Relational Subgroup Discovery for Descriptive Analysis of Microarray Data

Igor Trajkovski[1], Filip Železný[2], Jakub Tolar[3], and Nada Lavrač[1]

[1] Department of Knowledge Technologies, Jozef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia,
{igor.trajkovski, nada.lavrac}@ijs.si
[2] Department of Cybernetics, Czech Technical University in Prague,
Technicka 2, 166 27 Praha 6, Czech Republic
zelezny@fel.cvut.cz
[3] Department of Pediatrics, University of Minnesota Medical School,
420 Delaware Street, 55455 Minneapolis, USA
tolar003@umn.edu

Abstract. This paper presents a method that uses gene ontologies, together with the paradigm of relational subgroup discovery, to help find description of groups of genes differentially expressed in specific cancers. The descriptions are represented by means of relational features, extracted from gene ontology information, and are straightforwardly interpretable by the medical experts. We applied the proposed method to two known data sets: acute lymphoblastic leukemia (ALL) vs. acute myeloid leukemia and classification of fourteen types of cancer. Significant number of discovered groups of genes had a description, confirmed by the medical expert, which highlighted the underlying biological process that is responsible for distinguishing one class from the other classes. We view our methodology not just as a prototypical example of applying sophisticated machine learning algorithms to microarray data, but also as a motivation for developing more sophisticated functional annotations and ontologies, that can be processed by such learning algorithms.

1 Introduction

Microarrays are at the center of a revolution in biotechnology, allowing researchers to simultaneously monitor the expression of tens of thousands of genes. Independent of the platform and the analysis methods used, the result of a microarray experiment is, in most cases, a list of genes found to be differentially expressed. 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. For example, experts routinely scan gene expression clusters to

see if any of the clusters are explained by a known biological function. Efficient interpretation of these 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. Including the ontologies as a direct knowledge source for any algorithmic strategy to approach such data may greatly facilitate the analysis.

Here we present an algorithm that for given multi-dimensional numerical data set, representing the expression of the genes under different conditions (that define the classes of examples) and ontology used for producing background knowledge about these genes, is able to identify groups of genes, described by conjunctions of first order features, whose expression is highly correlated with one of the classes. For example, one of the applications of this algorithm is to describe groups of genes that were selected as discriminative for some classification problem. Medical experts are usually not satisfied with having a separate description of every discriminative gene, but want to know the processes that are controlled by these genes. With our algorithm we are able to find these processes and the cellular components where they are "executed", indicating the genes from the preselected list of discriminative genes which are included in these processes.

These goals can be achieved by using the methodology of Relational Subgroup Discovery (RSD) [7]. With RSD we were able to induce set of discrimination rules between the different types (or subtypes) of cancers in terms of functional knowledge extracted from the gene ontology and information about gene interactions. In this way, we have succeeded to explain the differences between the types of cancer in terms of the functions of the genes that are differentially expressed in these types.

1.1 Analysis of gene expression data

Large scale gene expression data sets include thousands of genes measured at dozens of conditions. The number and diversity of genes make manual analysis difficult and automatic analysis methods necessary. Initial efforts to analyze these data sets began with the application of unsupervised machine learning, or clustering, to group genes according to similarity in gene expression [2]. Clustering provides a tool to reduce the size of the dataset to a simpler one that can more easily be manually examined. The analysis of gene expression data for various tissue samples has enabled researchers to determine gene expression profiles characteristic of the disease subtypes. The groups of genes involved in these genetic profiles are rather large and a deeper understanding of the functional distinction between the disease subtypes might help not only to select highly accurate "genetic signatures" of the various subtypes, but hopefully also to select potential targets for drug design. Most current approaches to microarray data analysis use (supervised or unsupervised) clustering algorithms to deal with the numerical expression data. While a clustering method reduces the dimensionality of the data to a size that a scientist can tackle, it does not identify the

critical background biological information that helps the researcher understand the significance of each cluster. However, that biological knowledge in terms of functional annotation of the genes is already available in public databases. Direct inclusion of this knowledge source can greatly improve the analysis, support (in term of user confidence) and explain obtained numerical results.

