Assessment of discretization techniques for relevant pattern discovery from gene expression data

Ruggero G. Pensa[1], Claire Leschi[1], Jérémy Besson[1,2] and Jean-François Boulicaut[1]
1: INSA Lyon, LIRIS CNRS FRE 2672, F-69621 Villeurbanne cedex, France
2: UMR INRA/INSERM 1235, F-69372 Lyon cedex 08, France
{ruggero.pensa,claire.leschi,jeremy.besson,jean-francois.boulicaut}@insa-lyon.fr

ABSTRACT

In the domain of gene expression data analysis, various researchers have recently emphasized the promising application of pattern discovery techniques like association rule mining or formal concept extraction from boolean matrices that encode gene properties. To take the most from these approaches, a needed step concerns gene property encoding (e.g., over-expression) and its need for the discretization of raw gene expression data. The impact of this preprocessing step on both the quantity and the relevancy of the extracted patterns is crucial. In this paper, we study the impact of discretization parameters by a sound comparison between the dendrograms, i.e., trees that are generated by a hierarchical clustering algorithm, computed from raw expression data and from the various derived boolean matrices. Thanks to a new similarity measure and practical validation over several gene expression data sets, we propose a method that supports the choice of a discretization technique and its parameters for each specific data set.

1. INTRODUCTION

Thanks to a huge research effort and technological breakthroughs, one of the challenges for molecular biologists is to discover knowledge from data generated at very high throughput. For instance, different techniques (including microarray [13] and SAGE [24]) enable to study the simultaneous expression of (tens of) thousands of genes in various biological situations. The data generated by those experiments can be seen as expression matrices in which the expression level of genes (rows) are recorded in various biological situations (columns). A toy example of some microarray data is the matrix in Tab. 1a.

Exploratory data mining techniques are needed that can, roughly speaking, be considered as the search for interesting bi-sets, i.e., sets of biological situations and sets of genes that are associated in some way. Indeed, it is interesting to look for groups of co-regulated genes, also known as *synexpression groups* [19], which, based on the guilt by association approach, are assumed to participate in a common function, or module, within the cell. A set of co-regulated genes and the set of biological situations that gives rise to this co-regulation is called a *transcription module*. Discovering transcription modules is one of the main goals in functional genomics.

Various techniques can be used to identify a priori interesting bi-sets. Biologists often use clustering techniques to identify sets of genes that have similar expression profiles (see, e.g., [14]). Statistical methods can be used as well (see, e.g., [16; 4]). It is also possible to look for these putative synexpression groups by computing the so-called frequent itemsets from boolean contexts that encode gene expression properties [1]. Deriving association rules from frequently co-regulated genes has been studied as well [3; 10]. Furthermore, putative transcription modules can be provided by computing the so-called formal concepts (see, e.g., [25]) in this kind of boolean data [21; 22].

A key issue for using these pattern discovery techniques from boolean data concerns gene expression property encoding. Let $\mathcal{O}$ denotes a set of biological situations and

	1	2	3	4	5
a	-1	6	0	12	9
b	3	-2	3	-3	1
c	0	5	-1	6	6
d	4	-1	2	-2	-1
e	-3	9	1	10	6
f	5	-3	3	-6	0
g	4	-4	3	-7	0
h	-2	2	-2	8	5

(a)

	1	2	3	4	5
a	0	1	0	1	1
b	1	0	1	0	1
c	0	1	0	1	1
d	1	0	1	0	0
e	0	1	0	1	1
f	1	0	1	0	1
g	1	0	1	0	1
h	0	0	0	1	1

(b)

	1	2	3	4	5
a	0	0	0	1	0
b	1	0	1	0	0
c	0	0	0	1	1
d	1	0	0	0	0
e	0	0	0	1	0
f	1	0	0	0	0
g	1	0	0	0	0
h	0	0	0	1	0

(c)

Table 1: An example of gene expression matrix (a) with two derived boolean matrices (b and c)

$\mathcal{P}$ denotes a set of genes. The expression properties can be encoded into $\mathbf{r} \subseteq \mathcal{O} \times \mathcal{P}$. $(o_i, g_j) \in \mathbf{r}$ denotes that gene j has the encoded expression property in situation i. Different expression properties might be considered like, e.g., over-expression, up or down regulation, strong variation. Generally, encoding is performed according to some discretization operators that, given user-defined parameters, transform each numerical value from raw gene expression data into one boolean value per gene property. Many operators can be used that typically compute thresholds from which it is possible to decide wether the true or the false value must be assigned. For instance, in Tab. 1b, an over-expression property has been encoded and genes a, c, and e are over-expressed together in situations 2, 4 and 5.

