KnoFusius – A New Knowledge Fusion System for Interpretation of Gene Expression Data

Pavel Smrž
Faculty of Information Technology, Brno University of Technology
Bozetechova 2, 61266 Brno, Czech Republic
E-mail: smrz@fit.vutbr.cz

Abstract

This paper introduces a new architecture that aims at combining molecular biology data with information automatically extracted from relevant scientific literature (using text mining techniques on PubMed abstracts and fulltext papers) to help biomedical experts to interpret experimental results in hand. The infrastructural level bears on semantic-web technologies and standards that facilitate the actual fusion of the multi-source knowledge.

1. Introduction

A microarray (or a gene chip/array) is a collection of microscopic spots on a solid surface which enables monitoring expressions of thousands genes (virtually the entire genome) simultaneously. The term "gene expression" refers to turning on and off the production of proteins by which a given organism responds to environmental and biological situations. Microarrays play a significant role in today's biomedicine. They have been successfully used in diagnoses and prognoses for various diseases, to plan treatment, to design new drugs, etc.

The amount of data produced by microarray analysis is large and it is not possible to analyze it manually. The gene expression matrix can contain a lot of noise, missing values or other irregular variations. Advanced statistical or machine learning methods are applied to normalize data and to identify and correct unacceptable expression values (see, e.g., (Yang et al., 2002, Hershey et al., 2008) for approaches to microarray data normalization and (Yoon et al., 2007, Wang et al., 2006) for missing value estimation). There is also a need for management of the huge amount of diverse data. The heterogeneity can present the major obstacle for the fusion mechanism as data from different laboratories are produced by widely varying experimental techniques and can be incomplete in many respects.

The next step in the automatic processing of microarray data consists in grouping genes with similar behavior (and, usually, filtering out the rest). Various clustering algorithms have been applied for identifying biologically relevant groups of genes and samples. A survey of the clustering methods used for gene expression data can be found in (Jiang et al., 2004).

The mentioned processes form the general part of present standard microarray data processing which is common for various experiments run by laboratories all over the world. Many biologists work directly with the output of the employed clustering techniques and try to find a connection between the current measurements and their knowledge on the domain. Advanced visualization techniques can be called for to support the interpretation process. Nevertheless, it is usually extremely difficult to find the proper links between the potentially significant genes from the produced clusters and to formulate and verify correct hypotheses on the results. Even the experts with a long experience in the field often miss significant clues that would support the interpretation. They can fail to remember an article on the topic presenting the evidence in another context, not be aware of the latest results studying another organism, miss a paper on a discovered protein interaction patterns etc. Without a help of information technologies, the interpretation presents a tedious work with many potential obstacles.

The aim of our research is to reduce the workload as much as possible and let biologists focus on the interpretation of the particular pieces of knowledge the system can automatically infer from available data. Even though the deep biological knowledge will be probably always necessary to uncover new principles of the processes in living organisms, today's computers can significantly reduce the need for manual search of relevant scientific literature, the evidence from previous experiments and additional information.

As there are potentially many sources of relevant biological knowledge, one has to think about general ways to integrate the different views provided by the different databases. The actual fusion mechanism we apply in our work takes advantage of the latest semantic-web technologies and standards that facilitate the definition of the data-integration semantics. KnoFusius benefits from RDF/OWL-based data communication and interchange and defines open interfaces for later extensions. In this respect, we continue the work of Ruttenberg et al. (2007), Badea (2006) and many others that show the significance of the semantic web vision for the biomedical research. An important trend that helps us to speed up the development of the tool is "the opening" of the large biomedical databases. More and more sites are changing from former data silos to fully interoperable services and they can be therefore easily plugged into the semantics-aware platforms.

2. System architecture

A schema of the KnoFusius system is given in Figure 1. The process starts with experimental data prepared with the help of advanced biomedical methods and tools. The user can provide a normalized description of the experimental setup which can be then used to retrieve additional data from various databases. The quality of metadata is crucial for the success of subsequent processing steps.

