# Biological database access and integration using web services in GRID technology

Merelli I.(1) Landenna M.(2) Milanesi L.(1)

(1) Istituto di Tecnologie Biomediche, Consiglio Nazionale delle Ricerche, Milano (2)Dipartimento di Scienze dell'Informazione, Università degli studi di Milano, Milano

## Motivation

Data integration is a fundamental process in Bioinformatics because the enormous quantity of information available is often difficult to interpret. However, using a high performance platform such as GRID, it will be possible to complete important studies to improve the understanding of the biological process. Through facing this challenge, the importance of creating a data management system that guarantees efficiency on a distributed platform has emerged. This study concerns the definition of an innovative tool for the databases management in GRID technology and the implementation of a concrete case of use in integrating biological data. The core software is a Web Service that allows the execution of SQL query on a series of distributed databases, available on different computer, through the SOAP protocol. In this way, through the client, that can be run from a GRID Computing Element, it is possible to interact with the database. The data extracted from each local database are then integrated by the Web Server and sent to the application that asks for them, optimizing the communication times. Through this software it is possible to perform elaborations that involve data access and integration on GRID easily. This Web Service has been tested during the development of an integration pipeline among two important biological databases as UNIPROT and ENSEMBL in order to coordinate different information about a certain protein sequences. Thanks to this distributed system of data access it has been possible to conduct a systematic GRID analysis of the kinase sequence references in these important databases.

## Methods

The choice of a Web Service in order to interrogate the huge biological database emerges from the necessity of elaborating information making the most of the GRID platform's high performance. The tools that are available at the moment are not designed for performing SQL queries on distributed databases and therefore the solution to use a Web Service seems the most efficient to be implement without interfering with the GRID structure. The necessity of a client performing on GRID nodes, on which only standard libraries are installed, has forced the choice of a JAVA implementation for the whole system. In fact, using JAVA, it is possible to create a real portable solution using few optional packages. Among the different options available for the Web Service's implementation, the use of the TOMCAT container has been chosen, in the combination with the AXIS package, that provide JAVA classes for the SOAP communications. The Web Server picks up the incoming SOAP message from the client that contains information about the database to be used and the SQL query to be performed. The requests are therefore forwarded to the different database nodes connected through the JAVA driver and, according to the query requested, the results are integrated by the Web Service. If the application involves a SELECT procedure the data are memorized in a temporary table, that is analyzed at the end of the calculation to eliminate the replicated information. Using this system a real case of data integration between two important databases like UNIPROT and ENSEMBL, in their BIOMART version, was implemented. Submitting a sequence to each GRID node using a PERL script, it is

possible to check if a match exists between the input and the proteins present in UNIPROT. If positive matches are found the UNIPROT identifier of the corresponding proteins are recorded and the database is scanned for extracting information on them. The results are therefore stored in a special database directly from the GRID nodes through the execution of an INSERT query. The real data integration is then performed interrogating the ENSEMBL database in relation to the sequences found in the UNIPROT database. For all the positive matches the ENSEMBL identifier is recovered and the information concerning the entry are stored in a local database using the Web Services.

## Results

Through this data management system, developed to manage the database interrogation on the GRID platform efficiently, it has been possible to integrate all the information present in the UNIPROT and ENSEMBL databases in reference to a specific protein family sequence. The dataset chosen to perform the pipeline developed contains all the human protein kinase sequence. Thanks to parallelism that has been implemented using different GRID nodes this task has been carried out with particular success. Data management access and integration will become more important within bioinformatics applications because the problem of the understanding biological processes is tightly correlated with the possibility to meaningfully interpret the enormous information flow deriving from the high-through sequencing of new genomes.

**Contact email:** ivan.merelli@itb.cnr.it