Gene Analytics: Discovery and Contextualization of Enriched Gene Sets

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

[1] Jožef Stefan Institute, Jamova 39, Ljubljana, Slovenia
{nada.lavrac, igor.mozetic, vid.podpecan, petra.kralj}@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

Abstract. The paper present a preliminary study of creative knowledge discovery through bisociative data analysis. Bisociative reasoning is at the heart of creative, accidental discovery (serendipity), and is focused on finding unexpected links by crossing different contexts. Contextualization and linking between highly diverse and distributed data and knowledge sources is therefore crucial for implementation of bisociative reasoning. In the paper 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. Results of two case studies are used to illustrate the approach.

1 Introduction

Biologists collect large quantities of data from wet lab experiments and high-throughput platforms. Public biological databases, like Gene Ontology, Kyoto Encyclopedia of Genes and Genomes and ENTREZ, 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. However, scientific discovery requires creative thinking to connect seemingly unrelated information, for example, by using metaphors or analogical reasoning. These modes of thinking allow the mixing of conceptual categories and contexts, which are normally separated. The functional basis for these modes is a mechanism called *bisociation* [8]:

> "The pattern underlying ... 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."

From the computational point of view, we say that two concepts are bisociated [14] 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]. Thus, contextualization is one of the fundamental mechanisms in bisociative reasoning. In this paper we present an approach to discovery and contextualization of genes which should help in analysis of microarray data. The approach is based on information fusion, semantic subgroup discovery (by using ontologies as background knowledge in microarray data analysis), and the linking of various publicly available bioinformatics databases. We first explain the basic notions: information fusion, subgroup discovery and semantic subgroup discovery.

1.1 Information fusion

Information fusion can be defined as the study of efficient methods for automatically or semi-automatically transforming information from different sources and different points in time into a representation that provides effective support for human and automated decision making [1]. Recent investigations in using information fusion to support scientific decision making within bioinformatics include [3, 9]. Smirnov et al. [12] exploit the idea of formulating an ontology-based model of the problem to be solved by the user and interpreting it as a constraint satisfaction problem taking into account information from a dynamic environment.

An approach to the integration of biological databases GO, KEGG and ENTREZ is implemented in the SEGS information fusion engine (Searching for Enriched Gene Sets, [16]). Another, much larger, integrated annotated bioinformatics information source is Biomine [11].

1.2 Subgroup discovery

Subgroup discovery techniques are used to generate explicit knowledge in the form of rules that allow the user to recognize important relationships in a set of class labeled training instances, describing the target property of interest. Consider two applications. In the first one, the induced subgroup describing rules suggest the general practitioner how to select individuals for population screening, concerning high risk for coronary heart disease (CHD) [4]. The rule below describes a group of overweight female patients older than 63 years:

$$\text{High\_CHD\_Risk} \leftarrow \text{sex} = \text{female} \;\&\; \text{age} > 63 \text{ years} \;\&\; \text{body\_mass\_index} > 25\ kgm^{-2}$$

In the second application [5], subgroup describing rules suggest genes that are characteristic for a given cancer type (i.e., leukemia cancer) in an application of distinguishing among 14 different cancer types: leukemia, CNS, lung cancer, etc.:

$$\text{Leukemia} \leftarrow \text{KIAA0128 is diff\_expressed} \;\&\; \text{prostaglandin\_d2\_synthase is not diff\_expressed}$$

1.3 Semantic subgroup discovery

Semantic subgroup discovery refers to subgroup discovery, where semantically annotated knowledge sources (ontologies) are used as background knowledge in the data mining process. Using the technology of relational subgroup discovery [17], we have developed an approach to information fusion and semantic data mining, enabling background knowledge in the form of ontologies to be used in relational machine learning. The relational subgroup discovery approach, which was successfully adapted and applied to mining of bioinformatics data [15], and further refined in the SEGS algorithm (Searching for Enriched Gene Sets, [16]), is used in the information fusion and semantic subgroup discovery technology described in this paper. Example rules below are induced by a semantic knowledge discovery engine for two cancer types (ALL and AML) and ranked according to the enrichment score. The rules are a conjunction of ontology terms from the GO, KEGG and ENTREZ ontologies:

$$\text{ALL} \leftarrow \text{Func('zinc ion binding')} \;\&\; \text{Comp('chromosomal part')}$$
$$\text{AML} \leftarrow \text{Func('metal ion binding')} \;\&\; \text{Comp('cell surface')} \;\&\; \text{Proc('response to pest,pathogen,parasite')}$$