We consider that mining boolean gene expression data sets is extremely useful thanks to the patterns that can be extracted now with efficient algorithms (e.g., frequent closed set [7; 20; 26] or concept extractors [5]). In this context, the critical step of gene expression data discretization has not been studied enough while its impact on both the quantity and the relevancy of the extracted patterns is crucial. For instance, the density of the discretized data depends on the discretization parameters and the cardinalities of the resulting sets (collections of itemsets, association rules or formal concepts) can be very different.

In this paper, we propose a method that supports both the choice for a discretization technique and an informed decision about its parameters. We cooperate with molecular biologists that are used to collect important information about putative synexpression groups and transcription modules by using the hierarchical clustering technique that has been popularized by the Eisen software [14]. We decided to study the impact of discretization parameters by a sound comparison between the dendrograms that are generated by the same hierarchical clustering algorithm applied to both the raw expression data and various derived boolean matrices. Comparing trees by means of ad-hoc similarity measures has been studied a lot, including in the bioinformatics domain for the analysis of phylogenies (see, e.g., [18; 23; 15]). Other measures evaluate the quality of partitions w.r.t. a reference partition of the data set. The difficulty is then to identify on dendrograms the cut levels at which we can compare the partition on the real data set with the one on boolean data set.

The contribution of this paper is twofold. First, we propose a new similarity measure for binary trees (such as dendrograms generated by any hierarchical clustering algorithm), that is level independent, and depends for each node on its subtree structure. Next, we have studied the behavior of our measure on several gene expression data sets in order to support the choice a discretization technique and the discretization parameters that have to be used when encoding boolean gene expression properties in order to perform efficient pattern discovery techniques like association rule mining or formal concept discovery.

In Section 2, we define our similarity measure between two binary trees. In Section 3, we study the behavior of this measure for different gene expression data sets. Finally we consider in Section 4 the impact of our technique on a KDD process which finds biologically relevant information in a well-studied gene expression data set. Section 5 concludes.

2. COMPARING BINARY TREES

The problem of finding the best comparison method for trees is still open even though it has been considered in various application domains. Considering the analysis of phylogenies, distance measures between both rooted and unrooted trees have been designed to compare different phylogenetic trees concerning the same set of individuals (e.g., different species of animals having a common ancestor). Various distance metrics between trees have been proposed. The `nni` (nearest neighbor interchange) and the `mast` (maximum agreement subtree) are two of the most used metrics. `nni` has been introduced independently in [18] and [23] and its NP-completeness has been recently proved [11; 12]. `mast` has been proposed in [15], and [9] describes an efficient algorithm for computing this metrics on binary trees. These two approaches are tailored for the problem of comparing phylogenies where the goal is to measure some degree of isomorphism between two dendrograms representing the same species of biological organisms.

In our data mining problem, we have sets of objects (vectors of expression values for genes in various biological situations), that we want to process with a hierarchical clustering algorithm. Depending on the different discretization operations on raw expression data, a same clustering algorithm working on encoded boolean gene expression data can return (very) different results. We are looking for a method that supports the comparison of these various gene and/or situation dendrograms obtained on boolean data w.r.t. the common reference dendrogram that has been computed from the raw data. We need to measure both the degree of similarity of their structures and the similarity between the contents of their associated collections of clusters. We designed a simple measure which is also easy to compute. Intuitively, it depends on the number of matching nodes between the two trees we have to compare.

2.1 Definition of similarity scores

Let $\mathcal{O} = \{o_1, \ldots, o_n\}$ denote a set of n objects. Let T denote a binary tree built on $\mathcal{O}$. Let $\mathcal{L} = \{l_1, \ldots, l_n\}$ denote the set of n leaves of T associated to $\mathcal{O}$ for which, $\forall i \in [1 \ldots n], l_i \equiv o_i$. Let $\mathcal{B} = \{b_1 \ldots b_{n-1}\}$ denote the set of $n-1$ nodes of T generated by a hierarchical clustering algorithm starting from $\mathcal{L}$. By construction, we consider $b_{n-1} = r$, where r denotes the root of T. We define the two sets:

$$\delta(b_i) = \{b_j \in \mathcal{B} \mid b_j \textit{ is a descendent of } b_i\},$$

$$\tau(b_i) = \{l_j \in \mathcal{L} \mid l_j \textit{ is a descendent of } b_i\}.$$

An example of a tree for a set containing 8 objects (i.e., the genes from Tab. 1a) is given in Fig. 1. In this example, $\tau(b_3) = \{b, d, f, g\}$ and $\delta(b_3) = \{b_1, b_2\}$.

