DAWIS-M.D. - a data warehouse system for metabolic data

Klaus Hippe, Benjamin Kormeier, Thoralf Töpel, Sebastian Janowski and Ralf Hofestädt

Bioinformatics Department
Bielefeld University
Universitätsstraße 25
33501 Bielefeld

khippe@techfak.uni-bielefeld.de

Abstract: Today, experiments like microarrays and other high-throughput methods generate a large number of life sciences data. Biologists need to be supported in their research by tools and applications that can analyze and interpret experimental data while considering external data sources or data from experiments of other researchers. Moreover, relationships and interactions between different data sets and biomedical domains have to be detected and represented in a clear and understandable manner. These questions are some of the major challenges in bioinfomatics. Therefore, we present DAWIS-M.D., a platform-independent data warehouse system for metabolic data. This data warehouse information system provides an integrated and consistent view of large scale biomedical data. DAWIS-M.D. is a publicly available web-based system that integrates data from 11 different biomedical databases.

1 Introduction

Biomedical research generates a large amount of heterogeneous data. This data is usually stored in several heterogeneous biomedical databases. Typically, these databases are distributed over the entire world. The information of these databases comes from various experiments and from literature as well. Today, approximately 1230 publicly available databases and information systems for life science data are listed in the NAR catalogue [CG10]. Therefore, data integration is necessary to provide a consistent view of the distributed, heterogeneous and duplicated data.

Scientists and research groups usually use several data sources and information systems simultaneously, e.g. in parallel windows of their web browser. It is then difficult for the scientist to understand complex biological questions and find missing links, because of lacking information. Another problem is the incompatibility between the different information systems and software solutions.

In this paper we present DAWIS-M.D., a platform-independent data warehouse approach for metabolic data. DAWIS-M.D (*Data Warehouse Information System for Metabolic Data*). The information system contains information from 11 different databases. The following data sources are integrated into the data warehouse: BRENDA, EMBL, HPRD,

KEGG, OMIM, SCOP, Transfac, Transpath, ENZYME, GO and UniProt. The data in DAWIS-M.D. is divided into 12 various biological domains, which can be accessed via the web-based graphical user interface (GUI). The application provides search forms for the biological domains *Compound*, *Disease*, *Drug*, *Transcription Factor*, *Enzyme*, *Gene*, *Glycan*, *Gene Ontology*, *Pathway*, *Protein*, *Reaction* and *Reaction Pair* domain.

A wide range of biomedical information in DAWIS-M.D. supports scientists in understanding complex biological systems and their properties. In addition, DAWIS-M.D. identifies relationships and interactions spanning multiple biological domains and is able to display this information.

Networks can be displayed, edited and extended by using the VANESA (*Visualization and Analysis of Networks in System Biology Applications*) [JKT+10] software application. All information and biological knowledge in VANESA is provided by the DAWIS-M.D. data warehouse that is connected via web service to the VANESA network editor.

2 Related Works

One of the major topics in bioinformatics is data integration. The integration of heterogeneous, autonomous and distributed data can be realized using different approaches. Generally, in bioinformatics, integration approaches are divided into following classes: text indexing systems (e.g. SRS [EUA96]), multi database and federated database systems (e.g. DiscoveryLink [HSK+01]) and data warehouses systems (e.g. BioWarehouse [LPW+06], ONDEX [KBT+06] and SYSTOMONAS [CMB+07]). Data warehouse systems are widely used in bioinformatics for data integration. Advantages of the data warehouse techniques are high performance, data cleansing and full and easy access to the integrated data that is pivotal for bioinformatics.

In bioinformatics data warehouse approaches can be divided into two groups: General software infrastructures that can be locally installed and configured like BioWarehouse. And project-oriented data warehouse systems that are implemented for specific biological questions like SYSTOMONAS. Some software solutions cannot be categorized, because they have characteristics of both groups. These solutions are hybrid data warehouse approaches.

All these existing systems have several disadvantages: they are not platform-independent, the local installation is quite time-consuming or they are not available via the web. Furthermore, they are restricted to an operating system concerning their specific programming language. Therefore, the applications are not extensible or flexible enough for other developers. Another problem is the need for up-to-date content in the data warehouse. Most software solutions have to be updated manually, and this is a time-consuming process. Additionally, all systems are restricted to a specific relational database management system (RDBMS) and therefore are not independent from manufacturers such as MySQL, PostreSQL and Oracle. These systems cannot be widely used, because of their limitation to one RDBMS. Further problems are the usability, the project-specific data and missing interfaces for more flexibility.

Table 1 illustrates the main data warehouse systems used in bioinformatics. This table also displays the advantages and disadvantages of each software solution. Based on the advantages and disadvantages of existing data warehouse approaches we developed the concept of DAWIS-M.D. The main goal was to develop a platform-independent and flexible data warehouse system for metabolic data that integrates multiple heterogeneous data sources into a local database. The system should enable an intuitive search of integrated life science data, simple navigation to related information as well as visualization of biological domains and their relationship.

