BIOINFORMATICS

Vol. 19 no. 9 2003, pages 1116–1123
DOI: 10.1093/bioinformatics/btg047

Learning rule-based models of biological process from gene expression time profiles using Gene Ontology

Torgeir R. Hvidsten[1, 3], *Astrid Lægreid*[2] *and Jan Komorowski*[1, 3,*]

[1]*Department of Computer and Information Science, Norwegian University of Science and Technology, N-7491 Trondheim, Norway,* [2]*Department of Clinical and Molecular Medicine, Norwegian University of Science and Technology, N-7489 Trondheim, Norway and* [3]*The Linnaeus Centre for Bioinformatics, Uppsala University, Uppsala, Sweden*

Received on April 30, 2002; revised on November 8, 2002; accepted on November 12, 2002

ABSTRACT

Motivation: Microarray technology enables large-scale inference of the participation of genes in biological process from similar expression profiles. Our aim is to induce classificatory models from expression data and biological knowledge that can automatically associate genes with novel hypotheses of biological process.

Results: We report a systematic supervised learning approach to predicting biological process from time series of gene expression data and biological knowledge. Biological knowledge is expressed using gene ontology and this knowledge is associated with discriminatory expression-based features to form minimal decision rules. The resulting rule model is first evaluated on genes coding for proteins with known biological process roles using cross validation. Then it is used to generate hypotheses for genes for which no knowledge of participation in biological process could be found. The theoretical foundation for the methodology based on rough sets is outlined in the paper, and its practical application demonstrated on a data set previously published by Cho *et al.* (*Nat. Genet.*, **27**, 48–54, 2001).

Availability: The Rosetta system is available at http://www.idi.ntnu.no/∼aleks/rosetta

Contact: Jan.Komorowski@lcb.uu.se

Supplementary information: http://www.lcb.uu.se/∼hvidsten/bioinf_cho/

INTRODUCTION

Microarray technology (Schena *et al.*, 1995) makes it possible to measure levels of gene expression (mRNA abundance) for tens of thousands of genes in parallel. This enables large-scale inference of the participation of genes in biological processes from similar expression profiles. In this research, Gene Ontology (The Gene Ontology Consortium, 2000, http://genome-www.stanford.edu/GO/) is an important source of structured knowledge of biological roles of proteins (gene products). Gene ontology divides the general notion of 'function' into *molecular function*, *biological process* and *cellular component*. Experimental work shows that among these three categories, biological process agrees best with the hypothesis that similar expressions indicate a functional relation (Brown *et al.*, 2000). Genes encoding proteins involved in the same process tend to be co-regulated and Pilpel *et al.* (2001) show that there is a correlation in expression profiles between genes with the same motifs in their promoter region (Lockhart and Winzeler, 2000).

Techniques for extracting biological knowledge of process from microarray expression data can conceptually be divided into *class discovery* and *class prediction* or, equivalently, unsupervised and supervised learning. Unsupervised methods organize expression data by clustering genes with similar patterns of expression. Supervised methods use genes coding for proteins with known biological process roles as training examples. A model is induced from these examples, defining the relationship between gene expression and biological process.

The most used algorithm for studying biological process from microarray data is agglomerative hierarchical clustering (e.g. Eisen *et al.*, 1998; Iyer *et al.*, 1999), which starts with the individual genes in separate clusters, and successively merges the two most similar clusters until all genes have been grouped into one large cluster. The result is visualized using a binary tree called a dendrogram combined with a heat plot showing how genes with similar expression profiles are clustered together. These plots are often accompanied with heat plots of genes coding

[*] To whom correspondence should be addressed.

for proteins participating in similar biological processes (Eisen *et al.*, 1998; Iyer *et al.*, 1999). Although these studies have been successful in showing that genes participating in the same biological processes have similar expression profiles, there are several reasons why clustering analysis cannot solve the core issues of modeling biological process from gene expression data (Shatkay *et al.*, 2000). Genes that are biologically related often show a strong anti-correlation in their expression profiles and hence will not be clustered together. Also, clustering genes into disjoint clusters will not capture the fact that many gene products participate in more than one biological process. Sherlock (2000) also observes that most studies using clustering techniques do not report any measure of whether the overlap between biologically related genes and genes in expression clusters is greater than what would be expected by chance. Hence, no quantitative measures are given indicating to what degree we can trust process assignments to unknown gene products using these clusters. Recently, Cho *et al.* (2001) have reported statistically significant over- and under-representation of biologically related genes in expression clusters. Brown *et al.* (2000) used supervised learning with support vector machines to learn six different classes from annotated yeast genes. The classification quality was estimated using cross validation, and the model was used to provide hypotheses of participation in biological processes for 15 unknown genes.

