**Project I**

**Hetionet**

**Tal Bachar**

**Farhan Chowdhury**

**CSCI 49371**

**Prof. Lei Xie**

**4.4.2020**

**Rational of Database Selection**

For the first part of the project, we chose to use MongoDB as our data base. We chose MongoDB for several reasons:

1. Prior Knowledge: We felt comfortable working with MongoDB since we had sufficient knowledge and some extent of experience working with this system.
2. Helper Tools: The fact that you can use Compass as a helper tool to “visualize” your DB made it easier to work with. Also, it helped understanding the data given to us (.tsv files)
3. Python Compatibility: MongoDB works very smoothly with Python, which made it easier for us to query the data and print results with ease.

**Cassandra**: for the second part.

1. As we are designing the database based on multiple columns, it adds a more powerful “column family” data model.
2. Cassandra is known to be It is scalable, fault-tolerant, and consistent.
3. Data storage and distribution is easy and flexible.
4. Supports ACID properties
5. It can perform fast writes and can store bulk amount of data, without sacrificing the read efficiency.

**Design Diagram: Part 1**

**Design Diagram: Part 2**

**Potential Improvements**

1. Adding more functions such as: look up specific body part and get back disease which effect that specific part.
2. Instead of the program just print tens or sometimes hundreds of lines with all data, maybe build a menu for what the user needs specifically. If the user only wants to see the genes that cause a disease, there is not point of showing him all the drugs the can treat a disease.
3. Use cloud service – This project was all done locally on our personal computers, which means if the user has a slow computer, querying might take a while. If the data was all stored on a cloud with strong servers to support it, I believe it would make querying faster and more efficient.