Feature	**BioWarehouse**	**ONDEX**	**SYSTOMONAS**
Integration	Close integration, ready-made relational schemas	Loose integration	Unknown
DBMS	MySQL, Oracle	PostgreSQL	PostgreSQL
Programming language	Java, C	Java	PHP
Architecture	Software infrastructure	Software infrastructure, Web application	Web application
Platform independence	Unix-based	Yes	Yes
Up-to-dateness	Manually	High	Unknown
Maintenance and further development	Regular maintenance and further development	Regular maintenance and further development	Regular maintenance and further development
License	MPL	GNU General Public License	Database dump available, web application free
Open source	Yes	Yes	Unknown

Table 1: Comparison of three exemplary bioinformatics data warehouses systems. For each software solution the advantages are highlighted green and the disadvantages are red.

3 Design and Implementation

The system architecture of DAWIS-M.D. consists of five different layers. The source layer contains the multiple data sources such as BRENDA, EMBL, GO, Enzyme, KEGG, HPRD, OMIM, SCOP, Transfac, Transpath and UniProt.

The different biomedical databases, which are integrated in DAWIS-M.D., are available in the data sources. Most of these databases provide parseable flat files, XML files or Structured Query Language (SQL) dumps that can be processed by the BioDWH data

warehouse infrastructure [TKKH08] that is used for the integration process. A monitor component that is part of the integration layer controls the different data sources. It recognizes changes in the original sources and starts the download if necessary. In a defined cycle the parser will be activated to start the Extraction-Transform-Load (ETL) process. Furthermore, BioDWH supports the technique of object-relational mapping (ORM).

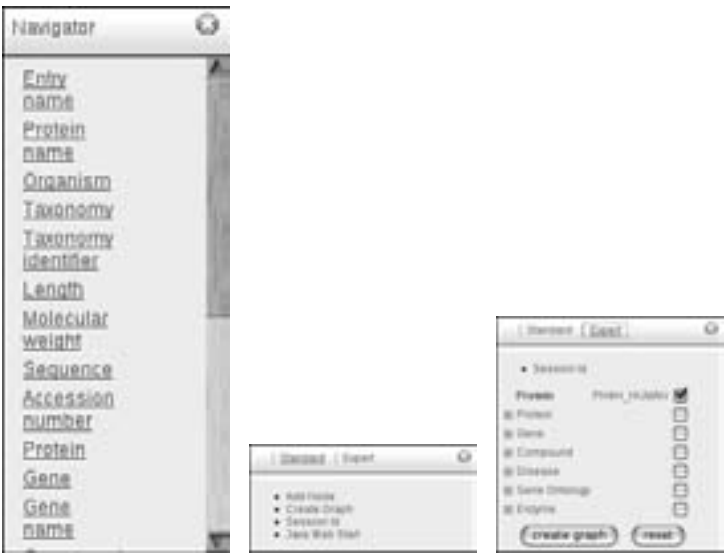
The database layer is essential for the DAWIS-M.D. information system. It contains all information of the integrated data. Furthermore, modules for administration and meta data for the 12 biological domains are located in this layer. As a relational backend for DAWIS-M.D. data warehouse the open source RDBMS MySQL is used. The databases layer communicates via JDBC with the persistence layer. The persistence layer provides all components for the ORM, whereby the application layer is independent. Thus, it is possible to support different RDBMS such as MySQL, Oracle or PostgreSQL without changing the application logic. Accordingly, the ORM technique enables a high level of flexibility and independence between database layer and application logic. The persistence layer was realized by using the open source framework Hibernate.

The application layer communicates with the persistence layer. A user can interact via the application layer with the system. Therefore, the web application provides a homogenous and integrated view of the data. A web service controls the communication between the server and external tools and the network editor VANESA. The web service communicates with external applications by using SOAP. For the implementation of the web service the open source framework Apache Axis2 is used. A user is able to send a single element or all elements of a DAWIS-M.D. entry to the VANESA network editor. Then, the VANESA software is able to generate a network of the entry, i.e. an entry of the web application. Furthermore, the resulting network can be extended and edited in VANESA by the scientist.

The web-based graphical user interface of DAWIS-M.D. is implemented with JSP that is based on the programming language Java and runs on an Apache Tomcat web server. For dynamic and interactive websites a web server with JSP-/Servlet engine and additional Web 2.0 technologies are required. Each domain provides a special search form to formulate simple queries to the data warehouse. The results of a query to the database are presented on a new web site. The information is presented as a table. Information that is larger than a certain length is hidden to keep a clear and well-structured overview. This information can be displayed by clicking a specific keyword. Using this ”toggle” function it is possible to show important information and also to hide information which is not of interest.

