UPR – RECINTO DE RÍO PIEDRAS
FACULTAD DE CIENCIAS NATURALES
DEPARTAMENTO DE BIOLOGÍA

LABORATORIO DE BIOTECNOLOGÍA (BIOL. 3365)

Análisis del gen AmyE mediante el uso de herramientas de bioinformática

A. Objetivos generales:

Analizar secuencias de ADN utilizando herramientas de bioinformática que se encuentran en la página electrónica del *National Center for Biotechnology Information* (NCBI).

Analizar estructura primaria para predecir estructura secundaria y terciaria

B. Objetivos específicos:

1. Hallar la secuencia del gen AmyE de *B. licheniformis* utilizando el programa GenBank®.
2. Identificar genes homólogos al gen AmyE utilizando la herramienta *Basic Local Alignment Search Tool* (BLAST).
3. Encontrar el "Open Reading Frame" (ORF) para el gen AmyE.

C. Introducción

La bioinformática es un campo de la ciencia que integra en una sola disciplina la biología, las ciencias de cómputos y la tecnología de información. El manejo y la organización de la información biológica han sido de vital importancia debido al incremento anual en la adquisición de información genética.

El *National Center for Biotechnology Information* (NCBI) fue establecido en noviembre de 1988 como una división del *National Library of Medicine* (NLM) en el *National Institutes of Health* (NIH) como una respuesta del gobierno de los Estados Unidos a la necesidad de más y mejores métodos de procesamiento de información biológica. Este juega un rol central en la creación de bases de datos públicas, investigaciones en el área de la biología computacional, desarrollo de programas para analizar datos de genoma y la diseminación de información biomédica mediante *PubMed* y *PubMed Central*. Algunas de las bases de datos contenidas en el NCBI son *GenBank*® y OMIM (*Online Mendelian Inheritance in Man*).

GenBank®

GenBank® es una colección pública e internacional de secuencias de ADN del NIH mantenido como un consorcio entre el NCBI, el *European Molecular Biology Laboratory* (EMBL) y el *DNA Data Bank of Japan* (DDBJ). Actualmente posee una base de datos de secuencias de ADN de 100 gigabases. Este sistema permite hacer búsquedas de secuencias de ADN y descargar archivos en computadoras personales para analizar posteriormente.

BLAST (Basic Local Alignment Search Tool)

BLAST es una herramienta para comparar secuencias de nucleótidos o proteínas con todas las secuencias contenidas en la base de datos del NCBI. Entre algunos de los programas de búsqueda que BLAST tiene se encuentran el blastn, el blastp y el blastx. El programa blastn compara el ADN de interés ("*query"*) con secuencias similares de ADN en la base de datos. El programa blastp compara la proteína de interés con secuencias de aminoácidos similares, mientras que el programa blastx compara la secuencia de nucleótidos de interés traducida con secuencias proteicas en la base de datos.

El análisis de homología que realiza este programa es de vital importancia para la investigación porque, si ciertas secuencias son similares (≥ 25% de homología para proteínas y ≥ 70% de homología para nucleótidos en amino ácidos mayores de 100 unidades), se pueden establecer relaciones evolutivas entre genes.

ORF Finder (Open Reading Frame)

Un gen es una secuencia de ADN que codifica para proteínas o ARN (i. e., tRNAs, rRNAs y snRNAs). Los genes se componen principalmente de regiones reguladoras, promotores, exones, intrones, "open reading frames", entre otros. Un ORF es una secuencia de ADN dentro de un gen que puede traducirse en una proteína. Este se encuentra entre la secuencia que codifica para el codón de iniciación (ATG) y uno de los tres posibles codones de terminación (UAA, UAG o UGA). El programa ORF Finder se especializa en identificar las posibles regiones que codifican para proteínas. Este es el primer paso para comprender la función de un genoma.

En este experimento utilizaremos algunas de las herramientas de bioinformática que se encuentran en la página electrónica del NCBI. Primero, se buscará la secuencia del gen AmyE en *B. licheniformis*, luego, se determinará si este gen es homólogo a cualquier otro gen u organismo usando el programa BLAST y, finalmente, mediante el programa ORF Finder, hallaremos todos los posibles ORFs para nuestra secuencia de interés.

