**Project Overview**

Description:

This project revolves around the creation of a database of protein-protein interactions for all genes that are listed as genetic contributors to a particular phenotype of Parkinson’s Disease on the database **OMIM** (*Online Mendelian Inheritance in Man).* It includes the acquisition, refiniement, and integration of clinical, genetic, and protein information from three different biological databases (**OMIM**, **STRING**, and **Uniprot)**, allowing for mysql queries of protein-protein interaction data relating to a particular Parkinson’s Phenotype – Gene pair featured on OMIM’s database.

Motivation:

The current focus of most databases dedicated to Parkinson’s research revolves around the archiving of genetic variants associated with the disease, but there does not appear to be any existing database that aims to compile protein interaction data for genetic factors that already have strong experimentally established associations with the various phenotypes of Parkinson’s.

Relatively few research publications exist that have attempted to generate networks of protein interactions for any form of Parkinson’s Disease. Developing a database for storing PPI’s from an established pool of PD-associated genes could serve as a foundation for any research that aims to identify and further understand molecular processes involved with Parkinson's.

Queries:

Based on the design and the structure of the database, queries can arise from two different situations:

* An individual already has data on a particular phenotype of Parkinson’s and/or its associated gene, and desires to further explore any protein-interaction data associated with that gene (irrespective of its different isoforms), based on some specific criteria.
* An individual has a specific protein (or set of proteins), and wants to determine if there is any significant interaction data that exists between that protein and one associated with a phenotype of Parkinson’s.

Objectives:

* Acquire all genes featured within the Parkinson’s *Phenotypic Series* list on **OMIM** and map them to their corresponding STRING\_IDs.
* Generate a comprehensive list of all protein-protein interactions (observed and predicted) for each Parkinson’s gene using the **STRING** database.
* Generate a list of each protein that interacts with at least one of the Parkinson’s genes, and supplement individual protein entries with information from **Uniprot**.
* Create a database using ***MySql*** that integrates all the above-mentioned data obtained from **OMIM,** **STRING,** and **Uniprot**.

Goal:

The Over-arching goal of this project was to provide a database that attempted to act as a basic support structure for research aimed at using Protein-Protein Interaction networks for identifying and understanding molecular mechanisms involved in the development of Parkinson’s Disease.

**Final status**: All in all, the completed project achieved each of the objectives listed above, and can function (in a very simple capacity) as database resource for any research involved in Parkinson's protein interactions,

**Data**

Data Sources/Specs:

The different data needed for this project were split up between 4 different categories:

* Parkinson's Disease and Parkinson’s Genes data (**OMIM**)
  + file type – tsv
  + size – 3 kB
  + Method – Website download
* Parkinson’s Gene data mapped to STRING identifiers and given additional annotations (**STRING**)
  + file type – json
  + size – 10 kB
  + Method - API
* Parkinson's Protein-Protein Interaction data (**STRING**)
* file type – tsv
  + size – 144 kB
* Method - API
* None\_Parkinson’s Protein data (**Uniprot**)
  + file type – tsv
  + size – 232kB
  + Method - API

Pitfalls:

Numerous problems and difficulties were encountered, both with the use of the API for **STRING**, and the importing of data into ***MySql.***

API Issues:

* **OMIM –** OMIM does not allow unregistered use of its API, so data was obtained by manually downloading it from the webpage link for Parkinson's Disease. Because all the required information was contained within one download file, it was probably just as efficient (if not faster) to do it this way, instead of trying to register, write scripts and access the data in a programmatic way.
* **STRING** – When using STRING’s API to obtain data on each of the Parkinson's genes from OMIM, 4 out of 30 of the Parkinson’s genes did not register with STRING, nor did they have any matching synonyms or other common names that could be mapped to a STRING ID. As a result of this, 4 different Parkinson's Phenotype-Gene pairs were omitted from the database, including:
  + **{Parkinson disease 16} – PARK16**
  + **{Parkinson disease 3} - PARK 3**
  + **{Parkinson disease, susceptibility to} – ADH1C**
  + **{Parkinson disease 12} – PARK12**

MySql data upload/import:

* At first the ‘load data local infile’ command was not permitted, and attempts at using a python script to upload the data into the MySql database were denied as well. Staff from the *Systems Support* department were able to clarify that the command can be enabled client-side by including the option ‘--local-infile’ with the mysql start-up command.
* Certain records from the Parkinson's Phenotype table had identical values for their primary key, which was not known when the data was first obtained, and as a result these records were not imported into the table the first time around, since their primary key was not unique. To get around this, a composite primary key of multiple candidate keys had to be created and so that all records could be inserted.
* The ‘Annotations’ field for the Parkinson's Protein table contains summaries of the gene’s function, which can be more than 150 words in length. Consequently, any query that returns the ‘Annotations’ column from the Parkinson's Protein table will distort the output of the table-format.

