GeneGrid: a workflow system for sequences analysis - (session: Database: Ontology and Integration)

Roberto Specchio, Andrea Caprera, John Hatton, Luciano Milanesi

Istituto di Tecnologie Biomediche, CNR, Segrate (Mi)

Here we present a project concerned with the realization of a new infrastructure for bionformatics computing.
The infrastructure consists in a workflow system, controlled by a job manager, where a pipeline analysis is loaded into a database and executed on a parallel complex. The aim of this system is to perform complex bionformatics analysis, where the single analysis tools are concatenated in an automatized procedure. The system is constituted by the following components:
- Server Host, which has the function of controlling the analysis process, and where the pipeline is loaded into a database.
- Master Host, which has the function of controlling the process. The jobs are first submitted to this host, where they are scheduled and sent to the Execution hosts.
- Execution Hosts are the nodes of the complex where the jobs are executed. When a job is completed, the Master host communicates with the Server host, where the status of the job is updated into the table Job of the database corresponding to the pipeline analysis, and the next jobs in the analysis process (if present) are sent to execution.
- Database Repository, where the information is retrieved from the programs in execution.

We included in our system tools for similarity analysis on nucleotide or protein sequence databases, such as BLAST, PSI-BLAST, and BLAT, several methods implemented for gene prediction and automatic annotations, and analysis of ESTs distribution in tissues and organs.
We are further including in our analysis system, tools for both structure and function comparative analysis. We are also implementing tools for protein interaction similarity analysis in order to complement the analysis based on sequence similarity search programs with procedures for the analysis of molecular interactions fields. We are also integrating this infrastructure into a Virtual Organization, in order to share and interface our resources in a collaborative environment. This will be realized by means of the Globus toolkit, which is a package providing a set of Grid services.