



UNIVERSIDAD DE MÁLAGA



Graduado en Ingeniería de la Salud

Base de Datos Biológicas

Realizado por
David Cubillos del Toro
Nerea Martín Serrano

Tutorizado por
José Enrique Gallardo Ruiz

Departamento
Lenguajes y Ciencias de la Computación

MÁLAGA, abril de 2022



UNIVERSIDAD
DE MÁLAGA



ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA INFORMÁTICA
GRADUADO EN INGENIERÍA DE LA SALUD

**Base de datos de mutaciones en Homo
Sapiens**

Realizado por
David Cubillos del Toro
Nerea Martín Serrano

Tutorizado por
José Enrique Gallardo Ruíz

Departamento
Lenguajes y Ciencias de la Computación

UNIVERSIDAD DE MÁLAGA
MÁLAGA, 2022

Fecha defensa: julio de 2020

Resumen

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Palabras clave: A, B, C

Índice

1. Introducción	5
1.1. Motivación	5
1.2. Objetivos	5
1.3. Estructura del documento	5
1.4. Tecnologías usadas	5
2. Modelo Relacional MySQL	7
2.1. Entidades	7
2.2. Relaciones	9
3. Inserción de datos	11
3.1. Genes	11
3.2. Mutación	13
3.3. Efecto	13
3.4. Proteína	14
4. Conclusiones y Líneas Futuras	15
4.1. Conclusiones	15
4.2. Líneas Futuras	15
Apéndice A. Manual de Instalación	17

1

Introducción

1.1. Motivación

Las bases de datos que existen referentes a mutaciones permiten buscar un gen, y a partir de esa búsqueda qué mutaciones tiene ese gen. Pero no hay bases de datos, que introduciendo la mutación, diga cuál es el gen original.

Gracias a esta base de datos se puede introducir una mutación de un gen y ver cual era el gen original, además de ver a a qué proteína a afectado y la enfermedad que ha podido provocar este mutación.

1.2. Objetivos

El objetivo es crear una base da datos concisa para poder consultar de forma rápida. El usuario podrá introducir, por ejemplo, una mutación y ver sus consecuencias (la enfermedad que ha podido provocar, el gen que ha modificado, y, si ese gen era codificante, a la proteína a la que ha afectado).

1.3. Estructura del documento

¿?

1.4. Tecnologías usadas

¿?

2

Modelo Relacional MySQL

En esta sección vamos presentar y a explicar el Modelo ERR (Entidad-Relación Extendido) realizado en MySQL, describiendo cada una de la identidades, así como los atributos de cada una y las relaciones entre estas.

El ERR Diagram obtenido en MySQL es el siguiente:

