Contribution ID: **102** Type: **Oral**

Setup and usage of an e-BioScience infrastructure for high throughput DNA sequence analysis

Tuesday, 13 April 2010 14:40 (20 minutes)

Current DNA sequencers produce a large amount of data. Because the amount of data is growing and the computation time for analysis is increasing, we initiated a pilot to run applications on the Dutch Grid (Big Grid, part of EGEE). We used the software platform that was developed for medical imaging in the VL-e project (e-BioInfra), and applied it to DNA sequence analysis. With the knowledge gained in this pilot, we now use this platform routinely for computational intensive analyses.

Detailed analysis

We ported sequence analysis applications to the grid with the e-BioInfra platform. The analysis tools (executables) are wrapped as workflow components with the Generic Application Service Wrapper (GASW) service. The workflows are described in the Scufl language of the Taverna workbench and executed on the grid with the MOTEUR workflow engine. Computation and data management are based on the gLite middleware. The analysis tools, GASW descriptions, workflow descriptions and data are located on grid storage. For the analysis of sequence data the user transfers files from local resources to grid storage with the Virtual Resource Browser (VBrowser). Thereafter, the user can start workflows that are enacted on the Dutch Grid by the MOTEUR Web Service from the VBrowser. The user monitors workflows and grid jobs from the same front-end, using specialized grid services and web interfaces. When the analysis is complete, the user recovers or browses results with the VBrowser.

Conclusions and Future Work

A generalized layer on top of the gLite middleware (MOTEUR, GASW) can help end-users to interact with grid resources from a user friendly interface (VBrowser). We observed that besides the hard- and software tools, in-house expertise about workflows and grid infrastructures is needed, because porting applications to the grid is not a trivial task. In the future we need to improve data security and error handling, which are known challenges for the deployment of grids in practice.

Impact

The platform has been used for two different sequencing projects. In a pilot study, a metagenomics study of viruses in human samples, we identified the experimental sequences using BLAST, a popular sequence alignment program in the bioinformatics community. For the second project, identification of alternative splice variants, we ported in-house developed Perl and R scripts to the grid to perform all versus all comparisons between 400,000 sequences per experiment. This has become a routine tool in the Bioinformatics Laboratory, facilitating the analysis of sequence data.

We expect that the number of applications and the throughput of the DNA sequencers will increase fast in the next years. The workflow technology can help us to reuse earlier developed software components and we need grid infrastructure to cope with the increasing data flow and analysis. The workflows are available for members of the same Virtual Organization and via myExperiment.org

Keywords

Next generation sequencing, e-BioInfra, Bioinformatics, Grid workflows, Porting, MOTEUR, VBrowser

URL for further information

http://www.bioinformaticslaboratory.nl/ (EBioScience infrastructure)

Primary authors: Mrs LUYF, Angela CM (Academic Medical Center); Mrs VAN SCHAIK, Barbera DC (Academic Medical Center); Dr OLABARRIAGA, Silvia D (Academic Medical Center)

Co-authors: Prof. VAN KAMPEN, Antoine HC (Academic Medical Center); Prof. BAAS, Frank (Academic Medical Center); Mrs RITZ, Katja (Academic Medical Center); Mr DE VRIES, Michel (Academic Medical Center)

Presenter: Mrs VAN SCHAIK, Barbera DC (Academic Medical Center)

Session Classification: Bioinformatics

Track Classification: End-user environments, scientific gateways and portal technologies