Reconstruction of biological networks based on life science data integration

Benjamin Kormeier[1,*], Klaus Hippe[1], Patrizio Arrigo[2], Thoralf Töpel[1], Sebastian Janowski[1] and Ralf Hofestädt[1]

[1]Bielefeld University, Bioinformatics Department PO Box 100131, D-33501 Bielefeld, Germany

[2]CNR ISMAC, Section of Genoa, Via De Marini 6, 16149 Genova, Italy

Summary

For the implementation of the virtual cell, the fundamental question is how to model and simulate complex biological networks. Therefore, based on relevant molecular database and information systems, biological data integration is an essential step in constructing biological networks. In this paper, we will motivate the applications BioDWH - an integration toolkit for building life science data warehouses, CardioVINEdb - a information system for biological data in cardiovascular-disease and VANESA - a network editor for modeling and simulation of biological networks. Based on this integration process, the system supports the generation of biological network models. A case study of a cardiovascular-disease related gene-regulated biological network is also presented.

1 Introduction

Large amounts of high dimensional biological data are generated from different high-throughput experiments and from literature. The rapidly growing number of databases and data types leads to the challenge of integrating the heterogeneous data, particularly in biology. Currently there are about 1170 important molecular biology databases [1] on different aspects of biological systems like sequences, proteins or pathways. Thus, the challenge is to capture, model, integrate and analyze the data in a consistent way to provide a new and deeper insight into complex biological systems.

High-throughput sequence investigation tools, array technologies for gene or protein analysis and the expanding electrical infrastructure for the study of molecular data represent the initiation of a virtual cell. Today the vision of implementation of a virtual cell has united bioinformatics and systems biology. However, we are still a long way from implementing even a simple virtual cell. The first step in reaching this goal is to understand metabolism, which is based on gene-controlled biochemical reactions. Therefore, modeling and simulation of metabolic networks is important. Different methods of modeling biological networks have been introduced in many publications. Another problem is the quantitative simulation of these processes. Therefore, it is still an open question to find the most useful method for the simulation of biological networks, which will represent the backbone of a virtual cell. We will present a new tool which creates a large-scale biological network using data integration and data warehousing methods.

*To whom correspondence should be addressed. E-mail: bkormeie@techfak.uni-bielefeld.de

doi:10.2390/biecoll-jib-2010-146

2 Related software platforms

The integration of life science data from heterogeneous, autonomous and distributed data sources is an important research field. One of the major challenges in data integration is the large heterogeneity of the databases on the semantic and technical level. Existing systems are based on different data integration techniques:

- indexing systems like SRS [2],
- multi database and federated database systems such as Discovery Link [3],
- ontology-based integration like CoryneRegNet [4] or ONDEX [5],
- data warehouse systems such Atlas [6], BioMart [7], BioWarehouse [8], Columba [9] and SYSTOMONAS [10].

The data warehouse is one of the famous architectures of materialized integration. In bioinformatics the data warehouse is usually used for data integration. For this paper, data warehouses are essential. Therefore, we will focus on two data warehouse approaches, Atlas and BioWarehouse.

The goal of Atlas is to provide data, as well as a software infrastructure for bioinformatics research and development. The biological data warehouse locally stores and integrates from biological sequences, molecular interactions, homology information, functional annotations of genes, and biological ontologies. The data sources of Atlas are categorized into four classes: sequence, molecular interactions, gene related resources and ontology. A full list of the Atlas data sources could be found in [6].

Each mentioned category has its own database schema in the Atlas relational database model. Depending on the level coupling, this approach can be categorized as a tightly coupled system. As relational backend, Atlas uses the open source software DBMS MySQL.

The different APIs of Atlas are developed in three programming languages C++, Java and Perl. But not every API is available in the respective programming language. The Atlas system provides many Unix command line tools, such as ac2seq, which is able to find a sequence in FASTA format on the basis of accession numbers. Moreover, the user is able to send direct SQL queries via MySQL client to the data warehouse.

