Mining Plausible Patterns from Genomic Data

Jiří Kléma, Arnaud Soulet, Bruno Crémilleux
GREYC, CNRS UMR 6072
Université de Caen, Campus Côte de Nacre,
F-14032 Caen Cédex France
forename.surname@info.unicaen.fr

Sylvain Blachon, Olivier Gandrillon
Centre de Génétique Moléculaire et Cellulaire,
CNRS UMR 5534, Univ. Claude Bernard Lyon 1
F-69622 Villeurbanne Cédex France
surname@cgmc.univ-lyon1.fr

Abstract

The discovery of biologically interpretable knowledge from gene expression data is one of the largest contemporary genomic challenges. As large volumes of expression data are being generated, there is a great need for automated tools that provide the means to analyze them. However, the same tools can provide an overwhelming number of candidate hypotheses which can hardly be manually exploited by an expert. An additional knowledge helping to focus automatically on the most plausible candidates only can up-value the experiment significantly. Background knowledge available in literature databases, biological ontologies and other sources can be used for this purpose. In this paper we propose and verify a methodology that enables to effectively mine and represent meaningful over-expression patterns. Each pattern represents a bi-set of a gene group over-expressed in a set of biological situations. The originality of the framework consists in its constraint-based nature and an effective cross-fertilization of constraints based on expression data and background knowledge. The result is a limited set of candidate patterns that are most likely interpretable by biologists. Supplemental automatic interpretations serve to ease this process. Various constraints can generate plausible pattern sets of different characteristics.

1 Introduction

The most usual way of gene-expression analysis is based on clustering [12]. The genes are grouped according to similarity in their gene expression profiles, the clusters are searched to find those containing genes with common biological properties, such as the presence of common upstream promoter regions or involvement in the same biological processes. However, genes rarely exhibit a similar expression profile across a wide range of biological situations which is the underlying idea of hierarchical or K-means clustering. Biclustering of gene expression data (also called co-clustering or two-way clustering) is a later methodology for the identification of *bi-sets*, i.e., gene groups that show a coherent expression profile across a subset of situations [25]. It can be understood as a natural methodology screening for genes that are functionally related, affected by the same drug or pathological condition, or genes that participate in the same pathways being potentially co-regulated by a small group of transcription factors [31].

Recent works on pattern mining improve the discovery of similar bi-sets by using significant constraints to focus the search [7, 13]. The raw expression data are binarized, typically in order to encode the over-expression. However, various binary properties such as under-expression or strong variability can also be studied. Although this binarization seems to cause an indispensable loss of information, when dealing with SAGE (Serial Analysis of Gene Expression) [30] data it can be rather advantageous. The obvious reason is a significant tag frequency error rate which turns out to be especially high for low frequency tags [7].

In this text we understand the term *pattern* as a set of tags over-expressed in a set of biological situations. This definition is quite vague and covers an extreme number of trivial patterns that are a priori uninteresting, e.g., a single tag over-expressed in a single situation. That is why we define additional constraints helping to focus on the most interesting fraction of the pattern space only, so-called *plausible patterns*. First, we try to discover patterns that are large enough. They must have a certain minimum length (i.e., contain a certain minimum number of tags) and frequency (i.e., cover a certain minimum number of situations). These are typical constraints used in the field of association rule mining [13]. They have also been applied in the domain of genomics [21]. As they are based purely on the boolean data being mined we will refer to them as the *internal constraints*. But, in this paper, to better achieve a cross-fertilization with the background knowledge, we do not restrict the search to maximum rectangles (i.e., closed patterns [13]) as it is usually done. This requires powerful data mining techniques (see Section 3).

Second, our interest lies in patterns that are biologically relevant. Generally speaking, every pattern interpretable by biologists might be considered interesting. More precisely, the interesting patterns are those exhibiting a general characteristic common for the tags and/or situations concerned (or at least their sub-sets). These constraints can in no case be inferred from the expression data. We refer to them as the *external constraints* and infer them from the external knowledge (in this text we consider the terms background and external knowledge as synonyms). An effective use of background knowledge in analysis and interpretation of expression data is a popular research topic nowadays. However, the main effort is aimed at clustering and thus integration of the biological knowledge into the statistical data analysis framework. The background knowledge is typically used to annotate the expression based clusters for statistically over-represented (or under-expressed) terms or categories [14, 19]. It can also be used to cluster the genes immediately [10] or to perform meta-clustering where the expression and external datasets are combined prior to clustering [15]. Among the approaches distinct from clustering, [33] deals with gene annotations as with relational logic features or [27] uses text mining to filter the most promising disease gene candidates.

