Practico BiolMol 09
Análisis de secuencias

NCBI

El NCBI (National Center for Biotechnology Information) se estableció en 1988 con el fin de mantener y distribuir rápidamente la cantidad de información sobre biología molecular que estaba emergiendo en ese momento y que continuó creciendo exponencialmente hasta hoy. A partir de ese momento, la creación de este tipo de bases de datos se hizo muy importante ya que el trabajo experimental en biología depende fuertemente de estas herramientas esenciales.

Ingreso: www.ncbi.nlm.nih.gov

Una de las bases de datos que vamos a probar es el GenBank. Aquí se depositan todas las secuencias de ADN que son sustento de publicaciones en revistas internacionales arbitradas asi como otras secuencias generadas que no necesariamente terminan en una publicación. Una de las responsabilidades del NCBI es GenBank® que es la base de datos de secuencias en el NIH (Nacional Institutes of Health). El NCBI produce y distribuye GenBank desde 1992, base que junto con el EBI (European Bioinformatics Institute) y el DDBJ (DNA DataBase of Japan) contienen la mayor colección de secuencias del mundo.
Podemos utilizar el genbank para buscar secuencias del gen *prnA* de *Aspergillus nidulans*. Introduzca estas palabras clave en el campo de busqueda.
Cuantas secuencias encuentran?
Desde un cierto punto de vista todas se ajustan a las condiciones de busqueda realizada, pero que fue lo que realmente busco? Todas las secuencias pertenecen a *A. nidulans*?

Observe la solapa *details*:
Que puede ver alli? Como es una búsqueda booleana? Que son las palabras entre corchetes?
Como restringiría la búsqueda a el *prnA* de *A. nidulans*?

Seleccione la secuencia con *accession* AJ223459.2, la información que se despliega se encuentra en el llamado formato genbank. Analice la información correspondiente a esta secuencia.

Como puede mejorar la visualización?
Guarde esta secuencia en formato genbank en su maquina.

Identificación de una secuencia problema

En la parte experimental del práctico uds. obtuvieron una secuencia de ADN, como primera etapa vamos a intentar determinar, utilizando métodos informáticos, a que puede corresponder. Para ello vamos a realizar una

búsqueda en las bases de datos de secuencias, con la intención de encontrar algo similar.

BLAST:

La herramienta BLAST (Basic Local Alingment Search Tool) es una de más conocidas y ampliamente utilizadas en bioinformática.
BLAST se utiliza para comparar dos secuencias nucleotídicas o proteicas y encontrar regiones de similitud local entre ellas. Esta herramienta resulta muy útil ya que: 1) la similitud de secuencias es una forma muy poderosa de encontrar e identificar secuencias desconocidas, 2) es una herramienta muy rápida y, debido a que la cantidad de secuencias aumenta enormemente todo el tiempo, la velocidad es muy importante, 3) BLAST es confiable desde el punto de vista estadístico y de desarrollo del software y 4) BLAST es flexible y se puede adaptar a prácticamente cualquier escenario.
BLAST permite localizar secuencias en una base de datos mediante el alineamiento de las mismas buscando regiones de similaridad. Una vez que el BLAST se completó, se puede calcular la singificancia de los matches para evaluar si el alineamiento ha sido estadísticamente significativo. El algoritmo básico que utiliza el BLAST para realizar dicho alineamiento es el algoritmo Smith-Waterman que realiza alineamientos locales de dos secuencias.
El BLAST crea una palabra de tres aminoácidos (en el caso de proteínas) que tendrá un vecindario determinado con las palabras que alineadas con ella tengan un score mayor a un determinado valor umbral. El BLAST construye un diccionario con todas las palabras y sus vecindarios. Cada punto de la matriz será un alineamiento de la palabra o de alguna del vecindario con la base permitiendo un pequeño solapamiento. En la matriz se busca un juego de diagonales consistentes y se realiza programación dinámica. Esto es muy eficaz ya que reduce el espacio para realizar programación dinámica lo cual disminuye el costo del método. De esta manera BLAST no es un método de alineamiento exacto sino que es un método heurístico ya que asegura una solución que está dentro de las mejores pero no garantiza que sea la mejor posible.
El nivel de significancia estadística se deriva en parte de las matrices de puntaje que utiliza.
Una matriz de puntajes es de dos dimensiones y contiene todos los posibles puntajes (scores) de los pares de aminoácidos según si su similaridad es más probable que sea por ancestría común o por azar. También se llaman matrices de sustitución ya que los puntajes representan las tasas de sustitución.
Para secuencias proteicas, estas matrices se llaman BLOSUM (Blocks Substitution Matrix) y PAM (Percent Accepted Mutations). Para secuencias de DNA se utilizan matrices específicas para nucleótgidos.
Los dos tipos de matrices tienen significados diferentes. Por ejempo, una matriz BLOSUM30 es una matriz de probabilidad estimada a partir de secuencias con un 30% de similitud y una BLOSUM62 está estimada a partir de secuencias con un 62% de similitud. En una matriz PAM, si deseo buscar niveles de similitud altos, me convendrá utilizar una PAM con número bajo (ej. 50) y si quiero niveles altos de similitud será conveniente utilizar una PAM250 que será más sensible en dicho caso.

