Semantic Subgroup Discovery and Cross-Context Linking for Microarray Data Analysis

Igor Mozetič[1], Nada Lavrač[1,2], Vid Podpečan[1], Petra Kralj Novak[1], Helena Motaln[3], Marko Petek[3], Kristina Gruden[3], Hannu Toivonen[4], and Kimmo Kulovesi[4]

[1] Jožef Stefan Institute, Jamova 39, Ljubljana, Slovenia
{igor.mozetic,nada.lavrac,vid.podpecan,petra.kralj.novak}@ijs.si
[2] University of Nova Gorica, Vipavska 13, Nova Gorica, Slovenia
[3] National Institute of Biology, Večna pot 111, Ljubljana, Slovenia
{helena.motaln,marko.petek,kristina.gruden}@nib.si
[4] Department of Computer Science, University of Helsinki, Finland
{hannu.toivonen,kimmo.kulovesi}@cs.helsinki.fi

Abstract. The article presents an approach to computational knowledge discovery through the mechanism of *bisociation*. Bisociative reasoning is at the heart of creative, accidental discovery (e.g., serendipity), and is focused on finding unexpected links by crossing contexts. Contextualization and linking between highly diverse and distributed data and knowledge sources is therefore crucial for the implementation of bisociative reasoning. In the article we explore these ideas on the problem of analysis of microarray data. We show how enriched gene sets are found by using ontology information as background knowledge in semantic subgroup discovery. These genes are then contextualized by the computation of probabilistic links to diverse bioinformatics resources. Preliminary experiments with microarray data illustrate the approach.

1 Introduction

Systems biology studies and models complex interactions in biological systems with the goal of understanding the underlying mechanisms. Biologists collect large quantities of data from wet lab experiments and high-throughput platforms. Public biological databases, like Gene Ontology and Kyoto Encyclopedia of Genes and Genomes, are sources of biological knowledge. Since the growing amounts of available knowledge and data exceed human analytical capabilities, technologies that help analyzing and extracting useful information from such large amounts of data need to be developed and used.

The concept of association is at the heart of many of today's ICT technologies such as information retrieval and data mining (for example, association rule learning is an established data mining technology, [1]). However, scientific discovery requires creative thinking to connect seemingly unrelated information, for example, by using metaphors or analogies between concepts from different domains. These modes of thinking allow the mixing of conceptual categories and

M.R. Berthold (Ed.): Bisociative Knowledge Discovery, LNAI 7250, pp. 379–389, 2012.

contexts, which are normally separated. One of the functional basis for these modes is the idea of *bisociation*, coined by Artur Koestler half a century ago [8]:

> "The pattern ... is the perceiving of a situation or idea, L, in two self-consistent but habitually incompatible frames of reference, M_1 and M_2. The event L, in which the two intersect, is made to vibrate simultaneously on two different wavelengths, as it were. While this unusual situation lasts, L is not merely linked to one associative context but *bisociated* with two."

Koestler found bisociation to be the basis for human creativity in seemingly diverse human endeavors, such as humor, science, and arts. The concept of bisociation in science is discussed in depth in [2]. Here we take a more restricted and focused view (illustrated in Figure 1).