We want to measure the similarity between a tree T and a reference tree T_{ref} built on the same set of objects $\mathcal{O}$. For each node b_i of T, we define the following score (denoted

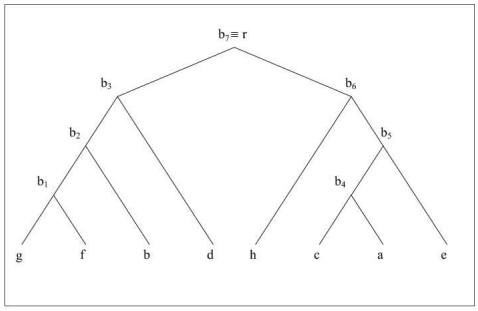


Figure 1: An example of binary tree

S_B and called **BScore**):

$$S_B(b_i, T_{ref}) = \sum_{b_j \in \delta(b_i)} a_j$$

$$a_j = \begin{cases} \frac{1}{|\tau(b_j)|}, & if \\ & \exists b_k \in T_{ref} \mid \tau(b_j) = \tau(b_k) \\ 0, & otherwise \end{cases} \quad (1)$$

In other terms, for a node b in T, its score depends both on the number of its matching nodes in T_{ref} ($b_k \in T_{ref}$ is a matching node for b if $\tau(b) = \tau(b_k)$) and $|\tau(b)|$. To obtain the similarity score of T w.r.t. T_{ref} (denoted S_T and called **TScore**), we consider the **BScore** value on the root, i.e.:

$$S_T(T, T_{ref}) = S_B(r, T_{ref}) \quad (2)$$

As usually, it is interesting to normalize the measure to get a score between 0 (for a tree which is totally different from the reference) and 1 (for a tree which is equal to the reference). For the **TScore** measure, since its max value depends on the tree morphology, we can normalize by $S_T(T_{ref}, T_{ref})$:

$$\overline{S_T}(T, T_{ref}) = \frac{S_T(T, T_{ref})}{S_T(T_{ref}, T_{ref})} \quad (3)$$

$\overline{S_T}(T, T_{ref}) = 0$ means that T is totally different from T_{ref}, i.e., there are no matching nodes between T and T_{ref}. Indeed, $\overline{S_T}(T, T_{ref}) = 1$ means that T is totally similar to T_{ref}, i.e., every node in T matches with a node in T_{ref}. Given two trees T_1 and T_2 and a reference T_{ref}, if $\overline{S_T}(T_1, T_{ref}) < \overline{S_T}(T_2, T_{ref})$, then T_2 is said to be more similar to T_{ref} than T_1 according to **TScore**.

Let us now provide a constructive definition to compute the **BScore**s for every node of the tree, and retrieve its value for the whole tree. Assume that functions $c_l(b_i)$ and $c_r(b_i)$ respectively return the left and the right child of b_i. In Fig. 1 $c_l(b_7) = b_3$ et $c_r(b_7) = b_6$. The **BScore** measure can be redefined as follows:

$$S_B(b_i, T_{ref}) = \sigma(c_l(b_i), T_{ref}) + \sigma(c_r(b_i), T_{ref}) \quad (4)$$

where

$$\sigma(b_k, T_{ref}) = \begin{cases} \frac{1}{|\tau(b_k)|} + S_B(b_k, T_{ref}), & if \\ \quad \exists b_j \in T_{ref} \mid \tau(b_k) = \tau(b_j) \\ S_B(b_k, T_{ref}), & otherwise \end{cases}$$

$$\sigma(l_k, T_{ref}) = 0, \ \forall l_k \in \mathcal{L}$$

This definition emphasizes that the **BScore** for each node depends on the **BScore** values of its children. The fact that each node "inherits" the similarity information of its children is useful when comparing two trees that result from a hierarchical clustering algorithm.

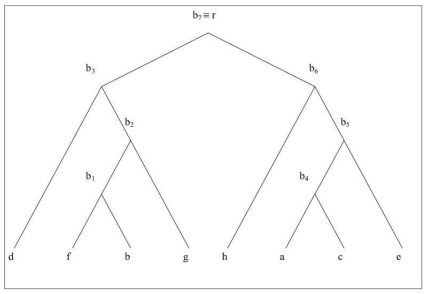
2.2 Comparison between gene dendrograms

Tab. 1a is a toy example of a gene expression matrix. Each row represents a gene vector, and each column represents a biological sample vector. Each cell contains an expression value for a given gene and a given sample. In this example, we have $\mathcal{O} = \{a, b, c, d, e, f, g, h\}$. A hierarchical clustering using the Pearson's correlation coefficient and the average linkage method (see, e.g., [14]) on the data from Tab. 1a leads to the dendrogram in Fig. 1.