1.2 Gene Ontologies

One of the most important tools for the representation and processing of information about gene products and functions is the Gene Ontology (GO). It provides a controlled vocabulary for the description of cellular components, molecular functions, and biological processes. As of January 2006 (www.geneontology.org), GO contains 1681 component, 7386 function and 10392 process terms. Terms are organized in parent-child hierarchies, indicating either that one term is more specific than another (is_a) or that the entity denoted by one term is part of the entity denoted by another (part_of). Typically, such annotations are first of all established electronically and later validated by a process of manual verification.

Recently, an automatic ontological analysis approach using GO has been proposed to help solving the task of interpreting the results of gene expression data analysis [5]. From 2003 to 2005, 13 other tools have been proposed for this type of analysis and more tools continue to appear daily. Although these tools use the same general approach, identifying statistically significant GO terms that cover a selected list of genes, they differ greatly in many respects that influence in an essential way the results of the analysis. A general idea and comparison of those tools is presented in [6]. Another approach to descriptive analysis of gene expression data is [11]. They present a method that uses text analysis to help find meaningful gene expression patterns that correlate with the underlying biology as described in the scientific literature.

2 Descriptive analysis of gene expression data

The fundamental idea of this paper is as follows. First, we construct a set of discriminative genes, $G_C(c)$, for every class $c \in C$. These sets can be constructed in several ways. For example: $G_C(c)$ can be the set of k ($k > 0$) most correlated genes with class c, computed by, for example, Pearson's correlation. $G_C(c)$ can also be the set of best k single gene predictors, using the recall values from a microarray experiment (absent/present/marginal) as the expression value of the gene. These predictors can look like this: **If** $gene_i$ = present **Then** class = c. In our experiments we used a measure of correlation, $P(g, c)$, that emphasizes the "signal-to-noise" ratio in using gene g as predictor for class c. The definition and analysis of $P(g, c)$ is presented in later section.

The second step aims at improving the interpretability of G_C. Informally, we do this by identifying groups of genes in $G_C(c)$ (for each $c \in C$) which can be summarized in a compact way. Put differently, for each $c_i \in C$ we search

for compact descriptions of group of genes which correlate strongly with c_i and weakly with all $c_j \in C; j \neq i$.

Searching for these groups of genes, together with their description, is defined as a separate supervised machine learning task. This secondary task is, in a way, orthogonal to the primary discovery process in that the original attributes (genes) now become training examples, each of which has a class label $c \in C$. To apply a discovery algorithm, information about relevant features of the new examples is required. No such features are usually present in the gene expression data sets themselves. However, this information can be extracted from a public database of gene annotations. For each gene we extracted its molecular functions, biological processes and cellular components where its protein products are located. Next, using GO, in the gene's background knowledge we also included its generalized GO terms and information about its interaction with other genes.

In traditional machine learning, examples are expected to be described by a tuple of values corresponding to some predefined, fixed set of attributes. Note that a gene annotation does not straightforwardly correspond to a fixed attribute set, as it has an inherently relational character. For example, a gene may be related to a variable number of cell processes, can play a role in variable number of regulatory pathways etc. This imposes 1-to-many relations which are hard to be elegantly captured within an attribute set of a fixed size. Furthermore, a useful piece of information about a gene g may for instance be expressed by the following feature:

gene g interacts with another gene whose functions include protein binding.

In summary, we are approaching the task of subgroup discovery from a relational data domain. For this purpose we employ the methodology of relational subgroup discovery proposed in [7, 13] and implemented in the RSD[4] algorithm. Using RSD, we were able to discover knowledge such as

The expression of genes coding for proteins located in the integral-to-membrane cell component, whose functions include receptor activity, has a high correlation with the BCR class of acute lymphoblastic leukemia (ALL) and a low correlation with the other classes of ALL.

The RSD algorithm proceeds in two steps. First, it constructs a set of relational features in the form of conjunctions of first order logic atoms. The entire set of features is then viewed as an attribute set, where an attribute has the value true for a gene (example) if the gene has the feature corresponding to the attribute. As a result, by means of relational feature construction we achieve the conversion of relational data into attribute-value descriptions. In the second step, groups of genes are searched, such that each group is represented as a conjunction of selected features. The subgroup discovery algorithm employed in this second step is an adaptation of the popular propositional rule learning algorithm CN2 [1].

