GenMiner: mining informative association rules from genomic data

Ricardo Martinez, Claude Pasquier, Nicolas Pasquier

GENMINER : Mining Informative Association Rules from Genomic Data

Ricardo Martinez
I3S Laboratory
UNSA/CNRS UMR-6070
2000 route des Lucioles
06903 Valbonne, France
rmartine@i3s.unice.fr

Claude Pasquier
ISDBC
UNSA/CNRS UMR-6543
Parc Valrose
06108 Nice, France
claude.pasquier@unice.fr

Nicolas Pasquier
I3S Laboratory
UNSA/CNRS UMR-6070
2000 route des Lucioles
06903 Valbonne, France
pasquier@i3s.unice.fr

Abstract

GENMINER *is a smart adaptation of closed itemsets based association rules extraction to genomic data. It takes advantage of the novel* NORDI *discretization method and of the* CLOSE *[27] algorithm to efficiently generate minimal non-redundant association rules.* GENMINER *facilitates the integration of numerous sources of biological information such as gene expressions and annotations, and can tacitly integrate qualitative information on biological conditions (age, sex, etc.). We validated this approach analyzing the microarray datasets used by Eisen* et al. *[10] with several sources of biological annotations. Extracted associations revealed significant co-annotated and co-expressed gene patterns, showing important biological relationships between genes and their features. Several of these relationships are supported by recent biological literature.*

1. Introduction

One of the main goals of gene expression analysis is to discover information about biological processes that govern cell behavior. An important task in this goal is the interpretation of gene expression profiles in the light of biological knowledge represented as gene annotations in biological databases. This task consists in detecting gene groups that are both co-expressed, i.e. sharing similar expression profiles, and co-annotated, i.e. sharing the same annotations such as function, regulatory mechanism, etc.

The volume of biological knowledge is rapidly increasing in gene expression databases (GEO, Arrayexpress, etc.), information on microarray experiments (spotted probes, data processing protocols, etc.), molecular databases (GenBank, Embl, Unigene, etc.), semantic sources as thesaurus, ontologies or semantic networks (UMLS, GO, etc.), bibliographical databases (Medline, Biosis, etc.) and gene/protein related specific sources (KEGG, OMIM, etc.). This important increase in data volume leads to several problems: How to integrate these data with gene expression data? How to efficiently analyze such amounts of data? How to detect the most relevant information patterns among the results?

Existing approaches dealing with the interpretation problem can be classified in three axes [21]: *Expression-based approaches*, such as Thea [26] or Generator [29], *knowledge-based approaches*, such as Page [16] or CGGA [22], and *co-clustering approaches*, such as Co-Cluster [14] or BiCluster [20]. The *expression-based* axis, that gives more weight to gene expression profiles than the other two interpretation axis, is the most currently used. However, approaches in this axis present many well-known drawbacks. First, genes are clustered if they have similar expression profiles across all biological conditions, but gene involved in the same biological process might be co-expressed in only a subset of conditions [2]. Second, genes may be conditionally co-expressed with different sets of genes, reflecting the different biological roles that genes can play in the cell. Most of the commonly used clustering methods group genes into single clusters only, masking more complex relationships between different sets of conditionally regulated genes [11]. Third, even when similar expression profiles are related to similar biological roles, discovering these biological connections among co-expressed genes is not a trivial task and requires a lot of additional work [30].

To overcome these drawbacks, we propose the use of association rule discovery (ARD). ARD is an unsupervised data mining technique used to discover links among sets of *items* (variable values) such as gene expression profiles or gene annotations from very large data relations. Association rules identify groups of items that frequently co-occur in data lines, establishing relationships between them with the form: $A \Rightarrow B$ which means that when A occurs it is likely that B occurs. ARD has the following advantages:

1. ARD can generate rules containing genes that are co-expressed only in a subset of the biological conditions.

2. Any gene can be assigned to any number of association rules as long as its expression profile fulfills the assignation criteria. This means that a gene involved in many co-expressed groups will appear in each and every one of those groups, without limitation.
3. ARD generates orientated knowledge patterns ***if*** *condition* ***then*** *consequent*, describing directed relationships. Thus, any type of relationship between expression measures and gene annotations can be discovered.
4. ARD facilitates the integration of various heterogeneous biological sources of information.

In the past years, ARD has been used for analysing gene expression data in order to discover frequent gene patterns among a subset of biological conditions: [6, 31, 12]. Association rules generated by these approaches are of the following form: *gene g1*$\downarrow \Rightarrow$ *gene g2*$\uparrow$, *gene g3*$\downarrow$, meaning that in a significant number of biological conditions, when gene $g1$ is under-expresssed, it is likely to observe an over-expression of gene $g2$ and an under-expression of gene $g3$. This technique has been successfully applied for clustering gene expression profiles, avoiding some drawbacks of standard clustering techniques [12]. However, these algorithms use exclusively gene expression measures without taking into account biological knowledge. The task of discovering and interpreting biological similarities hidden within gene groups is thus left to the expert.