Phyre2 (Protein structural analysis) [Kelly and Sternberg (2009)]

La información genética nos provee la base para la síntesis de proteínas. El trabajo de laboratorio hasta este momento ha sido dirigido al estudio de la función proteica de una amylasa. La función de una proteína esta determinada por la expresión, estructura y localización subcelular. La expresión del gen lleva la información para la generación de la estructura primaria (i.e. secuencia polipéptica). La energía dada por la secuencia de amino ácidos incide directamente en la formación de estructura secundaria. La interacción entre estructura secundarias genera la formación de la estructura terciaria y funcional de la proteína. En algunos casos dos proteínas en su estructura terciaria interaccionan para formar la estructura terciaria. Entender o conocer la estructura de proteínas es fundamental en el desarrollo de drogas inhibidoras o activadoras de actividad enzimática. Además, compañías biotecnológicas utilizan el conocimiento disponible sobre la estructura de la proteína para el diseño de estructuras que puedan aumentar la actividad de las proteínas. Pero, la resolución de estructura proteica es un proceso que

toma tiempo. Por tanto, técnicas bioinformáticas integran conocimiento previo para predecir la estructura de proteínas.

El desarrollo de algoritmos ha ayudado a poder predecir la estructura secundaria y terciaria de proteínas usando la estructura primaria (o secuencia) de la proteína. El algoritmo Phyre2 es utilizado en la actualidad para predecir la estructura secundaria y terciaria de una proteína. Utilizaremos la secuencia proteica que ustedes encuentren para modelar o predecir la estructura secundaria y terciaria de la proteína.

D. Procedimiento I. Obtener secuencia.

1. Acceder la página electrónica del *National Center for Biotechnology Information* (NCBI):
   http://www.ncbi.nlm.nih.gov/

2. Para acceder GenBank®, vaya hasta la parte inferior, ver "Featured" y seleccionar GeneBank.

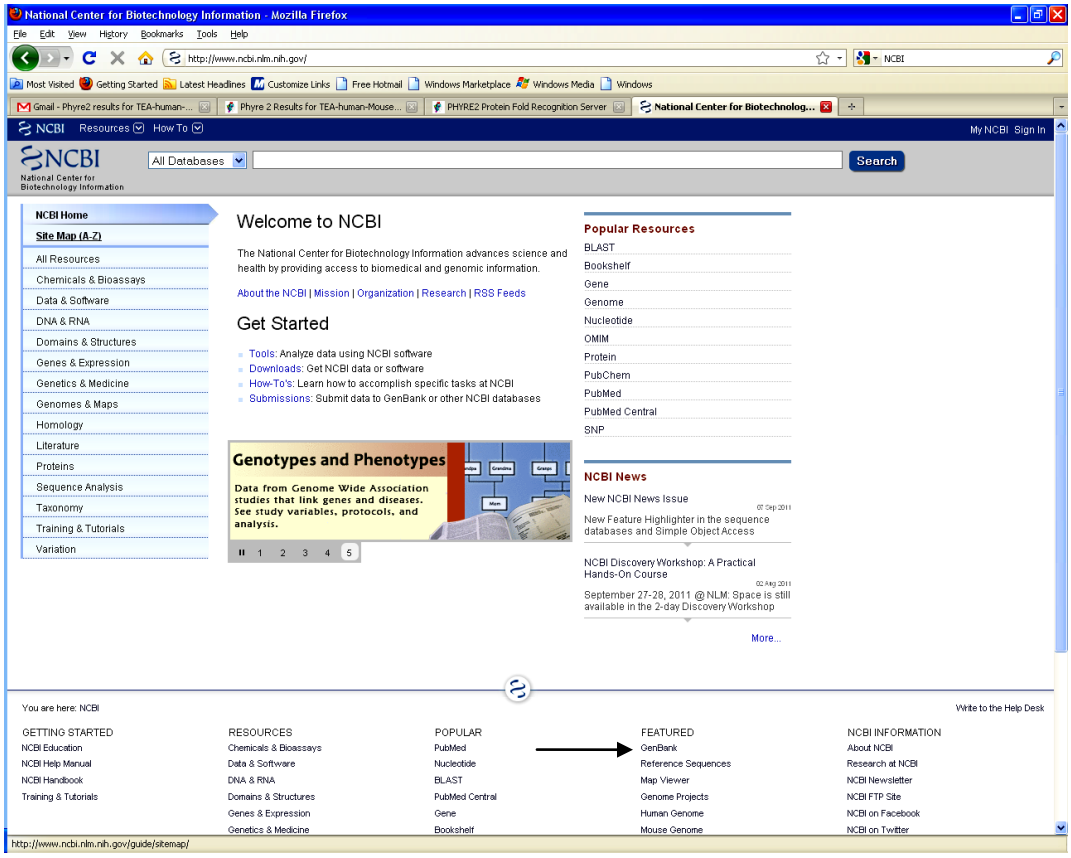


Figura 1. Página principal del *National Center for Biotechnology Information* (NCBI). La flecha de la figura indica el enlace hacia GenBank®.