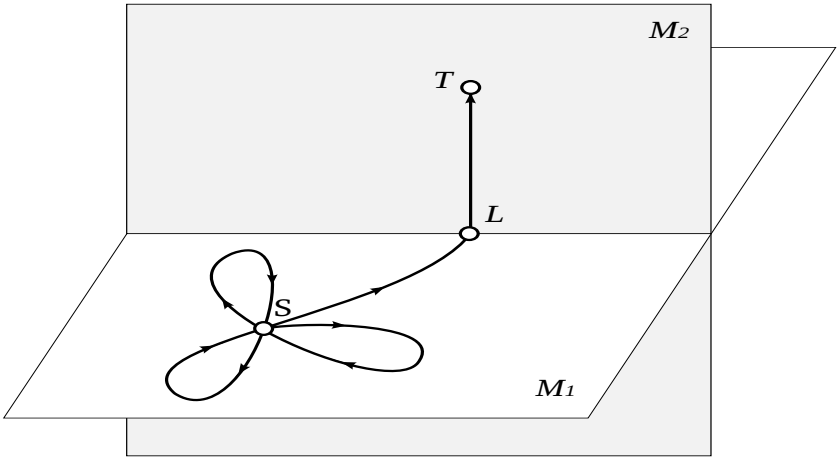

Fig. 1. Koestler's schema of bisociative discovery in science ([8], p. 107)

We are interested in creative discoveries in science, and in particular in computational support for knowledge discovery from large and diverse sources of data and knowledge. The computational realization of bisociative reasoning is based on the following, somewhat simplified, assumptions:

- A bisociative information network (named BisoNet) can be created from available resources. BisoNet is a large graph, where nodes are concepts and edges are probabilistic relations. Unlike semantic nets or ontologies, the graph is easy to construct automatically since it carries little semantics. To a large extent it encodes just circumstantial evidence that concepts are somehow related through edges with some probability.
- Different subgraphs can be assigned to different contexts (frames of reference).

– Graph analysis algorithms can be used to compute links between distant nodes and subgraphs in a BisoNet.
– A bisociative link is a link between nodes (or subgraphs) from different contexts.

In this article we thus explore one specific pattern of bisociation: long-range links between nodes (or subgraph) which belong to different contexts. A long-range link is a special case of a bridging graph [2] since it has the form of a path between two nodes. More precisely, we say that two concepts are bisociated if:

– there is no direct, obvious evidence linking them,
– one has to cross contexts to find the link, and
– this new link provides some novel insight into the problem domain.

We have to emphasize that context crossing is subjective, since the user has to move from his 'normal' context (frame of reference) to an *habitually incompatible context* to find the bisociative link [2]. In Koestler's terms (Figure 1), a habitual frame of reference (plane M_1) corresponds to a BisoNet subgraph as defined by a user or his profile. The rest of the BisoNet represents different, habitually incompatible contexts (in general, there may be several planes M_2). The creative act here is to find links (from S to target T) which lead 'out-of-the-plane' via intermediate, bridging concepts (L). Thus, contextualization and link discovery are two of the fundamental mechanisms in bisociative reasoning.

Finding links between seemingly unrelated concepts from texts was already addressed by Swanson [12]. The Swanson's approach implements *closed discovery*, the so-called A-B-C process, where A and C are given and one searches for intermediate B concepts. On the other hand, in *open discovery* [18], only A is given. One approach to open discovery, RaJoLink [9], is based on the idea to find C via B terms which are rare (and therefore potentially interesting) in conjunction with A. Rarity might therefore be one of the criteria to select links which lead out of the habitual context (around A) to known, but non-obviously related concepts C via B.

In this article we present an approach to bisociative discovery and contextualization of genes which helps in the analysis of microarray data. The approach is based on semantic subgroup discovery (by using ontologies as background knowledge in microarray data analysis), and the linking of various publicly available bioinformatics databases. This is an ongoing work, where some elements of bisociative reasoning are already implemented: creation of the BisoNet graph, identification of relevant nodes in a BisoNet, and computation of links to indirectly related concepts. Currently, we are expanding the BisoNet with textual resources from PubMed, and implementing open discovery from texts through BisoNet graph mining. We envision that the open discovery process will identify potentially interesting concepts from different contexts which will act as the target nodes for the link discovery algorithms. Links discovered in this way, crossing contexts, might provide instances of bisociative discoveries.

The currently implemented steps of bisociative reasoning are the following. The *semantic subgroup discovery* step is implemented by the SEGS system [16].