Recently, Carmona et al. [5] proposed to integrate gene expression profiles and gene annotations to extract rule with the form : *annotation* $\Rightarrow C1\,[\downarrow],\ C2\,[\uparrow]$ meaning that a group of genes annotated by *annotation* is likely to be under-expressed in biological condition $C1$ and over-expressed in condition $C2$. However, this approach presents several weaknesses. First, it uses the Apriori ARD algorithm [1] that is time and memory-consuming in the case of correlated data. Moreover, it generates a huge number of rules among which many are redundant thus complexifying results interpretation. This is a well-known major limitation of the Apriori algorithm for correlated data [5, 31]. Second, extracted rules are restricted to a single form: Annotations in the left-hand-side and expression profiles in the right-hand-side. However, all rules containing annotations and/or expression profiles, regardless of the side, bring important information for the biologist. Third, it uses the two-fold change cut-off method for discretizing expression measures in three intervals, a dangerous simplification that presents several drawbacks [25].

The GENMINER approach was developed to address these weaknesses and fully exploit ARD capabilities. It enables the integration of gene annotations and gene expression data to discover intrinsic associations between them. Gene annotations can be integrated from any source of biological information, such as semantic sources, bibiographic databases or gene expression databases for instance. It uses a novel method, called NORDI , for discretizing gene expression measures and generate gene expression profiles.

GENMINER takes advantage of the CLOSE [27] ARD algorithm to efficiently generate low support and high confidence non-redundant association rules. When data is dense or correlated, such as genomic data, CLOSE reduces both execution times and memory space usage compared with Apriori, thus enabling the analysis of huge datasets. Furthermore, it improves the result's relevance by extracting a minimal set of rules containing only non-redundant rules, hence reducing the number of rules and facilitating their interpretation by the biologists. These features make GENMINER an ARD approach adequate to biologists requirements for genomic data analysis.

ARD basics, the NORDI method and the GENMINER approach are presented in section 2. The extended Eisen dataset used to validate the approach and experimental results are presented in section 3 and 4 respectively. A brief discussion in section 5 concludes the paper.

2. Association rules extraction

Association rules are knowledge patterns expressing correlations between occurrences of attribute values as directed relationships between *itemsets* (sets of items). For each rule, the *support* and *confidence* statistics measure the scope and the precision of the rule respectively. For instance, an association rule *Event(A), Event(B)* $\Rightarrow$ *Event(C), support=20%, confidence=70%* states that when events *A* and *B* occur, event *C* also occurs in 70% of cases, and that all three events occur together in 20% of all situations. In this context, *Event(A), Event(B)* and *Event(C)* are items and situations are data objects, i.e. the lines of the dataset, describing co-occurring events. To extract only statistically significant associations, extraction is restricted to rules with support and confidence exceeding some user defined minimum support *minsupp* and minimum confidence *minconf* thresholds.

Association rules are extracted from a dataset that is a triplet $\mathcal{D} = \{\mathcal{O}, \mathcal{I}, \mathcal{R}\}$, where $\mathcal{O}$ and $\mathcal{I}$ are finite sets of objects (lines) and items (columns) respectively, and $\mathcal{R} \subseteq \mathcal{O} \times \mathcal{I}$ is a binary relation. Each item represents an attribute value or a set of attribute values and each couple $(o, i) \in \mathcal{R}$ denotes the fact that the object $o \in O$ is related to the item $i \in I$. When the attribute is numeric and continuous, each item represents an interval of values. If an object *o* is in relation with all items of an itemset *I* we say that *o contains I*. The *support* of an itemset *I* is the proportion of objects containing *I* and an itemset is *frequent* if its support is greater or equal to *minsupp*.

The natural decomposition of the ARD problem is: (1) Extract frequent itemsets and their supports from the dataset; (2) Generate all valid association rules from fre-

quent itemsets and their supports. The first phase is the most computationally expensive part of the process, since the number of potential frequent itemsets is exponential ($2^{|I|}$) in the size of the set of items and several dataset scans, that are time-consuming, are required.

Levelwise algorithms for extracting frequent itemsets are iterative algorithms that consider all itemsets of a given size at a time. They are based on the following properties: (i) all supersets of an infrequent itemset are infrequent; (ii) all subsets of a frequent itemset are frequent. These properties enable the use of previous iteration results to reduce the search space of the next iteration, and the total number of iterations, that is the number of dataset scans, is equal to the size of the largest frequent itemsets. This approach was proposed in the well-known Apriori [1] algorithm. Several optimisations have been proposed to improve the extraction efficiency, by avoiding several dataset scans, but they all give response times of the same order of magnitude, depending mainly on data correlation.

