In Silico Biology 7, 0033 (2007); ©2007, Bioinformation Systems e.V.

Clustering formal concepts to discover biologically relevant knowledge from gene expression data

Sylvain Blachon[1,2], Ruggero G. Pensa[2], Jérémy Besson[2], Céline Robardet[2], Jean-François Boulicaut[2] and Olivier Gandrillon[1]*

[1] Equipe "Bases Moléculaires de l'Autorenouvellement et de ses Altérations"
Université de Lyon, Lyon, F-69003, France; Université Lyon 1, Villeurbanne, F-69622, France; CNRS UMR5534, CGMC, Villeurba
F-69622, France
Telephone: +33-4 72 44 81 90 - Fax: +33-4 72 43 26 85
Email: Gandrillon@cgmc.univ-lyon1.fr; sylvain.blachon@gmail.com

[2] INSA-Lyon, LIRIS CNRS UMR5205
Bâtiment Blaise Pascal
F-69621 Villeurbanne cedex, France
Telephone: +33-4 72 43 89 05 - Fax: +33-4 72 43 87 13
Email: Ruggero.Pensa@insa-lyon.fr; Jeremy.Besson@insa-lyon.fr; Celine.Robardet@insa-lyon.fr; Jean-Francois.Boulicaut}@insa-lyon.fr

* Corresponding author

Edited by E. Wingender; received October 27, 2006; revised February 01 and May 29, 2007; accepted June 16, 2007; published July 16, 2007

Abstract

The production of high-throughput gene expression data has generated a crucial need for bioinformatics tools to generate biologically interesting hypotheses. Whereas many tools are available for extracting global patterns, less attention has been focused on local pattern discovery. We propose here an original way to discover knowledge from gene expression data by means of the so-called formal concepts which hold in derived Boolean gene expression datasets. We first encoded the over-expression properties of genes in human cells using human SAGE data. It has given rise to a Boolean matrix from which we extracted the complete collection of formal concepts, i.e., all the largest sets of over-expressed genes associated to a largest set of biological situations in which their over-expression is observed. Complete collections of suc patterns tend to be huge. Since their interpretation is a time-consuming task, we propose a new method to rapidly visualize clusters of formal concepts. This designates a reasonable number Quasi-Synexpression-Groups (QSGs) for further analysis. The interest of our approach is illustrated using human SAGE data and interpreting one of the extracted QSGs. The assessment of its biological relevancy leads to the formulation of both previously proposed and new biological hypotheses.

Keywords: transcriptome, SAGE, pattern discovery, formal concepts, closed sets, clustering

Introduction

Producing massive amounts of gene expression data is an everyday task for biologists involved in OMI programs. The critical bottleneck is now to derive knowledge from such huge datasets.

An important area of research deals with the identification of co-expressed gene sets known as

synexpression groups [Lee *et al.*, 2004] or transcription modules [Stuart *et al.*, 2003]. It is motivated by a consensual hypothesis in molecular biology which states that co-expressed genes interact together to perform the same biological function [Lee *et al.*, 2002].

Clustering or co-clustering are popular techniques for the identification of candidate synexpression groups from gene expression matrices. A partition is a typical global pattern that defines a similarity structure which is observed in the whole data set. As a result, partitions can ignore locally strong associations between th expression profiles of subsets of genes within subsets of samples. Also, most of the clustering algorithms do not allow clusters to overlap. This is obviously a problem given that many genes have different functions and thus might be involved in several synexpression groups.

To overcome these limitations, several local pattern discovery techniques have been designed. First, researchers have been considering the so-called biclustering methods which enable to identify subsets of genes sharing compatible expression patterns across subsets of biological samples [see Madeira and Oliveira, 2004; Prelic *et al.*, 2006, for surveys]. These patterns are local (i.e., they hold in a subset of the data) and the discovered bi-clusters can overlap. However, the mining algorithms are generally heuristic ones: some a priori interesting patterns are computed but complete, say exhaustive, methods can not be designed for most of the numerical data analysis tasks.

On another hand, many researchers have been considering complete algorithms for local set pattern mining from Boolean data. These methods are complete in the sense that all the patterns which satisfy a giv user-defined constraint can be extracted. This has been applied with some success to gene expression analysis. For instance, a few authors have investigated the use of association rule discovery (ARD) to generate biologically relevant hypotheses from gene expression Boolean matrices [Becquet *et al.*, 2002; Creighton and Hanash 2003; Li *et al.*, 2003; Elati *et al.*, 2005; Georgii *et al.*, 2005]. Boolean gene expression properties, such as over-expressions, are encoded for genes in given biological samples. Then, frequent and valid association rules, i.e. local patterns that can inform investigators about potential associations between sets of genes or sets of samples, can be generated. A major problem with the ARD technique is that hu collections of rules can be generated in even quite small datasets because they are all satisfying the user-defined constraints (e.g., the conjunction of a minimal frequency value and a minimal confidence value). The redundancy in classical association rule collections is now well understood and one solution is to consider some non redundant subsets, e.g., exploiting the properties of closed sets. For instance, the so-called formal concepts are built on closed sets. Each formal concept is a maximal set of genes satisfying the encoded property (e.g., over-expression) associated to the maximal set of biological samples (see Fig. 1) in which this expression property occurs.

	T1	T2	T3	T4	T5	T6	T7	T8	T9
S1	1	1	1	0	0	0	0	0	0
S2	1	1	1	0	0	0	0	0	0
S3	1	1	1	0	0	0	0	0	0
S4	0	0	0	1	1	1	0	1	0
S5	0	0	0	0	0	0	1	0	0
S6	0	0	0	1	1	1	0	1	0

Click on the thumbnail to enlarge the picture

Figure 1: Definition of a formal concept. More generally, a formal concept is a maximal set of attributes (columns or tags) that are verified (value = 1) by a maximal set of objects (lines or SAGE libraries). In this toy Boolean gene expression matrix, the over-expression of a gene was appropriately encoded: when a tag T_i is found over-expressed in the SAGE library S_j, its value is 1. In this context, 3 formal concepts can be extracted: {($S1$,$S2$,$S3$);($T1$,$T2$,$T3$)} (in black), {($S4$,$S6$);($T4$,$T5$,$T6$,$T8$)} (light grey), {($S5$);($T7$)} (dark grey). We can notice that a formal concept can also be considered as a bicluster with a constant value (the true value) which also satisfy two maximality constraints.

Thanks to the huge effort for Boolean data mining the last 5 years, efficient algorithms are available for mining formal concepts from large matrices. Complete extractions are tractable even for the tens of thousands attributes that might denote genes in Boolean gene expression datasets. Indeed, the number of formal concepts is exponential in the smallest dimension of the matrix, which can be rather small in the context of gene expression data where a few tens of samples are common. While feasible, the number of extracted

formal concepts in real datasets can be huge, up to millions.

To limit the number of such local patterns, while enforcing their relevancy for a given task, a first approach is to exploit constraint-based mining. In this context, the analyst specifies constraints that have to be satisfied by the patterns, e.g., the extracted formal concepts must be "large enough" *and* contain a given subset of genes *or* a given subset of samples, etc. Pushing such constraints into mining algorithms has been studied seriously over the last few years and it enables to speed up the computation while reducing the size of the extra collections. In turn, from the end-user point of view, this increases the a priori relevance and this reduces number of patterns to be interpreted. We [Klema *et al.*, 2006] and others [Carmona-Saez *et al.*, 2006] recently described the use of external constraints (stemming from the literature or from the GO ontology) to reduce and annotate the extracted patterns. Nevertheless the computed collections still remain huge, up to ten thousands.