3. Existen 2 formas de realizar la búsqueda de una secuencia de ADN. Una es escribiendo la descripción del gen o nombre del organismo. Esta es una búsqueda abarcadora que generará todas las secuencias relacionadas al término utilizado. Otro método para realizar la búsqueda es escribiendo el número de acceso de la secuencia (*accession number*) si este se conoce. Este método es más conveniente debido a que limita la búsqueda a la información de ese gen. En este caso utilizaremos la secuencia del gen AmyE con el número de acceso **X03236**.

   Para encontrar la secuencia que estas buscando, localiza en la parte superior de la página electrónica la barra de búsqueda. En el área identificada como *Search* debes seleccionar *Nucleotide* en vez de *Entrez* y en el área identificada como *for* escribe el número de acceso de la secuencia. Finalmente selecciona *Go* (Figura 2).

Figure 2. Nucleotide search

4. Una vez accedes el archivo encontrarás la descripción del gen. En la tabla I encontrarás una breve descripción del tipo de información del gen que se encuentra en el archivo. Al final de la página electrónica encontrarás la secuencia del gen.

Tabla I. Descripción de términos encontrados en el archivo que contiene la información del gen.

Términos	Descripción
Locus	Línea de identificación, contiene: nombre arbitrario, largo de la secuencia, tipo de molécula utilizada para obtener la secuencia y la fecha en que se hizo público el archivo.
Definition	Describe brevemente el contenido biológico del archivo.
Version	Muestra el número de acceso de la secuencia y luego del punto indica la versión. Además muestra el GI (*geninfo identifier*). El GI es un número asignado por NCBI, este nunca se repite y se utiliza para identificar un gen.
Source	Indica la fuente del organismo utilizado.
Organism	Indica la clasificación taxonómica del organismo.
Reference	Contiene algunas referencias en las que se basaron para realizar el trabajo, se incluye: autor, revista y titulo del artículo. Además contiene un enlace a la referencia del artículo en PubMed.
Features	Características del gen. Indica la posición de los nucleótidos que componen las secuencias importantes del gen como: promotores, secuencia codificadora (CDS) e inicio de trascripción entre otras cosas. En esta sección pueden encontrar la secuencia de amino ácidos que codifica el gen de interés.

5. Selecciona la secuencia resaltando la misma con el cursor desde el número 1 luego de *origin* hasta la última letra de la secuencia. Luego copia la misma. Para esto ve a *edit* en la barra que está en la parte superior de la ventana, luego selecciona el comando *copy*. Esta secuencia la utilizarás en la segunda parte del procedimiento.

   Es recomendable que guardes la secuencia que copiaste creando un archivo en *Word, Note Pad* o algún otro programa al cual tengas acceso.

E. Procedimiento II. Determinar homología.

1. Una vez obtenemos la secuencia podemos determinar cuáles son otras secuencias de genes en las que podemos encontrar una secuencia similar al gen de interés. Para esto utilizaremos la herramienta llamada *Basic Local Alignment Search Tool* (BLAST). Para acceder BLAST debes ir la página principal de NCBI.

   Si aún te encuentras en la página electrónica de la descripción del archivo del gen, puedes pulsar el logo de NCBI que se encuentra en la parte superior izquierda de la página electrónica. Por otro lado si decides comenzar el trabajo en otro momento, accede la página principal de NCBI en la dirección: http://www.ncbi.nlm.nih.gov/

En la parte superior de la página electrónica hay una barra que contiene diferentes enlaces, ahí debes hacer clic a BLAST (Figura 3).

Figura 3. Página principal de NCBI. La flecha a la izquierda de la figura muestra la barra que contiene el enlace hacia BLAST.

2. Para comparar la secuencia de ADN de interés con otras secuencias de ADN debes hacer clic en *nucleotide BLAST* (blastn) (Figura 4).