In this paper, we work with two principal external data sources, freetexts and gene ontologies (GOs). In the area of freetexts we have been inspired mainly by [10, 15]. Both of them deal with the term-frequency vector representation which is a simple however prevailing representation of texts. This representation allows for an annotation of a gene group as well as a straightforward definition of gene similarity. In the area of gene ontologies we stem from [19], the gene similarity results from the genes' positions in the molecular functional, biological process or cellular component ontology. The gene similarity is instrumental to the definition of constraints coming later on.

To our best knowledge, there is no substantial work on constraint-based pattern mining with the constraints inferred from the genomic background knowledge. Using external constraints in the context of pattern mining as well as a synergic combination of internal and external constraints are the main contributions of this paper.

To sum up, a promising, i.e., plausible, candidate pattern could have the following structure: "Tags A, B, C and D are jointly over-expressed in the biological situations 1, 2, 3 and 4." The supplemental interpretation can be such as: "Tags A, B and C share the biological function F and in the literature they often co-occur with terms T1, T2 and T3, while tag D has no information attached yet. The biological situations 1-4 represent cancerous tissues taken from organs O1 or O2." Potentially new knowledge gained from the plausible pattern can be two-fold. Firstly, function F (and/or terms T1, T2, T3) may truly interact with the cancerous context of the situations and secondly, tag D can prospectively share the same function F with the other tags. The main goal is to transform these verbal statements into formal constraints (and vice versa) and employ them within an effective pattern mining framework. Section 4 depicts such plausible patterns.

Section 2 summarizes the datasets, the way they were generated, preprocessed and utilized. It also provides the overall links among these datasets. Section 3 briefly introduces the constraint-based pattern mining tool MUSIC. The section also shows how to formalize constraints within the genomic domain under consideration. Section 4 demonstrates that the pattern sets reduced by the external constraints can serve as an enriched source of potential biological "nuggets". They are highlighted and interpreted. Finally, Section 5 summarizes the whole methodology, discusses its strengths and weaknesses and future work.

2 Raw data treatment and interaction

In this paper we stem from several different data sources. They greatly vary in their origin, a way of preprocessing and application. First of all, there is a SAGE dataset representing the data to be mined (a rectangular matrix: tags vs. situations). The background knowledge for the tags is extracted from textual resources and gene ontologies. Both of them define a presumed tag similarity as well as they provide an explanation of the selected patterns. Finally, there is a brief textual information on each SAGE library, i.e., each biological situation contained in SAGE. Such information can help to assume similarity among situations covered by the patterns and also to interpret them.

2.1 SAGE data

Like microarrays, the SAGE technique aims to measure the expression levels of genes in a cell population [30]. It is performed by sequencing tags (short sequences of 14 to 21 base pairs (bps)) which are theoretically specific of each mRNA. 207 SAGE libraries (i.e. 207 biological situations or experiments) were downloaded from the NCBI web site [3]. To eliminate putative sequencing errors, a pretreatment of the data described in [7] was applied, giving a set of 125985 14 bp tags. Tags were identified thanks to Identitag [16], using RefSeq mRNA sequences. The unambiguous tags identified with RefSeq were selected, leaving a set of 11082 tags. A 207x11082 gene expression matrix was built. There is also its sub-matrix which confines to the tags belonging to the minimal transcriptome [29]. It is based on 447 tags found and we refer to it as the minimum transcriptome (expression) matrix. Both the matrices were binarized to encode the over-expression of each tag using the MidRange method described in [7].

