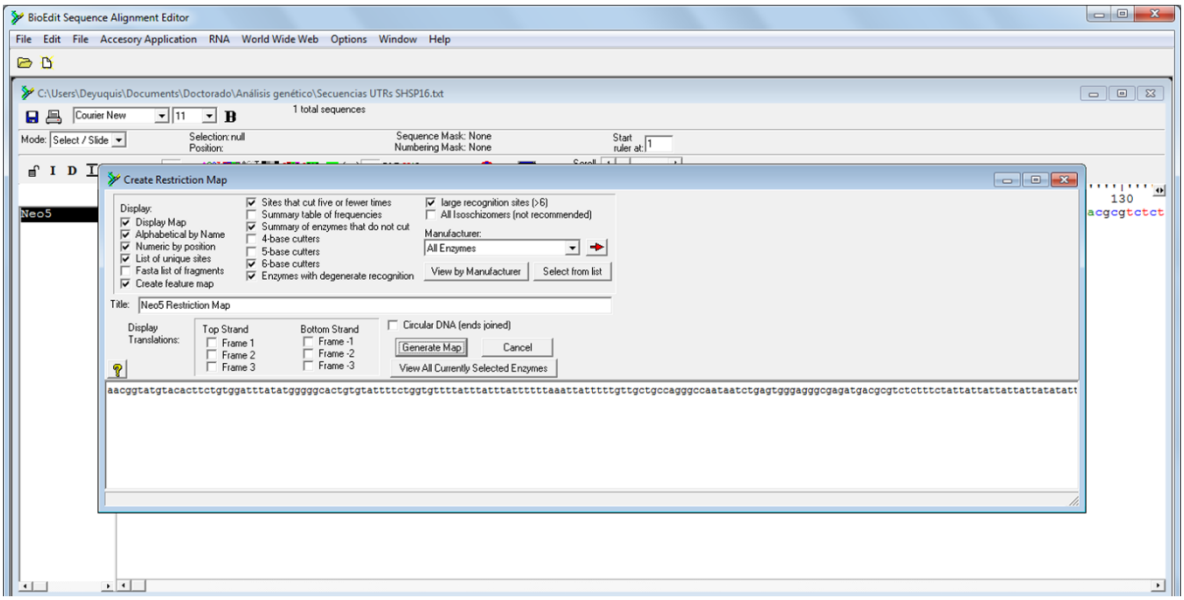


Para generar el patrón de restricción de tu secuencia, en la pestaña **"*Sequence*"**, selecciona la opción **"*Nucleic Acid*"** y enseguida **"*Restriction Map*"**.

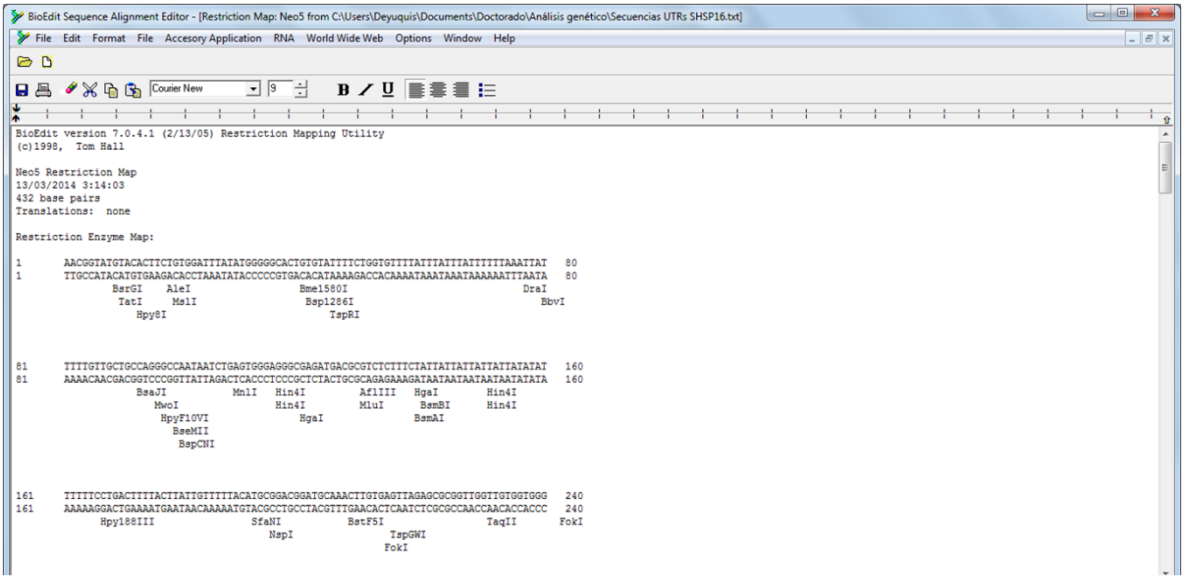


Una vez que diste click, aparecerá la siguiente ventana, donde puedes escoger las enzimas de restricción con las cuales quieres generar el patrón de restricción. Por ejemplo, si sólo te interesa saber si tu secuencia tiene sitios reconocidos por enzimas de

restricción que reconocen secuencias de 6 nucleótidos, puedes escoger esta opción. De lo contario, puedes escoger la opción que te muestre todas las enzimas. También puedes seleccionar que te muestre el nombre de la secuencia, en qué sitio de la secuencia se encuentra, la lista de aquéllas enzimas que no se encuentran en tu secuencia, etc.

