Report: Phase-1: BioJoin Basic

BiS332 course by Professor Doheon Lee

Lana Abu Hassan (20170842), Surayouth Phuksawattanachai (20180813), Alexander Semenov (20226159)

Introduction

This report aims to design an SNP-Gene-Disease database by integrating insight on biological information into database design. This database will allow us to understand the interaction among distinct genetic compositions. By implementing a python-based program, we may obtain the data by joining the three different databases. To examine the association among the SNPs, diseases, and genes, we can investigate various databases to get essential data. For instance, we may obtain a list of conditions associated with an SNP by introducing its key to the database. We may conveniently access the data among the three databases through their associative features. In particular, we will utilize data from dbSNP, Entrez gene, and OMIM database for our project. The primary database structure is illustrated in this ER diagram and the final attributes, their data types and the actual SQL query implementation will be shown later in this report.

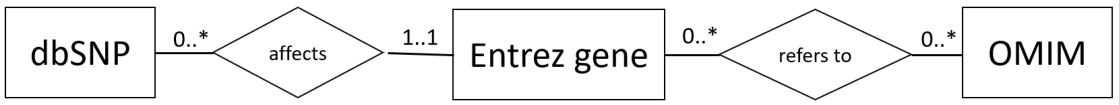
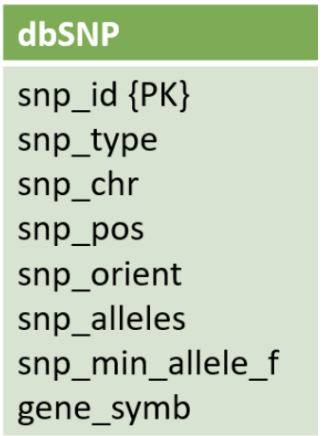


Figure 1: Primary ER diagram

Attributes



Attribute	Data type	Description
snp_id	integer (primary key)	Unique identifier for SNP entry
snp_type	character varying	Variant type of SNP e.g. SNV, DELINS
snp_chr	integer	Number of chromosome the gene is positioned on
snp_pos	character varying	Exact position of the SNP on given Chr.
snp_orient	boolean	Alleles orientation: 0 - forward, 1 - backward
snp_alleles	character varying	Major and minor alleles e.g. A>G
snp_min_allele_f	numeric(5,4)	Frequency of the minor allele
gene_symb	character varying	Relation of the SNP to the Gene

Table 1: dbSNP attributes with their type and description

Gene

gene_symb {PK}
gene_id
gene_type
gene_sum
gene_chr
gene_pos
gene_org
gene_mod_date
gene_syn
snp_id {FK}

Attribute	Data type	Description
gene_symb	character varying (primary key)	Official symbol provided by HGNC
gene_id	integer	ID for that particular locus in that organism
gene_type	character varying	Type of gene e.g. protein coding
gene_sum	text	Description of the gene
gene_chr	integer	Number of chromosome the gene is positioned on
gene_pos	character varying	Exact position of the Gene on a given Chr.
gene_org	character varying	Gene corresponding organism
gene_mod_date	date	Date of last modification of the entry
gene_syn	text []	Alternative gene symbols, synonyms
snp_id	integer (foreign key to table dbSNP)	Association of the gene entry to an SNP

Table 2: Gene attributes with their type and description

OMIM

omim_id {PK}
omim_name
omim_sum
omim_chr
gene_symb {FK}

Attribute	Data type	Description
omim_id	integer (auto-incremented) (primary key)	Unique identification of OMIM entry
omim_name	character varying	Name of the human gene and/or genetic disorder
omim_sum	text	Description of the disease
omim_chr	integer	Chromosome where it is positioned on
gene_symb	character varying (foreign key to table Gene)	Related gene symbol. Association with a concrete gene from the Gene table.