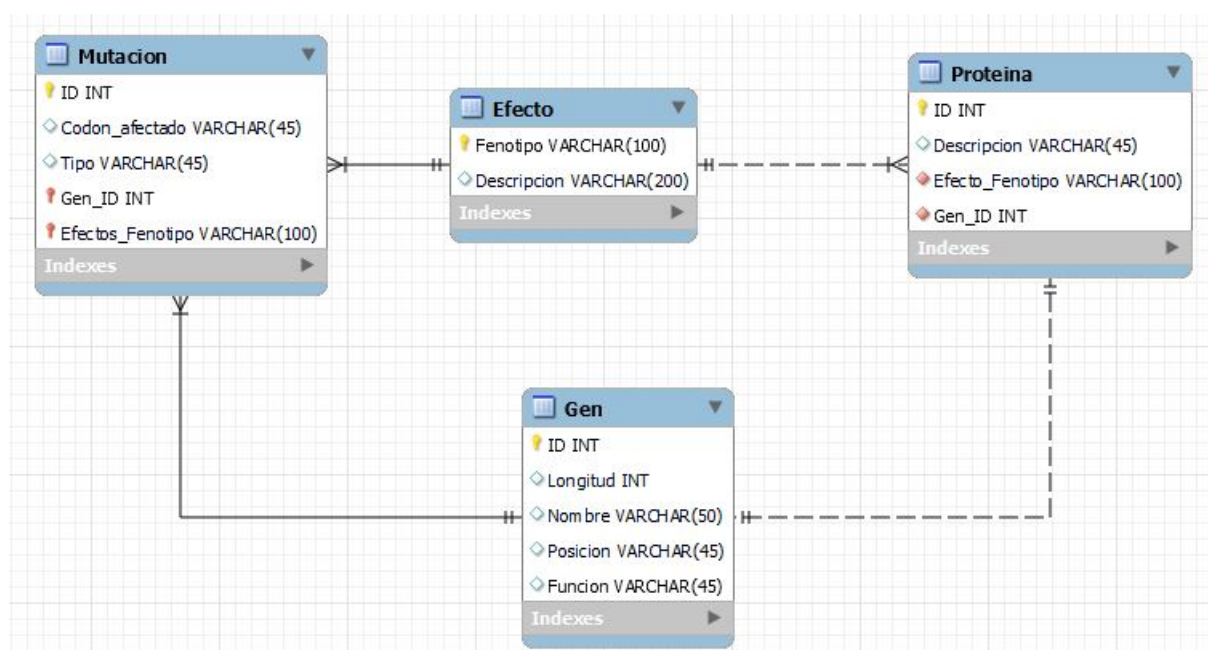


Figura 1: ERR Diagram

2.1. Entidades

Empezando describiendo las identidades que componen el diagrama:

- Gen: en esta entidad se va a encontrar información acerca del gen. Vamos a tener los

siguientes atributos:

- ID gen: esta será la clave primaria de esta clase. EL ID será el OMIM number.
 - Longitud: la longitud del gen en nucleótidos.
 - Nombre: nombre del gen, especificado por HUGO Nomenclature Committee.
 - Posición: posición cromosómica del gen.
 - Función: el tipo de función que realiza (si es un gen codificante, regulador,...)
- Proteína: esta entidad guarda la información relativa a las proteínas que codifican los genes que no tienen mutaciones. Es una identidad débil, pues podemos encontrar genes cuya función no se codifica a una proteína, pues se pueden encontrar genes cuya función sea reguladora y se encarguen de regular la transcripción de otros genes. Por lo tanto, es una entidad débil al depender de la función del gen (al depender de la entidad Gen).

Las proteínas van a ser identificadas por el código que tengan en UniProt, es decir que la clave principal, el ID, será este código. También se hará una descripción breve de la función de esta proteína. Como claves externas tendrá el fenotipo que es causado si la proteína no funciona correctamente, es decir, si una mutación impide que se transcriba de manera correcta, y por lo tanto pierda su función. La otra clave externa es el ID del gen que codifica a la proteína.

- Mutación: las mutaciones van a ser identificadas por el ID, coincidiendo este ID con el Accession Number de la base HGMD. Entre las claves externas encontramos el ID del gen (el OMIM number del gen original, antes de sufrir la mutación) y el fenotipo (la enfermedad que pueda causar la mutación). Por cada mutación se podrá ver a qué codón a afectado. Se podrán clasificar por tipos, siendo estos:
- Missense/nonsense: mutaciones que sustituyen solo un par de bases en la región codificante del gen.
 - Splicing: son mutaciones provocadas por el splicing del mRNA
 - Regulatory: sustituciones que causan anomalías reguladoras, se registran con treinta nucleótidos que flanquean el sitio de la mutación en ambos lados.

- Deletions: mutaciones que provocan la eliminación de pares de bases (pb). Estas se dividen en: small (20 pb o menos) y gross.
 - Insertions: introducción de pares de bases en la secuencia. También se dividen en small y gross.
 - Complex rearrangements
 - Repeat variations
- Efecto: hace referencia al efecto fenotípico que tenga la mutación. Este efecto es identificado por el fenotipo, es decir, el nombre de la enfermedad que cause la mutación. Además, esta identidad contendrá una breve descripción de dicha enfermedad.

2.2. Relaciones

Las identidades anteriores se relacionan entre si mediante las líneas que podemos ver en la figura 1. A En esta sección vamos a describir cada una de las relaciones:

- Gen-Mutación: es una relación 1:n pues un mismo gen pueden atener asociadas a varias mutaciones.
- Mutación-efecto: también es una relación 1:n porque varias mutaciones pueden tener el mismo efecto, es decir, pueden causar la misma enfermedad.
- Efecto-proteína: en este caso la relación esta representada por una línea discontinua al tratarse Proteína de una identidad débil. La relación es 1:n porque una enfermedad misma enfermedad puede ser causada por una misma proteína
- Gen-proteína: esta relación es 1:1 porque un gen codifica a una sola proteína (no vamos a tener en cuenta el splicing alternativo, por el cual un gen puede codificar a varias proteínas).

