Constraint-Based Knowledge Discovery from SAGE Data

Jiří Kléma[ac], Sylvain Blachon[b], Arnaud Soulet[d], Bruno Crémilleux[a] and Olivier Gandrillon[*b]

[a]GREYC, CNRS UMR 6072, Université de Caen,Campus Côte de Nacre, F-14032 Caen Cédex France
[b]Université de Lyon, Lyon, F-69003, France; Université Lyon 1, Lyon, F-69003, France;
Centre de Génétique Moléculaire et Cellulaire, CNRS UMR 5534, F-69622 Villeurbanne Cédex France
[c]Department of Cybernetics, Czech Technical University in Prague, Technická 2, Prague 166 27, Czech Republic
[d]LI, Université de Tours, 3 place Jean Jaure's, F-41029 Blois France

Email: Jiří Kléma - klema@labe.felk.cvut.cz; Sylvain Blachon - blachon@cgmc.univ-lyon1.fr; Arnaud Soulet - arnaud.soulet@univ-tours.fr; Bruno Crémilleux - bruno.cremilleux@info.unicaen.fr; Olivier Gandrillon*- gandrillon@cgmc.univ-lyon1.fr, tel.: (33) 4-72-44-81-90, fax: (33) 4-72-43-26-85;

*Corresponding author

Abstract

Current analyses of co-expressed genes are often based on global approaches such as clustering or bi-clustering. An alternative way is to employ local methods and search for patterns – sets of genes displaying specific expression properties in a set of situations. The main bottleneck of this type of analysis is twofold – computational costs and an overwhelming number of candidate patterns which can hardly be further exploited. A timely application of background knowledge available in literature databases, biological ontologies and other sources can help to focus on the most plausible patterns only. The paper proposes, implements and tests a flexible constraint-based framework that enables the effective mining and representation of meaningful over-expression patterns representing intrinsic associations among genes and biological situations. The framework can be simultaneously applied to a wide spectrum of genomic data and we demonstrate that it allows to generate new biological hypotheses with clinical implications.

Keywords: Functional genomics, SAGE, local pattern, background knowledge, gene ontology, biomedical literature, constraint.

In Silico Biology 8, 0014 (2008)

Introduction

The generation of very large gene expression databases by high-throughput technologies like microrarray [1] or SAGE [2] calls for similarly high-throughput exploration of possible functional links between genes and gene products. The link analysis is based upon similar expression properties, as well as possible relationships between co-expression patterns and sub-sets of biological situations.

Various techniques have been used for exploring such relationships, including global techniques like hierarchical clustering or K-means, or local pattern extraction such as association rule discovery (ARD) [3–6] or formal concepts [7].

Local patterns are groups of genes that harbor a specific expression property which can be over- or under-expression, either related to a single baseline (more/less expressed in situation A than in situation B) or related to the gene expression regime across multiple situations (more/less expressed in situation A than across multiple other situations). They provide the biologist with a list of genes that, through the "guilt by association" hypothesis [8], are supposed to vary together due to a genuine biological principle, such as common function within the cell.

Extraction of local patterns is justified by the limitations of the global methods (see [3]) as well as by the need to explore gene-to-gene relationships that would be too subtle (i.e. occurring in too small a number of situations, or in very heterogeneous situations) for detection by global approaches.

One of the main drawbacks of every local pattern approach is the huge number of extracted patterns. This is especially true in noisy data, such as transcriptomic data. At least three research directions can be explored for solving this problem. The first one relies upon the use of fault-tolerant pattern extraction (see e.g. [9]) – a difficult task whose tractability is to date still uncertain. The second direction tries to regroup patterns through hierarchical clustering [10]. In this paper, we propose a third solution using external sources to introduce constraints that focus on the most meaningful patterns. Different types of sources can be used, including Gene Ontology and literature-based evidence extracted through text-mining.

A *constraint* is a function evaluating whether a pattern is interesting, and can be used to streamline the pattern search. Gene expression data represent a new challenge for constraint-based pattern mining since the overall complexity of exhaustive pattern search is exponential with the number of genes (i.e., items) which itself is typically large. A simple approach can be decomposed into two distinct steps. Firstly, to mine all potentially interesting patterns satisfying an anti-monotone constraint (e.g., the usual constraint of minimum frequency) because this class of constraints can be efficiently pushed (to eliminate irrelevant itemsets/sets of genes early and minimize the number of itemsets to be examined). Secondly, to filter the resulting set of patterns by the remaining constraints. However, this naïve filtering approach performed by an ordinary level-wise algorithm is intractable due to the huge number of patterns [11]. Existing scalable techniques [12, 13] are limited to particular kinds of constraints (closed patterns, δ-free patterns). Integration of arbitrary background knowledge in the mining process in order to focus on the most plausible patterns requires more powerful data mining techniques.

Background knowledge is available in relational and literature databases, ontological descriptions and other sources. Its effective use in analysis and interpretation of expression data is a popular research topic

nowadays. However, the main effort is aimed at clustering and consequent integration of biological knowledge into the statistical data analysis framework. Background knowledge is typically used to annotate the expression based clusters for statistically over-represented (or under-represented) terms or categories [14, 15]. The same knowledge can also be employed to directly cluster genes [16] or to perform meta-clustering on pre-merged expression and external datasets [17]. Among the approaches distinct from clustering, [18] converts gene annotations into relational logic features, while [19] uses text mining to filter the most promising disease gene candidates. Recently an ARD-based approach has been proposed in which the authors search for associations among several data sources based on co-occurrence [20]. The resulting rules express e.g. a relation between a metabolic pathway and gene over(under)-expression in a group of biological conditions.

In this paper we introduce and apply a more general depth-first search framework which is based on a rich declarative language of *primitive-based constraints* enabling effective internal *pruning* and a condensed output representation based on *intervals*. This framework is implemented within the constraint-based pattern mining tool MUSIC (Mining with a User-SpecifIed Constraint). The first version of the tool was described in [21], this paper extends it towards utilization of external sources and the depth-first search. We demonstrate that our procedure leads to a very effective reduction of the number of patterns, together with an "interpretation" of the patterns in the form of a list of words related to the function of the genes involved in the pattern. To the best of our knowledge, there is no other constraint-based tool to efficiently discover patterns from large data under a broad set of constraints linking the information distributed in various knowledge sources. Using external constraints in the context of pattern mining as well as the integration of internal and external constraints are therefore the main contributions of this paper.

Materials and Methods

Constraint-based pattern mining through several datasets

Usual data-mining tasks rarely deal with a single dataset. Often it is necessary to connect knowledge scattered in several heterogeneous sources. In constraint-based mining, the constraints should effectively link different datasets and knowledge types. In the domain of genomics, there is a natural need to derive constraints both from expression data and descriptions of the genes and/or biological situations under consideration. Such constraints require an analysis of various data types - transcriptome data and background knowledge may be stored in the boolean, numeric, symbolic or textual format. This section presents our framework (and the declarative language) enabling the user to set flexible and meaningful constraints.

Let us consider the genomic mining context given in Figure 1. Firstly, the data involved include a boolean transcriptome dataset also called internal data where the items correspond to genes, the transactions represent biological situations and the binary values indicate gene over-expression. Secondly, external data – a similarity matrix and textual resources – are considered. They summarize background knowledge that contains various information on items (i.e., genes). This knowledge is transformed into a similarity matrix and a set of texts. Each field of the triangular matrix $s_{ij} \in [0, 1]$ gives a similarity measure between the

items i and j. The textual dataset provides a description of genes. Each row of this dataset contains a list of phrases characterizing the given gene. The mined patterns are composed of items of the internal data (the corresponding transactions are usually also added). The external data are used to further specify constraints in order to focus on meaningful patterns. In other words, the constraints may stem from all the datasets (see the example of q in Figure 1, the experimental section provides examples of other constraints). Let $\mathcal{I}$ be a set of items. A pattern is a non-empty subset of $\mathcal{I}$. $\mathcal{D}$ is a boolean matrix composed of patterns usually called transactions. The constraint-based mining task aims to discover all the patterns present in $\mathcal{D}$ and satisfying a constraint q. A pattern X is present in $\mathcal{D}$ whenever it is included in at least one transaction of $\mathcal{D}$. A distinctive point of our framework is its flexibility. Constraints are freely built of a large set of primitives representing a rich query language which allows to integrate various data/knowledge sources and to develop iteratively meaningful constraints.

