A Methodology for Biologically Relevant Pattern Discovery from Gene Expression Data

Ruggero G. Pensa[1], Jérémy Besson[1,2], and Jean-François Boulicaut[1]

[1] INSA Lyon, LIRIS CNRS FRE 2672, F-69621 Villeurbanne cedex, France
{Ruggero.Pensa,Jeremy.Besson,Jean-Francois.Boulicaut}@insa-lyon.fr
[2] UMR INRA/INSERM 1235, F-69372 Lyon cedex 08, France

Abstract. One of the most exciting scientific challenges in functional genomics concerns the discovery of biologically relevant patterns from gene expression data. For instance, it is extremely useful to provide putative synexpression groups or transcription modules to molecular biologists. We propose a methodology that has been proved useful in real cases. It is described as a prototypical KDD scenario which starts from raw expression data selection until useful patterns are delivered. Our conceptual contribution is (a) to emphasize how to take the most from recent progress in constraint-based mining of set patterns, and (b) to propose a generic approach for gene expression data enrichment. The methodology has been validated on real data sets.

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. This is true not only for genomic data but also for gene expression data. Indeed, different techniques (e.g., microarray [1]) enable to study the simultaneous expression of (tens of) thousands of genes in various biological situations. Such data can be seen as expression matrices in which the expression level of genes (the attributes or columns) are recorded in various biological situations (the objects or lines). A toy example of a gene expression matrix is in Fig. 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* [2], which, based on the guilt by association approach, are assumed to participate in a common function, or module, within the cell. Such an association between a set of co-regulated genes and the set of biological situations that gives rise to this co-regulation is called a *transcription module* and their discovery is a major goal 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., [3]). Statistical methods can be used as well (see, e.g., [4, 5]). Interesting pattern discovery techniques can be applied on boolean matrices that encode

E. Suzuki and S. Arikawa (Eds.): DS 2004, LNAI 3245, pp. 230–241, 2004.

expression properties of genes. Let $\mathcal{O}$ denote a set of biological situations and $\mathcal{P}$ denotes a set of genes. Expression properties, e.g., over-expression, 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. For deriving a boolean context from raw gene expression data, we generally apply discretization operators that, depending on the chosen expression property, compute thresholds from which it is possible to decide between whether the true or the false value must be assigned. On our toy example, a value "1" for a biological situation and a gene means that the gene is up (greater than $|t|$) or down (lower than $-|t|$) regulated in this situation. Using threshold $t = 0.3$ leads to the boolean matrix in Fig. 1b. It is then possible to look for putative synexpression groups by computing the so-called frequent itemsets from the derived boolean contexts [6]. In our boolean toy example, the genes g_3 and g_5 have the same encoded expression property in situations o_1 and o_4. This observation might lead us to derive the bi-set $(\{o_1, o_4\}, \{g_3, g_5\})$ as being potentially interesting. Notice that sets of genes that are frequently co-regulated can be post-processed into association rules [7, 8]. Stronger relationships between the components of a bi-set can increase their relevancy. For instance, $(\{o_1, o_4\}, \{g_2, g_3, g_5\})$ is one of the formal concepts (see, e.g., [9]) in the data from Fig. 1b. Informally, it means that $\{g_2, g_3, g_5\}$ is a maximal set of genes that have the recorded expression property in every situation from $\{o_1, o_4\}$ and that $\{o_1, o_4\}$ is a maximal set of situations which share the true value for every gene from $\{g_2, g_3, g_5\}$. Clearly, discovered concepts in this kind of boolean data provide putative transcription modules [10, 11].

	Genes				
Sit.	g_1	g_2	g_3	g_4	g_5
o_1	0.36	0.42	0.56	0.124	0.35
o_2	-0.24	0.01	0.28	0.02	-0.32
o_3	0.25	0.35	0.55	0.012	-0.21
o_4	0.27	0.89	-1.02	0.71	0.52
o_5	0.53	0.24	0.64	-0.6	-0.01

(a)

	Genes				
Sit.	g_1	g_2	g_3	g_4	g_5
o_1	1	1	1	0	1
o_2	0	0	0	0	1
o_3	0	1	1	0	0
o_4	0	1	1	1	1
o_5	1	0	1	1	0

(b)

Fig. 1. A gene expression matrix (a) and a derived boolean context (b).