SEGS uses as background knowledge data from three publicly available, semantically annotated biological data repositories, GO, KEGG and Entrez. Based on the background knowledge, it automatically formulates biological hypotheses: rules which define groups of differentially expressed genes. Finally, it estimates the relevance (or significance) of the automatically formulated hypotheses on experimental microarray data. The *BisoNet creation* and the *link discovery* steps are implemented by the Biomine system [3,11]. Biomine weakly integrates a large number of biomedical resources, and computes most probable links between elements of diverse sources. It thus complements the semantic subgroup discovery technology, due to the explanatory potential of additional link discovery and Biomine graph visualization. While this link discovery process is already implemented, our current work is devoted to the contextualization of Biomine nodes for bisociative link discovery.

The article is structured as follows. Section 2 gives an overview of five steps in exploratory analysis of gene expression data. Section 3 describes an approach to the analysis of microarray data, using semantic subgroup discovery in the context of gene set enrichment. A novel approach, a first attempt at bisociative discovery through contextualization, composed of using SEGS and Biomine (SegMine, for short) is in Section 4. An ongoing experimental case study is presented in Section 5. We conclude in Section 6 with plans for future work.

2 Exploratory Gene Analytics

This section describes the steps which support bisociative discovery, targeted at the analysis of differentially expressed gene sets: gene ranking, the SEGS method for enriched gene set construction, linking of the discovered gene set to related biomedical databases, and finally visualization in Biomine. The schematic overview is in Figure 2.

The proposed method consists of the following five steps:

1. **Ranking of genes.** In the first step, class-labeled microarray data is processed and analyzed, resulting in a list of genes, ranked according to differential expression.
2. **Ontology information fusion.** A unified database, consisting of GO[1] (biological processes, functions and components), KEGG[2] (biological pathways) and Entrez[3] (gene-gene interactions) terms and relationships is constructed by a set of scripts, enabling easy updating of the integrated database (details are discussed by [14]).
3. **Discovering groups of differentially expressed genes.** The ranked list of genes is used as input to the SEGS algorithm [16], an upgrade of the RSD relational subgroup discovery algorithm [4, 5, 15], specially adapted to microarray data analysis. The result is a list of most relevant gene groups that

[1] `http://www.geneontology.org/`
[2] `http://www.genome.jp/kegg/`
[3] `ftp://ftp.ncbi.nlm.nih.gov/gene/GeneRIF/interaction_sources`

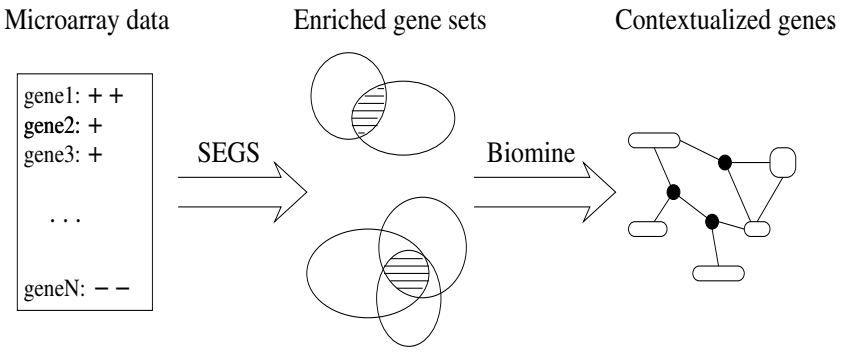

Fig. 2. Microarray gene analytics proceeds by first finding candidate enriched gene sets, expressed as intersections of GO, KEGG and Entrez gene-gene interaction sets. Selected enriched genes are then put in the context of different bioinformatic resources, as computed by the Biomine link discovery engine. The '+' and '-' signs under Microarray data indicate over- and under-expression values of genes, respectively.

semantically explain differential gene expression in terms of gene functions, components, processes, and pathways as annotated in biological ontologies.

