Toward the discovery of itemsets with significant variations in gene expression matrices

Mehdi Kaytoue-Uberall, Sébastien Duplessis, Amedeo Napoli

Toward the discovery of itemsets with significant variations in gene expression matrices

Mehdi Kaytoue-Uberall, Sébastien Duplessis and Amedeo Napoli

Abstract This paper presents new syntactic constraints for itemset mining in gene expression matrices. Biologists are interested in identifying gene expression profiles which present similar quantitative variation features. A two dimensional gene expression profile representation is introduced and adapted to itemset mining allowing to control gene expression. Syntactic constraints introduce expert knowledge at the beginning of the Knowledge Discovery in Databases process and are used to discover itemsets with significant expression variations from a large collection of gene expression profiles.

1 Introduction

Each gene in a cell is expressed with respect to ongoing cellular processes or in response to particular stresses. The expression of a gene affects the production of *RNAm* and generally of proteins. Recent microarray technologies allow simultaneous quantitifcation of the expression of each single gene of a genome in a given biological situation. The expression of a gene in several biological situations determines its *gene expression profile*. Gene expression matrices present thousands of profiles: numerical clustering methods, e.g. hierarchical clustering, K-Means and Self Organizing Maps (see [3] for a survey), are generally applied to find profile clusters.

Meanwhile, symbolic data mining methods such as itemset search [6, 1] and association rules extraction [2] are emerging thanks to their ease of result interpre-

Mehdi Kaytoue-Uberall
LORIA, Campus Scientifique, Vandoeuvre-les-Nancy, e-mail: mehdikaytoueuberall@loria.fr

Sébastien Duplessis
INRA, Champenoux e-mail: duplessi@nancy.inra.fr

Amedeo Napoli
LORIA, Campus Scientifique, Vandoeuvre-les-Nancy, e-mail: amedeo.napoli@loria.fr

tation and strong local association of clusters. These methods take binary data as input and the following problems may appear: (i) data need to be binarized introducing biases and loss of information, (ii) computed itemset collections are huge, and (iii) a few proportion of itemsets is biologically relevant for a given biological study.

This paper presents a set of new syntactic constraints for managing numerical expression variations by introducing a two dimensional gene representation (see Section 2) also applied to itemsets (Section 3). This representation allows a biologist to easily determine with the help of basic operations what sort of gene expression profile may be of interest. Furthermore, constraints ensures that only itemsets presenting desired variations are discovered, thus answering problems (ii) and (iii). A solution for (i) is discussed in the research perspectives.

2 Gene representation

A gene expression profile (GEP) is considered as a vector of numerical values such as $(1050, 1950, 1503)$ for Gene 1 in Table 1 describing the expression of the gene in given situations (S_a, S_b, S_c). Modalities are used for describing the expression of the gene, e.g. W (Weak), N (Normal), and S (Strong). Thus a gene can be represented as a 2D vector $g = \{(a_1, n_1), ..., (a_p, n_p)\}$ where a_k is a biological situation and n_k is an integer associated to the modality qualifying the situation. In our example, the set of modalities is $\{W, N, S\}$ corresponding to values of GEP respectively ranging in the intervals $[0, 5000[$, $[5000, 10000[$ and $[10000, 20000]$. The integer 0 is associated to W, 1 to N and 2 to S. The bounds of the intervals are dependent on expert knowledge. Figure 1 shows the histogram associated to Table 1. A bar indicates the number of genes that have expression in the same interval in a situation. For example, two genes have modality W (Gene 1, Gene 2), two others have modality N (Gene 3, Gene 4) and one has modality S (Gene 5). Now, the 2D-GEP vector for, e.g. Gene 5, is $\{(S_a, 2), (S_b, 0), (S_c, 2)\}$.

	S_a	S_b	S_c
Gene 1	1050	1950	1503
Gene 2	3025	4100	1708
Gene 3	8057	5057	1500
Gene 4	7392	6020	1300
Gene 5	16070	3021	17548

Table 1 An example of Microarray gene expression dataset.

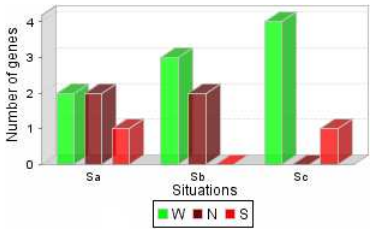

Fig. 1 The histogram associated to Table 1.

Biologists are interested in groups of genes showing similar expression levels. They can infer similar biological functions or membership to the same cellular process [4]. Two types of groups are distinguished: varying and constant groups. Typically, the biologist focuses on small and homogeneous gene groups presenting the most important variations simultaneously. Interpretation of variations leads after ex-

perimental validations to the discovery of gene functions. Large variations are important and allow to discriminate genes responsible of a particular cellular process or diseases [4].

In a 2D-GEP vector, an amplitude is defined as the largest difference between all values n_k ($k \in [2,p]$, not defined otherwise). The higher the amplitude is the higher the biological relevance is (see Table 2). Different groups of genes can be distinguished depending on their amplitude: null or non null amplitude. Genes presenting the same 2D-GEP are considered to be in a same group. For example, Gene 1 and Gene 2 have the same 2D-GEP.

Gene	2D-GEP	Maximal Amplitude	Biological relevance
{Gene 1, Gene 2}	$\{(S_a,0),(S_b,0),(S_c,0)\}$	$0-0=0$	low
{Gene 3, Gene 4}	$\{(S_a,1),(S_b,1),(S_c,0)\}$	$1-0=1$	medium
{Gene 5}	$\{(S_a,2),(S_b,0),(S_c,2)\}$	$2-0=2$	high

Table 2 2D-GEP vectorial representation.