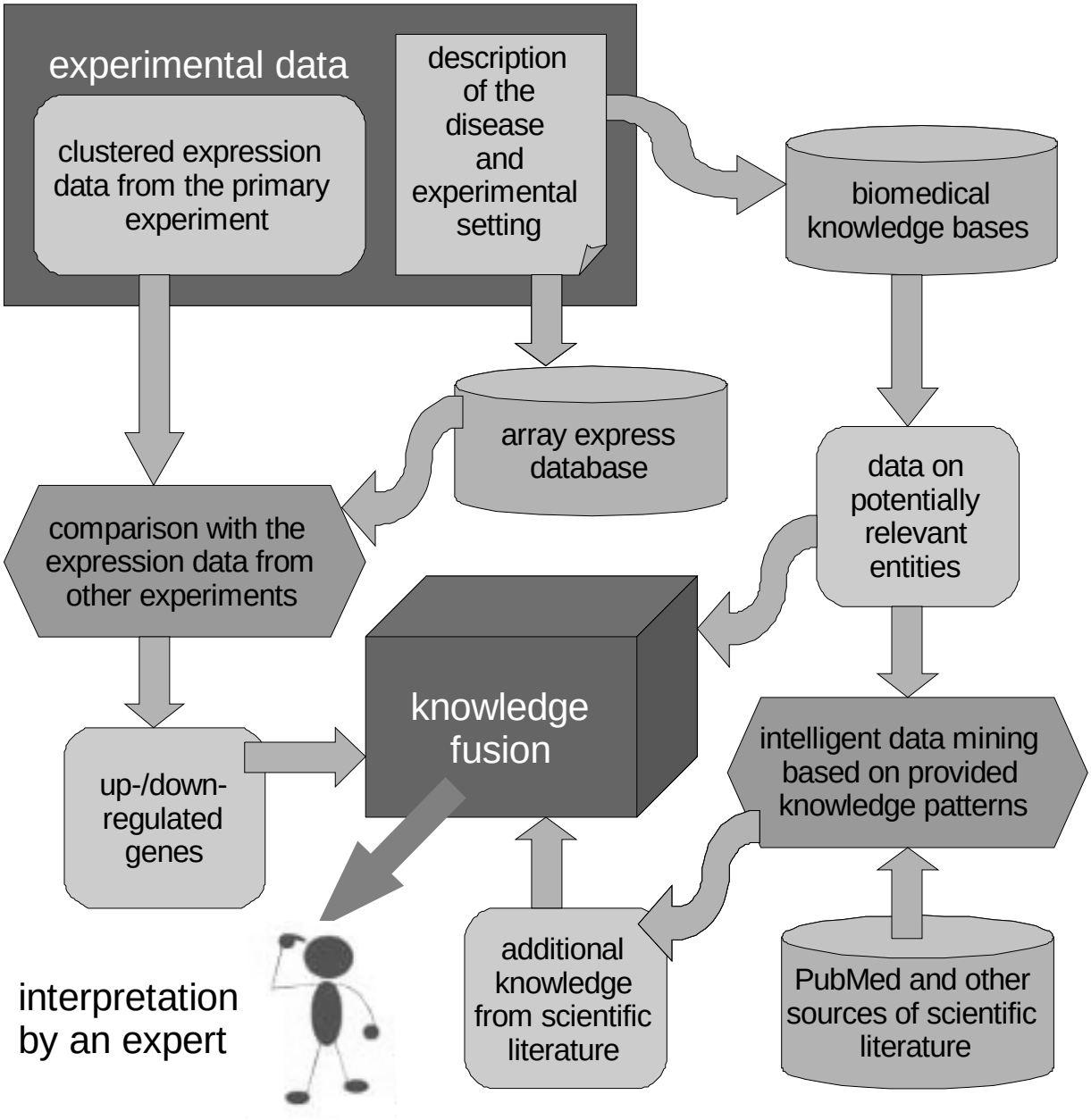

Figure 1. Processing schema of KnoFusius

Array express databases containing the results of other groups around the world are searched next. As it is extremely difficult (if not impossible) to compare the primary expression data across various experimental settings, arrays used etc., the system counts upon metainformation, provided by the original experimenters and stored together with the primary data in the array expression databases. We are currently trying to provide wrapper components that should enable combining data from two most populated databases – ArrayExpress (www.arrayexpress.com) and STNK (www.stanford.edu). The fusion on this level is rather problematic as the two databases differ significantly in their content as well as the functions supported.

The experimental data are then combined with relevant information from biomedical knowledge bases. They include various ontologies such as GO – the Gene Ontology (www.geneontology.org) or OBI – the Ontology for Biomedical Investigations (obi.sourceforge.net), pathway maps (representing the knowledge on the molecular interaction and reaction networks) such as KEGG – the Kyoto Encyclopedia of Genes and Genomes Pathway collection (www.genome.jp/kegg/pathway.html), Biocarta (biocarta.com) or the BioCyc collection (biocyc.org), protein knowledge bases such as UniProt – the Universal Protein Resource (uniprot.org) and many other resources.

Even though the biomedical knowledge bases try to scan journals and conference proceedings regularly to embrace as much information as possible, one still cannot rely on their full coverage. This is partially due to the shallow text analysis techniques employed and also due to the limited scope of the primary resources. Moreover, those knowledge bases that are "curated" by an individual or a small group of people have to tackle the issues of subjectivity and availability of the curators. On the other hand, the approach followed in our work reduces the work of personal judges to the definition of a declarative set of

extraction patterns for particular pieces of knowledge, and, if necessary, to semi-automatic evaluation of the source reliability. The text mining is applied not only to the content of the PubMed database, but also to the additional sources of scientific publications that can be stored locally (recent conference proceedings, various reports with restricted access rights…). An important point of the direct analysis of the scientific articles and papers (instead of taking benefit just from the pre-processed biomedical databases) is our ability to consider different weights (influence, reliability) of various pieces of information from various sources. The reasoning within the knowledge fusion system deals with explicitly represented uncertainty (see (Novacek & Smrz, 2006) for the details of our approach) and the source reliability is one of the important factors participating in the process.

