Contribution ID: **121** Type: **Oral**

High-throughput GRID application for Life Sciences: The Gene Analogue Finder

Tuesday, 12 February 2008 14:40 (20 minutes)

The algorithm is a very high data and data-access intensive application. The results of the functional analogous search demonstrates that the information contained by the GO is adequate to run such analysis using the gene production description. For example most of the homologous gene products of most of the model organisms were assigned as functional analogues, although the annotations were done independently. This result also proofs that the algorithm assigns functional analogues in the right way. More important, the algorithm finds significant functional analogous gene products within the same or within different organisms, also non-model organisms, which have such a low sequence similarity so that with conventional methods those assignments would not have been found.
Functional analogous associations is a very important information for scientists in the laboratory which are able to find new information and hints about the functionality of the gene product they are working on.

3. Impact

The GO and GOA repositories are updated in a monthly frequency improving the annotation quality but also increasing the number of annotated gene products.
The results show that most of the gene products from non-model organisms are poorly annotated and therefore were not considered within this search or produced low level informatin. For that reason the algorithm is highly dependent on new releases of the GO and GOA and the functional analogous search needs to be updated as frequent as possible. Only by using the GRID technology we are able to fulfill this need and are able to offer the best results to the scientific community by recalculating the whole search results using each new monthly release of GO and GOA.

4. Conclusions / Future plans

The algorithm is a very high data and data-access intensive application. To avoid the problem of concurrent accesses to the data, the system temporally distributes both the analysis tool and the data on WNs where the tool has to operate. The jobs were distributed over the EGEE grid infrastructure within the VO biomed using about 300 WNs. The input data is in the size of 600MB and the results in the order of 2GB. The process was terminated within a day instead of about 60 days using one CPU.

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

bioinformatics, life science, temporal data distribution,

1. Short overview

Up to now, researchers have compared genes looking at their sequence similarity. However the correlation "sequence –function"is only partially applicable. Descriptive annotations, such the one provided by the Gene Ontology (GO) and its associations with the gene products (GOA), offer information for a way of comparing genes according to their functional description.
The application consists of an algorithm that uses the data of GO and GOA to find functional analogous gene products, i.e. gene

Primary author: DE SARIO, Giulia (Istituto di Tecnologie Biomediche, CNR)

Co-authors: Dr GISEL, Andreas (Istituto di Tecnologie Biomediche, CNR); TULIPANO, Angelica (INFN Bari); DONVITO, Giacinto (INFN Bari); Prof. MAGGI, Giorgio (INFN Bari)

Presenter: DE SARIO, Giulia (Istituto di Tecnologie Biomediche, CNR)

Session Classification: Life Sciences

Track Classification: Scientific Results Obtained Using Grid Technology