It is now well understood that the number of formal concepts increases dramatically in noisy data sets (i.e., a 0 value has been recorded instead of a 1 value or *vice versa*). This is particularly true with Boolean gene expression data which are intrinsically noisy. Therefore, an idea is to extend formal concepts towar fault-tolerance (see, e.g., Besson *et al.*, 2006a; Besson *et al.*, 2006b). Such fault-tolerant patterns (FTPs) can be viewed as formal concepts in which a limited number of exceptions are tolerated (e.g., one tolerates that a few genes are not over-expressed in a small subset of situations, in the final synexpression group). Whethe FTPs might generate interesting hypotheses in real-life situations remains unclear: the mining task computationally much harder than for formal concept discovery. Indeed in large and/or dense data sets exhaustive generation of FTPs becomes intractable [Besson *et al.*, 2006b]. When considering biclustering techniques [Madeira and Oliveira 2004], it is also possible to look for some kind of fault-tolerant patterns like biclusters that contain almost constant values. However, let us recall that in that case, only heuristic techniques are available. Notice also that such a computation that seek bi-sets based on similar expressio levels, disregarding specific gene expression properties like over-expression, has to be considered as a different type of analysis task.

In this paper, we present a pragmatic solution to support the interpretation of collections formal concepts which hold in Boolean gene expression datasets. We present an original post-processing technique that groups together similar formal concepts. Each cluster of sufficiently similar formal concepts can be represented by a bi-set (i.e. a set of genes associated to a set of samples) and these bi-sets denote strong local associations between sets of genes and sets of samples. These representative bi-sets a Quasi-Synexpression Groups (QSGs): they tend to capture maximal associations that tolerate a few false values as exceptions.

Most of the algorithms used here have been described in much more generic settings separately (Rioult *et al.*, 2003; Robardet *et al.*, 2004 for tools usable in the application BioMiner). This is the first paper in which they are combined, and used non-trivial application with respect to a real-life problem, namely human SAGE data analysis. We sketch complete KDD (Knowledge Discovery from Databases) process on human SAGE data, i.e., data pre-processing, formal concept extraction, formal concept selection, clustering of formal concepts and biological interpretation. For the sake of brevity, only one resulting QSG is discussed.

Methods

From a computational perspective, three major steps are involved in our KDD process:

1. Pre-processing the data;
2. Computing formal concepts;
3. Post-processing formal concepts.

Data pre-processing

The human SAGE libraries were downloaded in December 2002 from the NCBI web site (NCBI). In order eliminate putative sequencing errors, all tags appearing only once in one library [Keime *et al.*, 2007] were discarded. The libraries were normalized by dividing the frequency of each tag by the total number of t composing that library. This frequency was multiplied by 300,000 (the estimated number of mRNAs molecules per cell [Velculescu *et al.*, 1999]) to give an integer value which means "the number of copies of that mRNA per cell". The tags were identified by Céline Keime in February 2005 using Identitag [Keime *et al.*, 2004] and the updated RefSeq database dated 2/2/2005. A database containing all the information on the tags, the libraries, and level of expression was created.

Applying the same treatments as those described in [Becquet *et al.*, 2002], a gene expression matrix displaying the expression level of 27,679 genes (or tags) in 90 biological situations (or libraries) was generated.

Next, a feature selection process was performed: genes were filtered by means of an ANOVA analysis with the BioConductor Package of the R software (see http://www.bioconductor.org/repository/devel/vignette/howtogenefilter.pdf). This filter selects the tags that best discriminate the samples with respect to the type of organ of origin of the cells.

Among the 27,679 tags, 5327 were found to better discriminate the seven groups of biological situations. This allowed the construction of a matrix displaying the expression level of 5327 genes in 90 biological situations on which all further experiments have been performed.

To apply efficient local set pattern mining techniques on Boolean data, we must identify a specific gene expression property (in principle, several properties per gene could be encoded, e.g. over-expre under-expression). In this study, we decided to focus on over-expression. Thus if a gene is over-expressed in a situation then there will be a true value (1) in the corresponding Boolean matrix cell, otherwise the value is 0, i.e., false. Several ways exist for identifying gene over-expression [Becquet *et al.*, 2002]. The Middle-Range discretization technique was used: for this, the highest and lowest expression values were identified for each tag and the mid-range value was defined as being equidistant from these two numbers (their arithmetic mean). Then, all expression values below or equal to the mid-range threshold were set to 0, and all values strictly above the mid-range were set to 1. Mainly two reasons guided this choice: 1. since the level of discretization does not depend upon the value of one given parameter, it is more robust, and easier to use; and 2. this discretization method has been validated through an automated evaluation method [Pensa *et al.*, 2004].

Extraction of formal concepts

A formal concept in a Boolean matrix is a maximal set of columns associated to a maximal set of rows such that there are only true values between these lines and columns (Fig. 1). Intuitively, it is a maximal rectangle of true values modulo arbitrary permutations of rows and columns, i.e., combinatorial rectangles. In our Boolean gene expression data analysis context, a formal concept is a largest set of over-expressed ge associated to a largest set of biological situations in which their over-expression is observed. Notice that exception (false value) is tolerated here. Let us now formalize this pattern domain.

Assume G denotes the set of genes (tags in SAGE data, $|G|$ = 5327 in our concrete instance) and S denotes the set of samples (libraries for SAGE, $|S|$ = 90 in our concrete instance). The over-expression property can be encoded into a binary relation $r \subseteq G \times S$. $(g_i,s_j) \in r$ denotes that gene i is over-expressed in sample j. A formal concept in r is a bi-set $(X,T) \in 2^G \times 2^S$ such that $T = \psi(X,r)$ and $X = \varphi(T,r)$. (ψ,φ) is the so-called Galois connection and is defined as follows: $\varphi(T,r) = \{g \in G \mid \forall\ s \in T, (g,s) \in r\}$ and $\psi(X,r) = \{s \in S \mid \forall\ g \in X, (g,s) \in r\}$. By construction, when (X,T) is a formal concept, X and T are closed sets, i.e., sets which are equal to their closures given the two dual closure operators $\psi o \varphi$ and $\varphi o \psi$. Interestingly, it is possible to compute every formal concept by computing every closed set on the smallest dimension, say the samples, and then associate the corresponding closed set in the other dimension, say the genes, by using the Galois connection. Each formal concept can be computed this way [Rioult *et al.*, 2003].

From the computational point of view, computing formal concepts from small data sets is often tractable ev though the problem is exponential in the smallest dimension of the matrix. The problem is much harder as soon as large contexts (the smallest dimension is more than a few tens) and a high density (i.e., a high number of true values) are mined. The solution can come from the intensive research on set pattern mining. For instance, extremely efficient algorithms have been designed for computing the so-called σ – frequent closed sets, i. e., every closed set T of genes such that $|\varphi(T,r)| > \sigma$ or every closed set such that $|\psi(G,r)| > \sigma$, see among others [Pasquier *et al.*, 1999; Zaki and Hsiao, 2002]. A survey on such algorithms was made available thanks to the FIMI initiative [Goethals and Zaki 2004]. One should also note that the algorithms for computing the so-called concept lattices are relevant for mining frequent closed sets (see Kuznetsov and Obiedkov, 2002, for a review).