Table 1 provides the meaning of the primitives involved in q and also in the other constraints used in this text. As primitives on external data are derived from different datasets, the dataset identification is another parameter of the primitive (for clarity not shown in Table 1). The first part (a) of q addresses the internal data and means that the biologist is interested in patterns having a satisfactory size (i.e., a *minimal area*). Indeed, $area(X) = freq(X) \times length(X)$ is the product of the frequency of X and its length and means that the pattern must cover a minimum number of situations and contain a minimum number of genes. The other parts deal with the external data: (b) is used to discard ribosomal patterns (one gene exception per pattern is allowed), (c) to avoid patterns with prevailing items of an unknown function and (d) to ensure a minimal average similarity. Table 1 also indicates the values of these primitives in the context of Figure 1. Our framework supports a large set of primitives, other examples of primitives with evident semantics are $\{\wedge, \vee, \neg, <, \leq, \subset, \subseteq, +, -, \times, /, sum, max, min, \cup, \cap, \setminus\}$. The only theoretical property which is required on the primitives to belong to our framework is a property of monotonicity according to each variable of a primitive [21]. The constraints of this framework are called *primitive-based constraints*. Let us recall that the primitives and the constraints defined in [21] only address one boolean data set.

The framework is by no means restricted to the similarities and textual annotations discussed above. The requirement of monotonicity allows a wide range of data sources. In the genomic domain one can also implement constraints based directly on other resources such as interaction networks or lists of transcriptional regulators.

MUSIC tool and its efficiency

We use the tool MUSIC [21, 22] which discovers soundly and completely all the patterns satisfying a given set of input constraints. The efficiency of MUSIC lies in its depth-first search strategy and a safe pruning of the pattern space by pushing the constraints. The constraints are applied as early as possible. The pruning conditions are based on intervals representing several patterns. Whenever it is computed that all the patterns included in an interval simultaneously satisfy (or not) the constraint, the interval is positively (negatively) pruned without enumerating all its patterns [21]. The output of MUSIC enumerates the intervals satisfying the constraint. Such an interval condensed representation improves the output legibility

and enables to easily compute the *selectivity* of the constraint. Selectivity is a proportion of patterns satisfying the constraint, and constitutes one of its important characteristics.

We start with the key idea of the safe pruning process based on intervals. The idea is to exploit properties of the monotonicity of the primitives on the bounds of intervals to prune them. This new kind of pruning is called *interval pruning*. Given two patterns $X \subseteq Y$, the interval $[X, Y]$ corresponds to the set $\{Z \subseteq \mathcal{I} \mid X \subseteq Z \subseteq Y\}$. Figure 2 depicts an example with the interval $[AB, ABCD]$ and the values of the primitives *sumsim* and *svsim*.

Assume the constraint $sumsim(X)/svsim(X) \geq 0.25$. As the values associated to the similarities are positive, $sumsim(X)$ is a function increasing with X. Thus $sumsim(ABCD)$ is the highest *sumsim* value for the patterns in $[AB, ABCD]$. Similarly, all the patterns of this interval have a higher $svsim(X)$ value than $svsim(AB)$. Thereby, each pattern in $[AB, ABCD]$ has its average similarity lower or equal than $sumsim(ABCD)/svsim(AB) = 0.2/1$. As this fraction does not exceed 0.25, no pattern of $[AB, ABCD]$ can satisfy the constraint and this interval can be pruned. We say that this pruning is *negative* because no pattern satisfies the constraint. In the same way, if the upper bound of the constraint on an interval $[X, Y]$ increases the threshold, all the patterns in $[X, Y]$ satisfy the constraint. $[X, Y]$ is also pruned and this pruning is named *positive*. For instance, assuming that $sumsim(AB)/svsim(ABCD) \geq 0.02$, then all the patterns in $[AB, ABCD]$ satisfy the constraint.

In a more formal way, this approach is performed by two interval pruning operators $\lfloor . \rfloor$ and $\lceil . \rceil$ introduced in [21]. The main idea of these operators is to recursively decompose the constraint to take into account the monotone properties of the primitives and then to soundly prune intervals as depicted above. This process works straightforwardly with all the primitives tackling several kinds of datasets. This highlights the generic properties of our framework. Thereby, all the parts of the constraint q are pushed into the mining step.

Let us show the usefulness of the interval pruning strategy of MUSIC. The experiment was conducted on a 2.2 GHz Pentium IV processor with Linux operating system and 3GB of RAM memory. For this purpose, we compare MUSIC with its modification that does not prune. The modification, denoted MUSIC-filter, mines all the patterns that satisfy the frequency threshold first, the other primitives are applied in the post-processing step. We use two typical constraints needed in the genomic domain and requiring the external data. These constraints and the time comparison between MUSIC and MUSIC-filter are given in Figure 3. The results show that post-processing is feasible until the frequency threshold generates reasonable pattern sets. For lower frequency thresholds, the number of patterns explodes and large intervals to be pruned appear. The interval pruning strategy decreases runtime and scales up much better than the comparative version without interval pruning and MUSIC becomes by orders of magnitude faster. MUSIC prototype is available at [23].

SAGE data

The SAGE technique aims to measure the expression levels of genes in a cell population [2]. It is performed by sequencing tags (short sequences of 14 to 21 base pairs (bps) which are theoretically specific of each

mRNA). A SAGE library is a list of transcripts expressed at one given time point in one given biological situation. Both the identity (assessed through a tag-to-gene complex process, [24]) and the amount of each transcript is recorded. Analyzing such data is relevant since this SAGE data source has been largely under-exploited as of today, although it has the immense advantage over microarrays to produce datasets that can be directly compared between libraries without the need for external normalization. The human transcriptome can be seen as libraries that would be performed in each and every biologically relevant situations in the human body. This is clearly out of reach at the moment, and we deal in the present work with 207 very different situations ranging from embryonic stem cells to foreskin primary fibroblast cells. Biologists consider that useful knowledge about the transcriptome can be expressed as sets of genes and/or sets of biological situations that have some remarkable properties. Co-regulated genes, also known as synexpression groups, based on the guilt by association approach, are assumed to participate in a common function, or module, within the cell. The 207 SAGE libraries were downloaded from the NCBI web site as of October 2004 [25]. To eliminate putative sequencing errors, a pretreatment of the data described in [3] was applied, giving a set of 125985 14 bp tags. Tags were identified thanks to Identitag [24], using RefSeq mRNA sequences. The unambiguous tags (displaying a 1 to 1 tag to RefSeq relationship) 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 [26]. It is based on 447 tags found and we refer to it as the minimum transcriptome (expression) matrix. To apply efficient local set pattern mining techniques on expression data, we must first identify and encode a specific gene expression properties (in principle, several properties per gene could be encoded, e.g., over-expression and under-expression). In this work, we decided to focus on over-expression. Thus if a gene is over-expressed in a situation then there will be a 1 value in the corresponding Boolean matrix cell, otherwise the value is 0. Both the matrices were binarized to encode the over-expression of each tag using the MidRange method described in [3]. For a thorough discussion upon the impact of discretization see [10, 27].

Background knowledge

The section on constraint-based pattern mining introduces two principal kinds of external datasets, similarity matrices and textual files. The following three sections formalize the way in which they may be built. We use two principal external data sources, freetexts and gene ontologies (GOs), and preprocess them into the external datasets. In the area of freetexts we have been inspired mainly by [16, 17]. 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 [15], the gene similarity results from the genes' positions in the molecular functional, biological process or cellular component ontology.
However, alternative sources can also be used, e.g., [28] suggests an approach to discover links between entities in biological databases. Information extracted from available databases is represented as a graph, where vertices correspond to entities and edges represent annotated relationships among vertices. A link is manifested as a path or a sub-graph connecting the corresponding vertices. Link goodness is based on edge

reliability, relevance and rarity. Obviously, the graph itself or a corresponding similarity matrix based on the link goodness can serve as an external knowledge source.

