BIOINFORMATICS

Vol. 00 no. 00 2008
Pages 1–2

# GenMiner: Mining non-redundant association rules from integrated gene expression data and annotations

Ricardo Martinez [1], Nicolas Pasquier [1] and Claude Pasquier [2]

[1]Laboratoire I3S, UNSA/CNRS UMR-6070, 2000 route des Lucioles, 06903 Valbonne, France
[2]IDBC, UNSA/CNRS UMR-6543, Parc Valrose, 06108 Nice, France

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

**ABSTRACT**

**Summary:** GenMiner is an implementation of association rule discovery dedicated to the analysis of genomic data. It allows the analysis of datasets integrating multiple sources of biological data represented as both discrete values, such as gene annotations, and continuous values, such as gene expression measures. GenMiner implements the new NorDi algorithm for normalizing and discretizing continuous values and takes advantage of the Close algorithm to efficiently generate *minimal non-redundant association rules*. Experiments show that execution time and memory usage of GenMiner are significantly smaller than those of the standard Apriori based approach, as well as the number of extracted association rules.
**Availability:** The GenMiner software and supplementary materials are available at http://bioinfo.unice.fr/publications/genminer_article/.
**Contact:** pasquier@unice.fr

## 1 INTRODUCTION

Association rule discovery (ARD) is an unsupervised data mining technique for discovering links among sets of variable values (*items*) from very large datasets. Association rules (AR) identify groups of items that frequently co-occur in data lines. For each rule, the *support* and *confidence* statistics measure respectively the scope and the precision of the rule. For instance, the rule $AB \rightarrow C$, *support=20%*, *confidence=70%* states that when $A$ and $B$ occur, $C$ also occurs in 70% of cases, and that all three events occur together in 20% of all situations. To limit the extraction to statistically significant ARs, only rules with support and confidence exceeding some user defined minimum support (*minsupp*) and minimum confidence (*minconf*) thresholds are extracted.

In the past years, ARD has been used to find knowledge patterns hidden into various biological datasets. Most of these approaches are based on the Apriori algorithm (Agrawal and Srikant, 1994) that was developed for mining sparse and weakly correlated data. However, when mining dense or correlated data, like most biological data, the efficiency of this algorithm drastically decreases (Brin *et al.*, 1997). Moreover, with such data, a huge number of ARs are extracted and many of them are redundant, thus complexifying their interpretation (Carmona-Saez *et al.*, 2006). Another obstacle in the use of ARD for mining biological data is the need to discretize continuous data before mining.

GenMiner was conceived as a generic ARD tool to mine data originating from any source of biological information. It was designed to address the weaknesses of classical Apriori approaches and fully exploit ARD capabilities. Moreover, a preprocessing of continuous values is not required as GenMiner includes a normalization and discretization algorithm.

## 2 METHODS

GenMiner takes advantage of the Close algorithm to efficiently generate low support and high confidence non-redundant association rules. It also integrates a novel algorithm called NorDi for discretizing continuous values.

Close is a *frequent closed itemsets* based approach (Pasquier *et al.*, 2005) using the closure operator of the Galois connection for extracting minimal non-redundant AR. It was developed to address the efficiency and redundant association rules problems of ARD from dense and correlated data reported in many studies. It first extracts equivalence classes of itemsets, defined by *generators* and *frequent closed itemsets*, and generates from them the *Informative Basis* containing only minimal non-redundant ARs. This basis, or minimal cover, is a generating set for all ARs that captures all information brought by the set of all rules in a minimal number of rules, without information loss (Cristofor and Simovici, 2002). Experiments conducted on ARD benchmark datasets showed that the rule number reduction factor varies from 5 to 400 according to data density and correlation. Several algorithms for extracting frequent closed itemsets, using complex data structures to improve efficiency, have been proposed since Close. However, they do not extract generators, precluding the Informative Basis generation, and their response times, that depend mainly on data density and correlation, are of the same order of magnitude.

The *Normal Discretization* (NorDi) algorithm (Martinez *et al.*, 2007) was developed to improve the discretization of gene expression measures into items. This phase is essential to extract relevant association rules. This algorithm first removes oultiers, detected using the Grubbs outliers method and the Jaque-Bera normality test, to improve the normality of the distribution. Then, the distribution normality is assessed by comparing the resulting distribution against the original distribution using the QQ-plot and the Lilliefors normality tests, and the over-expressed and under-expressed cutoff thresholds are calculated using the *z-score* methodology (Yang *et al.*, 2002). Finally, these cutoffs are used to discretize all values of the initial sample.

