Visualization and Analysis of a Cardio Vascular Disease- and MUPP1-related Biological Network combining Text Mining and Data Warehouse Approaches

Björn Sommer[1*], Evgeny S. Tiys[2], Benjamin Kormeier[1], Klaus Hippe[1], Sebastian J. Janowski[1], Timofey V. Ivanisenko[2], Anatoly O. Bragin[2], Patrizio Arrigo[3], Pavel S. Demenkov[4], Alexey V. Kochetov[2], Vladimir A. Ivanisenko[2], Nikolay A. Kolchanov[2], Ralf Hofestädt[1]

[1]Bio-/Medical Informatics Department, Bielefeld University, Universitätsstraße 25, 33615 Bielefeld, Germany
[2] Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, Lavrentyeva 10, 630090 Novosibirsk, Russia
[3] CNR ISMAC, Via De Marini 6, Genoa, Italy
[4] Sobolev Institute of Mathematics, Siberian Branch of the Russian Academy of Sciences, 4 Acad. Koptyug avenue, 630090 Novosibirsk, Russia

Summary

Detailed investigation of socially important diseases with modern experimental methods has resulted in the generation of large volume of valuable data. However, analysis and interpretation of this data needs application of efficient computational techniques and systems biology approaches. In particular, the techniques allowing the reconstruction of associative networks of various biological objects and events can be useful. In this publication, the combination of different techniques to create such a network associated with an abstract cell environment is discussed in order to gain insights into the functional as well as spatial interrelationships. It is shown that experimentally gained knowledge enriched with data warehouse content and text mining data can be used for the reconstruction and localization of a cardiovascular disease developing network beginning with MUPP1/MPDZ (multi-PDZ domain protein).

1 Introduction

More than 4000 human diseases are known and defined [1]. Regarding the medical characteristics or main features one can see that any disease is defined or specified by particular symptoms and/or laboratory parameters. In practice the diagnosis problem is based on the fact that a lot of symptoms, such as fever, are related to many diseases. Therefore, the diagnostic procedure will always be a differential process which will produce a set of possible diseases. Furthermore, the so-called personalized medicine makes the problem of finding the patient-relevant diagnosis and recommendation for its treatment much more difficult. Based on the data of molecular biology, the development of new and more efficient tools for medical diagnosis and therapy process is becoming possible.

Today, more and more diseases can be reduced to simple metabolic processes, which more or less are based on mutations in related genes. OMIM [2] examplifies of well-known information systems which exactly represent this kind of knowledge. Overall, there are more than 1000 molecular database and information systems which represent various molecular and phenotypic data. These information resources were designed on the basis of either automatic data extraction or manual

[*] To whom correspondence should be addressed. E-mail: bjoern@CELLmicrocosmos.org

annotation and curation. Behind these information systems there is one more specific and powerful information system which will present molecular and medical disease knowledge. The MEDLINE information system represents all relevant publications (abstracts and in the near future a complete listing of papers) which are relevant for molecular medicine or biomedicine. Overall one have access to more than 1000 powerful database and information systems which will help identify molecular knowledge about any disease. Furthermore, this data can be supported, enriched or fused by the extension of text and data mining techniques which allow the automatic extraction of medical and molecular knowledge from the PubMed system, which includes all relevant scientific results. Therefore, it is possible to construct or predict the metabolic network for any disease. This kind of work is relatively new and during the last years different database integration and data mining systems have been implemented. However, the problem of all these systems is, that data integration and mining tools will produce networks, which are too complex. Therefore, the development of special filter systems or visualization tools is a necessary step in understanding and analyzing these complex metabolic disease-related networks. In this paper it will be demonstrated how the data integration and data mining tools can be used to gather the molecular knowledge on diseases.

The focus of this application is on Cardiovascular diseases (CVDs), and more precisely the dilated cardiomyopathy, which is the leading cause of death in developed countries. Based on the experimental identification of a CVD relevant protein, two protein-protein interaction networks were constructed by using the network visualization and analysis tool VANESA [3] and the text mining tool ANDVisio [4], which is also able to identify the localization of network components. This localization information was extended, combined with the created networks and finally visualized in 3D by the CELLmicrocosmos 4.2 PathwayIntegration (CmPI) [5].

2 Basics

2.1 Metabolic disease networks

Much attention has been recently focused on the metabolic aspects of Cardiovascular diseases (CVDs). The discovery of new CVDs specific molecular targets promoted the investigation of proteins functional roles in their specific pathways. It is quite complex to evaluate the weight of each trigger factor (metabolism, hormones, exogenous factors, etc.) on CVDs emergence. Epidemiological studies constitute the starting point for molecular medicine screening. The advent of high throughput analytical techniques (DNA chip, protein arrays, molecular imaging) has improved the capability to screen new candidate target proteins (genes). The relations of metabolic pathways of a sample coming from patients affected with dilated cardiomyopathy (DCM) was the basis of study for this publication. The proteome analysis is based on experimental data on which integrative bioinformatics approaches have been applied to characterize a specific functional pathway deregulated in the pathological sample. In this study, the combination of data warehouse with text mining approaches is demonstrated by using different software applications.

2.2 Data integration

Since industrial research of molecular biology questions starting with the Human Genome Project, one of the main challenges in bioinformatics is the integration of molecular data. Today high throughput analysis delivers data of complete genomes, for instance short sequences of all genes in an organism or expression patterns of thousands of a cell in shortest time. Analysis of these high throughput data by manual investigation using publications or relevant databases is no longer

possible. Consequently, biologist has to be supported by tools and methods that can accumulate experimental data with complementary data sources, estimate the data and compare or classify these data. This challenge leads us to the problem of database integration.

