**P2P Database**

**BCHB697: Databases for Bioinformatics**

**Business Statement**

The protein to phenotype database provides the user a further understanding of the relationships that proteins have to various diseases. The protein information will be extrapolated from UniProt while the various disease information will be taken from Disease Ontology.

**Additional Business Rules**

* A data collector may provide information about two or more target disease. Each target disease may have two more proteins.
* Every protein to phenotype relation is tracked using the “item #” column which is assigned a value based on the following sequential order 1,2,3,…N where is an integer.
* Each protein must only have one UniProtKB Entry which will be unique from all other proteins.
* Each protein must also have a UniProt URL, protein name, protein sequence length, and mass in Daltons. The latter 2 will be in the form of integers.
* Each protein may have a list of alternative names.
* Each protein must the following gene information: UniProt Gene Name, HGNC ID, HGNC URL, approved symbol, approved name, and chromosomal location.
* Each protein may have the following gene information: list of synonyms.
* Each protein may have the following disease information: One or more UniProt disease name, and one or more phenotype MIM, one or more DOID, disease name, definition, and relationship.
* Each UniProt disease name will correspond to one phenotype MIM. However the phenotype MIM may correspond to one or more DOID which in turn may be related to one or more disease name, definition, and relationship.
* A protein can have identical disease information as another protein.

**Conceptual Data Model**

Entities: Protein and Disease

Relationship: many to many

**Protein**

**Disease**

ddffdf

**Determinants**

* Item # 🡪 rest of columns
* UniProtKB Entry 🡪 P2P + Disease information
* UniProtKB Entry + Phenotype OMIM 🡪 Disease information
* UniProtKB Entry + Phenotype OMIM + DOID 🡪 Disease Ontology Information

**Normalization**

The short table notation is utilized to describe the P2P universal table, containing 56 columns, below.

**P2P** ( item #, Data Collected by, Targeted Disease, UniProtKB Entry, URL-UniProtKB Entry, Protein Name, Alternative Names, Sequence Length, Mass(Da), UniProt Gene Name, HGNC ID, URL-HGNC ID, Approved Symbol, Synonyms, Chromosomal Location, Phenotype MIM #1, UniProt Disease Name, DOID #1, Disease Name, Definition, Relationship, DOID #2, Disease Name, Definition, Relationship, Phenotype MIM #2, UniProt Disease Name, DOID #1, Disease Name, Definition, Relationship, DOID #2, Disease Name, Definition, Relationship, Phenotype MIM #3, UniProt Disease Name, DOID #1, Disease Name, Definition, Relationship, DOID #2, Disease Name, Definition, Relationship, Phenotype MIM #4, UniProt Disease Name, DOID #1, Disease Name, Definition, Relationship, DOID #2, Disease Name, Definition, Relationship)

**Stage 1: Organize the data such that the resulting tables are in First Normal Form (1NF)**

**1NF rule#1: Each column has one and only one fact (one fact in one place)**

There are multiple values in the UniProt Disease name columns. This is 1NF violation, specifically rule #1 which states that each column will have one and only one fact. In order to resolve this violation, the columns are split into UniProt Disease Shortened Name and UniProt Disease Name.

**P2P** ( item #, Data Collected by, Targeted Disease, UniProtKB Entry, URL-UniProtKB Entry, Protein Name, Alternative Names, Sequence Length, Mass(Da), UniProt Gene Name, HGNC ID, URL-HGNC ID, Approved Symbol, Synonyms, Chromosomal Location, Phenotype MIM #1, UniProt Disease Shortened Name, UniProt Disease Name, DOID #1, Disease Name, Definition, Relationship, DOID #2, Disease Name, Definition, Relationship, Phenotype MIM #2, UniProt Disease Shortened Name, UniProt Disease Name, DOID #1, Disease Name, Definition, Relationship, DOID #2, Disease Name, Definition, Relationship, Phenotype MIM #3, UniProt Disease Shortened Name, UniProt Disease Name, DOID #1, Disease Name, Definition, Relationship, DOID #2, Disease Name, Definition, Relationship, Phenotype MIM #4, UniProt Disease Shortened Name, UniProt Disease Name, DOID #1, Disease Name, Definition, Relationship, DOID #2, Disease Name, Definition, Relationship)

**1NF rule#2: Each row in a table is unique and can be identified by the primary key**

At this point in the normalization process the universal table does not have a primary key defined which would be a 1NF rule #2 violation. The primary key that will be selected for the P2P table is the UniProtKB Entry column. This will be underlined in the following short table notation and place as the first column.

