Chapter XII
Discovering Knowledge from Local Patterns in SAGE Data

Bruno Crémilleux
Université de Caen, France

Arnaud Soulet
Université François Rabelais de Tours, France

Jiri Kléma
Czech Technical University, Czech Republic

Céline Hébert
Université de Caen, France

Olivier Gandrillon
Université de Lyon, France

ABSTRACT

The discovery of biologically interpretable knowledge from gene expression data is a crucial issue. Current gene data analysis is often based on global approaches such as clustering. An alternative way is to utilize local pattern mining techniques for global modeling and knowledge discovery. Nevertheless, moving from local patterns to models and knowledge is still a challenge due to the overwhelming number of local patterns and their summarization remains an open issue. This chapter is an attempt to fulfill this need: thanks to recent progress in constraint-based paradigm, it proposes three data mining methods to deal with the use of local patterns by highlighting the most promising ones or summarizing them. Ideas at the core of these processes are removing redundancy, integrating background knowledge, and recursive mining. This approach is effective and useful in large and real-world data: from the case study of the SAGE gene expression data, we demonstrate that it allows generating new biological hypotheses with clinical applications.

INTRODUCTION

In many domains, such as gene expression data, the critical need is not to generate data, but to derive knowledge from huge and heterogeneous datasets produced at high throughput. It means that there is a great need for automated tools helping their analysis. There are various methods, including global techniques such as hierarchical clustering, K-means, or co-clustering (Madeira & Oliveira, 2004) and approaches based on local patterns (Blachon et al., 2007). In the context of genomic data, a local pattern is typically a set of genes displaying specific expression properties in a set of biological situations. A great interest of local patterns is to capture subtle relationships in the data which are not detected by global methods and leading to the discovery of precious nuggets of knowledge (Morik et al., 2005). But, the toughness of extraction of various local patterns is a substantial limitation of their use (Ng et al.,1998; Bayardo, 2005). As the search space of the local patterns exponentially grows according to the number of attributes (Mannila &Toivonen, 1997), this task is even more difficult in *large* datasets (i.e., datasets where objects having a large number of columns).This is typically the case in gene expression data: few biological situations (i.e., objects) are described by ten of thousands of gene expressions values (i.e., attributes) (Becquet et al. 2002). In such situations, naive methods or usual level-wise techniques are unfeasible (Pan et al.,2003; Rioult et al., 2003). Nevertheless, especially in the context of transactional data, the recent progress in constraint-based pattern mining (see for instance (Bonchi & Lucchese, 2006; De Raedt et al., 2002) enable to extract various kind of patterns even in large datasets (Soulet et al., 2007). But, this approach has still a limitation: it tends to produce an overwhelming number of local patterns. Pattern flooding follows data flooding: the output is often too large for an individual and global analysis performed by the end-user. This is especially true in noisy data,such as genomic data where the most significant patterns are lost among too many trivial, noisy and redundant information. Naive techniques such as tuning parameters of methods (e.g., increasing the frequency threshold) limit the output but only lead to produce trivial and useless information.

This paper tackles this challenge. Relying on recent progress in constraint-based paradigm, it presents three data mining methods to deal with the use of local patterns by highlighting the most promising ones or summarizing them. The practical usefulness of these methods are supported by the case study of the SAGE gene expression data (introduced in the next section). First, we provide a method to mine the set of the simplest characterization rules while having a controlled number of exceptions. Thanks to their property of minimal premise, this method limits the redundancy between rules. Second, we describe how to integrate in the mining process background knowledge available in literature databases and biological ontologies to focus on the most promising patterns only. Third, we propose a recursive pattern mining approach to summarize the contrasts of a dataset: only few patterns conveying a trade-off between significance and representativity are produced. All of these methods can be applied even on large data sets. The first method comes within the general framework of removing redundancy and providing lossless representations whereas the two others propose summarizations (all the information cannot be regenerated but the most meaningful features are produced). We think that these two general approaches are complementary. Finally, we sum up the main lessons coming from mining and using local patterns on SAGE data, both from the data mining and the biological points of view. It demonstrates the practical usefulness of these approaches enabling to infer new relevant biological hypotheses.

This paper abstracts our practice of local patterns discovery from SAGE data. We avoid technical details (references are given for in-depth information), but we emphasize the main principles and results and we provide a cross-fertilization of our “in silico” approaches for discovering knowledge in gene expression data from local patterns.