2.2 Texts and their preprocessing

To access the gene annotation data for every tag considered, RefSeq identifiers were translated into EntrezGene identifiers [2]. The mapping approached 1 to 1 relationship. There were only 11 unidentified RefSeqs, 24 RefSeqs mapped to more than 1 id and 203 ids appeared more than once. Knowing the gene identifiers, the annotations were automatically accessed through hypertext queries to the Entrez Gene database [3] and sequentially parsed by the method stemming from [33]. The non-trivial textual records were obtained for 6302 ids which makes 58% of the total amount of 10858 unique ids (3926 genes had a short summary, 5109 had one abstract attached at least).

The gene textual annotations were converted into the vector space model. A single gene corresponds to a single vector, whose components correspond to a frequency of a single term from the vocabulary. This representation is often referred to as *bag-of-words* [23]. The particular vocabulary consisted of all the *stemmed* terms [4] that appear in 5 different gene records at least. The most frequent terms were manually checked and too general terms (such as gene, protein, human etc.) were removed. The resulting vocabulary consisted of 19373 terms. The similarity between genes was defined as the cosine of the angle between the corresponding *term-frequency inverse-document-frequency* (TFIDF) [23] vectors. TFIDF representation statistically considers how important a term is to a gene record. A similarity matrix for all the tags was generated. The underlying idea is that a high value of two vectors' cosine (which means a low angle among two vectors and thus a similar occurrence of the terms) indicates a semantic connection between the corresponding gene records and consequently their presumable connection. Although this model is known to generate false positive relations for the sake of utilization of the same terms in a different context as well as false negative relations mainly for the sake of synonyms, it is feasible and surprisingly often fitting.

2.3 Gene ontology

The genes can also be functionally related on the basis of their GO terms. The rationale sustaining this method is that the more GO terms the genes share, and the more specific the terms are, the more likely the genes are to be functionally related. [19] defines a distance based on the Czekanowski-Dice formula, the methodology is implemented within the GOProxy tool of GOToolBox [1].

The original RefSeq tag identifiers were translated into UniProt ids [5]. Out of 11082 tags there were 7670 known ids. As this set is too large to be processed by GOToolBox we confined to the minimum transcriptome dataset, 366 RefSeqs could be translated here. The resulting ids have been used by GoToolBox to generate two tag similarity matrices. For the biological process ontology there were 254 valid entries while 271 tags could be diagnosed within the molecular function ontology.

The GO terms themselves could be parsed from the records obtained in the previous subsection.

2.4 Description of libraries

There is a short textual annotation of the length about 10 terms attached to each SAGE library. Although these annotations represent very short documents, their vocabulary is quite compact. Consequently, they can be processed in the same way as the tag textual documentation. In this case, when considering all the terms that appear in 3 and more libraries the vocabulary consists of 83 terms. The situation similarity matrix was also generated.

2.5 General interaction among datasets

One of the basic questions rising prior to mining for the patterns is whether the datasets described above are mutually interconnected. Can we say that a group of tags that are functionally similar also tends to be co-expressed? Is there any relation between GO and textual definitions of similarity? Do similarly annotated situations tend to have similar expression profiles? Although the interconnection between the expression and external data is not a necessary condition to start the mining process, the positive answers would support the overall logic of future experiments – the application of the similarity constraints should also lead to the compact expression data regions.

Correlation can serve as a general interconnection measure between expression and similarity data and also similarity datasets themselves. In order to get the matrices of the same dimension, the tag correlation matrix is derived from the expression data first. Then, its correlation with the tag similarity matrices is calculated. An analogical process is applied when dealing with the situations. Figure 1 shows that there is a statistically significant correlation among all the considered datasets[1]. Nevertheless, the correlation values suggest a weak relationship only. When comparing the individual values, SAGE seems to be most strongly linked to the variance in situations. The interpretation may be such that SAGE deals with very different biological conditions – normal, cancerous or AIDS samples from different organs and individuals of different gender and age. They consequently vary in their expression profiles. The influence of tag similarity seems to be less striking. The similarity measure based on texts does not seem to be less valuable nor redundant with respect to the GO similarities.

3 Constraint-based pattern mining

A *constraint* is a pattern restriction defining the focus of search, in our case interestingness. Gene expression data give rise to new problems w.r.t. the standard application of pattern mining with constraints since the overall complexity is exponential with the number of genes which is

[1] The minimum transcriptome matrix underlies this experiment.