Assume now that we discretize the expression matrix by applying two different methods used for over-expression encoding [3]. The first one considers the mean between the max and min values for each gene vector. Values that are greater than the average value are set to 1, 0 otherwise (Tab. 1b). A second method takes into account the max value for each gene vector. Values that are greater than 90% of the max value are set to 1, O otherwise (Tab. 1c).

Assume now that we use the same clustering algorithm on the two derived boolean data sets. The resulting dendrograms are shown in Fig. 2. Fig. 2a (resp. Fig. 2b) represents the gene dendrogram obtained by clustering the boolean matrix in Tab. 1b (resp. Tab. 1c).

a)



b)

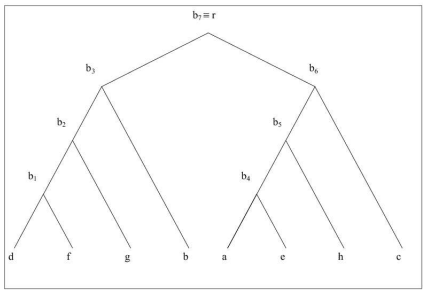


Figure 2: Gene trees built on two differently discretized matrices

We can now use the similarity score and decide which discretization is better for this gene expression data set, i.e., the one for which $\overline{S_T}(T, T_{ref})$ has the largest value. The common reference (T_{ref}) is the tree in Fig. 1. Let T_a and T_b denote the trees in Fig. 2a and 2b respectively. Using Equation 4, we obtain the results in Tab. 2.

To normalize the similarity scores, we just need to divide the **BScore**s of the root of the first two dendrograms,

	T_a			
Node	*Match*	τ	σ	S_B
b_1	-	$\{b, f\}$	0	0
b_2	b_2	$\{b, f, g\}$	0.33	0
b_3	b_3	$\{b, d, f, g\}$	0.58	0.33
b_4	b_4	$\{a, c\}$	0.5	0
b_5	b_5	$\{a, c, e\}$	0.83	0.5
b_6	b_6	$\{a, c, e, h\}$	1.08	0.83
b_7	b_7	$\mathcal{O}$	-	1.67

	T_b			
Node	*Match*	τ	σ	S_B
b_1	-	$\{d, f\}$	0	0
b_2	-	$\{d, f, g\}$	0	0
b_3	b_3	$\{b, d, f, g\}$	0.25	0
b_4	-	$\{a, e\}$	0	0
b_5	-	$\{a, e, h\}$	0	0
b_6	b_6	$\{a, c, e, h\}$	0.25	0
b_7	b_7	$\mathcal{O}$	-	0.5

	T_{ref}			
Node	*Match*	τ	σ	S_B
b_1	b_1	$\{f, g\}$	0.5	0
b_2	b_2	$\{b, f, g\}$	0.83	0.5
b_3	b_3	$\{b, d, f, g\}$	1.08	0.83
b_4	b_4	$\{a, c\}$	0.5	0
b_5	b_5	$\{a, c, e\}$	0.83	0.5
b_6	b_6	$\{a, c, e, h\}$	1.08	0.83
b_7	b_7	$\mathcal{O}$	-	2.17

Table 2: **BScore** values. Nodes matching in T_{ref} are in the *Match* columns.

by the **BScore** of the root of the reference dendrogram (Equation 3):

$$\overline{S_T}\,(T_a, T_{ref}) = \frac{S_T\,(T_a, T_{ref})}{S_T\,(T_{ref}, T_{ref})} = \frac{1.67}{2.17} = 0.77$$

$$\overline{S_T}\,(T_b, T_{ref}) = \frac{S_T\,(T_b, T_{ref})}{S_T\,(T_{ref}, T_{ref})} = \frac{0.5}{2.17} = 0.23$$

Since $\overline{S_T}\,(T_a, T_{ref}) > \overline{S_T}\,(T_b, T_{ref})$, the first discretization method is considered better for this data set w.r.t. the performed hierarchical clustering. In fact, in T_a, only node b_1 does not match (i.e., it does not share the same set of leaves) with any node in T_{ref}, while in T_b, there are only two nodes (b_3 and b_6) that match with some nodes in T_{ref}.
The same process can be applied to situation dendrograms by considering now that the objects are the situations. In practice, we perform both processes to support the choice of a discretization technique as illustrated in the next section.