4. **Finding links between gene group elements**. The elements of the discovered gene groups (GO and KEGG terms or individual genes) are used to formulate queries for the Biomine link discovery engine. Biomine then computes most probable links between these elements and entities from a number of public biological databases. These links help the experts to uncover unexpected relations and biological mechanisms potentially characteristic for the underlying biological system.
5. **Gene group visualization.** Finally, in order to help in explaining the discovered out-of-the-context links, the discovered gene relations are visualized using the Biomine visualization tools.

3 SEGS: Search for Enriched Gene Sets

The goal of the gene set enrichment analysis is to find gene sets which form coherent groups and are different from the remaining genes. More precisely, a gene set is *enriched* if the member genes are semantically coherent and statistically significantly differentially expressed as compared to the rest of the genes. Two methods for testing the enrichment of gene sets were developed: Gene set enrichment analysis (GSEA) [13] and Parametric analysis of gene set enrichment (PAGE) [7]. Originally, these methods take individual terms from GO and KEGG (which annotate gene sets), and test whether the genes that are annotated by a specific term are statistically significantly differentially expressed in the given microarray dataset.

The novelty of the SEGS method, developed by Trajkovski et al. [14, 16] and used in this study, is that the method does not only test existing gene sets for

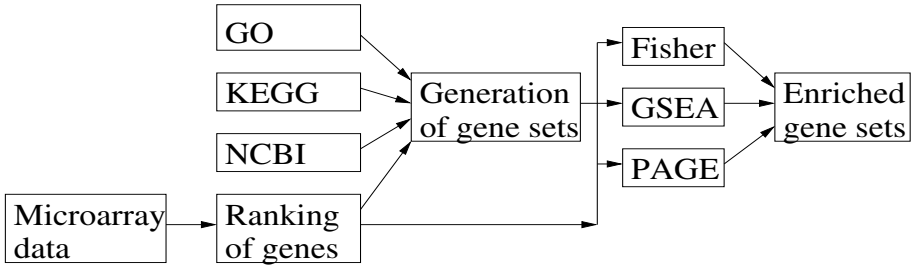

Fig. 3. Schematic representation of the SEGS method

differential expression but it also generates new gene sets that represent novel biological hypotheses. In short, in addition to testing the enrichment of individual GO and KEGG terms, this method tests the enrichment of newly defined gene sets constructed by the intersection of GO terms, KEGG terms and gene sets defined by taking into account also the gene-gene interaction data from Entrez.

The SEGS method has four main components:

- the background knowledge (the GO, KEGG and Entrez databases),
- the SEGS hypothesis language (the GO, KEGG and interaction terms, and their conjunctions),
- the SEGS hypothesis generation procedure (generated hypotheses in the SEGS language correspond to gene sets), and
- the hypothesis evaluation procedure (the Fisher, GSEA and PAGE tests).

The schematic workflow of the SEGS method is shown in Figure 3.

4 SegMine: Contextualization of genes

We made an attempt at exploiting bisociative discoveries within the biomedical domain by explicit contextualization of enriched gene sets. We applied two methods that use publicly available background knowledge for supporting the work of biologists: the SEGS method for searching for enriched gene sets [16] and the Biomine method for contextualization by finding links between genes and other biomedical databases [3,11]. We combined the two methods in a novel way. We used SEGS for hypothesis generation in the form of interesting gene sets, which are constructed as intersections of terms from different ontologies (different contexts). Queries are then formulated to Biomine for out-of-the-context link discovery and visualization (see Figure 4). We believe that this combination provides an easier interpretation of the biological mechanisms underlying differential gene expression for biologists.