Texts and their preprocessing

To access the gene annotation data for every tag considered, RefSeq identifiers were translated into EntrezGene identifiers [29]. 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 EntrezGene database [25] and sequentially parsed by the method stemming from [18]. 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* [30]. The particular vocabulary consisted of all the *stemmed* terms [31] that appear in 5 different gene records at least. The most frequent terms were manually checked and insufficiently precise 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) [30] 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 because of synonyms, it is feasible and surprisingly often faithful.

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. [15] defines a distance based on the Czekanowski-Dice formula, the methodology is implemented within the GOProxy tool of GOToolBox [32].

The original RefSeq tag identifiers were translated into UniProt ids [33]. Out of 11082 tags there were 7670 known ids. As this set is too large to be processed by GOToolBox we confined the analysis 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 whereas 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.

Description of libraries

There is a short textual annotation of 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.

This similarity matrix does not refer to items but transactions. The constraints are not inferred from it immediately but the matrix can be used in the latest phase of pattern annotation or filtration when the focus is on the most homogeneous transaction sets only.

Results

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, 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 4 shows that there is a statistically significant correlation among all the considered datasets. 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.

Altogether this demonstrates the potential utility of using external sources for applying constraints, since all data sets are neither fully redundant, nor entirely disconnected.

How many patterns are statistically relevant?

One obvious source for noise in transcriptomic data lies within the experimental limitations of the techniques used. For example, SAGE is by essence a pooling strategy, and it has obvious limitations, especially for low to medium-sequenced libraries. Second, there is an intrinsic biological variation in the expression level of genes that has to be dealt with. Third, the binarization strategy cuts the expression

values at a given threshold. Along with the use of formal concepts for generating patterns it can amplify the original experimental noise [34]. We therefore wanted to estimate the amount of patterns that were spurious, i.e., occurring randomly.

We generated 10 (pseudo)random datasets having the same properties as the original SAGE data: the same size (11082×207), the same density (the number of 1s is 53511) and the same gene frequencies. The gene frequencies are very uneven, some of the genes are over-expressed in one situation only, others can be over-expressed in tens of situations. The roulette wheel technique [35] was used to keep the original gene frequencies. The generated datasets were searched for patterns of large areas. As the genes are mutually independent, all of the patterns are necessarily spurious. Figure 5 shows their mean number as the function of the area and compares it with the number of patterns in the real dataset. The experiment proved that the random datasets contain no (spurious) patterns longer than 3 and more frequent than 5. The first spurious patterns (2.1±0.9) tend to appear when the frequency threshold is decreased by one, i.e., the constraints are $length \geq 4$, $freq \geq 5$ and thus $area \geq 20$. These patterns contain exclusively the most frequent genes. In the real dataset we observe 490267 patterns satisfying the same constraints. The experiment suggests that we may encounter at least about half a million non-random and thus large patterns.

The number of spurious patterns can also be theoretically estimated. Under assumption of gene independence and considering the prior frequency of genes, the probability that the pattern occurs at random is given by the multidimensional hypergeometric distribution:

$$p_s = \prod_{i=1}^{l} \frac{\binom{k_i}{f}\binom{m-k_i}{f-f}}{\binom{m}{f}} = \prod_{i=1}^{l} \frac{\binom{k_i}{f}}{\binom{m}{f}}$$

where l is the pattern length, f is the pattern frequency, m is the total number of situations and k_i is the frequency of i-th gene contained in the pattern. The probability p_s concerns specific biological context, i.e., it gives the chance that the pattern appears in a single set of situations. The total spurious occurrence of the pattern can be estimated as follows:

$$n_s = \binom{m}{f} p_s = \binom{m}{f} \prod_{i=1}^{l} \frac{\binom{k_i}{f}}{\binom{m}{f}} = \frac{\prod_{i=1}^{l} \binom{k_i}{f}}{\binom{m}{f}^{l-1}}$$

The more real pattern occurrence exceeds n_s or the smaller its p_s, the more surprising and interesting pattern. The patterns of small area based on non-frequent genes can prove to be more interesting than their larger counterparts composed of the frequent genes. Consequently, the best internal constraint would be based on n_s or p_s respectively. However, this constraint is difficult to calculate repeatedly during the pruning process. We have introduced it mainly to show that we deal with a large number of potentially meaningful patterns and they can be found even among patterns of a limited area.

The theoretical analysis confirmed that the final number of large patterns is even larger than mentioned in the experimental paragraph. Taken together these results clearly establish that the immense majority of the patterns that were generated could not by any means be attributed to noise, and have to be considered

as potential source of biologically-relevant information. As the biologists prefer output sets with tens of patterns at most, one of the main tasks is to make this large number of potentially relevant patterns accessible to the expert in a friendly and interactive way.

Internal and external constraints to reach a meaningful limited pattern set

Since we have to deal with an explosion of putatively interesting relevant patterns, we tried to estimate the impact of applying various constraints during the extraction process. Traditional pattern mining deals with constraints that we refer to as internal. Their characteristic property is that they are inferred from the mined dataset. In our case, it is the binarized expression dataset. The main goal is usually to identify the itemsets (the sets of genes) that tend to co-occur frequently. The larger the itemsets, i.e., the more genes they contain, the better. Speaking of patterns, the most meaningful internal constraint regards their area, i.e., product of size/number of genes and frequency. It can be also understood as the number of ones that the pattern covers in the binarized expression dataset. Subjectively, the large patterns can be simply all the patterns that are larger than a certain threshold. However, we will define them as all the patterns that are large enough not to be spurious, i.e., occurring randomly. The goal is to find the optimal area threshold to distinguish between spurious and meaningful patterns.

Figure 5 shows how many patterns and intervals satisfy the increasing area constraint. In order to reduce the number of extracted patterns, the minimum pattern length was set to 4 and frequency to 5. Even using such a constraint, the number of patterns above a given area were still 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 parameters generate rather sparse matrices. For other binarization types the explosion of patterns can be even faster.

There are two straightforward ways to treat the explosion of patterns. Firstly, one may try to focus on very large patterns only and increase the value of the area constraint. It is easy to show that this approach is rather counter-productive. The previous subsection on statistical pattern relevance clearly expresses that the more frequent genes are more likely to form very large patterns. In practice, the increase of the area threshold in order to get a reasonable number of patterns leads to a small but uniform set that is flooded by the ribosomal genes which represent the most frequent genes in our dataset. Biologists rated these patterns as valid but since they were found earlier ([3]) they chose to discard them. Apparently, the area constraint helps to distinguish between spurious and real (random and non-random) patterns, but it does not hold that the larger the better. The pattern reduction by means of a stronger area restriction is unsound.

The second way relies upon a condensed representation of patterns. Comprehensibility increases as the human expert deals with fewer and more compact condensed sets of similar patterns. As the patterns tend to "overlap" greatly, let us try to test how far they can be condensed. Let us define the *maximal pattern* as such a pattern that no other pattern that satisfies constraints is its super-set. For example, having the set of patterns

$s = \{b_1 = \{\{A, B, C\}, \{1, 2, 3\}\}, b_2 = \{\{A\}, \{2\}\}, b_3 = \{\{A, B\}, \{1, 2, 3\}\}, b_4 = \{\{A, C\}, \{1, 2, 4\}\}\}$, b_1 and

b_4 are the maximal patterns while b_2 is a subset of b_1 and b_4, b_3 is a subset of b_1. Let us search for all the non-trivial patterns having area$\geq$24. The search results in 46671 different patterns which can be condensed into 2274 maximal patterns. It is fewer than the original number, but still too high to be manually inspected. Moreover, this maximal representation is incomplete and the original set of patterns cannot be restored from it. The interval condensed representation generated by MUSIC is complete, the number of intervals is usually higher than the number of maximal patterns (in this case we would have 9335 intervals). Fundamentally different representation is a hierarchy of patterns [10]. The hierarchy is a result of clustering, whose partitions can speed up orientation among patterns, however, their number has to be decreased by external constraints again before the clustering is started. To sum up, the usual condensed representations of patterns are still too extensive to be surveyed by humans.