Table 3: OMIM attributes with their type and description

Final ER diagram

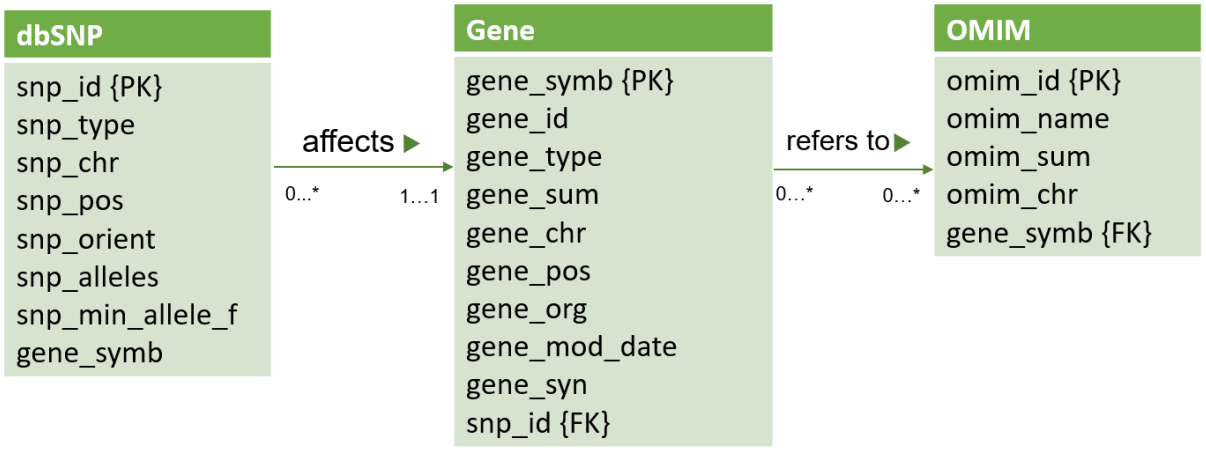


Figure 2: Final ER diagram

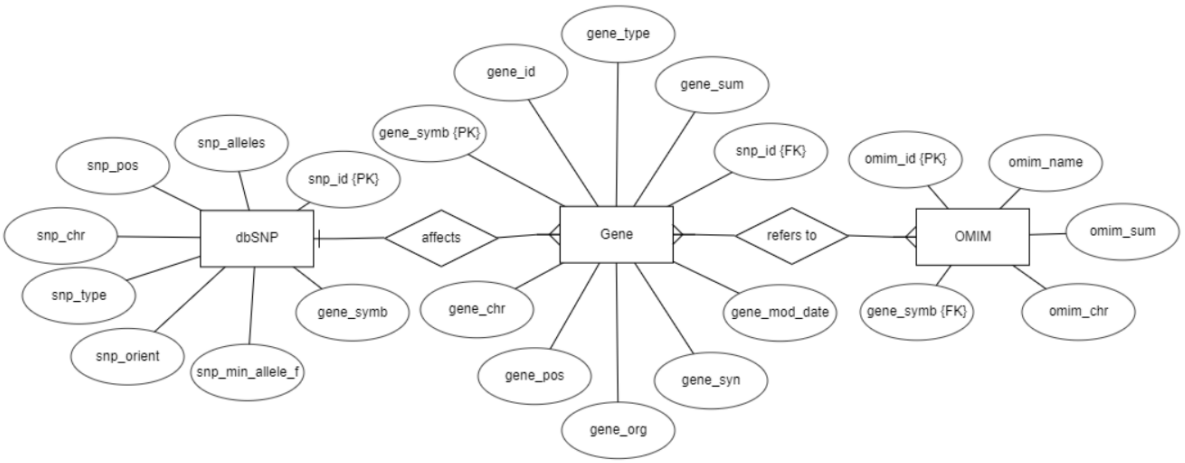


Figure 3: Combined ER diagram with symbolic description

Relational Schema

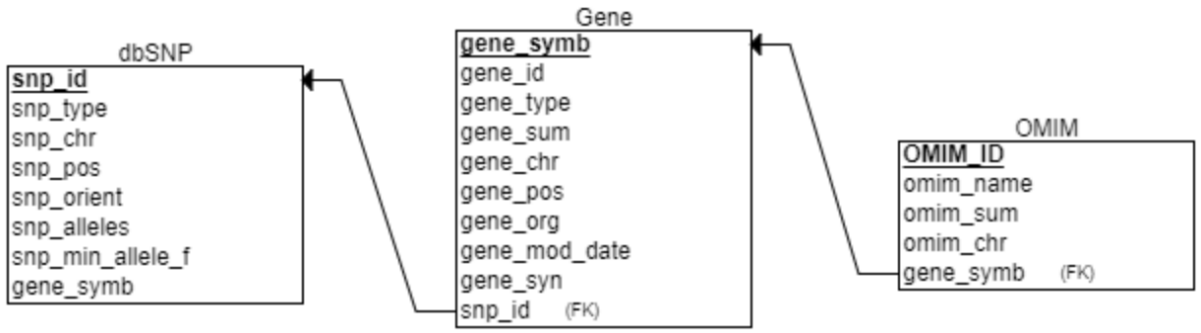


Figure 4: Relational schema with primary and foreign key assignment

DDL implementation

```
CREATE TABLE dbSNP (
  snp_id int NOT NULL,
  snp_type varchar,
  snp_chr int,
  snp_pos varchar,
  snp_orient boolean,
  snp_alleles varchar,
  snp_min_allele_f decimal(5,4),
  gene_symb varchar,
  CHECK (snp_min_allele_f <= 1 AND snp_min_allele_f >= 0),
  PRIMARY KEY (snp_id)
);

CREATE TABLE Gene (
  gene_symb varchar NOT NULL,
  gene_id int UNIQUE,
  gene_type varchar,
  gene_sum text,
  gene_chr int,
  gene_pos varchar,
  gene_org varchar,
  gene_mod_date date NOT NULL,
  gene_syn text [],
  snp_id int,
  PRIMARY KEY (gene_symb),
  FOREIGN KEY (snp_id) REFERENCES dbSNP(snp_id) ON DELETE SET NULL
);

CREATE TABLE OMIM (
  omim_id serial PRIMARY KEY,
  omim_name varchar,
  omim_sum text,
  omim_chr int,
  gene_symb varchar,
  FOREIGN KEY (gene_symb) REFERENCES Gene(gene_symb) ON DELETE SET NULL
);
```

Explanation of the database architecture

We have three relational databases for our ER diagram to organize the SNP, genes, and diseases related data. This enables us to search for conditions highly likely to occur with a specific gene through a mediator dbSNP table and the other way round.

