BCHB697: Databases for Bioinformatics
Homework 3
Jiayao Wang (jw1509)

(1) Inspect your own universal table carefully. Identify as many entities and relationships as you can and write down their definitions or business rules in a Word document.

Entities:
Uniprot Protein; MIM phenotype; Disease ontology.
New entities after 2&3 NF:
HGNC gene, gene names, protein names
Relationships:
Protein many to one gene
Gene one to many Protein
Protein many to many MIM phenotype
MIM phenotype many to many disease ontology
New relationships after 2&3 NF:
Protein one to many protein names
Gene one to many gene names
Uniprot Disease name one to one Uniprot Disease short name
Business statements:
This database is to track the various issues that are related to several diseases and their related proteins as well as related genes.
Business rules:

1. Each protein have a target disease, a length, uniprot entry, a mass (Da), an author, a link to Uniprot, HGNC id, a link to HGNC, an approved name, an approved symbol, and a location on chromosome.
2. Each protein may have zero to four MIM phenotype
3. Each MIM phenotype have a MIM id, Uniprot disease name, short name.
4. Each MIM phenotype may have zero to two Disease Ontology
5. Each Disease Ontology have a DOID, a diease name, a definition and a relashionship, MIM id, Uniprot disease name, short name.

New Business rules after 2&3 NF

6. Each protein have a target disease, a length, uniprot entry, a mass (Da), an author, a link to Uniprot and may have a gene.
7. Each protein may have zero to four MIM phenotype
8. Each MIM phenotype have a MIM id, Uniprot disease name, short name.
9. Each MIM phenotype may have zero to two Disease Ontology
10. Each Disease Ontology have a DOID, a diease name, a definition and a relashionship.
11. Each Disease Ontology may have many related MIM phenotype.
12. Each MIM phenotype may have relation with many Proteins.
13. Each protein have one primary name but may have many alt names
14. Each protein name related to one protein, no matter it is an alt name or primary name.
15. Each gene have a HGNC id, a link to HGNC, an approved name, an approved symbol, and a location on chromosome.
16. Each gene have one primary name and may have many Synonyms.

17. Each gene name related to one gene, no matter it is an approved name or synonyms.

(2) Describe the universal table using the short table notation technique. List three non-trivial determinants in the universal table with this format "determinant -> columns dependent on this determinant".

Short table notation of the universal table:
ProteinToDeseases (Item #, Creator, Targeted Disease, UniProtKB Entry, URL(uniform resource locator or website address)-UniProtKB Entry, Protein Name, Alternative Names, Sequence Length, Mass(Da), UniProt Gene Name, HGNC ID, URL-HGNC ID, Approved Symbol, Approved Name, Synonyms, Chromosomal Location, Phenotype MIM #1, UniPro Disease Name, DOID #1, Disease Name, Definition, Relationship DOID #2, Disease Name, Definition, Relationship, Phenotype MIM #2, UniPro Disease Name, DOID #1, Disease Name, Definition, Relationship DOID #2, Disease Name, Definition, Relationship, Phenotype MIM #3, UniPro Disease Name, DOID #1, Disease Name, Definition, Relationship DOID #2, Disease Name, Definition, Relationship, Phenotype MIM #4, UniPro Disease Name, DOID #1, Disease Name, Definition, Relationship DOID #2, Disease Name, Definition, Relationship)

Three non-trivial determinants:
Item # or Uniprot entry -> the rest of columns
Phenotype MIM and UniPro Disease Name -> DOID #1; Disease Name; Definition; Relationship; DOID #2; Disease Name; Definition; Relationship.
HGNC ID -> URL-HGNC ID; Approved Symbol, Approved Name, Synonyms, Chromosomal Location.

(3) Nominate one of the candidate keys as the primary key for the universal table and explain your choice.

Uniprot entry can be a good choice of primary key.
Each of other value no matter from which database all start from the certain uniprot entry id so at least Uniprot entry has the ability to determine rest of others.

(4) Normalize the universal table to 1NF. Describe the 1NF violations in a Word document. Use the short table notation to describe all the new tables.

1NF violations:
#1: each column has one and only one fact (one fact in one place)
Uniprot disease name include the name and short name, divided into Uniprot disease name and Uniprot disease short name.

#2: Each record is unique and can be identified by a primary key
Uniprot gene name and approved name (of HGNC) are the same so we need to delete one of them.

#3: no repeating groups or duplicate fields

There are two repeating groups in the table. First is Phenotype MIM #1, UniPro Disease Name, DOID #1, Disease Name, Definition, Relationship DOID #2, Disease Name, Definition, and Relationship. We have this fields repeat four times, we should divide them out of the protein table and create a “MIM phenotype” table, containing Phenotype MIM#, related Uniprot entry, UniPro Disease Name, DOID and its related things.
Second filed is DOID #, Disease Name, Definition, Relationship, it still repeat in according to each MIM phenotype, so it need to move out the MIM table and form a new table named “Disease Ontology” contains DOID, Uniprot entry, UniPro Disease Name,MIM phenotype #, Disease Name, Definition, Relationship.