In this paper we take a systematic supervised learning approach to predict the participation of gene products in biological process from time series of gene expression data labeled using gene ontology (GO). Known genes are grouped into classes extracted from their GO annotations and constitute training examples. Numerical values of expression levels are replaced by templates describing qualitative change over sub-intervals. A rule model is induced based on this language of templates using the rough set framework (Pawlak, 1982, 1991). The model is evaluated using cross validation and finally used to provide novel hypotheses of participation in biological processes for both known and unknown genes. The methodology is demonstrated on a publicly available data set published by Cho *et al.* (2001, http://www.salk.edu/docs/labs/chipdata/). We can report high classification quality in terms of cross validation estimates. The method is fully implemented in the Rosetta system (Komorowski *et al.*, 2002, http://www.idi.ntnu.no/∼aleks/rosetta/), a publicly available toolkit for data mining and knowledge discovery (Fayyad *et al.*, 1996) using rough sets.

METHOD

Pawlak's rough set theory constitutes a mathematically sound framework for inducing minimal decision rules from labeled examples. Each ‘if-then’ rule identifies a minimal set of features discriminating one particular example from examples in all other classes. The set of rules from all examples constitutes a classificatory model capable of classifying new examples. A detailed, relatively formal, description of each step in the method is given in the next four sections.

Combining microarray data and knowledge of biological process

Time series microarray experiments provide snapshots of the state of a cell in terms of quantitative measures of gene expression levels (i.e. relative mRNA level) during some biological response. Let U be the universe of participating genes and let $T = \{t_1, t_2, \ldots, t_m\}$ be a set of functions such that $t_i : U \to \mathcal{R}$ maps each gene to a numerical value measuring its expression level at time point i. Furthermore, let an *information system* $\mathcal{M} = (U, T)$ be a table with this data. We now wish to combine the data with annotations of biological processes.

Gene ontology provides a structured language for protein function and is therefore a natural tool for representing such knowledge. Formally, Gene ontology is a directed acyclic graph (DAG) $GO = (V, E)$, where V is a set of protein function descriptions (*GO terms*) and E is a binary relation on V such that proteins with functions described by v_j are a subset of proteins with functions described by v_i, denoted $v_j \preccurlyeq v_i$, if and only if there exists a path $(v_i, v_{i+1}, \ldots, v_{j-1}, v_j)$ such that $(v_{m-1}, v_m) \in E$ for $m = i + 1, i + 2, \ldots, j - 1, j$.

Again let U be the universe of genes and let $a : U \to V^k$ be a function annotating each gene with a set of GO-terms at the most specific level of the biological process part of gene ontology. We will then assume that we want to model a set of predefined GO terms $G = \{v_1, v_2, \ldots, v_m\}$. Hence we move the annotations to the appropriate level of generalization to obtain a set of gene–GO term pairs $A = \{(x, v) \mid x \in U \text{ and } u \in a(x) \text{ and } u \preccurlyeq v \text{ and } v \in G\}$. Next, we add this biological knowledge to our information system to obtain a *decision system* $\mathcal{M} = (U^d, T \cup \{d\})$. U^d and $d : U^d \to G$ is defined such that $U^d = \{x \in U \mid (x, u) \in A \text{ and } d(x) = u\}$. Elements of U^d are henceforth called *known genes* (or simply examples) while elements of $U^{-d} = \{x \in U \mid x \notin U^d\}$ are called *unknown genes*. Unknown genes are either genes without annotations or genes with annotation outside the scope of our predefined set G.

Given a decision system $\mathcal{A} = (U, A \cup \{d\})$ in general, A is called the set of *conditional attributes* and $d \notin A$ is called the *decision attribute*. The elements of the universe U are called *objects* and sets of objects with similar decision are called *decision classes*. Note that genes in U with more than one annotation in G are represented by one object per annotation in U^d.

Feature synthesis

Machine learning algorithms deal with the examples in a purely syntactical fashion, i.e. as points in a *m*-dimensional space spanned by the measured features (i.e. feature space). However, semantical interpretations of these features can greatly increase the quality of the induced model by using this knowledge to map the data into a new more appropriate feature space.

We know that genes serve as recipes in the synthesis of proteins and that the mRNA levels reflect the amount of protein being produced. However, also other factors influence this amount, such as, how long the mRNA exists in the cell before it is decomposed and how fast it is translated into proteins. Even if proteins *are* being produced, some proteins, as for instance enzymes, might be switched off. Nevertheless, since protein synthesis is an energy consuming task, cells are arranged so that only strictly needed proteins are produced. Changes in expression level should therefore indicate whether more of the relevant gene product is needed or not, and, consequently, whether the relevant biological process is more or less active.