**Methodology**

Databasedata:

* All database data pertaining to proteins (ID mapping, interaction scores, annotations, etc) were obtained from **STRING/Uniprot** API’s using the HTTP request method *urlopen* (POST), from the python module ***urllib.request***
  + Any data returned as a json file was processed using the ***json*** pythonpackage
* Once all data had been obtained (including both Parkinson’s phenotype and protein data), python code was used to process and extract all the desired data fields from each file.
* In order to create files that would be compatible with mysql ‘load data’ statements, processed data were output to a final tab-separated file that contained each field in the appropriate column for each record. This process was carried out for 4 different tsv files, corresponding to the 4 different data categories described in the ‘Data’ section.

MySql data:

Loki server upload:

* + - Each .tsv file was transferred to the home directory of the user *ectroudt* on the *loki.ist.unomaha.edu* server using the linux ‘scp’ program.

Mysql commands:

* ‘create table’ was used to create 4 different tables for each of the desired data categories.
* ‘load data local infile’ was used to import each .tsv file straight into its corresponding sql table
* ‘alter table’ was used to modify column names, change column values, and add primary keys

PHP code:

* All commands and code written in PHP were heavily based off of the ‘Assignment 5 tutorial’ which demonstrated how to use PHP:

* Connect to the mysql database on the loki server - ‘mysql\_connect()’
* Select database to use - ‘mysql\_select\_db()’
* execute and retrieve the query results - ‘mysql\_query()’ & ‘mysql\_fetch\_array’
* display it to a public-facing web-page - <http://loki.ist.unomaha.edu/~username/>’php program’
  + (running ‘loki-publish-html’) from bash in loki allowed public access to the database link

**Planned vs Actual Implementation**

Planning the implementation of the project and the creation of the database was a speculative process that turned out to be completely wrong in a lot ways, in large part because so many of the project’s aspects had never been encountered before. Ultimately, the actual timeline differed in a lot of ways from the initial planned schedule of the project:

Assumptions of planned timeline:

* Data acquisition would proceed in a week-by-week manner, with any errors being worked through in a way that wouldn’t be time-consuming or problematic (ie API’s would be easy to learn how to use and wouldn’t take much time)
* Mapping of data between databases was a simple process that involved just looking at the different ID’s and assuming any differences would be minor
* Mysql database could be constructed in a matter of one-two days because all data would already being processed and refined before hand

Actual timeline:

* Overall, the use of APIs was the most efficient part of the project’s implementation, and even though problems were encountered, troubleshooting of the code was not a major hindrance. Both **STRING** and **Uniprot** have APIs that are straightforward with helpful tutorials.
* One of the biggest unforeseen issues of the project’s implementation revolved around trying to correctly map different IDs between the different databases. While the use of the APIs was easy and intuitive, providing the correct data as input was the exact opposite. This was especially true for **Uniprot**, which makes basic distinctions between a *STRING ID* (9606.ENSP00000340278) and an *Ensembl Protein ID* (ENSP00000340278), which while simple, took additional time to figure out, and as a result, this portion of the project was much more time-consuming.
* The actual construction of the mysql database was smooth, but the initial process of learning how to enable certain commands, and revising the design of the schema based on data issues caused this phase to also become drawn out, and ultimately the completion of the database itself finished much later than anticipated.

**References**

Rakshit, H., Rathi, N., & Roy, D. (2014). Construction and analysis of the protein-protein interaction networks based on gene expression profiles of Parkinson's disease. *PLoS One*, *9*(8), e103047.

Rezaei-Tavirani, M., Zamanian-Azodi, M., Rajabi, S., Masoudi-Nejad, A., Rostami-Nejad, M., & Rahmatirad, S. (2016). Protein Clustering and Interactome Analysis in Parkinson and Alzheimer's Diseases. *Archives of Iranian Medicine (AIM)*, *19*(2).

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., ... & Duesbury, M. (2013). The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic acids research*, *42*(D1), D358-D363.

Orchard, S., Kerrien, S., Abbani, S., Aranda, B., Bhate, J., Bidwell, S., … Hermjakob, H. (2012). Protein Interaction Data Curation - The International Molecular Exchange Consortium (IMEx). *Nature Methods*, *9*(4), 345–350. <http://doi.org/10.1038/nmeth.1931>

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., ... & Jensen, L. J. (2016). The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic acids research*, gkw937.

Porras, P., Duesbury, M., Fabregat, A., Ueffing, M., Orchard, S., Gloeckner, C. J., & Hermjakob, H. (2015). A visual review of the interactome of LRRK2: Using deep‐curated molecular interaction data to represent biology. *Proteomics*, *15*(8), 1390-1404.

Manzoni, C., Denny, P., Lovering, R. C., & Lewis, P. A. (2015). Computational analysis of the LRRK2 interactome. *PeerJ*, *3*, e778.

Meixner, A., Boldt, K., Van Troys, M., Askenazi, M., Gloeckner, C. J., Bauer, M., … Ueffing, M. (2011). A QUICK Screen for Lrrk2 Interaction Partners – Leucine-rich Repeat Kinase 2 is Involved in Actin Cytoskeleton Dynamics. *Molecular & Cellular Proteomics : MCP*, *10*(1), M110.001172. <http://doi.org/10.1074/mcp.M110.001172>

Menon, R., & Farina, C. (2011). Shared molecular and functional frameworks among five complex human disorders: a comparative study on interactomes linked to susceptibility genes. *PloS one*, *6*(4), e18660.