This paper is a methodological paper. It abstracts our practice in several real-life gene expression data analysis projects in order to disseminate a promising practice within the scientific community. Our methodology covers the whole KDD process and not just the mining phase. Starting from raw gene expression data, it supports the analysis and the discovery of relevant biological information via a constraint-based bi-set mining approach from computed boolean data sets. The generic process is described within the framework of inductive databases, i.e., each step of the process can be formalized as a query on data and/or patterns that satisfy some constraints [12, 13]. It leads us to a formalization of *boolean gene expression data enrichment*. We already experimented a couple of prac-

tical instances of this approach and it has turned to be crucial for increasing the biological relevancy of the extracted patterns. An original validation of the methodology on a real data set w.r.t. a non trivial biological problem is provided.

In Section 2, the methodology is described by means of the definition of a prototypical KDD scenario. Each critical step is specified and difficulties for its execution are emphasized. In Section 3, we consider our recent contributions for supporting some of these steps. In other terms, we explain how we can execute specific instances of the given prototypical scenario by using our own data preprocessing tools (e.g., [14]), mining algorithms (e.g., [15]), and post-processing software (e.g., [16]). In Section 4, we illustrate an original application of the methodology for a real gene expression data analysis task. Section 5 concludes.

2 A Prototypical KDD Scenario

We assume that raw expression data, i.e., a function that assigns a real expression value to each couple $(o, g) \in \mathcal{O} \times \mathcal{P}$ is available and that some open problems have been selected by the molecular biologists. A typical example concerns the discovery of putative transcription modules that involve at least a given set of genes that are already known to be co-regulated in some class of biological situations, e.g., cancerous ones.

Due to the lack of space, we do not consider the typical data manipulation statements that are needed, e.g., for data normalization, data cleaning, gene and/or biological situation selection according to some background knowledge (e.g., removing the so-called housekeeping genes from consideration).

Discretization. The discretization step concerns gene expression property encoding and is obviously crucial. The simplest case concerns the computation of a boolean matrix $\mathbf{r} \subset \mathcal{O} \times \mathcal{P}$ which encode a simple expression property for each gene in each situation, e.g., over-expression. Different algorithms can be applied and parameters like thresholds have to be be chosen. For instance, [7] introduces three techniques for encoding gene over-expression:

- "Mid-Ranged". The highest and lowest expression values in a biological situation are identified for each gene and the mid-range value is defined. Then, 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 deduce 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 biological situations in which its level of expression is in X% of the highest values. These genes are assigned to value 1, 0 for the others.

The impact of the chosen algorithm and the used parameters 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. Therefore, we need to evaluate the goodness of a discretization process. Our thesis is that a good discretization might preserve some properties that can be already observed from raw data (see Section 3).

Boolean Gene Expression Data Enrichment. We can mine boolean gene expression matrices for frequent sets of genes and/or situations, association rules between genes and/or situations, formal concepts, etc. In the following, we focus on mining phases that compute concepts. When the extractions are feasible, many patterns are discovered (up to several millions) while only a few of them are interesting. It is however extremely hard to decide of the interestingness characteristics a priori. We now propose an extremely powerful approach for improving the relevancy of the extracted concepts by boolean data enrichment. It can be done a priori with some complementary information related to genes and/or situations. For instance, we can add information about the known functions of genes as it is recorded in various sources like Gene Ontology [17]. Other information can be considered like the associated transcription factors. A simple way to encode this kind of knowledge consists in adding a row to $\mathbf{r}$ for each gene property. Dually, we can add some properties to the situations vectors. For instance, if we know the class of a group of situations (e.g. cancerous vs. non cancerous cells) we can add a column to $\mathbf{r}$. We can also add boolean properties about, e.g., cell type or environmental features. Enrichment of boolean data can be performed by more or less trivial data manipulation queries from various bioinformatics databases. $\mathbf{r}' \subset \mathcal{O}' \times \mathcal{P}'$ will denote the relation of the enriched boolean context.