These techniques can be used for computing part of the collection of formal concepts, e. g., formal concept whose one of the set component has a minimal size. However, when looking for complete collections and when the Boolean data turns to be dense and/or large, these approaches fail. In many application domains, it is clear that large enough patterns are useful but this imposes to consider not one dimension only. Intuitively, an area constraint which would enforce that $|X|*|T|$ is over a user-defined threshold will avoid to capture patterns that are not significant. Tackling such constraints is difficult because the specialization relations samples (respectively on genes) are in opposite directions (subset *vs.* superset). Considering constraints on both dimensions and pushing them efficiently during the data mining phase is thus challenging but extreme useful. This has been studied for constraint-based data mining of formal concepts in [Besson *et al.*, 2005] where the D-Miner algorithm is described. This algorithm not only enables to compute every formal concept in rather dense Boolean matrices but also it exploits other user-defined constraints for pruning the search space (e.g., pushing minimal size constraints on both sets that constitute the patterns). The input of D-Miner is description of a Boolean matrix and the optional specification of constraints on the desired formal concepts. The D-Miner software is freely available as part of the BioMiner software package (BioMiner, see availab section).

Post processing collections of formal concepts

Let us assume now that well-specified collections of formal concepts have been extracted and stored. In other terms, complete collections of patterns satisfying a given constraint were stored into pattern databases. Now, the challenge is to support the subjective search for relevant patterns according to a specific analysis tas biological question. Basically, two types of tools were implemented. First, querying tools over collections of formal concepts were designed, which enables the operator to ask questions with respect to the collection formal concepts. Next, a technique for grouping similar formal concepts was designed.

These tools allowed the identification of formal concepts that appear to be interesting with respect to three criteria: homogeneity in the library description, presence of a keyword in the gene description, and finally size of the set components. Clearly, these selection criteria can be arbitrarily combined.

In order to cluster formal concepts, the possibility of performing a classical agglomerative hierarchical clustering operation on formal concepts instead of genes or samples was considered. The main difficulty is define a similarity measure, which can take into account both genes and biological samples. The intuition is that the overlap between two formal concepts with respect to common genes and situations can be measured and used as a distance.

Definition. Let c_i and c_j be two formal concepts containing respectively X_i and X_j as sets of over-expressed tags and T_i and T_j as sets of libraries, the distance between c_i and c_j is defined as follows:

$$d_{ij} = \frac{1}{2}\frac{|X_i \Delta X_j|}{|X_i \cup X_j|} + \frac{1}{2}\frac{|T_i \Delta T_j|}{|T_i \cup T_j|}$$

where Δ is the symmetrical set difference between S_i and S_j: $S_i \Delta S_j = |(S_i \cup S_j) \setminus (S_i \cap S_j)|$.

Now, distances have to be computed between two clusters of formal concepts. Since there is usually much

more formal concepts than tags or libraries, we decided to use pseudo concepts, as suggested in [Robardet *et al.*, 2004]. The idea is to associate to each cluster a pseudo concept summing up the main characteristics of the formal concepts it contains. A pseudo concept is composed of two fuzzy sets: one for tags and the other one for libraries. A fuzzy set is a set whose element membership is quantified. For example, for a fuzzy set of tags, a parameter α_i (a real number between 0 and 1) is used to measure the degree of membership of the i^{th} tag. When α_i equals 0, it means the i^{th} tag is never present in the fuzzy set, and thus in the pseudo concept. Symmetrically, when α_i equals 1, it means the i^{th} tag is always present in the fuzzy set, and thus in the pseudo concept. Biologically, α_i evaluates the probability that the i^{th} tag is over-expressed in the situations *contained* in the pseudo concept. The same principle can be used for a fuzzy set of libraries.

Definition. Let $S = \{s_1, s_2, ..., s_R\}$ be the set of R libraries and $G = \{a_1, a_2, ..., a_M\}$ be the set of M tags, a pseudo concept is (X', T', N) such that $X' = \{(a_1,n_1), ..., (a_i,n_i), ..., (a_M,n_M)\}$, m>T' = $\{(s_1,m_1),..., (s_j,m_j), ..., (s_R,m_R)\}$ and N is the number of formal concepts represented by the pseudo concept where for all $i \in \{1,...,M\}$ and $j \in \{1,...,R\}$ then $a_i \in [0;1]$ and $s_j \in [0;1]$.

The pseudo concept (X',T',N) of a formal concept (X,T) is defined as follows:

$$\begin{cases} X' = \{(g,\beta) | g \in G, \beta = 1 \text{ if } g \in X, \beta = 0 \text{ otherwise}\} \\ T' = \{(s,\alpha) | s \in S, \alpha = 1 \text{ if } s \in T, \alpha = 0 \text{ otherwise}\} \\ N = 1 \end{cases}$$

The pseudo concept (X', T', N) of two pseudo concepts (X'1, T'1, N1) and (X'2, T'2, N2) is defined as follows:

$$\begin{cases} X' = \left\{ \left(g, \frac{N_1\beta_1 + N_2\beta_2}{N_1 + N_2} \right) \middle| \; g \in G, (g,\beta_1) \in X'_1 \text{ and } (g,\beta_2) \in X'_2 \right\} \\ T' = \left\{ \left(s, \frac{N_1\alpha_1 + N_2\alpha_2}{N_1 + N_2} \right) \middle| \; s \in S, (s,\alpha_1) = T'_1 \text{ and } (s,\alpha_2) \in T'_2 \right\} \\ N = N_1 + N_2 \end{cases}$$

Now, the typical merging phase of two clusters for a hierarchical clustering (like UPGMA) is efficiently computed as a merge between two pseudo concepts. The clustering result can be represented on a graph familiar to most biologists: the TreeView algorithm proposed in [Eisen *et al.*, 1998] was used. A portable implementation is available (see TreeView, http://genetics.stanford.edu/~alok/TreeView/). This helps the biologist deciding at which depth a cluster of formal concept is analyzed, and the resulting bi-set will now be called a Quasi-Synexpression-Group (QSG).

Every QSG can be viewed by representing the grouping of formal concepts it harbors either as a function of genes or as a function of biological situations. In any case, a color-coding approach supports the visu identification of potentially interesting QSGs. An option enables cluster selection: graphically, it is possible to select an area of the TreeView output. This area corresponds to a formal concept set for which several biological situations or tags are over-represented. The output is a representation of a new matrix showing biological situations in line and the genes in columns. At the intersection, there is the number of formal concepts in which the gene is found over-expressed within the QSG. Using this representation, it is easy to identify which genes or situations are really in the QSG such that one can remove a marginal gene or situation.