As mentioned above, we employ rather deep text processing to extract as much relevant information as possible. After the standard preprocessing steps – transformation of the input formats, tokenization and sentence boundary detection, we employ POS tagging, syntactic analysis and pronominal anaphora resolution. We benefit from the available domain-specific terminological thesauri and ontologies to define particular categories of interest. The results form an input for our pattern-based semantic-relation extractor. It takes advantage of general-purpose language resources, namely WordNet, to expand pre-defined knowledge patterns (and transfer terms to concepts in general). The set of extracted relations (such as "protein-A inhibits protein-B") is then merged with the related information from biomedical knowledge bases and the output is used to filter and interpret the experimental data in hand.

A significant attention has been recently paid to the aggregation and integration of data drawn from diverse sources in the field of life sciences. A unifying view on these activities can be provided by the vision of the semantic web – an extension of the current web that enables automatic processing of the various resources. It is based on common formats (RDF, OWL, RIF…) and related technologies. For example, the above-mentioned knowledge bases have been recently transformed from many proprietary formats (often focusing on the visual representation suitable for humans) into RDF/OWL appropriate for machine processing.

There are many limitations of the current semantic web technologies due to their immaturity. The major issue connected to the huge knowledge bases and complex ontologies typical for the biomedical field is the low performance and limited scalability of the available automatic reasoners. That is why we currently employ ad-hoc mechanisms for the interpretation of experimental data based on a simple fuzzy-rule chaining. However, as the overall architecture is modular enough to allow easy replacement of the inferring engine, we plan to evaluate various recently proposed solutions (Stracia, 2001, Simou & Kollias, 2007) in terms of their performance and scalability and to integrate the module that will best meet our needs.

3. Experiments and results

It is always difficult to evaluate systems aiming at finding new, potentially interesting links among given pieces of knowledge that should help to understand a complex problem. To evaluate the functionality of KnoFusius, we defined two tasks that are described in the following text.

The first set of input data consists of two lists of gene identifiers – 22 genes that showed to be up-regulated and 29 down-regulated in an experiment. No additional knowledge on the disease in question, experimental settings or other characteristics has been provided to the system. The test should show whether the system is able to find appropriate hints to identify the type of disease.

KnoFusius calls FABLE (Fast Automated Biomedical Literature Extraction – fable.chop.edu) to locate papers dealing with the particular genes. The service returns PubMed identifiers. A local copy of the PubMed archive indexed by Lucene (lucene.apache.org) is called next. This configuration allows us to evaluate even complex queries in a short time. The system tries to identify articles containing one or more genes from the given sets. Table 1 shows an excerpt of such an output.

set(['TCF7', 'ZAP70']) – 15325098 Different gene expression in immunoglobulin-mutated and immunoglobulin-unmutated forms of chronic lymphocytic leukemia.
set(['EGR1', 'TNFRSF1A']) – 14735464 Antisense abrogation of DENN expression induces apoptosis of leukemia cells in vitro, causes tumor regression in vivo and alters the transcription of genes involved in apoptosis and the cell cycle.
set(['ABCA6', 'FMOD']) – 12651908 Identification of a global gene expression signature of B-chronic lymphocytic leukemia.

Table 1: A list of papers containing combinations of genes from the first dataset.

Next, KnoFusius extracts the relevant GeneRIFs (Gene Reference Into Function) from NCBI (ncbi.nlm.nih.gov). It computes an intersection of the functional annotation and the most frequent terms from the articles obtained in the previous step.

Table 2 shows an example of the resulting list for our testing data. The most relevant terms include "leukemia" – the disease that was really confirmed in the examined patients. It shows that KnoFusius is potentially able to provide hints for the expert interpretation of the microarray measurement based on the term extraction.