In the Biomine[4] project [3,11], data from several publicly available databases were merged into a large graph, a BisoNet, and a method for link discovery

[4] `http://biomine.cs.helsinki.fi/`

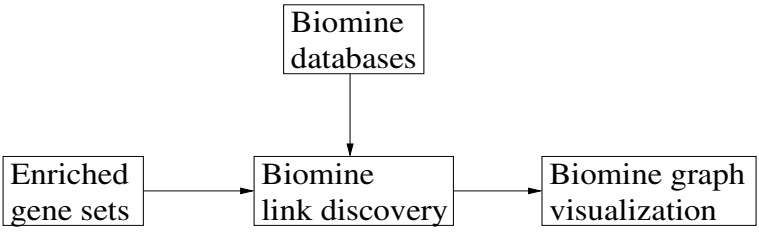

Fig. 4. SegMine workflow

between entities in queries was developed. In the Biomine framework nodes correspond to entities and concepts (e.g., genes, proteins, GO terms), and edges represent known, probabilistic relationships between nodes. A link (a relation between two entities) is manifested as a path or a subgraph connecting the corresponding nodes.

Table 1. Databases included in the Biomine snapshot used in the experiments

Vertex Type	Source Database	Nodes	Degree
Article	PubMed	330,970	6.92
Biological process	GO	10,744	6.76
Cellular component	GO	1,807	16.21
Molecular function	GO	7,922	7.28
Conserved domain	ENTREZ Domains	15,727	99.82
Structural property	ENTREZ Structure	26,425	3.33
Gene Entrez	Gene	395,611	6.09
Gene cluster	UniGene	362,155	2.36
Homology group	HomoloGene	35,478	14.68
OMIM entry	OMIM	15,253	34.35
Protein Entrez	Protein	741,856	5.36
Total		1,968,951	

The Biomine graph data model consists of various biological entities and annotated relations between them. Large, annotated biological data sets can be readily acquired from several public databases and imported into the graph model in a relatively straightforward manner. Some of the databases used in Biomine are summarized in Table 1. The snapshot of Biomine we use consists of a total of 1,968,951 nodes and 7,008,607 edges. This particular collection of data sets is not meant to be complete, but it certainly is sufficiently large and versatile for real link discovery.

5 A Case Study

In the systems biology domain, our goal is to computationally help the experts to find a creative interpretation of wet lab experiment results. In the particular

experiment, the task was to analyze microarray data in order to distinguish between fast and slowly growing cell lines through differential expression of gene sets, responsible for cell growth.

Table 2. Top SEGS rules found in the cell growth experiment. The second rule states that one possible distinction between the slow and fast growing cells is in genes participating in the process of DNA replication which are located in the cell nucleus and which interact with genes that participate in the cell cycle pathway.

Enriched Gene Sets
1. SLOW-vs-FAST ← GO_Proc('DNA metabolic process') &
INTERACT(GO_Comp('cyclin-dep. protein kinase holoenzyme complex'))
2. SLOW-vs-FAST ← GO_Proc('DNA replication') &
GO_Comp('nucleus') &
INTERACT(KEGG_Path('Cell cycle'))
3. SLOW-vs-FAST ← . . .

Table 2 gives the top rules resulting from the SEGS search for enriched gene sets. For each rule, there is a corresponding set of over expressed genes from the experimental data. Figure 5 shows a part of the Biomine graph which links a selected subset of enriched gene set to the rest of the nodes in the Biomine graph.

