Towards Knowledge Discovery from cDNA Microarray Gene Expression Data

Jan Komorowski[1], Torgeir R. Hvidsten[1], Tor-Kristian Jenssen[1],
Dyre Tjeldvoll[1], Eivind Hovig[2], Arne K. Sandvik[3], and Astrid Lægreid[3]

[1] Knowledge Systems Group, Department of Information and Computer Science,
Norwegian University of Science and Technology, 7491 Trondheim, Norway
jan.komorowski@idi.ntnu.no, http://www.idi.ntnu.no/grupper/KS-grp/
[2] Department of Tumor Biology, Institute for Cancer Research,
The Norwegian Radium Hospital, Oslo, Norway
[3] Department of Physiology and Biomedical Engineering,
Norwegian University of Science and Technology, Trondheim, Norway

Abstract. The advent of the so-called cDNA microarrays has offered the first possibility to obtain a global understanding of biological processes in living organisms by simultaneous readouts of tens of thousands of genes. Initial experiments suggest that genes with similar function have similar expression patterns in microarray experiments. Until now, most approaches to computational analysis of gene expressions have used unsupervised learning. Although in some cases unsupervised methods may be sufficient, the complexity of the biological processes is so high that it is unlikely that purely syntactical analyses are capable of fully exploiting the richness of the microarray data. In addition, it seems natural to re-use the existing biological (background) knowledge.

In this paper, we present some elements of a methodology for knowledge discovery from microarray experiments. Two source of bio-medical knowledge are used: Ashburner's gene ontology and our own literature-derived network of gene-gene relations obtained by analysing Medline citation records. Predictive models can be induced and their classification quality validated through the ROC/AUC analysis and applied to provide hypotheses regarding the function of unclassified genes. The methodology has been so far tested on publicly available gene expression data and its results evaluated by molecular biologists and medical researchers.

1 Introduction

Until recently, molecular biological research has been investigating the genetic and biochemical mechanisms that underly inter- and intra-cellular organisation using experimental methods that produce extremely low levels of information in relation to the huge number of parameters that govern living systems. With the advent of the so-called microarray technology [1], it is now possible to observe in parallel several thousands of genes. Coupled with the large scale data generating programs, such as the Human Genome Project, very large sources of information about the living systems are becoming available. However, the avalanche of data

D.A. Zighed, J. Komorowski, and J. Żytkow (Eds.): PKDD 2000, LNAI 1910, pp. 470–475, 2000.

is overwhelming and it is now realized that new methods to process this data are needed.

The microarrays provide a view onto cellular organisation of life through quantitative data on mRNA expression levels. These may be used to *discover* the biological function of genes [2]. It has been shown in early experiments [4] that genes of similar function tend to produce similar patterns in microarray experiments. With one exception, [3], most of the approaches uses unsupervised learning to discover the functional classes in expression data. In the unsupervised methods, a distance function is used to define similarity of gene expressions and a clustering algorithm is applied to find groups of genes. Hierarchical clustering [4] and self-organising maps [5] have been used.

The main thesis of this paper is that knowledge discovery from gene expressions requires new approaches to knowledge discovery, if the complexity barrier is to be ever broken. In our approach we move away from strictly unsupervised methods that seem to dominate the current state of art in this area, and show how to use various forms of background knowledge to discover and validate new knowledge. Individual genes are first *annotated* using available (background) knowledge such as various genomic databases (e.g. SWISSPROT), literature repositories (e.g. Medline) and ontologies (e.g. Ashburner's Gene Ontology [6]). Gene expressions are clustered according to some method, preferably in a way that lends itself to a biological interpretation. Having clusters with annotations, it is possible to validate their correctness. The annotations also effectively provide a training sample from which a model can be induced and whose quality of classification can be evaluated in some standard way (e.g. with the ROC/AUC analysis) and applied to classify unknown genes.