MOTIVATIONS AND CONTEXT

Motivations

There is a huge research effort to discover knowledge from genomics data and mining local patterns such as relevant synexpression groups or characterization rules is requested by biologists. It is a way to better understand the role and the links between genes. Elucidating the association between a set of co-regulated genes and the set of biological situations that gives rise to a transcription module is a major goal in functional genomics. Different techniques including microarray (DeRisi et al., 1997) and SAGE (Velculescu et al., 1995) enable to study the simultaneous expression of thousands of genes in various biological situations. The SAGE technique aims to measure the expression levels of genes in a cell population. 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 micro-arrays to produce datasets that can be directly compared between libraries without the need for external normalization. In our work, we use publicly available human serial analysis of gene expression SAGE libraries. We built a 207x11082 data set made up of 207 biological situations described by 11,082 gene expressions (i.e., a set of genes identified without ambiguous tags which will be useful for the techniques integrating the background knowledge) and a 90x27679 data set gathering 90 biological situations for 27,679 gene expressions (i.e., all the available transcriptomic information from these libraries).

As said in introduction, local pattern discovery has become a rapidly growing field (Blachon et al., 2007) and a range of techniques is available for producing extensive collections of patterns. Because of the exhaustive nature of most such techniques, the so-called local patterns provide a fairly complete picture of the information embedded in the database. But, as these patterns are extracted on the basis of their individual merits, this results in large sets of local patterns, potentially highly redundant. Moreover, the collections of local patterns represent fragmented knowledge and their huge size prevents a manual investigation. A major challenge is their combination and summarization for global modeling and knowledge discovery. It is a key issue because a useful global model, such a classifier or a co-clustering, is often the expected result of a data mining process. As well as their exhaustive nature and their ability to catch subtle relationships, summarizations of local patterns can capture their joint effect and reveal a knowledge not conveying by the usual kinds of patterns. The next section provides a few attempts in this general direction.

Related Work

Several approaches have been proposed to reduce the number of local patterns irrespective of their subsequent use. Examples include condensed representations (Calders et al., 2005), compression of the dataset by exploiting the Minimum Description Length Principle (Siebes et al., 2006) or the constraint-based paradigm (Ng et al., 1998; De Raedt et al., 2002). Constraints provide a focus that allows to reduce the number of extracted patterns to those of a potential interest given by the user. Unfortunately, even if these approaches enable us to reduce the number of produced patterns, the output still remains too large for an individual and global analysis performed by the end-user. Recently, two approaches appeared in the literature, which explicitly have the goal of combining and selecting patterns on the basis of their usefulness in the context of the other selected patterns: these pattern set discovery methods are constraint-based pattern set mining (De Raedt & Zimmermann, 2007), and pattern teams (Knobbe

& Ho, 2006). Constraint-based pattern set mining is based on the notion of constraints defined on the level of pattern sets (rather than individual patterns). These constraints capture qualities of the set such as size or representativeness. In the pattern team approach, only a single subset of patterns is returned. Pattern sets are implicitly ranked on the basis of a quality measure, and the best-performing set (the pattern team) is reported. Even if these approaches explicitly compare the qualities of patterns between them, they are mainly based on the reduction of the redundancy.

On the other hand, we think that it should be a pity to consider the summarization of local patterns only from the point of view of the redundancy. Local patterns can be fruitfully gathered for global modeling and knowledge discovery. Interestingly, such global models or patterns can capture the joint effect of local patterns such as co-classification performs. This approach is a way of conceptual clustering and provides a limited collection of bi-clusters. These bi-clusters are linked for both objects (i.e., biological situations) and attributes (i.e., genes). Tackling genomic data, Pensa et al. (Pensa et al., 2005) show that the bi-clusters of the final bi-partition are not necessary elements of the initial set of the local patterns. The bi-partition may come from a reconstruction of the biological situations and genes defining the local patterns. Except for particular kinds of local patterns (e.g., closed patterns (Blachon et al., 2007)), due to their large number of attributes, there are few works on discovery knowledge from SAGE data (Klema et al.).

Constraint-Based Pattern Mining

As said in introduction, methods presented in this paper stem from recent progress in constraint-based paradigm. A constraint is a way to express a potential interest given by the user. Due to the huge search space of candidate patterns, a challenge is to push constraints in the core of the mining process by automatically inferring powerful and safe pruning conditions in order to get patterns satisfying a constraint. At least in transactional domains, there are now generic approaches to discover *local patterns* under constraints (De Raedt et al., 2002; Soulet & Crémilleux, 2005) even in large datasets (Soulet et al., 2007). A survey of the primitive-based framework (Soulet & Crémilleux, 2005) is provided below. This framework is at the basis of our method integrating background knowledge. We give now basic definitions used among the paper.