All the new tables:
Protein (**Uniprot entry ID**, Target Disease ,Author ,Uniprot URL, Protein Name, Sequence Length, Mass (Da), HGNC ID, HGNC URL, Approved Symbol, Approved Name, Chromosome Location)
MIM phenotype (**Uniprot entry ID**, **MIM phenotype ID**, **UniPro Disease Name**, short Name)
Disease Ontology (**DOID**, Disease Name, Definition, Relationship, **MIM phenotype ID**, **UniPro Disease Name**)

(5) Normalize all 1NF tables to 2NF then to 3NF tables. Use the short table notation to describe all the new tables that are created for resolving violations to the normal forms.

2NF:
Rule #1: It must be in First Normal Form.
Rule #2: No part of the primary key can determine any non-key columns.
In table of MIM phenotype
(**Uniprot entry ID**, **MIM phenotype ID**, **UniPro Disease Name**, short Name)
Short name can be determined by **UniPro Disease Name** whictch is part of PK. So we need to split it out to from a new table:
Disease short name (**UniPro Disease Name,** short name)
PK: **UniPro Disease Name** FK: None
MIM phenotype (**MIM phenotype ID**, **UniPro Disease Name**, **Uniprot entry ID**)
PK: **MIM phenotype ID**, **UniPro Disease Name**, **Uniprot entry ID** FK: **Uniprot entry ID**

In table of Disease ontology table:
Disease Ontology (**DOID**, Disease Name, Definition, Relationship, **Uniprot entry ID**, **MIM phenotype ID**, **UniPro Disease Name**)
DOID as a part of PK can determine non key value of Disease Name, Definition, and Relationship. So we need to split these columns out and form a new table.
Disease Ontology (**DOID**, Disease Name, Definition, Relationship)
Disease Ontology to MIM (**DOID**, **Uniprot entry ID**, **MIM phenotype ID**, **UniPro Disease Name**)

3NF:
Rule #1: It must be in Second Normal Form.
Rule #2: The determinant of a non-key column is the primary key.
In the table of Proteins:

Protein (**Uniprot entry ID**, Target Disease ,Author ,Uniprot URL, Protein Name, Sequence Length, Mass (Da), HGNC ID, HGNC URL, Approved Symbol, Approved Name, Chromosome Location)
HGNC URL, Approved Symbol, Approved Name, Chromosome Location can be determine by HGNC ID, which violate the Rule #2. So we need to split out the columns.
Protein (**Uniprot entry ID**, Target Disease ,Author ,Uniprot URL, Protein Name, Sequence Length, Mass (Da), HGNC ID)
HGNC gene (**HGNC ID**, HGNC URL, Approved Symbol, Approved Name, Chromosome Location)

More violations:
There are two columns violates the 1NF which we didn't fix previously. In Our new table of protein and HGNC gene:
Protein (**Uniprot entry ID**, Target Disease ,Author ,Uniprot URL, Protein Name, Alternative Names, Sequence Length, Mass (Da), HGNC ID)
HGNC gene (**HGNC ID**, HGNC URL, Approved Symbol, Approved Name, Synonyms, Chromosome Location)
Alternative Names in Protein and Synonyms in Genes violate the rule of one thing at one place, they have many names at one place. So we need to split them in to new name tables:
Protein (**Uniprot entry ID**, Target Disease ,Author ,Uniprot URL, Protein Name, Sequence Length, Mass (Da), HGNC ID)
HGNC gene (**HGNC ID**, HGNC URL, Approved Symbol, Approved Name, Chromosome Location)
Protein name (**name id**, Uniprot entry ID, name) PK: name id (new assigned surrogate key) FK: Uniprot entry ID.
Gene name (**name id**, HGNC ID, name) PK: name id (new assigned surrogate key) FK: HGNC ID.

Tables in short table notations after 2&3 NF:

Protein (**Uniprot entry ID**, Target Disease ,Author ,Uniprot URL, Protein Name, Sequence Length, Mass (Da), HGNC ID)
PK: **Uniprot entry ID** FK: HGNC ID to Gene

Protein name (**Protein name id**, Uniprot entry ID, Name)
PK: **Protein name id** FK: Uniprot entry ID to Protein

Gene (**HGNC ID**, HGNC URL, Approved Symbol, Approved Name, Chromosome Location)
PK: **HGNC ID** FK: None

Gene name (**Gene name id**, HGNC ID, Name)
PK: **Gene name id** FK: HGNC ID to Gene

MIM phenotype (**MIM phenotype ID**, **UniPro Disease Name**, **Uniprot entry ID**)
PK: **MIM phenotype ID**, **UniPro Disease Name**, **Uniprot entry ID** FK: **Uniprot entry ID** to ***Protein,*** **UniPro Disease Name** to ***Disease short name***

Disease Ontology (**DOID**, Disease Name, Definition, Relationship)
PK: **DOID** FK: None

Disease Ontology to MIM (**DOID, MIM phenotype ID**, **UniPro Disease Name**, **Uniprot entry ID**)
PK: **DOID, MIM phenotype ID**, **UniPro Disease Name**, **Uniprot entry ID** FK: DOID to ***Disease Ontology***, **MIM phenotype ID**, **UniPro Disease Name**, **Uniprot entry ID** to ***MIM phenotype***

Disease short name (**UniPro Disease Name,** short name)
PK: **UniPro Disease Name** FK: None