These algorithms are efficient when data is weakly correlated and sparse, such as sales data, but performance drastically decrease when data is correlated or dense, such as census data [4]. Moreover, with such data, a huge number of association rules are extracted, even for high *minsupp* and *minconf* values, and a majority of these rules are redundant (bring the same information). For instance, consider the nine rules presented below that all have the same support and confidence and the item *annotation1* in the antecedent:
1. *annotation1* $\Rightarrow$ *gene g1*$\uparrow$
2. *annotation1* $\Rightarrow$ *gene g1*$\uparrow$, *gene g2*$\uparrow$
3. *annotation1* $\Rightarrow$ *gene g1*$\uparrow$, *gene g3*$\uparrow$
4. *annotation1* $\Rightarrow$ *gene g1*$\uparrow$, *gene g2*$\uparrow$, *gene g3*$\uparrow$
5. *annotation1, gene g2*$\uparrow$ $\Rightarrow$ *gene g1*$\uparrow$
6. *annotation1, gene g2*$\uparrow$ $\Rightarrow$ *gene g1*$\uparrow$, *gene g3*$\uparrow$
7. *annotation1, gene g3*$\uparrow$ $\Rightarrow$ *gene g1*$\uparrow$
8. *annotation1, gene g3*$\uparrow$ $\Rightarrow$ *gene g1*$\uparrow$, *gene g2*$\uparrow$
9. *annotation1, gene g2*$\uparrow$, *gene g3*$\uparrow$ $\Rightarrow$ *gene g1*$\uparrow$

The most relevant rule from the user's viewpoint is rule 4 since all other rules can be deduced from this one, including support and confidence (but the reverse does not hold). Information brougth by all other rules are summed up in rule 4, that is a *non-redundant association rule with minimal antecedent and maximal consequent*, or *minimal non-redundant rule* for short.

CLOSE algorithm The *frequent closed itemsets based approach* [27] is based on the closure operator of the Galois connection. This operator γ associates with an itemset X the maximal set of items common to all the objects containing X, i.e. the intersection of these objects. *Frequent closed itemsets* are frequent itemsets with $\gamma(X) = X$. An itemset X is a frequent closed itemset if no other item $i \in X$ is common to all objects containing X. *Generators* of a frequent closed itemsets X are minimal (by inclusion) itemsets which closure is X. The frequent closed itemsets constitute a *generating set* for all frequent itemsets and thus for all association rules [27]. This relies on the following properties: (i) The support of a frequent itemset is equal to the support of its closure; (ii) The maximal frequent itemsets are maximal frequent closed itemsets. Using these properties, a new approach for mining association rules was proposed: (1) Extract frequent closed itemsets and their supports; (2) Derive frequent itemsets and their supports; (3) Generate all valid association rules. The search space of the first phase is then reduced to the closed itemsets. The first algorithm based on this approach is CLOSE [27]. Several algorithms for extracting frequent closed itemsets, using complex data structures to improve efficiency, have been proposed. However, they do not extract generators and their response times, depending mainly of data density and correlation, are of the same order of magnitude.

An association rule is *redundant* if it brings the same or less general information than is brought by another rule with identical support and confidence [7, 28]. Then, an association rule R is a minimal non-redundant association rules if there is no association rule R' with same support and confidence, which antecedent is a subset of the antecedent of R and which consequent is a superset of the consequent of R. Using generators and frequent closed itemsets, CLOSE can generate a basis (a minimal set) for association rules containing only non-redundant minimal rules. This basis contains: (1) Exact association rules $G \Rightarrow \gamma(G) \setminus G$ between a generator G and its closure $\gamma(G)$ such that $\gamma(G) \neq G$; (2) Approximate association rules $G \Rightarrow \gamma(H) \setminus G$ between a generator G and a closure $\gamma(H)$ that is a superset of the closure $\gamma(G)$. This basis called *Informative* or *Min-max basis* is a generating set for all association rules [28]. It captures all the information brought by the set of all rules in a minimal number of rules, without information loss [7]. Experiments conducted on benchmark datasets show that the reduction factor varies from 5 to 400 according to data density and correlation [28].

GENMINER approach GENMINER is a co-clustering and bi-clustering approach that integrates gene annotations and gene expressions to discover intrinsic associations among both data sources based on co-ocurrence patterns. It is a co-clustering approach that integrates co-expressed and co-annotated gene groups at the same time. Furthermore, it is a bi-clustering approach that finds co-annotated and co-expressed gene groups even in a small subset of biological conditions.

GENMINER follows the four steps of the ARD process: data selection and pretreatment, frequent itemsets extraction, association rules generation and interpretation of extracted rules. It uses the NORDI algorithm for gene expres-

sion data discretization and the CLOSE algorithm for minimal non-redundant rules extraction.

NORDI algorithm The *Normal Discretization* (NORDI) algorithm was developed to improve gene expression measures discretization into items. This phase is essential to extract relevant association rules. This algorithm is based on statistical detection of outliers and the continuous application of normality tests for transforming the initial sample distribution "almost normal" to a "more normal" one. The term "almost" means that the sample distribution can be normally distributed without the outlier's presence.