Availability

The implemented software prototypes used in this study are either online (BioMiner available on http://liris.cnrs.fr/dmidb/BioMiner/index.php) or available for free upon request to the authors. We have

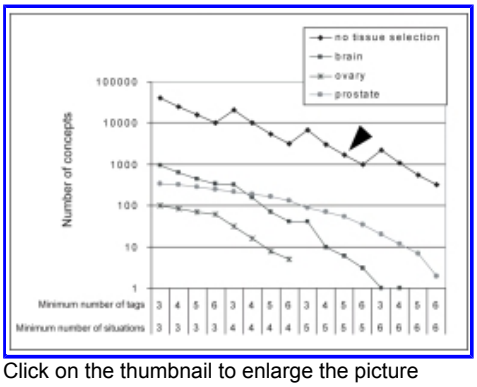
implemented a web-based database, called SQUAT (for "SAGE querying and analysis tools"), which allo biologist to query raw SAGE data as well as formal concepts and QSGs. This database is available at: http://bsmc.insa-lyon.fr/squat and will be described elsewhere (Leyritz et al., in preparation).

Results

Generation of QSGs

64,836 formal concepts holding in the 5327×90 Boolean matrix were extracted. As a reminder, in such a matrix, there are theoretically 2^{90} possible formal concepts (~10^{30}). Even if the end result is a very small proportion (~6.5×10^{-26}) of all possible combinations, this nevertheless represents an unmanageable amount of information for the biologist.

In order to reduce it, formal concepts were first selected according to the tissue homogeneity and the size of the formal concepts. The evolution of the number of formal concepts with respect to these criteria is shown on Fig. 2.



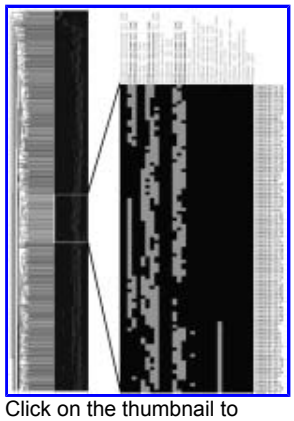
Click on the thumbnail to enlarge the picture

Figure 2: Selection of formal concepts according to different criteria. The graph sums up the variations of the number of formal concepts according to different values of the parameters. "No tissue selection" means that all the formal concepts have the required minimal size (at least 3 tags and 3 situations as the starting point). The black arrowhead points toward the formal concept set which was aggregated by clustering.

It is interesting to note that a large majority of the formal concepts involve libraries from different organ types, since using the organ homogeneity criterion severely reduced the number of formal concepts. By using the word "Brain" as a selection filter (note that 25% of our libraries are derived from brain), one drops from 41,114 formal concepts (obtained from a selection filter using *no* keyword, *at least* 3 tags and *at least* 3 libraries) to 961 ("brain" keyword, *at least* 3 tags and *at least* 3 libraries - cf. Fig. 2). The method therefore captures the simultaneous over-expression of a limited number of genes that are mostly not tissue specific. This is expected because of the local property of the extracted patterns.

This first selection procedure reduces the number of formal concepts but a second selection step can furtl simplify the biological interpretation. Therefore, the potential of the formal concept clustering method wa assessed.

For this, from the original 64,836 concepts, the 1669 concepts involving at least 5 libraries and at least 5 tags (see arrowhead on Fig. 2) were first selected. Those concepts were then grouped according to the hierarchical clustering approach. The resulting clusters were visualized using Tree view [Eisen *et al.*, 1998]. This leads to 50 QSGs than can be identified visually (Fig. 3). Numerous QSGs were analyzed (not shown). Among those, a representative QSG associating 7 SAGE libraries to 13 tags (Tab. 1) was selected for further studies.



Click on the thumbnail to enlarge the picture

Figure 3: Visual selection of a quasi-synexpression-group (QSG). This classification is obtained from the 1669 formal concepts selected with the following selection criteria: the formal concepts associate at least five libraries with at least five tags. The visually selected QSG is composed in arrays by all 95 formal concepts and in columns by 7 SAGE libraries (shown in bold) mainly present in the QSG. On the left, one can see the formal concept classification (only half of the classification is shown). A similar classification can be obtained by displaying the formal concepts as a function of tags.

Table 1: One sample of a biologically interesting QSG.

A. Repartition of the formal concepts in the QSG with respect to the tags and the libraries of the cluster

Tag sequence	Library number						
	20	**37**	**45**	**46**	**48**	**51**	**57**
AGATCCTACT	14	13	0	9	0	13	13
AGCTCTCCCT	22	19	27	21	21	27	0
AGGCTACGGA	13	12	18	14	13	18	0
AGGGTGAAAC	16	14	19	0	16	17	11
CAGCTCACTG	18	17	15	19	0	15	17
CCCATCCGAA	67	60	71	65	45	61	47
CCTCCACCTA	14	14	16	16	11	0	13
CGGTTTGCAG	66	60	69	62	44	58	47
CTCTTCGAGA	0	15	17	16	11	14	9
GCGTGATCCT	18	18	20	17	16	16	16
GGCAAGAAGA	29	0	29	26	20	26	16
GTTGCTGCCC	56	51	54	50	35	40	46
TGGGCAAAGC	14	12	14	15	11	11	0

It should be read as follows: there are 14 formal concepts in the QSG that associates the tag AGATCCTACT to the library no. 20. The zeros points to tags/situations associations that were not found in any of the 95 formal concepts.

B. Identification of the tags contained in the QSG.

Tag sequence	Tag identification	Gene symbol
AGATCCTACT	farnesyl-diphosphate farnesyltransferase 1	*FDFT1*
AGCTCTCCCT	ribosomal protein L17	*RPL17*
AGGCTACGGA	ribosomal protein L13a	*RPL13A*
AGGGTGAAAC	splicing factor, arginine/serine-rich 9	*SFRS9*
CAGCTCACTG	ribosomal protein L14	*RPL14*
CCCATCCGAA	Transcribed sequence with strong similarity to protein sp:Q02877 (H.sapiens) RL26_HUMAN 60S ribosomal protein L26	?
CCTCCACCTA	peroxiredoxin 2	*PRDX2*
CGGTTTGCAG	Nit protein 2	*NIT2*
CTCTTCGAGA	glutathione peroxidase 1	*GPX1*
GCGTGATCCT	aldo-keto reductase family 1, member A1 (aldehyde reductase)	*AKR1A1*

GGCAAGAAGA	ribosomal protein L27	*RPL27*
GTTGCTGCCC	seven transmembrane domain protein	*NIFIE14*
TGGGCAAAGC	eukaryotic translation elongation factor 1 gamma	*EEF1G*

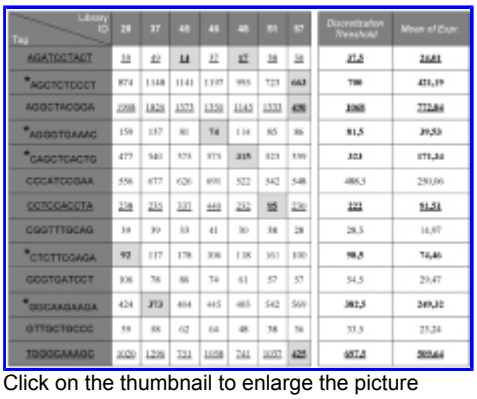
C. Description of the libraries contained in the QSG.