**Table 1.** Examples of association rules generated by GenMiner

| Rule | Antecedent | Consequent | Supp. (#) | Conf. (%) |
|---|---|---|---|---|
| 1 | go:0006412 (translation) kegg:sce03010 (ribosome pathway) | heat3↓ | 95 | 74 |
| 2 | go:0005198 (structural molecule activity) go:0005840 (ribosome) | heat3↓ | 96 | 56 |
| 3 | heat3↓ heat4↓ heat5↓ | go:0006412 (translation) | 35 | 88 |
| 4 | heat2↓ | go:0006996 (organelle organization and biogenesis) | 41 | 69 |
| 5 | heat2↓ | go:0042254 (ribosome biogenesis and assembly) | 39 | 66 |
| 6 | heat2↑ heat3↑ heat4↑ | go:0006950 (response to stress) | 15 | 52 |
| 7 | cold4↓ | heat3↓ | 66 | 68 |
| 8 | kegg:sce00190 (purine metabolism) | go:0005737 (cytoplasm) | 52 | 96 |
| 9 | pubmed:16155567 | phenotype:inviable | 168 | 93 |
| 10 | regulator:FHL1 | regulator:RAP1 | 114 | 86 |
| 11 | regulator:RAP1 | regulator:FHL1 | 114 | 61 |

In this table, heat1 to heat6 and cold1 to cold4 refer to the different time points of the heat shock and cold shock experiments respectively. ↑ denotes an over-expression while ↓ denotes an under-expression. The prefixes *go*, *kegg*, *pubmed*, *phenotype* and *regulator* identify gene annotations consisting of GO terms, KEGG pathways, PUBMED ids, phenotype descriptions and name of transcriptional regulators respectively. Supports are given in number of gene and confidences are percentages.

## 3 RESULTS

For demonstration purposes, GenMiner was applied to the well-known Eisen *et al.* (1998) genomic dataset containing expression measures of 2465 *Saccharomyces cerevisiae* genes for 79 biological conditions. Each yeast gene was annotated with its associated terms in *Yeast GO Slim* (a yeast-specific cut-down version of Gene Ontology), its associations with research papers, the KEGG pathways in which it is involved, its phenotypes and the transcriptional regulators that bind its promoter regions. The resulting dataset was a matrix of 2464 lines representing yeast genes and 737 columns representing discretized gene expression measures and gene annotations. GenMiner processed this dataset in 16 minutes on a standard desktop machine using a *minsupp* of 0.5%, a *minconf* of 50% and a discretization confidence level of 95%.

Table 1 shows some examples of rules extracted by GenMiner. Rules 1 and 2 highlight a general reduction of protein synthesis, ribosomal organization and cell maintenance following a heat shock, leading to cellular damages. Rules 3 to 6 show that genes under-expressed during the heat shock experiment are involved in protein synthesis, cellular organization and ribosomal organization while over-expressed genes are involved in stress response. Rule 7 highlights a group of genes that are under-expressed after both a heat shock and a cold shock. Rules 8 and 9 reveal possible links between annotations from different sources like the relationship between the pathway 'purine metabolism' and the gene ontology term 'cytoplasm' or the strong association between the genes cited in a scientific paper presenting a review of the essential yeast genes and the phenotype inviable. Rules 10 and 11 enable to state strong relationship between transcriptional regulators *FHL1* and *RAP1* as genes regulated by *FHL1* are also regulated by *RAP1*. The reverse is less true since there is a considerable proportion of genes regulated by *RAP1* and not by *FHL1*. These two rules reflect the known fact that *RAP1* binding is essential for the recruitment of *FHL1* as described in many articles.

These experiments confirmed that GenMiner allows the use of smaller *minsupp* thresholds than Apriori based approaches and significantly reduces the number of extracted association rules, facilitating their exploration and interpretation by the end-user.

## 4 CONCLUSION

The GenMiner ARD approach was developed for processing high-dimensional biological datasets integrating both gene expression measures and biological annotations. It was designed considering the specific characteristics of such biological data that are noisy, dense and highly correlated. It uses a frequent closed itemsets based approach that allows to discover low-support, or rare, rules representing associations between very small groups of genes and to reduce the number of extracted association rules without information loss. The application of GenMiner to a dataset integrating the well-known Eisen *et al.* (1998) gene expression dataset with corresponding gene annotations demonstrated its capability to extract meaningful associations between gene expression profiles and annotations. Furthermore, it showed the potential of this approach to integrate several heterogeneous sources of information. This is only an example of GenMiner capabilities, as it can easily integrate every kind of gene annotations obtained from any source of biological information. With these features, GenMiner is an ARD approach that is adequate to biologist requirements for the analysis of genomic data.

## REFERENCES


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

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

Carmona-Saez, P., Chagoyen, M., Rodríguez, A., Trelles, O., Carazo, J., and Pascual-Montano, A. (2006). Integrated analysis of gene expression by association rules discovery. *BMC Bioinformatics*, **7**(1):54–69.

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

Eisen, M., Spellman, P., Brown, P., and Botsein, D. (1998). Cluster analysis and display of genome wide expression patterns. In *Proc. Nat. Aca. Sci. USA*, **95**, 14863–8.

Martinez, R., Pasquier, C., and Pasquier, N. (2007). GenMiner: Mining informative association rules from genomic data. In *Proc. IEEE BIBM conf.*, pages 15–22.

Pasquier, N., Taouil, R., Bastide, Y., Stumme, G., and Lakhal, L. (2005). Generating a condensed representation for association rules. *J. Intell. Inf. Syst.*, **24**(1), 29–60.

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