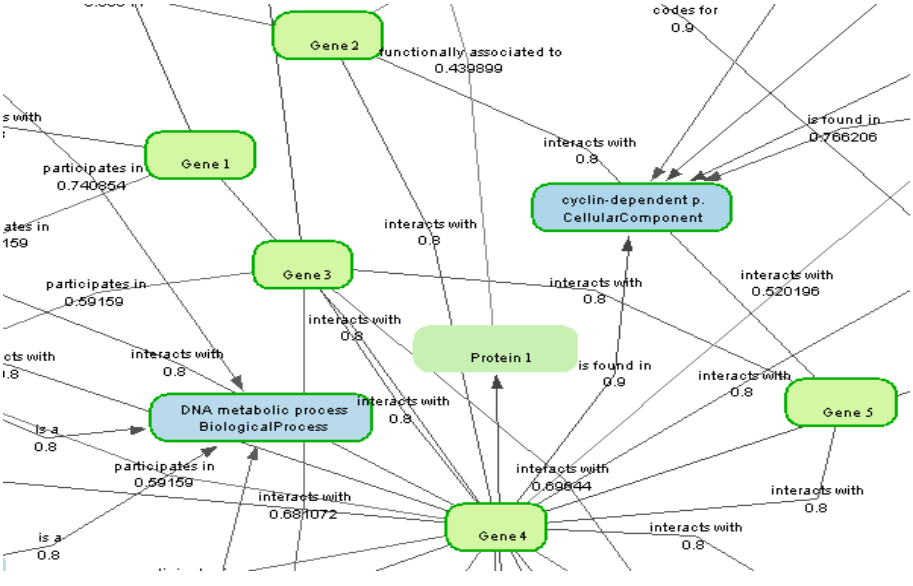

Fig. 5. Biomine subgraph related to five genes from the enriched gene set produced by SEGS. Note that the gene and protein names are not explicitly presented, due to the preliminary nature of these results.

The wet lab scientists have assessed that SegMine, SEGS in combination with Biomine, provides additional hints on what to focus on when comparing the expression data of cells. In subsequent analysis of senescence in human stem cells, the use of SegMine resulted in formulation of three novel research hypotheses which could improve understanding of the underlying mechanisms and identification of candidate marker genes [10].

In principle, such an in-silico analysis can considerably lower the costs of in-vitro experiments with which the researchers in the wet lab are trying to get a hint of a novel process or phenomena observed. This may be especially true for situations when one cannot find deeper explanation for drug effects, organ functions, or diseases from surface observations only. Namely, the gross, yet important characteristics of the cells (organ function) are not directly accessible (since they do not affect visual morphology) or could not be identified soon enough. An initial requirement for this approach is wide accessibility and low costs of high throughput microarray analysis which generate appropriate data for in-silico analysis.

6 Conclusions

We presented SegMine, a bisociation discovery system for exploratory gene analytics. It is based on the non-trivial steps of subgroup discovery (SEGS) and link discovery (Biomine). The goal of SegMine is to enhance the creation of novel biological hypotheses about sets of genes. An implementation of the gene analytics software, which enhances SEGS and creates links to Biomine queries and graphs, is available as a set of workflows in the Orange4WS[5] service-oriented platform at `http://segmine.ijs.si/`.

In the future we plan to enhance the contextualization of genes with contexts discovered by biomedical literature mining. We will add PubMed articles data into the BisoNet graph structure. In particular, we already have a preliminary implementation of software, called Texas [6], which creates a probabilistic network (BisoNet, compatible to Biomine) from textual sources. By focusing on different types of links between terms (e.g., frequent and rare co-ocurances) we expect to get hints at some unexpected relations between concepts from different contexts.

Our long term goal is to help biologists better understand inter-contextual links between genes and their role in explaining (at least qualitatively) underlying mechanisms which regulate gene expression. The proposed approach is considered a first step at computational realization of bisociative reasoning for creative knowledge discovery in systems biology.

Acknowledgements. The work presented in this article was supported by the European Commission under the 7th Framework Programme FP7-ICT-2007-C FET-Open project BISON-211898, by the Slovenian Research Agency grants P2-0103, J4-2228, P4-0165, and by the Algorithmic Data Analysis (Algodan)

[5] `http://orange4ws.ijs.si/`

Centre of Excellence of the Academy of Finland. We thank Igor Trajkovski for his previous work on SEGS, and Filip Železný and Jakub Tolar for their earlier contributions leading to SEGS.