Library number	Library name	Library description
20	SAGE_LNCaP	Prostate adenocarcinoma cell line
37	SAGE_A+	Prostate carcinoma cell line
45	SAGE_Chen_LNCaP	Prostate carcinoma cell line
46	SAGE_Chen_LNCaP_no-DHT	Prostate carcinoma cell line
48	SAGE_Chen_Tumor_Pr	Prostate carcinoma bulk
51	SAGE_HS766T	Pancreas adenocarcinoma cell line
57	SAGE_CPDR_LNCaP-C	Prostate carcinoma cell line

These SAGE libraries were all constructed from carcinoma and almost all from prostate carcinoma.

Part A of Tab. 1 shows the main advantage of the clustering of concepts: all of the 7 libraries have a null value for at least one tag. This means that no concept associates this tag with this given library. In order to understand the origin of these null values, which can be considered as noise, it is possible to look back at the original gene expression levels in these libraries.

The Fig. 4 shows two origins for this noise: first the employed discretization which can sometimes be too strict (tags labeled with an asterisk in Fig. 4 would have been considered as over-expressed using a slightly lower threshold). The second one is experimental or biological: from their raw expression values, four tags cannot be considered as over-expressed in some libraries (tags underlined in Fig. 4).



Click on the thumbnail to enlarge the picture

Figure 4: Origin of the noise in the concepts composing the presented QSG. Shown is a submatrix of the raw gene expression matrix with the tags and libraries composing the presented QSG. The expression values are expressed as "copies per cell" (see methods section). On the right side is shown the discretization thresholds used for each gene, as well as the mean expression value among all 90 situations. The shaded expression values correspond to the 0 values in Part A of Tab. 1. For a correspondence between tag sequence sand gene names, and between Library numbers and biological situations, see Tab. 1B and C, respectively.

Biological interpretation of one QSG

The QSG contains 7 SAGE libraries (Tab. 1). All these libraries are derived from prostate carcinoma (two of them are adenocarcinoma). Six of them are cell lines and one of them is bulk. It is quite interesting to note that this tissue homogeneity was not selected beforehand, and that even when using locally strong associations, tissue-specific gene over-expression patterns have emerged.

The QSG contains 13 tags (Tab. 1). 12 of them are clearly identified: *FDFT1*, *RPL17*, *RPL13A*, *SFRS9*,

RPL14, *PRDX2*, *NIT2*, *GPX1*, *AKR1A1*, *RPL27*, *NIFIE14* and *EEF1G*.

Gene Ontology annotates *FDFT1* as a gene involved in cholesterol biosynthesis. In particular, it is essential for squalene synthesis [Tansey and Shechter 2000]. It was found amplified in oesophagus carcinomas [Hughes *et al.*, 1998].

RPL17
is annotated by Gene Ontology as a gene involved in protein biosynthesis and in signal transduction in the NF-κB cascade. It was found in association with NF-κB during the differentiation of adherent blood cells as macrophages [Day *et al.*, 2004]. Other ribosomal proteins (RPL13A, RPL14 and RPL27) are present. RPL13A was found to be involved in cell proliferation [Chen and Ioannou 1999]. It was also reported to be a translation regulator [Mazumder *et al.*, 2003] similarly to EEF1G [Zimmermann, 2003]. RPL14 was found over-expressed in glioma [Qi *et al.*, 2002].

Gene Ontology annotates EEF1G as a translational elongation factor involved in protein biosynthesis. It is highly expressed in pancreatic cancers [Lew *et al.*, 1992].

SFRS9 (alias SRp30c) is involved in splice site selection [Raffetseder *et al.*, 2003]. It was found to be over-expressed in cancer (Hela cell line) [Screaton *et al.*, 1995]. It was also associated with the response to oxidative stress by regulating the splicing of the glucocorticoid receptor of neutrophils [Xu *et al.*, 2003] or interacting with HSPs [Metz *et al.*, 2004].

PRDX2 (alias *NKEFB*, *PRP*, *PRXII*, *TDPX1*, *TSA*) is annotated by Gene Ontology as a gene coding for a protein involved in electron transport and response to oxidative stress. It was found in benign vascular tumours of the skin [Lee *et al.*, 2003] but also in breast cancers [Noh *et al.*, 2001; Karihtala *et al.*, 2003] and prostate tumour cell lines [Shen and Nathan 2002]. Its function during oxidative stress seems to depend upon environmental conditions, especially affecting the nitric oxide concentrations [Simzar *et al.*, 2000]. Interestingly, it was found over-expressed in the AML-2/DX100 cell line, which is derived from the AML-2/WT cell line and is more resistant to endogenous oxidative stress in spite of a catalase inhibition. This suggests that changes in several gene expression levels - including an increase in the expression of *PRDX2* - characterize the adaptation of the cell line exposed to endogenous oxidative substances [Oh *et al.*, 2004]. Finally, *PRDX2* was found to be up-regulated in apoptosis resistant cell lines [Crowley-Weber *et al.*, 2002].

GPX1
is annotated by Gene Ontology as a gene involved in response to oxidative stress. Like PRDX2, its product has a peroxidase activity. It is also involved in the adaptation of cells to oxidative stresses [Anuszewska *et al.*, 1997].

AKR1A1
is a gene coding for a protein involved in oxidative molecule degradation (reductase), especially in the reduction of biogenic and xenobiotic aldehydes [Barski *et al.*, 1999].

Gene Ontology annotates *NIT2* as a gene coding for a protein involved in nitrogen metabolism. This protein is well known in fungi (especially *Neurospora crassa*) as a member of the GATA family of transcription factors. It is a positive global regulator which controls the expression of entire sets of nitrogen-catabolyzing genes, especially nitrate reduction [Mo and Marzluf, 2003]. For *Homo sapiens*, *NIT2* cDNA has first been found in embryonal carcinoma of human testis [Strausberg *et al.*, 2002].

Altogether, three major tendencies emerge from this list of genes. First, there is a very strong connectivity between most of these genes with cancer; second there are numerous genes coding for proteins involved oxidative stress response, and third there are a significant number of genes coding for proteins involved translation regulation.

It is known that exposition to an oxidative stress is a factor that favors development of different types of tumors [Valko *et al.*, 2004]. It is therefore reasonable to suggest that these genes are co-over-expressed respond to an oxidative stress to which cells have been exposed. This is in good agreement with

epidemiological, experimental and clinical studies which have, over the last decade, implicated oxidative stress in development and progression of prostate cancer [Pathak *et al.*, 2005].

Another feature is the over-expression of proteins involved in translational regulation. We had already observed the co-over-expression of various mRNAs coding for ribosomal proteins [Becquet *et al.*, 2002]. The over-expression of ribosomal proteins in several interesting contexts, including prostate cancers wa independently reported [Vaarala *et al.*, 1998]. The biological role for such an over-expression is still a matter of debate [Naora 1999].

Statistical analysis of one QSG

In order to estimate the statistical significance of the regrouping of the genes in the QSG, two different web tools were used: L2L [Newman and Weiner 2005] and GOToolBox [Martin *et al.*, 2004]. Both tools provide, given a gene list, GO categories that are statistically overrepresented as compared to a gene random sampling. Using the genes contained in the previously described QSG, both sites returned as the first hit t "protein biosynthesis" category:

- $p = 1.1\times10^{-6}$ with L2L;
- $p = 9.35\times10^{-8}$
  with GOToolBox (using an hypergeometric test with a Benjamini-Hochberg correction for multiple testing and all UniProt identifiers generated from the RefSeq identifiers).