1.4 Overview of the paper

This paper describes first steps in creative data and knowledge exploration through *semantic subgroup discovery* and contextualization through *link discovery* between diverse bioinformatics databases. The described approach to semantic subgroup discovery employs semantically annotated knowledge sources as background knowledge for subgroup discovery. In this paper we investigate a special subgroup discovery task: the *gene set enrichment* analysis task. A gene set is *enriched* if the genes that are members of the set are statistically significantly differentially expressed compared to the rest of the genes.

The SEGS method [16] 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 Biomine service [11] provides links to a large number of biomedical resources, complementing

our semantic subgroup discovery technology, due to the explanatory potential of additional link discovery and Biomine graph visualization.

The paper 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 methodology, a first attempt at bisociative discovery through contextualization, composed of using SEGS and Biomine (SEGS+Biomine, for short) is in Section 4. Two preliminary case studies are presented in Section 5.

2 Exploratory gene analytics

This section describes the methodological ingredients of the semantic subgroup discovery technology, 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 shematic overview is in Figure 1.

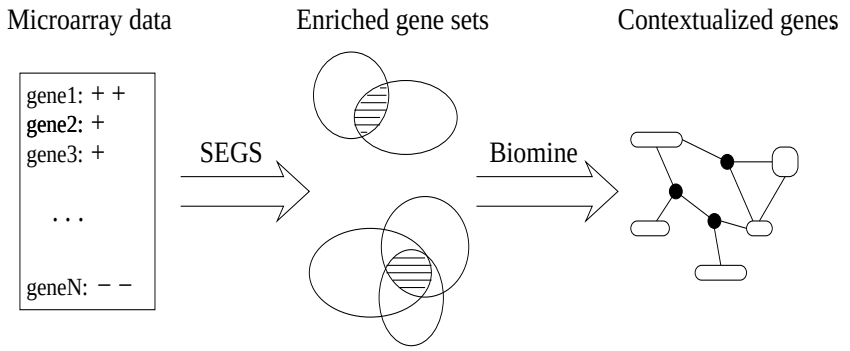

Fig. 1. Microarray gene analytics proceeds by first finding candidate enriched gene sets, expressed as intersections of GO, KEGG and gene-gene interaction sets. Selected enriched genes are then put in context of different bioinformatic resources, as computed by Biomine link discovery engine.

The proposed method consists of the following five steps:

1. **Ranking of genes.** In the first step, class-labeled microarray data is processed and analysed, resulting in a list of genes, ranked according to differential expression.
2. **Ontology information fusion.** A unified database, consisting of GO (processes, functions and components), KEGG (biological pathways) and ENTREZ (gene-gene interactions) terms and relationships is constructed. To this end, a set of scripts was written, enabling easy updating of the integrated database.

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 [15], specially adapted to microarray data analysis. The result is a list of most relevant gene groups that semantically explain differential gene expression in terms of gene functions, components and processes 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 entered as queries to the Biomine crawler. Biomine computes most probable links between these elements and a number of public biological databases. These links help the experts to uncover unexpected relations and biological mechanisms potentially characteristic for the underlying biological processes.
5. **Gene group visualization.** Finally, in order to help in explaining the discovered ontological relationships, the discovered gene relations are visualized using Biomine visualization toolbox.

3 SEGS: Search for Enriched Gene Sets

The goal of gene set enrichment analysis is to find groups of genes—the so-called gene sets—that are enriched. A gene set is *enriched* if the genes that are members of that gene set are statistically significantly differentially expressed 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 terms (gene sets) from the Gene Ontology (GO), the Kyoto Encyclopedia of Genes and Genomes (KEGG) and ENTREZ interactions, and test whether the genes that are annotated by a specific term are statistically significantly differentially expressed in the given dataset.

