Virtual Integration of Biological Databases Through Web Services

S.A. Silveira[1], J.C.D. Lopes[2] , C.H. da Silveira[3] , M. M. Santoro[4], W. Meira Jr[1]
{sabrinas, meira}@dcc.ufmg.br, {jcdlopes , carlos.silveira@gmail.com}, santoro@icb.ufmg.br
[1]Departamento de Ciência da Computação - Universidade Federal de Minas Gerais
[2]Departamento de Química - Universidade Federal de Minas Gerais
[3]Departamento de Matemática e Computação - Universidade Federal de Itajubá
[4]Departamento de Bioquímica-Imunologia – Universidade Federal de Minas Gerais.

Abstract

There are several public databases of biological data available on the Internet. In general, each of them stores a particular type of biological data, such as nucleic acid sequences, primary, secondary and tertiary structures of proteins, compound libraries, therapeutic molecular targets, potential active sites, cellular processes involving metabolism, DNA and protein expression analysis among others. Since such data are on vast and different databases, their integration becomes a real challenge. Biological data integration is a widely discussed issue because researchers need to access and correlate biological information from different repositories.

Two main approaches are found in the literature [1] to address the biological data integration: physical integration and virtual integration. The former approach proposes to create a single but complex data model which will be fed with data from various sources. The latter approach proposes to build a data model in which user has an abstract view of the set of data resources at hand, whatever architecture is used to tie these data together.

A new paradigm able to explore this virtual integration [2, 4, 5] may be through Web Service (WS) systems [3], also forming the so-called “cloud computations”. WS can be seen as a collection of methods that form a Web API (Application Programming Interface), designed to support a network accessible interface for interoperability among distributed applications. The proposal of this work is to verify the possibility of use WS to promote virtual integration of biological databases. The idea is to build a set of Web Services that will have well-defined methods capable of manipulating different sources of data in an integrated way, so that programmers may access data through its applications without knowing details of databases and of the implementation behind WS. The Web APIs could form an abstract standardized layer centered in a middleware, separating technical components such as databases, platform and programming-language specific details from the application client. Scalability, reliability, fault tolerance, load balance and parallelism could also be obtained by replicating this middleware through the Internet, composing mirrored or redundant sites.

With this model in mind, we are implementing a WS prototype using Axis (a Java platform for creating and deploying web services applications that is an implementation of the SOAP protocol) and Tomcat web server, in a first phase involving data from PDB (Protein Data Bank). For our initial tests, we built an experimental relational version of PDB data using the DBLoader tool provided by RCSB PDB. We developed a Java WS method that synchronizes the local version of PDB with official PDB site, as a way to implement coherence between the middleware and the data source. Other sets of methods are under development, such as those responsible for making unified keyword queries on all virtual integrated databases.

References

[1] Lesk, M. A. Database Annotation in Molecular Biology. England: John Wiley and Sons, 2005.
[2] Pillai, S. et al. SOAP-based services provided by the European Bioinformatics Institute. Nucleic Acids Research. Oxford, v. 33, p. 25-28, 2005.
[3] W3C Web Services Architecture, 2004. http://www.w3.org/TR/ws-arch/
[4] Web Services at the EBI. http://www.ebi.ac.uk/Tools/webservices
[5] RCSB PDB WebServices (BETA). http://www.rcsb.org/robohelp_f/#webservices/summary.htmare