Contribution ID: **65** Type: **poster**

Protein Clustering on the EGEE

Wednesday, 9 May 2007 17:30 (20 minutes)

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

As protein databases are growing up day after day, the clustering process on
interesting datasets in a single machine is not feasible due to memory constrains. A
Grid environment allows an adaptive database distribution in order to optimize its
overall analysis. The complexity of the workflow inherent to "CD-HIT"needs a robust
framework able to handle it. In addition, this framework may be successfully used in
other applications which result in a same type of workflow.

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).

The target of this activity is the Bioinformatic scientific community, and in
particular, those members who use a tool called "CD-HIT"which performs protein
clustering on a protein sequence database. This consists in removing redundant
sequences at a given sequence similarity level and generating a new database with the
representatives only. This activity was proposed by CNIO (Spanish National Cancer
Research Centre) and started in the context of the BioGridNet Program
(www.biogridnet.org).

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.

For porting this application onto the Grid, we used the GridWay Metascheduler
(http://www.gridway.org/) and relied on the execution services offered by BIOMED VO.
GridWay is being used as a Resource Broker because its workflow management
capabilities and interoperability have been proved to be very valuable. On the other
hand, we are working with its Open Grid Forum

(http://www.ogf.org/) DRMAA Standard
(http://drmaa.org/) implementation (both C and JAVA bindings).
Finally, GridWay has
also been chosen because it allows interoperability with EGEE and
GRIDIMadrid
(http://www.gridimadrid.org/), which is a Globus-based regional
testbed.

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 plan to start with production input data proposed by CNIO. In
particular, with the
analysis of various meta-genomes, starting with the first
published one from Sargasso
Sea.

Primary authors: Dr VALENCIA HERRERA, Alfonso (CNIO (Spain)); Dr HUEDO CUESTA, Eduardo (Universidad Complutense de Madrid (Spain)); Dr MARTIN LLORENTE, Ignacio (Universidad Complutense de Madrid (Spain)); Mr VAZQUEZ-POLETTI, Jose Luis (Universidad Complutense de Madrid (Spain)); Dr FERNANDEZ GONZALEZ, Jose Maria (CNIO (Spain)); Dr SANTIAGO MONTERO, Ruben (Universidad Complutense de Madrid (Spain))

Presenter: Mr VAZQUEZ-POLETTI, Jose Luis (Universidad Complutense de Madrid (Spain))

Session Classification: Poster and Demo Session