Gaelle Sop-Kamga Emmanuelle

Protein P2PU Database
For
BCHB697: Databases for Bioinformatics
Spring 2016

Business Statement

This database is to track the protein information and related diseases associated with 2 diseases of interest, Schizophrenia and Malaria. Data acquired from UniprotKB, HGNC and the disease ontology resource are store in this database. This data include protein and gene information, and definitions of other diseases that have protein in common with our diseases of interest.

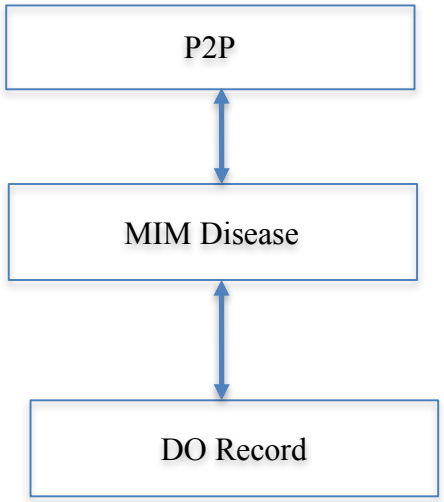
Additional Business Rules

- Multiple piece of data (rows) can be collected by a person, but a piece of data (a row) can be collected by only one person.
- A targeted disease can yield multiple uniprotKB entries. A uniprotKB entry number is always unique.
- A uniprot entry can correspond to multiple diseases, but in this table, to only one targeted disease.
- Each uniprot entry has exactly only one uniprot url. This url can be derived from the uniprot entry number. It is unique to uniprot entry.
- Each uniprot entry must have exactly one full protein name. This is unique to a uniprot entry. It is also required for any uniprot entry.
- Each uniprot entry may have one set of Alternative names unique to the protein.
- Each uniprot entry must have one sequence length and one mass. Multiple entries may have the same sequence length or the same mass.
- Each uniprot entry must have one gene name. This is unique to a uniprot protein.
- Each uniprot entry must have exactly one HGNC ID. This is unique to the entry. It is accompanied by the HGNC URL, which is derivable from the HGNC ID, and also unique to one entry.
- Each uniprot entry must have a unique HGNC approved symbol and approved name. Each must also have a unique set of HGNC synonyms.

- Each uniprot entry must a Chromosomal location. Multiple uniprot entries may have the same chromosomal location.
- Each uniprot entry may have an MIM ID Identifying a disease. Multiple uniprot entries may have the same MIM ID.
- Each MIM ID must have a uniprot disease name. Each MIM disease is tracked by the column item # and the uniprotID.
- Each MIM ID may have one or more DO records. DO records are tracked by the uniprot ID and the MIM ID.
- Each DO record must have a unique DO ID and a disease name. Multiple uniprot entries may have the same DO ID and record.
- Each DO entry must be related to an MIM entry. We are more interested in DO entries related to schizophrenia disease.

Conceptual Data Model

- Entities: Uniprot Entry, MIM disease, DO entry
- Relationships: many to many, many to many.



Determinants

- Item # -> the rest of columns
- UniprotKB Entry-> The rest of columns
- UniprotKB Entry+ Phenotype MIM ID -> MIM disease columns
- UniprotKB Entry+Phenotype MIM ID+DO ID-> DO columns