Contribution ID : **16** Type : **Poster**

A parallel data mining application for Gene Ontology term prediction

Tuesday, 12 February 2008 16:00 (0:00)

1. Short overview

Protein classification is one of the most commonly discussed problems in bioinformatics. One of the latest tools for protein function annotation is the Gene Ontology (GO) project which provides a controlled vocabulary to describe gene and gene product attributes in organisms. Although there are several cases of automated annotation, the bulk of the annotation process is performed by human curators. We present a parrallel algorithm for GO term prediction, deployed over the EGEE grid environment.

2. Analysis

Gene ontology can be thought of as a database of expert-based terms. The application presented utilizes the motifs that exist in already annotated protein sequences in order to model the corresponding GO terms. The input data set is created in a semi-automatic way, using the unique (UNIPROT) code of each protein and the InterProScan tool so that all available sequence databases (such as PRODOM, PFAM etc) will be taken under consideration. For each GO term that appears in the original protein set, a new training set is created, which contains all the protein sequences that have been annotated with the specific GO term. Based on the motifs present in the new data sets, a finite state automaton model is created for each GO term. In order to predict the annotation of an unknown protein, its motif sequence is run through each GO model thus producing similarity scores for every term. Results have shown that the algorithm is both efficient and accurate in predicting the correct GO term.

3. Impact

The methodology has been implemented so that it can be used both as a standalone or as a grid-based application. The algorithm however is by design an embarassingly parallel one allowing for multiple models to be trained simultaneously, thus making the Grid the ideal environment for execution. In fact, it has been shown experimentally that the time to process the entire dataset on a single processor is prohibitively long. In an MPI-enabled application the utilization of the clusters available over the Grid provides a significant reduction of the processing time. The Grid also enables the seamless integration of the training process with the actual model evaluation, by allowing the concurrent retraining of GO models from different input sources or experts and the use of the existing ones.

4. Conclusions / Future plans

The initial dataset is stored and replicated as a single compressed file on multiple storage elements (SEs). The application was executed on available clusters using from 4 to 32 processors in different experiment configurations. In all cases a significant speedup was observed. Overall, the utilization of the Grid as the application platform has provided both a reduction in processing time and a seamless environment for running simultaneously different experiments.

Provide a set of generic keywords that define your contribution (e.g. Data Management, Workflows, High Energy Physics)

Bioinformatics, Protein Classification, Data Mining, Parallel Algorithms, Gene Ontology

URL for further information:

If demonstration is requested please explain what visual or interactive aspects of the contribution necessitate a demonstration rather than a presentation or poster?

Primary author(s) : Mr. PSOMOPOULOS, Fotis (Aristotle University of Thessaloniki)

Co-author(s) : Mr. GKEKAS, Christos (Aristotle University of Thessaloniki); Prof. MITKAS, Pericles (Aristotle University of Thessaloniki)

Presenter(s) : Mr. PSOMOPOULOS, Fotis (Aristotle University of Thessaloniki)

Session Classification : Posters

Track Classification : Scientific Results Obtained Using Grid Technology