Ryan Verdon

3/10/12

Validation Framework

My thesis involves examining all of the distributed databases to find the best match for bioinformatics data. Not only for storing data but for processing the data for useful features or information by using several well known bioinformatics algorithms like BLAST. From this I see three essential parts to any validation framework that I use.

Part 1: Which distributed database

The first part will consist of a logical proof of why the chosen distributed database is a good choice for storing biological data. This will consist of a cost/benefit analysis and some logical reasoning of why the choice is better than others. The goal of this validation is to prove the chosen database was well thought out and not picked randomly.

Part 2: Query results

The second part will entail comparing query results from the chosen database setup to commonly used biological databases such as the one found on nih.gov and flybase.org. Things to examine are query time, how detailed queries can get, performance effects of large queries, etc. This section also includes examining some issues with distributed databases. For example, inserting into some distributed databases followed by a query right after has a chance to return old results. Basically this boils down to examining the effects of losing the ACID consistency model and moving towards a BASE model and seeing what affect that would have on the end user if any.

Part 3: Algorithm tests

The last section will examine how fast bioinformatics algorithms perform on a distributed database. There are a variety of common algorithms but at a minimum I want to fully examine BLAST. BLAST is an extremely common and important biological algorithm to compare a sequence of amino acids/proteins to a library of amino acids/proteins. In the end any algorithms I look at will be comparing run times on my system to runtimes on common databases.