Typically, data of genomes, genes, proteins, enzymes, chemical compounds, diseases, etcetera is stored in databases with worldwide availability. A good overview of important databases is provided by the annual special issue of Nucleic Acids Research [6]. The number of molecular databases is continuously increasing in the last decade. Molecular biological data has a high semantic heterogeneity that is caused by (experimental) data extracted from a series of experiments. Molecular biology deals with complex problems, hence enormous and versatile data is produced. The total number of databases, as well the data itself, is continuously increasing, as is the distribution and heterogeneity of the data. Particularly, data heterogeneity causes big problems in molecular biological data integration. Technical heterogeneity is caused by a high number of different formats and interfaces of the different data sources. Furthermore, the data is usually not available in a standard format which causes structural heterogeneity. Moreover, there is a level of semantic heterogeneity, because of missing standards and consensus for basic biological terms. In addition to the problems of molecular biological databases there are some more in data integration. Usually, data sources of an integrated system are distributed. That means, each and every source is located on separate systems and different locations. The distribution of several data sources leads automatically to the problem of autonomy. Regarding data integration, autonomy means independence of the data source that refers to access, configuration, development and administration.

The major problem of data integration is heterogeneity that is caused by autonomy. Moreover, distribution can also cause heterogeneity, but not generally. The development of an integrated database system is a complex task. Particularly, if a large number of heterogeneous databases have to be integrated. Data warehouses (DWH) are one of the widely used structures for database integration. For that purpose a software infrastructure for building life science data warehouses using different common relational database management systems is introduced. The BioDWH [7] system is realized as a Java-based open source application that is supported on different platforms with an installed Java Runtime Environment (JRE). BioDWH is a flexible DWH infrastructure for bioinformatics; it is independent from the underlying RDBMS. Furthermore, the data warehouse approach provides an easy-to-use graphical user interface for administration and configuration. The main feature of the BioDWH tool-kit is the automatic storage and visualization of data content and information from different public databases into a homogeneous and consistent data warehouse. It provides integrated data from different widely-used life science databases, such as BRENDA [8], EMBL [9], ENZYME [10], GO [11], HPRD [12], KEGG [13], OMIM [2], Reactome [14], SCOP [15], Transfac [16], Transpath [17] and UniProt [18] and microarray data. Additionally, configuration of the infrastructure and its tools is also possible via XML, because it is human readable, well-formatted, easy accessible and standardized. A logging mechanism observes the integration process and begins a simple recovery process to guarantee a consistent state of the data warehouse. The data warehouse BioDWH addresses the aforementioned aspects of data integration.

Based on the data from the warehouse infrastructure, the CardioVINEdb [19], a data warehouse approach, was developed to browse and explore life science data. Furthermore, a DWH system to search integrated life science data and simple navigation called DAWIS-M.D. was implemented based on the life science data from the BioDWH toolkit. In addition, the network editor VANESA uses the data from DAWIS-M.D to generate biological networks and enrich them with additional information.

2.3 Text mining

Work with scientific literature and factual databases is required for research in every knowledge area. The size of this information pool is immense and expands exponentially. The PubMed database alone contains over 19 million abstracts [20] and their number increases annually by 1 million. Thus, the development of computer algorithms for automated text analysis (text-mining) becomes a timely task.

In microbiology and biomedicine the most important type of interactions is molecular-genetic. Extraction of facts concerning such interactions from literature and providing access to them can be divided in two directions: manual analysis of literature and automated data analysis with text-mining techniques. Manual curation is most accurate but also is a highly time-consuming task [21]. The automated data analysis methods are not so accurate but allow the processing of larger amounts of data in less time and usually are used in three main tasks: extraction of data on molecular-genetic interactions between biological objects, discovery of new associations between different sources of biological information and biological data classification.

For the automated extraction of molecular-genetic interactions between biological objects and new associations from the literature various methods exist:

- methods based on the co-occurrence of objects,
- natural language methods based on the deep syntactic analysis of single sentences (full parsing),
- natural language methods based on rules and templates analysis (shallow parsing).

Co-occurrence is based on the statistically important values of the joint frequency of names of biological objects in texts. The main advantage of this method is that it is easy to implement and achieves good results in regards to search completeness, but it is not so accurate. Moreover such approach does not allow detection of different parameters of interconnections between objects, such as type of interaction and its direction. PUBGENE [22] and FACTA [23] are based on this approach. In BioGene [24] it was implemented for prioritization of genes.

“Full-parsing” is based on the definition of the language with formal grammar. There are many various types of grammars as well as descriptions of the complete sentence structure. The main limitation of this approach is its low time efficiency so it cannot be used for all tasks. MedScan from the PathwayStudio [25], GeneScene [26] are examples of a full parsing-based system, also it has been used by Fayruzov [27] for protein relation extraction. A text analysis algorithm based on formal grammar implemented in this system shows high accuracy but is also very time-consuming.

“Shallow parsing” (deep parsing) is based on the extraction of information from sentences by using the partial connection between words in the sentence with the help of specified rules and templates. A SUISEKI system is based on it [28]. In the Chilibot system [29], the deep parsing method was implemented for the classification of extracted proteins (genes) from PubMed abstracts. The relations between two proteins may result from the existence or non-existence of an interaction or co-location.

The biological data classification task is based on the idea of classification of various sources of data by user-specified features. For the solution of this task, different algorithms are applicable, such as hidden Markov models [30] or Bayesian networks. The BioBayesNet server [31], based on Bayesian networks can be used for such classification.

Most of the modern text-mining systems are combining various methods. The ANDVisio, ALI BABA [32] and PolySearch [33] systems are based on the combination of co-ocurrence and shallow-parsing methods. The co-occurrence method is used for mapping of the biological objects