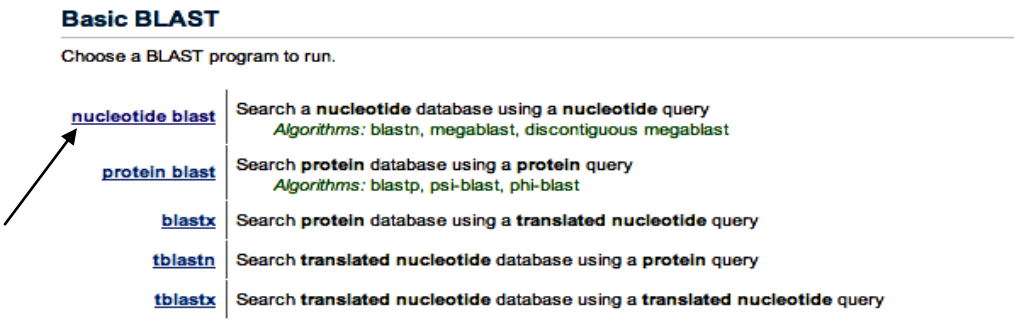
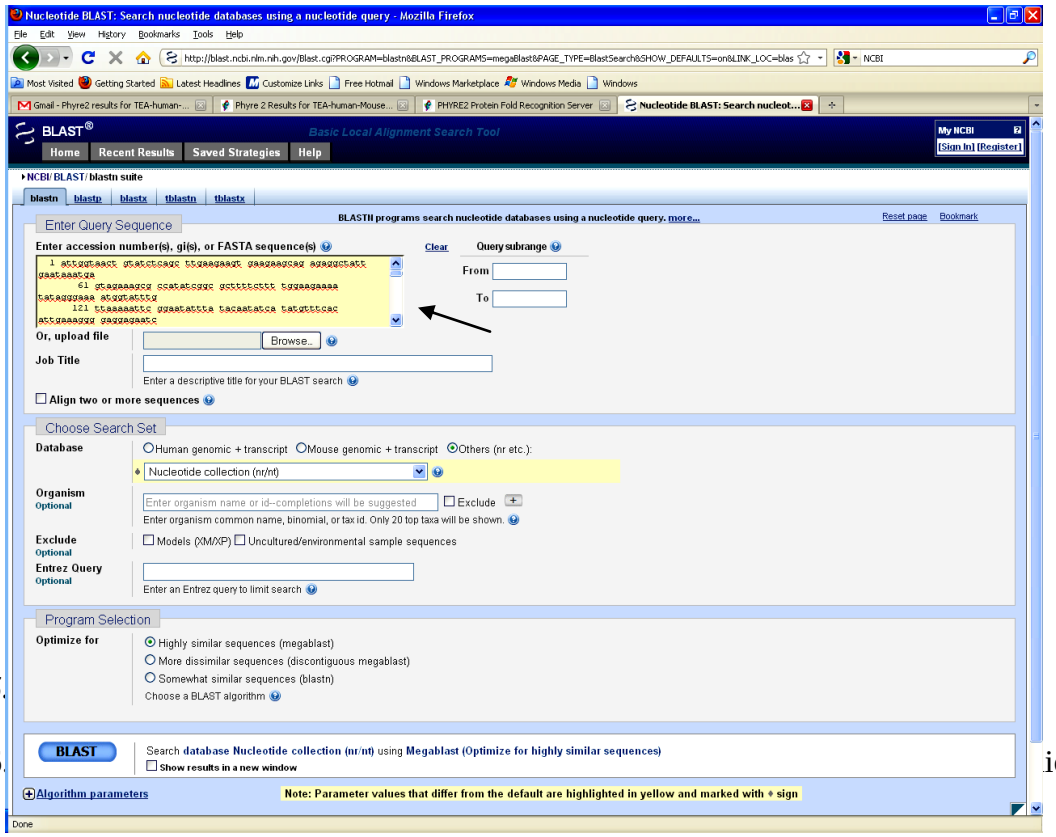


Figura 4. Página principal NCBI BLAST. La flecha dentro de la figura señala el enlace hacia la página electrónica de *nucleotide-nucleotide BLAST*.

3. Si no estas llevando a cabo el procedimiento el mismo día, debes buscar la secuencia del gen de AmyE que guardaste antes de proseguir. Debes copiar la misma utilizando la aplicación *copy* en el programa que estés utilizando.

4. Ve a *edit* en la barra superior de la ventana del *browser* y utilizando la aplicación *paste* inserta la secuencia del gen en el recuadro provisto en la parte superior de la página identificado como *Enter Query Sequence* (Figura 5).



5.