3. COMPARING DIFFERENT DISCRETIZATION TECHNIQUES

Many discretization techniques can be used to encode gene expression properties from expression values that are either integer values (case for SAGE data [24]) or real values (case for microarray data [13]). In this paper, we consider for our experimental study only three techniques that have been used for encoding the over-expression of genes in [3]:

- "Mid-Ranged". The highest and lowest expression values are identified for each gene and the mid-range value is defined. For a given gene, all expression values that are strictly above the mid-range value give rise to value 1, 0 otherwise.
- "Max - X% Max". The cut off is fixed w.r.t. the maximal expression value observed for each gene. From this value, we remove a percentage X of this value. All expression values that are greater than the $(100 - X)\%$ of the Max value give rise to value 1, 0 otherwise.
- "X% Max". For each gene, we consider the situations in which its level of expression is in X% of the highest values. These genes are assigned to value 1, 0 otherwise.

We want to evaluate the relevancy of a discretization algorithm and its parameters according to the preserved properties w.r.t. a hierarchical clustering of the raw data. So, we have to compare the dendrograms obtained from the three different boolean matrices with the reference dendrogram.
We have considered three gene expression data sets: two microarray data sets and a SAGE data set. The first data set (CAMDA [8]) concerns the transcriptome of the intraerythrocytic developmental cycle of the plasmodium falciparum, a parasite that is responsible for a very frequent form of malaria. We have the expression values for 3 719 genes in 46 different time points. The second data set (Drosophila [2]) concerns the gene expression of drosophila melanogaster during its life cycle. We have the expression values for 3 030 genes and 81 biological samples, including both male and female adult individuals. The third one (human SAGE data from NCBI, see also [17; 22]) contains the expression values for 5 327 human genes in 90 different cancerous and not cancerous cellular samples belonging to different human organs.
In Tab. 3, we report the densities (i.e., the ratio of true values) of the boolean matrices produced with the "Mid-Ranged" method. In Fig. 3, we provide the density curves for the three data sets and depending on different thresholds for the "Max - X% Max" method (densities for the "X% Max" method are quite similar).

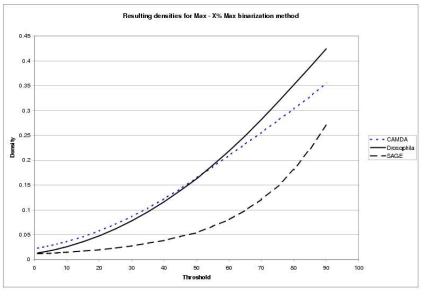


Figure 3: Density values for different "Max - X% Max" thresholds

We processed all the computed boolean matrices with a hierarchical clustering algorithm based on the centered Pearson's correlation coefficient and the average linkage method. The same algorithm with the same options has been applied to the three original matrices. Finally, for each data set, we have compared all the genes and situations trees derived from the boolean matrices with the reference trees. The results in terms of **TScore** (Equation 4) for the "Mid-Ranged" method, are summarized in Tab. 3. For the "Max - X% Max" and "X% Max" meth-

ods we summarize the results depending on the variation of the threshold X for the gene dendrograms in Fig. 4a and Fig. 4c, for the situation dendrograms in Fig. 4b and Fig. 4d. It is important to observe that, for each data set, we obtained the highest values of similarity scores for both the genes and the situations for almost the same discretization thresholds.

		Similarity score	
Data set	**Density**	*Genes*	*Situations*
CAMDA	0.313665	0.034155	0.746437
Drosophila	0.441146	0.059570	0.591343
SAGE	0.053958	0.110131	0.776736

Table 3: Similarity scores for clustering trees on Mid-Ranged discretized matrices

We have also applied the same clustering algorithm on various randomly generated boolean matrices based on the same sets of objects. Then, we have compared the resulting dendrograms with the reference. In the first two data sets (CAMDA and Drosophila), the similarity scores of the randomly generated boolean matrices are always very low or equal to 0. In the SAGE data set, given a density value, the gene scores resulting from randomly generated matrices are always lower than the ones obtained by any discretization method (while the situation scores are always negligible). One possible reason is that the discretized matrices are here very sparse compared to the first two data sets (see Fig. 3). Using a low threshold to discretize such a matrix does not make sense: obtained scores are similar to the scores which are computed on random boolean matrices. Moreover, using a high threshold value X for the "X% Max" discretization method leads to similarity scores that are close to those obtained for randomly generated matrices, though still higher. We can observe the behavior of this particular SAGE data set in Fig. 5.