This fully confirms that the regrouping of genes in the QSG is highly statistically significant, and further back up our biological analysis: the over-expression of proteins involved in translational regulation is a significar feature captured by our analysis. In order to compare the statistical significance of the individual concepts that composed the QSG, the 95 formal concepts composing the QSG were analyzed using a local version of L2L. It appears that only 2 of them (2%) have lower p-values than the QSG for the GO category 'protein biosynthesis'. Using GOToolBox, we also compared those formal concepts with the QSG and obtained similar results. Altogether, these results show that a QSG is able to summarize the main biological functions contained in the various formal concepts composing it. In this sense, a QSG is more informative than the formal concepts separately.

A second important feature of the presented QSG was the appearance of the "response to oxidative stre category. This category was indeed found over-represented, both with L2L ($p = 1.09\times10^{-4}$), and GOToolBox (p = 0.0011). This demonstrates that the statistical significance is somewhat less marked that for the "p biosynthesis" category.

We also examined the 95 individual concepts for this category. 75 (80 %) of them had a p-value that was not statistically significant ($p > 0.05$) for this term. The only one concept, among the 95, which associates the 4 genes that were identified as participating to the response to oxidative stress was then compared. It has a lower p-value (L2L p-value=3.71×10^{-5}; GOToolBox p-value= 3.02×10^{-4}) than the QSG but the 'protein biosynthesis' category is entirely absent. This therefore demonstrates one of the main interests of the QSG that is to associate to significant extent different functions that were distributed to a sub-significant level in individual concepts.

We furthermore assessed the importance of locality of pattern extracted. For this we performed hierarcl clustering, SOM and k-means using Cluster, implemented by Michael Eisen (see http://rana.lbl.gov/EisenSoftware.htm). For each method, several sets of parameters were tested. None of these methods were able to capture the co-overexpression patterns that we biologically validated, since the genes composing the QSG were scattered throughout the resulting clusters (data not shown).

Discussion

The main advantage of local pattern discovery techniques is that they assume very little background knowledge and fit well with exploratory unsupervised data mining processes. Complete collections of locally strong associations can be extracted and presented for interpretation. As a result, unexpected associations can be discovered.

The main drawback of these approaches is that huge numbers of patterns are extracted (up to millions). Therefore, the required interpretation process has to be supported by means of sophisticated post-processing techniques. In this paper, we propose a solution that severely reduces the number of patterns (i.e., our technique performs an heuristic pattern aggregation) to be examined by the end user when dealing with intrinsically noisy data. Interestingly, at any moment, the analyst can however go back not only to the data (including the original gene expression data) but also to each of the local patterns that has been used to build a given cluster or QSG.

For gene expression data analysis, we believe that formal concepts allow an easier biological interpretation than the popular association rules that were recently studied by several groups. Indeed, each formal concept provides a "summary" of the information contained in many association rules because various (but similar) association rules can be generated from a unique formal concept [Rioult *et al.*, 2003]. The number of patterns is significantly reduced. Formal concepts are also easier to interpret since association rules only associate genes (or samples) whereas formal concepts associate both genes and the biological samples in which they are co-expressed.

However, generating biologically interesting hypotheses based on local patterns, like formal concepts, is confronted with several difficulties. The necessity to encode Boolean gene expression properties is a key issue. There is no single method to encode the over-expression property and it is sure that this choice has a major impact on extracted patterns. Some recent efforts have been made to guide the choice of discretization techniques and parameters [Pensa *et al.*, 2004].

A second problem concerns formal concept extraction and comes paradoxically from its added value: the strong locality of these patterns does not help to distinguish easily between valid formal concepts, say true positives, and spurious patterns due to noise. This has to be moderated when considering the over-expression of genes since it is clear that SAGE data is much noisier for low expression levels that it is for high expression ones. Again, recent techniques can be used for tackling these issues. For instance, we can use user-defined constraints to look for large enough patterns and thus to avoid small patterns that are indeed due to noise. It is also possible to use randomization techniques to remove some false positive patterns [see, e.g., Gionis *et al.*, 2006]. We decided to tackle this aspect by proposing a new clustering method to group similar formal concepts. By applying such an approach on human SAGE data, a manageable number (about 50) of interesting clusters of formal concepts or QSGs have been extracted and one of them has been presented here in detail.

Without clustering patterns, the information contained in a QSG would have been distributed in 95 concepts scattered among 1669. This demonstrates the noise tolerance of the method and its power to condense complex information into readable and understandable patterns. One should note that very heterogeneous sources of noise are combined to produce the final "noise" observed in gene expression matrices. Some of this noise should be disregarded (because of its artifactual origin) and some might be of biological interest [Kaern *et al.*, 2005]. Our clustering technique allows biologists to investigate this question since it can be used to study the reason for the absence of a gene over-expression value and go back to the whole set of formal concepts contained in a QSG.

QGS can be ranked according to statistical criteria (like the p-value calculated by web-based tools, see above). Nevertheless, this p-value can not be the only criterion. Functionally heterogeneous QSGs might be biologically interesting. In that perspective, using a tool to aggregate formal concepts and visually select them by a biologist - because they are supported by several biological situations of interest - avoids to eliminate some of the patterns which may be interesting but do not hold the statistical test.

A careful analysis of one of the QSGs shows that it is possible to formulate very relevant biological hypotheses. It was very rewarding that some of these hypotheses (such as the role of oxidative stress i

generation of prostate cancer, or some proteins-proteins interactions) could be validated *a posteriori* through a literature search. Nevertheless some novel hypotheses were also raised by this approach: what might be the function of over-expression of some genes of the translation machinery in the generation of prostate cancers? Ultimately, only biological experiments will be able to answer those questions.

We mentioned some fundamental limitations of popular global patterns used for gene expression data analysis, i.e., clusters of genes or non-overlapping bi-clusters linking gene sets to sample sets. It is however clear that these techniques, biclustering techniques and our approach are complementary. The former provide feedback on global similarity structures. Biclustering techniques, that somehow include the computation of formal concepts from derived Boolean matrices, might give rise to the discovery of unexpected locally strong associations. It turns out that software platforms dedicated to the analysis of expression data, whatever their origin (i.e. SAGE, microarray, MSPSS, etc…), might support both clustering and local pattern discovery. For instance, it is possible to use local patterns like formal concepts or association rules when characterizing clusters. It is also possible to compute clusters from collections of l patterns (e.g., biclusters, formal concepts). Last but not the least, it would make sense to integrate the tools that work on Boolean data with the more classical techniques that compute biclusters from numerical g expression data sets (see, e.g., the BicAT toolbox by Barkow *et al.*, 2005). As a result, it means that we have to support both the computation of various types of patterns and querying or more generally post-processing on materialized collections of patterns. These observations motivate our current effort for a software integration of many different solvers into a unique platform.

Acknowledgements

We thank Céline Keime for her help in identifying tags and for helpful discussions. Sylvain Blachon was a fellow from the Comité de Saône et Loire de la Ligue Contre le Cancer. We are indebdtful to Edmund Derrington (CGMC UMR 5534) for his critical and thoughtful reading of the manuscript, as well as for his sense of acronyms.