6. ice

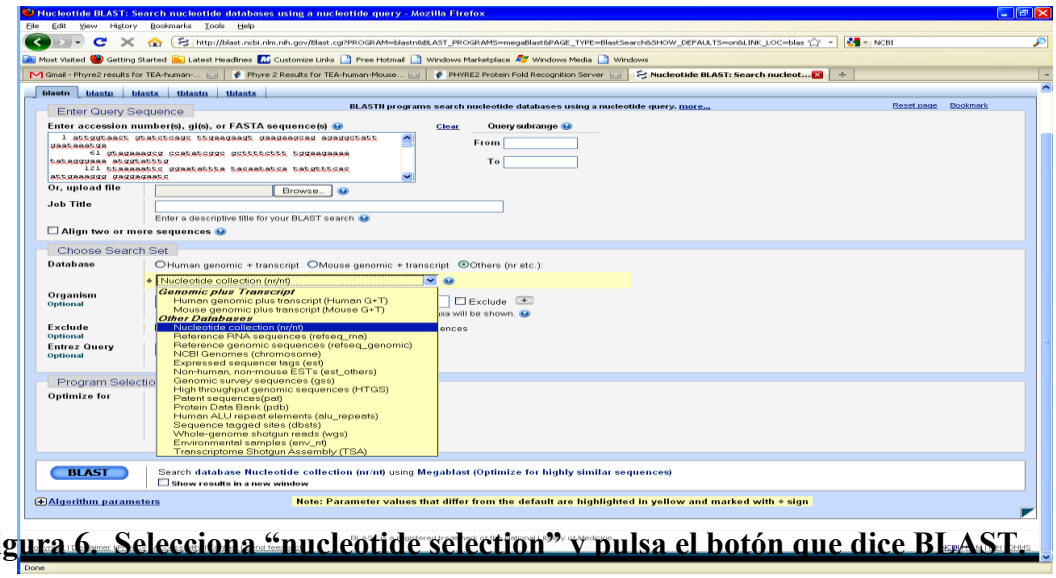


Figura 6. Selecciona "nucleotide selection" y pulsa el botón que dice BLAST.

7. Una vez te presiones el botón de BLAST aparecerá una página igual a la de la figura 7. Ahí deberás presionar el botón de *Formatting Options*. Luego en la siguiente página que aparece se deberá seleccionar el botón de *View report* (Figura 8). También, puede esperar hasta que el análisis termine.

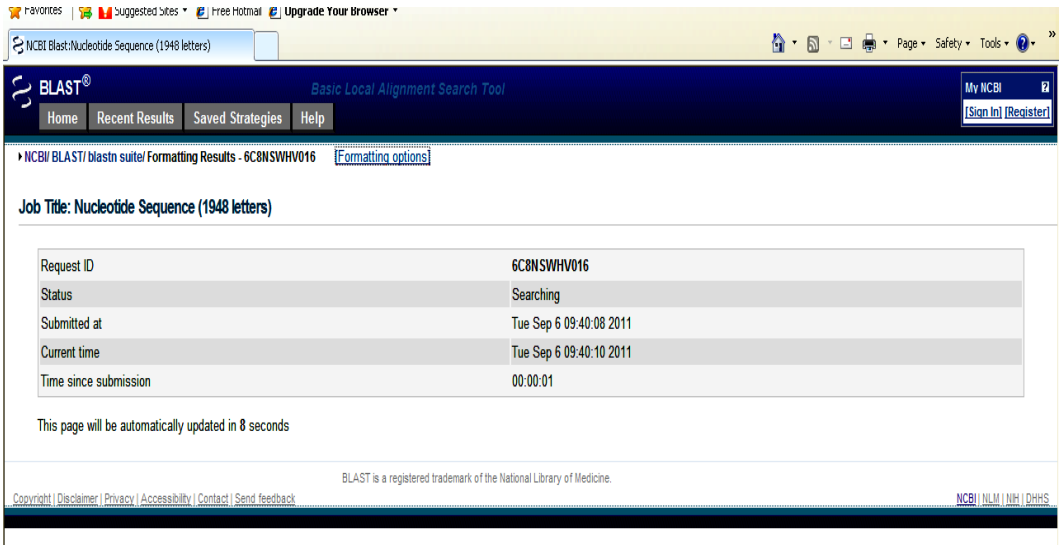


Figura 7: Descripción del estado de la búsqueda. La flecha a la derecha de la figura señala el botón *Formatting Options*

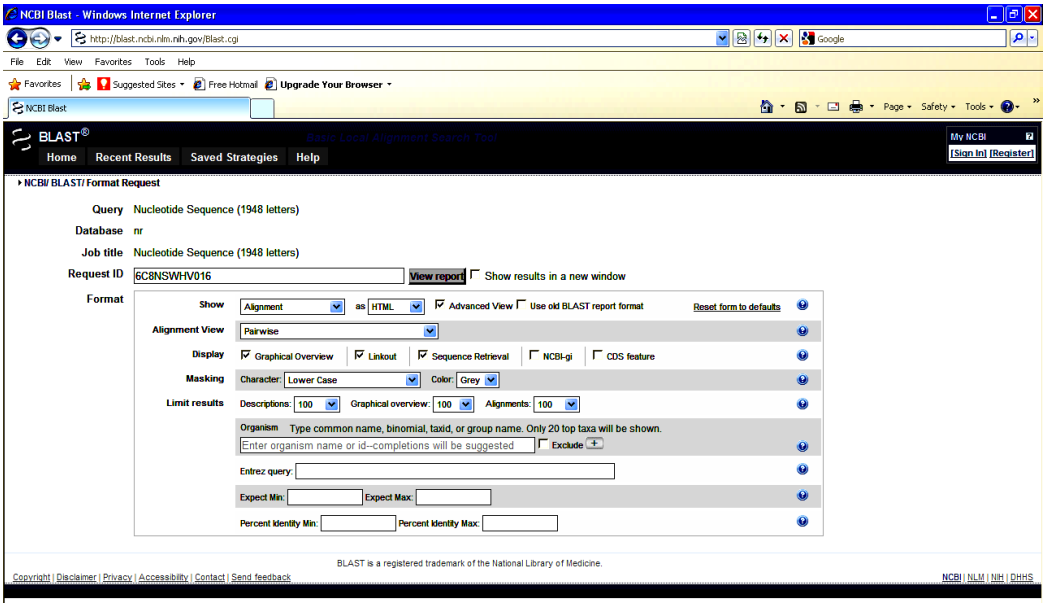


Figura 8: Descripción del estado de búsqueda. La flecha a la derecha de la figura señala el botón de *View Feport*