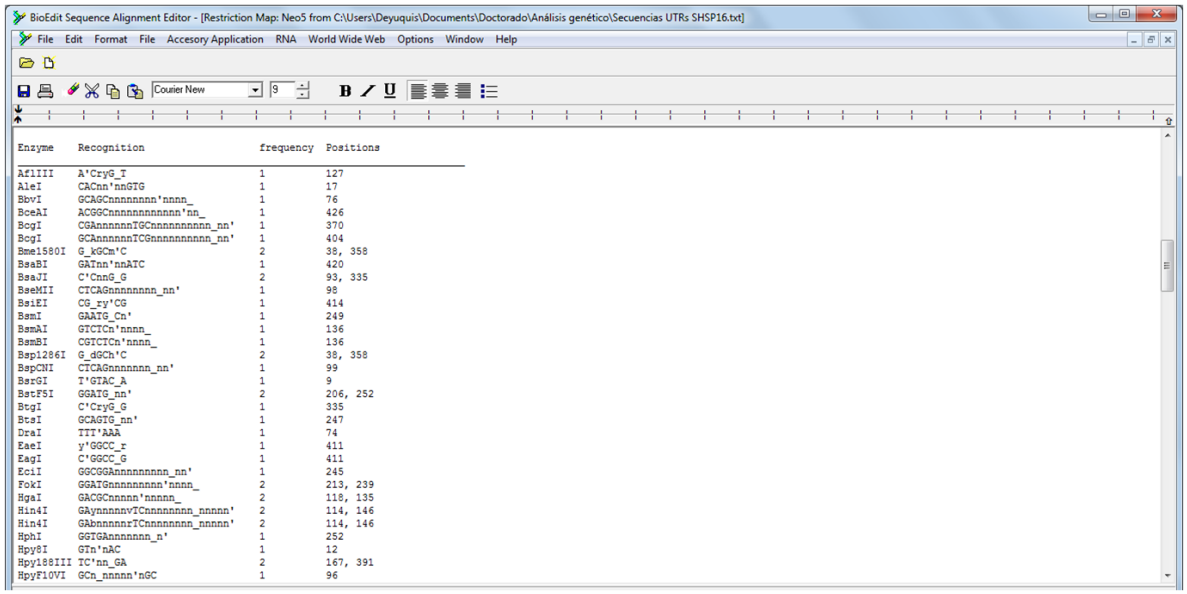


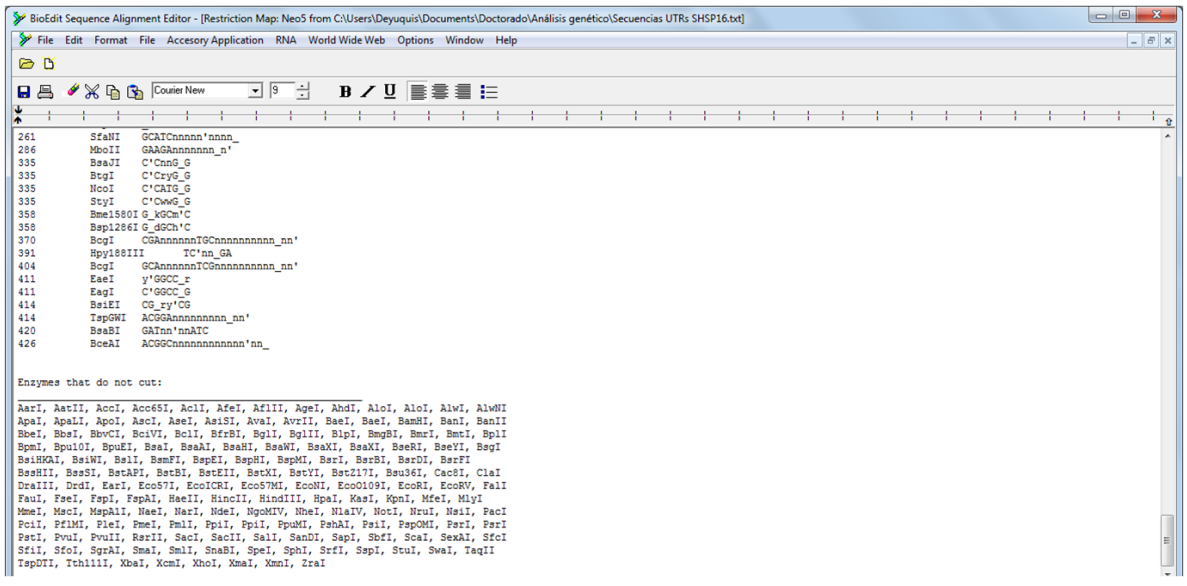
El mapa de restricción generado se muestra como sigue, indicando el sitio de corte y el nombre de la enzima que reconoce esa secuencia.



También se muestra la frecuencia de corte de cada enzima a lo largo de la secuencia, indicando su posición.

Al final se muestra la lista de aquéllas enzimas que no reconocieron su secuencia de corte.