2 Datamining Methods

Our methodology is illustrated on a publicly available data set [7] previously analysed by Iyer et al. [8]. Iyer et al. studied the human fibroblast response to serum. This cell has a pivotal structural role in connective tissue and in important processes such as wound healing. The temporal changes in mRNA level of 8613 human genes were measured at 12 time points in the time period between 0 minutes and 24 hours after serum stimulation. A subset of 517 genes whose expression changed substantially in response to serum was selected for further analysis. An agglomerative implementation of the hierarchical clustering method was used to cluster the 517 genes into groups on the basis of the similarity of their expression profiles over the entire period of 24 hours. Ten clusters were identified containing 452 of the 517 genes.

Template-based clustering. Requiring that gene expressions be similar over the entire period of 24 hours seems to be somewhat unrealistic for the process at hand. Instead, we proposed a clustering that is defined on time sub-intervals. Since the same gene may be involved in different processes, we wanted to obtain clusters which include genes with expression profiles that show similarity in one time period but be rather different in another one. To this end we introduced a method called *template-based clustering.* Constraints defining the exact properties of a gene matching one template were prepared according to the uncertainties of the data. These templates have been used on all possible combinations of subintervals in order to assign the genes to clusters (see Fig. 1). According to domain experts, the resulting set of clusters describes well the dynamics in the data. Most importantly, the clusters display characteristic temporal changes of gene expression levels during a very large number of subintervals. This large variety of patterns of changes in gene expression reflects the expected complexity of molecular biological events during the fibroblast serum and thus, significantly simplifies the task of biological interpretation.

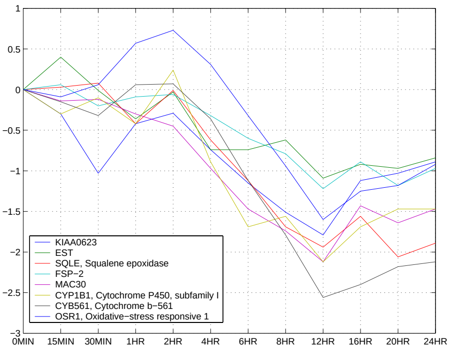

Fig. 1. A sample template-based cluster: Genes with a *decreasing* expression profile between 2 hours and 12 hours.

From clusters to decision classes. The need for a structured way of relating genes to each other based on their function or process has prompted the creation of *gene ontologies.* A gene ontology is a hierarchical structure where parent nodes give a more general description of a gene than their children. The leaf nodes give an accurate description of each individual gene. An ontology is a formalised source of high quality biological knowledge that will be important in much of the research on gene expressions. In particular, Ashburner's Gene Ontology [6] was developed using two model organisms, fruit fly (*Drosophila melanogaster*) and ordinary yeast (*Saccharomyces cerevisiae*). Surprisingly many genes in these simple organisms are homologous to human genes. This can be exploited simply by placing the human gene in the same location in the ontology as its homologous counterpart in these organisms. The information associated with a gene as a result of its location in the ontology is referred to as a gene *annotation.*

The annotations provide essentially a way of testing the quality, or rather classificatory power of a cluster before making assumptions about unknown genes located in it. We prove the point by means of the example in the earlier Fig. 1. This cluster contains eight genes of which four are named. Two of the named genes, Squalene epoxidase and Cytochrome P540, have annotation "cholesterol bio-synthesis". Another, CYB561, has annotation "electron transport" which may indicate a possible role in cholesterol synthesis. The gene OSR1, Oxidative-stress responsive-1, encodes a protein induced by oxidative stress, which is a process related to the redox processes involved in cholesterol bio-synthesis. EST, KIAA0623, FSP-2 and MAC30 encode proteins of unknown functions. To the biological domain experts, the relation between the annotated functions of the named genes is striking. The clustering of genes involved in redox processes and more specifically cholesterol synthesis, which is detected by this method may suggest that some of the genes with unknown function may also be involved in similar processes. This and several other related results are new in comparison with Iyer et al. [8].