To conclude this section, comparing dendrograms resulting from the clustering of different types of derived boolean matrices enables to choose the "best" discretization method and parameters for a given data set. If we analyze the graphics of similarity scores w.r.t. the thresholds used in the "Max - X% Max" and "X% Max" methods (see Fig. 4), we observe the presence of either a max or an asymptotic behavior. It means that the best choice for the discretization threshold could be a trade-off between the value for which we get the best similarity score, and the value for which the data mining task remains tractable.

4. AN APPLICATION TO A REAL PROBLEM

We have applied the proposed preprocessing technique to a real gene expression data set to validate our approach throughout a complete KDD process. We have decided to mine the data described in [2]. It concerns the gene expression of the Drosophila melanogaster during its life cycle. The expression levels of 4 028 genes are evaluated for 66 sequential time periods from the embryonic state till the adulthood. The total number of samples is 81 since the gene expression during the adult state is measured for male and female individuals. For our experiment we have used only a set of 20 time periods for adult individuals. This set is composed of 8 male adult samples, 8 female adult samples, 2 male and 2 female tudor samples. The set of genes we have used is derived from the original set from which we removed those genes that are under-expressed in all the 20 situations and over-expressed in at least 11 biological situations. We have obtained a 3 433 $\times$ 20 matrix which has been processed according to our methodology. The raw expression matrix has been discretized using the "Mid-Ranged" and "Max - X% Max" methods. The resulting boolean matrices and the original matrix have been processed with the same ascendant hierarchical clustering algorithm using Pearson's correlation coefficient and average linkage. Then, using our tree comparison technique, we have compared the gene and situation dendrograms. The similarity scores are presented in Fig. 6.

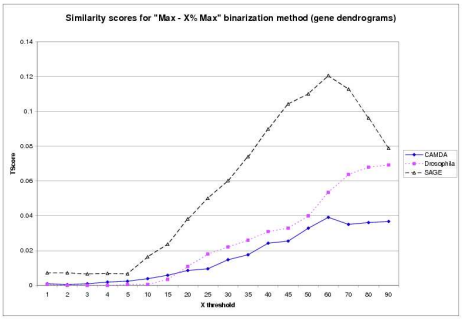
Our goal was to identify a particular class of genes, the so-called "male somatic genes", that characterizes the male adult individuals (see Table S30 in [2]). 31 of these 37 genes are present in our data set and we tried to search them by mining formal concepts in the various derived boolean matrices. Intuitively, formal concepts are maximal rectangles of true values in boolean matrices. For instance, in the boolean context from Tab. 1b, ({a,c,e},{2,4,5} is a formal concept, i.e., a strong association between two closed sets. We used the D-Miner algorithm [5; 6]) which extracts all the concepts satisfying some user-defined monotonic constraints. We extracted all the concepts with at least 3 situations and at least 20 genes. Then we have post-processed the extracted collection to keep those which concern only male individuals. Finally, we measured the number of male somatic genes which appear in the different sets of the post-processed concepts. To better evaluate the results, we also built two other sets of concepts: the collection of concepts which concern only female individuals, and the collection of concepts which involve at least one female individual. We summarize the results in Fig. 7.

The discretization threshold that gives the best similarity score and that we identify in both graphs from Fig. 6 (X = 54% for the "Max - X% Max" method), enables to retrieve 25 of the 31 male somatic genes from the concepts that concern only male individuals. Moreover, even though higher thresholds enable to retrieve more somatic genes, the slope of the curve, after the optimum, begins to decrease, while the slope of the curves of male somatic genes identified in concepts concerning female individuals starts to increase. Choosing the discretization threshold enables to control the trade-off between extraction completeness and noise impact.

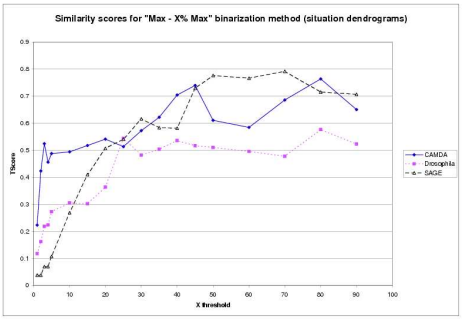
5. CONCLUSION

We defined a new pre-processing technique that supports the evaluation and assessment of different discretization techniques for a given gene expression data set. The evaluation is based on the comparison of dendrograms obtained by clustering various derived boolean matrices with the one obtained on the raw matrix while using the same clustering algorithm. The defined metrics is simple and we have validated its relevancy on different real data sets and on a biological problem. This is a step towards a better understanding of a crucial pre-processing step when we want to apply the extremely promising pattern discovery techniques based on set patterns.