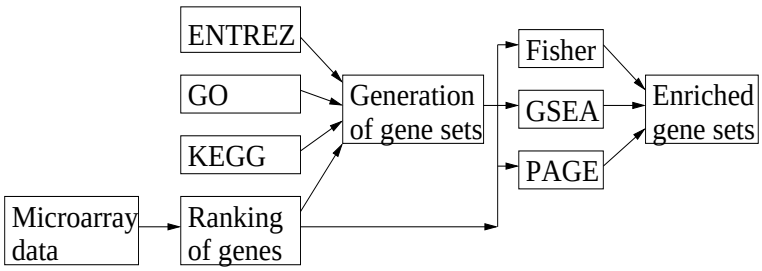

Fig. 2. Schematic representation of the SEGS method.

The novelty of our SEGS method, developed by Trajkovski et al. [16] and used in this study, is that the method does not only test existing gene sets for 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 hypothesis language, the hypothesis generation procedure and the hypothesis evaluation procedure. The schematic workflow of the SEGS method is shown in Figure 2.

4 SEGS+Biomine: 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 [11]. We combined the two methods in a novel way: we used SEGS for hypothesis generation and evaluation from microarray experimental data, and then input the SEGS results into Biomine for inter-context link discovery and visualization (see Figure 3). We believe that by forming hypotheses with SEGS, constructed as conjunctions of terms from different ontologies (different contexts), discovering links between them by Biomine, and visualizing the SEGS hypotheses and the discovered links by the Biomine graph visualization engine, the interpretation of the biological mechanisms underlying differential gene expression is easier.

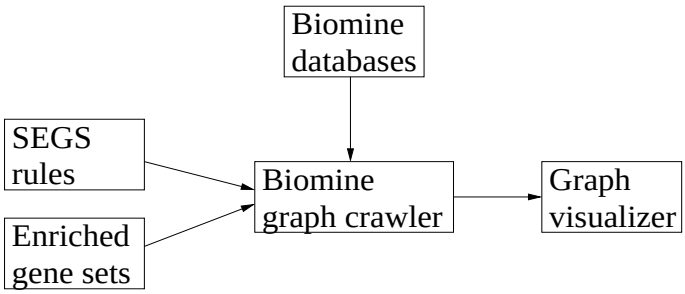

Fig. 3. SEGS+Biomine workflow.

In the Biomine project [11], data from several publicly available databases were merged into a large graph and a method for link discovery between entities in queries was developed. In the Biomine framework vertices correspond to entities and concepts, and edges represent known, annotated relationships between

vertices. A link (a relation between two entities) is manifested as a path or a subgraph connecting the corresponding vertices.

Vertex Type	Source Database	Vertices	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

Table 1. Databases included in Biomine.

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 straightforward manner. Some of the databases used in Biomine are summarized in Table 1. Currently, Biomine consists of a total of 1,968,951 vertices 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 Two case studies

In the first case study, SEGS was applied to find enriched gene sets for distinguishing between two cancer types. In the second case, SEGS and Biomine were combined in order to find an underlying mechanism which might explain why some specific cells are growing faster then the others, in terms of genetic markers.

5.1 Functional genomics

In functional genomics, gene expression monitoring by DNA microarrays (gene chips) provides an important source of information that can help in understanding many biological processes. The database we analyzed consists of a set of gene expression measurements (examples), each corresponding to a large number of measured expression values of a predefined family of genes (attributes). Each measurement in the database was extracted from a tissue of a patient with

a specific disease; this disease is the class for the given example. The domain, described in [5, 10] and used in our experiments, is a typical scientific discovery domain characterised by a large number of attributes compared to the number of available examples. As such, this domain is especially prone to overfitting, as it has two different cancer classes and a few training examples, where the examples are described by thousands of attributes presenting gene expression values. While the standard goal of machine learning is to start from the labeled examples and construct models/classifiers that can successfully classify new, previously unseen examples, our main goal is to uncover interesting patterns/rules that can help to better understand the dependencies between classes (diseases) and attributes (gene expressions values).