The previous paragraphs explain the motivation for using background knowledge to formalize constraints. It has been experimentally proven that the number of large patterns is so high that they cannot be effectively surveyed by a human expert. Simultaneous application of internal and external constraints, such as interestingness or expressiveness, may help to further reduce the patterns while keeping the interesting ones. The selectivity of selected external constraints is shown in Figure 6. They capture the amount of similarity in given patterns through the measurement of the similarity of all tags pairs within that given pattern. $sumsim(x)/svsim(x)$ expresses 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. The pruning starts with 46671 patterns that are larger than 3 genes and more frequent than 5 libraries. The graphs depict that if both similarity (sumsim or insim) and existence (svsim) are thresholded, very compact sets of patterns can be reached. The next section gives a demonstration that these sets also gather biologically meaningful patterns.

Biological interpretation of patterns

The experimental setting started with all the large patterns that have a satisfactory average textual similarity among mostly known tags (see the measures $sim1(x) \geq 0.025$ and $sim3(x) \geq 0.7$ in Figure 6. It was immediately apparent that most of the extracted patterns 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 [3], and we therefore decided to focus on some other biological functions. We further focused on patterns that did not harbor ribosomal proteins. This left us with a set of 19 patterns that were manually inspected. On the basis of their automatic explanation, we found the following pattern: B1={(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 [36–38]. The term "RNA-bind" appears in the list of terms associated with this pattern. 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 [39]. It is interesting to note that the NONO gene product was shown to have a role in DNA unwinding [37], 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 [40]. 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 [41].
• A recent paper using a microarray has demonstrated the importance of RNA splicing processes for adult neurogenesis [42]. 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 over-expressed (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 [43, 44].
Altogether the biological hypothesis that can be made from this pattern 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 nervous system stem cells. Targeting this RNA binding activity, might prove beneficial for medulloblastoma treatment, just as topoisomerase II inhibition has proven to be.
We then tried to assess the efficiency of using the GO-based external knowledge (annotations plus similarity), instead of the text-based one. We have constructed on principle a similar constraint to that mentioned at the beginning of this section. It is very interesting to note that the very same pattern that we previously analyzed (B1) was also found using this constraint. This clearly illustrated the level of redundancy that we previously described (see Figure 7) and it demonstrates that some patterns are very robust.
We then focused on the following pattern: B2={(EIF3S5, MRPL23, RPL18, EEF1G) & (6, 30, 31, 116, 150, 171)}. This pattern is very homogeneous in term of the function of the genes since all of the genes participate to the translation machinery: EIF3S5 encodes the eukaryotic translation initiation factor 3, MRPL23 encodes the mitochondrial ribosomal protein L23, RPL18 encodes the ribosomal protein L18 and EEF1G encodes the eukaryotic translation elongation factor 1 gamma. It is interesting to note that at the time we built our dataset, the gene MRPL23 had no GO record attached. Therefore, it belongs to this pattern only by virtue of its expression pattern, although it encodes a mitochondrial ribosomal protein, and therefore also participates to the same function that the rest of the genes in this pattern. It is interesting to note that, although ribosomal genes were explicitly filtered out, one nevertheless obtained a pattern displaying such homogeneous, translation-related, functions. The nature of the situations harboring this set of simultaneously over-expressed genes is very heterogeneous, although some display stem cell characteristics (fibroblast cells immortalized by telomerase over-expression, CD34+ haematopoietic stem cells), and some do not (lung normal cell line). It is therefore difficult to understand why those situations have in common an over-expression of part of their translation machinery. One should

nevertheless note that a preferential expression of translation-associated genes has just been described in murine haematopoietic stem cells [45]. In any case, this illustrates the power of local patterns to highlight gene expression patterns appearing though very different conditions, and that would not be captured by global tools like hierarchical clustering.
The example given in Figure 7 gives another evidence that background constraints can effectively reduce the number of patterns, they can express various kinds of interest and the patterns that tend to reappear are likely to be recognized as interesting by an expert.

Gene function prediction

The proposed framework clearly serves knowledge discovery and the patterns correspond to descriptive models. Contrary to predictive models such as support vector machines they do not directly classify biological samples nor explicitly assign functions to genes. In this section we demonstrate an intelligible application of patterns for gene function prediction. Its motivation is twofold. Firstly, the descriptive models are hard to evaluate objectively. One can think of the manual evaluation of patterns done in the previous subsection as data fishing. The predictive experiment provides means to objectively assess the pattern sets en bloc. Secondly, the experiment implicitly outlines one of the ways the patterns can be interpreted by the biologists. On the other hand, the experiment does not outline the way to routinely and automatically predict gene functions. It is well known (see e.g. [46]) that similar gene expression profiles do not immediately imply similar tissue functions.
Let us assume the hypothesis that there is a relationship between the functional similarity of genes and their co-occurrence in patterns. Let us suppose that we have an expression dataset that mixes genes with known and unknown functions (annotations). Under our hypothesis, patterns can be applied to predict an unknown gene function in the following manner. Having a gene g with an unknown function, all the plausible patterns containing g are mined. The function of g is likely to relate to the function of the annotated genes that appear in the same patterns as g.
Let us experimentally verify our hypothesis. Obviously, gene co-occurrence in patterns does not imply gene functional similarity logically/immediately, the implication under consideration is probabilistic. That is why the predictive experiment tests all the genes that are frequent in the given expression dataset (and likely to appear in a sufficient number of patterns) and their annotation is known (the annotation is not used during pattern mining, only to evaluate the predictions). The hypothesis holds when the tested genes show a significantly higher functional similarity within their patterns than with other genes. The experiment pseudocode is as follows:

1. $\mathbb{E} : B \times G \to \{0, 1\}$ stands for a binary expression matrix, B is a set of m biological situations, G is a set of n genes, $\mathbb{S} : G \times G \to \langle 0, 1\rangle \cup \{NA\}$ is a gene similarity matrix (derived e.g. from the gene function ontology, NA stands for the undefined/missing similarity value).

2. Find a subset of frequent and annotated genes $F \subseteq G$ such that
   F=$\{f \in G | freq(f) \geq thres \wedge \exists i \neq f : S_{fi} \neq NA\}$, where $freq(f) = \sum_{b \in B} e_{bf}$. Frequent genes are

likelier to appear in patterns, annotations are needed to make assumptions on the similarity among genes.

3. Select a minimum pattern frequency $pfreq \leq thres$.

4. For each $f \in F$ calculate the weighted mean similarity to the other genes in the expression matrix:

$$msim_f = \frac{\sum_{g \in G, freq(g) > thres}^{S_{fg} \neq NA, g \neq f} (freq(g) - pfreq) S_{fg}}{\sum_{g \in G, freq(g) > thres}^{S_{fg} \neq NA, g \neq f} (freq(g) - pfreq)}$$

5. Choose a minimum pattern area $parea \geq pfreq$. In $\mathbb{E}$ search for the set of all the large patterns $LPS \subseteq 2^G$ such that $LPS = \{P \subseteq G | freq(P) \geq pfreq \wedge area(P) \geq parea\}$, where $freq(P) = supp(P, \mathbb{E})$, $area(P) = freq(P) \times length(P)$.

6. For each $f \in F$ find a subset LPS_f of large patterns LPS that contains f: $LPS_f = \{P \in LPS | f \in P\}$. Enumerate gene occurrence in LPS_f, every single occurrence of a gene is counted. GF_f is a set of gene occurrences in LPS_f such that: $GF_f = \{(g, g_{freq}) | g \in P \in LPS_f, g \neq f, g_{freq} = |\{P \in LPS_f | g \in P\}|\}$

7. For each $f \in F$ calculate the weighted mean similarity to the genes co-occurring in the large patterns:

$$psim_f = \frac{\sum_{g \in \{(g, g_{freq}) \in GF_f\}} g_{freq} S_{fg}}{\sum_{g \in \{(g, g_{freq}) \in GF_f\}} g_{freq}}$$

8. Do a paired test between msim and psim vectors. The null hypothesis is that genes (the frequent and annotated) show no difference in their similarity to all the other genes and the genes that co-occur in their patterns. The alternative hypothesis states that the genes that co-occur in patterns tend to be more similar than randomly taken genes.

