Gene Expression Data Mining Guided by Genomic Background Knowledge

Pavel Smrž[1], Jana Šilhavá[1], Jiří Kléma[2], and Filip Železný[2]

[1] Faculty of Information Technology, Brno University of Technology
Božetěchova 2, 612 66 Brno, Czech Republic
E-mail: {silhava,smrz}@fit.vutbr.cz
[2] Department of Cybernetics, Czech Technical University in Prague
Technická 2, 166 27 Praha 6, Czech Republic
E-mail: {klema,zelezny}@labe.felk.cvut.cz

Abstract. Microarray data represents valuable information resources, nevertheless the knowledge is hidden inside the data and it is not easy to mine. Background knowledge is also stored in various formats and it is challenging to automatically infer the biological meaning from existing repositories. This paper deals with a new gene-expression knowledge-fusion system that combines molecular biology data from various sources — the experiment in hand, gene expression data from similar experiments stored in array expression databases, additional knowledge on the most significant genes and their products from specialised services (e.g., pathway databases), and automatically derived results provided by relevant scientific literature. The design of the proposed system is rather complex. We take advantage of recent semantic web technologies to integrate the various modules of the system. Some of the components described in the paper have already taken part in the end-user applications, others still wait for their implementation in the form of software tools.

1 Introduction

Gene chips or microarrays enable monitoring the expression of tens of thousands genes (virtually the entire genome) simultaneously. They play a significant role in today's biomedicine as they improve diagnosis and prognosis, plan treatment or drug design.

The term "gene expression" means turning on and off the production of proteins by which a given organism responds to enviromental and biological situations. Genes are contained in DNA. Proteins are produced from the corresponding genes in a two-step process. The first step consists in the transcription of the gene from DNA into RNA. Then, RNA is further translated into a corresponding protein. A crucial property of DNA and RNA is the complementarity. The advantage of complementarity lies in detecting specific sequences of bases within strands of DNA and RNA. In theory, it is done by synthesizing a probe, a piece of DNA or RNA that is the reverse complement in the sample. The act of binding between probe and sample is called hybridization. DNA probe

technology has been adapted for detection of, not just a sequence, but tens of thousands sequences simultaneously. This is done by synthesizing a large number of different probes and their placing at a specific position on a glass slide (a spotted array) or by attaching the probes to specific positions on some surface. The crucial steps in the processing of a microarray consist in labeling samples with fluorescent dyes or radioactive isotopes, hybridization, washing to remove non-specific binding, scanning and data analysis. The gene expression matrix is obtained by scanning and several image analysis algorithms including segmentation and registration. The same set of genes can be measured under various circumstances or at various time points. The measured expression values can be organized in several ways. Gene expression data is stored in tables where rows (columns) represent genes and columns (rows) represent experimental conditions. The number associated with array items represents an expression level of a specific gene under specific conditions.

DNA microarrays come in several different types. The most common are Affymetrix arrays (GeneChips), spotted oligonucleotide arrays and spotted cDNA arrays. Each microarray has a unique layout. So-called gene list describes the configuration of a particular array. It is critical to identify the right layout as any error could lead to a total misinterpretation of the results.

The amount of data produced by microarray analysis is large and it is not possible to analyze it manually. Thus, there is a need for automated processing. The gene expression matrix contains expression numbers covering noise, missing values, arising nonsystematic variations. The data is further processed with the aim to normalize the scale and location. Various methods to identify unacceptable expressed genes or corrective procedures are also applied (see, e.g., [1], [2], [3] for the approaches to data normalization and [4] for missing value estimation). There is also a need for the managing the huge amounts of diverse data. Heterogenous data are produced by different labs using widely different experimental techniques. Data could contain wrong values or be incomplete. The correct diagnosis is vital as each patient is unique. Consequently, there are difficulties in mircoarray data interpretation and comparing. Additional available knowledge in data analysis has to be used to obtain more accurate results or to make it easier to find similar cases. Extracted knowledge should be in an understable form. Microarray data represents valuable resources for therapeutic process, however, the knowledge and information is not usually easy to mine because it is hidden inside the data. Background knowledge is stored in various formats and it is challenging to automatically infer the biological meaning from existing repositories.