In Fig. 2a, we add two gene properties p_1 and p_2. A value "1" assigned to a property for some gene means that this gene has the property. For instance, p_1 could mean that the gene has a given function or is regulated by a given transcription factor. Dually, we consider two classes of situations c_1 and c_2.

	Genes						
Sit.	g_1	g_2	g_3	g_4	g_5	c_1	c_2
o_1	1	1	1	0	1	1	0
o_2	0	0	0	0	1	1	0
o_3	0	1	1	0	0	0	1
o_4	0	1	1	1	1	0	1
o_5	1	0	1	1	0	0	1
p_1	1	1	0	1	0	1	1
p_2	1	0	0	1	1	1	1

(a)

	Genes						
Sit.	g_1	g_2	g_3	g_4	g_5	c_1	c_2
o_1	1	1	1	0	1	1	0
o_2	0	0	0	0	1	1	0
o_3	0	1	1	0	0	0	1
o_4	0	1	1	1	1	0	1
o_5	1	0	1	1	0	0	1
p_1	1	1	0	1	0	1	1
p_2	1	0	0	1	1	1	1
p_3	1	0	1	1	0	1	1
p_4	1	1	1	0	1	1	1

(b)

Fig. 2. Two examples of enriched boolean microarray contexts.

A value "1" for a situation and a class means that this situation belongs to the class but this could be interpreted in terms of situation properties as well. For instance, c_1 could mean whether biological situations are cancerous ones or not. In the data in Fig. 2a, a formal concept like $(\{o_4, o_5\}, \{g_3, g_4, c_2\})$ informs us about a maximal rectangle that involves two genes in two situations that are of class c_2. This could reveal sets of genes that are co-regulated in non cancerous situations but not in cancerous ones. We discuss later how iterative enrichment is a key technique for improving the relevancy of the extracted patterns.

Constraint-Based Extraction of Formal Concepts. We consider here only formal concept discovery from eventually enriched boolean contexts. A formal concept is a maximal rectangle of "1" (1-rectangle) in the boolean matrix, and it can be represented as a bi-set of genes (eventually with situation properties) and situations (eventually with gene properties).

Definition 1. *(Concept and $\mathcal{C}_{Concept}$ constraint) A bi-set $(T, G) \in \mathcal{O} \times \mathcal{P}$ is a concept in $\mathbf{r}$ when it satisfies constraint $\mathcal{C}_{Concept}$ in $\mathbf{r}$ and $\mathcal{C}_{Concept}(T, G, \mathbf{r}) \equiv (T = \psi(G, \mathbf{r})) \wedge (G = \phi(T, \mathbf{r}))$ where ψ and ϕ are the Galois operators [9]. Let us recall that we have $\phi(T, \mathbf{r}) = \{g \in \mathcal{P} \mid \forall o \in O,\ (o, g) \in \mathbf{r}\}$ and $\psi(G, \mathbf{r}) = \{o \in \mathcal{O} \mid \forall g \in G,\ (o, g) \in \mathbf{r}\}$. (ϕ, ψ) is the Galois connection between $\mathcal{O}$ and $\mathcal{P}$.*

It is now possible to apply an algorithm for concept extraction to obtain the whole set of concepts and thus putative transcription modules. Notice that by construction, concepts are built on closed sets. It means that every algorithm that compute closed sets can be used to compute the concepts (see, e.g., [11] for the use of frequent closed set computation algorithms). Given Fig. 2a, $(\{o_1, o_4\}, \{g_2, g_3, g_5, \})$ and $(\{o_1, o_2, p_2\}, \{g_5, c_1\})$ are among the 29 concepts.