[4] http://labe.felk.cvut.cz/železny/rsd/rsd.pdf

2.1 Relational feature construction

The feature construction component of RSD aims at generating a set of relational features in the form of relational logic atom conjunctions. For example, the feature exemplified informally in the previous section has the following relational logic form:

$$\text{interaction}(g,G),\ \text{function}(G,\text{protein\_binding})$$

Here, upper cases denote existentially quantified variables and g is the key term that binds a feature to a specific example (here a gene). The user specifies a grammar declaration which constraints the resulting set of constructed features. RSD accepts feature language declarations similar to those used in the inductive logic programming system Progol [9].

A remark is needed concerning the way constants (such as *protein_binding*) are employed in features. RSD extracts them automatically from the training data. For each constant-free feature, a number of different features are generated, each corresponding to a possible replacement of the combination of the indicated variables with constants. RSD then only proceeds with those combinations of constants which make the feature true for at least a pre-specified number of examples. Finally, to evaluate the truth value of each feature for each example for generating the attribute-value representation of the relational data, the first-order logic resolution procedure is used, provided by a Prolog language engine.

2.2 Subgroup Discovery

Subgroup discovery aims at finding population subgroups that are statistically "most interesting", e.g., are as large as possible and have the most unusual statistical characteristics with respect to the property of interest [12].

Rule learning, as implemented in RSD, involves two main procedures: the search procedure that performs search to find a single subgroup discovery rule, and the control procedure (the weighted covering algorithm) that repeatedly executes the search in order to induce a set of rules. Description of the two procedures, described in [7], are omitted for space constraints.

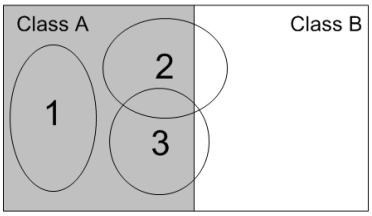

Fig. 1. Description of discovered subgroups can cover: a) individuals from only one class (1), b) individuals from other classes (2) and c) individuals covered also by other subgroups (2,3)

3 Experiments

This section presents a statistical validation of the proposed methodology. We aim at evaluating the properties of the secondary, descriptive learning task. Namely, we wish to determine if the high descriptive capacity pertaining to the incorporation of the expressive relational logic language incurs a risk of *descriptive overfitting*, i.e., a risk of discovering fluke subgroups. We thus aim at measuring the discrepancy of the quality of discovered subgroups on the training data set on one hand and an independent test set on the other hand. We will do this through the standard 10-fold stratified cross-validation regime. The specific qualities measured for each set of subgroups produced for a given class are average *precision* (PRE) and *recall* (REC) values among all subgroups in the subgroup set.

3.1 Materials and methods

We apply the proposed methodology on two problems of predictive classification from gene expression data.

The first was introduced in [3] and aims at distinguishing between samples of ALL and AML from gene expression profiles obtained by the Affymetrix HU6800 microarray chip, containing probes for 6817 genes. The data contains 73 class-labeled samples of expression vectors. The second was defined in [10]. Here one tries to distinguish among 14 classes of cancers from gene expression profiles obtained by the Affymetrix Hu6800 and Hu35KsubA microarray chip, containing probes for 16,063 genes. The data set contains 198 class-labeled samples.

To access the annotation data for every gene considered, it was necessary to obtain unique gene identifiers from the microarray probe identifiers available in the original data. We achieved this by querying Affymetrix site[5] for translating probe ID's into unique gene ID's. Knowing the gene identifiers, information about gene annotations and gene interactions can be extracted from Entrez gene information database[6]. We developed a program script [7] in the Python language, which extracts gene annotations and gene interactions from this database, and produces their structured, relational logic representations which can be used as input to RSD.