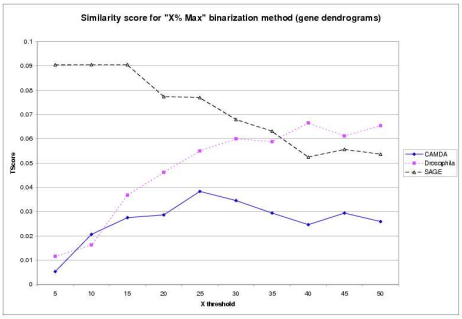
a)



b)



c)



d)

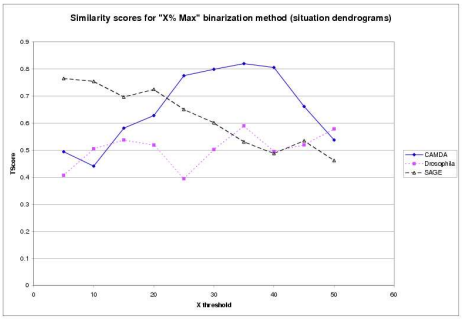


Figure 4: Similarity scores w.r.t. different thresholds for "Max - X%Max" (a and b) and "X%Max" (c and d) discretization methods

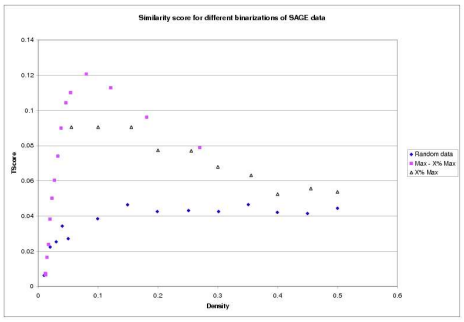
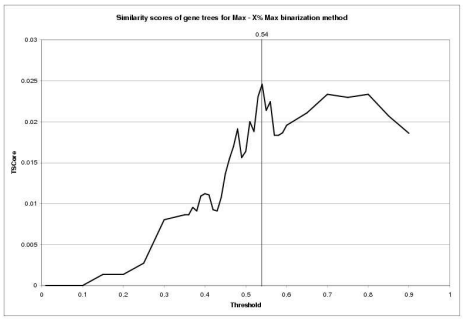


Figure 5: Similarity scores depending on density for "Max - X%Max", "X%Max" and random discretization methods applied to SAGE data



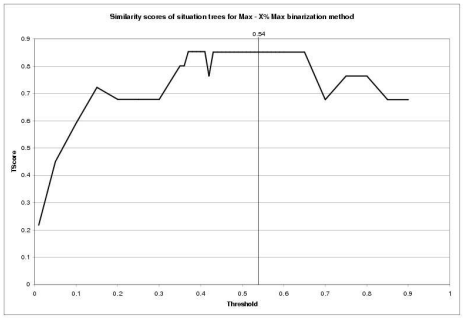


Figure 6: Similarity scores depending on various thresholds for "Max - X%Max" discretization method

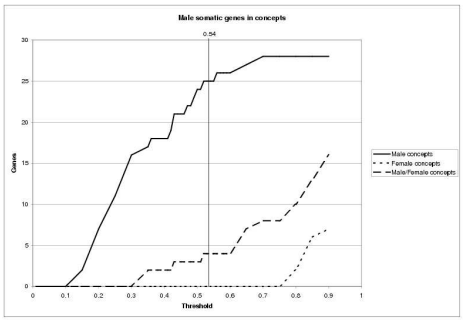


Figure 7: Number of identified male somatic genes w.r.t. discretization thresholds for different sets of concepts

6. ACKNOWLEDGEMENTS

The authors want to thank Céline Robardet, Sylvain Blachon and Olivier Gandrillon for the pre-processing of the SAGE data set, and Sophie Rome for stimulating discussions and her participation to the Drosophila application.

7. REFERENCES

[1] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A.I. Verkamo. Fast discovery of association rules. In *Advances in Knowledge Discovery and Data Mining*, pages 307–328. AAAI Press, 1996.

[2] M.N. Arbeitman, E.E. Furlong, F. Imam, E. Johnson, B.H. Null, B.S. Baker, M.A. Krasnow, M.P. Scott, R.W. Davis, and K.P. White. Gene expression during the life cycle of drosophila melanogaster. *Science*, 297:2270–2275, september 2002.

[3] C. Becquet, S. Blachon, B. Jeudy, J-F. Boulicaut, and O. Gandrillon. Strong-association-rule mining for large-scale gene-expression data analysis: a case study on human SAGE data. *Genome Biology*, 12, November 2002.

[4] S. Bergmann, J. Ihmels, and N. Barkai. Iterative signature algorithm for the analysis of large-scale gene expression data. *Physical Review*, 67, March 2003.