Table 2 summarizes the results for $thres = 15, area = 15, pfreq = 5$. It clearly shows that the intra-pattern functional gene similarity is significantly higher than the similarity among randomly sampled genes. The conclusion of this experiment is that the patterns actually generalize to the "unseen" cases, i.e., the patterns enable to draw attention to the function of yet unknown genes.

Discussion

The goal of our work was to enhance the applicability of local pattern discovery for specific end users, such as biologists. For this we first verified that the immense majority of local pattern generated from human SAGE dataset were not attributable to random noise. This therefore clearly reinforces the need of automatic tools for navigating among the huge amount of potentially biologically relevant local associations among genes and situations.

We then verified that the external sources like Medline and Gene Ontology were at the same time sufficiently correlated and not too redundant so that their use would provide an add on value for selecting

among the whole lists of patterns. We then applied a general filtering strategy based upon a new constraint-based mining algorithm, called MUSIC. Applying this algorithm on SAGE data could effectively lead to a very significant reduction in the amount of patterns the end user has to deal with. Furthermore, the "labeling" through lists of words rendered the selection of patterns for future exploration more easy. Since the biological interpretation of a given pattern still has to be done manually, and is very time consuming, it is critical that such patterns are presented to the end-user in a way where he/she can choose rapidly which pattern is worth further investigation.

We applied this general strategy to a gene expression dataset displaying the expression of 11082 genes in 207 different situations. We explored the patterns generated and found that some patterns are sufficiently robust to be generated through different types of constraints, either based upon GO-terms or upon text-based evidence. Compared to a recently published related work [20], our approach adheres to local patterns satisfying user-defined background properties specified by constraints. The fact that such constraints may be derived from current literature rather than through the use of an ontology makes it a more versatile tool, allowing recent evidence, available only in the literature, to be used as constraints. One pattern obtained by the use of different constraints was further explored in detail. It led to an interesting hypothesis regarding the role of RNA-binding activities in the generation and/or maintenance of medulloblastomas. Another pattern pointed toward a role for the over-expression of part of the translation machinery in heterogeneous situations. Altogether this work demonstrates the usefulness of applying external constraints, and reinforces the potential impact of automated tools for analyzing large matrices of gene expression.

The predictive experiment confirmed the hypothesis that there is a relationship between the functional similarity of genes (and their products) and their co-occurrence in patterns. As a consequence, patterns enable us to draw attention to the function (and presumably other properties) of yet unknown genes.

In summary, constraints provide a human understandable way to extract valuable knowledge from potentially large and heterogeneous data. Provided they are computationally efficient, they enable interactive knowledge discovery resulting in the user-optimal set of constraints and consequently the set of desired patterns. We demonstrate the feasibility and usefulness of such an approach.

Authors contributions

Jiří Kléma designed the general strategy, collected and processed background data, and generated the patterns. Sylvain Blachon generated the initial SAGE data. Arnaud Soulet is an author of the constraint-based tool MUSIC. Together with Jiří Kléma they designed the problem specific primitive constraints. Olivier Gandrillon and Sylvain Blachon analyzed the biological meaning of the extracted patterns. Olivier Gandrillon and Bruno Crémilleux supervised the whole work. Jiří Kléma and Olivier Gandrillon wrote the manuscript. All authors read and approved the manuscript.

Acknowledgements

This work has been supported by the ANR (French Research National Agency) project BINGO2 ANR-07-MDCO-014 which is a follow-up of the first BINGO project (2004-007). The work of Jiří Kléma was partly funded by the Czech Ministry of Education in terms of the research programme Transdisciplinary Research in the Area of Biomedical Engineering II, MSM 6840770012. Sylvain Blachon was a fellow from the Comité de Saône et Loire de la Ligue Contre le Cancer. 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). Travels were covered by Czech-French PHC Barrande project "Fusion de données hétérogenes pour la découverte de connaissances en génomique". We thank Céline Keime for her help in identifying tags and for helpful discussions, Edmund Derrington (CGMC UMR 5534) for his critical and thoughtful reading of the manuscript and all members of the BINGO2 project for stimulating discussions.

References

1. Gershon D: **Microarray technology: an array of opportunities**. *Nature* 2002, **416**:885–891.
2. Velculescu V, Zhang L, Vogelstein B, Kinzler K: **Serial Analysis of Gene Expression**. *Science* 1995, **270**:484–7.
3. Becquet C, Blachon S, Jeudy B, Boulicaut JF, Gandrillon O: **Strong Association Rule Mining for Large Gene Expression Data Analysis: A Case Study on Human** SAGE **Data**. *Genome Biology* 2002, **3**(12):16 pages.
4. Creighton C, Hanash S: **Mining gene expression databases for association rules**. *Bioinformatics* 2003, **19**:79–86.
5. Georgii E, Richter L, Ruckert U, Kramer S: **Analyzing microarray data using quantitative association rules**. *Bioinformatics* 2005, **21**(Suppl 2.):ii123–ii129.
6. Li J, Liu H, Downing J, Yeoh A, Wong L: **Simple rules underlying gene expression profiles of more than six subtypes of acute lymphoblastic leukemia (ALL) patients**. *Bioinformatics* 2003, **19**:71–78.
7. Rioult F, Robardet C, Blachon S, Crémilleux B, Gandrillon O, Boulicaut JF: **Mining Concepts from Large SAGE Gene Expression Matrices.** In *KDID*. Edited by Boulicaut JF, Dzeroski S, Rudjer Boskovic Institute, Zagreb, Croatia 2003:107–118.
8. Wolfe C, Kohane I, Butte A: **Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks.** *BMC Bioinformatics* 2005, **6**:227.
9. Besson J, Robardet C, Boulicaut JF: **Mining a new fault-tolerant pattern type as an alternative to formal concept discovery**. In *14th International Conference on Conceptual Structures (ICCS'06)*, Aalborg, Denmark: Springer-Verlag 2006.
10. Blachon S, Pensa R, Besson J, Robardet C, Boulicaut JF, Gandrillon O: **Clustering formal concepts to discover biologically relevant knowledge from gene expression data**. *In Silico Biol.* 2007, **7**:0033.
11. Jeudy B, Rioult F: **Database Transposition for Constrained (Closed) Pattern Mining**. In *Post-Proceedings of the Workshop on Knowledge Discovery in Inductive Databases*, *Volume 3377 of* Lecture Notes in Computer Science. Edited by Goethals B, Siebes A, Springer-Verlag 2005:89–107.
12. Pan F, Cong G, Tung AKH, Yang Y, Zaki MJ: **CARPENTER: finding closed patterns in long biological datasets**. In *In 9th ACM SIGKDD KDD conf.*, Washington, DC, USA: ACM Press 2003:637–642.
13. Rioult F, Boulicaut JF, Crémilleux B, Besson J: **Using transposition for pattern discovery from microarray data**. In *8th ACM SIGMOD DMKD Workshop*, San Diego, CA 2003:73–79.
14. Glenisson P, Coessens B, Van Vooren S, Mathys J, Moreau Y, De Moor B: TXTG**ate: profiling gene groups with text-based information**. *Genome Biology* 2004, **5(6)**:R43.

15. Martin D, Brun C, Remy E, Mouren P, Thieffry D, Jacq B: GOT**oolB**o**x : functional investigation of gene datasets based on Gene Ontology**. *Genome Biology* 2004, **5(12)**:R101.

16. Chaussabel D, Sher A: **Mining microarray expression data by literature profiling**. *Genome Biology* 2002, **3**.

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

18. Zelezny F, Tolar J, Lavrac N, Stepankova O: **Relational Subgroup Discovery for Gene Expression Data Mining**. In *EMBEC: 3rd IFMBE European Medical & Biological Engineering Conf.* 2005.