3 Definition of 2D-itemsets for GEP

In this section, standard itemset search [5] is adapted to 2D-GEP. A 2D-itemset $X = \{(a_1,n_1),(a_2,n_2),...,(a_p,n_p)\}$ is a set of pairs where p is the length of the 2D-itemset, a_k is an attribute name, and n_k is the value of a_k ($k \in [1,p]$).

Given a set of genes G, a set of attributes M (actually, attributes denote pairs (a_i,n_i)), a relation $I \subseteq G \times M$ is defined as follows: $(g,m) \in I$ means that the gene g include m in its 2D-GEP. For example, $\{(S_a,0)\}$ is a 2D-itemset of length 1, $\{(S_a,1),(S_b,1)(S_c,0)\}$ is a 2D-itemset of length 3. The image of an itemset X is the set of objects including the itemset. The cardinality of the image of X is called *support*. An itemset is *frequent* if its support is greater than a given minimal frequency. An itemset is *closed* if adding an item changes its support.

Gene	$(S_a,0)$	$(S_a,1)$	$(S_a,2)$	$(S_b,0)$	$(S_b,1)$	$(S_b,2)$	$(S_c,0)$	$(S_c,1)$	$(S_c,2)$
Gene 1	×			×			×		
Gene 2	×			×			×		
Gene 3		×			×		×		
Gene 4		×			×		×		
Gene 5			×	×					×

Table 3 Binary table derived from 2D-GEP. Attributes denote pairs.

Biological application. A biologist may be interested in mining 2D-itemsets from a set of 2D-GEP with respect to a set of constraints. It allow biologists to introduce their knowledge and preferences for a given biological study at the beginning of a Knowledge Discovery in Databases process. A lot of combinations are possible in order to formalize expert knowledge [7].

Frequent itemsets (support)	$\{(S_a,0)\}(2)$, $\{(S_a,1)\}(2)$, $\{(S_b,1)\}(2)$, $\{(S_a,0),(S_b,0)\}(2)$, $\{(S_a,0),(S_c,0)\}(2)$, $\{(S_a,1),(S_b,1)\}(2)$, $\{(S_a,1),(S_c,0)\}(2)$, $\{(S_b,0),(S_c,0)\}(2)$, $\{(S_b,1),(S_c,0)\}(2)$
Frequent closed itemsets (support) (maximal amplitude)	$\{(S_b,0)\}(3)$(-), $\{(S_c,0)\}(4)$(-), $\{(S_a,0),(S_b,0),(S_c,0)\}(2)(0)$, $\{(S_a,1),(S_b,1),(S_c,0)\}(2)(1)$

Table 4 Frequent and frequent closed itemsets from Table 3. Minimal frequency is 2/5.

For example, the biologist can search for itemsets presenting high variations between a normal (a_1) and a cancerous cell (a_2). A 2D-itemset $Z = \{(a_1,n_1),(a_2,n_2)\}$ is interesting iff $|n_1 - n_2| \geqslant \delta$, where δ is a minimum threshold. In the same way, a biologist can search for itemsets with high variations, i.e. $max\{|n_i - n_j|, 0 \leqslant j < i \leqslant p\} \geqslant \delta$. Some itemsets have a high global expression whereas others have a low one. The latter itemsets are generally patterns of specific genes such as those encoding receptors or transcription factors associated to low expression levels. The average on all n_k may be used in the constraint.

4 Conclusion and perspectives

In this paper, we have presented a method for transforming complex data describing gene expression profiles into standard binary tables. In this way, we are able to apply classical itemset search algorithms (e.g. Apriori or Charm [8]) or association rule extraction algorithms to these complex data. This transformation is based on 2D-itemsets, an extension of classical symbolic itemset. Moreover, knowledge domain can be embedded into constraints to be used within the data mining process.

References

1. Blachon, S., Pensa, R.G., Besson, J., Robardet, C., Boulicaut, J-F., Gandrillon, O.: Clustering formal concepts to discover biologically relevant knowledge from gene expression data. In: Silico Biology **7**, 0033 (2007).
2. Creighton C., Hanash S.: Mining gene expression databases for association rules. In: Bioinformatics, **19**(1), 79–86 (2003).
3. Jiang, D., Chun, T., Tang, Zhang, A.: Cluster Analysis for Gene Expression Data: A Survey, IEEE Transactions on Knowledge and Data Engineering , **16**(11), 1370–1386 (2004)
4. Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., Zeitlinger, J., Jennings, E. G., Murray, H. L., Gordon, D. B., Ren, B., Wyrick, J. J., Tagne, J. B., Volkert, T. L., Fraenkel, E., Gifford, D. K., Young, R. A.: Transcriptional regulatory networks in Saccharomyces cerevisiae. In: Science, 298, 799–804 (2002).
5. Pasquier, N., Bastide, Y., Taouil, R., Lakhal, L.: Discovering frequent closed itemsets for association rules. In Beeri and Buneman editors, Database Theory, 398–416, Springer (1999).
6. Pensa, R.G., Besson, J., Boulicaut, J-F.: A methodology for biologically relevant pattern discovery from gene expression data. In : Proc. of the 7th Int. Conf. on Discovery Science (2004).
7. Perng, C.S., Wang, H., Ma, S., Hellerstein, J.L.: Discovery in multi-attribute data with user-defined constraints. In: SIGKDD Explorations, 4, 1, 56 – 64 (2002).
8. Zaki, M.J., Hsiao, C.J.: CHARM: An Efficient Algorithm for Closed Itemset Mining. In: Proc. SIAM Intl Conf. Data Mining, 457–473, (2002).