Contribution ID: **154** Type: **Oral**

GRID distribution supporting chaotic map clustering on large mixed microarray data sets

Tuesday, February 12, 2008 12:00 PM (20 minutes)

To find correlation between genes within different experiments, clustering is a good and challenging analysis method for data sets of such size and complexity. We have chosen an unsupervised hierarchical clustering algorithm based on the cooperative behaviour of an inhomogeneous lattice of coupled chaotic maps, the Chaotic Map Clustering.
Analyzing data sets of 587 samples we were able to retrieve stable groups of genes. Using the biological knowledge of the gene ontology, we could show, applying a Fisher exact test, that each of the clusters have a set of over-represented functionalities and in most of the cases also clearly different functionalities from cluster to cluster.
In order to evaluate the vast number of clusters found by this process we use a cluster validation method based on resampling subsets of the data under investigation are constructed randomly, and the cluster algorithm is applied to each subset.
Measures of sensitivity and of positive predictive value are pro

3. Impact

The clustering of each resampled subset is a very time-consuming process and it is not possible to retrieve the results within a reasonable time using one CPU.
To validate the clusters by resampling, only a distribution of the task over several processing units will solve the problem of processing time. Since the task can be easily splitted into several smaller, independent sub-task we chose the GRID infrastructure to distribute the calculation.
After performing the initial clustering and calculating of the resampled matrices on a single machine, each resampled matrix was clustered on a different WN. The clustering of one matrix takes about 2 hours and therefore a resampling validation with 100 matrices about 200 hours, or 8days. Using the GRID, the whole set of the 100 resampled matrices were clustered in 4 hours instead of about 8 days. The improvement in processing time allows the user to increase the number of resampled matrices and therefore improve the precision of the positiv

4. Conclusions / Future plans

The whole set of the 100 resampled matrices were distributed over 100 WN of the EGEE infrastructure within the VO biomed and processed totally in parallel, clustering those matrices in a time slightly longer than a clustering of one matrix. The process used mainly CPU since the output data file are small. The only problem we were confronted with is the size of the RAM usage. The clustering process occupies about 1.5 GB of the WN's RAM which in certain cases lead to the failure of the job which t

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

bioinformatics, life science, clustering, cluster validation

1. Short overview

Microarray data are a rich source of information because they contain the expression values of thousands of genes and in addition, especially in public repositories, hundreds of experiments with the same array design are available. Comparing expression levels over a wide range of experiments can reveal new and valuable information about behaviours of genes. Furthermore, because of the vast amount of experiments available, technical errors can be filtered out.

Primary author: TULIPANO, Angelica (INFN Bari)

Co-authors: Dr GISEL, Andreas (Istituto di Tecnologie Biomediche, CNR); Dr MARANGI, Carmela (Istituto per le Applicazioni del Calcolo, CNR); DONVITO, Giacinto (INFN Bari); Prof. MAGGI, Giorgio (INFN Bari); DE SARIO, Giulia (Istituto di Tecnologie Biomediche, CNR); Prof. ANGELINI, Leonardo (INFN Bari)

Presenter: Dr GISEL, Andreas (Istituto di Tecnologie Biomediche, CNR)

Session Classification: Life Sciences

Track Classification: Scientific Results Obtained Using Grid Technology