En el reporte de los resultados, se informa sobre el score y el e value de cada alineamiento. El score es computado a partir de la matriz de puntajes y las penalidades de gap. Un score más alto indica mayor similaridad. El e value (expected value) es el número de alineamientos esperados por azar dado el tamaño del espacio de búsqueda, la matriz de puntajes, y las penalidades de gap. Cuanto más bajo es el e value, es menos probable que se una similitud por azar.
A lo largo de los años, el algoritmo original de BLAST fue expandido a otras variantes del mismo. De esta manera se lograron formas especializadas de BLAST que pueden utilizarse en casos más variados. Algunas de ellas son:

Program	Database	Query	Typical uses
BLASTN	Nucleotide	Nucleotide	Mapping oligonucleotides, cDNAs, and PCR products to a genome; screening repetitive elements; cross-species sequence exploration; annotating genomic DNA; clustering sequencing reads; vector clipping
BLASTP	Protein	Protein	Identifying common regions between proteins; collecting related proteins for phylogenetic analyses
BLASTX	Protein	Nucleotide translated into protein	Finding protein-coding genes in genomic DNA; determining if a cDNA corresponds to a known protein
TBLASTN	Nucleotide translated into protein	Protein	Identifying transcripts, potentially from multiple organisms, similar to a given protein; mapping a protein to genomic DNA
TBLASTX	Nucleotide translated into protein	Nucleotide translated into protein	Cross-species gene prediction at the genome or transcript level; searching for genes missed by traditional methods or not yet in protein databases

El algoritmo de BLAST fue desarrollado en primer lugar en el NCBI (National Center of Biotechnology Information) que forma parte del National Institute of Health. Cualquier persona puede ingresar al sitio del NCBI y correr el BLAST especificando los parámetros y la base de datos (http://www.ncbi.nlm.nih.gov).

- Utilice el producto de PCR silvestre para realizar el BLAST ¿Qué tipo de variante del BLAST convendrá utilizar?
- ¿Qué parámetros de BLAST utilizó?
- Analice la salida de BLAST.

Apunte su navegador hacia la secuencia que aparece como el primer hit. Esta es una entrada genbank. Analícela.

- ¿Qué es la secuencia?
- ¿Qué puede decir de la estructura del gen encontrado?

Visualización del cluster que contiene a prnA en A. nidulans

Una forma de visualizar la información de un archibo de tipo genbank, anotado, es utilizar el programa Artemis.
Para ejecutar el programa Artemis deberan obtenerlo de la pagina del curso en la sección del práctico y guardarlo en algun lugar de l maquina en la que estan trabajando e.g el escritorio. Luego abrir una terminal de linea de comando de Windows (Inicio -> ejecutar) y escribir cmd. Esto abrira una terminal de comandos tipo DOS, aquí deberan llegar hasta el lugar donde guardaron el Artemis., si fue en el escritorio, escriban
cd escritorio
y luego pulsen ***enter***, después escriban la órden:
java -jar artemis_v11.jar

Abrir ahora la secuencia guardada de *A. nidulans*
Que indican las diferentes tiras que observan? Y los colores?
Que indican las barritas verticales negras?

Cuantos genes estamos observando? Que podemos decir de la estructura génica de cada uno?

A que distancia se encuentran uno de otro?

Grafique el contenido de acuerdo a una ventana deslizante (Graph) se observa algo particular?

Guarde el la secuencia codificante para proteinas (CDS) en formato fasta (utilizar File->Write)

Análisis de los mutantes:

Para visualizar la mutación del producto de PCR realizaremos un alineamiento con el programa seaview. Para esto cargamos el archivo salvmut.txt que contiene la secuencia del producto de PCR salvaje y las mutantes. Las secuencias se encuentran en formato FASTA, que es probablemente el formato más simple. En primera instancia vamos a realizar un alineamiento de nucleotidos.
- ¿Cual es la mutación?
- ¿En que zona se encuentra?
- ¿Produce un cambio de un aminoácido por otro? En caso de que si, cual por cual?