8. Los resultados de la búsqueda de homologías aparecerán en otra pantalla. Ahí encontraras un mapa con las alineaciones de diferentes archivos de GenBank® con los que el programa encontró homología con la secuencia del gen que

sometiste. Puedes mover el cursor sobre las líneas del mapa para identificar las secuencias (Figura 9).

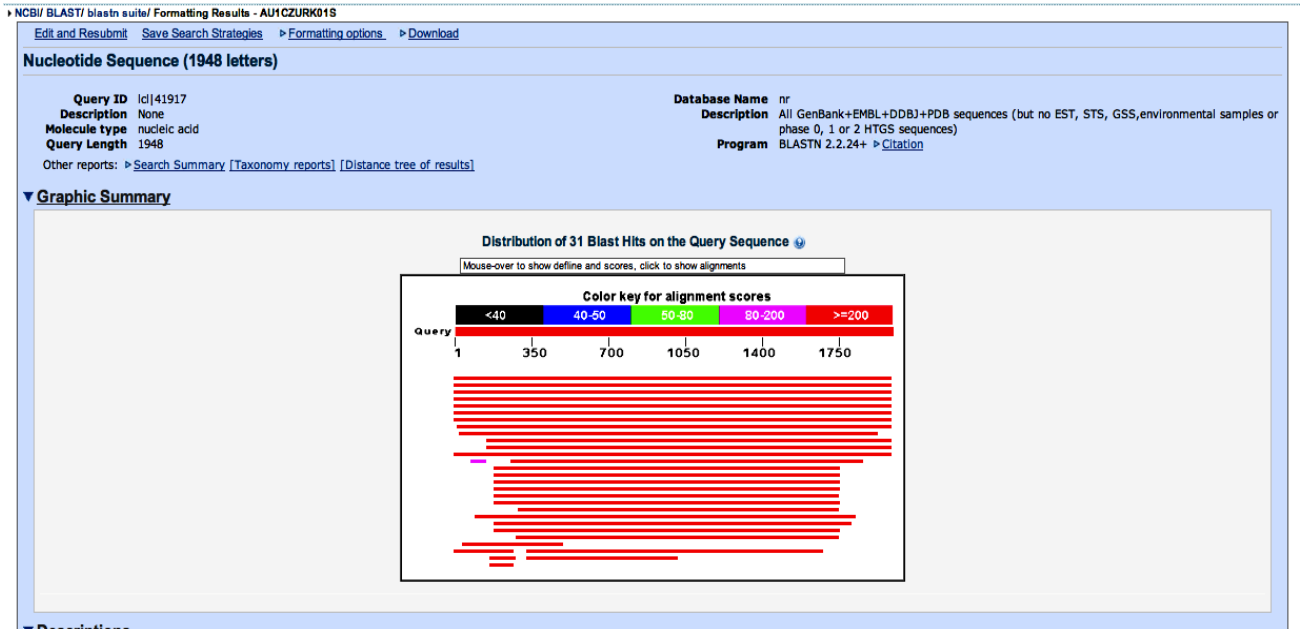


Figura 9. Resultados de la búsqueda de homología.

9. Para ver en detalle la descripción de los archivos que contienen homología, puedes pulsar una línea del mapa. Esto te llevará a ver el alineamiento de tu secuencia con la encontrada o puedes "roll down" con el curso para ver la descripción de los resultados.

F. Procedimiento III. Encontrar un Open Reading Frame (Marco de lectura abierto)

1. Accede a *google.com* y escribe " ORF finder" o accede a este enlace http://www.ncbi.nlm.nih.gov/gorf/gorf.html

2. Existen dos métodos para realizar la búsqueda, uno es escribiendo el número de acceso de la secuencia de DNA (*accession number)* o el número de identificación de información del gen (GI) en el espacio provisto. El otro método se realiza escribiendo o insertando la secuencia del gen en el espacio provisto para ese propósito identificado como *sequence in FASTA format*. Ambos métodos son igualmente efectivos. El método a utilizarse depende de la información que se tenga acerca del gen de interés.

   En este caso conocemos el número de acceso para la secuencia de AmyE **(X03236)**. Debes escribir el mismo en el lugar que está identificado como *Enter GI or ACCESSION* (Figura 10). Luego pulsar "OrtFind."