**P2P** (UniProtKB Entry, item #, Data Collected by, Targeted Disease, URL-UniProtKB Entry, Protein Name, Alternative Names, Sequence Length, Mass(Da), UniProt Gene Name, HGNC ID, URL-HGNC ID, Approved Symbol, Synonyms, Chromosomal Location, Phenotype MIM #1, UniProt Disease Shortened Name, UniProt Disease Name, DOID #1, Disease Name, Definition, Relationship, DOID #2, Disease Name, Definition, Relationship, Phenotype MIM #2, UniProt Disease Shortened Name, UniProt Disease Name, DOID #1, Disease Name, Definition, Relationship, DOID #2, Disease Name, Definition, Relationship, Phenotype MIM #3, UniProt Disease Shortened Name, UniProt Disease Name, DOID #1, Disease Name, Definition, Relationship, DOID #2, Disease Name, Definition, Relationship, Phenotype MIM #4, UniProt Disease Shortened Name, UniProt Disease Name, DOID #1, Disease Name, Definition, Relationship, DOID #2, Disease Name, Definition, Relationship)

**1NF rule#3: There are no duplicate columns or repeating groups of columns**

After creating the primary key, the next step is to ensure that there are no more 1NF violations. After analyzing the short table notation above, it appears that the columns pertaining to the disease information are repeated four times and in addition there are duplicate column names being used. This violates the 1NF rule #3 which states that there should be no duplicate columns or repeating groups of columns. In addition to there being a 1NF violation, it should be addressed that a limit of four repeating disease groups is not sufficient because there could be more than four UniProt Disease names associated with each protein.

In order to resolve the 1NF rule #3 violation the repeating groups have to be identified; this is illustrated below with the repeated rows colored red.

**P2P** (UniProtKB Entry, item #, Data Collected by, Targeted Disease, URL-UniProtKB Entry, Protein Name, Alternative Names, Sequence Length, Mass(Da), UniProt Gene Name, HGNC ID, URL-HGNC ID, Approved Symbol, Synonyms, Chromosomal Location, Phenotype MIM #1, UniProt Disease Shortened Name, UniProt Disease Name, DOID #1, Disease Name, Definition, Relationship, DOID #2, Disease Name, Definition, Relationship, Phenotype MIM #2, UniProt Disease Shortened Name, UniProt Disease Name, DOID #1, Disease Name, Definition, Relationship, DOID #2, Disease Name, Definition, Relationship, Phenotype MIM #3, UniProt Disease Shortened Name, UniProt Disease Name, DOID #1, Disease Name, Definition, Relationship, DOID #2, Disease Name, Definition, Relationship, Phenotype MIM #4, UniProt Disease Shortened Name, UniProt Disease Name, DOID #1, Disease Name, Definition, Relationship, DOID #2, Disease Name, Definition, Relationship)

The repeated groups of columns can be combined together to minimize the repetion. This can be seen below, with the combination displayed in red.

**P2P** (UniProtKB Entry, item #, Data Collected by, Targeted Disease, URL-UniProtKB Entry, Protein Name, Alternative Names, Sequence Length, Mass(Da), UniProt Gene Name, HGNC ID, URL-HGNC ID, Approved Symbol, Synonyms, Chromosomal Location,

{Phenotype MIM, UniProt Disease Shortened Name, UniProt Disease Name, DOID #1, Disease Name, Definition, Relationship, DOID #2, Disease Name, Definition, Relationship})

After consolidating the repeated groups into a new combined set of columns it appears that the P2P table should be split into two separate tables: P2P and Disease. The process in creating this new table can be seen in the short table notations provided below. The new table will be named Disease.

**P2P** (UniProtKB Entry, item #, Data Collected by, Targeted Disease, URL-UniProtKB Entry, Protein Name, Alternative Names, Sequence Length, Mass(Da), UniProt Gene Name, HGNC ID, URL-HGNC ID, Approved Symbol, Synonyms, Chromosomal Location)

**Disease** (Phenotype MIM, UniProt Disease Shortened Name, UniProt Disease Name, DOID #1, Disease Name, Definition, Relationship, DOID #2, Disease Name, Definition, Relationship)

The next step after splitting the original table into two tables is to include the primary key, UniProtKB Entry, from the P2P table as a foreign key in the new Disease table and is illustrated below.