The BioWarehouse is part of the Bio-SPICE (Biological Simulation Program for Intra- and Inter-Cellular Evaluation) project, an open source framework and software toolset for systems biology. BioWarehouse is an open source toolkit that integrates a multiplicity of biological databases. BioWarehouse facilitates creating user defined and user specific data warehouse instances. Several data sources for the toolkit are available, such as ENZYME, KEGG, GO and UniProt, which are also part of this work. Similar to the Atlas system, different relational database schema exist according to the different data types. Therefore, this approach can be characterized as loose coupled system. BioWarehouse supports MySQL and Oracle database management systems. Integration of the different data sources is realized by a specific loader. Each loader is adapted for a particular data source. Due to heterogeneities between different data sources, the data is transformed into a consistent format and afterwards transferred into

doi:10.2390/biecoll-jib-2010-146

the database schema. Loaders have been implemented in programming languages C and Java. A special feature of the loaders is the error-tolerance during integration. In case an error occurs the integration process will be completed and the incorrect data sets will be marked. Furthermore, the BioWarehouse implementation provides a set of java utility classes that are useful for developers who would like to construct their own loaders or applications. A publicly available version of BioWarehouse called PublicHouse is available via internet. The second alternative to using the system is to install the software locally on the computer and select the required data sources. In the locally installed version, users are able to manage the data sources and do not have only-read access.

However, the disadvantage of all these approaches is that they are time-consuming when installed locally or they are only available via the web. Usually, they are restricted to an operating system such as Unix/Linux or to specific programming languages. Another problem is update strategies; the data is either old or the data warehouse has to be updated manually. Figure 1 illustrates these main restrictions of current bioinformatics data warehouse projects. BioDWH data warehouse application intends to increase customization of the data warehouse concept with the advantages of better performance, scalability, up-to-dateness and data quality.

3 Data integration with BioDWH

BioDWH [11] is implemented in Java and uses a relational database management system in its backend, e.g., Oracle or MySQL. It provides an easy-to-use Java application for parsing and loading the source data into the data warehouse. Several ready-to-use parsers for popular life science information systems are already available, such as: UniProt, KEGG, OMIM, GO, Enzyme, BRENDA, MINT, IntAct, SCOP, EMBL-Bank, TRANSFAC / TRANSPATH, Reactome and IProClass. Furthermore, a configurable monitor for data source updates is part of the system. For status requests to the data warehouse, we have developed a graphical user interface based on Java that works on every system which is installed with the Java Runtime Environment.

A well-engineered, object-relational mapping tool called Hibernate was used as a persistence layer, which performs well and is independent from manufacturers like MySQL or Oracle. Additionally, the Hibernate framework fits perfectly into the Java-based infrastructure of the data warehouse. A Java interface and the object-relational mapping using Hibernate or Java Persistence Architecture (JPA) constitute an easy plug-in architecture for the integration of a new parser. This object-relational mapping (ORM) is an automated and transparent persistence method of Java application for tables in a relational database system, whereas a mapping between objects and metadata of the database is described. In principle, ORM works with reversible transformation of data from one representation into another.

An ORM solution consists of four parts: first, an application programming interface (API) that executes simple CRUD (create, retrieve, update, delete) operations using objects of persistent classes. Secondly, a programming language or API to formulate queries that depend on Java entity classes or properties of classes. Thirdly, a facility for mapping metadata. And finally, techniques of an ORM implementation to handle interactions of dirty checking, lazy association fetching and other optimization functions of transactional objects.

The different features of BioDWH are usable with a graphical user interface. It enables the

doi:10.2390/biecoll-jib-2010-146

configuration of the monitor and parser for the different public life science data sources as well as the local database management system.