Mining every concept is not always tractable. If it is tractable, it provides potentially huge collections of patterns that have to be materialized for further post-processing guided by the molecular biologists. When the computation of every concept is not tractable, it is possible that pushing other user-defined constraints leads to tractable computations. For instance, we can extract concepts that contains a minimal or a maximal number of situations and/or genes, or that contains some particular situation and/or genes and/or their associated properties in the case of enriched contexts. Let us formalize such constraints:

Definition 2. *(Constraints on concept) A concept (T, G) is called frequent when it satisfies constraint $\mathcal{C}_t(\mathbf{r}, \sigma_1, T) \equiv |T| \geq \sigma_1$ (resp. $\mathcal{C}_g(\mathbf{r}, \sigma_2, G) \equiv |G| \geq \sigma_2$). A concept (T, G) satisfies a syntactical constraint of inclusion $\mathcal{C}_{Inclusion}(\mathbf{r}, X, G) \equiv X \subseteq G$ (resp. exclusion $\mathcal{C}_{Exclusion}(\mathbf{r}, X, G) \equiv X \not\subseteq G$). Dually, we can use $\mathcal{C}_{Inclusion}(\mathbf{r}, Y, T) \equiv Y \subseteq T$ (resp. $\mathcal{C}_{Exclusion}(\mathbf{r}, Y, T) \equiv Y \not\subseteq T$).*

It is quite useful to use these constraints in enriched contexts. For instance, we can specify that we want concepts whose situations belong to Class c_2 (say non cancerous cells) and such that the gene set contain some genes that are already known to participate to the studied regulatory way (e.g., g_1). It can be specified as the following inductive query:

$$q_1 : \mathcal{C}_{Concept}(T, G, \mathbf{r}) \wedge \mathcal{C}_{Inclusion}(\mathbf{r}, c_2, G) \wedge \mathcal{C}_{Inclusion}(\mathbf{r}, g_1, G).$$

Then, we can ask for a second collection with all the concepts (T, G) such that the class attribute c_1 is included in T:

$$q_2 : \mathcal{C}_{Concept}(T, G, \mathbf{r}) \wedge \mathcal{C}_{Inclusion}(\mathbf{r}, c_1, G).$$

Post-processing and Iteration. Concept extraction, even constraint-based mining, can produce large numbers of patterns, especially in the first iterations of the KDD process, i.e., when very few information can be used to further constrain the bi-sets to be delivered. Notice also that from a practical perspective, not all the specified constraints can be pushed into concept mining algorithms, in which case some of these constraints have to be checked in a post-processing phase.

KDD processes are clearly complex iterative processes for which every result can give rise to new ideas for more relevant constraint-based mining phases (inductive queries) or data manipulations. When a collection of patterns has been computed, it can be used for deriving new boolean properties. In particular, we have obtained two sets of patterns that can characterize two classes of genes and, dually, two classes of situations. Therefore, we can define two new class properties related to genes and their dual class properties related to situations. The boolean context $\mathbf{r}'$ can then be extended towards $\mathbf{r}'' \subset \mathcal{O}'' \times \mathcal{P}''$. Considering our running example, we can associate a new property p_3 (resp. p_4) for the genes belonging to the concepts which are returned by q_1 (resp. q_2). It leads to a new enriched boolean context given in Fig. 2b. New constraints on the classes can be used for the next mining phase. New set size constraints can be defined as well to avoid results due to noise. A new iteration will provide a new set of concepts. Each time a collection of concepts is available, we can decide either to analyze it by hand, e.g., studying each gene separately, or looking for new boolean data enrichment and new constraints for a new iteration.

In any case, at the end of the process, we have a set of putative interesting genes and a set of putative interesting situations. When considering our running example of putative transcription module discovery from an initial gene set (here $\{g_1\}$, called hereafter the seed set), it is interesting to stop iterations when the sets of genes include (almost) all the genes from the seed set and when the total number of genes which are not in the seed gene set can be studied in a reasonable time by means of biological experiments.

In our toy example (Fig. 2b), let us enforce the absence in T of p_4, i.e., those genes that are contained in the concepts concerning the situation belonging to Class c_1. The result is a single concept $(\{o_4, o_5, p_1, p_2, p_3\}, \{g_4, c_2\})$. The gene g_4 is co-regulated in two situations associated to Class c_2 but it is not in the seed set of genes known to be involved in the studied transcription process. It means that g_4 is a putative interesting gene that can be studied further to verify if its function is related to the studied biological problem. Notice also that genes to which we can associate new functions appear as interesting candidates for performing new iterations and take advantage of larger seed gene sets.

3 About Scenario Practical Executions

The prototypical scenario we have presented in the previous section can be executed in different ways, depending on available algorithms and tools. In this section, we explain how we can execute it on practical cases by taking the most from some recent advances on constraint-based set mining and gene expression data analysis. We do not provide here new results but evidence that such a prototypical scenario can be used by practitioners.

