Milestone 4 – Gene Ontology and Rosetta Relation

Overview:

The objective of this milestone is for you to retrieve protein function from GO (gene ontology). Infer additional protein function from the Rosetta relation. When sufficient protein function is already known among the three proteins involved in a Rosetta relationship, validate the Rosetta methodology by comparing computed function with that already found in GO.

Deviation from original scope of the project.

1) We are not asking you to collect additional proteins from external data sources.
2) Per class discussion, GO has been mirrored here at U.T. Since GO is MySQL based you will need to load an additional JBDC driver.
3) Since there are many GO terms associated with each gene product. (i.e. many functional annotations for each protein) in lieu of an attribute importedFunction in the protein table you will need to replace the attribute with a table ImportedFunction which forms an identifying aggregation with the protein table.

Specific Instructions:

1) Update your Rose model, update your database using the SQL script provided
2) Load drivers and test GO/MySQL database connectivity from your Java environment.
3) Write a Java method that, given a gene product symbol, fetches from GO all the GO terms associated with that symbol, and its ancestral (up the ontology) terms. Write these terms to your Postgres database, in effect annotating the proteins in your database with their function.
4) Write a Java method (or a SQL query) that computes the validity of the Rosetta concept.

Turn in your code and the results of step 4.

Details for step 4

For each Rosetta relationship stored in your database, if more than one of the three proteins already has GO annotations, **compute the intersection of the set of annotations among the annotated proteins**, i.e. if P1 and P2 are annotated with GO terms, find the terms they have in common. If P1, P2 and Pf are all annotated, **find the terms in common among all three**. **Compute the number of terms in common and divide it by the total number of distinct terms associated with the proteins being compared**. This number represents the "overlap" among the proteins. Print out a table of the amount of overlap for each Rosetta relation.