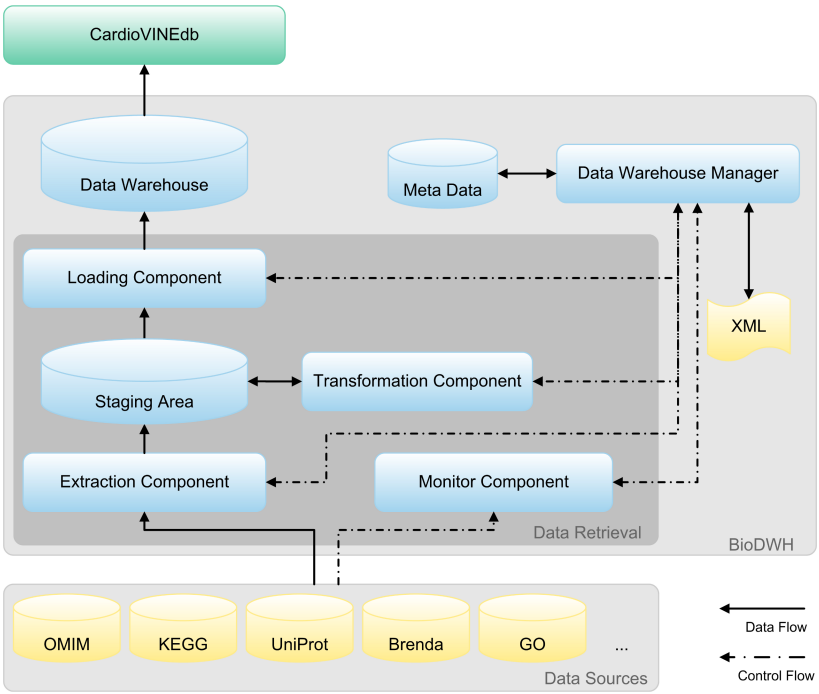

Figure 1: Schematic illustration of the BioDWH system architecture following the general data warehouse design. The data source layer comprises parseable flat files, XML files or Structured Query Language (SQL) dumps of the original heterogeneous data sources. BioDWH provides a number of ready-to-use parsers to extract and transform the data from data sources and to load the content into a data warehouse (i.e. relational database system). The CardioVINEdb information system uses the integrated data of the data warehouse. CardioVINEdb is realized as a web application, which offers advantages like platform independence and high usability.

4 Web-based information system CardioVINEdb

Based on the CardioWorkBench EU project, we implemented a platform-independent information system that integrates multiple heterogeneous data sources into a local database enriched with protein microarrays from human smooth muscle cells that are related to cardiovascular diseases. Based on our VINEdb [12] information system, we extended CardioVINEdb with more data sources, better data warehouse infrastructure (BioDWH) and microarray data. In addition, we upgraded the visualization components and web pages for better navigation and exploration. To ensure maximum up-to-dateness of the integrated data, we used the data warehouse infrastructure BioDWH to collect and to merge the data from different molecular biology data sources.

The CardioVINEdb system architecture consists of a 3-layer architecture that is illustrated in Figure 2. The source layer contains the multiple data sources BRENDA, EMBL, GO, IntAct, KEGG, MINT, OMIM, PubChem, SCOP, TRANSFAC / TRANSPATH and UniProt. In addition to the publicly available databases, we integrated experimental microarray data of human smooth muscle cells that are associated with cardiovascular diseases. Most of these databases provide parseable flat files that can be processed by data warehouse infrastructure BioDWH. A monitor component that is part of the integration layer controls the different data sources.

It recognizes changes in the original sources and begins downloading if files have changed. In a defined cycle the parser will be activated to start the ETL (Extraction-Transform-Load) process. ETL means that data is extracted from the source data, transformed into the data warehouse schema and loaded into the data warehouse. Data marts for specific analysis applications can easily be constructed by the database layer, i.e. the data warehouse.