Moreover, the web application provides a local navigation bar that is shown in figure 1. It is possible to navigate easily and fast within the result page. The navigation bar is only displayed for large result pages. Furthermore, the relationships and interactions between the different biological elements such as genes and proteins are presented and highlighted for each entry, if available. For proteins and enzymes, a graphical representation of a protein classification, using SCOP hierarchy, is available in DAWIS-M.D.

One of the most important features of DAWIS-M.D. is the interaction with the network editor VANESA. The web application provides two different views for the interaction.



(a) Navigation bar (b) Standard view (c) Expert view

Figure 1: Graphical representation of the navigation and the communication bridge between DAWIS-M.D. and VANESA.

Figure 1 shows the Standard and Expert view of the communication bridge. Using the standard view a user is able to display the root element of the entry in the web application as a node in the network editor. The functionality of the expert view is nearly the same, but in this case the user is able to select elements more precisely. Therefore, it is possible to create user-defined networks in VANESA to the current entry.

Finally, the data warehouse information system provides additional information about the integrated databases such as release or update information. The statistics are presented as table or diagram. Based on the statistics it is clear how the database of DAWIS-M.D. is composed.

4 Summary

One of the major challenges in bioinfomatics is the integration and management of data from different sources and their presentation in a user-friendly format. Therefore, in this paper we presented DAWIS-M.D., a platform-independent data warehouse information system for metabolic data. The information system integrates data from 11 widely-used life science databases. The information of integrated databases is divided into 12 various biological domains, which are available via the graphical user interface of the web application.

The data warehouse architecture provides a platform-independent web-interface that can

be used with any common web browser. The system enables intuitive search of integrated life science data, simple navigation to related information as well as visualization of biological domains and their relationships. To ensure maximum up-to-dateness of the integrated data the BioDWH data warehouse infrastructure including a monitor component is used. The persistence layer of DAWIS-M.D. uses the ORM technique whereby the application layer is independent from database layer. Thereby, it is possible to support different database management systems.

The DAWIS-M.D. data warehouse incorporates the advantages of a navigation and informational system and builds a bridge to the network editor approach VANESA. Hence, it is possible to browse through the integrated life science data and bring the information into a modeling and visualization environment. Therefore, it is easy for the scientists to search information of interest, find relationships and interactions between different biomedical domains and bring them for editing, manipulation and analyzing directly into the VANESA network editor. Finally, the scientists gain a better understanding of complex biological problems and are able to develop new theoretical models for further experiments. DAWIS-M.D. is available at http://agbi.techfak.uni-bielefeld.de/DAWISMD/.

References

[CG10] G. R. Cochrane and M. Y. Galperin. The 2010 Nucleic Acids Research Database Issue and online Database Collection: a community of data resources. *Nucleic Acids Research*, 38:D1–D4, 2010.

[CMB+07] C. Choi, R. Münch, B. Bunk, J. Barthelmes, C. Ebeling, D. Schomburg, M. Schobert, and D. Jahn. Combination of a data warehouse concept with web services for the establishment of the Pseudomonas systems biology database SYSTOMONAS. *Journal of Integrative Bioinformatics*, 4(1), 2007.

[EUA96] T. Etzold, A. Ulyanov, and P. Argos. SRS: Information retrieval system for molecular biology data banks. *Methods in Enzymology*, 266:114–128, 1996.

[HSK+01] L. M. Haas, P. M. Schwarz, P. Kodali, E. Kotlar, J. E. Rice, and W. C. Swope. DiscoveryLink: a system for integrated access to life sciences data sources. *IBM Systems Journal*, 40:489–511, 2001.

[JKT+10] S. Janowski, B. Kormeier, T. Töpel, K. Hippe, R. Hofestädt, N. Willassen, R. Friesen, S. Rubert, D. Borck, P. Haugen, and M. Chen. Modeling of cell-to-cell communication processes with Petri nets using the example of quorum sensing. *In Silico Biology*, 10:0003, 2010.

[KBT+06] J. Köhler, J. Baumbach, J. Taubert, M. Specht, A. Skusa, A. Rüegg, C. Rawlings, P. Verrier, and S. Philippi. Graph-based analysis and visualization of experimental results with ONDEX. *Bioinformatics*, 22(11):1383–1390, 2006.

[LPW+06] T. J. Lee, Y. Pouliot, V. Wagner, P. Gupta, D. W. J. Stringer-Calvert, J. D. Tenenbaum, and P. D. Karp. BioWarehouse: a bioinformatics database warehouse toolkit. *BMC Bioinformatics*, 7, 2006.

[TKKH08] T. Töpel, B. Kormeier, A. Klassen, and R. Hofestädt. BioDWH: A Data Warehouse Kit for Life Science Data Integration. *Journal of Integrative Bioinformatics*, 5(2), 2008.