After preprocessing steps, such as mentioned data filtering and normalization, the task is to choose the groups of genes that are significant for the current case. The principal target consists in improving the gene groups interpretability.

This paper deals with a new architecture of a gene-expression knowledge fusion system that combines molecular biology data from various sources — a particular experiment in question, gene expression data from similar experiments stored in array expression databases, additional knowledge on the most signifi-

cant genes and their products from specialised services (e.g., pathway databases) and automatically derived results provided by relevant scientific literature (using text mining techniques on PubMed abstracts and fulltext papers). To easy such a combination of heterogenous data, we employ semantic-web technologies and standards that provide the infrastructural level for the system.

The reader of this paper can find methods that can be used for gene groups selection in Section 2. Mining transcriptomic data is described in Section 3. Section 4 introduces major building blocks of our system from a conceptual point of view. The overall architecture of the integrated system as well as the fuctionality of particular components is discussed next. Section 6 concludes the paper and provides directions of our future research.

2 Gene Groups Selection

There can be as much as 20,000 spots on a microarray chip. The goal of the gene selection phase can be seen as eliminating redundancy in the resulting data. It can be done by means of clustering or unsupervised machine learning. The goal of clustering is to determine the groups in a set of unlabeled data. It is possible to find groups of experiments with similar gene expression profiles. Various clustering algorithms have been proven to be useful for identifying biologically relevant groups of genes and samples. A survey of clustering methods used for gene expression data can be found in [5]. Some of the usual clustering problems are that gene clusters are previously unknown, it is needed to choose distance function, the results of the clustering algorithm can be interpreted in different ways, cluster gene expression patterns are based on their similarities. Another known drawback which is highly relevant for the gene expression data is the sensitivity of the algorithms to noise. The data can be analysed from many different viewpoints. A connection data with some background knowledge can have influence on a selection of more significant clusters.

Supervised machine learning is employed to class prediction and gene selecton, based on gene expression profiles, generally. The information about classes is known (e.g. cases vs. controls) and the objective is to select genes differentially expressed or to try to predict class membership based on the corresponding gene expression profiles. Support vector machines (SVM) are widely used in this area. According to [6], SVM can provide near-prefect classification accuracy on a particular data set. Gene selection can also take advantage of information-theoretic methods [22].

The basic two-class hypothesis can be evaluated by the standard t-test:

$$t = \frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}, \tag{1}$$

where $\overline{Y}_1$ and $\overline{Y}_2$ represent sample means of data in each of the two classes, s_1 and s_2 are standard deviations for each class that are divided by the numbers of genes in each class N_1 and N_2.

Anova [23] can be used for multi-class cases.

3 Mining Transcriptomic Data

Gene-expression data analysis represents a difficult task as the data usually shows an inconvenient rate of samples (biological situations) and variables (genes). Datasets are often noisy and they contain a great part of variables irrelevant in the context under consideration. Independent of the platform and the analysis methods used, the result of a gene-expression experiment should be driven, annotated or at least verified against genomic background knowledge.

As an example, let us consider a list of genes found to be differentially expressed in different types of tissues. A common challenge faced by the researchers is to translate such gene lists into a better understanding of the underlying biological phenomena. Manual or semi-automated analysis of large-scale biological data sets typically requires biological experts with vast knowledge of many genes, to decipher the known biology accounting for genes with correlated experimental patterns. The goal is to identify the relevant functions, or the global cellular activities, at work in the experiment. Experts routinely scan gene expression clusters to see if any of the clusters are explained by a known biological function. Efficient interpretation of this data is challenging because the number and diversity of genes exceed the ability of any single researcher to track the complex relationships hidden in the data sets. However, much of the information relevant to the data is contained in the publicly available gene ontologies and annotations. Including this additional data as a direct knowledge source for any algorithmic strategy may greatly facilitate the analysis.