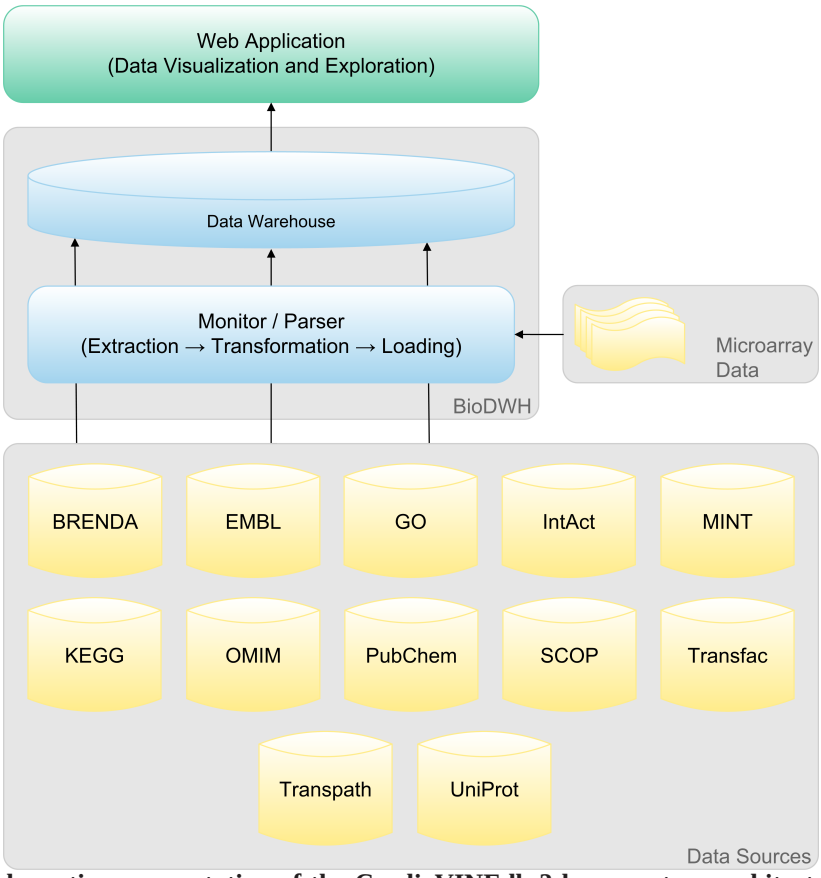

Figure 2: Schematic representation of the CardioVINEdb 3-layer system architecture from the original heterogeneous data sources to the web application layer. The data source layer contains flat files of the different data sources. The data warehouse infrastructure BioDWH is used to extract, transform and to load the data from different molecular biology data sources into the data warehouse (ETL-Process). Also microarray data can be integrated into the data warehouse by BioDWH. The web application layer represents the web-based graphical user interface of CardioVINEdb that allows the user to find information of interest from the different databases illustrated in the data source layer. Furthermore, a network-based visualization enables intuitive and comfortable exploration of the integrated data.

The web-based graphical user interface of CardioVINEdb is implemented with JavaServer Pages (JSP) and runs on an Apache Tomcat web server. Each data entry has detailed information and a further link to the original data source. A comprehensive search engine allows the user to find information of interest spanning multiple domains, such as proteins, enzymes, genes, compounds etc. Additionally, each domain has its own specific search engine to find required information for research.

For a better understanding of the relationships between the biological objects, the network-based visualization enables intuitive and comfortable exploration of the integrated data. The system produces a PNG image file with the graphical visualization of biological objects in different domains and their linkage. Finally, this image is embedded in the HTML pages and

doi:10.2390/biecoll-jib-2010-146

displayed by the web browser. The dynamic component is using a Java Applet. In this case, the graph is directly generated and displayed within the applet. For more interactive navigation and exploration the applet has a zoom function, different graph layouts and a picking function to move and select nodes within the graph. Therefore, the applet is embedded in the HTML pages and can be displayed by the web browser if Java Runtime Environment is installed on the computer. The database management for integrated data is realized in MySQL using Java Database Connectivity (JDBC). For data integration we used the data warehouse infrastructure BioDWH which is described in Section 3.

5 Modeling of biological networks with VANESA

The software solution VANESA [13] is outstanding because of its comprehensive graphical network representation of biological research data. Information is visualized in a clear and understandable manner to meet the purposes of underlying research activities. By an intuitive graphical user interface, the user is enabled to record research results and thoughts in the form of a digital network model. The user is not limited to any kind of biological model; moreover it is possible to create an individual system well-suited to the wishes and requirements of each research activity.

Another important point is access to external biomedical data sources. The software solution consults different kinds of databases to support the user with useful information. VANESA provides 11 different databases like KEGG and Brenda that are available based on data integration. The communication between VANESA and the biomedical data sources is realized by a web service. The user merely needs an internet connection to access the information of interest. There is no need for a local data repository.