Let us assume that the expression data measures are presented as an nXm matrix: $\boldsymbol{E}$ with n genes (rows) and m samples or biological conditions (columns). Each matrix entry, $e_{i,j}$ represents the gene expression measure of gene i in sample j where $e_{i,j}$ is continuous in all real numbers. Let's suppose that the gene expression matrix $\boldsymbol{E}$ accomplishes the following assumptions:

1. All data is well cleaned (minimal noise).
2. Number of genes is largely enough.
3. The samples of the matrix S_j for every $j = 1, 2, ..., m$ are independent from each other and they are "almost" normally distributed $S_j \sim N(\mu_j, \sigma_j)$.
4. Missing values are no significant in relation to the number of genes.

The NORDI algorithm states that every sample of the expression matrix S_j can be "more" normally distributed $S_j^k \sim N(\mu_j, \sigma_j)$ if all outliers of each sample are momentarily removed (that is keeping a list of the k removed outliers for each sample, i.e. L_j^k) by Grubbs outliers method [13]. Each time an outlier k is removed, a Jaque-Bera normality test [3] has to be accomplished for the remaining sample S_j^k, where k is the number of removed outliers at each step in sample S_j and $k = 0, 1, 2, \ldots, clean$ ($k = clean$ means that there are no more outliers in the sample according to the Grubbs criterium). So, for every sample, we obtain the remaining sample S_j^{clean} that is "more normally" distributed than the original sample S_j. To verify this assertion we compare S_j^{clean} against S_j using the QQ-plot [24] and Lilliefors [18] normality tests. Then, we calculate the over-expressed, Ot, and under-expressed, Ut, cutoff thresholds using the $z - score$ methodology [32] over the cleaned sample S_j^{clean}.

Supposing the four precedent assumptions with $S_j^{clean} \sim N(\mu_j, \sigma_j)$ normal distributed and a certain degree of predetermined confidence $1 - \alpha$, the $z - score$ threshold cutoffs for three intervals are defined as:

- $Z_j = \frac{e_{i,j} - \mu_j}{\sigma_j} \geq z_{\alpha/2} = Ot \Rightarrow e_{i,j}$: over-expressed ($\uparrow$)
- $Z_j = \frac{e_{i,j} - \mu_j}{\sigma_j} \leq z_{\alpha/2} = Ut \Rightarrow e_{i,j}$: under-expressed ($\downarrow$)
- $Ut < e_{i,j} > Ot \Rightarrow e_{i,j}$: unexpressed

where $z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$, if the cumulative distribution function is $\Phi(z_{\alpha/2}) = P(S_j^{clean} \leq z_{\alpha/2}) = 1 - \alpha/2$.

It is important to notice that this procedure for computing the threshold cutoffs is done over all the m cleaned samples S_j^{clean} contained in the expression matrix $\boldsymbol{E}$. Once the computation of threshold cutoffs is done, the k elements in each sample's outliers list L_j^k are integrated to the original sample S_j and the discretization procedure is calculated for all values in S_j. The main reason is that outliers values cannot be removed from the analysis because they may contain relevant information of the biological experiment.

3. Presentation of the dataset

To validate the GENMINER approach we applied it to the well-known Eisen et al. genomic dataset [10]. This dataset contains expression measures of 2465 yeast genes under 79 biological conditions extracted from a collection of four independent microarray studies about the *Saccharomyces cerevisiae* during several biological processes: cell cycle experiments, sporulation experiments, temperature shock experiments, and diauxic shift.

Gene expression measures The Eisen dataset was pretreated by taking the log_2 ratios (to consider cellular inductions and repressions in a numerically equal way) and applying the imputation algorithm of k-nearest neighbors [19] in order to treat the missing values (1.9% of the total). This dataset was discretized using the NORDI algorithm at a 95% confidence level.

Gene annotations *S. cerevisiae* genes were annotated using five sources of biological information:

- Gene Ontology (GO) annotations, describing molecular fonctions, biological process and locations of gene products,
- bibliographic annotations, representing associations between research papers and genes (data manually curated from the literature by *Saccharomyces cerevisiae database* (SGD) staff),
- pathway annotations from Kyoto Encyclopedia of Genes and Genomes (KEGG), identifying the metabolic pathways in which each gene is involved,
- phenotypic annotations, describing visible traits or characteristics of genes (extracted from SGD's file),
- transcriptional regulators (TR) annotations, identifying protein that bind to promoter regions in order to either increase or decrease the transcription of genes (the data result from a study of Lee et al. [17], by using a *p-value* threshold of 0.0005).

All gene annotations were taken as boolean variables, i.e. $i \in \{0, 1\}$, indicating if an annotation pertains, $i = 1$, or not, $i = 0$, to a given gene. The prefixes *go:*, *path:*, *pmid:*, *pr:* are used to identify Gene Ontology terms, KEGG pathways, Pubmed identifiers and promoters respectively.