This work has been partially funded by the ACI "Masse de données" (MD46, Bingo (BINGO). The work in Olivier Gandrillon's laboratory is supported by the Ligue contre le Cancer (Comité Départemental du Rhône), the UCBL, the CNRS, the Région Rhône Alpes (Thématique prioritaire) and the Association pour la Recherche contre le Cancer (ARC). The work by Jérémy Besson is supported by an ASC INRA.

References

- Anuszewska, E. L., Gruber, B. M. and Koziorowska, J. H. (1997). Studies on adaptation to adriamycin in cells pretreated with hydrogen peroxide. Biochem. Pharmacol. **54**, 597-603.
- Barski, O. A., Gabbay, K. H. and Bohren, K. M. (1999). Characterization of the human aldehyde reductase gene and promoter. Genomics **60**, 188-198.
- Barkow, S., Bleuler, S., Prelic, A., Zimmermann, P. and Zitzler, E. (2006). BicAT: a biclustering analysis toolbox. Bioinformatics **22**, 1282-1283.
- Becquet, C., Blachon, S., Jeudy, B., Boulicaut, J.-F. and Gandrillon, O. (2002). Strong-association-rule mining for large-scale gene-expression data analysis: a case study on human SAGE data. Genome Biol. **3**, RESEARCH0067.
- Besson, J., Robardet, C., Boulicaut, J.-F. and Rome, S. (2005). Constraint-based concept mining and its application to microarray data analysis. Intelligent Data Analysis **9**, 59-82.
- Besson, J., Pensa, R., Robardet, C. and Boulicaut, J.-F. (2006a). Constraint-based mining of fault-tolerant patterns from Boolean data. Knowledge Discovery in Inductive Databases. KDID'05, Revised Selected and Invited papers.Lecture Notes in Computer Science **3933**, 55-71.

- Besson, J., Robardet, C. and Boulicaut, J.-F. (2006b). Mining a new fault-tolerant pattern type as an alternative to formal concept discovery. Conceptual Structures: Inspiration and Application. 14th International Conference on Conceptual Structures ICCS'06, Proceedings. Lecture Notes in Artificial Intelligence **4068**, 144-157.
- Carmona-Saez, P., Chagoyen, M., Rodriguez, A., Trelles, O., Carazo, J. M. and Pascual-Montano, A. (2006). Integrated analysis of gene expression by Association Rules Discovery. BMC Bioinformatics **7**, 54.
- Chen, F. W. and Ioannou, Y. A. (1999). Ribosomal proteins in cell proliferation and apoptosis. Int. Rev. Immunol. **18**, 429-448.
- Creighton, C. and Hanash, S. (2003). Mining gene expression databases for association rules. Bioinformatics **19**, 79-86.
- Crowley-Weber, C. L., Payne, C. M., Gleason-Guzman, M., Watts, G. S., Futscher, B., Waltmire, C. N., Crowley, C., Dvorakova, K., Bernstein, C., Craven, M., Garewal, H. and Bernstein, H. (2002). Development and molecular characterization of HCT-116 cell lines resistant to the tumor promoter and multiple stress-inducer, deoxycholate. Carcinogenesis **23**, 2063-2080.
- Day, C. J., Kim, M. S., Stephens, S. R., Simcock, W. E., Aitken, C. J., Nicholson, G. C. and Morrison, N. A. (2004). Gene array identification of osteoclast genes: differential inhibition of osteoclastogenesis by cyclosporin A and granulocyte macrophage colony stimulating factor. J. Cell. Biochem. **91**, 303-315.
- Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci USA **95**, 14863-14868.
- Elati, M., Radvanyi, F. and Rouveirol, C. (2005). Mining transcriptional regulation from expression data. Actes Journées Ouvertes de Biologie Informatique et Mathématiques (JOBIM), Lyon.
- Georgii, E., Richter, L., Rückert, U. and Kramer, S. (2005). Analyzing microarray data using quantitative association rules. Bioinformatics **21 Suppl. 2**, ii123-ii129.
- Gionis, A., Mannila, H., Mielikäinen, T. and Tsaparas, P. (2006). Assessing data mining results via swap randomization. 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD'06. ACM, 157-176.
- Goethals, B. and Zaki, M. J. (2004). Advances in frequent itemset mining implementations: report on FIMI'03. SIGKDD Explorations **6**, 109-117.
- Hughes, S. J., Glover, T. W., Zhu, X.-X., Kuick, R., Thoraval, D., Orringer, M. B., Beer, D. G. and Hanash, S. (1998). A novel amplicon at 8p22-23 results in over-expression of cathepsin B in esophageal adenocarcinoma. Proc. Natl. Acad. Sci. USA **95**, 12410-12415.
- Kaern, M., Elston, T. C., Blake, W. J. and Collins, J. J. (2005). Stochasticity in gene expression: from theories to phenotypes. Nat. Rev. Genet. **6**, 451-64.
- Karihtala, P., Mäntyniemi, A., Kang, S. W., Kinnula, V. L. and Soini, Y. (2003). Peroxiredoxins in breast carcinoma. Clin. Cancer Res. **9**, 3418-3424.
- Keime, C., Damiola, F., Mouchiroud, D., Duret, L. and Gandrillon, O. (2004). Identitag, a relational database for SAGE tag identification and interspecies comparison of SAGE libraries. BMC Bioinformatics **5**, 143.
- Keime, C., Sémon, M., Mouchiroud, D., Duret, L. and Gandrillon, O. (2007). Unexpected observations after mapping LongSAGE tags to the human genome. BMC Bioinformatics **8**, 154.
- Klema, J., Soulet, A., Crémilleux, B., Blachon, S. and Gandrillon, O. (2006). Mining Plausible Patterns from Genomic Data. 19th IEEE International Symposium on Computer-Based Medical Systems CBMS'06, IEEE Computer Society, 183-190.
- Kuznetsov, S. O. and Obiedkov, S. A. (2002). Comparing performance of algorithms for generating concept lattices. Journal of Experimental and Theoretical Artificial Intelligence **14**, 189-216.
- Lee, H. K., Hsu, A. K., Sajdak, J., Qin, J. and Pavlidis, P. (2004). Coexpression analysis of human genes across many microarray data sets. Genome Res. **14**, 1085-1094.