The mentioned data sources provide an established basis for the modeling and the characterization of a biomedical system. Moreover, the information from these data sources can aid in finding missing links in a system. The data integration is a powerful feature of VANESA that provides many possibilities. Furthermore, graph comparison and graph theory functions support the user in a better understanding of biological circumstances. Highlighting and comparison functions point out important facts in a set of different models. In order to make the graphical representation and the analysis on the networks more legible, graph layout transformations and animation algorithms are considered as well. Furthermore, an interface for experimental data like PCR (Polymerase Chain Reaction) and microarray are also taken into consideration. Export formats such as SBML, CSML and GraphML are provided by the software solution.

The software solution is a unit that consists of many individual elements which are described as follows (Figure 3). The main components concern the data modeling, loading and transformation. Those elements compose the framework of the software solution.

On the one hand they realize the representation of the network models and on the other hand they make a data and information integration possible. The connection to the external data sources BRENDA and KEGG is realized by an Axis2 Web service. This service connects to a data warehouse and the underlying data sources to gather biological and medical information of interest. The database management for integrated data is realized in MySQL. For data integration we used the data warehouse infrastructure BioDWH which is described in Section 3.

doi:10.2390/biecoll-jib-2010-146

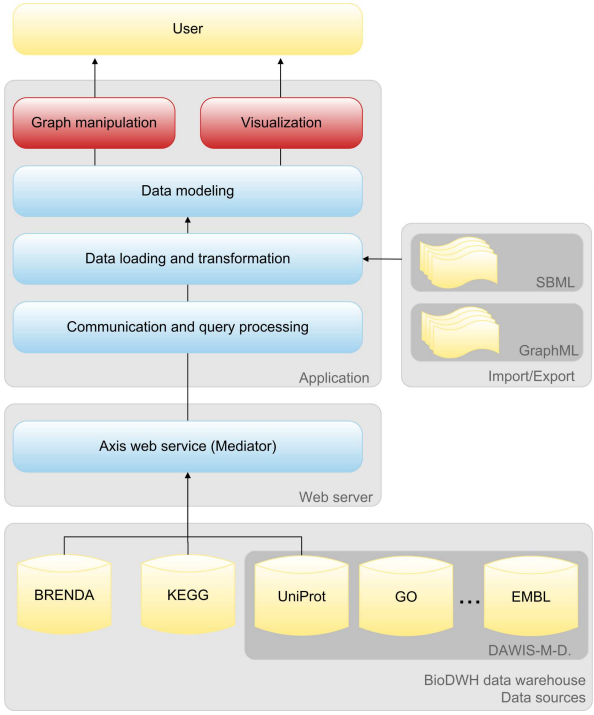

Figure 3: Schematic representation of the VANESA 3-layer system architecture.

The user interacts with the software through the visualization and graph manipulation components. With these components the user has at hand the possibilities to create, edit and examine network models. The final important component concerns the export and import functions. These functions also realize an inter-exchange of network models among different users and different software solutions.

6 Case-study of cardiovascular disease

Cardiovascular diseases (CVDs) are the leading causes of death in developed countries. The CVDs are classified as multifactorial diseases. A multifactorial disease is caused by a complex set of synergistic actions that involve endogenous factors (genetic susceptibility) and exogenous factors (diet, environmental contaminants, life-style and other causes). It is quite complex estimating the role of the main factors at the onset of the disease. In many cases, the currently available biomarkers are not sufficient for diagnostic purposes. The advent of high-throughput genomic and proteomic techniques have enhanced the capability to find new potential biomarkers at the cellular level. Proteomic techniques are the selected methods for biomarker screening. The majority of biomarkers are screened by using electrophoretic methods. In the last decade new proteomic tools have been developed. The most promising proteomic tool is the antibody array. This technique consists of obtaining a semi-quantitative proteomic profile by using an array of protein specific monoclonal antibodies. The protein expression level is estimated on the basis of antigen-antibody reaction intensity. The main advantage of this approach is the possibility of detecting hundreds of proteins at a time. The outcome of antibody array is easy to integrate with other databases, for instance KEGG, because it does not need high quality image processing tools such as a 2D-page. The major disadvantage of this method is that it screens only known proteins. The analysis of antibody array is similar to that applied to transcript array.

The two applications of exploratory data analysis differ in terms of the spread of fluorescence data. The range of intensity of antibody array is more restricted than that of transcript array. This difference requires the application of modified data preprocessing tools.