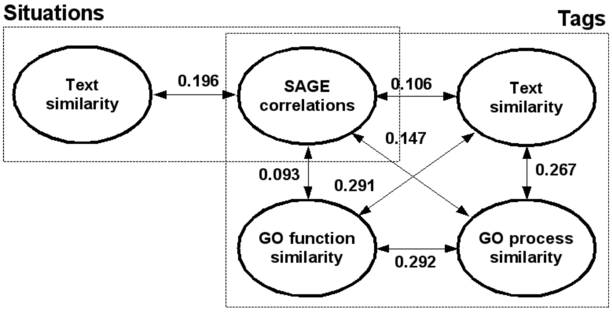

Figure 1. Correlations among the datasets.

large. A simple approach is first to mine all potentially interesting patterns satisfying an anti-monotone constraint (e.g., the usual constraint of frequency) because this class of constraints can be efficiently pushed and second to filter the resulting set of patterns by the other constraints. However, this naïve filtering approach fails due to the huge number of patterns. Existing scalable techniques [20, 22] are limited to particular kinds of constraints (closed patterns, δ-free patterns). In this paper we apply a more general framework which is based on a rich declarative language of *primitive-based constraints* enabling an effective internal *pruning* and a condensed output representation based on *intervals*. Experiments (Section 4) show its effectiveness.

We use the tool MUSIC (Mining with a User-SpecifIed Constraint) [26] which discovers soundly and completely all the patterns satisfying the specified set of constraints. As MUSIC's theoretical properties cannot be given here the main attention is paid to its basic features applicable within the genomic context. A set of syntactic and aggregate primitives on a pattern x enables to specify a broad spectrum of constraints in a flexible way. These primitives and their combinations form the necessary condition of pattern interestingness. For instance, the product of two primitives $length(x) \times frequency(x)$ may address the patterns having a certain minimum length (i.e., containing a minimum number of tags) and frequency (i.e., covering a minimum number of situations). We refer to it as *area(x)*.

The tag similarity matrices provide a transparent background for external interestingness definitions. We deal with primitives such as $sumsim(x)$ denoting the similarity sum over the set of tags x or $insim(x, min, max)$ for the number of tag pairs whose similarity lies between min and max. As we deal with a certain portion of tags without any information, there are primitives that distinguish between zero similarity and its missing value. The primitive $svsim(x)$ gives the number of tag pairs belonging to x whose mutual similarity is valid and $mvsim(x)$ stands for its counterpart, i.e., the missing interactions when one of the tags has an empty record within the given external representation.

The primitives can make compounds. Among many others, $sumsim(x)/svsim(x)$ makes the average similarity, $insim(x, thres, 1)/svsim(x)$ gives a proportion of the strong interactions (similarity higher than the threshold) within the set of tags, $svsim(x)/(svsim(x) + mvsim(x))$ can avoid patterns with prevailing tags of an unknown function. Relational and logical operators enable to create the final constraint, e.g., $C_1 \geq thres1$ and $C_2 \neq thres2$ where C_i stands for an arbitrary compound or primitive. Constraints can also be simultaneously derived from different external tag and/or situation datasets.

The efficiency of MUSIC lies in a safe pruning of the pattern space by pushing the constraints. The pruning conditions are based on intervals gathering several patterns. Whenever it is computed that all the patterns included in the interval $[x, y]$ simultaneously satisfy (or not) the constraint, the interval is positively (negatively) pruned without enumerating all its patterns [26]. The output of MUSIC enumerates the intervals satisfying the constraint. Such a condensed representation improves the output lucidity and enables to easily compute the *selectivity* of the constraint. Selectivity is a proportion of patterns satisfying the constraint. It constitutes one of its key characteristics.

4 Experiments

This section familiarizes with the practical impact of the external constraints. It demonstrates why a purely internal constraint set can hardly deliver a set of patterns that is compact and interpretable. It formulates examples of external constraints which can localize biologically interesting patterns. One of the outcomes is biologically interpreted in depth. Finally a general scope of the external constraints is discussed.