We emphasize the potential of genomic background knowledge stored in various formats such as free texts, ontologies, pathways, links among biological entities, etc. We deal with various ways in which heterogeneous background knowledge can be preprocessed and subsequently applied to improve various learning and data mining techniques. In particular, we focus on background knowledge in the following tasks:

- feature selection and construction (and its impact on classification accuracy);
- constraint-based knowledge discovery;
- quantitative association rule mining;
- relational descriptive analysis.

Improving Classification Accuracy with Background Knowledge The traditional attribute-valued classification searches for a mapping from unlabelled instances to discrete classes. When dealing with a large number of attributes and a small number of instances, the resulting classifier is likely to overfit the training data, a wide range of classifiers may show comparable testing performance and the

classifiers may be hardly explainable. In order to increase the predictive power of the classifier and its understandability, it is advisable to incorporate background knowledge into the learning process. In our previous work [15,16,17], we studied and tested several simple ways to improve a genomic classifier that results from gene expression data as well as textual and gene ontology annotations available both for the genes and the biological situations.

Constraint-Based Knowledge Discovery Current analyses of co-expressed genes are often based on global approaches such as clustering or bi-clustering. An alternative way is to employ local methods and search for patterns – sets of genes displaying specific expression properties in a set of situations. The main bottleneck of this type of analysis is the computational cost and the overwhelming number of candidate patterns which can hardly be further exploited. A timely application of background knowledge available in literature databases, biological ontologies and other sources can help to focus on the most plausible patterns only. In [14], we discussed a flexible constraint-based framework that enables the effective mining and representation of meaningful over-expression patterns representing intrinsic associations among genes and biological situations. The framework can be simultaneously applied to a wide spectrum of genomic data. It has been demonstrated that it allows generating new biological hypotheses with clinical implications.

Quantitative association rule mining in genomics using apriori knowledge Regarding association rules, transcriptomic data represent a difficult mining context. First, the data are high-dimensional which asks for an algorithm scalable in the number of variables. Second, expression values are typically quantitative variables. This variable type further increases computational demands and may result in the output with a prohibitive number of redundant rules. Third, the data are often noisy which may also cause a large number of rules of little significance. We tackle the above-mentioned bottlenecks with an alternative approach to the quantitative association rule mining [18,19]. The approach is based on simple arithmetic operations with variables and it outputs rules that do not syntactically differentiate from classical association rules. Apriori genomic knowledge can be used to prune the search space and reduce the amount of derived rules.

Learning Relational Descriptions of Differentially Expressed Gene Groups A method that uses gene ontologies, together with the paradigm of relational subgroup discovery, to find compactly described groups of genes differentially expressed in specific cancers was described in [20,21]. The groups are described by means of relational logic features, extracted from publicly available gene ontology information, and are straightforwardly interpretable by medical experts. We applied the proposed method to three gene expression data sets with the following respective sets of sample classes: (i) acute lymphoblastic leukemia (ALL) vs. acute myeloid leukemia (AML), (ii) seven subtypes of ALL, and (iii) fourteen different types of cancers. Significant number of discovered groups of genes

had a description which highlighted the underlying biological process that is responsible for distinguishing one class from the other classes. The quality of the discovered descriptions was also verified by crossvalidation. The presented approach significantly contributes to the application of relational machine learning to gene expression analysis, given the expected increase in both the quality and quantity of gene/protein annotations.

4 System Architecture

As mentioned above, microarray technology brough completely new possibilities to the field of molecular biology. However, it also became evident that it is not a panacea that would help to understand gene-related mechanisms on its own. Today, it is impossible to interpret data from microarray experiments without deep biological knowledge on the particular data in question, relevant pathways, significant coverage of scientific literature for the particular disease and a manual search for relevant additional information.

The aim of our research is to reduce the tedious work 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. Figure 1 demonstrates the proposed architecture of such a knowledge fusion system.

The process starts with experimental data prepared with the help of methods discussed in the previous section. The user also provides a normalized description of the experimental setup that can be used to retrieve additional data from various databases (the means to avoid ambiguities are discussed later in this section). The comprehensive metadata is crucial for the success of subsequent processing steps.