Given decision system $\mathcal{M} = (U^d, T \cup \{d\})$ of labeled expression profiles, we define three templates, *increasing* (*incr*), *decreasing* (*decr*) and *constant* (*const*). Given a gene $x \in U^d$, a template $t \in \{incr, decr, const\}$ and an interval $i \in I = \{(t_i, t_j) \mid 1 \leq i < j \leq |T| \text{ and } t_i, t_j \in T\}$, we define the predicate $\texttt{match}(x, t, i)$ to be true if x matches t in i and false otherwise. We use this to revise our previous decision system to obtain $\mathcal{M} = (U^d, I \cup \{d\})$, where

$$i(x) = \begin{cases} t \in \{incr, decr, const\} & \text{if } \texttt{match}(x, t, i) \\ \varnothing & \text{otherwise} \end{cases} \quad (1)$$

Hence, we represent each expression profile in terms of increasing, decreasing or constant expression levels over sub-intervals. This equips us with a data representation of the desired generality and with the flexibility suitable to describe the complex relationship between gene expression and biological process. The templates and the predicate `match/3` should be tailor made to each application data set, e.g. a gene should match the template *increasing* in some interval if the mRNA level of this gene increases significantly during the interval.

Model induction

Rough set based models are founded on the concept of *indiscernibility*. Given a decision system $\mathcal{A} = (U, A \cup \{d\})$, we define $IND_{\mathcal{A}}(A, x, d)$ to be the set of objects that are indiscernible from x with respect to the attribute set A or are equal to x with respect to the decision attribute d (We do not bother to discern objects from the same class).

From the definition of indiscernibility we derive for each object $x \in U$ the set of *reducts* $RED_{\mathcal{A}}(x, d)$ to be the set of minimal sets of attributes $B \subseteq A$ such that $IND_{\mathcal{A}}(B, x, d) = IND_{\mathcal{A}}(A, x, d)$. Hence, a reduct of x is a minimal set of attributes B with the same discriminatory power as A. Finding the set of all reducts is NP-hard (Skowron and Rauszer, 1992), however, there are heuristics that compute a sufficient number of reducts in an acceptable time. Since real-world data almost always is polluted with noise, methods finding approximate reducts that reveal the underlying, general pattern in the data have also been developed. Two such approaches are *dynamic reducts* (Bazan *et al.*, 1994) and α*-reducts* (Skowron and Nguyen, 1999).

Reducts serve the purpose of synthesizing minimal decision rules of the form $\alpha \to \beta$. The fundamental building blocks for assembling such rules are called *descriptors*. A descriptor is an expression $a = a(x)$, where $a \in (A \cup \{d\})$. Descriptors may be combined in a recursive manner to form more complex formulae such as $F_A(x) = \bigwedge_{a \in A} (a = a(x))$ and $G_A(x) = \bigvee_{j \in \delta_A(x)} (d = j)$. Here, $\delta_A(x) = \{i \mid (\exists y \in U)(y \in IND_{\mathcal{A}}(A, x, d) \text{ and } d(y) = i)\}$ is called the *generalized decision* of x. Minimum decision rules from the decision system $\mathcal{A}$ constitute the set $RUL_{\mathcal{A}} = \bigcup_{x \in U} \{F_B(x) \to G_B(x) \mid B \in RED_{\mathcal{A}}(x, d)\}$.

The decision system $\mathcal{M} = (U^d, I \cup \{d\})$ of expression data and biological knowledge can now be plugged into the above described rough set framework to obtain the predictive rule model $\kappa = RUL_{\mathcal{M}}$. We should, however, be careful when defining indiscernibility. The classical definition of *decision relative* indiscernibility would be $IND_{\mathcal{M}}(I, x, d) = \{y \in U^d \mid (\forall i \in I)(i(x) = i(y)) \text{ or } d(x) = d(y)\}$. However, $\mathcal{M}$ includes empty entries (Equation 1) with an undefined interpretation, and it seems unsatisfactory to induce rules based on these empty entries. We therefore use as an indiscernibility definition $IND_{\mathcal{M}}(I, x, d) = \{y \in U^d \mid [(\forall i \in I)(i(x) = i(y) \text{ or } i(x) = \varnothing)] \text{ or } d(x) = d(y)$. Hence, the empty entry is considered unfit to discern x from other objects. Note that this definition is not symmetric, i.e. $x \in IND_{\mathcal{M}}(I, y, d)$ does not necessarily imply that $y \in IND_{\mathcal{M}}(I, x, d)$.