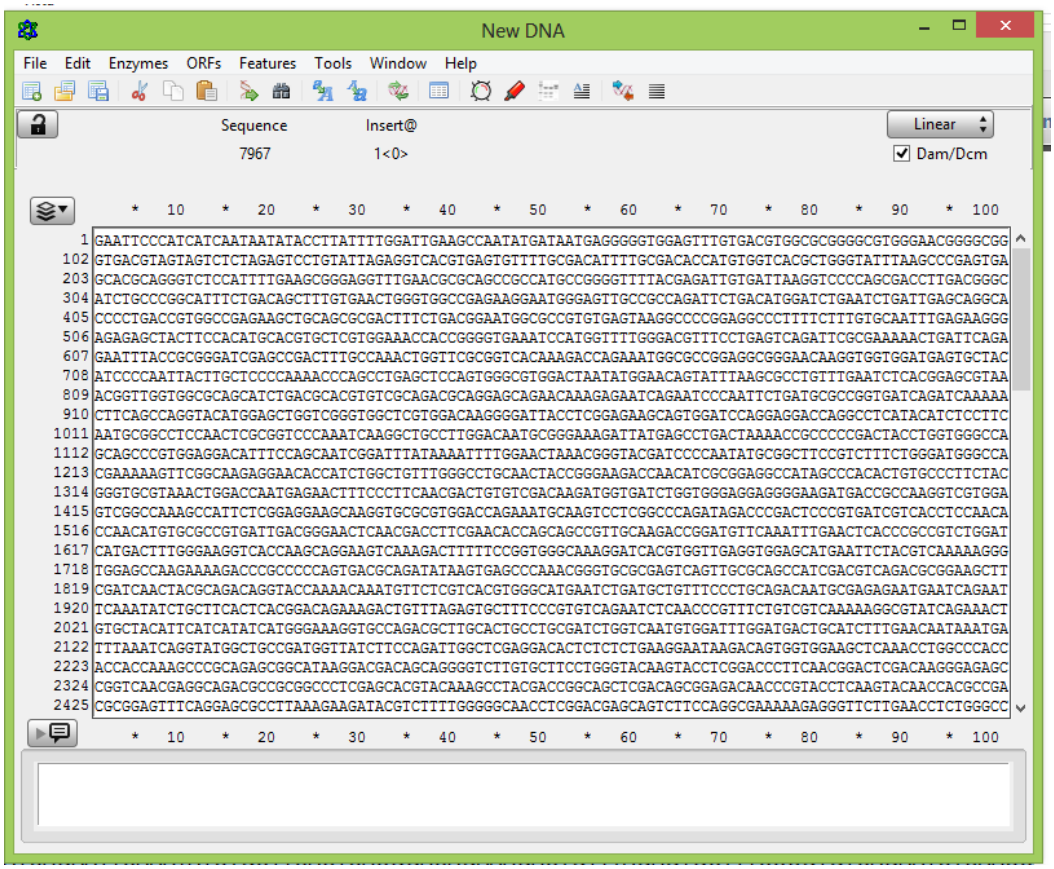


Otro ejemplo útil para el análisis de plásmidos es el software denominado *A Plasmid Editor* (***ApE***) el cual fue creado por Wayne Davis en la Universidad de Utah con el cual se puede analizar un plásmido a partir de su secuencia.

Este programa se puede descargar en la siguiente dirección de internet:
http://biologylabs.utah.edu/jorgensen/wayned/ape/

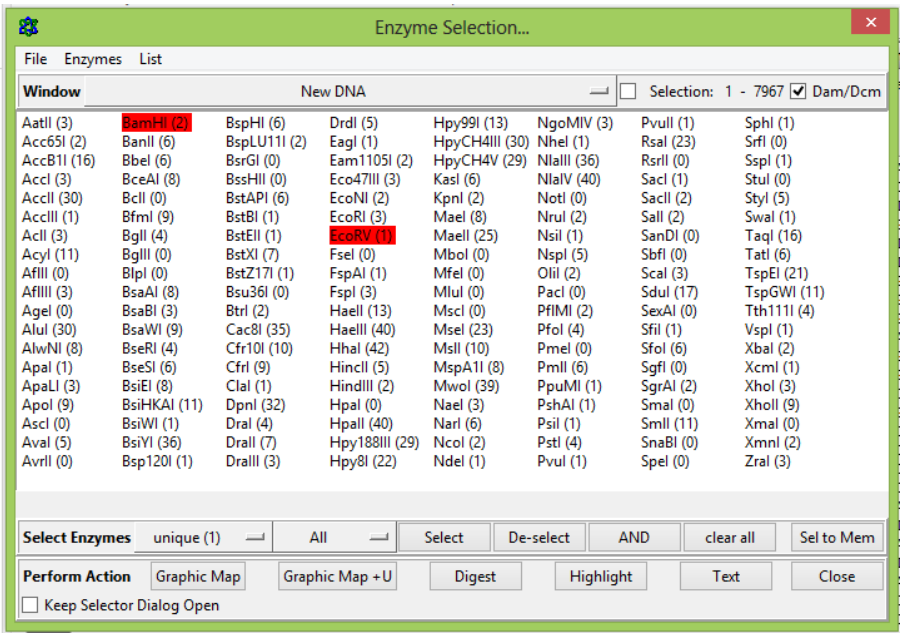
A continuación se muestran algunos de los análisis que se pueden realizar en el programa ***ApE***, utilizando la secuencia del plásmido pBR322 (la secuencia de este plásmido la puedes descargar de la plataforma).