19. Tiffin N, Kelso JF, Powell AR, Pan H, Bajic VB, Hide WA: **Integration of text- and data-mining using ontologies successfully selects disease gene candidates**. *Nucleic Acids Res* 2005, **33**(5):1544–1552.

20. Carmona-Saez P, Chagoyen M, Rodríguez A, Trelles O, Carazo JM, Pascual-Montano AD: **Integrated analysis of gene expression by association rules discovery.** *BMC Bioinformatics* 2006, **7**:54.

21. Soulet A, Crémilleux B: **An Efficient Framework for Mining Flexible Constraints.** In *PAKDD*, *Volume 3518 of* Lecture Notes in Computer Science. Edited by Ho TB, Cheung D, Liu H, Springer 2005:661–671.

22. Soulet A, Kléma J, Crémilleux B: *Efficient Mining Under Rich Constraints Derived from Various Datasets*, Springer Berlin / Heidelberg, *Volume 4747 of* Lecture Notes in Computer Science 2007 chap. 14, :223–239. [Knowledge Discovery in Inductive Databases].

23. M**ining with a** U**ser-**S**pecifIed** C**onstraint** [http://www.sir.blois.univ-tours.fr/~soulet/music-dfs/music-dfs.html].

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

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

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

27. Pensa RG, Leschi C, Besson J, Boulicaut JF: **Assessment of discretization techniques for relevant pattern discovery from gene expression data.** In *BIOKDD*. Edited by Zaki MJ, Morishita S, Rigoutsos I 2004:24–30.

28. Sevon P, Eronen L, Hintsanen P, Kulovesi K, Toivonen H: **Link discovery in graphs derived from biological databases**. In *3rd International Workshop on Data Integration in the Life Sciences 2006 (DILS'06)*, Hinxton, UK July 2006:35–49.

29. **Match**M**iner** [http://discover.nci.nih.gov/matchminer/].

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

31. **Porter stemmer** [http://www.tartarus.org/$\sim$martin/PorterStemmer/].

32. GOT**oolB**o**x** [http://crfb.univ-mrs.fr/GOToolBox/].

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

34. Robardet C, Pensa RG, Besson J, Boulicaut JF: **Using Classification and Visualization on Pattern Databases for Gene Expression Data Analysis.** In *PaRMa*, *Volume 96 of* CEUR Workshop Proceedings. Edited by Theodoridis Y, Vassiliadis P, CEUR-WS.org 2004.

35. Baker JE: **Reducing Bias and Inefficiency in the Selection Algorithm.** In *Grefenstette, J. J. (ed.): Proceedings of the Second International Conference on Genetic Algorithms and their Application*, Hillsdale, New Jersey, USA: Lawrence Erlbaum Associates 1987:14–21.

36. Lukong KE, Richard S: **Sam68, the KH domain-containing superSTAR**. *Biochim Biophys Acta* 2003, **1653**:73–86.

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

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

39. Champoux JJ: **DNA topoisomerases: structure, function, and mechanism**. *Annu Rev Biochem* 2001, **70**:369–413.

40. Boon K, Edwards JB, Siu IM, et al: **Comparison of medulloblastoma and normal neural transcriptomes identifies a restricted set of activated genes**. *Oncogene* 2003, **23**:7687–7694.

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

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

43. Al-Hajj M, Clarke MF: **Self-renewal and solid tumor stem cells**. *Oncogene* 2004, **23**:7274–7282.

44. Derrington EA, Dufay N, Rudkin BB, Belin MF: **Human primitive neuroectodermal tumour cells behave as multipotent neural precursors in response to FGF2**. *Annu Rev Biochem* 1998, **17**:1663–1672.

45. Huttmann A, Duhrsen U, Heydarian K, Klein-Hitpass L, Boes T, Boyd A, Li C: **Gene expression profiles in murine hematopoietic stem cells revisited: analysis of cDNA libraries reveals high levels of translational and metabolic activities.** *Stem Cells* 2006, **24**:1719–1727.

46. Yanai I, Korbel J, Boue S, McWeeney S, Bork P, Lercher M: **Similar gene expression profiles do not imply similar tissue functions.** *Trends Genet* 2006, **22**:132–138.

Figures

Figure 1 - An overview of constraint-based mining through several heterogeneous datasets

A toy example of the mining context along with a possible constraint. The figure shows various data types addressed by various sets of primitives. The constraint q addresses the large patterns (a) which are not composed of more than one ribosomal gene (b) and contain mainly annotated genes (c) with a minimal average similarity (d). The primitives are detailed in Table 1. The overall process can be viewed as a simultaneous query on data and on patterns. The combination of the primitive constraints can be therefore seen as an inductive query.

Figure 2 - Illustration of the interval pruning

The figure depicts an example of a pruning applied to the interval $[AB, ABCD]$. The pruning is exemplified with values of the primitives *sumsim* and *svsim*. The key idea is to exploit properties of the monotonicity of the primitives on the bounds of intervals. Whole intervals can be pruned at once.

Figure 3 - Efficiency of the interval pruning

The efficiency of interval pruning with decreasing frequency primitive threshold is shown. The left image deals with the constraint

$freq(X) \geq thres \wedge lenght(X) \geq 4 \wedge sumsim(X)/svsim(X) \geq 0.9 \wedge svsim(X)/(svsim(X) + mvsim(X)) \geq 0.9$. The right image deals with the constraint $freq(X) \geq thres \wedge length(regexp(X, '*ribosom*', \texttt{GO\_terms})) = 0$.

Figure 4 - Correlations among the datasets

The degree of correlation among the considered datasets. Similarity among gene profiles (or profiles of biological situations) is calculated within the individual datasets first. Then, the correlation between similarity matrices is determined. The higher the correlation between two datasets the more they agree in gene similarity. This experiment was performed on the minimum transcriptome matrix (207x447).

Figure 5 - Selectivity of the area constraint

The number of patterns larger than the given area. This experiment was performed on the complete 207x11082 matrix.

Figure 6 - Pattern pruning by the external constraints

Simultaneous application of internal and external constraints helps to arbitrarily reduce the number of patterns while attempting to conserve the potentially interesting ones. The figures show the decreasing number of patterns with increasing threshold of selected external constraints. The effect of six different constraints of various complexity is shown. This experiment was performed on the complete 207x11082 matrix.

Figure 7 - Demonstration of selectivity and possible overlap among various constraints

The gradual reduction of patterns by background constraint is shown. The individual constraints are applied in conjunction. The figure demonstrates that background constraint can effectively reduce the number of patterns, it can define various domains of interest and the patterns that emerge are likely to be recognized as interesting by an expert. The example demonstrates three different ways to obtain a concise output that can be easily surveyed by a human because it consists of 9, 2 or 5 patterns only. An interesting observation is that the pattern that was identified by the expert as one of the "nuggets" (shown at the bottom of the image) can be obtained by several alternative ways. The first way uses NCBI textual resources (gene summaries and adjoined PubMed abstracts), the second way relies only on functional GO, while the third way utilizes similarities among biological situations too. Note that syntactically identical constraints aiming at textual and GO resources result in output of different quantity (3881 vs. 1633 patterns). Considering the datasets of different origin but the same format and purpose, the expert can decide whether to use them independently, unify or intersect them during pre-processing or via constraints. These experiments were performed on the complete 207x11082 matrix.

Tables

Table 1 - Examples of primitives and their values in the data mining context of Figure 1

Table provides the meaning of primitives as well as their values in the context of Figure 1.