4.1 Constraint selectivity

Figure 2 shows how many patterns and intervals satisfy the increasing $area$ constraint. To reject useless "longish" patterns, the minimum pattern length was set to 4 and frequency to 5 (patterns such as 1 × 50, 30 × 2, etc. are not considered). Obviously, the patterns of a reasonable area are too numerous to be manually explored. For an example, there are 2090 intervals and 73378 patterns having their area larger than 50. Let us note that the largest area patterns are very likely to be trivial, bringing no new knowledge, and it makes little sense to focus purely on them. At the same time, the selected binarization is modest and generates rather sparse matrices. For other binarization types the explosion of patterns can be even faster.

Simultaneous application of internal and external constraints may help to further reduce the patterns while keeping the interesting ones. The selectivity of selected external constraints is shown in Figure 3. The pruning starts

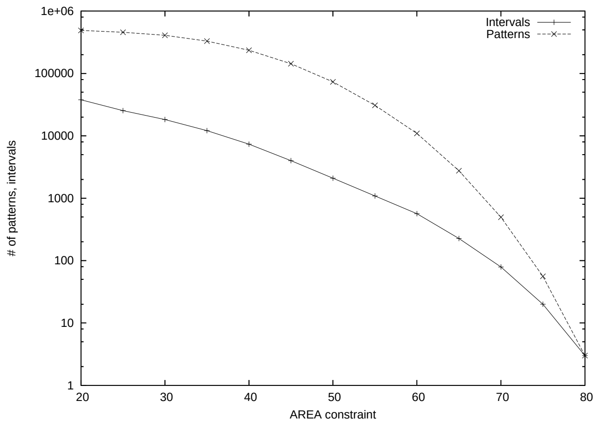

Figure 2. Selectivity of the area constraint.

with 46671 patterns that are longer than 3 and more frequent than 5. The graphs demonstrate that if both similarity ($sumsim$ or $insim$) and existence ($svsim$) are thresholded, very compact sets of patterns can be reached.

4.2 Results and their interpretation

The experimental setting started with all the large patterns that have a satisfactory average similarity among mostly known tags (see the measures $sim1(x) \geq 0.025$ and $sim3(x) \geq 0.7$ in Figure 3). It was immediately apparent that most of the extracted bi-sets were harboring genes encoding ribosomal proteins, and proteins involved in the translation process. Such a trend has already been described, although in a different dataset [7], and we therefore decided to focus on some other biological functions. We further focused on bi-sets that did not harbor ribosomal proteins. This left us with a set of 19 bi-sets that were manually inspected. On the basis of their automatic explanation, we found the following bi-set: (KHDRBS1, NONO, TOP2B, FMR1) & (48, 52, 54, 56, 62, 65). There were 74 characteristic terms adjoined to genes, 8 terms characterized the situations. It is of biological interest for these reasons:

• Three out of the four genes (KHDRBS1, NONO and FMR1) have been shown to encode proteins that display an RNA-binding activity [18, 24, 32]. The term "RNA-bind" appears in the list of terms associated with this bi-set. Of those genes, two (KHDRBS1 and NONO) have been more specifically shown to be involved in RNA splicing.

• The fourth gene (TOP2B) encodes a topoisomerase [9]. It is interesting to note that the NONO gene product was shown to have a role in DNA unwinding [24], an activity where it is known to interact functionally with Topoisomerase 1 (a member of the family to which TOP2B belongs). Moreover an isoform of TOP2B, TOP2A, has also been found differentially expressed in medulloblastoma versus normal SAGE libraries [8]. The authors also note the existence of various anticancer drugs directed against TOP2A. These drugs might have an effect on the TOP2B isoform, enhancing the anticancer effect. A topoisomerase II inhibitor was also shown to display a significant antitumor activity in a medulloblastoma xenograft [28].

• A recent paper using microarray has demonstrated the importance of RNA splicing processes for adult neurogenesis [17]. The KHDRBS1 gene was found in this study among the genes important for adult neural stem cells.

• All of the situations in which these genes are overexpressed (48, 52, etc.) are medulloblastomas. These are very aggressive brain tumors in children. There is an increasing body of evidence that the most aggressive cells within a medulloblastoma behave as brain stem cells [6, 11].