**P2P** (UniProtKB Entry, item #, Data Collected by, Targeted Disease, URL-UniProtKB Entry, Protein Name, Alternative Names, Sequence Length, Mass(Da), UniProt Gene Name, HGNC ID, URL-HGNC ID, Approved Symbol, Synonyms, Chromosomal Location)

**Disease** (UniProtKB Entry\*, Phenotype MIM, UniProt Disease Shortened Name, UniProt Disease Name, DOID #1, Disease Name, Definition, Relationship, DOID #2, Disease Name, Definition, Relationship)

After adding in the foreign key, a primary key has to be assigned for the newly created Disease table. The primary key will be a two-column key, including the UniProtKB Entry and the Phenotype MIM. By using these two columns, it will allow for the each Disease entry to be uniquely identified from each other because the same the same disease could be present in multiple proteins. The new primary key is displayed in red below.

**P2P** (UniProtKB Entry, item #, Data Collected by, Targeted Disease, URL-UniProtKB Entry, Protein Name, Alternative Names, Sequence Length, Mass(Da), UniProt Gene Name, HGNC ID, URL-HGNC ID, Approved Symbol, Synonyms, Chromosomal Location)

**Disease** (UniProtKB Entry\*, Phenotype MIM, UniProt Disease Shortened Name, UniProt Disease Name, DOID #1, Disease Name, Definition, Relationship, DOID #2, Disease Name, Definition, Relationship)

Now that there are two distinct tables, P2P and Disease, the 1NF rules are checked once again to ensure that there are no new violations. It appears that there is another repeating group of columns present in the newly formed Disease table. This includes the information pertaining to the disease ontology and include the following columns: DOID #1, Disease Name, Definition, Relationship, DOID #2, Disease Name, Definition, and Relationship. In order to resolve this issue the disease table has to be split up in order to eliminate the repeating group of columns. In addition, there are only two DOID groups present in the table, however there could be more than two groups of DOID information present in the data collection process. The repeated groups are colored in red below.

**P2P** (UniProtKB Entry, item #, Data Collected by, Targeted Disease, URL-UniProtKB Entry, Protein Name, Alternative Names, Sequence Length, Mass(Da), UniProt Gene Name, HGNC ID, URL-HGNC ID, Approved Symbol, Synonyms, Chromosomal Location)

**Disease** (UniProtKB Entry\*, Phenotype MIM, UniProt Disease Shortened Name, UniProt Disease Name, DOID #1, Disease Name, Definition, Relationship, DOID #2, Disease Name, Definition, Relationship)

The repeated columns in red can be combined into one group, to minimize the repetition and is illustrated below.

**P2P** (UniProtKB Entry, item #, Data Collected by, Targeted Disease, URL-UniProtKB Entry, Protein Name, Alternative Names, Sequence Length, Mass(Da), UniProt Gene Name, HGNC ID, URL-HGNC ID, Approved Symbol, Synonyms, Chromosomal Location)

**Disease** (UniProtKB Entry\*, Phenotype MIM, UniProt Disease Shortened Name, UniProt Disease Name,

{DOID, Disease Name, Definition, Relationship})

With the newly consolidated group, the Disease table can now be split into two table. The new table will be labeled DO, which stands for Disease Ontology.

**P2P** (UniProtKB Entry, item #, Data Collected by, Targeted Disease, URL-UniProtKB Entry, Protein Name, Alternative Names, Sequence Length, Mass(Da), UniProt Gene Name, HGNC ID, URL-HGNC ID, Approved Symbol, Synonyms, Chromosomal Location)

**Disease** (UniProtKB Entry\*, Phenotype MIM, UniProt Disease Shortened Name, UniProt Disease Name)

**DO** (DOID, Disease Name, Definition, Relationship)

The next step in creating this new DO table is to transfer the primary keys from the Disease table to the DO table and is displayed in red below.

**P2P** (UniProtKB Entry, item #, Data Collected by, Targeted Disease, URL-UniProtKB Entry, Protein Name, Alternative Names, Sequence Length, Mass(Da), UniProt Gene Name, HGNC ID, URL-HGNC ID, Approved Symbol, Synonyms, Chromosomal Location)

**Disease** (UniProtKB Entry\*, Phenotype MIM, UniProt Disease Shortened Name, UniProt Disease Name)

**DO** (UniProtKB Entry\*, Phenotype MIM\*, DOID, Disease Name, Definition, Relationship)

In order to satisfy the 1NF rule #2, a primary key has to be created for the DO table in order to uniquely identify each row of the table. The primary key will be a three-column primary key incorporating the foreign key as well as the DOID column. The reason for this is because each OMIM entry may have one or more disease ontology entries. The newly assigned primary key is illustrated below in red.