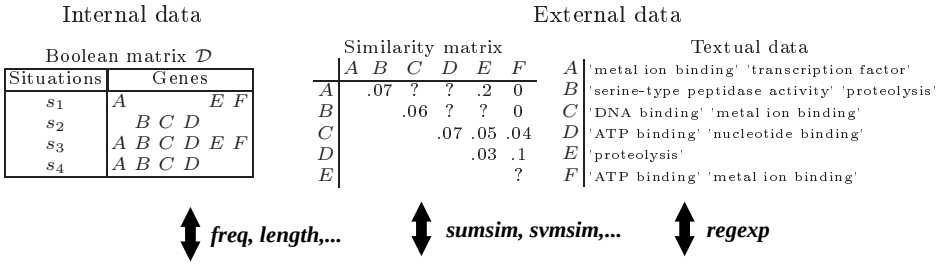
primitives		values
Boolean matrix		
$freq(X)$	frequency of X	$freq(ABC) = 2$
$length(X)$	length of X	$length(ABC) = 3$
Textual data		
$regexp(X, RE)$	items of X whose associated phrases match the regular expression RE	$regexp(ABC, '*ion*') = AC$
Similarity matrix		
$sumsim(X)$	similarity sum over the set of item pairs of X	$sumsim(ABC) = 0.13$
$svsim(X)$	number of stated item pairs belonging to X	$svsim(ABC) = 2$
$mvsim(X)$	number of missing item pairs belonging to X	$mvsim(ABC) = 1$
$insim(X, min, max)$	number of item pairs whose similarity lies between min and max	$insim(ABC, 0.07, 1) = 1$

Table 2 - Relation between gene co-occurrence in patterns and their similarity (in terms of molecular function and biological process)

Table shows mean similarity among genes. It averages over all the genes that are frequent enough (the expression matrix) and annotated (the similarity matrix). The value of $msim$ estimates the similarity regardless patterns (it postulates that patterns do not correlate with gene annotation at all). The value of $psim$ gives an estimate of the real gene similarity withinside patterns. The first row considers the similarity in terms of the molecular function, the second row concerns the biological process. The similarity is derived of the respective GO annotations. F is a number of the frequent and annotated genes, $+/0/-$ give numbers of genes out of F whose $msim < psim / msim = psim / msim > psim$.

Annotation type	F	$+/0/-$	$\overline{msim}$	$\overline{psim}$	p-value (paired t-test)
molecular function	290	137/35/118	0.27	0.31	1.8E-7
biological process	274	135/33/106	0.32	0.36	2.4E-8

Figure 1



$$
\begin{aligned}
q(X) \equiv\ & freq(X) \times length(X) \geq 24 && \text{(a)}\\
& \wedge length(regexp(X, '*ribosom*', \texttt{TEXT\_terms})) \leq 1 && \text{(b)}\\
& \wedge svsim(X, \texttt{TEXT})/(svsim(X, \texttt{TEXT}) + mvsim(X, \texttt{TEXT})) \geq 0.7 && \text{(c)}\\
& \wedge sumsim(X, \texttt{TEXT})/svsim(X, \texttt{TEXT}) \geq 0.025 && \text{(d)}
\end{aligned}
$$

Figure 2

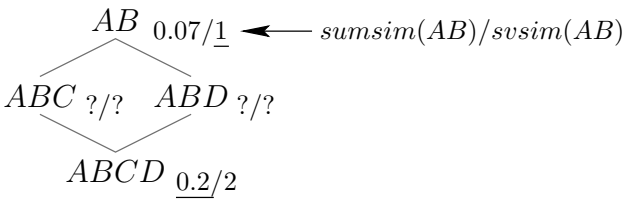

AB $0.07/\underline{1}$
$sumsim(AB)/svsim(AB)$
ABC ?/?
ABD ?/?
$ABCD$ $\underline{0.2}/2$

Figure 3

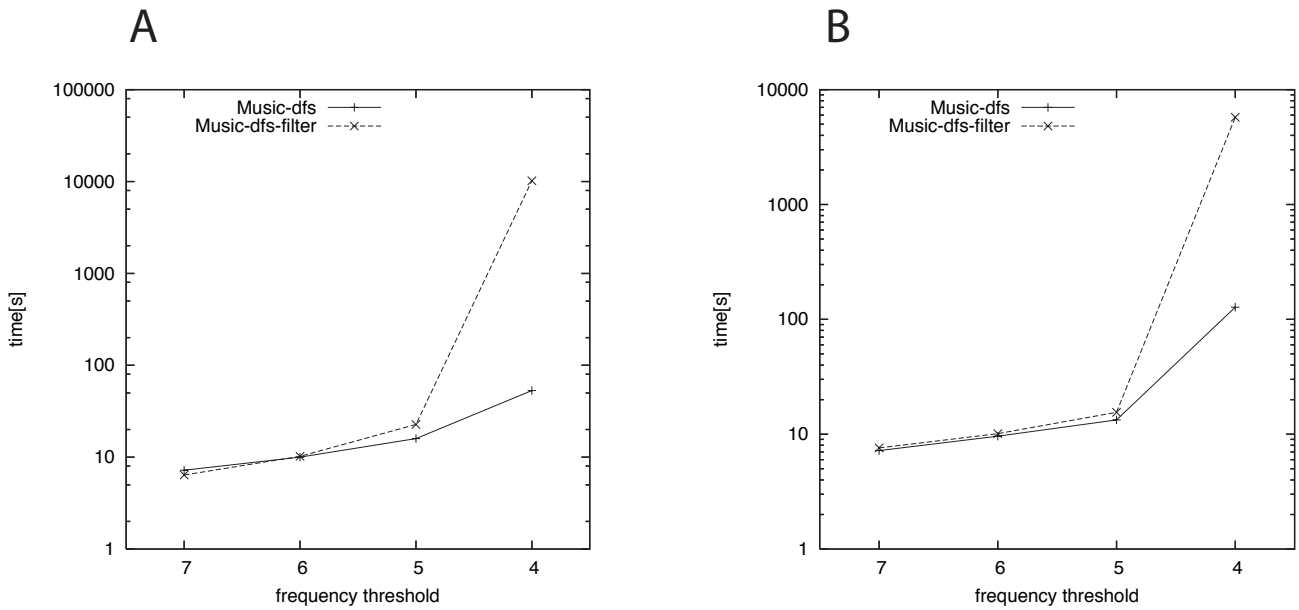

A
B
Music-dfs
Music-dfs-filter
time[s]
frequency threshold
100000
10000
1000
100
10
1
7
6
5
4

Figure 4

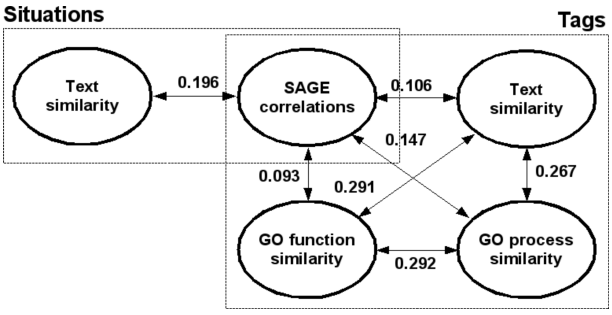

Situations
Tags
Text similarity
0.196
SAGE correlations
0.106
Text similarity
0.147
0.093
0.291
0.267
GO function similarity
0.292
GO process similarity

Figure 5

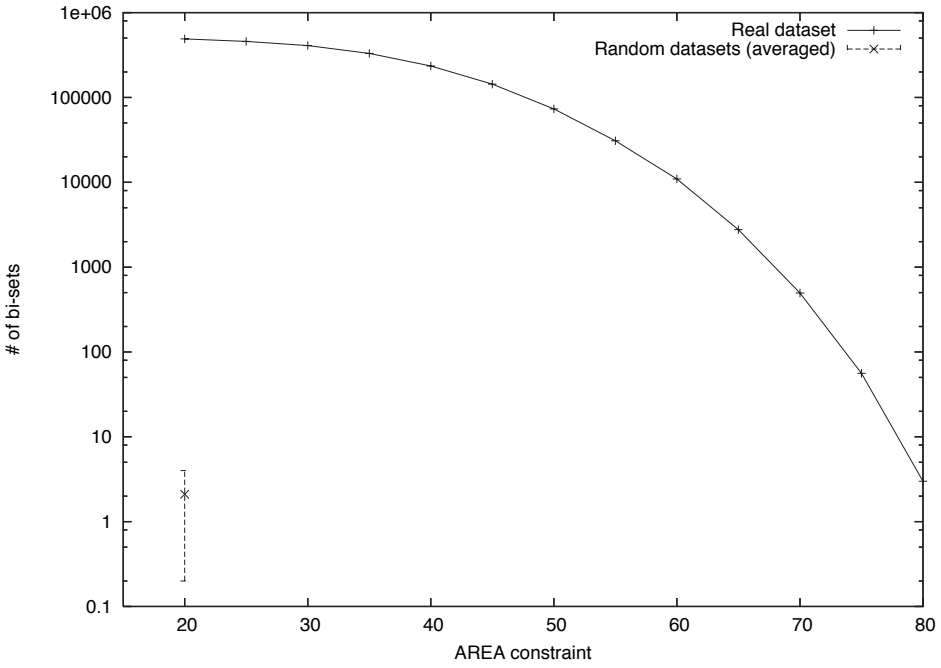

Real dataset
Random datasets (averaged)
# of bi-sets
AREA constraint
1e+06
100000
10000
1000
100
10
1
0.1
20
30
40
50
60
70
80

