Contribution ID: **169** Type: **oral presentation**

CSTGrid: a whole genome comparison tool for the identification of coding and non-coding conserved sequences.

Thursday 10 May 2007 14:20 (20 minutes)

Describe the scientific/technical community and the scientific/technical activity using (planning to use) the EGEE infrastructure. A high-level description is needed (neither a detailed specialist report nor a list of references).

One of the major task of genomic analysis is the identification of genes and regulatory elements of gene expression. Among the different methods used to solve this task, comparative analysis have shown their strength and reliability.
CSTminer is a comparative analysis tool developed to compare two or more sequences to identify conserved sequences and to classify them as coding (likely genes) or non-coding (potential regulatory regions) upon the computation of a coding potential score (CPS).

Report on the experience (or the proposed activity). It would be very important to mention key services which are essential for the success of your activity on the EGEE infrastructure.

In order to have a more efficient jobs we grouped together a lot of comparison (about 1000), in this way we can reduce the overhead do to files transfer and environment setting-up.
We have developed a job submission tool (JST) that allows the submission of large number of jobs in an almost unattended way. It is based on the concept of "task"to be executed. Only if all steps are correctly executed, the status of that particular task on the central DB is updated to "Done". In this way the central DB provides a monitoring of the task execution and no manual intervention is required to manage the resubmission of the failed tasks.
The most important services needed in order to run this application is WMS in order to submit jobs optimizing the utilization of each resource available.
We also exploit successfully the data transport services using gsiftp in order to copy files both for input and output.

With a forward look to future evolution, discuss the issues you have encountered (or that you expect) in using the EGEE infrastructure. Wherever possible, point out the experience limitations (both in terms of existing services or missing functionality)

We can foresee to build a web interface in which the user can easily submit its own comparison with between a given genome and a new one (provided by the user itself).
In this way this application can be established as a service for the user even if not expert of the EGEE infrastructure.

Describe the added value of the Grid for the scientific/technical activity you (plan to) do on the Grid. This should include the scale of the activity and of the potential user community and the relevance for other scientific or business applications

The availability of a collection of evolutionary conserved sequences among different organisms could be of great importance for scientific community for the functional annotation of genomes and for the identification of regulatory elements.
Given the reliability and sensitivity of CSTminer software it would be very useful to perform genome wide comparisons of several genomes. Unfortunately given the complexity of CSTminer alignment step and the size of the genomes (3 and 2.6 Gbp for Human and Mouse genomes respectively) it would be impossible to carry out such analysis even on very powerful servers.
This is the reason why we exploited the possibility to use the EGEE infrastructure and the power of a Grid approach.

Primary authors: Dr MIGNONE, Flavio (Università di Milano); Dr DONVITO, Giacinto (UNIVERSITà DEGLI STUDI DI BARI); Dr VICARIO, Saverio (CNR - ITB Bari + EEB - Yale University)

Co-authors: Dr TURI, Antonio (University of Salento); Dr JEAN, Atul (INFN-BARI + Politecnico di Bari); Prof. SACCONE, Cecilia (CNR - ITB Bari + Università di Bari); Dr GRILLO, Giorgio (CNR-ITB); Prof. MAGGI, Giorgio Pietro (INFN-BARI + Politecnico di Bari); Prof. ALOISIO, Giovanni (SPACI Consortium and University of Salento, Italy); Prof. PESOLE, Graziano (Università di Bari); Dr MIRTO, Maria (SPACI Consortium and University of Salento, Italy); Dr LIUNI, Sabino (CNR-ITB); Dr MY, Salvatore (INFN-BARI + Politecnico di Bari)

Presenter: Dr MIGNONE, Flavio (Università di Milano)

Session Classification: Experience with application domains