Gene Set	ES
Enriched in ALL	
1. ALL ← GO_Func('zinc ion binding') & GO_Comp('chromosomal part') & GO_Proc('interphase of mitotic cell cycle')	0.60
2. ALL ← GO_Proc('DNA metabolism')	0.59
3. ALL ← GO_Func('ATP binding') & GO_Comp('chromosomal part') & GO_Proc('DNA replication')	0.55
Enriched in AML	
1. AML ← GO_Func('metal ion binding') & GO_Comp('cell surface') & GO_Proc('response to pest,pathogen,parasite')	0.54
2. AML ← GO_Comp('lysosome')	0.53
3. AML ← GO_Proc('inflammatory response') & GO_Comp('cell surface')	0.51

Table 2. The top most enriched gene sets found in the leukemia dataset with the p-value ≤ 0.001.

Sample top-ranked rules, induced by a semantic knowledge discovery engine for two cancer types (ALL and AML), ranked according to enrichment score (ES), are listed in Table 2. Note that in Table 2 a term *enrichment* is used, meaning the enrichment of differential expression of a set of genes, annotated by the given conjunction of GO, KEGG and/or ENTREZ terms.

5.2 Systems biology

In the systems biology domain, our goal is to help the expert to find a biological interpretation of wet lab experiment results. In the particular experiment, the task is to analyse microarray data in order to distinguish between fast and slowly growing cell lines. The aim of this study was to explain the differences between

the cases of fast and slowly growing cell lines through differential expression of gene sets, responsible for cell growth.

Gene Set
1. SLOW-vs-FAST ← GO_Proc('DNA metabolic process') & INTERACT(GO_Comp('cyclin-dependent 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 3. 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.

Table 3 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 4 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.

We believe that SEGS in combination with Biomine may give a wet lab scientist additional hints on what to focus on when comparing the expression data of cells. Additionally, 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 just knowing the final outcome one cannot explain the drug effect, organ function, or disesase satisfactory, since the gross, yet important characteristics of the cells (organ function) are hidden (do not affect visual morphology) or could not be recognized soon enough. An initial predisposition for this approach is wide accessibility and low costs of high throughput microarray analyses which generate appropriate data for in-silico analyses.

6 Conclusions

A prototype version of the gene analytics software, which enchances SEGS and creates links to Biomine queries and graphs is available as a web application at `http://zulu.ijs.si/web/segs_ga/`.

In the future work we plan to enchance the contextualization of genes with biomedical literature as available in PubMed. To this end, we already have a preliminary implementation of software, called Texas [6], which createas a probabilistic network (BisoNet, compatible to Biomine) from textual sources. By

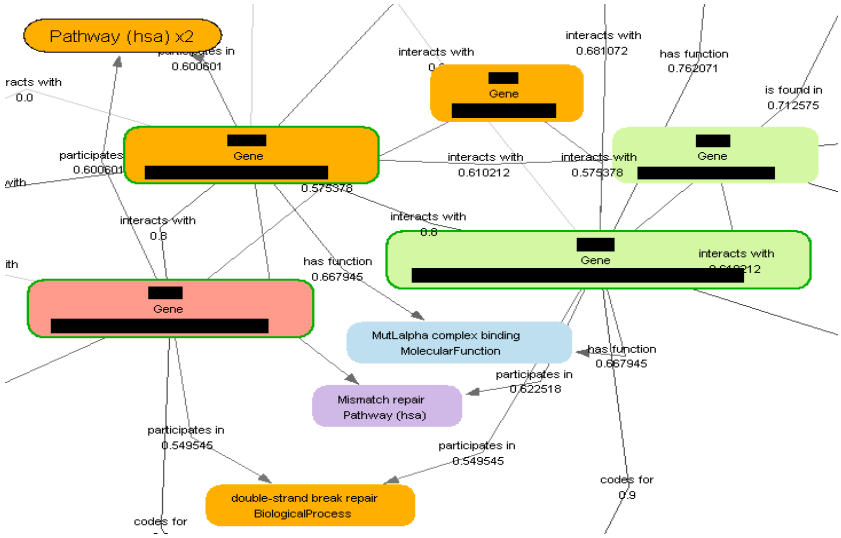

Fig. 4. Biomine subgraph related to three genes from the enriched gene set produced by SEGS. Note that some information is hidden, due to preliminary nature of these results.

focusing on different types of links between terms (e.g., frequent and rare coocurances) we expect to get hints at some unexpected relations between concepts.