Altogether the biological hypothesis that can be made from this bi-set is as follows: RNA binding in general and RNA splicing in particular, somehow connected with genomic DNA conformation via TOP2B, is as essential for medulloblastomas as it is for normal adult nervous stem cells. Targeting this RNA binding activity, might prove beneficial for medulloblastoma treatment, just like topoisomerase II inhibition was shown to be.

5 Conclusion

The paper describes a flexible environment for analysis of plausible biological patterns. It is computationally effective and enables interactive mining that exploits available background knowledge. The paper demonstrates a practical example of interaction that proved to be able to generate a new biological hypothesis that can be tested experimentally, and that may have clinical implications.

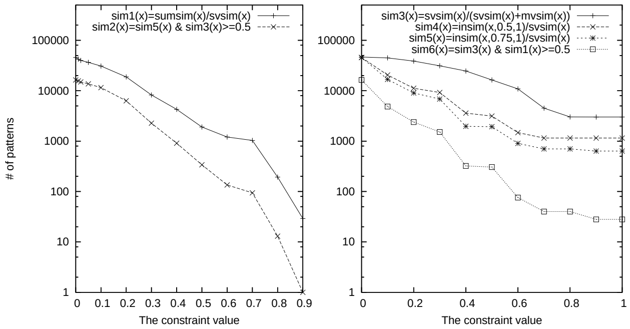

Figure 3. Pruning by the external constraints.

Future work covers several possible areas. A larger text corpus as well as a more sophisticated textual representation can be used. Switching from abstracts to full texts and employment of lexical parsing or a graph representation of texts may improve the notion of similarity. The direct textual constraints as e.g., annotation(x)<>'ribosomal' can help to avoid a specific group of tags that may flood the pattern set being out of scope of the biological interest. Fault tolerant patterns that accept a small portion of zeroes in the binarized over-expression rectangles could help to deal with noise and avoid breaking the natural large patterns. Clustering of patterns can ease the biological analysis.

Acknowledgements

This work has been supported by the French Ministry of Research project BINGO. We thank Céline Keime for her help in identifying tags and for helpful discussions.

References

[1] GOToolBox website: http://crfb.univ-mrs.fr/gotoolbox/.

[2] MatchMiner website: http://discover.nci.nih.gov/matchminer/.

[3] NCBI website: http://www.ncbi.nlm.nih.gov/.

[4] Porter stemmer website: http://www.tartarus.org/∼martin/porterstemmer/.

[5] UCL website: http://www.gene.ucl.ac.uk/nomenclature/data/gdlw_index.html.

[6] M. Al-Hajj and M. F. Clarke. Self-renewal and solid tumor stem cells. *Oncogene*, 23:7274–7282, 2004.

[7] C. Becquet, S. Blachon, B. Jeudy, J.-F. Boulicaut, and O. Gandrillon. Strong association rule mining for large gene expression data analysis: A case study on human SAGE data. *Genome Biology*, 3(12):16 pages, 2002.

[8] K. Boon, J. B. Edwards, I. M. Siu, and et al. Comparison of medulloblastoma and normal neural transcriptomes identifies a restricted set of activated genes. *Oncogene*, 23:7687–7694, 2003.

[9] J. J. Champoux. DNA topoisomerases: structure, function, and mechanism. *Annu Rev Biochem*, 70:369–413, 2001.

[10] D. Chaussabel and A. Sher. Mining microarray expression data by literature profiling. *Genome Biology*, 3, Sept. 2002.

[11] E. A. Derrington, N. Dufay, B. B. Rudkin, and M. F. Belin. Human primitive neuroectodermal tumour cells behave as multipotent neural precursors in response to FGF2. *Annu Rev Biochem*, 17:1663–1672, 1998.

[12] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of National Academy of Science of the USA 95*, pages 14863–14868, 1998.

[13] K. Gade, J. Wang, and G. Karypis. Efficient closed pattern mining in the presence of tough block constraints. In W. Kim, R. Kohavi, J. Gehrke, and W. DuMouchel, editors, *KDD*, pages 138–147. ACM, 2004.