Model evaluation and application

Let $\hat{d}_\kappa(x)$ be the classification assigned to x by model κ and let $d(x)$ be the true actual classification of x. Then $\hat{d}_\kappa(x)$ takes the form $\hat{d}_\kappa : U \overset{\Phi}{\to} [0, 1] \overset{\theta_\tau}{\to} \{0, 1\}$, where 1 represents a fixed decision class X_1 and 0 represents all other decision classes X_0. $\Phi(x)$ is the certainty of κ that x belongs to the fixed class X_1, while $\theta_\tau(x)$ is a simple threshold function that evaluates to 0 if $\Phi(x) < \tau$, and

1 otherwise. $\Phi(x)$ is realized by a voting procedure that lets each matching rule cast a number of votes in favor of the decision class the rule indicates. The number of votes given to a class is proportional to the support of the rule, i.e. to the number of examples matching the rule.

Most metrics measuring the classification quality of models are τ-dependent, e.g. accuracy, sensitivity and specificity. The *receiver operating characteristic* (ROC) curve (Hanley and McNeil, 1982), however, results from plotting *sensitivity* against (*1 - specificity*) while letting τ vary across the full spectrum of possible values (i.e. $[0, 1]$). The ROC curve may be collapsed into one value by computing the *area under the ROC curve* denoted AUC. This value is threshold-independent, and hence independent of both error costs and prevalence of classes. The AUC value also has a nice and intuitive probabilistic interpretation called the *c-index* (Harrel *et al.*, 1982). Let $S_\kappa(\square) = \{(x_0, x_1) \in X_0 \times X_1 \mid \Phi(x_0) \;\square\; \Phi(x_1)\}$, where $\square \in \{<, >, =\}$. Now,

$$\text{c-index} = AUC = \frac{|S_\kappa(<)| + \frac{1}{2}|S_\kappa(=)|}{|X_0||X_1|} \qquad (2)$$

Hence, the c-index equals the probability that given a pair (x_0, x_1) randomly drawn from $X_0 \times X_1$, the certainty function Φ realized by the classifier κ will rank x_0 and x_1 correctly.

The rule model κ induced from $\mathcal{M} = (U^d, I \cup \{d\})$ will be used to classify the unknown genes. Let $\mathcal{M}^{-d} = (U^{-d}, I)$ be the information system of unknown genes. For each $x \in U^{-d}$, κ will produce a vector $\boldsymbol{\Phi}(\boldsymbol{x}) = \langle \Phi_1, \Phi_2, \ldots, \Phi_{|G|}\rangle$ during classification, where Φ_i is the certainty of κ that class number i is a correct classification of x. The set of classification for x is $C_{\boldsymbol{\Phi},\boldsymbol{\tau}}(x) = \{v_i \in G \mid \Phi_i(x) > \tau_i\}$, where $\boldsymbol{\tau} = \langle \tau_1, \tau_2, \ldots, \tau_{|G|}\rangle$ is a vector of the 'best' thresholds for each class of $\mathcal{M}$. The 'best' thresholds are found by collecting the Φ values for each object and each class for each cross validation iteration and then performing ROC-analysis for each class. If the cost of false negatives and false positives are equally weighted, the thresholds that produced the point closest to (0, 1) on the ROC curves should be used as the 'best' thresholds. Let (*sensitivity*(τ), (1 - *specificity*(τ))) be a point produced by τ on some ROC curve. Minimizing the expression

$$c * (1 - \mathit{specificity}(\tau)) + (1 - \mathit{sensitivity}(\tau)) \qquad (3)$$

with respect to τ for each class of $\mathcal{M}$ makes it possible to choose the 'best' $\boldsymbol{\tau}$ relative to a cost c on false positives. A larger c implies a larger cost on false positives and hence fewer predictions with higher precision (i.e. lower sensitivity and higher specificity). A smaller c implies a smaller cost on false positives and hence more predictions with lower precision (i.e. higher sensitivity, lower specificity). This flexibility is important for biologists who view classifications as hypotheses subject to further experiments.

ALGORITHMS AND IMPLEMENTATION

Our method is fully implemented in the Rosetta system, a publicly available toolkit for data mining and knowledge discovery using rough sets. The system consists of a computational kernel and a front end. The computational kernel is a general C++ open source class library compilable on a large variety of platforms. The front end is developed using the Microsoft Foundation Classes (MFC).

Next, we will outline some algorithmic details about the generation of reducts in Rosetta. Let $\mathcal{A} = (U, A \cup \{d\})$ be a decision system with $n = |U|$ objects. The discernibility matrix D of $\mathcal{A}$ is the $n \times n$ matrix with entries $d_{ij} = \{a \in A \mid x_j \notin IND_{\mathcal{A}}(\{a\}, x_i, d)\}$. That is, d_{ij} is the set of attributes that discern x_i from x_j (Note that in general this is not the same set as the set that discerns x_j from x_i). A *hitting set* of the multiset $\mathcal{S}_D(x_i) = \{d_{ij} \mid d_{ij} \neq \emptyset, j = 1, 2, \ldots, n\}$ is a set $B \subseteq A$ that forms a non-empty intersection with every set in $\mathcal{S}_D(x_i)$. Furthermore, B is a *minimal* hitting set of $\mathcal{S}_D(x_i)$ or a reduct if it ceases to be a hitting set if any of its elements are removed. Searching for minimal hitting sets can easily be implemented using a genetic algorithm with the following fitness function:

$$\begin{aligned} f(B, x_i) &= (1 - \epsilon) \times \frac{A - B}{A} + \\ \epsilon &\times \min\left\{\alpha, \frac{|\{d_{ij} \mid d_{ij} \cap B \neq \emptyset, j = 1, \ldots, n\}|}{n}\right\} \end{aligned} \qquad (4)$$

The function results in approximate reducts (so called α-reducts) having a hitting fraction of at least α. More details about the genetic algorithm implemented in Rosetta can be found in Vinterbo and Øhrn (2000).

Having to construct the discernibility matrix obviously gives the method a time complexity of $O(n^2)$, although n can be reduced from the number of genes to the number of discernible genes. Since the matrix entries are computed on the fly when needed, the memory usage is quite modest.

RESULTS AND DISCUSSION

Cho *et al.* (2001) report the transcriptional profiling of the cell cycle in human fibroblasts using microarray analysis. Duplicate experiments were carried out for 6800 genes every other hour from 0 to 24 hours. The two time series were individually normalized to a unit standard deviation with a mean of zero and then averaged to form one time series (data available on the web: http://www.salk.edu/docs/labs/chipdata/). Clustering employing the Pearson correlation as similarity measure was used to organize the time series into cell cycle-regulated

Table 1. The 27 classes of biological process used as a basis for learning a model

GO term	GO no.	AUC	SE
Apoptosis*	GO:0006915	0.81	0.013
Carbohydrate metabolism	GO:0005975	0.72	0.021
Cell adhesion*	GO:0007155	0.77	0.015
Cell cycle control*	GO:0000074	0.83	0.012
Cell motility*	GO:0006928	0.81	0.011
Cell proliferation	GO:0008283	0.80	0.009
Cell surface receptor linked signal transduction	GO:0007166	0.79	0.008
Cell-cell signaling	GO:0007267	0.80	0.010
DNA metabolism	GO:0006259	0.78	0.015
Energy pathways	GO:0006091	0.76	0.020
Humoral immune response	GO:0006959	0.77	0.017
Immune response	GO:0006955	0.81	0.012
Intracellular signaling cascade	GO:0007242	0.81	0.015
Lipid metabolism	GO:0006629	0.71	0.017
Mesoderm development	GO:0007498	0.77	0.015
Mitotic cell cycle*	GO:0000278	0.84	0.014
Neurogenesis	GO:0007399	0.78	0.014
Oncogenesis	GO:0007048	0.77	0.012
Phototransduction	GO:0007602	0.85	0.011
Physiological processes	GO:0007582	0.77	0.011
Protein biosynthesis	GO:0006412	0.80	0.017
Protein metabolism and modification	GO:0006411	0.77	0.008
Protein amino acid phosphorylation	GO:0006468	0.82	0.014
Proteolysis and peptidolysis	GO:0006508	0.80	0.017
Transcription	GO:0006350	0.71	0.011
Transport	GO:0006810	0.71	0.011
Vision	GO:0007601	0.83	0.013
Average		0.78	0.014
		Precision	Coverage
	$c = 1$	0.31	0.66
	$c = 2$	0.61	0.58
	$c = 3$	0.65	0.56

50-fold cross validation AUC (Area Under (ROC) Curve) estimates are given together with their standard errors. Also given are precision (true positives/(true positives + false positives)) and coverage (true positives/(true positives + false negatives)) for $c = 1$, $c = 2$ and $c = 3$ (Equation 3). Classes marked with * correspond to the classes Cho *et al.* (2001) found to be statistically over-represented in one or more expression clusters.

expression clusters (see Methods in Cho *et al.*, 2001, for details). The binomial distribution was used to show statistically significant over- and under-representation of genes participating in the same biological process in the expression clusters. Of 160 biological processes, seven showed an over-representation in one or more clusters.