We have explained that discretization of raw gene expression data is a crucial phase. We clearly need a method to evaluate different boolean encoding (different techniques and/or various parameters) of the same raw data and thus a framework to support user decision about the discretization from which the mining process can start. Let E denote a gene expression matrix. Let $\{Bin_i,\ i = 1..b\}$ denote a set of different discretization operators and $\{\mathbf{r}_i,\ i = 1..b\}$ a set of boolean contexts obtained by applying these operators, i.e. $\forall i = 1..b,\ \mathbf{r}_i = Bin_i(E)$. Let $S : \mathbb{R}^{n,m} \longmapsto \mathbb{R}$ denote an evaluation function that measure the quality of the discretization of a gene expression matrix. We say that a boolean context $\mathbf{r}_i$ is more valid than another context $\mathbf{r}_j$ w.r.t the S measure if $S(\mathbf{r}_i) > S(\mathbf{r}_j)$. In [14], we recently studied an original method for such an evaluation. We suggest to compare the similarity between the dendrogram generated by a hierarchical clustering algorithm (e.g., [3]) applied to the raw expression data and the dendrograms generated by the same algorithm applied to each derived boolean matrix. Given a gene expression matrix E and two derived boolean contexts $\mathbf{r}_i$ and $\mathbf{r}_j$ for two distinct discretizations, we can choose the discretization that leads to the dendrogram which is the most similar to the one built on E. The idea is that a discretization that preserves the expression profile similarities is considered more relevant. In [14], a simple measure of similarity between dendrograms has been studied and experimentally validated on various gene expression data sets. It is used in Section 4 for our original application to the drosophila data set.

A second major problem concerns constraint-based mining of concepts. In our applications to gene expression data, we can get rather dense boolean contexts that are hard to process if further user-defined constraints are not only specified but also pushed deeply into the extraction algorithms. Using user-defined constraints enables to produce putative interesting concepts before any post-processing phase. Indeed, concept discovery techniques can provide huge collection of patterns and supporting post-processing on such collections is hard or even impossible. It motivates the a priori use of constraints on both $2^{\mathcal{O}}$ and $2^{\mathcal{P}}$. We saw typical examples of constraints on the size of T and G. The recent algorithm D-Miner introduced in [15] computes concepts under monotonic constraints and can be used in dense boolean data sets when the previous algorithms (concept lattice discovery algorithms or frequent closed set computation algorithms) generally fail.

Another important problem concerns the postprocessing of concept collections. As the number of concepts to analyze starts to be huge, we need efficient exploration techniques to support the subjective search of interesting concepts. In [16], we propose an "Eisen-like" visualization technique, that allows to group

similar concepts by means of a hierarchial clustering algorithm using an original definition of distance. Thanks to this approach we can reduce the effects of concept multiplication due to noise in data and support the post-processing of thousands of concepts with a graphical approach.

4 An Application

We have used our methodology on a real gene expression data analysis problem for which it was possible to evaluate the relevancy of the results thanks to the available documentation [18]. It concerns the gene expression of the Drosophila melanogaster during its life cycle. The paper considers the expression level of 4 028 genes for 66 sequential time periods from the embryonic state till the adulthood. The related data set is available on line[1]. The total number of samples is 81 since the gene expression during the adult state is measured for male and female individuals and the expression level of more genes is available. For our experiment we have selected a set of 4 137 genes and 20 time periods concerning adult individuals. This set is composed of 8 male adult samples, 8 female adult samples, 2 male and 2 female tudor samples. We selected 4 of the 4 137 genes which are known to be "male somatic genes", i.e., a class of genes that characterize the male individuals (Genes CG2858, CG2267, CG17843, and CG2082). Let us denote this set by $KG = \{kg_1, kg_2, kg_3, kg_4\}$. We want to discover new knowledge about this group, i.e., adding other genes to the seed set KG by applying our methodology. Notice that the genes from KG have been selected randomly among the known male somatic genes. In this experiment, our goal is to demonstrate that, given a small gene set, we are able to increase our knowledge with two simple iterations of the method. In other terms, we do not claim that we want to find all "male somatic genes" but we want to rediscover part of this knowledge thanks to the available biological results from [18].