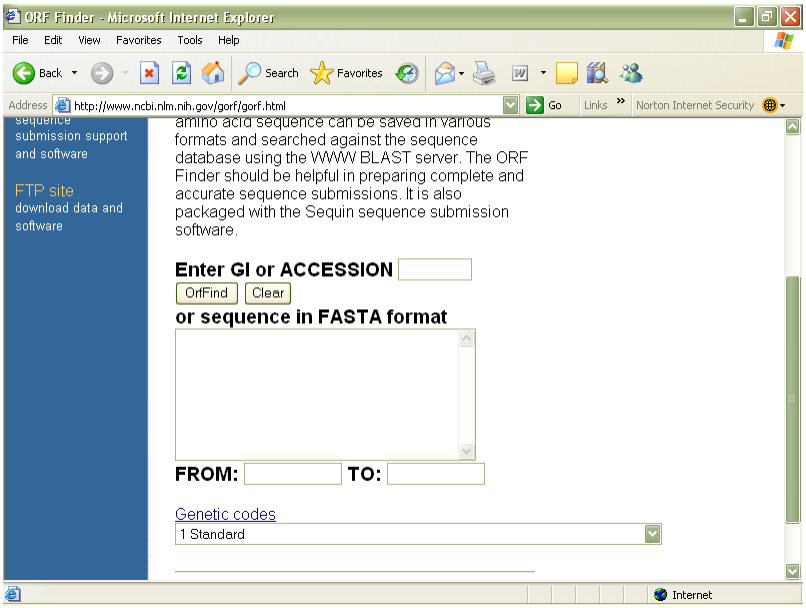


Figura 10. Página electrónica *ORF Finder* en NCBI. La figura muestra el área provista para llevar a cabo la búsqueda del gen de interés

3. La próxima página electrónica te muestra un mapa del gen que tiene sombreadas las partes del gen que posiblemente codifiquen para una proteína. A la izquierda del mapa se muestra una lista en la cual encontrarás el marco de referencia en el que puedes encontrar ese ORF, la localización de las bases y la cantidad de bases que incluye.

   Puedes seleccionar la secuencia que entiendas que codifica para la proteína y hacer clic para ver más detalles de la misma. Para seleccionar el "open-reading frame" correcto puedes utilizar la información obtenida en el ***Procedimiento I***, en la parte de "FEATURES, CDS."

   Determina cuantos amino ácidos codifica el gen y copiar la secuencia primaria de la proteína. Necesitaras la secuencia de la proteína para el Procedimiento IV.

4. Al seleccionar uno de los fragmentos, aparecerá toda la secuencia para ese segmento. Podrás ver la misma desplazando la página electrónica hacia abajo.

   Si te interesa ver la información del gen puedes hacer clic al botón que dice *View* justo sobre el mapa. Este enlace te llevará al documento que guarda la información del gen en *GenBank.*

G. Procedimiento IV. Análisis de secuencia de proteínas: Estructura secundaria y terciaria

1. El análisis de la secuencia de proteína se hará usando el algoritmo Phyre2 que pueden accesar en la página: http://www.sbg.bio.ic.ac.uk/phyre2/ (Figura 11)
2. Al llegar a la página indicada llenarán la información requerida que incluye, dirección de "email", nombre del trabajo y la secuencia de la proteína (Figura 11).
3. En la sección descrita como "Modelling Mode" marcaran "Intensive" (Figura 11)
4. Luego de entrada toda la información, pulsaran el botón que dice "Phyre Search"

Figura 11. Visual de la página del algoritmo Phyre2

5. Luego de pulsar "Phyre Search" saldrá una nueva página que le indicará el proceso de análisis. La pantalla se auto-renovará cada 30 segundos hasta que el algoritmo termine el ensayo. Usted puede dejar esta pantalla abierta hasta que el proceso termine o cerrarla y esperar a que le llegue un correo electrónico con la dirección en donde puede encontrar los resultados.

6. Al terminar el análisis, el servidor le enviará un correo electrónico a la dirección que uste registro. En ese mensaje habrá un "link" que le llevará a ver los resultados. (Figura 12)