Inserción de datos

Para la inserción de datos lo primero que hemos hecho ha sido elaborar una lista de los símbolo de los genes (según HUGO Nomenclature Committee) que vamos a incluir en nuestra base de datos, teniendo en cuenta que todos los genes de nuestra lista tiene que tener una entrada en la base de datos HGMD (The Human Gene Mutation Database). A partir de esos genes de referencia hemos ido buscando en distintas bases de datos la información de interés y rellenado cada tabla.

3.1. Genes

Para el caso de la tabla genes, debemos de consultar en la base de datos HGMD por el Gene Symbol. En esta base de datos podemos encontrar el OMIM number, el cual iría en la columna ID. Para rellenar la columna sobre la longitud de la secuencia debemos de acceder a la base de datos Uniprot. Se puede acceder directamente a la referencia del gen que estamos buscando en Uniprot gracias a un enlace que se encuentra en HGMD. Cuando visitemos las base de datos de UniProt hay que fijarse que el gen corresponde a humanos (lo cual se hace mirando la columna Organism y fijándonos que ponga Homo Sapiens). El nombre de la secuencia se corresponde con el Gen symbol. La información relativa a la localización cromosómica se encuentra directamente en la base de datos HGMD. Por último, para saber que función tiene el gen debemos de acceder al NCBI con el OMIM number del gen en cuestión y buscar qué función realiza.

Por ejemplo, uno de los genes de la lista mencionada tiene el símbolo ABAT (este símbolo lo incluimos en la columna nombre). Hacemos una búsqueda en HGMD. En la página que nos sale tras esta búsqueda podemos encontrar casi todos los datos de interés, el OMIM number y la localización. Falta la longitud, y como hemos dicho antes, al final de la página, a la derecha, hay un enlace a la referencia de ese gen en la base de datos de UniProt. En la figura 1 se muestra

una captura de HGMD cuando hacemos la búsqueda del gen ABAT, subrayado en verde donde se encuentran los datos mencionados.

Gene Symbol	Chromosomal location	Gene name	cDNA sequence	Extended cDNA	Mutation viewer
ABAT (Aliases: available to subscribers)	16p13.2	4-aminobutyrate aminotransferase (Aliases: available to subscribers)	NM_020686.6	Not available	Available to subscribers
Mutation type		Number of mutations	Mutation data by type (register or log in)		
Missense/nonsense		9	Get mutations		
Splicing		0	No mutations		
Regulatory		0	No mutations		
Small deletions		0	No mutations		
Small insertions		0	No mutations		
Small indels		0	No mutations		
Gross deletions		2	Get mutations		
Gross insertions/duplications		1	Get mutations		
Complex rearrangements		0	No mutations		
Repeat variations		0	No mutations		
Get all mutations by type			Available to subscribers		
Public total (HGMD Professional 2021.4 total)		12 (17)			
Disease/phenotype		Number of mutations	Mutation data by disease/phenotype		
GABA-transaminase deficiency		10	Available to subscribers		
Mental retardation		1	Available to subscribers		
Neurodegeneration		1	Available to subscribers		
First published mutation report		PubMed	External links - ABAT		
Available to subscribers			OMIM 137150		
			GDB 581658		
Related genes			Entrez Gene entry		
Available to subscribers			Nomenclature Committee 23		
			SwissProt entry		
Gene ontology for ABAT			GeneCards entry		
Available to subscribers			GenAtlas entry		
			JSNP entry		
			COSMIC entry		
			GAD database		

Figura 2: Captura HGMD

Hacemos lo mismo para el resto de genes de la lista de genes, obteniendo la tabla de la figura 2.

ID	Longitud	Nombre	Posicion	Funcion
102545	376	ACTG2	2p13.1	coding gene
137150	500	ABAT	16p13.2	coding gene
167040	807	ARHDA1	11q12.1	coding gene
300011	1500	ATP7A	Xq13.2-q13.3	coding gene
601902	867	ORC1	1p32	coding gene
601925	204	OSBP	17q25.3	coding gene

Figura 3: Tabla genes

3.2. Mutación

Por cada gen existen varios tipos de mutaciones, por lo cual se han seleccionado dos de cada para ejemplificar el uso de la base de datos. En la figura 3 se muestra el repertorio de mutaciones de la pagina HGMD.