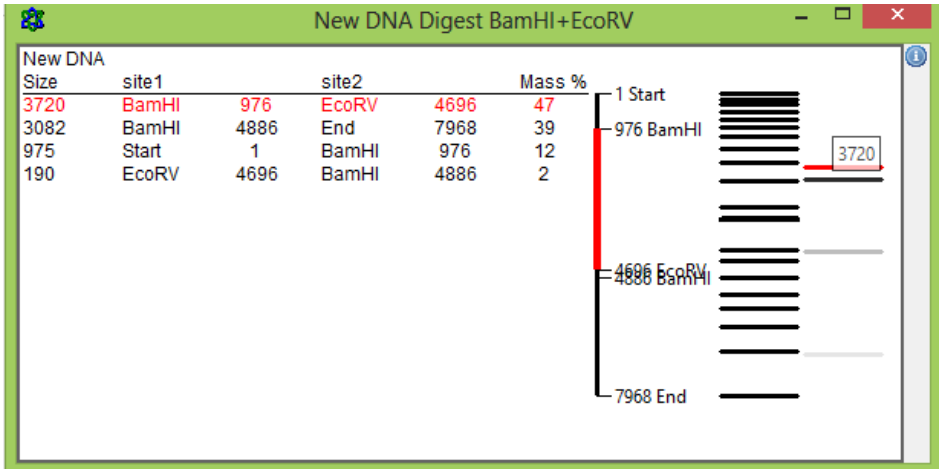
En la primera interfaz se muestra la secuencia del plásmido que se agregó a la ventana del programa.



Si seleccionas la pestaña de “Enzymes” >> “Enzyme Selector” se abre la ventana que muestra una lista con el nombre de diversas enzimas de restricción acompañado de un número que denota el número de sitios de reconocimiento que esa enzima en la secuencia del plásmido que se está analizando, si seleccionamos por ejemplo, EcoRV y BamHI y posteriormente la opción de digerir “Digest”, podemos observar el patrón de corrimiento electroforético en un gel de acuerdo a los tamaños de los fragmentos generados a partir de la digestión con las enzimas. De acuerdo al ejemplo, la selección en

rojo corresponde al fragmento de 3720 pb que se genera del corte entre los sitios de EcoRV (Posición 4696) y BamHI (Posición 976).