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 currently 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 (http://www.arrayexpress.com) and STNK (http://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 (http://www.genome.jp/kegg/pathway.html) or the BioCyc collection (biocyc.org), protein knowledge bases such as UniProt the Universal Protein Resource (uniprot.org) and many other resources.

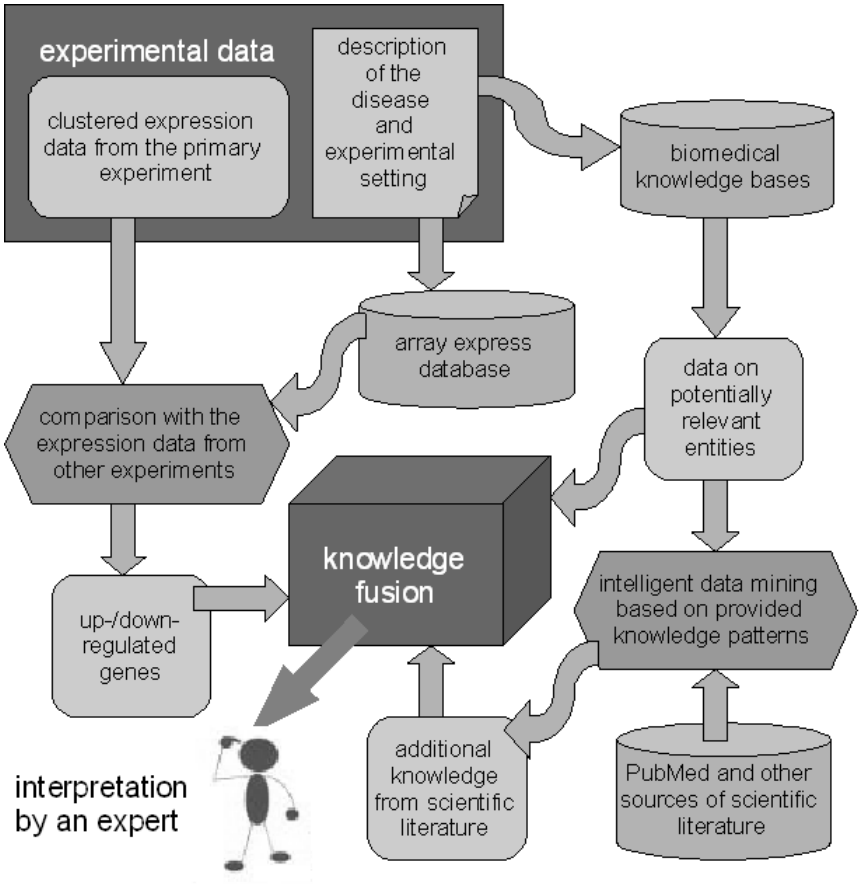

Fig. 1. Schema of the gene-expression knowledge fusion system

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 (just the Pubmed database in many cases). 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 (see below). 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 re-