[5] J. Besson, C. Robardet, and J-F. Boulicaut. Constraint-based mining of formal concepts in transactional data. In *Proceedings PaKDD'04*, volume 3056 of *LNAI*, pages 615–624, Sydney, Australia, May 2004. Springer-Verlag.

[6] J. Besson, C. Robardet, J-F. Boulicaut, and S. Rome. Constraint-based concept mining and its application to microarray data analysis. *Intelligent Data Analysis journal*, 9(1), 2004. To appear.

[7] J-F. Boulicaut and A. Bykowski. Frequent closures as a concise representation for binary data mining. In *Proceedings PAKDD'00*, volume 1805 of *LNAI*, pages 62–73, Kyoto, Japan, April 2000. Springer-Verlag.

[8] Z. Bozdech, M. Llinás, B. Lee Pulliam, E.D. Wong, J. Zhu, and J.L. DeRisi. The transcriptome of the intraerythrocytic developmental cycle of plasmodium falciparum. *PLoS Biology*, 1(1):1–16, October 2003.

[9] R. Cole and R. Hariharan. An o(n log n) algorithm for the maximum agreement subtree problem for binary trees. In *Proceedings ACM-SIAM Symposium SODA'96*, pages 323–332, Atlanta, USA, January 1996.

[10] C. Creighton and S. Hanash. Mining gene expression databases for association rules. *Bioinformatics*, 19(1):79 – 86, 2003.

[11] B. DasGupta, X. He, T. Jiang, M. Li, J. Tromp, and L. Zhang. On distances between phylogenetic trees. In *Proceedings ACM-SIAM Symposium SODA'97*, volume 55, pages 427–436. 1997.

[12] B. DasGupta, X. He, T. Jiang, M. Li, J. Tromp, and L. Zhang. On computing the nearest neighbor interchange distance. In *Discrete mathematical problems with medical applications*, pages 125–143, Providence, USA, 2000. Amer. Math. Soc.

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

[14] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 95:14863–14868, December 1998.

[15] C.R. Finden and A.D. Gordon. Obtaining common pruned trees. *Journal of Classification*, 2:255–276, 1985.

[16] J. Ihmels, G. Friedlander, S. Bergmann, O. Sarig, Y. Ziv, and N. Barkai. Revealing modular organization in the yeast transcriptional network. *Nature Genetics*, 31:370–377, august 2002.

[17] A.E. Lash, C.M. Tolstoshev, L. Wagner, G.D. Schuler, R.L. Strausberg, G.J. Riggins, and S.F. Altschul. SAGEmap: A public gene expression resource. *Genome Research*, 10(7):1051–1060, July 2000.

[18] G. W. Moore, M. Goodman, and J. Barnabas. An iterative approach from the standpoint of the additive hypothesis to the dendrogram problem posed by molecular data sets. *Journal of Theoretical Biology*, 38:423–457, 1973.

[19] C. Niehrs and N. Pollet. Synexpression groups in eukaryotes. *Nature*, 402:483–487, 1999.

[20] J. Pei, J. Han, and R. Mao. CLOSET an efficient algorithm for mining frequent closed itemsets. In *Proceedings ACM SIGMOD Workshop DMKD'00*, pages 21–30, Dallas, USA, May 2000.

[21] F. Rioult, J-F. Boulicaut, B. Crémilleux, and J. Besson. Using transposition for pattern discovery from microarray data. In *Proceedings ACM SIGMOD Workshop DMKD'03*, pages 73–79, San Diego, USA, June 2003.

[22] F. Rioult, C. Robardet, S. Blachon, B. Crémilleux, O. Gandrillon, and J-F. Boulicaut. Mining concepts from large sage gene expression matrices. In *Proceedings KDID'03 co-located with ECML-PKDD 2003*, pages 107–118, Catvat-Dubrovnik, Croatia, September 2003.

[23] D. F. Robinsons. Comparison of labeled trees with valency three. *Journal of Combinatorial Theory, Series B*, 11:105–119, 1971.

[24] V.E. Velculescu, L. Zhang, B. Vogelstein, and K.W. Kinzler. Serial analysis of gene expression. *Science*, 270:484–487, 1995.

[25] R. Wille. Restructuring lattice theory: an approach based on hierarchies of concepts. In I. Rival, editor, *Ordered sets*, pages 445–470. Reidel, 1982.

[26] M. J. Zaki and C. J. Hsiao. Charm: An efficient algorithm for closed itemset mining. In *Proceedings SDM'02*, Arlington, USA, Avril 2002.