A Computational Support for the Access to Integrated Molecular Biology Data

M. Lange[1], A. Freier, U. Scholz, A. Stephanik

Otto-von-Guericke-University of Magdeburg

Department of Computer Science

Institute of Technical and Business Information Systems

Bioinformatics Research Group

P.O. Box 4120, D-39016 Magdeburg, Germany

Tel. ++49-391-67-11291 Fax. ++49-391-67-12020

email: mlange;freier;uscholz;stephani@iti.cs.uni-magdeburg.de

Introduction

The internet is developing into the most powerful medium for information retrieval. This fact is consequently reflected in molecular biology. Thus majority of databases are accessible using the internet. With regard to persistent data storage three general techniques are used: *Web-pages, flat files* and *database systems* (DBS). The public access and querying is mostly achieved by a WWW-server, which acts as middleware between the user interface and the database.

In order to take advantage of the potential of these valuable databases it has to be considered that Bioinformatics is an inherently integrative discipline, requiring standardized access to data from a wide range of sources using powerful query languages e.g. SQL. Without the ability to combine the existing data sources in new and interesting ways, the field of Bioinformatics would be severely limited in scope. Consequently, the integration of databases and the offering of a declarative query language can help to detect new information and coherence.

Most biologists experience database integration by internet researches. In this context, free WWW-based systems like SRS or KEGG are popular. Among others, these two systems are examples for existing database integration solutions available via the WWW.

The very popular SRS is a database querying/navigation system including a large amount of public domain databases (~140 at last count). The databases are locally stored flat files, which are indexed and accessible using a WWW-interface or a special API and retrieval commands. The result of this approach are homogeneous and separately accessible data sources, which can be limited integrated by views or implicit links to cross-reference entries. If a result set contains entries from different databases, the subsets are listed unmerged.

This system and further ones like e.g. KEGG, Entrez, TAMBIS, OPM, Kleisli, DiscoveryLink are often limited in complex declarative query possibilities and the flexible, system independed communication to several heterogeneous data sources.

Mediator Based Molecular Database Integration

Consequently, the idea of a mediator based database integration approach has been resulted in the system architecture of the *BioDataServer* (BDS), which is shown in .

The proposed architecture is realized as a client-server system, where the BDS is the server and any molecular biology application including database import modules could act as clients. In this way several users and related global integrated data schemas can be managed.

The key idea of our approach for molecular database integration is a mediator architecture in conjunction with specific adapters. The BDS is a JAVA application and provides remote services over a TCP/IP socket. Beside read-only SQL queries, additional commands, e.g. for

[1] To whom correspondence should be addressed.

administrative tasks, can be sent as character codes. In order to give JAVA clients the capability to use a standardized access interface to integrated data, a JDBC driver was implemented. A demo JDBC based JAVA applet is available using the URL http://www-bm.cs.uni-magdeburg.de/BDSDemo.

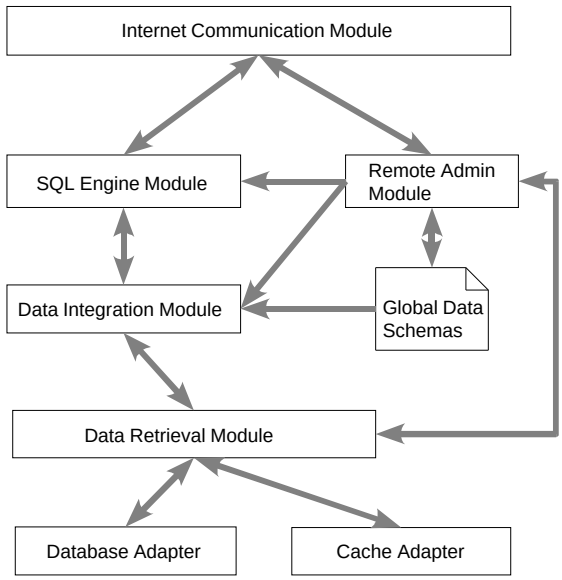

Figure 1 Architecture for our mediator based database integration

In addition, an ODBC driver for Microsoft Windows OS and a WWW-interface with XML capabilities are currently under development. The BDS was designed as a universally applicable component for a homogeneous data acquisition in close context to molecular biology. The attachable software ranges from simple analysis tools (e.g. structural metabolism analysis: http://www-bm.cs.uni-magdeburg.de/phpMetatool) via various molecular information systems up to complete frameworks for complex problems like simulations (e.g. our project MARGBench).

Acknowledgement

This work is based on results of the *Research Group Information Fusion* and the project *Modeling and Animation of Regulatory Gene Networks*, which are kindly supported by the German Research Council (DFG).

Selected References

Baxevanis, A. D., *The Molecular Biology Database Collection: an updated compilation of biological database resources*, Nucleic Acids Research, 2001, vol. 29, no. 1, pp. 1-10.

Roos, D. S., *Bioinformatics--Trying to Swim in a Sea of Data*, Science, 2001, vol. 291, no. 5507, pp. 1260-1261.

Özsu, M. T. and Valduriez, P., *Principles of Distributed Database Systems*, London et al.: Prentice-Hall, 2nd international edition, 1999.

Etzold, T. et al., *SRS: Information Retrieval System for Molecular Biology Data Banks*, Methods in Enzymolgy, 1996, vol. 266, pp. 114-128.

Freier, A. et al, *MARGBench - An Approach for Integration, Modeling and Animation of Metabolic Networks*, in Proceedings of the German Conference on Bioinformatics (GCB '99), Hannover, Germany, 1999, pp. 190-194.

Codd, E. F., *A Relational Model of Data for Large Shared Data Banks*, Communications of the ACM, 1970, vol. 13, no. 6, pp. 377-387.

Hofestaedt, R. et al., *Information Processing for the Analysis of Metabolic Pathways and Inborn Errors*, BioSystems, 1998, vol. 47, pp. 91-102.