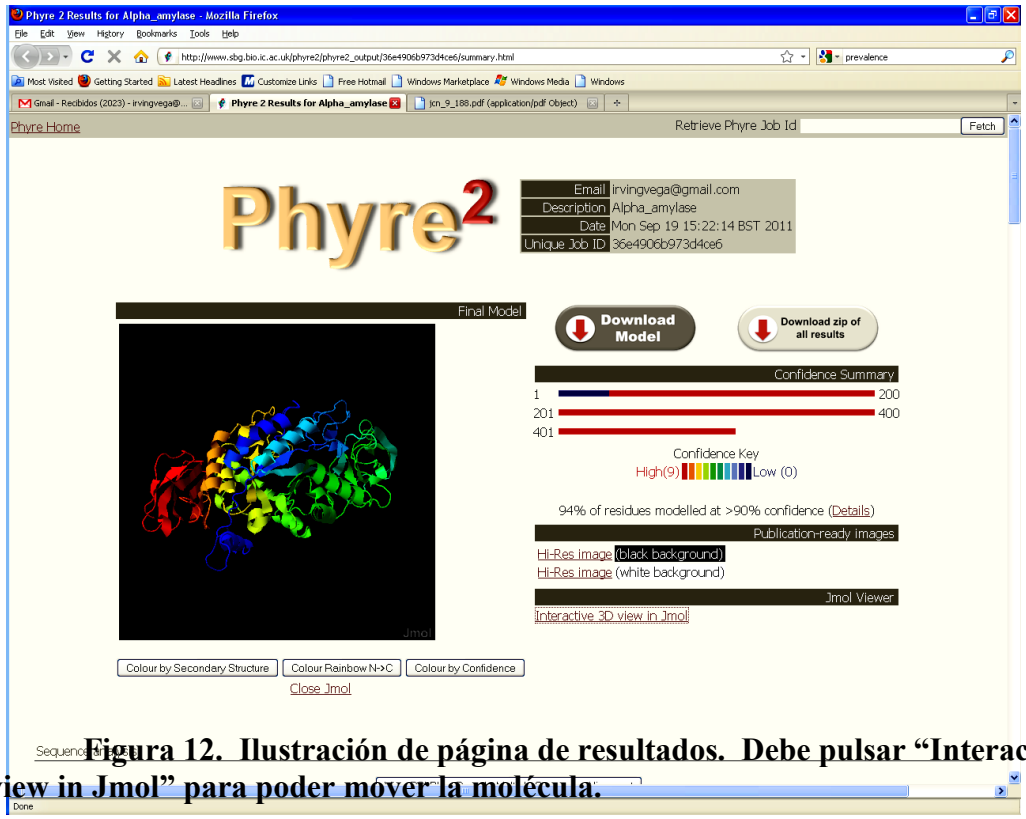


Figura 12. Ilustración de página de resultados. Debe pulsar "Interactive 3D view in Jmol" para poder mover la molécula.

7. "Scroll down" en la pagina de resultados para ver la predicción de estructura secundaria y alineamiento con otras proteínas análogas.

Figura 13. Predicción de estructuras secundarias.

H. Preguntas guías para análisis

Procedimiento I y II

1. ¿A que organismo pertenece el gen?

2. ¿Cuántos pares de base tiene el gen?

3. ¿Cuántos amino ácidos codifica este gen?

4. ¿Con cuál otros genes es homologo AmyE? Indique el por ciento de Identidad entre ellos.

5. ¿Qué indica o implica que haya genes homólogos a AmyE?

Procedimiento III

1. ¿Por qué hay diferentes "open reading frames"?

2. ¿Qué componentes determinan el "open reading frame"?

3. ¿Por qué el "open reading frame" tiene que ser continuo? (i.e. sin espacios o secuencias que interrumpan el "open reading frame")

4. ¿Cuántos codones contiene el "open reading frame" seleccionado?

5. ¿Cuántos amino ácidos codifica el "open reading frame" seleccionado?

6. ¿Existe diferencia entre la cantidad de codones y el número de amino ácidos codificados? ¿Por qué?

Procedimiento IV

1. Haga un listado de estructuras secundarias según la predicción hecha por el algoritmo Phyre2

2. ¿Cuál es el por ciento de confiabilidad ("confidence") en la estructura terciara modelada para alpha amylase?

3. ¿Cómo usted describe pudiera describir la estructura de la proteína?

4. Determine donde se encuentra localizados el N y C-terminal en la estructura de alpha amylase.

5. ¿Puedes determinar donde se encuentra el sitio activo de la proteína?

6. Si modificas amino ácidos catalíticos, ¿afectarías la estructura de la proteína?

I. Referencias

Thiel, Teresa, et al. (2002) Biotechnology: DNA to protein: a laboratory project in Molecular Biology. McGraw-Hill; pp. 73-84, 163-169.

Clavarie Jean-Michael and Cedric Notredame. (2003) Bioinformatics for Dummies. Wiley Publishing, Inc; pp. 73, 160, 215-217.

National Center for Biotechnology Information. "Our Mission."
Disponible en: http://www.ncbi.nlm.nih.gov/About/index.html
Revisado: 09/19/2011

Kelley LA, and Sternberg MJ. Protein structure prediction on the Web: a case study using the Phyre server. Nat Protoc 2009; 4(3):363-71.