In both data sets, for each class c we first extracted a set of discriminative genes $G_C(c)$. In our experiments we used a measure of correlation $P(g, c)$, used by [3], that emphasizes the "signal-to-noise" ratio in using gene g as predictor for class c. $P(g, c)$ is computed by the following procedure:

Let $[\mu_1(g), \sigma_1(g)]$ and $[\mu_2(g), \sigma_2(g)]$ denote the means and standard deviations of log of the expression levels of gene g for the samples in class c and samples in all other classes, respectively.

[5] www.affymetrix.com/analysis/netaffx/
[6] ftp://ftp.ncbi.nlm.nih.gov/gene/
[7] This script is available on request to the first author

Let $P(g,c) = \frac{\mu_1(g)-\mu_2(g)}{\sigma_1(g)+\sigma_2(g)}$, which reflects the difference between the classes relative to the standard deviation within the classes. Large values of $|P(g,c)|$ indicate a strong correlation between the gene expression and the class distinction, while the sign of $P(g,c)$ being positive or negative correspond to g being more highly expressed in class c or in other classes. Unlike a standard Pearsonćorrelation coefficient, $P(g,c)$ is not confined to the range $[-1,+1]$. The set of informative genes for class c, $G_C(c)$ of size n, consists of the n genes having the highest $|P(g,c)|$ value. If we have only two classes, then $G_C(c_1)$ consist of genes having the highes $P(g,c_1)$ values, and $G_C(c_2)$ consists of genes having the highest $P(g,c_2)$ values.

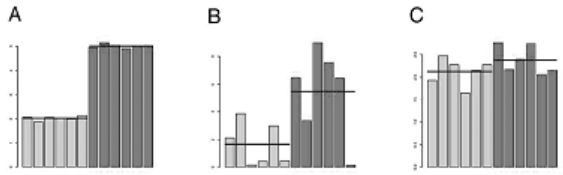

Fig. 2. Expression of three genes (A, B and C) for five patients of class 1 and five patients of class 2. The class distinction is represented by an idealized gene A, in which the expression level is uniformly low in class 1 and uniformly high in class 2. Gene B is well correlated with the class distinction. Gene C is also well correlated, but we are interested in genes that have significant difference in its expression between classes, like A and B.

For both problems we selected 50 discriminative genes[8]. The average value of correlation coefficient, $|P(g,c)|$, of selected discriminatory genes for each class/problem are listed in Table 1. The usage of the gene correlation coefficient is twofold. In the first part of the analysis, for a given class it is used for selection of discriminative genes, and in the second part as initial weight of the example-genes for the meta-mining procedure where we try to describe these discriminative genes. In the second mining task RSD will prefer to group genes with large weights, so these genes will have enough weight to be grouped in several groups with different descriptions.

After the selection of sets of discriminatory genes, $G_C(c)$ for each $c \in C$, these sets were merged and every gene coming from $G_C(c)$ was class labeled as c. Now RSD was run on these data, with the aim to find as large as possible and as pure (in terms of class labels) as possible subgroups of this population of example-genes, described by relational features constructed from GO and gene interaction data.

[8] This is a usual number of selected genes in the context of microarray data classification with SVM or voting algorithms.

Table 1. Average(AVG), maximal(MAX) and minimal(MIN) value of $\{|P(g,c)|g \in G_C(c)\}$ for each task and class c.

TASK	CLASS	AVG	MAX	MIN
ALL-AML	ALL	0.75	1.25	0.61
	AML	0.76	1.44	0.59
MULTI	BREAST	1.06	1.30	0.98
CLASS	PROSTATE	0.91	1.23	0.80
	LUNG	0.70	0.99	0.60
	COLORECTAL	0.98	1.87	0.73
	LYMPHOMA	1.14	2.52	0.87
	BLADDER	0.88	1.15	0.81
	MELANOMA	1.00	2.60	0.73
	UTERUS	0.82	1.32	0.71
	LEUKEMIA	1.35	1.75	1.18
	RENAL	0.81	1.20	0.70
	PANCREAS	0.75	1.08	0.67
	OVARY	0.70	1.04	0.60
	MESOTHELIOMA	0.90	1.91	0.74
	CNS	1.38	2.17	1.21

3.2 Results

The discovered regularities have very interesting biological interpretations.