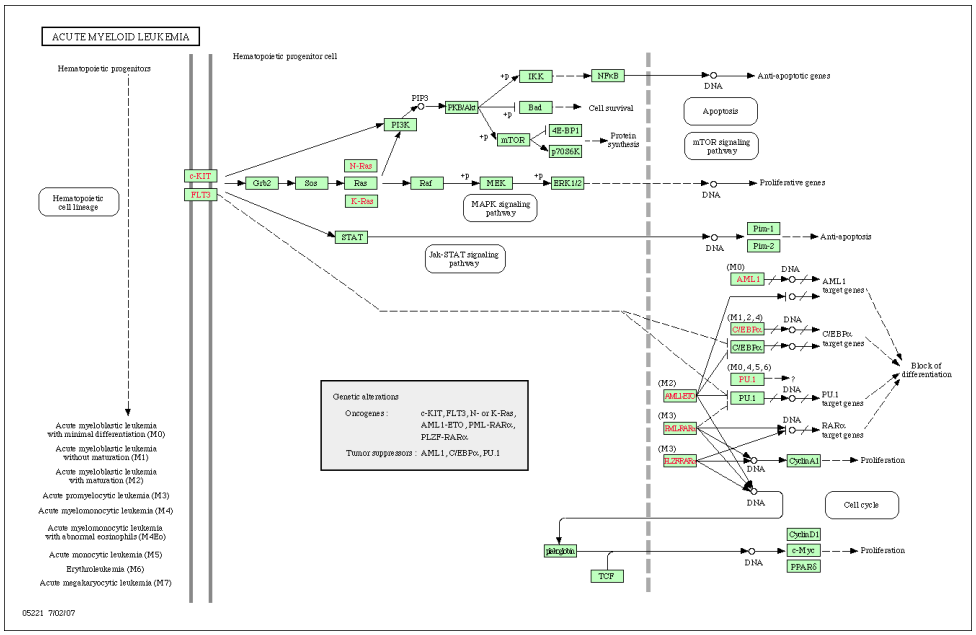

Fig. 2. An example of a pathway relevant to the experimental data (from http://www.genome.jp/kegg/)

stricted 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 [7] 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 [8], 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. Figure 2 shows such an example of a pathway that is represented visually for biologists but, at the same time, can be downloaded or even directly accessed by automatic methods.

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 inferencing engine, we plan to evaluate various recently proposed solutions (e.g., [9]) in terms of their performance and scalability and to integrate the module that will best meet our needs.

5 Related work

There are very many scientific papers dealing with the interdisciplinary field discussed in this paper. In this section, we reference just the sources that directly inspired our presented solutions.

The advantages as well as shortcomings of the current semantic web technologies for the field of biomedical domain in general are tackled by the Semantic Web Health Care and Life Sciences Interest Group operating within the framework of W3C. Even though the outcomes of the group as a whole are rather general and infrastructural in the sense of providing common formats such RDF representation of the biomedical data or core vocabularies and ontologies, various activities of particular members are highly relevant for our research (see, e.g., [10] for a report on joint activities).

Another valuable source of ideas for our research comes from large European projects, either on the national level (e.g., the UK e-Science project myGrid [11]) or an international one (e.g., REWERSE http://rewerse.net). Let us particularly mention the recent work of L. Badea [12] within the last mentioned project which proposes very similar architecture to that discussed in this paper. In contrast to his work, we focus much less on the dynamicity issues and stress rather the aspect of processing efficiency. Especially due to the relatively deep analysis employed in the preprocessing phase, we prefer local replica of the available resources (plus the mechanisms for their regular updates). This schema also simplifies the quality checking and reliability estimation procedures.

Many researchers actually develop sophisticated methods for an automatic processing of biomedical data. The presented architecture allows easy integration of various techniques, especially those that can be characterized as machine learning procedures. For example, the next step of our research will explore the possibility to plug in a recent ILP (inductive logic programming) method to relation mining described in [13].

6 Conclusions and Future Directions

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

Acknowledgements

The work of Pavel Smrž and Jana Šilhavá was partly supported by the Czech Ministry of Education research grants 2B06052 and MSM0021630528. The work of Jiří Kléma and Filip Železný was supported by Czech Academy of Sciences under Grant 1ET101210513. We are also grateful to the team of the Center for Biological Analysis, Masaryk University, Brno, which kindly provided us with all the experimental data we refer to and shared with us their comprehensive knowledge in the field.

References

[1] A. Hill, E. Brown, M. Whitley, G. Tucker-Kellogg, C. Hunter, and D. Slonim, Evaluation of Normalization Procedures for Oligonucleotide Array Data Based on Spiked cRNA Contros, Genome Biology, vol. 2, no. 12, pp. research0055.-1-0055.13, 2001.

[2] J. Schuchhardt, D. Beule, A. Malik, E. Wolski, H. Eickhoff, H. Lehrach, and H. Herzel, Normalization Strategies for cDNA Microarrays, Nucleic Acids Research, vol. 28, no. 10, 2000.

[3] C. Ball, Stanford MicroArray Data Analysis Tutorial: http://genome-www5.stanford.edu/help/TUTORIALS/SMD-Analysis.htm.