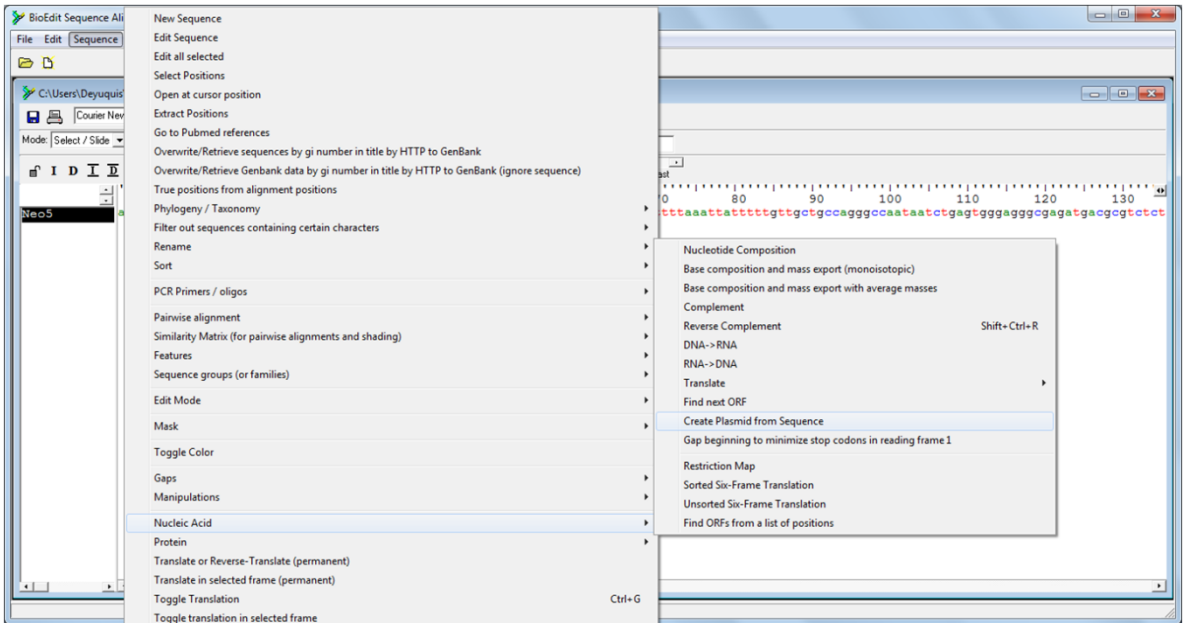




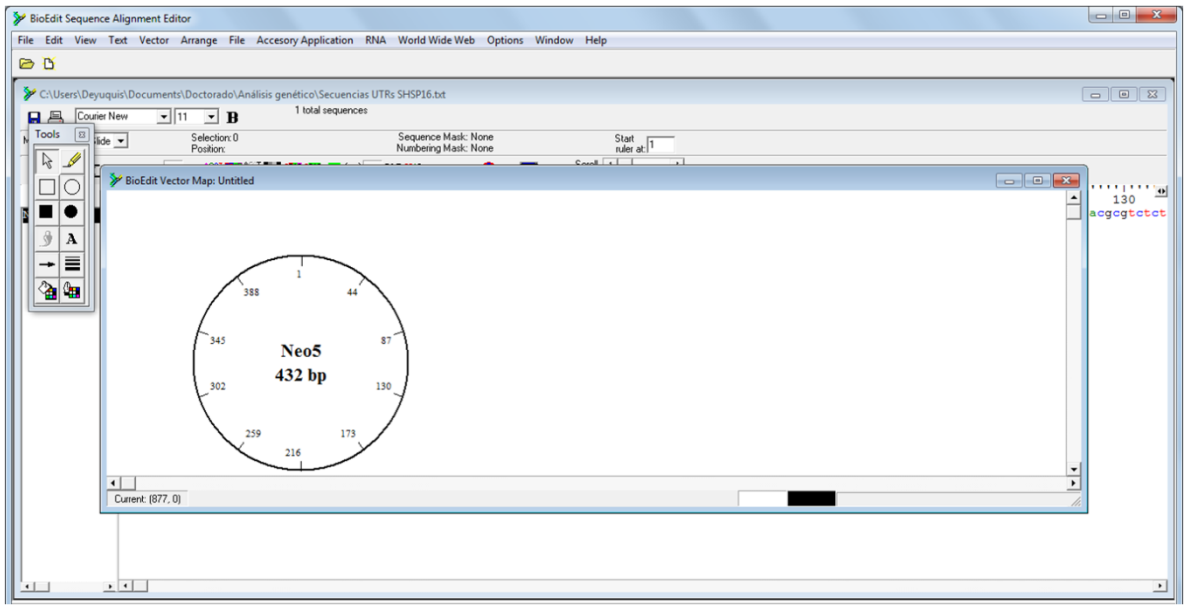
Estas herramientas son muy útiles para poder predecir los tamaños de los fragmentos esperados después de hacer la digestión con enzimas de restricción y para poder analizar plásmidos que ya tengan un inserto de algún gen.

Otra forma de visualizar un mapa de restricción es convirtiéndolo a **un mapa circular**. Por ejemplo, si queremos clonar algún gen en un vector de clonación, podemos generar nuestro propio plásmido utilizando al secuencia del vector más nuestra secuencia de clonación, con el objeto de tener un mapa más fácil de visualizar.

Así, regresando al programa BioEdit, crearemos un mapa circular con la misma secuencia anterior. Nuevamente abre la secuencia de interés con la opción abrir documento. En la pestaña **"*Sequence*"**, selecciona la opción **"*Create Plasmid from Sequence*"**.



Enseguida se mostrará el mapa circular generado, abarcando toda tu secuencia. Se muestra el nombre de tu secuencia y la longitud de ésta. Intenta cambiar los colores del mapa circular y el tipo de letra, entre otras acciones que puedes realizar con la herramienta **"*Tools*"** que se encuentra en una pestaña de lado izquierdo.



¡Muy bien!, acabas de crear un mapa circular con una secuencia genética. Dicho mapa podría representar tu vector de clonación con el que harás diversos experimentos, al cual le puedes introducir distinto tipo de información y agregar a tu bitácora de trabajo. Piensa acerca de qué otros usos podrías darle a la creación de mapas circulares.

2.4. Transcriptómica

Para finalizar con esta unidad en donde hemos analizado con diferentes herramientas secuencias de ADN, ahora revisaremos el tema de transcriptómica que involucra, como lo viste en la unidad 1, el estudio de la molécula de RNAm.

Antes de la secuenciación de los genomas de varios organismos incluidos los del nemátodo *C. elegans* y del humano, se consideraba que el número de genes era uno de los factores que se relacionaba con mayor complejidad del organismo. En este caso comparando el genoma de *C. elegans*, un gusano microscópico, con el genoma del humano la predicción era que el humano tendría mayor número de genes. Sin embargo, después de la secuenciación del genoma del nemátodo y el primer bosquejo del genoma humano, se descubrió que la predicción era incorrecta y que la mayor parte del genoma no codificaba para proteínas. El hecho de que el genoma humano esté compuesto de una proporción de entre el 66-69% de elementos repetitivos, es decir secuencias de kilo bases de nucleótidos que se repiten a través del genoma (Koning *et al.*, 2011), la comunidad científica vio que su predicción había fallado e incluso se llegó a llamar “ADN basura” a los nucleótidos no codificantes.

Se sabe que todas las células de un organismo tienen el mismo ADN, sin embargo un hepatocito es diferente morfológica y fisiológicamente a una neurona, ¿Si tienen el mismo ADN, por qué son diferentes? La respuesta es porque no expresan los mismos genes, ni al mismo tiempo. Para que un gen se exprese, necesita de secuencias accesorias (promotores, represores, sitios de unión de remodeladores de la cromatina, etc.) que recluten a la maquinaria de transcripción cuando se requiere. Del esfuerzo de comprender como se regula la transcripción de los genes, nacieron aproximaciones que se apoyaron en la nueva tecnología de secuenciación masiva. Uno de estos esfuerzos es la proteómica, que tiene como objetivo analizar las proteínas de un tipo de célula, tejido u organismo (Orengo, Jones y Thornton, 2003). Esta aproximación puede dar información parcial sobre la expresión génica, ya que el ARNm es necesario para la síntesis de proteínas, sin embargo como ya se había mencionado, las proteínas pueden sufrir modificaciones post-traduccionales, las cuales no están contempladas en la información del ARNm.

