

The collaboration between ITC–irst and ICTP E–GRID Project

Here is described the collaboration between ITC–irst and ICTP E–GRID Project.

ITC–irst(<http://www.itc.it/irst>) investigated the problem of developing a complete setup for predictive molecular profiling and its implementation as a GRID enabled application. Class prediction of high dimensional gene expression data requires access to a complex learning process coupled with its complete validation. Complete validation design (Simon et al 2004) requires two loops of replicated experiments of high computational costs. The inner loop, typically based on cross–validation, is needed for model tuning and feature selection which is replicated at each run of the external loop. In addition to the learning process, the preprocessing of the current training subset of data has to be performed from scratch at each step of the internal loop to avoid selection bias effects. Thus, intensive computational resources and a careful experimental setup are required in order to verify that a prognostic molecular signature is not due to overfitting, but is confirmed by an honest evaluation of the expected accuracy on novel cases.

The E–RFE complete validation setup developed at ITC–irst for Support Vector Machine classifiers was redesigned to obtain BioDCV, a distributed version for virtual GRID facilities. A specific requirement for the BioDCV system was to make it portable on a wide range of computational platforms: from single workstations to local Linux clusters on computational Grids.

To guarantee speed, slim and robust code, and a relational access to data and model descriptions, BioDCV was written in C and interfaced with SQLite (<http://www.sqlite.org>) that supports concurrent access and transactions which are useful in a distributed environment where the learning, tuning and evaluation tasks may be replicated for up to a few millions of models. We recently ported our application on Grid systems, namely the Egrid (<http://www.egrid.it>) computational grids. The Egrid infrastructure is based on Globus/EDG/LCG2 middleware and is integrated as independent virtual organization within the Grid.it, the INFN production grid. The porting requires just two wrappers, one shell script to submit jobs and one C MPI program. Three basic elements of a Grid.it infrastructure are used: Storage Element (SE) – stores user data in the grid and makes it available for subsequent elaboration; Computing Element (CE) – where user tasks are delivered for elaboration: this is usually a front end to several Worker Node (WN) machines – where the grid user programs are actually executed. When the user submits a BioDCV job to the Grid, the GRID middleware looks for the CE and the WNs required to run the parallel program. As soon as the resources (CPUs in WNs) are available, the shell script wrapper is executed on the assigned CE. This script distributes the microarray dataset from the SE to all the involved WNs. It then starts the C MPI wrapper which spawns several instances of the BioDCV program itself. When all BioDCV instances are completed, the wrapper copies all outputs, including model and diagnostic data, from the WNs to the starting SE. Finally, the process outputs are returned allowing the reconstruction of a complete data archive for the study.

The BioDCV analysis tool may run without modifications on Linux workstations, Linux clusters and on Grid systems by using the following elements: SE access, interaction with the PBS/LSF queuing system in the CE, and WNs usage in a parallel fashion through the MPI library.

The application was tested on different microarray datasets to provide classification models and lists of prognostically useful genes. The analysis was replicated with gene expression data from the Wistar Institute and from INT–IFOM. Moreover, the data stored in the final SQLite database have been used to search for molecular cancer subtypes using a new semi–supervised method (Furlanello et al, 2005). Both ITC–irst/MPBA and ICTP/Democritos clusters were used to first test the BioDCV system on these tasks, with up to several hundred thousands SVM models on datasets containing up to 20,000 genes. The BioDCV grid application was then run on the same datasets on the Egrid's virtual organization (V.0.) within the Grid.it production grid. The application (data, classification software) were correctly distributed amongst computing resources in all the cases. The tests showed scalable behaviour of our application for increasing numbers of

CPUs. This is fundamental to take full advantage of the available computing power provided by the Egrid V.O. within the grid facility, i.e. more than 100 cpus on 'best effort basis'. The same classification accuracy was found on the different platforms in the suite of functional genomics studies performed.

A second aim of this collaboration was to install a grid site at ITC–irst. Here five machines were dedicated to the task. One of them was installed with Egrid's supernode live cd while the other ones with Egrid's worker node software. The ITC–irst site is linking with Egrid's testbed and it is configured to accept MPI jobs. Now we are also running BioDCV on this test facility.

AUTHORS: Paoli S [1,2] , Leto A [2] , Zoicas C [2]

[1] ITC–irst, Trento

[2] ICTP E–GRID Project, Trieste