The second test set deals with human diabetes. It comes from http://www.broad.mit.edu/mpg/oxphos and contains 43 microarray samples containing 22,283 genes taken in skeletal muscle biopsy from males (17 with normal glucose tolerance (NGT), 8 with impaired glucose tolerance (IGT) and 18 with type 2 diabetes mellitus (DM2)). GEPAS (Gene Expression Profile Analysis Suite – gepas.bioinfo.cipf.es/) has been used to preprocess data. KnoFusius is called to help biologists find the best match between their measurements and knowledge stored in the form of published pathway databases. The pathways are

Likelihood	OBO terms
32.283984	Leukemia, Lymphocytic, Chronic
26.351238	Leukemia of unspecified cell type
18.497584	Leukemia
12.519245	Precursor Lymphoblastic Lymphoma
8.417293	Megakaryocytic leukemia
7.956313	Urticaria
7.031021	Multiple Sclerosis
6.549814	Demyelinating Diseases

Table 2: The most relevant terms for the first dataset.

taken as graphs (see Figure 2) stored in a local database which combines data from KEGG and Biocarta sites mentioned above. Results are presented in the form of pathway lists ordered according to their estimated relevance to the experimental data on input.

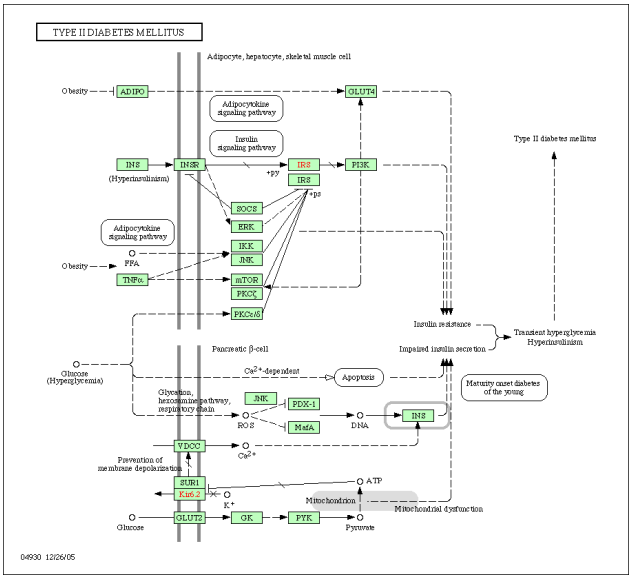

Figure 1: Type II Diabetes Mellitus Pathway from KEGG

Table 3 demonstrates a part of the results on the second set. Biological assessment confirmed that such an output can significantly improve the interpretability of microarray data and potentially lead to finding new facts on various diseases.

Pathway name	From	Score
Phosphatidylinositol signaling system	KEGG	0.3966
Insulin signaling pathway	Biocarta	0.3121
Melanogenesis	KEGG	0.2895
Maturity onset diabetes of the young	KEGG	0.1374

Table 3: A list of pathways related to the second dataset.

4. Conclusions and Future Directions

Despite recent efforts to overcome the fragmented nature of biomedical knowledge on the current web, the problem of information fusion from various resources has not been solved to a sufficient extent till now. The presented work can be seen as our contribution to this research. The modular architecture enables easy integration of various components and methods and the semantic web context simplifies the data integration procedures.

A lot of work needs to be done to realize the full potential of KnoFusius. We will focus on "opening" all the tools by means of web services that will make them available to a broad community of interested users. We will also pay attention to the input part of the system that currently suffers from a relatively high number of unidentified gene IDs. Last but not least, we are going to improve the user interface of the system accordingly to the feedback from its real use.

5. Acknowledgements

This work was supported by the Czech Ministry of Education, research grant 2B06052.

6. References

Badea L. Semantic web reasoning for analyzing gene expression profiles. Proc. Principles and Practice of Semantic Web Reasoning, PPSWR, LNCS 4187, pp. 78-89, Springer Verlag, 2006.

Hershey A. D., Burdine D., Liu C., Nick T. G. , Gilbert D. L. Glauser T. A. Assessing quality and normalization of microarrays: case studies using neurological genomic data. Acta Neurol Scand, 2008.

Jiang, D., Tang, C., Zhang, A. Cluster analysis for gene expression data: A survey. IEEE TKDE, 16(11), 2004.

Nováček, V., Smrž, P. Empirical merging of ontologies: A proposal of universal uncertainty representation framework. In: Proceedings of ESWC, 2006, pp. 65-79.

Ruttenberg, A. et al. Advancing translational research with the Semantic Web, BMC Bioinformatics, 8(3), 2007.

Simou, N, Kollias, S. FiRE: A fuzzy reasoning engine for impecise knowledge, K-Space PhD Students Workshop, Berlin, Germany, 2007.

Straccia, U. Reasoning within fuzzy description logics. Journal of Artificial Intelligence Research, 14:137–166, 2001.

Wang X., Li A., Jiang Z., Feng H. Missing value estimation for DNA microarray gene expression data by Support Vector Regression imputation and orthogonal coding scheme. BMC Bioinformatics, 7:32, 2006.

Yang Y. H. , Dudoit S., Luu P., Lin D. M., Peng V., Ngai J., Speed T. P. Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res., 30(4), 2002.

Yoon D., Lee E. K., Park T. Robust imputation method for missing values in microarray data BMC Bioinformatics 8, Suppl 2, 2007