Una aproximación para conocer los genes que codifican proteínas presentes en el genoma es el **transcriptoma**. El transcriptoma tiene como meta analizar la totalidad de

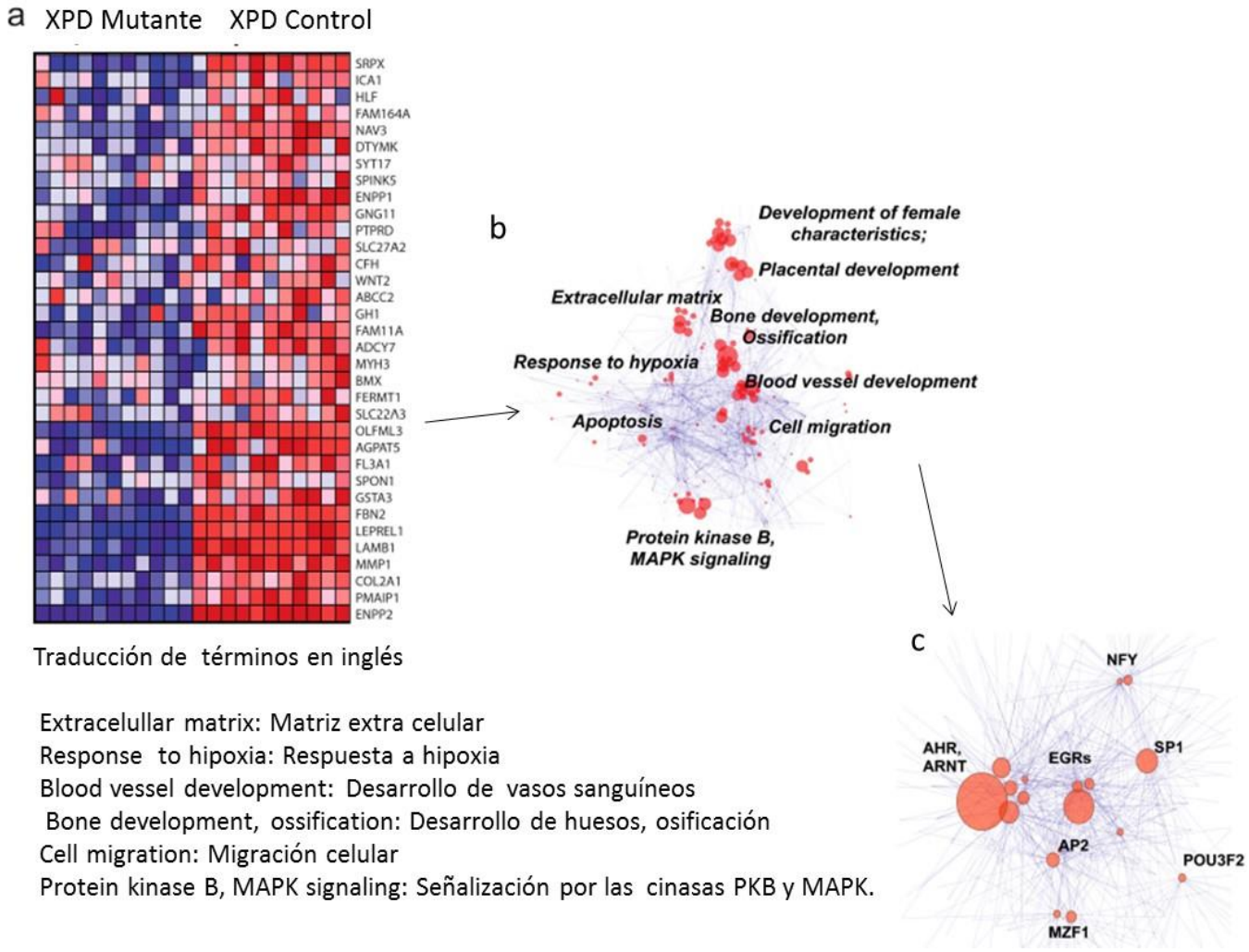
los ARNm que se expresan en una célula o tejido en una etapa específica del desarrollo (Vallin, 2007). Al tomar en cuenta la totalidad de los ARNm, se puede inferir en que momento del desarrollo y en qué tejido se activa la transcripción de un gen. Esto es de utilidad en el diagnóstico y estudio de padecimientos ya que los cambios en la transcripción de genes pueden estar asociados a enfermedades como el cáncer o infecciones virales.

Los ARNm se pueden diferenciar del resto de ARN, debido a que tienen añadida una cadena de adeninas, la cual se puede usar como sitio para diseñar los cebadores. Lo anterior resulta bastante conveniente ya que una vez que se puede diferenciar los ARNm del resto de los ARN y ADN, se puede amplificar con la ayuda de la enzima viral llamada transcriptasa reversa, que sintetiza ADNc a partir de ARN. El ADNc puede ser amplificado por PCR e identificado y de esta manera se pueden comparar las secuencias codificantes entre tejidos u organismos e investigar homología, filogenia, equivalencia entre animales modelo y el humano, etc.