Using human gene annotations collected from the euGenes database (Gilbert, 2002, http://eugenes.org:8089) we could label 3620 genes with 7679 specific terms from the 'biological process'-part of gene ontology. By moving the most specific annotations upwards in the ontology we extracted a set of 40 GO terms with at least 100 genes each. Since some of these processes completely covered others, we finally extracted 27 partly overlapping processes as our classes (Table 1)[†]. These included 3043 genes with 5521 annotations. Next, we transformed the numerical expression data using our language of templates over sub-intervals and employed the rough set machinery to obtain a rule model of 11 630 decision rules. Parameter settings are given in Table 2. The model was used to re-classify the known genes and to provide hypotheses of biological process roles for the remaining unknown genes (Figure 1). Table 1 shows the cross validation AUC estimates foreach classes together with *precision* and *coverage* over all classes for different costs on false positives. For $c = 2$ (see Equation 3) we predict 58% of the annotations correctly during cross validation (coverage), while 61% of our prediction were correct (precision). A 50-fold cross validation was performed in which the examples were divided into 50 equally sized subsets and each subset was used as a test set once and as a part of the training set 49 times. The cross validation

[†] Midelfart *et al.* (2001) reports a different approach in which classes are selected iteratively according to their learnability.

Table 2. Parameter values used in the methodology to obtain cross validation estimates and final classifications

Approximate reducts (α in Equation 4)	0.99
Weighting between subset cost and hitting fraction (ϵ in Equation 4)	0.40
Template *increasing*	
Required increase over the full interval	0.60
Required increase during the first and last atomic interval	0.10
Maximum decrease from one time point to the next:	1.00
Template *decreasing*	
Required decrease over the full interval	0.60
Required decrease during the first and last atomic interval	0.10
Maximum increase from one time point to the next:	1.00
Template *constant*	
Maximum difference between largest/smallest value and average value:	0.20

The parameters were choosen experimentally to maximize average AUC during cross validation. To reduce the risk of overfitting, the final estimates in Table 1 were obtained from different computational trials.

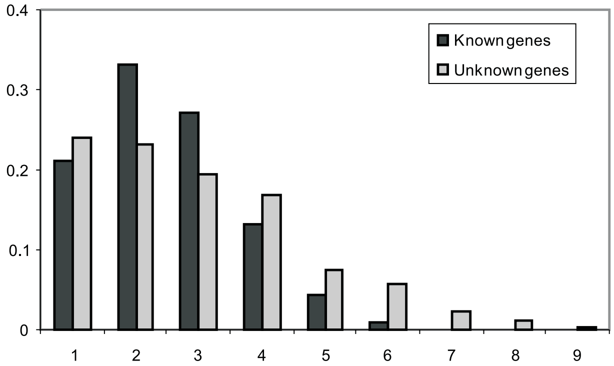


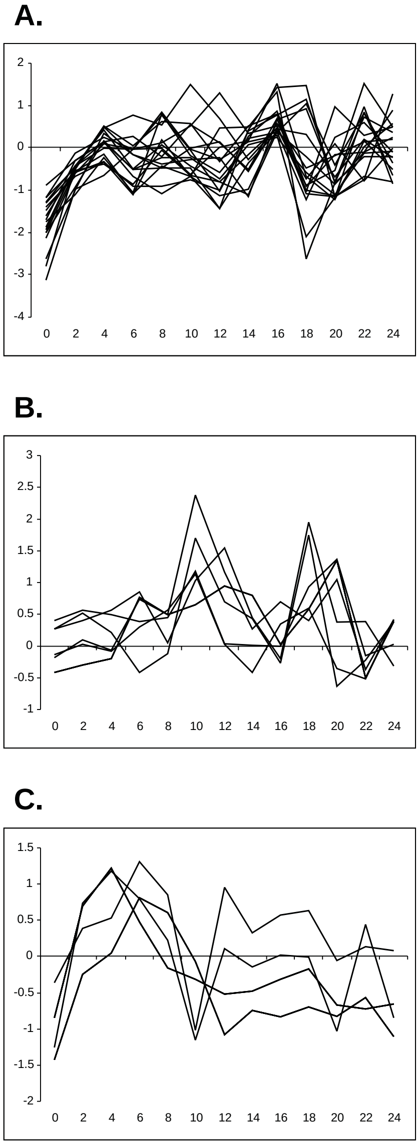
Fig. 1. Distributions (i.e. histograms) for the number of classifications assigned to one gene by the final classifier. As it can be seen, the classifier in general produces several hypotheses of participation in biological process per gene.

estimates are normally interpreted as the classification quality one would get if a model was induced from the full set of examples and used to classify new examples. Hence, precision and coverage in Table 1 are estimates of the quality we may expect when classifying unknown genes.

