PRUEBA PRÁCTICA-BIOINFORMÁTICA

BÚSQUEDA DE SECUENCIAS DE PROTEÍNAS EN BASES DE DATOS

INTRODUCCIÓN

Las bases de datos biológicas se han convertido en un instrumento importante para ayudar a los científicos a comprender y explicar los fenómenos biológicos, desde la estructura biomolecular y su interacción, hasta el metabolismo completo de los organismos y la comprensión de la evolución de las especies. Este conocimiento ayuda al diagnóstico de patologías, al desarrollo de medicamentos, a la lucha contra las enfermedades, al descubrimiento de las relaciones básicas entre las especies en la historia de la vida...

El conocimiento biológico se distribuye entre múltiples bases de datos generales y especializadas. Uno de los tipos de bases de datos biológicas más usados son las **bases de datos de secuencias**. Estas son una gran colección de secuencias de ADN, proteínas y otras, que se almacenan en computadoras. Se denomina 'secuencia' al orden en que los nucleótidos (en el ADN) o los aminoácidos (en los péptidos y proteínas) se encadenan. La secuencia de aminoácidos de una proteína viene determinada por la secuencia de nucleótidos del ADN del gen que la codifica.

Una base de datos puede incluir secuencias de un sólo organismo, como las bases de datos que contienen todas las proteínas de la bacteria *Escherichia coli*, o de la levadura de la cerveza *Saccharomyces cerevisiae* o la de *Homo sapiens*, o puede incluir secuencias de todos los organismos cuyo ADN ha sido secuenciado. Las bases de datos biológicas también tienen referencias cruzadas con otras bases de datos con el número de acceso como una forma de vincular sus conocimientos relacionados con el conjunto.

Existen **bases de datos primarias**, que contienen información directa de la secuencia, estructura o patrón de expresión de ADN o proteína y, **secundarias** que contienen datos e hipótesis derivados del análisis de las bases de datos primarias, como mutaciones, relaciones evolutivas, agrupación por familias o funciones, implicación en enfermedades, etc.

La identificación y el análisis de las secuencias de nucleótidos y proteínas es un requerimiento básico para la investigación bioquímica, biomédica y biotecnológica. Para ello, el desarrollo y uso de herramientas bioinformáticas, así como el manejo de las bases de datos biológicas, son indispensables.

Las principales **bases de datos de ADN** son: EMBL-BANK en el Instituto europeo de Bioinformática (EBI); DNA Data Bank of Japan (DDBJ) en el Centro de Información Biológica (CIB); GenBank en el Centro Nacional de Información Biotecnológica (NCBI).

Las principales **bases de datos de proteínas** son: Swiss-Prot, que contiene secuencias anotadas o comentadas, es decir, cada secuencia ha sido revisada, documentada y enlazada a otras bases de datos; PROSITE que contiene información sobre la estructura secundaria de proteínas, familias, dominios, etc; Protein Data Bank (PDB) que es la base de datos de estructura terciaria 3-D de proteínas que han sido cristalizadas; InterPro que integra la información de diversas bases de datos de estructura secundaria como PROSITE, proporcionando enlaces a otras bases de datos e información más extensa.

Las principales **bases de datos de genomas** son: Ensembl que integra genomas eucariotas grandes como el genoma humano, ratón, rata, pez cebra, mosquito, *Drosophila melanogaster*, *Saccharomyces cerevisiae*.

También existen motores de búsqueda como PubMedo BLAST que dan libre acceso a diferentes tipos de bases de datos.

MATERIAL

- Ordenadores con conexión a internet

DESARROLLO DE LA PRÁCTICA

1. Accede a la base de datos **Uniprot/Swiss-Prota** través del enlace http://www.uniprot.org/

2. Busca la secuencia de la proteína con código de acceso **P69891**. Para ello introduce este código en la ventana "Query" y asegúrate que la búsqueda la haces en UniprotKB.

3. Explora la ficha resultado. En ella podrás encontrar el nombre de la proteína, el organismo al que pertenece, su clasificación taxonómica, la longitud de su secuencia de aminoácidos, su peso molecular, su función, su localización, etc. También puedes tener acceso al análisis de su estructura primaria y a la secuencia completa de aminoácidos entre otras informaciones.

4. Busca otras proteínas que pertenezcan a la misma familia de P69891. Para ello accede a la familia de proteínas a la que pertenece esta proteína pulsando en el correspondiente enlace que aparece en la ficha de resultados.

5. Las bases de datos biológicas suelen estar conectadas a motores de búsqueda que pueden proporcionar información adicional. Con la herramienta bioinformática BLAST podemos identificar proteínas idénticas u homólogas en otros organismos.
Pincha en BLAST y podrás identificar esta proteína en otros organismos. El resultado de la búsqueda está ordenado por el porcentaje de identidad que hay entre la secuencia problema que has introducido y las secuencias resultantes de la búsqueda.

Los resultados de la búsqueda pueden aparecer en varias páginas. Para moverte por estas páginas usa 'next' y 'previous'.

6. Estos motores de búsqueda también pueden proporcionar información adicional taxonómica. Accede a la ventana 'taxonomy' y podrás encontrar los organismos en los que se encuentra esta proteína agrupados por especies.

7. La consulta de otras bases de datos puede proporcionar información relevante.

A) **Protein Data Bank** (PDB) (http://www.rcsb.org) proporciona información estructural 3D de las proteínas cuya estructura ha sido determinada, indicando el método que se ha empleado para determinar la estructura y la resolución atómica a la que ha sido resuelta en angstroms (Å). Este dato es importante para conocer la calidad de la estructura. Las proteínas tienen una estructura primaria, secundaria, terciaria y cuaternaria. La estructura primaria se refiere a la secuencia de aminoácidos. La estructura secundaria es la disposición espacial local del esqueleto proteico, debido a la formación de puentes de hidrógeno entre los átomos que forman el enlace peptídico, sin hacer referencia a la cadena lateral. Los elementos de estructura secundaria pueden ser ordenados como las hélices alfa y láminas beta (u hojas plegadas beta) o desordenados. La estructura terciaria es el modo en que la cadena proteíca se pliega en el espacio. La estructura cuaternaria deriva de la conjunción de varias cadenas peptídicas que, asociadas, conforman un multímero, que posee propiedades distintas a la de sus monómeros componentes.

B) **GenBank** (http://www.ncbi.nlm.nih.gov) contiene información de todos los genomas secuenciados así como de su distribución cromosómica.