Microarray data infrastructure using Gebbalab based on Alfresco technology

ID - 133

Giuliani Silvia[1], Rossi Simona[2], D'ascia Sergio[3], Rossi Elda[1], Emerson Andrew[1], Fiameni Giuseppe[1], Tagliavini Luca[2], Frangiamone Giuseppe[3], Volinia Stefano[2]

[1]High Performance Systems Division, CINECA, Casalecchio di Reno (Bologna)

[2]DAMA, University of Ferrara, Ferrara

[3]NSI - Nier Soluzioni Informatiche, Castel Maggiore (Bologna)

Motivation

Great progress has been made in recent years in integrating technologies and innovations in computer science with those of the life sciences. However, many activities in biological and especially clinical research still do not have access to the necessary computer technology. Hospitals, for example, often perform outstanding research but lack the bioinformatics tools which could fully exploit the activities carried out. The GeBBALab project is addressing these problems by creating a virtual laboratory' with contributions from both scientific and technological/industrial partners. The project has identified two key areas:

1. Microarray data infrastructure and analysis
2. Integration of patient and clinical data with genomics information

Although the latter objective is of critical importance in health care this is still under discussion. This abstract will therefore concentrate on the GebbaLab (Genetics, Biotecnologie and Bioinformatica Applicata) infrastructure for microarray storage and analysis, also because one of the consortium members already provides a microarray analysis service and so can guide and test its design. We emphasise, however, that the system has been designed to allow the addition of clinical data

Methods

The efficient storage and analysis of microarray data is of considerable interest and there is much activity worldwide. In general most researchers adopt a single workstation approach' for data management and analysing expression data. However this method is rapidly becoming inconvenient for many reasons:

- There is no provision for the systematic recording of experimental information.
- Current PCs are not sufficiently powerful for analysing data.
- Comparison with data from other researchers or public repositories is difficult.

Careful consideration of these points has suggested the following criteria for the design of the microarray infrastructure.

- Users must be given the opportunity to use a wide range of common and user-friendly tools for data entry and for the different platforms available, e.g.

Affymetrix, Agilent, Illumina etc.

- Data should be distributed.
- Data must be recorded in a format which allows interoperability of all the data sources.
- User-friendly portals or clients are required to access resources and powerful computational facilities to process datasets.

To satisfy these criteria the infrastructure was structured into two distinct levels:

1. The data entry and storage level.
2. The application level for running analysis applications.

The system consists of a central node and many satellite nodes, each of which with its own data store, potentially virtualized. The system has been designed in a modular way in order to work even in case of unavailability of the central node. In fact in our schema central' merely indicates a central registry for distributed indexing and querying. Data is stored, analyzed and exchanged through a complex architecture build upon Alfresco, an advanced Open Source Enterprise Content Management that provides a common interface and access to distributed data sources.

Alfresco also includes user authentication and various levels of access privileges, thus allowing many degrees of data security and privacy. Our effort have been lead to build upon the Alfresco structure many software modules in order to manipulate MAGE-ML files, extract metadata from MAGE-MLs, to store and index metadata into the repository for querying microarray data according to different search criteria. All the modules are provided through a SOA (Service Oriented Architecture) layer among different web applications that allow users to choose, through a web portal, the appropriate software from those available that transparently invokes algorithms to fetch the data for analysis. High performance servers are available for CPU or memory intensive calculations

Results
We have demonstrated a user-oriented, powerful infrastructure for microarray data management and analysis. It allows the user to enter data and being distributed avoids the limitations of a centralised server. A prototype using Alfresco is already available and microarray researchers are invited to contact the authors if they wish to experiment with the system. Future enhancements to Gebbalab will include analysis applications and, crucially, the possibility of integrating patient data.
Availability: http://www.gebbalab.it
Email: silvia.giuliani@cineca.it