In ALL, RSD has identified a group of 23 genes, described as: component(G, 'nucleus') AND interaction(G,B),process(B,'regulation of transcription, DNA-dependent'). The products of these genes, proteins, are located in the nucleus of the cell, and they interact with genes that are included in the process of regulation of transcription. In AML, RSD has identified several groups of overexpressed genes, located in the membrane, that interact with genes that have 'metal ion transport' as one of their functions.

In breast cancer, RSD has identified a group of genes (described as process(G,'regulation of transcription'), function(G,'zinc ion binding')) containing five genes (Entrez Gene id's: 4297, 51592, 91612, 92379, 115426) whose under expression is a good predictor for that class. These genes are simultaneously involved in regulation of transcription and in zinc ion binding. Zinc is a cofactor in protein-DNA binding, via a "zinc finger" domain (id 92379). Second, zinc is an essential growth factor and a zinc transporter associated with metastatic potential of estrogen positive breast cancer, termed LIV-1, has been described [4]. A separate group of genes involved in ubiquitin cycle (process(G,'ubiquitin cycle')) was identified in breast cancer, (Entrez id's: 3093, 10910, 23014, 23032, 25831, 51592, 115426). The role of ubiquitin in a cell is to recycle proteins. This is of a paramount importance to the overall cellular homeostasis, since inappropriately active proteins can cause cancer [8]. This is the example where one gene, id: 115426, was included in two groups with different descriptions.

In CNS (central nervous system) cancer, we discovered two important groups concerning neurodevelopment (description: process(G,'nervous system develop-

ment'), Entrez Gene id's: 333, 1400, 2173, 2596, 2824, 3785, 4440, 6664, 7545, 10439, 50861), and immune surveillance (Entrez Gene id's: 199, 1675, 3001, 3108, 3507, 3543, 3561, 3588, 3683, 4046, 5698, 5699, 5721, 6352, 9111, 28299, 50848, 59307). The genes in the first/second group are over/under-expressed (respectively) in CNS. As for the former, reactivation of genes relevant to early development (i. e., ineffective recapitulation of embryonal or fetal neural growth at a wrong time) is a hallmark of the most rapidly growing tumors (id's 3785 and 10439 specific to neuroblastoma). The latter illustrates the common clinical observation that immune deficiency (subnormal expression of genes active in immune response shown in this work) creates a permissive environment for cancer persistence. Thus, both major themes of malignant growth are represented in this example: active unregulated growth and passive inability to clear the abnormal cells.

In addition, we subjected the RSD algorithm to 10-fold stratified cross-validation on both classification tasks. Table 2 shows the PRE and REC values (with standard deviation figures) results for the two respective classification tasks. Overall, the results show only a small drop from the training to the

Table 2. Precision-recall figures and average sizes of found subgroups of differentially expressed genes (DEG), for the ALL/AML and multi-class problem, obtained through 10-fold cross-validation.

TASK	DATA	PRE(st.dev.)	REC(st.dev.)	AVG. SIZE
ALL-AML DEG	Train	0.96(0.01)	0.18(0.02)	12.07
	Test	0.76(0.06)	0.12(0.04)	
MULTI-CLASS DEG	Train	0.51(0.03)	0.15(0.01)	8.35
	Test	0.42(0.10)	0.10(0.02)	

testing set in terms of both PRE and REC, suggesting that the number of discriminant genes selected (Table 1) was sufficient to prevent overfitting. In terms of *total* coverage, RSD covered more then $\frac{2}{3}$ of the preselected discriminative genes (in both problems), while $\frac{1}{3}$ of the preselected gene were not included in any group. One interpretation is that they are not functionally connected with the other genes, but were selected by chance. This information can be used in the first phase of the classification problem, feature selection, by choosing genes that were covered by some subgroup. That will be the next step in our future work, using the proposed methodology as feature (gene) selection mechanism.