Preprocessing. We marked a group of 351 genes as being always under-expressed (in all the 20 situations). Another group of 353 genes has been marked since it is over-expressed in more than 10 biological situations. We performed a projection on non-marked genes and we obtain at the end a $20 \times 3\,433$ expression matrix denoted E. To discretize E, we choose the "Max - X% Max" method:

$$\overline{b}_i = Bin(\overline{e}_i)$$

where, for each gene vector $\overline{b}_i$,

$$b_{ij} = \begin{cases} 1, & if\ (1-X)\max_j(e_{ij}) < e_{ij} < \max_j(e_{ij}) \\ 0, & otherwise \end{cases} \quad (1)$$

where $j = 1..20$ and $\overline{e}_i \in E$.

[1] http://genome-www5.stanford.edu/

Different values can be chosen for X and we applied the method described in [14] when considering X values between 0.01 and 0.9. The result of this analysis for gene dendrograms are summarized in Fig. 3. The best value for our similarity score is when $X = 0.54$. Consequently this is the threshold we used in order to derive the boolean gene expression data.

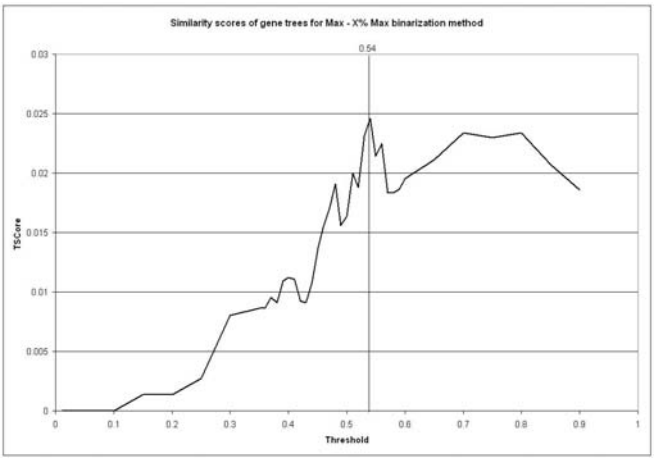

Fig. 3. Gene similarity scores for "Max - X%Max" on E when X varies.

Then, we associated two sex class properties to situations by adding two columns to the boolean matrix. The first property c_M is set to "1" for all male individuals while the second one c_F gets the "1" value for all female individuals. This enriched boolean context $\mathbf{r}$ has been the starting point for our concept mining process.

Extraction. We performed the sequence of operations described in Section 2. First we tried to get the whole collection of concepts:

$$GT = \left\{(T, G) \in 2^{\mathcal{O}} \times 2^{\mathcal{P}} \mid \mathcal{C}_{Concept}(T, G, \mathbf{r})\right\}.$$

It has been feasible in this context and $|GT| = 14\,884$ (excluding those containing only situation and gene properties).

The following step has been to further constrain the solution set. We decided to focus on the collection of concepts that concern male individuals and that contains at least one gene from KG. The associated constraint $\mathcal{C}_M$ is:

$$\mathcal{C}_M(T, G, \mathbf{r}) \equiv \mathcal{C}_{Concept}(T, G, \mathbf{r}) \wedge \mathcal{C}_{Inclusion}(\mathbf{r}, c_M, G) \wedge \mathcal{C}_a(\mathbf{r}, KG, G)$$

where $\mathcal{C}_a(\mathbf{r}, KG, G)$ is a "at-least-one" constraint, and it is satisfied if $\exists kg \in KG \mid \mathcal{C}_{Inclusion}(\mathbf{r}, kg, G)$.

Let GT_M denote this set, D-MINER can compute it and $|GT_M| = 440$.