Validation of annotations – pubgene. A large part of the biological knowledge needed to annotate genes is not organised in ontologies, but exists as text documents. We present here a method for mining gene-gene relations from such documents. These relations may be used both for validating existing annotations and clusters, and obtaining new annotations.

We have created a database of gene-gene relations by analysing Medline citation records. Two genes are related if they have been mentioned in the same article. Hence, the gene-gene links constitute a literature-derived network of genes. Our local list of human genes has been compiled by collecting and structuring data from public databases available through ftp or http. The main source of information has been the list of approved genes/symbols from the HUGO Nomenclature Committee [9,10].

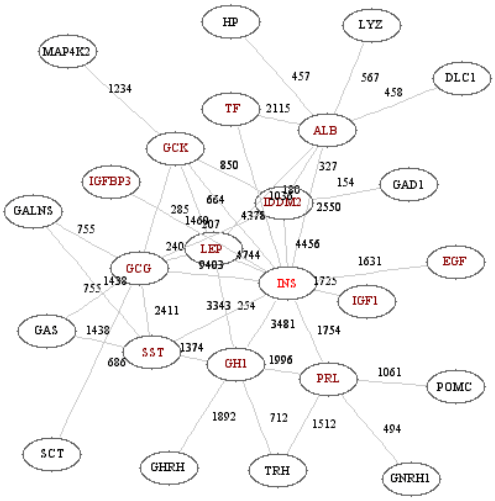

Fig. 2. A sample neighbourhood-network for the gene insulin (INS).

In the graphical display on our website [11] the neighbourhood of each gene is presented as a graph centered around the gene (see Fig. 2). Neighbours of the gene appear as nodes with edges weighted by the number of co-citations. The user can move around in the network by changing the gene in focus by clicking neighbours in the graph.

Altogether we compiled a list containing 14961 genes. Based on a subset of Medline containing all publications from the years 1992 - 1999, we found citations for 7561 (50%). Of these, 6978 had one or more co-citations (47%). The total number of articles was about 3 million, of which approximately 15% gave a hit for one or more genes.

Our work demonstrates that literature co-occurrence can be exploited to extract biologically meaningful information. Moreover, we also show that the approach can be carried out on a large scale. Details are described in [12].

Learning gene classifiers. Most processes that involve genes can be only described by a combination of several templates over a number of time sub-intervals. Consequently, we mine sets of template–interval pairs that together describe a process and obtain minimal sets of discerning pairs. Then, a set of IF-THEN rules over a training set of annotated genes is induced. This rule set holds both the descriptive knowledge about the temporal aspect of a given process and the predictive knowledge that can be used to classify genes for which no process is known (unknown genes). In order to induce a classifier, we use the rough set framework [13] for rule induction and voting for classification as implemented in the ROSETTA system [14,15]. The classification quality is validated through the ROC/AUC analysis in order to estimate the expected correctness of the classification of genes for which no process is known.

In the fibroblast data we have identified about 300 known genes, most of them having more than one annotation. These genes seem to be involved in over 20 different processes, of which 16 include sufficiently many genes (e.g. 10 genes) to induce a predictive model. Early results show that we in fact are capable of inducing a model that recognise these 16 processes. One example is the genes involved in *blood coagulation* which in a 5 × 3-fold cross-validation setting can be classified with an AUC value of 0.90. A thorough discussion of these results is to be found in [16].

3 Conclusions

A suite of knowledge discovery tools that support gene expression analysis, annotation and visualisation has been created (e.g. [17,18]). We have made significant steps towards an automatic use of background knowledge in discovery from gene expressions. It should be noticed that standard methods of validating clusters against background knowledge are not feasible here due to such factors as, for instance, a very large number of genes, hierarchical structure of this knowledge and the uncertainty of the genomic information.