Figure 6

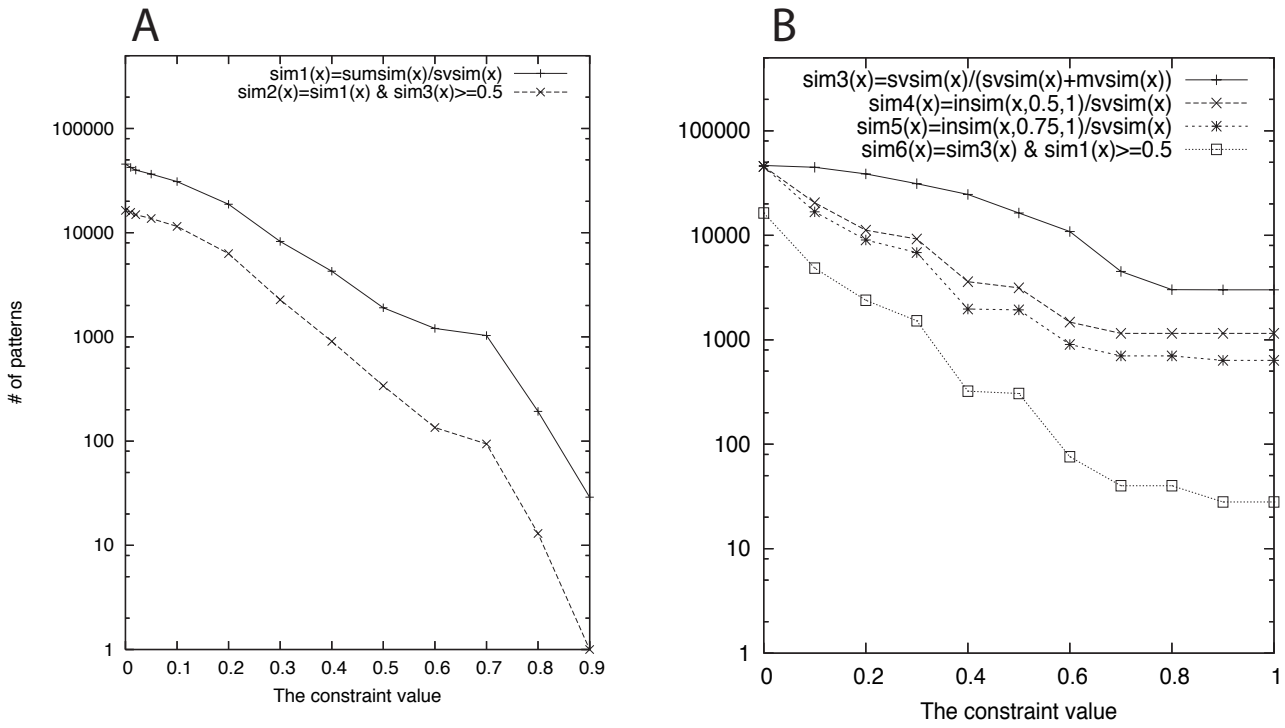
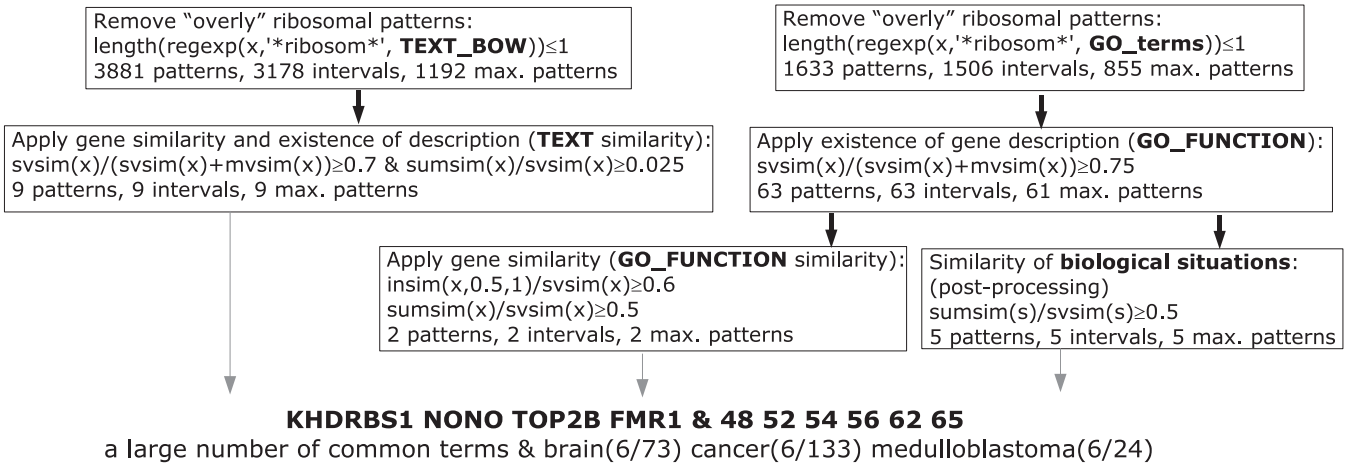

A
sim1(x)=sumsim(x)/svsim(x)
sim2(x)=sim1(x) & sim3(x)>=0.5
# of patterns
100000
10000
1000
100
10
1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
The constraint value
B
sim3(x)=svsim(x)/(svsim(x)+mvsim(x))
sim4(x)=insim(x,0.5,1)/svsim(x)
sim5(x)=insim(x,0.75,1)/svsim(x)
sim6(x)=sim3(x) & sim1(x)>=0.5
100000
10000
1000
100
10
1
0
0.2
0.4
0.6
0.8
1
The constraint value

Figure 7


Remove "overly" ribosomal patterns:
length(regexp(x,'*ribosom*', **TEXT_BOW**))≤1
3881 patterns, 3178 intervals, 1192 max. patterns
Remove "overly" ribosomal patterns:
length(regexp(x,'*ribosom*', **GO_terms**))≤1
1633 patterns, 1506 intervals, 855 max. patterns
Apply gene similarity and existence of description (**TEXT** similarity):
svsim(x)/(svsim(x)+mvsim(x))≥0.7 & sumsim(x)/svsim(x)≥0.025
9 patterns, 9 intervals, 9 max. patterns
Apply existence of gene description (**GO_FUNCTION**):
svsim(x)/(svsim(x)+mvsim(x))≥0.75
63 patterns, 63 intervals, 61 max. patterns
Apply gene similarity (**GO_FUNCTION** similarity):
insim(x,0.5,1)/svsim(x)≥0.6
sumsim(x)/svsim(x)≥0.5
2 patterns, 2 intervals, 2 max. patterns
Similarity of **biological situations**:
(post-processing)
sumsim(s)/svsim(s)≥0.5
5 patterns, 5 intervals, 5 max. patterns
KHDRBS1 NONO TOP2B FMR1 & 48 52 54 56 62 65
a large number of common terms & brain(6/73) cancer(6/133) medulloblastoma(6/24)