The aim of our proteomic analysis is the identification of deregulated proteins in human smooth muscle cell samples. Our analysis has been carried out in three cytoplasmic samples (S12 fraction) of cultured aortic smooth muscle cells. These cells have been extracted from three different DCM patients with additional pathologies.

From the perspective to identifying functional biomarkers we are taking the protein functional context into account.

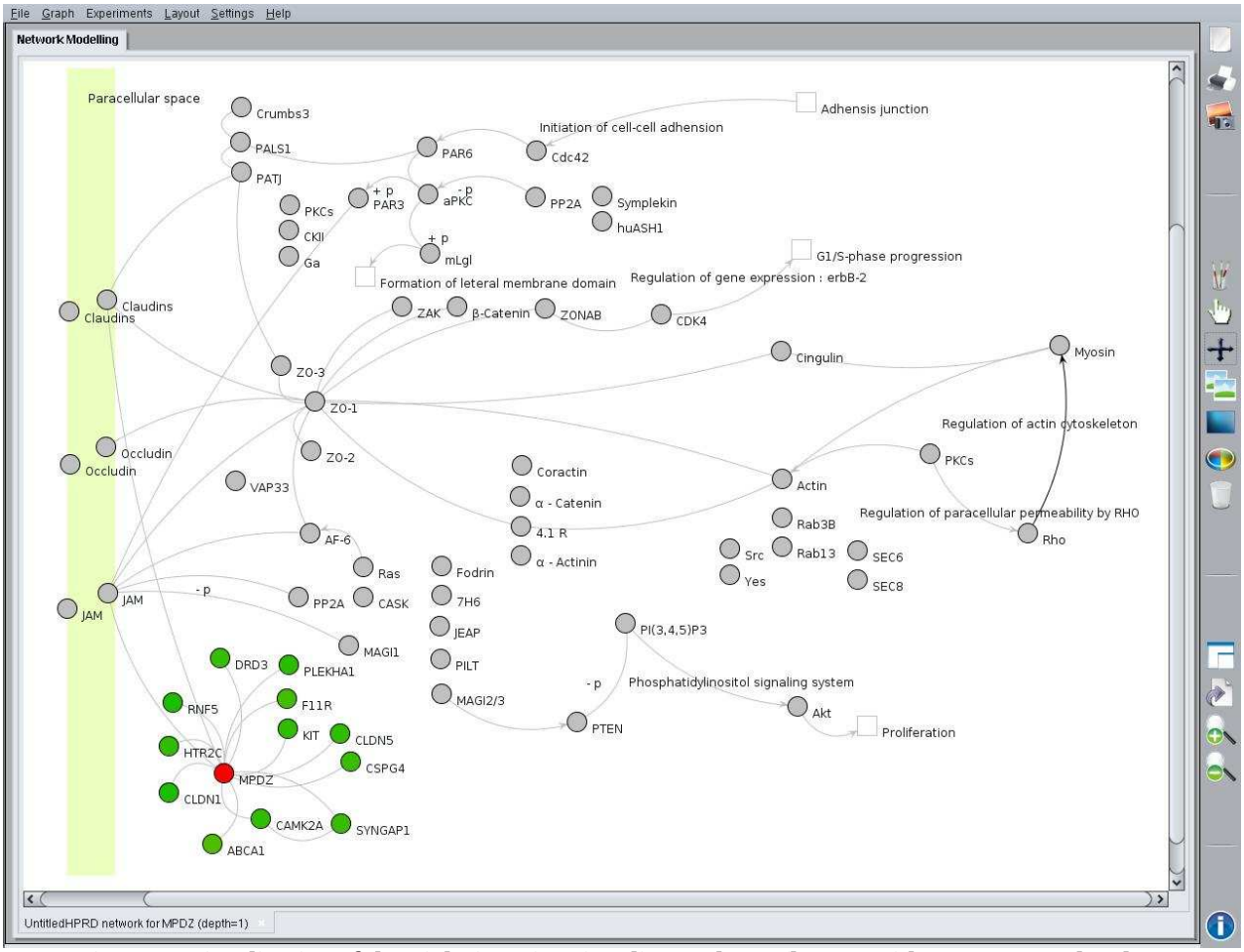

Figure 4: Visualization of the Tight junction signaling pathway (hsa04530) by VANESA. The place marked in red is the relevant protein of the microarray sample.