Then, we have been looking for concepts that concern only female individuals. Furthermore, to tackle noisy data in the boolean context, we specified also a constraint of minimal size for gene sets ($\sigma_g = 20$) and situation sets $\sigma_t = 5$:

$$\mathcal{C}_F(T, G, \mathbf{r}) \equiv \mathcal{C}_{Concept}(T, G, \mathbf{r}) \wedge \mathcal{C}_{Inclusion}(\mathbf{r}, c_F, G) \wedge \mathcal{C}_g(\mathbf{r}, \sigma_g, G) \wedge \mathcal{C}_t(\mathbf{r}, \sigma_t, T)$$

where $\mathcal{C}_t$ and $\mathcal{C}_g$ are constraints on minimal size that are efficiently pushed into the computation by the D-MINER algorithm.

The result denoted by GT_F is a collection of $|GT_F| = 515$ concepts.

Boolean Context Enrichment. To reduce the size of concepts and thus the number of genes to analyze, we have defined a new class for genes that are not in the GT_F set. The genes contained in such a set do not characterize male individuals and can be excluded from consideration at the next extraction task. We added a new row called r_M that is true (1) for all genes not contained in $\bigcup G \mid (T, G) \in GT_F$. 713 genes were contained in such a set and thus the r_M property is true for 2 720 genes. Let $\mathbf{r}'$ denote this new boolean context.

Second Extraction. We then processed the new boolean context using a new constraint $\mathcal{C}_{MG}$ using the r_M property:

$$\mathcal{C}_{MG}(T, G; \mathbf{r}') \equiv \mathcal{C}_{Concept}(T, G, \mathbf{r}') \wedge \mathcal{C}_{Inclusion}(\mathbf{r}', r_M, T) \wedge \mathcal{C}_a(\mathbf{r}', KG, G)$$

where $\mathcal{C}_a$ is defined as before. We obtained a reduced set GT'_M of 295 concepts.

We decided to further reduce the size of the collection of concepts by means of a minimal size constraint on situations. We wanted to keep only concepts that contains at least 6 situations, i.e., one more than 1/2 of the total number of male individuals:

$$\mathcal{C}'_{MG}(T, G, \mathbf{r}') \equiv \mathcal{C}_{MG}(T, G; \mathbf{r}') \wedge \mathcal{C}_t(\mathbf{r}', 6, T)$$

It has given a set GT''_{MG} of 83 concepts. This has been considered as a relatively small set for subjective exploration.

Final Post-processing. We finally performed some post-processing on GT''_{MG}. We selected the genes contained in the concepts of GT''_{MG} when they were appearing in at least $0.5 \cdot |GT''_{MG}|$ concepts, i.e. genes that were fairly represented. As result, we got a quite small collection of 11 genes. None of the genes from our seed set KG occurs in this collection. On the other hand, three of these genes are already described in [18] as belonging to the "male somatic gene" class. This result has been obtained by analyzing in detail only 11 genes among the 3 433 genes of the expression matrix. Another important result is the presence of a very interesting concept in the last set of concepts we built (Tab. 1). It concerns 8 male individuals and 14 genes, 5 of them being presented in [18] as "male somatic genes". Among these, only one was present in our seed set KG.

Table 1. A concept concerning 14 genes (5 somatic) and 8 male individuals. Each cell in the table contains the original expression value. Only somatic genes are represented.

	Genes						
Situations	...	CG17843	CG6761	CG10096	CG18284	CG7157	...
A03M	...	1.789	2.199	2.659	4.159	3.749	...
A05M	...	2.628	2.728	4.168	4.788	3.858	...
A10M	...	2.29	1.83	2.89	3.53	3.86	...
A15M	...	2.048	1.588	4.728	4.998	4.628	...
A20M	...	2.587	2.127	3.377	3.597	4.967	...
A25M	...	2.336	1.886	3.636	4.516	3.716	...
A30M	...	2.568	1.958	3.048	3.858	3.808	...
AT05M	...	3.505	1.925	5.125	5.535	5.385	...

5 Conclusion

We have designed a new data mining methodology to analyze gene expression data thanks to constraint-based mining of formal concepts. We have described a prototypical KDD scenario that has been proved useful in several real-life gene expression data analysis problems. Boolean data enrichment is a very powerful technique for supporting the iterative search of relevant patterns w.r.t. a given analysis task. It is indeed related to the many contribution to feature construction techniques. We are currently applying the whole method on the data from [19] to improve our understanding of insulino-regulation.