In order to define the essential attributes of the tables and the relations between them, we have explored the data provided by The National Center for Biotechnology Information (NCBI). NCBI houses a series of databases relevant to biotechnology and biomedicine and is approved and funded by the government of the United States. We will use the Entrez gene database (https://www.ncbi.nlm.nih.gov/gene) and use the gene identifier provided by the HGNC (https://www.genenames.org/) for our Gene table. The dbSNP table sources will come from The Single Nucleotide Polymorphism Database (https://www.ncbi.nlm.nih.gov/snp/), which is a free public archive for genetic variation within and across different species developed and hosted by NCBI. The OMIM table about human genes and genetic disorders and traits will be served with the NCBI data, particularly from the Online Mendelian Inheritance in Man catalog (https://omim.org/).

For SNP attributes, the identification of an SNP entry is according to its id, so we set it as a primary key since it is a unique identifier in the primary source. Some attributes describe the position of the SNP in which the SNP are designated and further used for gene mapping: snp_chr, snp_pos, and snp_orient. The type of SNP is included too. The alleles contain the information of nucleotides in which mutation occurs, including the minor and major alleles in the form of “<major> > <minor>”. The minor allele frequency is stored in snp_min_allele_f and should have a value between 0 and 1 (we implemented SQL CHECK constraint to ensure it). For each dbSNP entry, a gene is also assigned if it is listed in the NCBI database.

We assign the primary key to the gene symbol for Gene table attributes since it is a unique identifier that rarely changes. The gene id is more stable and could be a good candidate for the primary key, but in most cases, the attribute did not appear in the OMIM and the dbSNP databases of NCBI while the gene symbol did. Hence, we may combine the gene and OMIM data by assigning the foreign key to the gene symbol attribute in the OMIM database. The gene id is still a good attribute that requires identifying the gene entry in case it has changed its gene symbol. This explains why alternative characters are stored in the table as well. We want to track the changes in those symbols so users searching for old signs could still find the same gene with the new naming. This attribute will be kept optional since there can be a possibility that there were no changes in the naming of the gene. General information about the gene as its short description, the origin organism, and the type, are included in the table. To identify genes that might share a regulatory region, we want to store the information about the genomic context of our genes. Links to genomic DNAs will help us locate the chromosome on which the gene is located. So, as for the SNP table, we hold the information about the gene position with the attributes gene_chr and gene_pos.

There are no identification numbers found in the primary source for the diseases table that could be mapped with SNP and Gene tables too. So, we assign OMIM id for each entry by ourselves using an auto-incremented integer function provided by the SQL query language. This means that a unique number is generated automatically for every new entry and will be the primary key for the OMIM table entry. Furthermore, we include basic information about each genetic disorder or disease as the name, textual description, and chromosome attributes (omim_name, omim_sum, and omim_chr, respectively).

One gene can have many SNPs, so the dbSNP table's relation to the Gene table is many-to-one. On the other hand, various genes map different diseases, and different genes can be associated with the same disease. Therefore, the relation between the Gene table and the OMIM table is many-to-many. We connect the Genes with the SNPs with help of the foreign key snp_in the Genes table. For OMIM attributes, we assign the foreign key to the gene symbol in order to join this OMIM table to the gene table which consists of the gene symbol as the primary key.

References

[1] National Library of Medicine, National Center for Biotechnology Information. ***dbSNP.*** https://www.ncbi.nlm.nih.gov/snp/. Retrieved on 13rd April 2022.

[2] National Library of Medicine, National Center for Biotechnology Information. ***Gene.*** https://www.ncbi.nlm.nih.gov/gene. Retrieved on 13rd April 2022.

[3] Johns Hopkins University. ***OMIM.*** https://omim.org/. Retrieved on 13rd April 2022.

[4] ***Entity–relationship model***. Entity–relationship model - Wikipedia. Retrieved on 13rd April 2022.

[5] Supreya Saxena. ***Relation Schema in DBMS.*** Relation Schema in DBMS - GeeksforGeeks. Retrieved on 13rd April 2022.

[6] Alvaro Monge. ***Basic SQL statements: DDL and DML.*** Database Design - DDL & DML (csulb.edu). Retrieved on 13rd April 2022.