- Lee, S.-C., Na, Y.-P. and Lee, J.-B. (2003). Expression of peroxiredoxin II in vascular tumors of the skin: a novel vascular marker of endothelial cells. J. Am. Acad. Dermatol. **49**, 487-491.
- 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. and Young, R. A. (2002). Transcriptional regulatory networks in Saccharomyces cerevisiae. Science **298**, 799-804.
- Lew, Y., Jones, D. V., Mars, W. M., Evans, D., Byrd, D. and Frazier, M. L. (1992). Expression of elongation factor-1 gamma-related sequence in human pancreatic cancer. Pancreas **7**, 144-152.
- Li, J., Liu, H., Downing, J. R., Yeoh, A. E.-J. and Wong, L. (2003). Simple rules underlying gene expression profiles of more than six subtypes of acute lymphoblastic leukemia (ALL) patients. Bioinformatics **19**, 71-78.
- Madeira, S. C. and Oliveira, A. L. (2004). Biclustering algorithms for biological data analysis: a survey. IEEE/ACM Trans. Comput. Biol. Bioinform. **1**, 24-45.
- Martin, D., Brun, C., Remy, E., Mouren, P., Thieffry, D. and Jacq, B. (2004). GOToolBox: functional analysis of gene datasets based on Gene Ontology. Genome Biol. **5**, R101.
- Mazumder, B., Sampath, P., Seshadri, V., Maitra, R. K., DiCorleto, P. E. and Fox, P. L. (2003). Regulated release of L13a from the 60S ribosomal subunit as a mechanism of transcript-specific translational control. Cell **115**, 187-198.
- Metz, A., Soret, J., Vourc'h, C., Tazi, J. and Jolly, C. (2004). A key role for stress-induced satellite III transcripts in the relocalization of splicing factors into nuclear stress granules. J. Cell Sci. **117**, 4551-4558.
- Mo, X. and Marzluf, G. A. (2003). Cooperative action of the NIT2 and NIT4 transcription factors upon gene expression in Neurospora crassa. Curr. Genet. **42**, 260-267.
- Naora, H. (1999). Involvement of ribosomal proteins in regulating cell growth and apoptosis: translational modulation or recruitment for extraribosomal activity? Immunol. Cell Biol. **77**, 197-205.
- Newman, J. C. and Weiner, A. M. (2005). L2L: a simple tool for discovering the hidden significance in microarray expression data. Genome Biol. **6**, R81.
- Noh, D. Y., Ahn, S. J., Lee, R. A., Kim, S. W., Park, I. A. and Chae, H. Z. (2001). Over-expression of peroxiredoxin in human breast cancer. Anticancer Res. **21**, 2085-90.
- Oh, Y.-K., Lee, T.-B. and Choi, C.-H. (2004). Anti-oxidant adaptation in the AML cells supersensitive to hydrogen peroxide. Biochem. Biophys. Res. Commun. **319**, 41-45.
- Pasquier, N., Bastide, Y., Taouil, R. and Lakhal, L. (1999). Efficient mining of association rules using closed itemset lattices. Information Systems **24**, 25-46.
- Pathak, S. K., Sharma, R. A., Steward, W. P., Mellon, J. K., Griffiths, T. R. L. and Gescher, A. J. (2005). Oxidative stress and cyclooxygenase activity in prostate carcinogenesis: targets for chemopreventive strategies. Eur. J. Cancer **41**, 61-70.
- Pensa, R., Leschi, C., Besson, J. and Boulicaut, J. F. (2004). Assessment of discretization techniques for relevant pattern discovery from gene expression data. 4th ACM SIGKDD Workshop on Data Mining in Bioinformatics BIOKDD'04, ACM, 24-30.
- Prelic, A., Bleuler, S., Zimmermann, P., Wille, A., Bühlmann, P., Gruissem, W., Hennig, L., Thiele, L. and E. Zitzler (2006). A systematic comparison and evaluation of biclustering methods for gene expression data. Bioinformatics **22**, 1122-1129.
- Qi, Z.-Y., Li, Y., Ying, K., Wu, C.-Q., Tang, R., Zhou, Z.-X., Chen, Z.-P., Hui, G.-Z. and Xie, Y. (2002). Isolation of novel differentially expressed genes related to human glioma using cDNA microarray and characterizations of two novel full-length genes. J. Neurooncol. **56**, 197-208.
- Raffetseder, U., Frye, B., Rauen, T., Jürchott, K., Royer, H. D., Jansen, P. L. and Mertens, P. R. (2003). Splicing factor SRp30c interaction with Y-box protein-1 confers nuclear YB-1 shuttling and alternative splice site selection. J. Biol. Chem. **278**, 18241-18248.

- Rioult, F., Robardet, C., Blachon, S., Crémilleux, B., Gandrillon, O. and Boulicaut, J. F. (2003). Mining concepts from large SAGE gene expression matrices. 2nd Int. Workshop Knowledge Discovery in Inductive Databases KDID'03. pp. 107-118.

- Robardet, C., Pensa, R., Besson, J. and Boulicaut, J. F. (2004). Using classification and visualization on pattern databases for gene expression data analysis. Pattern Representation and Management PaRMa'04, CEUR Workshop Proceedings Vol. **16**, pp. 107-118.

- Screaton, G. R., Cáceres, J. F., Mayeda, A., Bell, M. V., Plebanski, M., Jackson, D. G., Bell, J. I. and Krainer, A. R. (1995). Identification and characterization of three members of the human SR family of pre-mRNA splicing factors. EMBO J. **14**, 4336-4349.

- Shen, C. and Nathan, C. (2002). Nonredundant antioxidant defense by multiple two-cysteine peroxiredoxins in human prostate cancer cells. Mol. Med. **8**, 95-102.

- Simzar, S., Ellyin, R., Shau, H. and Sarafian, T. A. (2000). Contrasting antioxidant and cytotoxic effects of peroxiredoxin I and II in PC12 and NIH3T3 cells. Neurochem. Res. **25**, 1613-1621.

- Strausberg, R. L., *et al.*; Mammalian Gene Collection Program Team (2002). Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. Proc. Natl. Acad. Sci. USA **99**, 16899-903.

- Stuart, J. M., Segal, E., Koller, D. and Kim, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. Science **302**, 249-255.

- Tansey, T. R. and Shechter, I. (2000). Structure and regulation of mammalian squalene synthase. Biochim. Biophys. Acta **1529**, 49-62.

- Vaarala, M. H., Porvari, K. S., Kyllönen, A. P., Mustonen, M. V., Lukkarinen, O. and Vihko, P. T. (1998). Several genes encoding ribosomal proteins are over-expressed in prostate-cancer cell lines: confirmation of L7a and L37 over-expression in prostate-cancer tissue samples. Int. J. Cancer **78, 27-32.**

- Valko, M., Izakovic, M., Mazur, M., Rhodes, C. J. and Telser, J. (2004). Role of oxygen radicals in DNA damage and cancer incidence. Mol. Cell. Biochem. **266**, 37-56.

- Velculescu, V. E., Madden, S. L., Zhang, L., Lash, A. E., Yu, J., Rago, C., Lal, A., Wang, C. J., Beaudry, G. A., Ciriello, K. M., Cook, B. P., Dufault, M. R., Ferguson, A. T., Gao, Y., He, T. C., Hermeking, H., Hiraldo, S. K., Hwang, P. M., Lopez, M. A., Luderer, H. F., Mathews, B., Petroziello, J. M., Polyak, K., Zawel, L., Zhang, W., Zhang, X., Zhou, W., Haluska, F. G., Jen, J., Sukumar, S., Landes, G. M., Riggins, G. J., Vogelstein, B. and Kinzler, K. W. (1999). Analysis of human transcriptomes. Nat. Genet. **23**, 387-378.

- Xu, Q., Leung, D. Y. and Kisich, K. O. (2003). Serine-arginine-rich protein p30 directs alternative splicing of glucocorticoid receptor pre-mRNA to glucocorticoid receptor beta in neutrophils. J. Biol. Chem. **278**, 27112-27118.

- Zaki, M. J. and Hsiao, C.-J. (2002). CHARM: An Efficient Algorithm for Closed Itemset Mining. Proceedings Second SIAM International Conference on Data Mining, Arlington, USA.

- Zimmermann, R. A. (2003). The double life of ribosomal proteins. Cell **115**, 130-132.