El transcriptoma a diferencia del genoma se modifica continuamente en respuesta a cambios en las condiciones microambientales celulares o de los tejidos, por lo tanto, la interpretación del transcriptoma require hacer comparación de una muestra control con el objetivo de estudiar para saber qué diferencias existen entre las diferentes condiciones (Vallin, 2007). El estudio mediante **microarreglos** permite esta comparación, esta tecnología está basada en la hibridación de secuencias. En primer lugar se debe aislar el ARNm de ambos tejidos que se quieran comparar y, a partir de cada uno de ellos, obtener sus correspondientes ADNc. Estas moléculas de ADNc deben marcarse con un compuesto fluorescente, que será diferente en los tejidos objeto de estudio y el control. Se incuban los ADNc con secuencias de ADN y cuanto mayor sea la hibridación de una especie determinada de ADNc marcado con el ADN del microarreglo, mayor será la expresión del ARNm que proviene de la muestra de estudio (Vallin, 2007).

Uno de los estudios más extensivos para conocer el transcriptoma fue hecho por el consorcio FANTOM del instituto RIKEN en Japón, que describió el transcriptoma de una cepa de ratón, haciendo uso de sus bibliotecas de ADN clonado. Para el análisis de la gran cantidad de información derivada de ya sea la secuenciación masiva o los micro arreglos es necesario utilizar herramientas bioinformáticas para la identificación de los genes y su clasificación (Okazaki, *et al.,* 2002).

En la figura 9 se ilustra un estudio del transcriptoma en el que comparan el impacto en la regulación de genes en muestras de fibroblastos por la proteína XPD. Con este tipo de análisis se puede determinar las causas detrás de cambios en el transcriptoma de una célula, tejido, organismo, etc. (Maslehi, et al., 2013).



Traducción de términos en inglés

Extracelullar matrix: Matriz extra celular
Response to hipoxia: Respuesta a hipoxia
Blood vessel development: Desarrollo de vasos sanguíneos
Bone development, ossification: Desarrollo de huesos, osificación
Cell migration: Migración celular
Protein kinase B, MAPK signaling: Señalización por las cinasas PKB y MAPK.

Figura 9. Ejemplo de un estudio de transcriptoma. A) Mapa de calor en el que se muestra la que hay una menor expresión de 34 genes en fibroblastos con una versión mutada de XPD. B) Análisis de ontología de genes derivado de los genes regulados a la baja descritos en el mapa de calor. C) Análisis de factores de la transcripción derivado de la información obtenida con el análisis de ontología de genes. (Modificada de Maslehi, et al., 2013).

Con este tema terminamos la Unidad 2, pero no olvides practicar y explorar los programas por tu cuenta, verás que es muy interesante descubrir toda la información que estas herramientas te pueden dar.

Actividades

La elaboración de las actividades estará guiada por tu figura académica, mismo que te indicará, a través de la Planificación de actividades, la dinámica que tú y tus compañeros (as) llevarán a cabo, así como los envíos que tendrán que realizar.

Para el envío de tus trabajos usarás la siguiente nomenclatura: BIIN_U2_A1_XXYZ, donde BIIN corresponde a las siglas de la asignatura, U2 es la etapa de conocimiento, A1 es el número de actividad, el cual debes sustituir considerando la actividad que se realices, XX son las primeras letras de tu nombre, Y la primera letra de tu apellido paterno y Z la primera letra de tu apellido materno.

Autorreflexiones

Para la parte de **autorreflexiones** debes responder las *Preguntas de Autorreflexión* indicadas por tu figura académica y enviar tu archivo. Cabe recordar que esta actividad tiene una ponderación del 10% de tu evaluación. Para el envío de tu autorreflexión utiliza la siguiente nomenclatura: BIIN_U2_ATR _XXYZ, donde BIIN corresponde a las siglas de la asignatura, U2 es la unidad de conocimiento, XX son las primeras letras de tu nombre, y la primera letra de tu apellido paterno y Z la primera letra de tu apellido materno

Cierre de la unidad

A lo largo de esta unidad, el alumno se percató de la utilidad de las bases de datos y software para el análisis de secuencias de ADN. Aprendió a acceder a secuencias de nucleótidos y analizarlas, interpretando los resultados obtenidos. Entre las herramientas bioinformáticas utilizadas se encuentran el alineamiento de secuencias, la determinación del porcentaje de identidad, la generación del patrón de restricción de una secuencia, diseño de cebadores, construcción de un mapa circular a partir de una secuencia, búsqueda y determinación del contenido GC. Los datos proveídos por estas herramientas permiten tener un mayor número de datos biológicos de esa secuencia particular, aún sin habernos trasladado a un laboratorio de investigación. Sin embargo, no debes olvidar que si bien las herramientas bioinformáticas son muy valiosas a ese respecto, la mayoría de las veces tendremos que comprobar nuestras hipótesis en el laboratorio. Por el momento nos enfocamos en secuencias de ADN. En la siguiente unidad, nos adentraremos en el análisis de secuencias de aminoácidos, un mundo también fascinante de la biología.

