Contribution ID: **104** Type: **not specified**

Challenges in data management and running DNA sequencing experiments on grid

Tuesday, 27 March 2012 11:30 (30 minutes)

Description of the Work

Data analysis involves several steps. Data has to be transferred to grid storage, workflows are set up for the analysis, the workflow is executed on a grid infrastructure, and once the results are complete they need to be transferred back and shared. We encountered several challenges while performing these steps, defined requirements and addressed some of them.
Before the analysis can be performed the raw data that is stored on a server from the sequencing facility needs to be transferred to some grid storage. We have installed software on this server to be able to transfer the data directly to grid storage, which minimizes data transfer time.
The user who analyses the data can choose to run an existing workflow with a command-line client (Moteur), desktop client (VBrowser with Moteur plugin) or web-based interface (e-BioInfra gateway). The progress of the workflow runs can be examined from the e-BioInfra gateway, which gives status updates and provides error information in case jobs or workflows fail. Job failures occur because of too large input files, too memory-intensive jobs, application failures and grid related errors, e.g. data transfer errors and job scheduling. Improvements to the e-BioInfra monitoring have been made to detect errors and distinguish their type. We have also implemented a mechanism to resume analysis for data sets that failed by keeping track of inputs and produced results for specific studies in a database. Currently improvements to the workflow design are made to solve this on the workflow level.
Finally, when the analysis is complete, results are transferred to a server where they can be shared with users without a grid certificate. At the moment we use a wiki to gather and to share information about each experiment and provide links to the results.

Conclusions

The e-BioInfra is used daily for the analysis of high throughput sequence data on the Dutch grid. It enables bioinformaticians and biomedical users to scale up the analysis and be flexible in their choice for analysis tools to use in a workflow. The data volume raises challenges for data transfer to and from the grid, as well as between grid storage and computing nodes. The run times of individual jobs are relatively long (in the order of hours) and it is important to reduce failures to avoid an increase of the total run time of a workflow due to resubmission of failed jobs. The infrastructure has been improved in several aspects for the end-user. Users with different profiles can make use of grid computing for their research with different interfaces. The error reporting has been improved and helps to identify problems at run time and set priorities for solving remaining issues.

Impact

The e-BioInfra provides generic services, such as workflow submission and monitoring services, which are used on a daily basis via easy to use interfaces by a growing number of users. At the end of 2008 we started a pilot for using this infrastructure for DNA sequence analysis. The number of implemented workflow components from existing and newly developed software for DNA sequence analysis has grown and the components are now used by bioinformatics and biomedical researchers for virus discovery, (partial) human genome resequencing, whole transcriptome sequencing and small RNA expression profiling. New discoveries were done in these areas and are verified with complementary laboratory experiments.
Using workflow technology in combination with grid computing changed the way researchers setup their experiments and expanded the possibilities for automated and large scale data analyses requiring resources exceeding local clusters. The separation of the analysis description in a workflow from the input parameters enables automatic data sweeps without additional programming. Furthermore, parallel analysis of datasets with different parameters is feasible and experiments can be repeated with different algorithms by simple replacement of workflow components. In this way larger studies can be performed more easily using a single interface.

URL

http://www.bioinformaticslaboratory.nl/

Overview (For the conference guide)

Modern DNA sequencing machines produce data in the range of 1-100GBytes per sample and with ongoing technological developments this amount is rapidly increasing. The majority of experiments involve resequencing of human genomes and exomes to find genomic regions that can be associated with disease. In the bioinformatics field the development of analysis software for DNA sequencing experiments progresses rapidly. There are many analysis tools freely available, e.g. for sequence alignment, quality control and variant detection, and frequently new tools are developed. We use workflow technology to allow easy incorporation of such software in our data analysis pipelines, and analyzing multiple data sets at once. Since the end of 2008 we use our framework, called e-BioInfra, for various DNA experiments on the Dutch grid. Here we will present our current procedure of analyzing DNA experiments, comment on the experiences and focus on improvements that were needed for analyzing genomics data.

Primary author: Ms VAN SCHAIK, Barbera (Academic Medical Center)

Co-authors: Dr JONGEJAN, Aldo (Academic Medical Center); Prof. VAN KAMPEN, Antoine (Academic Medical Center); Mr WILLEMSEN, Marcel (Academic Medical Center); Mr SANTCROOS, Mark (Academic Medical Center); Dr OLABARRIAGA, Silvia (Academic Medical Center); Dr KORKHOV, Vladimir (Academic Medical Center)

Presenter: Ms VAN SCHAIK, Barbera (Academic Medical Center)

Session Classification: Community-tailored Services