Rule	Antecedent	Consequent	Supp. (#)	Conf. (%)
1	go:0006412, go:0005840	heat3↓	103	51
2	go:0005840, go:0005198	heat3↓	96	56
3	go:0042254, go:0005840, go:0005198	heat4↓	15	52
4	go:0006412, go:0006996, go:0005198	heat3↓	30	64
5	path:sce03010	heat4↓	69	53
6	pr:RAP1, pr:FHL1	heat3↓	71	62
7	pmid:5542014, pmid:9649613, pmid:3533916	heat3↓	12	100
8	path:sce00190	dx6↑, dx7↑	14	26
9	path:sce00020	dx6↑, dx7↑	8	32
10	path:sce00630	dx7↑	6	55
11	pr:FHL1, pr:GAT3, pr:RAP1, path:sce03010	dx7↓	17	50

Table 1. Associations *Annotations* ⇒ *Expressions*

Dataset The resulting dataset is a matrix of 2465 lines, each one corresponding to a yeast gene, and 177 columns, each one corresponding to an expression level or an annotation. Each line contains expression profiles over the 79 biological conditions (values discretized by NORDI) and at most 98 gene annotations (24 GO annotations, 15 KEGG annotations, 25 transcriptional regulators, 14 phenotypes and 20 pubmed keywords).

4 Experimental results

To explore the full potential of the GENMINER approach, we applied it to the extended Eisen dataset integrating gene expression profiles and collected sources of biological information. Furthermore, we considered all possible types of rules, having either gene annotations or gene expression measures either or both in the antecedent and the consequent. We have selected and described meaningful biological rules, emphasizing the form of the rule in order to show the potentials of the GENMINER approach.

4.1 Associations *Annotations* ⇒ *Expressions*

Association rules with the form *gene annotations* ⇒ *gene expression profiles* mean that a group of gene associated with a specific set of annotations is likely to be over-expressed or under-expressed in a set of biological conditions. This type of association rules corresponds to the type of rules searched by Carmona *et al.* [5]. Selected association rules extracted with GENMINER are presented in Table 1. Supports are given in number of transactions and confidences are percentages.

Rules on the heat shock experiment (labels heat1 to heat6) have as antecedent various terms related to protein synthesis (go:0006412:*translation*, go:0042254:*ribosome biogenesis and assembly*, go:0005840:*ribosome* and path:sce03010:*ribosome pathway*) or cellular organization (go:0005198:*structural molecule activity*, and go:0006996:*organelle organization and biogenesis*) and as consequent, an under-expression at time points 3 and 4. This highlight a general reduction of protein synthesis and cell maintenance following a heat shock, leading to cellular damages. This is confirmed by rule 6 which shows that genes regulated by RAP1 and FHL1 promoters are under-expressed at time point 3. This reflects the known fact that RAP1 recruits FHL1 to activate transcription [33]. Rule 7 in Table 1 shows that all the genes cited in three different articles (which are all about the study of ribosome in yeast) are under-expressed at time point 3.

Examining results relative to the the yeast diauxic shift process only (labels dx1 to dx7), we have found almost all the rules presented by Carmona et al. [5]. The differences concern only the support and confidence measures, because Eisen data contains only a selection of 2465 genes of the 6199 genes used in DeRisi data. However, the same biological interpretation of the results can be drawn.

Rules 8 and 9 in Table 1 revealed marked alterations at biological conditions *Oxidative phosphorylation* (path:sce00190) and *Citrate cycle* (path:sce00020), which is in agreement with the curve of glucose concentration reported in the original paper [9].

Additionally, rule 10 shows that the genes involved in *glyoxylate and dicarboxylate metabolism* (path:sce00630) were also mainly over-expressed at the last time point which reflects the main metabolic changes associated to the diauxic shift in yeast, manually identified by DeRisi [9].

Rule 11 shows that ribosomal genes (annotated with path:sce03010) whose promoter regions were bound by GAT3, FHL1 and RAP1 presented an inhibition pattern in response to nutrient starvation. These associations were extracted with relatively high support values and suggest a connection among GAT3, FHL1 and RAP1 and the decrease in ribosomal gene transcription in response to glucose depletion. The connection among RAP1 and ribosomal gene transcription is well-known [23].

4.2 Associations *Expressions* ⇒ *Annotations*

Rules with the form *gene expression profiles* ⇒ *gene annotations* mean that when a group of genes is over-expressed or under-expressed in a set of biological conditions, these genes are likely to have the corresponding gene annotations. Selected association rules extracted with GENMINER are presented in Table 2. The antecedent of the rule contains the over-expression or under-expression in a set of biological conditions and the consequent is composed by their corresponding gene annotations.

Concerning the elutriation process (labels elu1 to elu14) that is part of the cell cycle experiment, we have found (rules 1-3 from Table 2) an over-expression of the responsi-

Rule	Antecedent	Consequent	Supp (#)	Conf (%)
1	elu5↑ elu6↑ elu7↑	go:0006412	26	87
2	elu4↑ elu5↑ elu6↑	go:0006412	18	86
3	elu2↓	go:0006996	12	55
4	spo4↓ spo5↓ spo6↓	go:0005975	12	52
5	spo2↓ spo3↓	go:0006412	27	57
6	spo3↓ spo4↓ spo5↓	path:sce00010	13	52
7	heat3↓ heat4↓ heat5↓	go:0006412	35	88
8	heat2↓	go:0006996	41	69
9	heat2↓	go:0042254	39	66
10	heat2↑ heat3↑ heat5↑	go:0006950	15	52
11	dx5↑ dx7↑	go:0006091	24	52
12	dx6↓ dx7↓	go:0006412	21	66