Fuentes de consulta

- Baxevanis, A. E. & Ouellette, B. F. F. (Eds.). (2001). *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins. Second* Edition. New York: John Wiley & Sons, Inc.
- Bolívar Zapata F.G. (Editor). (2007). *Fundamentos y casos exitosos de la biotecnología moderna.* México, El Colegio Nacional.
- Brown S.M. (2000). *Bioinformatics. A biologist´s guide to biocomputing and the internet.* New York: Eaton Publishing.
- Buehler L.K., Rashidi H.H. (Eds.) (2005). *Bioinformatics Basics. Applications in Biological Science and Medicine.* Second Edition. EUA: Taylor & Francis.
- Claverie, Jean-Michel & Notredame, Cedric. (2007). *Bioinformatics for Dummies*. Second Edition. Indiana. Wiley Publishing Inc.
- Dieffenbach, C.W., Lowe, T.M.J., Dveksler, G.S. *General Concepts for PCR Oligonucleótido Design, in PCR Oligonucleotides,* A Laboratory Manual, Dieffenbach, C.W, and Dveksler, G.S., Ed., Cold Spring Harbor Laboratory Press, New York, 1995, 133-155.
- Gutteling, E. et al. (2006). *Mapping phenotypic plasticity and genotype–environment interactions affecting life-history traits in* Caenorhabditis elegans*.* Heredity 98, 28–37.
- Hartl D. L., Jones E.W. (1998). *Genetics. Principles and Analysis (4º Edition).* EUA: Jones and Bartlett Publishers.

- Konietzny, U. & Greiner, R. (2003). *The application of PCR in the detection of mycotoxigenic fungi in foods.* Braz. J. Microbiol. 34:283-300

- Koning J., Wanjun G., Castoe T., Batzer M., Pollock D. (2011). *Repetitive elements may comprise over two-thirds of the human genome.* PLoS Genet 7(12): e1002384. doi:10.1371/journal.pgen.1002384.

- Koressaar T, Remm M (2007) *Enhancements and modifications of primer design program Primer3 Bioinformatics* 23(10):1289-9.
- Main-Hester, K. L., Colpitts, K. M., Thomas, G. A., Fang, F. C., Libby, S. J. (2008). Coordinate regulation of *Salmonella* pathogenicity island 1 (SPI1) and SPI4 in *Salmonella enterica* Serovar *Typhimurium*. Infect Immun. 76(3):1024-1035.

- Krebs, J. E., Goldstein, E. S., Kilpatrick, S. T. (2010). *Lewin's essential genes.* Second Edition. Massachusetts: Jones and Bartlett Publishiers.

- Maslehi R., Mills J., Signore C., Kumar A., Ambroggio X., Amiran D. (2013*). Integrative transcriptome analysis reveals dysregulations of canonical cancer molecular pathways in placenta leading to preclamapsia.* Scientific reports, 3:2407.

- Mount, David. (2001). *Bioinformatics. Sequence and Genome Analysis*. New York. Cold Spring Harbor Laboratory Press.

- Novagen. (2003). *pET System Manual.* Novagen Catalog. https://web.archive.org/web/20150226010933/http://richsingiser.com/4402/Novagen%20pET%20system%20manual.pdf

- Okazaki, Y., et al. (2002). *Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs*. Nature 420:563–573.

- Orengo C.A., Jones D.T., Thornton J.M. (2003). *Bioinformatics: Genes, proteins & computers.* New York: BIOS Scientific Publishers.

- Reece R.J. (2004). *Analysis of genes and genomes.* UK: John Wiley & Sons, Ldt.

- Sambrook J., Russell D. (2001). *Molecular cloning: A laboratory manual.* Cold spring harbor, tercera edición. Nueva York, USA. Segundo volumen, sección 8.13.

- Schuler, G. Sequence Alignement and Database searching en: Baxevanis A.F., Ouellette, F. (2001). *Bionformatics: A practical guide to the analysis of genes and proteins*. 3ª Ed. UK: Wiley.

- Sharrocks, A.D. *The design of oligonucleótidos for PCR*, in *PCR Technology, Current Innovations*, Griffin, H.G., and Griffin, A.M, Ed., CRC Press, London, 1994, 5-11.

- Shin D.H., Roberts A., Jancarik J., Yokota H., Kim R., David E. W. and Kim S. H. (2003). *Crystal structure of a phosphatase with a unique substrate binding domain from* Thermotoga marítima*.* Protein Science, 12:1464–1472.

- Untergrasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG (2012) *Primer3 - new capabilities and interfaces*. *Nucleic Acids Research* 40(15):e115.

- Vallin Plous C. (2007). *Microarreglos de ADN y sus aplicaciones en investigaciones biomédicas.* Revista CENIC. Ciencias Biológicas, Mayo-Agosto, 132-135.

- Westhead D.R., Parish J.H., Twyman R.M. (2002). *Instant Notes Bioinformatics.* New York. BIOS Scientific Publishers

- Yang, Xiaohan, Schffler, Brian, Weston., Leslie. (2006). *Recent developments in primer design for DNA polymorphism and mRNA profiling in higher plants*. Plants methods. 2:4.

- *http://kinase.com/blast/docs/newoptions.html*
- *https://web.archive.org/web/20171120205425/http://viroblast.dbi.udel.edu/CHO/parameters.php#*
- *https://www.institutoroche.es/biotecnologia/bioinformatica*
- *http://www.ebi.ac.uk/ena/about/about*