Taller grupal. Creación de una base de datos genómica

Objetivos

Los objetivos que se espera alcanzar en este taller grupal se centran en el proceso de desarrollo y creación de una base de datos. Tras la consecución de este taller, podrás imaginar y comprender la estructura y organización de las bases de datos genómicas, diseñar una base de datos genómica utilizando un modelo E-R, implementar una base de datos a partir del modelo de datos relacional en SQL y rellenar la base de datos con datos genómicos de ejemplo.

Pautas de elaboración

La actividad grupal consta de unas pautas de realización claras. Estas pautas son las lógicas a la hora de implementar una base de datos relacional, en este caso las adaptamos a una base de datos relacional genómica. La actividad, por tanto, consiste en desarrollar una base de datos en SQL (<https://dev.mysql.com/downloads/installer/>) sobre el dominio de bases de datos genómicas.

Los puntos sobre las que se tiene que desarrollar la base de datos genómica son los siguientes:

1. **Análisis de requerimientos:** antes de diseñar la base de datos, es esencial entender qué información se quiere almacenar y cómo se relaciona. Para este taller, se almacenarán genes, secuencias, variantes genéticas y anotaciones asociadas.
2. **Diseño E-R:** utilizando herramientas de diseño o papel y lápiz crea un diagrama E‑R que represente las entidades principales (por ejemplo, gen, secuencia, variante) y sus relaciones.
3. **Transformación a modelo relacional:** convierte tu diseño E-R en un conjunto de tablas, atributos y relaciones que puedan implementarse en MySQL.
4. **Creación en MySQL:** utiliza sentencias SQL para crear las tablas diseñadas en el paso anterior. Asegúrate de definir correctamente las claves primarias, las relaciones entre tablas y restricciones.
5. **Inserción de datos:** rellena la base de datos con datos de ejemplo. Puedes usar datos ficticios o reales, pero asegúrate de que representen adecuadamente la estructura y relaciones de tu diseño.

* El **enunciado** sobre el que hay que trabajar la base de datos es el siguiente:

Dentro del vasto mundo de la genómica, los científicos se esfuerzan por comprender la estructura y función de los genes. Para almacenar esta información, la unidad principal es un gen. Cada gen, identificado por un nombre único y una descripción, se localiza en una posición específica de un cromosoma. Estos genes están compuestos por distintas secuencias y cada secuencia tiene una representación de cadena de ADN, un tipo (como exones o intrones, CDS, promotores, etc.) y una posición relativa al gen.

A medida que se han ido explorando los genes se han identificado variaciones genéticas. Estas se han de almacenar también para llevar un registro de qué partes de una secuencia pueden variar. Cada variante se distingue por su posición relativa al gen, además presenta el alelo de referencia y el alelo mutado y un tipo específico de variante, como un cambio de un solo nucleótido, inserción o deleción.

Como extra, y para enriquecer la comprensión, los genes tienen anotaciones asociadas. Estas anotaciones ofrecen información adicional, como la función biológica de un gen o las implicaciones médicas de una variante. De cada anotación, por tanto, tenemos una descripción, el tipo de anotación (funcional, clínica, etc.) y su relación con los genes.

Por último, almacenamos también los estudios que trabajan con genes o variantes. Estos representan los estudios científicos que han identificado o caracterizados los genes o variantes. De esta manera, cada estudio puede haber trabajado con varias entidades genéticas y cada entidad genética puede haber sido estudiada en varios estudios. Para cada estudio nos interesa su título, su fecha de publicación y su referencia.

La propia base de datos tiene **restricciones intrínsecas,** las cuales deben identificar los estudiantes y aplicarlas en la creación de la base de datos. Sin embargo, además queremos **definir nuevas restricciones** para la creación final de la base de datos, estas son:

* La representación de ADN de una secuencia no debe exceder los mil caracteres y como mínimo debe tener diez caracteres.
* El nombre del gen y la descripción son obligatorios.
* La referencia de un estudio debe seguir el formato siguiente: cuatro letras, /, tres números. Por ejemplo: aaaa/111.
* El valor por defecto en los alelos es «-», que indica un hueco en la secuencia.
* La posición relativa de las secuencias y variantes no puede ser nula y debe ser un número entero positivo.

Hay que destacar que, debido al problema que se presenta, no hay una solución única, sino que muchas soluciones pueden ser válidas siempre que capten el mundo real y el enunciado que se ha descrito. Sin embargo, hay que cumplir claramente con las pautas de desarrollo de cada uno de los pasos, el diagrama de E-R tiene que estar bien diseñado, se tiene que hacer la transformación al modelo de datos relacional sin errores, en la creación de la base de datos en MySQL se deben de utilizar correctamente las sentencias de SQL y cuanto más correctas y completas estén mucho mejor (definiciones de restricciones, nomenclatura, identificadores, etc.) y en la inserción de datos se espera que se documenten tanto inserciones correctas como inserciones que permiten controlar los errores que impiden las restricciones.

Respecto a la inserción de datos, se espera que la inserción sea manual, insertando datos sintéticos pensados por los estudiantes. Por tanto, no es necesario insertar una gran cantidad de datos ni secuencias de una longitud grande, simplemente un ejemplo de funcionamiento.

Extensión y formato

La entrega constará de la documentación y resultado gráfico de cada una de las partes 2, 3, 4 y 5. Para el diagrama E-R es necesaria una visualización clara del diagrama con las entidades, relaciones, atributos y cardinalidad. Para las partes con SQL se esperan las propias sentencias, documentación sobre ellas y el resultado de configuración de las tablas para el punto 4 y el contenido de las tablas con datos insertados para el punto 5.

La extensión máxima para esta actividad es de veinte páginas, en fuente Calibri 12, con interlineado simple.

Rúbrica

|  |  |  |  |
| --- | --- | --- | --- |
| Título de la actividad | Descripción | Puntuación máxima  (puntos) | Peso  % |
| Criterio 1 | Diseño correcto del modelo E-R, definiendo las entidades, relaciones, atributos y cardinalidad correctas. | 3 | 30 % |
| Criterio 2 | Conversión correcta al modelo de datos relacional, utilizando la nomenclatura estándar y teniendo en cuenta normalización. | 2 | 20 % |
| Criterio 3 | Utilización correcta de las sentencias SQL para crear las tablas de datos. Dando importancia a sentencias profundas y con restricciones. | 3 | 30 % |
| Criterio 4 | Inserción correcta de los datos en las tablas. Buscando documentar tanto datos insertados correctos como datos incorrectos que el modelo no permite insertar. | 2 | 20 % |
|  |  | **10** | **100 %** |