Mutation type	Number of mutations	Mutation data by type (register or log in)
Missense/nonsense	6	Get mutations
Splicing	2	Get mutations
Regulatory	0	No mutations
Small deletions	0	No mutations
Small insertions	0	No mutations
Small indels	1	Get mutations
Gross deletions	0	No mutations
Gross insertions/duplications	0	No mutations
Complex rearrangements	0	No mutations
Repeat variations	0	No mutations
Get all mutations by type	Available to subscribers 	
Public total (HGMD Professional 2021.4 total)	9 (15)	

Figura 4: Tipos de mutaciones para un gen

La primera columna será la columna “tipo” de nuestra mutación, pudiendo haber varias mutaciones de un mismo tipo.

Al hacer click en el botón “get mutation” nos llevará a una página con todas la mutaciones del tipo escogido(ver en figura 4), aquí se muestran algunos de los datos imputables en las tablas como el ID de mutación (que es la columna de accession number) el fenotipo con el que se va a relacionar dicha mutación o el cambio de codones causante de la mutación.

3.3. Efecto

En la tabla Efecto, se recopila información sobre las enfermedades asociadas con las mutaciones, dando el nombre de la patología y una breve descripción de esta. El nombre se obtiene







Missense/nonsense	Splicing	Regulatory	Small deletions	Small insertions	Small indels	Gross deletions	Gross insertions	Complex	Repeats
10 mutations in HGMD professional 2021.4	2 mutations in HGMD professional 2021.4	No mutations	2 mutations in HGMD professional 2021.4	No mutations	1 mutation in HGMD professional 2021.4	No mutations	No mutations	No mutations	No mutations
Further options available in HGMD professional 2021.4									
Accession Number	Codon change	Amino acid change	Codon number	Genomic coordinates & HGVS nomenclature	Phenotype	Reference		Comments	
CM112288	TTT-TCT	Phe-Ser	89	Available to subscribers 	Microcephalic primordial dwarfism	Bicknell (2011) Nat Genet 43, 350 Additional phenotype report available to subscribers Functional characterisation report available to subscribers Additional phenotype report available to subscribers			
CM112289	CGG-CAG	Arg-Gln	105	Available to subscribers 	Microcephalic primordial dwarfism	Bicknell (2011) Nat Genet 43, 350 Additional phenotype report available to subscribers Additional report available to subscribers Additional report available to subscribers Functional characterisation report available to subscribers Functional characterisation report available to subscribers			
CM112287	GAG-GGG	Glu-Gly	127	Available to subscribers 	Microcephalic primordial dwarfism	Bicknell (2011) Nat Genet 43, 350 Additional phenotype report available to subscribers Functional characterisation report available to subscribers			
CM124031	ACG-ATG	Thr-Met	574	Available to subscribers 	Meier-Gorlin syndrome ?	de Munnik (2012) Eur J Hum Genet 20, 598			
CM112304	CGG-TGG	Arg-Trp	666	Available to subscribers 	Meier-Gorlin syndrome	Guernsey (2011) Nat Genet 43, 360 Additional report available to subscribers Additional phenotype report available to subscribers			
CM112290	CGA-CAA	Arg-Gln	720	Available to subscribers 	Microcephalic primordial dwarfism	Bicknell (2011) Nat Genet 43, 350 Additional report available to subscribers Additional phenotype report available to subscribers			

Figura 5: Mutaciones de tipo Missense/nonsense para el gen ORC1

de la columna phenotype de la tabla de mutaciones de un tipo determinado (vista en inserto de datos en Mutaciones), para todas ellas se ha buscado un breve resumen en paginas médicas como la pagina de la clínica Mayo.

3.4. Proteína

Para rellenar esta tabla necesitamos tener completas con antelación las tablas Gen y Efecto, debido a que esta tabla necesita del ID del gen que codifica a la proteína y del fenotipo que se produce si la proteína no realiza su función de forma normal. Por lo tanto, comenzamos rellenando esta tabla por el gen que codifica a la proteína

Conclusiones y Líneas Futuras

4.1. Conclusiones

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

4.2. Líneas Futuras

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Apéndice A

Manual de Instalación



UNIVERSIDAD
DE MÁLAGA

| **uma.es**

E.T.S. DE INGENIERÍA INFORMÁTICA

E.T.S de Ingeniería Informática
Bulevar Louis Pasteur, 35
Campus de Teatinos
29071 Málaga