Genes participating in the same biological process formalized by gene ontology show a great diversity in expression profiles. In addition, biologists have to assign more than one GO term to each gene to explain the biological process roles of its products. It is evident that such a complex relationship cannot be modeled using a few large non-overlapping expression clusters. Instead, a large number of small overlapping clusters would be needed, each modeling subsets of genes associated with one or a few classes. In essence, this is what rules in our method do. Each rule covers a small set of genes annotated to an even smaller number of processes. For this to be possible, a supervised rather than an unsupervised approach is advantageous where annotations form constraints guiding the search. Furthermore, our feature language of templates over subintervals makes it possible to base similarity on discriminatory features of the expression profiles. To illustrate how the rules model the data we have depicted three typical rules in Figure 2 together with the expression profiles of the covered known genes. During classification these rules will contribute to the final classification of unknown genes. Hence, we also included the unknown genes covered by each rule. As we can clearly see, one rule is not enough to pinpoint a class. The rule in Figure 2A indicates 10 different classes. This is partly a result of one gene participating in several biological processes (e.g. gene M27288 is associated with three different processes) and partly the result of genes participating in different biological processes exhibiting similar expression profiles. However, six of the 18 time profiles covered by the rule in Figure 2A are associated with genes annotated with the GO term *transport*.

Altogether, the rules constitute a model of the relationship between temporal transcript profiles and knowledge of biological process. Since each class normally includes genes with several different profiles, a rather large number of relatively specific rules are needed to describe each class. To prove that this model is not a specific definition only applicable to the genes already used to induce it, we tested its classificatory capabilities on unseen examples using cross validation. Clearly the results given in Table 1 are not random (a random classifier would produce AUC values of 0.5). Although each rule is quite specific (i.e. covers few examples), the rules capture important general patterns characterizing different subsets of the classes. The generalization claim is also confirmed applying a similar method to another data set (Hvidsten *et al.*, 2001).

As shown by the values obtained for precision and coverage in Table 1, we are not able to fully discriminate



Rule	Covered genes
0-2(Increasing) AND 2-16(Increasing) =>	
GO(cell proliferation) OR	M27288 U70663
GO(DNA metabolism) OR	U73737 X99586
GO(transcription from Pol II promoter) OR	AB005535
GO(cell-cell signaling) OR	M27288
GO(transport) OR	**U58130 X99586 U81006 M14758 U62434 U55936**
GO(physiological processes) OR	M27288
GO(energy pathways) OR	Y00764
GO(intracellular signaling cascade) OR	D86968
GO(mesoderm development) OR	X04143 U70663
GO(oncogenesis)	U73737
	U79267 D64158 U90910 X61373 U20230 (unknown)

Rule	Covered genes
0-4(Constant) AND 0-10(Increasing) =>	
GO(protein metabolism and modification) OR	**M35296 J02783 D13748 X05130**
GO(mesoderm development) OR	X60957
GO(protein biosynthesis)	D13748
	U90918 (unknown)

Rule	Covered genes
0-4(Increasing) AND 6-10(Decreasing) AND 14-18(Constant) =>	
GO(cell proliferation) OR	**Y07909 X58377 U66468**
GO(cell-cell signaling) OR	X58377
GO(intracellular signaling cascade) OR	X85106
GO(oncogenesis)	Y07909

Fig. 2. Three sample rules together with the expression profiles and the corresponding genes covered by these rules. Also given are unknown genes covered by the same rules. The examples illustrate how the template language enables induction of discriminatory rules. However, single rules alone cannot discriminate the classes. A large number of rules is needed. Also shown is how multiple process-assignments make it even more difficult to discriminate (e.g. M27288 in a). The main biological process is shown in bold for each rule.

the classes. False positives occur because genes from different classes cannot be discerned. False negatives occur because some genes show too little similarity with other genes from the same class. Although we have a less strict similarity definition than most clustering approaches (i.e. genes need to match the same templates over sub-intervals, rather than having to have similar numerical values over the whole time series), some degree of similarity within classes is of course necessary in order to induce rules that can generalize over more than one example. This may also be seen by comparing our cross validation results for classes that were found to be significantly over-represented in some expression clusters by Cho *et al.* (2001). For example, *cell cycle* was found to be over-represented in two expression clusters, and *mitotic cell cycle* and *cell cycle control* (both sub-classes of *cell cycle*) in Table 1 have high AUC values (0.84 and 0.83, respectively). This trend is also true for *DNA replication* (a sub-process of *mitotic cell cycle* with AUC = 0.84), *muscle contraction* (a sub-process of *cell motility* with

AUC = 0.81) and *apoptosis* (with AUC = 0.81). Other processes found to be over-represented in expression clusters were *cytoskeletal reorganization* (not related to any process in Table 1) and *cell-to cell adhesion* (a subprocess of *cell adhesion* with AUC = 0.77).

In summary, we can report a successful approach to modeling participation of gene products in biological processes from gene expression time series using a supervised learning approach. High precision hypotheses can be obtained for both known and unknown genes as demonstrated on the data set previously published by Cho *et al.* (2001). The full rule model and classifications for all known and unknown genes can be found on our web site: http://www.lcb.uu.se/∼hvidsten/bioinf_cho/. We believe that supervised learning algorithms have an important role as hypotheses generators in functional genomics.