[4] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R Tibshirani, D. Botstein, and R. Altman, Missing Value Estimation Methods for DNA Microarrays, Bioinformatics, in press.

[5] D. Jiang, Ch. Tang, A. Zhang, Cluster Analysis for Gene Expression Data: A Survay, IEEE, vol. 16, no. 11, November 2004.

[6] T. S. Furey, Support vector machine classification and validation of cancer tissue samples using microarray expression data, Bioinformatics 16, 906-914, 2000.

[7] V. Novacek and P. Smrz, Empirical merging of ontologies: A proposal of universal uncertainty representation framework. In: Proceedings of ESWC, 2006, pp. 65-79.

[8] G. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller, Five papers on wordnet. Technical Report CSL Report 43, Cognitive Science Laboratory, Princeton University, 1990.
[9] G. Stoilos, N. Simou, G. Stamou and S. Kollias, Uncertainty and the Semantic Web, IEEE Intelligent Systems, 21(5), p. 84-87, 2006.
[10] A. Ruttenberg et al. Advancing translational research with the Semantic Web, BMC Bioinformatics 2007, 8 (Suppl 3):S2.
[11] R. D. Stevens, A. J. Robinson, C. A. Goble, myGrid: Personalised bioinformatics on the information grid. Bioinformatics 2003, 19 (Suppl 1):i302-304.
[12] L. Badea, Semantic Web Reasoning for Analyzing Gene Expression Profiles. In: PPSWR 2006, LNCS 4187, pp. 78-89, 2006.
[13] F. Zelezny, N. Lavrac. Propositionalization-Based Relational Subgroup Discovery with RSD. Machine Learning 62(1-2):33-63, Springer, 2007
[14] I. Trajkovski, F. Zelezny, N. Lavrac and J. Tolar. Learning Relational Descriptions of Differentially Expressed Gene Groups. Accepted to IEEE Trans. Sys Man Cyb C, spec. issue on Intelligent Computation for Bioinformatics.
[15] J. Klema and F. Karel. Quantitative Association Rule Mining in Genomics Using Apriori Knowledge. PriCKL'07 - ECML/PKDD'07 Workshop on Prior Conceptual Knowledge in Machine Learning and Knowledge Discovery, University of Warsaw, Poland, pp. 53-64, 2007.
[16] J. Klema, S. Blachon, A. Soulet, B. Cremilleux, and O. Gandrilon. Constraint-Based Knowledge Discovery from SAGE Data. In Silico Biology, 8, 0014, 2008.
[17] J. Klema, A. Soulet, B. Cremilleux, S. Blachon and O. Gandrilon. Mining Plausible Patterns from Genomic Data. Proceedings of Nineteenth IEEE International Symposium on Computer-Based Medical Systems, pp. 183-188, 2006.
[18] A. Soulet, J. Klema and B. Cremilleux. Efficient Mining under Flexible Constraints through Several Datasets. Proceedings of 5th International Workshop on Knowledge Discovery in Inductive Databases, pp. 4-15, 2006.
[19] T. Charnois, N. Durand and J. Klema. Automated Information Extraction from Gene Summaries. Proceedings of The Workshop on Data and Text Mining for Integrative Biology, 2006.
[20] I. Trajkovski, F. Zelezny, J. Tolar and N. Lavrac. Relational Descriptive Analysis of Gene Expression Data. STAIRS-2006 (3rd European Starting AI Researcher Symposium) at the 17th European Conference on Artificial Intelligence, 2006.
[21] I. Trajkovski, F. Zelezny, J. Tolar and N. Lavrac. Relational Subgroup Discovery for Descriptive Analysis of Microarray Data. Proceedings of the 2n Int Sympos on Computational Life Science, 2006.
[22] W. Pan, A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray expriments, Bioinformatics, 18(4):546-554, 2002.
[23] P. Pavlidis, Using ANOVA for gene selection from microarray studies of the nervous system, Elsevier, Methods 31, 282-289, 2003.