Máster en Bioinformática 2014-2015 Escuela Nacional de Sanidad (*ISC*III), Madrid

Trabajo de Bases de Datos (v7.0).

Objetivo :

El trabajo consiste en crear una base de datos relacional (ya sea en PostgreSQL o SQLite3), que permita la integración de dos fuentes de datos externas, como son PFAM y DOE Joint Genome Institute (http://jgi.doe.gov).

El trabajo implicará la creación de varias tablas en SQL y de varios programas en Python 3, que serán necesarios tanto para insertar la información dentro de vuestra base de datos, cómo para obtener determinada información de la base de datos, una vez que la información esté guardada.

A cada grupo se os asignará un organismo distinto y se os enviará un fichero Fasta con todas las proteínas conocidas para ese organismo. A partir de esa secuencia en Fasta y usando hmmer3 (*Se recomienda revisar la ayuda de este programa para aumentar la velocidad de cálculo y para facilitar la lectura de los ficheros de salida)* frente a Pfam, se deberán encontrar y almacenar, los dominios resultantes en la base de datos.

Una vez que la información del organismo (procedente del fichero Fasta), la información de Pfam y la de los dominios funcionales (procedente de ejecutar hmmer3) esté debidamente integrada en la base de datos, se deberán contestar varias preguntas mediante programas realizados en Python 3.

Junto al resto de programas y ficheros que se solicitan, la entrega del trabajo debe incluir una memoria (en .txt o .pdf). Esta memoria debe detallar todos los pasos necesarios para ejecutar el trabajo de principio a fin, tanto por nosotros, como para que se entere otra persona sin conocimientos previos. Se valorará positivamente la inclusión en la memoria de explicaciones breves relacionadas con cada punto del trabajo que detallen los problemas encontrados y las soluciones aplicadas.

Detalles Pfam :

ftp://ftp.sanger.ac.uk/pub/databases/Pfam/releases/Pfam27.0/

Necesitaréis Pfam-A.hmm.gz (para construir la base de datos Pfam usable con Hmmer3) y Pfam-A.seed.gz (para extraer los datos interesantes y guardarlos en la base de datos relacional). La documentación del formato de este último fichero se encuentra en:

ftp://ftp.sanger.ac.uk/pub/databases/Pfam/releases/Pfam27.0/userman.txt

Detalles del trabajo :

1) Diseñar una base de datos para almacenar la siguiente información:

   a. **Procedente de JGI**, Tendréis que diseñar una o más tablas para guardar estos 5 campos: sinónimo del nombre del organismo, nombre científico del organismo del que procede, la secuencia de las proteínas, el identificador de esa proteína y por último, la descripción.

   Cada entrada del fichero FASTA tiene una cabecera como ésta:

   > *jgi|Selmo1|Selaginella moellendorffii|24961|e_gw1.1.94.1*

   Donde el primer elemento después del ángulo es una descripción del consorcio de secuenciación (**normalmente** pondrá jgi), el segundo es un sinónimo del nombre del organismo, el tercero es el nombre científico del organismo del que procede la secuencia de proteínas, el cuarto es el identificador de esa proteína. Por último hay una descripción.

   b. **Procedente de Pfam**. Tendréis que diseñar una o más tablas para guardar estos 4 campos: el identificador del dominio Pfam, el accession , la descripción y todas aquellas referencias a la base de datos InterPro que aparezcan.

   Ejemplo :
   *#=GF **ID** **1-cysPrx_C***
   *#=GF **AC** **PF10417.2***
   *#=GF **DE** **C-terminal domain of 1-Cys peroxiredoxin***
   *#=GF DR INTERPRO; **IPR019479**;*

   c. **Procedente de hmmer3.** Tendréis que diseñar una o más tablas para relacionar los resultados del hmmer con las proteínas del organismo.
   Tendréis además que almacenar coordenadas de inicio y fin de cada motivo encontrado dentro de cada secuencia del microorganismo, así como el *score* y el *independent-Evalue* (http://hmmer.janelia.org/help/result).

Las tablas creadas en éste apartado deben enviarse vacías y en formato SQL. Además deberán contener una explicación breve de qué es cada tabla y cada columna de cada tabla.

2) Diseñar y ejecutar los programas necesarios para realizar el trabajo práctico. La recomendación es un programa para cada fase de carga en la base de datos o recuperación de datos de la base de datos para proporcionar resultados.

   2.1. Programa que inserte los datos necesarios de Pfam en las tablas del punto 1.b.

   2.2. Programa que inserte los datos necesarios del microorganismo que os haya correspondido en las tablas creadas para el punto 1.a.

   2.3. Programa que ejecute `hmmscan` sobre Pfam y las proteínas del organismo y procese el fichero de resultados generado por `hmmscan` para almacenar los resultados en la(s) tabla(s) del punto 1.c. Pero, en vez de usar todas las proteínas del organismo, se usarán solo aquellas secuencias cuyo tamaño total de secuencia, sea mayor a la media del tamaño, de todas las secuencias del organismo.

3) Partiendo de que ya tenéis todos los datos almacenados en la base de datos, y ya está realizada la búsqueda con `hmmscan`, se deberán crear dos programas para responder a las siguientes preguntas:

   3.1. Para una secuencia concreta de vuestro microorganismo (ID de JGI o descripción) que vuestro programa aceptará como parámetro, si ha sido analizada tenéis que listar el dominio/os de PFam (ID y Acc) , junto con los identificadores de InterPro de cada uno de esos dominios, y si no, decir que no está analizada.

   3.2. Para un identificador de InterPro determinado, que se aceptara como parámetro, se deberá listar la(s) proteína(s) (ID de JGI) donde aparecen los dominios PFam anotados con ese identificador.

4) Partiendo de que ya tenéis todos los datos almacenados en la base de datos, y ya está realizada la búsqueda con `hmmscan`, con otro programa a crear, se responderá a la siguiente pregunta:

   Para el organismo que os haya correspondido, ver para las proteínas etiquetadas como kinasas, el número medio, el máximo, el mínimo y la desviación estándar del número de dominios para esa función. Esto deberá imprimirse además junto con las proteínas etiquetadas como lyasas, *ion channel*, *receptor* y *transport*.

Formato de entrega :

Un archivo comprimido con el modelo de la base de datos debidamente documentado, los programas requeridos por los diversos puntos y una memoria escrita (ya sea en .txt o en .pdf) que además de una explicación del procedimiento usado para realizar el trabajo explique la forma de ejecutar cada uno de los programas usados en cada paso.
Deberá enviarse por correo electrónico a las siguientes direcciones :

jmfernandez@cnio.es
eleon-ibis@us.es

Puntuación y plazos de entrega:

Pregunta 1) : 3 ptos
Pregunta 2) : 2.5 ptos
Pregunta 3) : 2.5 ptos
Pregunta 4) : 2 ptos

Plazo de entrega : 12 de Enero de 2015

Nota 1: Para todos los formatos os recomendamos que os miréis más de una entrada, para comprobar cuáles de los campos indicados son opcionales, o se repiten.

Nota 2: Se puede reutilizar todo aquello explicado, proporcionado y creado en las clases. No hace falta reinventar la rueda.

Nota 3: No tenéis que escribir una memoria con más páginas que Los Pilares de la Tierra. La puntuación no es “al peso”.