ACKNOWLEDGEMENT

We would like to thank Vladimir Yankovski for his work with the euGenes database.

REFERENCES

Bazan,J.G., Skowron,A. and Synak,P. (1994) Dynamic reducts as a tool for extracting laws from decision tables. In *Proceedings of the International Symposium on Methodologies for Intelligent Systems*, Lecture Notes in Artificial Intelligence, 869, Springer, New York, pp. 346–355.

Brown,M.P.S., Grundy,W.N., Cristianini,N., Sugnet,C.W., Furey,T.S., Ares,M. and Haussler,D. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl Acad. Sci. USA*, **97**, 262–267.

Cho,R.J., Huang,M., Campbell,M.J., Dong,H., Steinmetz,L., Sapinoso,L., Hampton,G., Elledge,S.J., Davis,R.W. and Lockhart,D.J. (2001) Transcriptional regulation and function during the human cell cycle. *Nat. Genet.*, **27**, 48–54.

Eisen,M., Spellman,P., Brown,P. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression pattern. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Fayyad,U., Piatetsky-Shapiro,G. and Smyth,P. (1996) The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, **39**, 27–34.

Gilbert,D.G. (2002) euGenes: a eukaryote genome information system. *Nucleic Acids Res.*, **30**, 145–148.

Hanley,J.A. and McNeil,B.J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**, 29–36.

Harrel,Jr,F.E., Califf,R.M., Pryor,D.B., Lee,K.L. and Rosati,R.A. (1982) Evaluating the yield of medical tests. *J. Am. Med. Assoc.*, **247**, 2543–2546.

Hvidsten,T.R., Komorowski,J., Sandvik,A.K. and Lægreid,A. (2001) Predicting gene function from gene expressions and ontologies. In Altman,R.B., Dunker,A.K., Hunter,L., Lauderdale,K. and Klein,T.E. (eds), *Pacific Symposium on Biocomputing*. World Scientific, Mauna Lani, HI, pp. 299–310.

Iyer,V.R., Eisen,M.B., Ross,D.T., Schuler,G., Moore,T., Lee,J.C.F., Trent,J.M., Staudt,L.M., Hudson,J.Jr. and Boguski,M.S. *et al.* (1999) The transcriptional program in the response of human fibroblasts to serum. *Science*, **283**, 83–87.

Komorowski,J., Øhrn,A. and Skowron,A. (2002) The rosetta software system. In Klösgen,W. and Żytkow,J. (eds), *Handbook of Data Mining and Knowledge Discovery*. Oxford University Press, pp. 554–559.

Lockhart,D.J. and Winzeler,E.A. (2000) Genomics, gene expression and DNA arrays. *Nature*, **405**, 827–836.

Midelfart,H., Lægreid,A. and Komorowski,J. (2001) Classification of Gene Expression Data in an Ontology. In Crespo,J., Maojo,V. and Martin,F. (eds), *Second International Symposium on Medical Data Analysis*. Springer, New York, pp. 186–194.

Pawlak,Z. (1982) Rough sets. *International Journal of Information and Computer Science*, **11**, 341–356.

Pawlak,Z. (1991) Rough sets: theoretical aspects of reasoning about data. *Series D: System Theory, Knowledge Engineering and Problem Solving, 9*. Kluwer, Dordrecht.

Pilpel,Y., Sudarsanam,P. and Church,G.M. (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.*, **29**, 153–159.

Schena,M., Shalon,D., Davis,R. and Brown,P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, **270**, 467–470.

Shatkay,H., Edwards,S., Wilbur,W. and Boguski,M. (2000) Genes, themes and microarrays—using information retrieval for large-scale gene analysis. In *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB2000)*. pp. 317–328.

Sherlock,G. (2000) Analysis of large-scale gene expression data. *Curr. Opin. Immunol.*, **12**, 201–205.

Skowron,A. and Nguyen,H.S. (1999) Boolean reasoning scheme with some applications in data mining. In Żytkow,J.M. and Rauch,J. (eds), *Proceedings of the Third European Symposium on Principles and Practice of Knowledge Discovery in Databases (PKDD'99)*, Lecture Notes in Artificial Intelligence, 4, Springer, Prague, pp. 107–115.

Skowron,A. and Rauszer,C. (1992) The discernibility matrices and functions in information systems. In Słowiński,R. (ed.), *Intelligent Decision Support: Handbook of Applications and Advances in Rough Sets Theory*, Series D: System Theory, Knowledge Engineering and Problem Solving, 11, Kluwer, Dordrecht, **Chapter III-2**, pp. 331–362.

The Gene Ontology Consortium (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Vinterbo,S. and Øhrn,A. (2000) Minimal approximate hitting sets and rule templates. *Int. J. Approximate Reasoning*, **25**, 123–143.