Preliminarily, we have created a tabular representation of all the proteins mapped onto the Abs-array. Table 1 gives information about the set of pathways that are common to all analyzed samples. The list below shows the list of common pathways obtained from KEGG database:

1. Genetic Information Processing (three out of twelve)
2. Cellular Processing (four out of twelve)
3. Environmental Information Processing (two out of twelve)

4. Metabolism (two out of twelve)

5. Human disease specific pathway (one pathway).

The arrangement of proteins on the basis of their KEGG pathway constitute the background knowledge for integrative bioinformatic analysis. The VANESA tool has been used to model and visualize the pathways by using the results of microarray proteomic profiles of analyzed samples. The results are summarized in Table 2 and Figure 4 illustrates the Tight junction signaling pathway with the multiple PDZ domain protein as a relevant protein from the first sample (S1). In addition to the general pathway, the network is enriched with sample specific, direct protein-protein interactions by mining specific database information of the data warehouse CardioVINEdb. VANESA has been able to predict sample specific pathways: the Tight junction signaling pathway and the Regulation of the actin cytoskeleton in sample 2 and the Calcium signaling pathway in sample 3.

Finally, the high throughput proteomic data has revealed a general framework of functional deregulation associated with DCM in the presence of concomitant pathologies. Our integrative approach can be helpful for molecular medicine decision-making. In a personalized medicine perspective, the identification of functional elements in their metabolic context is important in order to be able to select the most appropriate therapy for a specific patient.

Sample	**Number of biunivocal correspondences single pathway-single protein**	**Associated pathology**
S1	23/36	Dilatative cardiomyopathy with renal insufficiency
S2	24/36	Dilatative cardiomyopathy and Type 2 diabetes
S3	23/36	Dilatative cardiomyopathy with renal insufficiency and pulmonary disease

Table 1: The table outlines for each sample the relative number of pathways in which a single dysregulated protein is involved.

7 Summary

Data integration is one of the major problems in bioinformatics that primarily addresses the heterogeneity and increasing volume of information in biological databases. Based on life science data from different molecular biology databases and microarray experiments we constructed a comprehensive system for integration, modeling and analysis of biological networks. This software system consists of BioDWH data warehouse infrastructure for data integration, CardioVINEdb information system for web-based data access and VANESA for modeling and analyzing biological network data.

BioDWH is a Java-based open source toolkit for building life science data warehouses using common relational database management systems. Based on object-relational mapping technology, most relational database management systems can be used for local data storage.

Common perturbed pathways	Nr	Protein names	Swissprot code
hsa00350 Tyrosin metabolism	1	catechol-O-methyltransferase	P21964
hsa03022 Basal Transcription factor	2	general transcription factor II, i general transcription factor IIB	O15359 Q00403
hsa03060 Protein Export	1	signal recognition particle 54kDa	P13624
hsa04010 MAPK signalin pathway	1	arrestin, beta 1	P49407
hsa04110 Cell cycle	4	cyclin A1 MCM6 minichromosome maintenance deficient 6 (MIS5 homolog, S. pombe) (S. cerevisiae) cyclin-dependent kinase 7 (MO15 homolog, Xenopus laevis, cdk-activating kinase) cyclin-dependent kinase inhibitor 1C (p57, Kip2)	P20248 Q14566 P50613 P49918
hsa4120 Ubiquitin mediated proteolysis	1	ubiquitin-conjugating enzyme E2I (UBC9 homolog, yeast)	P50550
hsa04510 Focal adhesion	1	caveolin 1, caveolae protein, 22kDa	Q03135
hsa04612 Antigen processing and presentation	2	heat shock 90kDa protein 1, alpha calnexin	P07900 P27824
hsa05222 Small cell lung cancer	1	TNF receptor-associated factor 4	Q14848
hsa00860 Porphyrin and chlorophyll metabolism	1	heme oxygenase (decycling) 1	P09601
hsa04080 Neuroactive ligand receptor interaction	1	nuclear receptor subfamily 3, group C, member 1 (glucocorticoid receptor)	P04150
hsa04910 Insulin signaling pathway	1	flotillin 2	Q14254

Table 2: The table gives an outline of the names and of the Swissprot codes of the proteins, screened by Abs-array, involved in the common KEGG functional networks.