**P2P** (UniProtKB Entry, item #, Data Collected by, Targeted Disease, URL-UniProtKB Entry, Protein Name, Alternative Names, Sequence Length, Mass(Da), UniProt Gene Name, HGNC ID, URL-HGNC ID, Approved Symbol, Synonyms, Chromosomal Location)

**Disease** (UniProtKB Entry\*, Phenotype MIM, UniProt Disease Shortened Name, UniProt Disease Name)

**DO** (UniProtKB Entry\*, Phenotype MIM\*, DOID, Disease Name, Definition, Relationship)

After the completion of the 1NF process and validations, the original P2P Universal table has been normalized into three tables: P2P, Disease, and DO. Each of these table have their own primary keys and there is no redundancy within the data in the columns as well as no repetitions present. Below are the finalized tables after first normalization.

**P2P** (UniProtKB Entry, item #, Data Collected by, Targeted Disease, URL-UniProtKB Entry, Protein Name, Alternative Names, Sequence Length, Mass(Da), UniProt Gene Name, HGNC ID, URL-HGNC ID, Approved Symbol, Synonyms, Chromosomal Location)

**Disease** (UniProtKB Entry\*, Phenotype MIM, UniProt Disease Shortened Name, UniProt Disease Name)

**DO** (UniProtKB Entry\*, Phenotype MIM\*, DOID, Disease Name, Definition, Relationship)

**Stage 2: Organize the data such that the resulting tables are in Second Normal Form (2NF)**

**2NF Rule#1: Tables must be in First Normal Form (1NF)**

The three tables above, P2P, Disease, and DO are all in 1NF. (Look above)

**2NF Rule #2: No part of the primary key can determine any non-key columns. (All non-key columns depend on the whole primary key.)**

The P2P table does not have 2NF violations because the primary key (UniProtKB Entry) is a determinant for all of the other columns.

**P2P** (UniProtKB Entry, item #, Data Collected by, Targeted Disease, URL-UniProtKB Entry, Protein Name, Alternative Names, Sequence Length, Mass(Da), UniProt Gene Name, HGNC ID, URL-HGNC ID, Approved Symbol, Synonyms, Chromosomal Location)

UniProtKB Entry 🡪 Item#, Data Collected by, Targeted Disease, URL-UniPRotKB Entry, Protein Name, Alternative Names, Sequence Length, Mass(Da), UniProt Gene Name, HGNC ID, URL-HGNC ID, Approved Symbol, Synonyms, Chromosomal Location.

The Disease table appears to violate the second 2NF rule. One of the columns in the primary key determines non-key columns.

**Disease** (UniProtKB Entry\*, Phenotype MIM, UniProt Disease Shortened Name, UniProt Disease Name)

Phenotype MIM 🡪 UniProt Disease Shortened Name, UniProt Disease Name

In order to resolve this 2NF rule #2 violation, the Disease table has to be split. The following steps illustrate this process.

**Disease** (UniProtKB Entry\*, {Phenotype MIM, UniProt Disease Shortened Name, UniProt Disease Name})

The three columns listed in the brackets above will be used to create the new table that will be seen below.

**NewTable** (Phenotype MIM, UniProt Disease Shortened Name, UniProt Disease Name)

**Disease** (UniProtKB Entry\*, Phenotype MIM)

The newly created table will be called OMIM Informationand Phenotype MIM will now be a foreign key in the Disease table. The table splitting results in the new OMIM Information table as the Parent of the Disease table, which is the child.

**OMIM Information** (Phenotype MIM, UniProt Disease Shortened Name, UniProt Disease Name)

**Disease** (UniProtKB Entry\*, Phenotype MIM\*)

After resolving the 2NF violation in the Disease table the last table, DO, is checked for any violation. It appears that the DO table also violates the second rule; it has part of the primary key determining non-key columns.

**DO** (UniProtKB Entry\*, Phenotype MIM\*, DOID, Disease Name, Definition, Relationship)

DOID 🡪 Disease Name, Definition, Relationship

The DOID column of the primary key determines the non key columns listed above. In order to resolve this violation, the DO table has to be split.

**DO** (UniProtKB Entry\*, Phenotype MIM\*, {DOID, Disease Name, Definition, Relationship})

The columns listed above will be used to split the DO table and create a new table.