[14] P. Glenisson, B. Coessens, S. Van Vooren, J. Mathys, Y. Moreau, and B. De Moor. TXTGate: profiling gene groups with text-based information. *Genome Biology*, 5(6):R43, 28 May 2004.

[15] P. Glenisson, J. Mathys, and B. D. Moor. Meta-clustering of gene expression data and literature-based information. *SIGKDD Explor. Newsl.*, 5(2):101–112, 2003.

[16] C. Keime, F. Damiola, D. Mouchiroud, L. Duret, and O. Gandrillon. Identitag, a relational database for SAGE tag identification and interspecies comparison of SAGE libraries. *BMC Bioinformatics*, 5:143, October 2004.

[17] D. A. Lim, M. Suarez-Farinas, F. Naef, and et al. In vivo transcriptional profile analysis reveals RNA splicing and chromatin remodeling as prominent processes for adult neurogenesis. *Mol Cell Neurosci*, 31:131–148, 2006.

[18] K. E. Lukong and S. Richard. Sam68, the KH domain-containing superSTAR. *Biochim Biophys Acta*, 1653:73–86, 2003.

[19] D. Martin, C. Brun, E. Remy, P. Mouren, D. Thieffry, and B. Jacq. GOToolBox : functional investigation of gene datasets based on gene ontology. *Genome Biology*, 5(12):R101, 26 Nov. 2004.

[20] F. Pan, G. Cong, A. K. H. Tung, Y. Yang, and M. J. Zaki. CARPENTER: finding closed patterns in long biological datasets. In *In 9th ACM SIGKDD KDD conf.*, pages 637–642, Washington, DC, USA, 2003. ACM Press.

[21] R. G. Pensa, J. Besson, and J.-F. Boulicaut. A methodology for biologically relevant pattern discovery from gene expression data. In *Discovery Science*, pages 230–241, 2004.

[22] F. Rioult, J. F. Boulicaut, B. Crémilleux, and J. Besson. Using transposition for pattern discovery from microarray data. In *8th ACM SIGMOD DMKD Workshop*, pages 73–79, San Diego, CA, 2003.

[23] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing Management*, 24(5):513–523, 1988.

[24] Y. Shav-Tal and D. Zipori. PSF and p54(nrb)/NonO–multi-functional nuclear proteins. *FEBS Lett*, 531:109–114, 2002.

[25] Q. Sheng, Y. Moreau, and B. De Moor. Biclustering microarray data by Gibbs sampling. *Bioinformatics*, 19:II:196–205, 2003.

[26] A. Soulet and B. Crémilleux. An efficient framework for mining flexible constraints. In *PAKDD, LNCS 3518, Springer*, pages 661–671, 2005.

[27] N. Tiffin, J. F. Kelso, A. R. Powell, H. Pan, V. B. Bajic, and W. A. Hide. Integration of text- and data-mining using ontologies successfully selects disease gene candidates. *Nucleic Acids Res*, 33(5):1544–1552, 2005.

[28] G. Vassal, J. L. Merlin, M. J. Terrier-Lacombe, and et al. In vivo antitumor activity of S16020, a topoisomerase II inhibitor, and doxorubicin against human brain tumor xenografts. *Cancer Chemother Pharma*, 51:385–394, 2003.

[29] V. Velculescu, S. Madden, L. Zhang, and et al. Analysis of human transcriptomes. *Nat. Genet.*, 23:387–8, 1999.

[30] V. Velculescu, L. Zhang, B. Vogelstein, and K. Kinzler. Serial analysis of gene expression. *Science*, 270:484–7, 1995.

[31] C. J. Wu and S. Kasif. GEMS: a web server for biclustering analysis of expression data. *Nucleic Acids Res.*, 1:33:596–9, July 2005.

[32] F. Zalfa and C. Bagni. Molecular insights into mental retardation: multiple functions for the fragile X mental retardation protein? *Curr Issues Mol Biol*, 6:73–88, 2004.

[33] F. Zelezny, J. Tolar, N. Lavrac, and O. Stepankova. Relational subgroup discovery for gene expression data mining. In *EMBEC: 3rd IFMBE European Medical & Biological Engineering Conf.*, November 2005.