Table 2. Associations *Expressions ⇒ Annotations*

Rule	Antecedent	Consequent	Supp. (#)	Conf. (%)
1	path:sce00190	go:0005737	52	96
2	pr:FHL1	pr:RAP1	114	86
3	pr:RAP1, pr:FHL1	go:0005737, go:0006412, go:0005840	93	82
4	pmid:16155567	phenot:inviable	96	94
5	go:0005739	go:0005737	503	100
6	go:0005740	go:0005737, go:0005739	167	100

Table 3. Associations *Annotations ⇒ Annotations*

ble genes of the protein synthesis (go:0006412:*translation*) and an under-expression of the genes responsible of the cellular organization (go:0006996:*organelle organization and biogenesis*).

In the sporulation experiments (rules 4-6 from Table 2), we note an under-expression of the genes intervening in the sugar formation (go:0005975:*carbohydrate metabolic process*) and the protein synthesis (go:0006412:*translation*). This claim is confirmed by the under-expression of the genes belonging to the process of sugar transformation into energy (path:sce00010:*Glycolysis / Gluconeogenesis*).

In the Heat Shock process (rules 7-10 from Table 2), we note an under-expression of the genes responsible for the protein synthesis (go:0006412:*translation*), the cellular organization (go:0006996:*organelle organization and biogenesis*), the ribosomal organization (go:0042254:*ribosome biogenesis and assembly*) and an over-expression of the genes related to stress response (go:0006950:*response to stress*).

Concerning the diauxic shift process (rules 11-12 from Table 2), there is an over-expression of the genes responsible for the energy generation (go:0006091:*generation of precursor metabolites and energy*) and an under-expression of the genes responsible for the protein synthesis (go:0006412:*translation*).

4.3 Associations *Annotations ⇒ Annotations*

Independently from the gene expression levels, it is also possible to highlight existent relationships among gene annotations. Selected association rules extracted with GENMINER are presented in Table 3. Both antecedents and consequents of these rules contain gene annotations.

We identify associations between annotations from different sources like the relationship between the KEGG term sce00190 (*purine metabolism*) and the GO term go:0005737 (*cytoplasm*) (rule 1 of Table 3).

Concerning the transcriptional regulators, extracted rules enable to state strong relationship between promoters $FHL1$ and $RAP1$. For example the rule $FHL1 \Rightarrow RAP1$ with a high support of 114 and a confidence of 86% (rule 2 of Table 3) indicates that the genes activated by $FHL1$ are also activated by $RAP1$ This information is already known and was described in many articles. For example Zhao et al. [33] state that "$RAP1$ binding is essential for the recruitment of $FHL1$", and they explain the association between them in the following phrase: "based on recent work, a simple model for the transcription of RP (ribosomal proteins) genes is that $RAP1$ recruits $FHL1$, which in turn recruits the transcriptional activator $IFH1$". The last phrase confirms the results obtained in rule 3 of Table 3 where the promoters $RAP1$ and $FHL1$ are closely related to the Gene Ontology terms go:0005737 (*cytoplasm*), go:0006412 (*translation*) and go:0005840 (*ribosome*). The two last terms are closely related to protein synthesis and the cytoplasm activity shows us the transcriptional cellular activity while $RAP1$ and $FHL1$ transcription factors are activated.

We also detected rules which relate scientific articles with phenotypes as the rule 4 of Table 3 where *pmid:16155567 ⇒ phenot:inviable* with a support of 96 genes and a confidence of 94%. The PubMed article 1615567 'The synthetic genetic interaction spectrum of essential genes' [8] presents a review of the essential yeast genes. These genes are for the majority annotated as inviable, i.e. the organism does not survive when the corresponding gene is removed.

When analyzed data represent a hierarchy, it is possible, by examining the obtained rules, to reconstitute the original hierarchy. For example, rule 4 of Table 3, i.e. *go:0005739 ⇒ go:0005737* with a support of 503 and a confidence of 100% means that there are 503 genes annotated by go:0005739 and also by go:0005737. go:0005739 (*mitochondrion*) is either a sub-term of go:0005737 (*cytoplasm*) or it represents exactly the same concept. In Eisen data set, we have more than 1500 genes annotated with go:0005737. Therefore, go:0005739 is a sub-term or child of the parental term go:0005737.

The rule 6 of Table 3, i.e. *go:0005740 ⇒ go:0005737, go:0005739* with a support of 167 and a confidence of 100%, means that the terms annotated go:0005740 are also annotated by go:0005737 and go:0005739. Thus, we con-

tinue the unfolding of the hierarchy go:0005740 (*mitochondrial envelope*) is a sub-term of go:0005739 (*mitochondrion*) containing 167 genes.