Our long term goal is to help biologists at better understanding of inter-contextual links between genes and their role in explaining (at least qualitatively) underlying mechanisms which regulate gene expressions.

7 Acknowledgment

The work presented in this paper was supported by the Slovenian Research Agency grant Knowledge Technologies, and by the grant of the European Commission under the 7th Framework Programme FP7-ICT-2007-C FET-Open, contract no. BISON-211898. We thank Igor Trajkovski for his work on SEGS, and Hannu Toivonen and Kimmo Kulovesi for their help with using Biomine.

References

1. H. Bostrom et al. On the definition of information fusion as a field of research. Technical report, University of Skövde, School of Humanities and Informatics, 2007.
2. W. Dubitzky. Personal communication, FP7 BISON project review, Leuven, June 2009.
3. E. Dura, B. Gawronska, B. Olsson and B. Erlendsson, Towards Information Fusion in Pathway Evaluation: Encoding Relations in Biomedical Texts. Proc. of the 9th International Conference on Information Fusion, 2006.

4. D. Gamberger and N. Lavrač. Expert-Guided Subgroup Discovery: Methodology and Application. *Journal of Artificial Intelligence Research* 17:501–527, 2002.
5. D. Gamberger, N. Lavrač, F. Železný, and J. Tolar. Induction of comprehensible models for gene expression datasets by the subgroup discovery methodology. *Journal of Biomedical Informatics* 37:269–284, 2004.
6. M. Juršič, N. Lavrač, I. Mozetič, V. Podpečan, H. Toivonnen. Constructing information networks from text documents. ECML/PKDD 2009 Workshop "Explorative Analytics of Information Networks", Bled, 2009.
7. S.Y. Kim and D.J. Volsky. PAGE: Parametric Analysis of Gene Set Enrichment. *BMC Bioinformatics* 6:144, 2005.
8. A. Koestler. *The Act of Creation*, The Macmillan Co, New York, 1964.
9. S. Racunas and C. Griffin, Logical data fusion for biological hypothesis evaluation. Proc. of the 8th International Conference on Information Fusion, 2005.
10. S. Ramaswamy et al. Multiclass cancer diagnosis using tumor gene expression signitures. In *Proceedings of the National Academy of Science, USA*, 98(26): 15149–15154, 2001.
11. P. Sevon, L. Eronen, P. Hintsanen, K. Kulovesi, and H. Toivonen. Link discovery in graphs derived from biological databases. In *Proceedings of 3rd International Workshop on Data Integration in the Life Sciences, (DILS'06)*, July 2006. Springer.
12. Smirnov, M. Pashkin, N. Shilov, T. Levashova and A. Krizhanovsky, Intelligent Support for Distributed Operational Decision Making. In: Proceedings of the 9th International Conference on Information Fusion, 2006.
13. P. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, et al. Gene set enrichment analysis: A knowledge based approach for interpreting genome-wide expression profiles. In *Proceedings of the National Academy of Science, USA*, 102(43):15545–15550, 2005.
14. H. Toivonen. Personal communication, FP7 BISON project meeting, Ulster, Sep. 2008.
15. I. Trajkovski, F. Železny, N. Lavrač, and J. Tolar. 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, 2008a.
16. I. Trajkovski, N. Lavrač, and J. Tolar. SEGS: Search for enriched gene sets in microarray data. *Journal of Biomedical Informatics*, 41(4):588–601, 2008b.
17. F. Železny and N. Lavrač. Propositionalization-based relational subgroup discovery with RSD. *Machine Learning*, 62(1–2): 33–63, 2007.