4 Discussion

In this paper we presented a method that uses gene ontologies, together with the paradigm of relational subgroup discovery, to help find patterns of expression for genes with a common biological function that correlate with the underlying biology responsible for class differentiation. Our methodology proposes to

first select a set important discriminative genes for all classes and then finding compact, relational descriptions of subgroups among these genes.

Since genes frequently have multiple functions that they may be involved in, they may under some of the conditions exhibit the behavior of genes with one function and in other conditions exhibit the behavior of genes with a different function. Here subgroup discovery is effective at selecting a specific function. The same gene can be included in multiple subgroup descriptions (gene id: 115426 in breast cancer), each emphasizing the different biological process critical to the explanation of the underlying biology responsible for observed experimental results. Unlike other tools for analyzing gene expression data that use gene ontologies, which report statistically significant single GO terms and do not use gene interaction data, we are able to find a set of GO terms (the first reported group of genes, for breast cancer, is described with two GO terms), that cover the same set of genes, and we use available gene interaction data to describe features of genes that can not be represented with other approaches (the third reported group, for ALL).

However, this approach of translating a list of differentially expressed genes into subgroups of functional categories using annotation databases suffers from a few important limitations. The existing annotations databases are incomplete, only a subset of known genes is functionally annotated and most annotation databases are built by curators who manually review the existing literature. Although unlikely, it is possible that certain known facts might get temporarily overlooked. For instance, [6] found references in literature published in the early 90s, for 65 functional annotations that are yet not included in the current functional annotation databases. However, many such annotations are often made at very high-level GO terms, which limit their usefulness.

Despite the current imperfectness of the available ontological background knowledge, the presented methodology was able to discover and compactly describe several gene groups, associated to specific cancer types, with highly plausible biological interpretation. We thus strongly believe the presented approach will significantly contribute 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 in the near future.

Acknowledgment

The research of Igor Trajkovski and Nada Lavrač is supported by the Slovenian Ministry of Higher Education, Science and Technology. Filip Železný is supported by the Czech Academy of Sciences through the project KJB201210501 Logic Based Machine Learning for Analysis of Genomic Data. Jakub Tolar is supported by the Children's Cancer Research Fund, University of Minnesota Cancer Center and Department of Pediatrics.

References

1. Clark, P. & Niblett T. (1989). The CN2 induction algorithm. Machine Learning, pages 261-283.
2. Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA, 95:25, 14863-14868.
3. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M. et al. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science, 286:5439, 531-537.
4. Kasper, G. et al. (2005). Expression levels of the putative zinc transporter LIV-1 are associated with a better outcome of breast cancer patients. Int J Cancer. 20;117(6):961-73.
5. Khatri, P. et al. (2002). Profiling gene expression using Onto-Express. Genomics, 79, 266-270.
6. Khatri, P. & Draghici S. (2005). Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics, 21(18):3587-3595
7. Lavrač, N., Železný F. & Flach P. (2002). RSD: Relational subgroup discovery through first-order feature construction. In Proceedings of the 12th International Conference on Inductive Logic Programming, pages 149-165.
8. Mani, A. & Gelmann E.P. (2005). The ubiquitin-proteasome pathway and its role in cancer. Journal of Clinical Oncology. 23:4776-4789
9. Muggleton, S. (1995). Inverse entailment and Progol. New Generation Computing, Special issue on Inductive Logic Programming, 13(3-4):245-286.
10. Ramaswamy, S., Tamayo P., Rifkin R. et al. (2001). Multiclass cancer diagnosis using tumor gene expression signatures, PNAS 98 (26) 15149-15154.
11. Raychaudhuri, S., Schtze,H.S. & Altman,R.B. (2003). Inclusion of textual documentation in the analysis of multidimensional data sets: application to gene expression data. Machine Learn., 52, 119-145
12. Wrobel, S. (1997). An algorithm for multi-relational discovery of subgroups. In Jan Komorowski and Jan Zytkow, editors, Proceedings of the First European Symposion on Principles of Data Mining and Knowledge Discovery, 78-87.
13. Železný, F. & N. Lavrač. (2006). Propositionalization-Based Relational Subgroup Discovery with RSD. Machine Learning 62(1-2):33-63. Springer 2006.