DBOra: An Integrated Database for Context Based Search and Browsing in Protein Data

Matthias Lange, Vasudev Kumanduri, Thomas Münch, Roland Schnee, Uwe Scholz, Patrick Schweizer
Institute of Plant Genetics and Crop Plant Research (IPK)
Corrensstraße 3, D-06466 Gatersleben, Germany
lange@ipk-gatersleben.de

Abstract

DBOra is an integrated database of interlinked protein data. It is based on a well structured and normalized database schema. A large set of data was imported and enriched with suitable indexes. Thus, it provides a WWW-based user interface to formulate intuitive queries as a combination of simple search terms. Furthermore, it is possible to search for nucleotide and amino-acid sequences by simply switching to another query type. This tool enables a specific search with fuzzy queries and is freely available in the WWW: **http://www-bm.ipk-gatersleben.de/stable/php/idbb**.

Keywords. Integrated databases, Context search, Text similarity, Sequence similarity

1. Introduction

In life science a steadily increasing amount of experimental and computed data is being produced. For instance, [1] lists 386 via the internet accessible databases and categorizes them. Within this list, protein related databases take a high rank.

As of now, one will find information by explicitly cross linked data sources using hyperlinks. This situation forces the user to spend a lot of time for intelligent surfing. However, main problems of the manual Internet navigation are, e.g. that the interesting data is distributed over several pages and the data itself is poorly structured. In turn, this narrows the capabilities of specific searching for relevant data. For example, to retrieve all relevant proteins that are transcription factors it is not very valuable to use an "all text" search for the term "transcription factor" in an integrated life science database, like SRS [3]. This is because the results are in general not very specific for sensitive queries and vice versa. Thus, a user friendly tool for a context based search would be useful. It should be based on a well structured database with a minimum of unstructured free text. Furthermore, very it should support intuitive queries as well as complex composed ones, without the knowledge of query languages like SQL[1]. Finally, indexes should be embedded to deliver query results in a suitable time. To realize such a system, we suggest the subsequently presented system called DBOra.

2. The Integrated Database DBOra

DBOra is an integrated database on an ORACLE9i DBMS[2] in combination with an intiutive user interface. It enables the search for nucleotide and amino-acid sequences as well as text search. The provided data stems from the connected databases SWISS-PROT/TrEMBL [2], BRENDA [7], KEGG [6] and OMIM[5]. In a first step the scientific user may search for sequences or text and then navigate through the results and display further related data sets and information.

For this, an integrated database with 81 interconnected tables was modeled and implemented as a relational database schema. 50 tables alone representing database links, which makes the developed schema and thus the database a comprehensive basis to search in interconnected data. More detailed information on the underlying data schema are available using the

[1] Structured Query Language - standardized query language for relational database management systems
[2] Database Management System

URL:
http://www-bm.ipk-gatersleben.de/stable/php/idbb/Slices/schema_1.html

To fill DBOra with data, database integration was utilized to import raw data from the mentioned data sources to a locally integrated database - compareble to a Data Warehouse. The used database integration approach was developed within the MARG project at the University of Magdeburg. The developed software is called BioDataServer (BDS) [4] and implements a mediator based architecture in conjunction with semi-automatically generated database access adapters. This import is periodically executed to update the database. Currently, DBOra contains 836013 SWISS-PROT and TrEMBL entries that were relational restructured and normalized. Thus, the database contains about 35 million tuple distributed over the mentioned data schema. Including indexes, the overall size of the database is about 10 Gbyte.

3. Capabilities of Database Search

A straight forward search and browsing interface of the integrated database was implemented by using a PHP based web interface. Currently, three kinds of search queries are supported:

1. Text (including phonetic, word steam, fuzzy match and word neighborhood)
2. amino-acid sequence homology search (using local BlastP)
3. nucleotide sequence homology search (using local BlastX)

This tool and further information concerning the database schema, construction rules of search queries and examples are available at: http://www-bm.ipk-gatersleben.de/stable/php/idbb. The user is requested to register himself to use the full functionality or to login as guest.

4. Further Work

In further work DBOra is used for automatic functional annotation. The idea is that homology search is one entry point to finding functional annotation for unknown sequences. Herefore, it is necessary to query several databases. To combine the wide data spectrum of integrated databases with the advantage of a homogeneous, central entry point for homology searches, it is necessary to treat the integrated database as a virtual union of inter-linked data items. Thus, it will be possible to put a sequence at the entry of a maze of interconnected life science databases and search, the way to the treasure - e.g. the functional categorization.

Acknowledgement

This work is based on results of the InnoPlanta project “Elektronische Infrastruktur für die Pflanzenbiotechnologie”, which is kindly supported by the German Ministry of Research and Education.

References

[1] A. D. Baxevanis. The Molecular Biology Database Collection: 2003 update. *Nucleic Acids Research*, 31(1):1–12, 2003.

[2] B. Boeckmann, A. Bairoch, R. Apweiler, M.-C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and M. Schneider. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research*, 31(1):365–370, 2003.

[3] T. Etzold, A. Ulyanow, and P. Argos. SRS: Information Retrieval System for Molecular Biology Data Banks. *Methods in Enzymology*, 266:114–128, 1996.

[4] A. Freier, R. Hofestädt, M. Lange, U. Scholz, and A. Stephanik. BioDataServer: A SQL-based service for the online integration of life science data. *In Silico Biology*, 2(0005), 2002. *Online Journal*: `http://www.bioinfo.de/isb/2002/02/0005/`.

[5] A. Hamosh, A. F. Scott, J. Amberger, C. Bocchini, D. Valle, and V. A. McKusick. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 30(1):52–55, 2002.

[6] M. Kanehisa, S. Goto, S. Kawashima, and A. Nakaya. The KEGG database at GenomeNet. *Nucleic Acids Research*, 30(1):42–46, 2002. http://www.genome.ad.jp/kegg/.

[7] I. Schomburg, A. Chang, and D. Schomburg. BRENDA, enzyme data and metabolic information. *Nucleic Acids Research*, 30(1):47–49, 2002.