5 Discussion and conclusion

We presented the GENMINER ARD approach fulfilling the requirements of data obtained from gene expression technologies. This approach integrates gene expression profiles with gene annotations to discover intrinsic associations among both data sources; it is thus a co-clustering technique. It is also a bi-clustering technique that can find patterns of genes that are co-expressed in subsets of biological conditions. In opposition to most gene expression interpretation approaches, as well *expression-based* as *knowledge-based*, in which biological information and gene expression profiles are incorporated in an independent manner, our approach integrates both data sources in a single framework.

GENMINER takes advantage of the CLOSE algorithm [27] that was specifically designed for extracting association rules from highly correlated data. With such data, ARD execution time and memory space usage are high [4], limiting capabilities of classical algorithms, such as Apriori [1], to extract only associations concerning important groups of genes. CLOSE addresses this problem by limiting the search space and the number of dataset scans to reduce execution times and memory space usage. Moreover, the number of association rules extracted from correlated data is most often very important and many of these rules bring the same information, and are thus redundant [7, 28]. This is an important drawback for rules interpretation by the analysts. To address this problem, CLOSE extracts a basis for association rules that is a minimal set of non-redundant rules; All information is summarized in a minimal number of rules, each rule bringing as much information as possible, to improve the results relevance.

Gene expression technologies data, where several gene groups are expressed together in different biological conditions, are highly correlated data. Using the CLOSE algorithm, GENMINER can deal with very large datasets of genomic data and experiments show that its execution times and memory usage are significantly smaller than those of the Carmona et al. [5] Apriori-based approach. Furthermore, it enables the use of several heterogenous sources of annotations, including thousands of annotations related to studied genes.

GENMINER also implements a new discretization algorithm, called NORDI , specially designed for discretizing data issued from gene expression technologies in the case of independent biological conditions. Experiments conducted on the Eisen dataset show that NORDI algorithm results are relevant. However, the discretization issue is a delicate step when using data mining methods as ARD and we propose the use of several discretization scenarios, analyzing the pertinence of obtained results against expected results, to validate the discretization method. In a recent work, Pan et al. [25] suggested that "the robustness of biological conclusions made by using microarray analysis should be routinely assessed by examining the validity of the conclusions by using a range of threshold parameters issued from different discretization algorithms". Unfortunately, to our knowledge no discretization algorithm, specially designed for time process data, can integrate the time variable without an important loss of temporal information.

Another delicate issue in association rules discovery is the thresholds for selecting significative rules. Support and confidence are computed while rules are extracted from the dataset, and are, in many cases, the only ones used to point out its relevance. For genomic data, the minimum support threshold must be set low since if only a small set of genes are annotated into a very specific category, the support of rules containing this annotation will be quite low. Nevertheless, if these rules have a high confidence value, they reveal that this specific biological property is highly associated with an expression pattern of another gene annotation that appears in the consequent. However, an association rule with high support and confidence can be useless, if the consequent itemset of the rule is highly frequent in the dataset and is thus associated to many other itemsets. In other words, associations among weakly correlated elements can be generated using the support-confidence framework [15]. GENMINER is based on the support-confidence framework, but other statistical measures to evaluate correlation (or independence) between consequents and antecedents of rules can easily be integrated during the calculation phasis or the interpretation phasis, to filter rules between weakly correlated gene patterns and order other rules.

The analysis of the well-known gene expression datasets from Eisen [10] has demonstrated the capacity of GENMINER to extract meaningful associations among gene expression profiles and gene annotations. Furthermore, we have shown the potential of this approach to integrate several heterogeneous sources of information such as GO, KEGG, phenotype information, transcriptional regulators information and information of selected articles with gene expression profiles. This is only an example of GENMINER possibilities, that can easily integrate any kind of gene annotations obtained from any source of biological information. Therefore, the integration of different types of biological information is an essential consideration to fully understand the underlying biological processes. In addition, qualitative variables (gender, tissue, age, etc.) could easily be added to the analysis in order to extract association rules among these features and gene expression patterns.

Another important feature of GENMINER is its capacity to extract association rules containing itemsets composed

of both gene annotations and gene expression patterns in the antecedent and/or the consequent. Analysing association rules generated by GENMINER , we have found important relationships supported by recent biological literature. These results show that GENMINER is a promising tool for finding meaningful relationships between gene expression patterns and gene annotations.

References

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proc. VLDB conf.*, pages 478–499, 1994.

[2] R. Altman and S. Raychaudhuri. Whole-genome expression analysis: challenges beyond clustering. *Current Opinion Structural Biology*, 11:340–347, 2001.

[3] A. Bera and C. Jarque. Efficient tests for normality, homoscedasticity and serial independence of regression residuals: Monte carlo evidence. *Economics Letters*, 7:313–318, 1981.

[4] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. In *Proc. ACM SIGMOD conf.*, pages 255–264, 1997.

[5] P. Carmona-Saez, M. Chagoyen, A. Rodriguez, O. Trelles, J. Carazo, and A. Pascual-Montano. Integrated analyis of gene expression by association rules discovery. *BMC Bioinformatics*, 7(54), 2006.

[6] C. Creighton and S. Hanansh. Mining gene expression databases for association rules. *Bioinformatics*, 19:79–86, 2003.

