Documetación Base de datos

David Gómez Sánchez y Pablo Vicente Munuera

Abstract

Esto es un abstract. Y como tal, procederá a relatarle, un esbozo de lo que es este documento. En él, trataremos temas tan acuciantes como: ¿Qué es esto? ¿Cómo hemos trabajado? ¿Qué problemas hemos sobrepasado? y ¿Cómo se ejecuta nuestro *wonderful* programa?

1 Introducción

2 Getting started

Aunque está explicado como preparar el entorno para ejecutar nuestros programas, procederemos a hacerlo, igualmente, aquí.

3 Realización de la práctica

3.1 Herramientas utilizadas

- Sublime Text o geany.
- Texworks.
- Github.
- Python3.4.
- Postgresql.

3.2 Diseño de base de datos

El diseño de base de datos nos ha llevado de cabeza durante la mayor parte del desarrollo del trabajo, aunque eso trataremos más debidamente en la sección: problemas encontrados.

Ha habido dos tablas claras desde el primero momento: la tabla jgi y la tabla Pfam. La primera representa al punto 1.a "Procedente de jgi" y corresponde a las proteínas y sus correspondientes secuencias y datos. Tiene como clave primaria el id, lo cual parece lógico, ya que es el elemento más representativo de cada proteína. Aunque en un primer momento se puso también el organismo, posteriormente se vió que el id cambiaba con el organismo, con lo que no era necesaria la dupla id-organismo como clave primaria. Al no tener algo en lo que basarnos para decidir los tipos, se ha tomado algo que no

consumiera demasiado, pero que fuera más que suficiente. Por ello, en cosas que pueden ser muy largas, como la secuencia y la descripción, se han puesto *Text* como tipo de dato. Los *NOT NULL*, aún a sabiendas de que todo se debía meter, hay cosas que creemos que aunque no se metieran, no tendrían un efecto adverso, a la hora de las búsquedas y filtrados.

Por otro lado, la tabla Pfam también quedó bastante resuelta desde un principio. A parte de los datos que se nos informaba que debía contener esta tabla, lo único que había que pensar era la clave primaria y la posibilidad de claves alternativas. Las dos únicas cosas que no se pueden repetir es el ID y el accesion number. La combinación de ambas no era posible, ya que no para un mismo ID no va a haber 2 accesion number distintos, será siempre el mismo, con lo que se opta por poner una clave primaria (ID, pero podría haber sido perfectamente el accesion number) y una clave alternativa que será el accesion number. Los tipos de esta tabla, excepto el ID, que debió adecuarse a otra tabla, estaban específicados en la página web oficial de Pfam, así que no tuvimos que realizar el esfuerzo nosotros. Un dato curioso: intentando imitar lo que se dió en clase, se hizo una tabla para los accesion number y los posibles accesion numbers antiguos. Esta se acabó por eliminar, ya que no tenía mucho sentido en el entorno de la práctica.

El problema y la parte de las tablas que más cambios han sufrido han sido las correspondiente al punto 1.c "Datos provenientes de hmmer". Ha llegado a haber hasta 3 tablas, sólo para esta sección. Finalmente, se vió que con solo 2 tablas era suficiente. En primer lugar, tenemos la tabla hmmer, en la cual, guardamos la información de: el ID de la proteína con la que hacemos la query a hmmscan, que será un *INT* ya que es un número; la descripción de esta, que no tendrá tope de caracteres; y el e-value que debido a que puede ser muy pequeño, lo pondremos como *float*. El ID de esta tabla será clave primaria y hará referencia al mismo ID de JGI. Por lo tanto, en la tabla HMMER deberán existir solo los ID, los cuales se encuentren también almacenados en la tabla JGI.

3.3 Problemas encontrados

3.3.1 Problemas con el diseño

Como ya se ha mencionado anteriormente, la parte de las tablas correspondientes a la información sacada desde *hmmscan*, ha sido siempre la más compleja y cambiante a lo largo del proyecto.

4 Conclusión

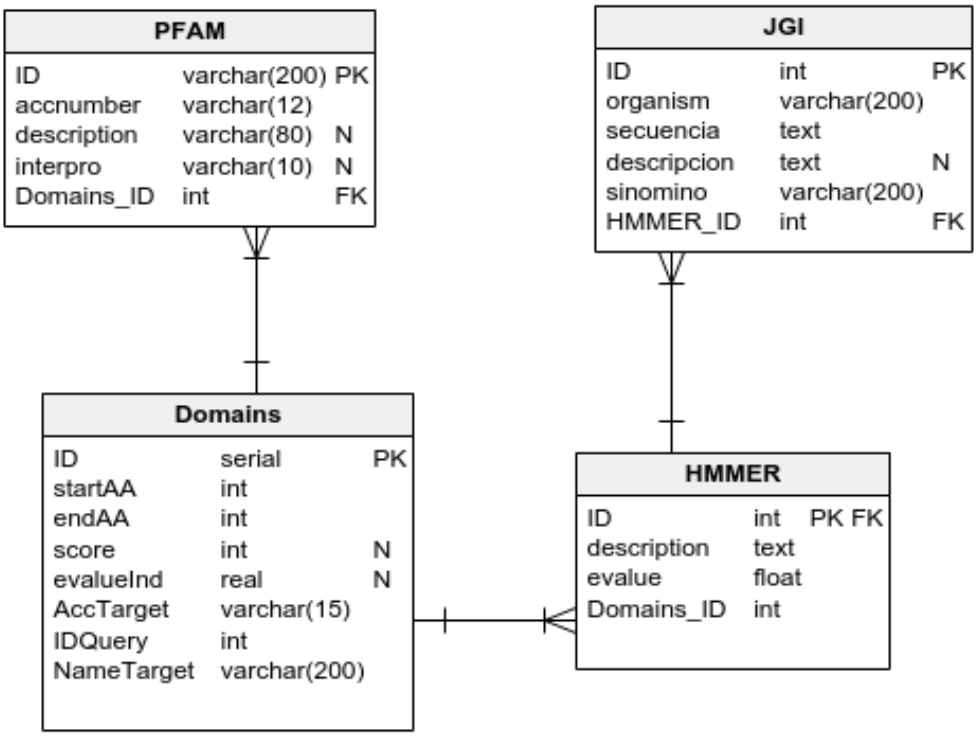

Figure 1: Diseño final de la base de datos