References

1. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proc. 20th Intl. Conf. on Very Large Data Bases, VLDB, Santiago, Chile, pp. 487–499 (1994)
2. Dubitzky, W., Kötter, T., Schmidt, O., Berthold, M.R.: Towards Creative Information Exploration Based on Koestler's Concept of Bisociation. In: Berthold, M.R. (ed.) Bisociative Knowledge Discovery. LNCS (LNAI), vol. 7250, pp. 11–32. Springer, Heidelberg (2012)
3. Eronen, L., Hintsanen, P., Toivonen, H.: Biomine: A Network-Structured Resource of Biological Entities for Link Prediction. In: Berthold, M.R. (ed.) Bisociative Knowledge Discovery. LNCS (LNAI), vol. 7250, pp. 364–378. Springer, Heidelberg (2012)
4. Gamberger, D., Lavrač, N.: Expert-guided subgroup discovery: Methodology and application. Journal of Artificial Intelligence Research 17, 501–527 (2002)
5. Gamberger, D., Lavrač, N., Železný, F., Tolar, J.: Induction of comprehensible models for gene expression datasets by the subgroup discovery methodology. Journal of Biomedical Informatics 37, 269–284 (2004)
6. Juršič, M., Lavrač, N., Mozetič, I., Podpečan, V., Toivonen., H.: Constructing information networks from text documents. In: ECML/PKDD 2009 Workshop on Explorative Analytics of Information Networks, Bled, Slovenia (2009)
7. Kim, S.Y., Volsky, D.J.: PAGE: Parametric Analysis of Gene Set Enrichment. BMC Bioinformatics 6, 144 (2005)
8. Koestler, A.: The Act of Creation. The Macmillan Co., New York (1964)
9. Petrič, I., Urbančičc, T., Cestnik, B., Macedoni-Lukšič, M.: Literature mining method RaJoLink for uncovering relations between biomedical concepts. Journal of Biomedical Informatics 42(2), 219–227 (2009)
10. Podpečan, V., Lavrač, N., Mozetič, I., Kralj Novak, P., Trajkovski, I., Langohr, L., Kulovesi, K., Toivonen, H., Petek, M., Motaln, H., Gruden, K.: SegMine workflows for semantic microarray data analysis in Orange4WS. BMC Bioinformatics 12, 416 (2011)
11. Sevon, P., Eronen, L., Hintsanen, P., Kulovesi, K., Toivonen, H.: Link Discovery in Graphs Derived from Biological Databases. In: Leser, U., Naumann, F., Eckman, B. (eds.) DILS 2006. LNCS (LNBI), vol. 4075, pp. 35–49. Springer, Heidelberg (2006)
12. Swanson, D.R., Smalheiser, N.R., Torvik, V.I.: Ranking indirect connections in literature-based discovery: The role of Medical Subject Headings (MeSH). JASIST 57(11), 1427–1439 (2006)

13. Subramanian, P., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A.: Gene set enrichment analysis: A knowledge based approach for interpreting genome-wide expression profiles. Proc. of the National Academy of Science, USA 102(43), 15545–15550 (2005)
14. Trajkovski., I.: Functional interpretation of gene expression data. Ph.D. Thesis, Jozef Stefan International Postgraduate School, Ljubljana, Slovenia (2007)
15. Trajkovski, I., Železny, F., Lavrač, N., Tolar, J.: Learning relational descriptions of differentially expressed gene groups. IEEE Transactions of Systems, Man and Cybernetics C, Special Issue on Intelligent Computation for Bioinformatics 38(1), 16–25 (2008)
16. Trajkovski, I., Lavrač, N., Tolar, J.: SEGS: Search for enriched gene sets in microarray data. Journal of Biomedical Informatics 41(4), 588–601 (2008)
17. Železny, F., Lavrač, N.: Propositionalization-based relational subgroup discovery with RSD. Machine Learning 62(1-2), 33–63 (2007)
18. Weeber, M., Klein, H., de Jong-van den Berg, L.T.W., Vos, R.: Using concepts in literature-based discovery: Simulating Swanson's Raynaud-fish oil and migraine-magnesium discoveries. J. Am. Soc. Inf. Sci. Tech. 52(7), 548–557 (2001)