[7] L. Cristofor and D. A. Simovici. Generating an informative cover for association rules. In *Proc. ICDM conf.*, pages 597–600, 2002.

[8] A. davierwala, J. Haynes, B. Li, R. Brost, M. Robinson, and L. e. a. Yu. The synthetic genetic interation spectrum of essential genes. *Nature Genetics*, 37(10):1147–52, 2005.

[9] J. DeRisi, L. Iyer, and V. Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278:680–686, 1997.

[10] M. Eisen, P. Spellman, P. Brown, and D. Botsein. Cluster analysis and display of genome wide expression patterns. In *Proc. National Academy of Sciences USA*, volume 95, pages 14863–8, 1998.

[11] A. Gasch and M. Eisen. Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biology*, 3:1–22, 2002.

[12] E. Georgi, L. Richter, U. Ruckert, and S. Kramer. Analyzing microarray data using quantitative association rules. *Bioinformatics*, 21:123–129, 2005.

[13] F. Grubbs. Procedures for detecting outlying observations in samples. *Technometrics*, 11:1–21, 1969.

[14] D. Hanisch, A. Zien, R. Zimmer, and T. Lengauer. Co-clustering of biological networks and gene expression data. *Bioinformatics*, 18:S145–S154, 2002.

[15] L. Ji and K. Tan. Mining gene expression data for positive and negative co-regulated gene clusters. *Bioinformatics*, 20:2711–2718, 2004.

[16] S. Kim and D. Volsky. Page: Parametric analysis of gene set enrichment. *BMC Bioinformatics*, 6:144, 2005.

[17] T. Lee, N. Rinaldi, F. Robert, D. Odom, Z. Bar-Joseph, G. Gerber, N. Hannett, C. Harbison, C. Thompson, I. Simon, J. Zeitlinger, G. Jennings, H. Murray, B. Gordon, and R. Young. Transcriptional regulatory networks in saccharomyces cerevisiae. *Science*, 298(5594):799–804, 2002.

[18] H. Lilliefors. On the kolmogorov-smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62, 1967.

[19] R. Little and D. Rubin. *Statistical Analysis with Missing Data*. Wiley and Sons, 2 edition, 2002.

[20] J. Liu, J. Yang, and W. Wang. Gene ontology friendly biclustering of expression profiles. In *Proc. CSB conf.*, pages 436–447, 2004.

[21] R. Martinez and M. Collard. Extracted knowledge: Interpretation in mining biological data, a survey. *International Journal of Comp. Science and Applications: Special issue in Research Challenges in Information Science*, 1:1–21, 2007.

[22] R. Martinez, N. Pasquier, C. Pasquier, M. Collard, and L. Lopez-Perez. Co-expressed gene groups analysis (cgga): An automatic tool for the interpretation of microarray experiments. *J. of Integrative Bioinformatics*, 3(11):1–12, 2006.

[23] R. H. Morse. Rap, rap, open up! new wrinkles for rapi in yeast. *Trends Genet*, 16:51–53, 2000.

[24] NIST. *e-Handbook of Statistical Methods*. SEMATECH, 2007. http://www.itl.nist.gov/div898/handbook/.

[25] K. Pan, C. Lih, and N. Cohen. Effects of threshold choice on biological conclusions reched during analysis of gene expression by dna microarrays. *National Academy of Sciences PNAS*, 102:8961–8965, 2005.

[26] C. Pasquier, F. Girardot, K. Jevardat, and R. Christen. Thea: Ontology-driven analysis of microarray data. *Bioinformatics*, 20(16), 2004.

[27] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Efficient mining of association rules using closed itemset lattices. *Information Systems*, 24(1):25–46, 1999.

[28] N. Pasquier, R. Taouil, Y. Bastide, G. Stumme, and L. Lakhal. Generating a condensed representation for association rules. *Journal of Intelligent Information Systems*, 24(1):29–60, January 2005.

[29] P. Pehkonen, G. Wong, and P. Toronen. Theme discovery from gene lists for identification and viewing of multiple functional groups. *BMC Bioinformatics*, 6:162, 2005.

[30] H. Shatkay, S. Edwards, W. W., and B. M. Genes, themes, microarrays: using information retrieval for large-scale gene analysis. In *Proc. ISMB conf.*, pages 340–347, 2000.

[31] A. Tuzhilin and G. Adomavicius. Handling very large numbers of association rules in the analysis of microarray data. In *Proc. ACM SIGKDD conf.*, pages 396–404, 2002.

[32] I. Yang, E. Chen, J. Hasseman, W. Liang, B. Frank, V. Sharov, and J. Quackenbush. Within the fold: assesing differential expression measures and reproducibility in microarray assays. *Genome Biology*, 3:11, 2002.

[33] Y. Zhao, K. McIntosh, D. Rudra, S. Schawalder, D. Shore, and J. Warner. Fine-structure analysis of ribosomal protein gene transcription. *Mollecular Cellular Biology*, 26(13):4853–62, 2006.