The tools are now in use in our project on developing genomic classifiers from microarray data and background knowledge. Biologists and medical researchers find the tools particularly interesting since paradigms of biomedical background knowledge are often well captured. We have corroborated our hypothesis that knowledge-based tools are likely to gain an edge in knowledge discovery in such complex field as molecular biology.

References

1. Schena M, Shalon D, Davis R and Brown PO, Quantitative monitoring of gene expression patterns with a cDNA microarray, *Science*, 270:467–470, 1995.
2. Deboucek and Goodfellow, *Nature Genetics*, 21 (1 Suppl):48–52, 1999.
3. Brown MPS, Grundy WN, Cristianini N, Sugnet CW, Furey TS, Ares M and Haussler D, Knowledge-based analysis of microarray gene expression data by using support vector machines, *PNAS*, No. 1, Vol. 97:262–267, 1999.
4. Eisen M, Spellman P, Brown P and Botstein D, Cluster analysis and display of genome-wide expression pattern, *Proc. Natl. Acad. Sci. USA*, 95:1464–1480, 1998.
5. Kohonen T, The Self-Organizing Map, *Proceedings of the IEEE*, Vol. 78, No. 9:1464–1480, 1990.
6. The Ashburner Gene Ontology homepage, `http://genome-www.stanford.edu/GO/`.
7. The transcriptional program in the response of human fibroblasts to serum on the WEB `http://genome-www.stanford.edu/serum/`.
8. Iyer VR, Eisen MB, Ross DT, Schuler G, Moore T, Lee JCF, Trent JM, Staudt LM, Dudson Jr. J, Boguski MS, Lashkari D, Shalon D, Botstein D and Brown PO, The transcriptional program in the response of human fibroblasts to serum, *Science*, 283:83–87, 1999.
9. White JA, et al., Guidelines for human gene nomenclature, *Genomics*, 45(2):468–471, Oct 15 1997.
10. White JA, et al., The HUGO Nomenclature Committee home page `http://www.gene.ucl.ac.uk/nomenclature`.
11. Jenssen TK, The PubGene home page `http://www.idi.ntnu.no/grupper/KS-grp/microarray/pubgen/genes.cgi`.
12. Jensen T-K, Lægreid A, Komorowski J and Hovig E, *A literature network of human genes for high-throuput gene-expression analysis*, submitted for publication, June 2000.
13. Pawlak Z, Rough Sets, *International Journal of Computer and Information Sciences*, Vol. 11:341–356,1982.
14. Komorowski J, Skowron A and Øhrn A, *The Rosetta system*, to appear in *Handbook of Data Mining and Knowledge Discovery*, (W. Klösgen, J. Zytkow, Eds.), Oxford University Press, 2000.
15. Komorowski J and Øhrn A, Modelling Prognostic Power of Cardiac Tests Using Rough Sets, *Artificial Intelligence in Medicine*, Vol. 15, No. 2:167–191, 1999.
16. Hvidsten TR, Komorowski J, Lægreid A and Sandvik, *Discovery of gene functions and processes from gene expressions and ontologies*, submitted for publication, July 2000.
17. Hvidsten TR, Jenssen T-K, Komorowski J, Lægreid A, Sandvik A and Tjeldvoll D, *Template-based gene expression analysis*, in "Currents in Computational Molecular Biology – RECOMB 2000", edited by S. Miyano, R. Shamir and T. Takagi, pp. 10–11, ISBN 4-946443-61-4, Universal Academy Press, Inc, April 8-11, 2000, Tokyo, Japan.
18. Jenssen T-K, Lægreid A, Komorowski J and Hovig E, *PubGene: Discovering and visualising gene-gene relations*, in "Currents in Computational Molecular Biology – RECOMB 2000", edited by S. Miyano, R. Shamir and T. Takagi, pp. 48–49, ISBN 4-946443-61-4, Universal Academy Press, Inc, April 8-11, 2000, Tokyo, Japan.