Moreover, BioDWH provides a number of ready-to-use parsers to extract data from public data sources and to store the content in a data warehouse.

CardioVINEdb provides integrated data from different popular life science databases and microarray data related to cardiovascular diseases from an EU project in a homogeneous web-based system. For building the data warehouse our integration toolkit BioDWH was used. Additionally, the CardioVINEdb information system enables intuitive search of integrated life science data, simple navigation to related information in addition to the visualization of biological domains and their relationships.

To efficiently explore and compare the heterogeneous datasets provided by the CardioVINEdb system, the software application VANESA was implemented. VANESA provides new bioinformatics methods and visualization approaches to analyze dynamic interacting networks. The data from the CardioVINEdb system is analyzed on a large scale and visualized in a biologically meaningful way. An important aspect of the visualization is the consideration of multi-dimensional data annotations in a way suitable for the knowledge discovery process. With the software application it was possible to trim down the data to a manageable yet relevant size and to analyze and identify new as well as altered versions of interaction patterns.

The application of an integrative bioinformatics approach on high throughput proteomic data has is shown, for the first time, a general framework of functional dysregulations associated with DCM in the presence of concomitant pathologies. Our proposed integrative approach can be a helpful support to clinicians in selecting a more appropriate therapy for a patient, taking a wide range of metabolisms into account.

References

[1] M. Y. Galperin and G. R. Cochrane. Nucleic Acids Research annual Database Issue and the NAR online Molecular Biology Database Collection in 2009.
Nucleic Acids Research, 37(Database issue):D1-D4, 2009.

[2] T. Etzold, A. Ulyanov, and P. Argos. SRS: Information retrieval system for molecular biology data banks.
Methods in Enzymology, 266:114128, 1996.

[3] 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 science data sources.
IBM Systems Journal, 40(2):489511, 2001.

[4] J. Baumbach. CoryneRegNet 4.0 - A reference database for corynebacterial gene regulatory networks.
BMC Bioinformatics, 8:429, 2007.

[5] J. Köhler, J. Baumbach, J. Taubert, M. Specht, A. Skusa, A. Rueegg, C. Rawlings, P. Verrier, and S. Philippi. Graph-based analysis and visualization of experimental results with ondex.
Bioinformatics, 22(11):13831390, 2006.

[6] S. P. Shah, Y. Huang, T. Xu, M. M. S. Yuen, J. Ling, and B. F. F. Ouellette. Atlas - a data warehouse for integrative bioinformatics.
BMC Bioinformatics, 6:34, 2005.

[7] S. Durinck, Y. Moreau, A. Kasprzyk, S. Davis, B. D. Moor, A. Brazma and W. Huber. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis.
Bioinformatics, 21(16):3439-40, 2005

[8] 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, 7:170, 2006.

[9] S. Tril, K. Rother, H. Müller, T. Steinke, I. Koch, R. Preissner, C. Frömmel, and U. Leser. Columba: an integrated database of proteins, structures, and annotations.
BMC Bioinformatics, 6:81, 2005.

[10] C. C. Choi, R. Münch, S. Leupold, J. Klein, I. Siegel, B. Thielen, B. Benkert, M. Kucklick, M. Schobert, J. Barthelmes, C. Ebeling, I. Haddad, M. Scheer, A. Grote, K. Hiller, B. Bunk, K. Schreiber, I. Retter, D. Schomburg, and D. Jahn. SYSTOMONAS - an integrated database for systems biology analysis of Pseudomonas.
Nucleic Acids Research, 35(Database issue):D533-D537, 2007.

[11] 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):93, 2008.
Online Journal: `http://journal.imbio.de/index.php?paper_id=93`

[12] S. Hariharaputran, T. Töpel, B. Brockschmidt and R. Hofestädt. VINEdb: a data warehouse for integration and interactive exploration of life science data.
Journal of Integrative Bioinformatics, 4(3):63, 2007.
Online Journal: `http://journal.imbio.de/index.php?paper_id=63`

[13] 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.
Online Journal: `http://www.bioinfo.de/isb/2010/10/0003/`