Acknowledgements

The authors want to thank Dr. Sophie Rome for stimulating discussions and her help for preparing the drosophila data set. Jérémy Besson is funded by INRA.

References

1. DeRisi, J., Iyer, V., Brown, P.: Exploring the metabolic and genetic control of gene expression on a genomic scale. Science **278** (1997) 680–686
2. Niehrs, C., Pollet, N.: Synexpression groups in eukaryotes. Nature **402** (1999) 483–487
3. Eisen, M., Spellman, P., Brown, P., Botstein, D.: Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA **95** (1998) 14863–14868
4. Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y., Barkai, N.: Revealing modular organization in the yeast transcriptional network. Nature Genetics **31** (2002) 370–377
5. Bergmann, S., Ihmels, J., Barkai, N.: Iterative signature algorithm for the analysis of large-scale gene expression data. Physical Review **67** (2003)
6. Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., Verkamo, A.: Fast discovery of association rules. In: Advances in Knowledge Discovery and Data Mining, AAAI Press (1996) 307–328

7. Becquet, C., Blachon, S., Jeudy, B., Boulicaut, J.F., Gandrillon, O.: Strong-association-rule mining for large-scale gene-expression data analysis: a case study on human sage data. Genome Biology **12** (2002)
8. Creighton, C., Hanash, S.: Mining gene expression databases for association rules. Bioinformatics **19** (2003) 79 – 86
9. Wille, R.: Restructuring lattice theory: an approach based on hierarchies of concepts. In Rival, I., ed.: Ordered sets. Reidel (1982) 445–470
10. Rioult, F., Boulicaut, J.F., Crémilleux, B., Besson, J.: Using transposition for pattern discovery from microarray data. In: Proceedings ACM SIGMOD Workshop DMKD'03, San Diego (USA) (2003) 73–79
11. Rioult, F., Robardet, C., Blachon, S., Crémilleux, B., Gandrillon, O., Boulicaut, J.F.: Mining concepts from large sage gene expression matrices. In: Proceedings KDID'03 co-located with ECML-PKDD 2003, Catvat-Dubrovnik (Croatia) (2003) 107–118
12. Boulicaut, J.F., Klemettinen, M., Mannila, H.: Modeling KDD processes within the inductive database framework. In: Proceedings DaWaK'99. Volume 1676 of LNCS., Florence, I, Springer-Verlag (1999) 293–302
13. De Raedt, L.: A perspective on inductive databases. SIGKDD Explorations **4** (2003) 69–77
14. Pensa, R.G., Leschi, C., Besson, J., Boulicaut, J.F.: Assessment of discretization techniques for relevant pattern discovery from gene expression data. In: Proceedings BIOKDD'04 co-located with SIGKDD'04, Seattle, USA (2004) To appear.
15. Besson, J., Robardet, C., Boulicaut, J.F.: Constraint-based mining of formal concepts in transactional data. In: Proceedings PAKDD'04. Volume 3056 of LNAI., Sydney (Australia), Springer-Verlag (2004) 615–624
16. Robardet, C., Pensa, R., Besson, J., Boulicaut, J.F.: Using classification and visualization on pattern databases for gene expression data analysis. In: Proceedings PaRMa'04 co-located with EDBT 2004. Volume 96 of CEUR Proceedings., Heraclion-Crete, Greece (2004) 107–118
17. Ashburnerand, M., Ball, C., Blake, J., Botstein, D., et al.: Gene ontology: tool for the unification of biology. the gene ontology consortium. Nature Genetics **25** (2000) 25–29
18. Arbeitman, M., Furlong, E., Imam, F., Johnson, E., Null, B., Baker, B., Krasnow, M., Scott, M., Davis, R., White, K.: Gene expression during the life cycle of drosophila melanogaster. Science **297** (2002) 2270–2275
19. Rome, S., Clément, K., Rabasa-Lhoret, R., Loizon, E., Poitou, C., Barsh, G.S., Riou, J.P., Laville, M., Vidal, H.: Microarray profiling of human skeletal muscle reveals that insulin regulates 800 genes during an hyperinsulinemic clamp. Journal of Biological Chemistry (2003) 278(20):18063-8.