**DO Information** (DOID, Disease Name, Definition, Relationship)

**DO** (UniProtKB Entry\*, Phenotype MIM\*, DOID\*)

The newly created table will be called DO Informationand DOID will now be a foreign key in the DO table. The table splitting results in the new DO Information table as the Parent of the DO table, which is the child.

After resolving all 2NF violations the resulting tables are listed below.

**P2P** (UniProtKB Entry, item #, Data Collected by, Targeted Disease, URL-UniProtKB Entry, Protein Name, Alternative Names, Sequence Length, Mass(Da), UniProt Gene Name, HGNC ID, URL-HGNC ID, Approved Symbol, Synonyms, Chromosomal Location)

**OMIM Information** (Phenotype MIM, UniProt Disease Shortened Name, UniProt Disease Name)

**Disease** (UniProtKB Entry\*, Phenotype MIM\*)

**DO Information** (DOID, Disease Name, Definition, Relationship)

**DO** (UniProtKB Entry\*, Phenotype MIM\*, DOID\*)

**Stage 3: Organize the data such that the resulting tables are in Third Normal Form (3NF)**

**3NF Rule#1: All tables must be in Second Normal Form (2NF)**

After the 2NF normalization, all tables above are in second normal form.

**3NF Rule#2:** **The determinant of a non-key column is the primary key. (The primary key is the only determinant)**

It appears that the there is a violation to the second 3NF rule in the P2P table.

**P2P** (UniProtKB Entry, item #, Data Collected by, Targeted Disease, URL-UniProtKB Entry, Protein Name, Alternative Names, Sequence Length, Mass(Da), UniProt Gene Name, HGNC ID, URL-HGNC ID, Approved Symbol, Synonyms, Chromosomal Location)

HGNC ID (non-key) 🡪 URL-HGNC ID (non-key), Approved Symbol (non-key), Synonyms (non-key), Chromosomal Location (non-key)

To resolve the above violation the P2P table has to be split into two tables.

**P2P** (UniProtKB Entry, item #, Data Collected by, Targeted Disease, URL-UniProtKB Entry, Protein Name, Alternative Names, Sequence Length, Mass(Da), UniProt Gene Name, {HGNC ID, URL-HGNC ID, Approved Symbol, Synonyms, Chromosomal Location})

Above the columns in the brackets signify the columns used to create the new tables.

**P2P** (UniProtKB Entry, item #, Data Collected by, Targeted Disease, URL-UniProtKB Entry, Protein Name, Alternative Names, Sequence Length, Mass(Da), UniProt Gene Name HGNC ID)

**NEWTABLE** (HGNC ID, URL-HGNC ID, Approved Symbol, Synonyms, Chromosomal Location)

The split above displays the newly formed table, which has the primary key as HGNC ID. The new table will be named Gene as displayed below. The P2P will have HGNC ID as a foreign key column because the Gene table will be a parent and the Gene table will be a child.

**P2P** (UniProtKB Entry, HGNC ID\*, item #, Data Collected by, Targeted Disease, URL-UniProtKB Entry, Protein Name, Alternative Names, Sequence Length, Mass(Da), UniProt Gene Name)

**Gene** (HGNC ID, URL-HGNC ID, Approved Symbol, Synonyms, Chromosomal Location)

At this point all of the tables appear to be in 3NF. They all satisfy the 1NF rules by having single entries in columns, all rows can be identified by a primary key and are unique, and there are no duplicate or repeating columns. The tables all satisfy 2NF because all primary keys as a whole determine the rest of the non-key columns. Finally they are all in 3NF because there are no non-key columns that determine any other columns, the primary key is the only column/columns that determines non-key columns. All the 3NF tables are listed below.

**P2P** (UniProtKB Entry, HGNC ID\*, item #, Data Collected by, Targeted Disease, URL-UniProtKB Entry, Protein Name, Alternative Names, Sequence Length, Mass(Da), UniProt Gene Name)

**Gene** (HGNC ID, URL-HGNC ID, Approved Symbol, Synonyms, Chromosomal Location)

**OMIM Information** (Phenotype MIM, UniProt Disease Shortened Name, UniProt Disease Name)

**Disease** (UniProtKB Entry\*, Phenotype MIM\*)

**DO Information** (DOID, Disease Name, Definition, Relationship)

**DO** (UniProtKB Entry\*, Phenotype MIM\*, DOID\*)