Contribution ID: **23** Type: **Oral**

A Grid Implementation For Genetic Linkage Analysis

Tuesday 13 April 2010 14:00 (20 minutes)

The Genetic Linkage Analysis of SNP markers aims to discover the genetic correlation in monogenic diseases by following their inheritance in families through the generations. The computational cost and memory requirements of the major algorithms in literature make large data sets very hard to be analyzed on a single CPU. The work here presented is a Grid implementation of a data pipeline application for Linkage Analysis, a web based tool for Grid submission, monitoring and retrieval of Linkage challenges with large pedigrees and big markers datasets derived from genotyping chips up to 1M SNPs.

Detailed analysis

Enabling the use of the EGEE Grid Infrastructure for the execution of linkage analysis on very large SNPs datasets is achieved through the implementation of a web application conceived in 3 main layers: the User Interface is designed to visually make up the pipeline for the linkage process, to set up the linkage challenges and the related otpions, and to monitor the workflow execution masking the complexity of low level interactions with the Grid middleware; the Application Layer executes data pre-processing operations like the retrieval of the SNPs data from the genotype database and the opportune formatting and splitting of the computational pipeline into Grid compliant jobs; the Submission layer, built on top of the grid middleware, monitors each single grid process and ensures the success of its computation by managing the automatic resubmission of failed jobs, bypassing certain Grid limitations such as the size of the input/output sandbox, while simultaneously optimizing the Grid storage and transfer overheads. When all tasks are computed, the results are retrieved, merged, showed to the user and made available for downloading through the web interface.

Conclusions and Future Work

This Linkage Application enables the exploitation of the Grid infrastructure to set up and run analysis without a dedicated cluster and without the need of mastering complex informatics skills, achieving ease of use through the implementation of a dynamic and visually interactive User Interface. The data splitting approach has been tested as a valid approximation and the distributed implementation is mostly useful in high-end challenges, where Grid overheads are negligible with respect to parallelization benefits, making very large analyses accessible even without dedicated hardware.

Impact

The Application is aimed at biologists with minor informatics skills, all the user interactions being implemented through an intuitive and interactive web interface where visual techniques are adopted to set up the system parameters and options. The system was tested with real case analyses to evaluate both the efficiency on computational time and the functionality of the data splitting approach, comparing its performances with a single 2 GHz CPU workstation and with a cluster composed by 280 CPU cores. The analysis, run using the Merlin software, involved different size datasets (10k - 1M of SNPs) and logged all phases duration of jobs life span. Results show that distributed analysis pipelines with big datasets can achieve a speedup up to more than 70x compared to a mid range dual-core 2 GHz CPU execution; the average performance shows a scaling trend comparable with the cluster due to the similar workload distribution technique adopted (not MPI). With

the increase of the challenges size in terms of number of jobs, the variability in the total lasting time increases, due to the higher job resubmission rate, nevertheless the benefit in the distribuited approach gets even higher.

Keywords

grid, distributed computing, linkage analysis

URL for further information

www.itb.cnr.it/linkage

Primary authors: Dr CALABRIA, Andrea (National Research Council - Institute of Biomedical Technologies); Dr DI PASQUALE, Davide (National Research Council - Institute of Biomedical Technologies); Dr TROMBETTI, Gabriele (National Research Council - Institute of Biomedical Technologies); Dr MILANESI, Luciano (National Research Council - Institute of Biomedical Technologies); Dr GNOCCHI, Matteo (National Research Council - Institute of Biomedical Technologies); Dr COZZI, Paolo (National Research Council - Institute of Biomedical Technologies)

Presenter: Dr MILANESI, Luciano (National Research Council - Institute of Biomedical Technologies)

Session Classification: Bioinformatics

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