BioInView: a Tool for Integration and Visualization of Heterogeneous Bioinformatics Data Sources

ID - 175

Capasso Pasquale[1], Antonio D'acierno[2], Antonio Picariello[1]
[1]DIS, Università degli Studi di Napoli 'Federico II', Napoli
[2]ISA-CNR, Avellino

Motivation

A recent annual review of molecular biology databases lists several hundreds structured and unstructured data banks containing information relative to the genomic domain. Many of these public sources are not really databases, considering the usual computer science meaning of the term, for a number of reasons and mainly because they do not have a separate schema containing meta-data, and clear data constraints (or, if they do, they are not publicly accessible). In addition, some of these archives are actually no more than tools, processes, or Internet flat files containing embedded kind of meta-data, with a limited set of services, that need to be accessed through suitable interfaces.

Moreover, these sources have complete autonomy, continually extending their coverage, are poorly integrated and, more important, very difficult to use together. As a result, the complexity of the usual retrieval tasks that the biologists need to perform, is strongly associated to the heterogeneity of these tools/data sources: thus, the need of an efficient and effective integration system has never been so pressing.

Methods

In this work we present a prototypical integration system, which is able to provide an effective and computable representation of data, together with efficient tools for performing essential querying tasks (such as homology searches, concept browsing, etc.), and - at the same time - for accessing data from different and heterogeneous data banks, making format heterogeneity completely invisible to the final users. A sketch of the system's architecture can be found at the following link: http://wpage.unina.it/pacapass/SystemArchitecture.jpg.

The system's design follows the well-known Mediator-Wrapper paradigm. The user communicates with the system through a visual interface, which provides a guided input mechanism and a result-browsing tool. Queries provided by the user are analyzed by the mediation module, which opportunely compiles them into a set of (source-dependent) queries to be passed to the wrappers at run time. The queries are then passed to a query planner, which is responsible of establishing the query execution order and of providing the wrappers with intermediate results, where needed. Eventually, the wrappers send the appropriate queries to the remote sources, in order to get the requested data; they also return the results to the mediator, for further analysis and presentation.

In our system we tried to automate as much as possible the source description and the data gathering processes. It is worth noticing that in the biological data realm it is really important the precision preservation and the correctness of the representation: the proposed system aims to guarantee the previous constraints.

Results

A prototypical system has been implemented in JAVA on a Windows 2000 platform. The system consists of approximately 7500 lines of code, and a reduced set of the core functionalities are implemented. Having in mind to firstly integrate data banks about proteins, the model ontology has been mainly developed around the relevant concept of Protein. In the current implementation, we have thus integrated two well known data banks: UniProtKB/Swiss-Prot (containing information about protein sequences) and Swiss-Prot (containing information about protein structures). The system will be extended for integrating different data banks. The system performances are still under evaluation, but preliminary results we have obtained seem to be very promising.

Availability: http://wpage.unina.it/pacapass/BioInView.pdf
Image: http://wpage.unina.it/pacapass/SystemArchitecture.jpg
Email: pacapass@unina.it