Let I be a set of distinct literals called *items*, an itemset (or pattern) corresponds to a non-null subset of I. These patterns are gathered together in the language L_I: $L_I = 2^I \setminus \emptyset$. A transactional dataset is a multi-set of patterns (i.e., transactions) of L_I. Each *transaction* is a database entry. More generally, transactions are called *objects* and items *attributes*. For instance, Table 1 gives a transactional dataset D with 8 objects o_1,..., o_8 (e.g., biological situations) described by 6 items A,…, F (e.g., gene expressions). This is a toy example which will be used throughout this paper. A value 1 for a biological situation and a gene expression means that this gene is over-expressed in this situation. In the SAGE data, each situation belongs to a class value (cancer versus no cancer) according to the biological origin of the tissue of the situation. For that reason, we divide D in two datasets D_1 and D_2 and a situation is labeled by the item C_1 (i.e., it belongs to D_1) or C_2 (i.e., it belongs to D_2).

Local patterns are regularities that hold for a particular part of the data. Let X be a pattern. We recall that the support of X in D denoted by $supp(X, D)$ is the proportion of objects in D containing X (we omit D when this data set is used by default). For instance, $supp(AB) = 3/8$. The constraint-based pattern mining framework D aims at discovering all the patterns of L_I satisfying a given predicate q, named *constraint*, and occurring in D. A well-known example is the *frequency* constraint focusing on

Table 1. Example of a transactional dataset

$\mathcal{D}$

	Gene expressions							
Situations	A	B	C	D	E	F		
o_1			1				C_1	
o_2	1	1		1		1	C_1	$\mathcal{D}_1$
o_3	1			1	1		C_1	
o_4	1	1		1			C_1	
o_5	1			1			C_2	
o_6	1	1	1			1	C_2	$\mathcal{D}_2$
o_7		1	1	1			C_2	
o_8			1		1		C_2	

patterns having a support exceeding a given minimal threshold *minsupp* > 0: *supp*(*X*, D) ≥ *minsupp*. For instance, *AB* is a frequent pattern with *minsupp* = 0.2. We will also use an absolute definition of the support , the frequency of *X* denoted *freq*(*X*) (*freq*(*X*, D) = *supp*(*X*,D) x |D|). As previously, we omit D when this data set is used by default. For instance, *freq*(*AB*) = 3. The frequency of the rule $X \rightarrow Y$ is *freq*(*XY*) and its *confidence* is *supp*(*XY*)/*supp*(*X*).

There are a lot of various constraints to evaluate the relevance of local patterns (Ng et al., 1998; Soulet & Crémilleux, 2005). The constraint-based paradigm also includes interestingness measures (the frequency is an example) to select local patterns. In the following, we will use the area of a pattern *area*(*X*): it is the frequency of a pattern times its length (i.e., *area*(*X*) = *freq*(*X*) x *count*(X) where *count*(X) denotes the cardinality of *X*. The area can be seen as the translation in the constraint paradigm of a synexpression group. For instance, the pattern *AB* (or *ABD*) satisfies the constraint *area*(*X*) ≥ 6 (as previously, if no data set is specified, it means that D is used). Emerging patterns (EPs) are another example. They are at the core of the summaries presented in the following. An EP is a pattern whose support strongly varies between two parts of a dataset (i.e., two classes),enabling to characterize classes (Dong & Li, 1999). The growth rate of *X* is $gr_i(X) = supp(X, D_i)/supp(X, D\backslash D_i)$. More formally, if we consider the two cancer and no cancer classes, a frequent emerging pattern X satisfies the constraint $supp(X,D) \geq minsupp \wedge (gr_{cancer}(X) \geq mingr \vee gr_{no\ cancer}(X) \geq mingr)$.

MINING A SYNTHESIS OF CLASSIFICATION RULES

There is an intense need of classification and classes characterization techniques to perform data mining tasks required on real-world databases. For instance, the biological situations in SAGE data are divided into two classes (cancer and no cancer) and biologists would like to better understand the relationships between the genes and these classes. For that purpose, we use the characterization rules previously introduced in (Crémilleux & Boulicaut, 2002). Thanks to a property of minimal premises, these characterization rules provide a kind of synthesis of the whole set of classification rules (i.e., all

the rules concluding on a class value). This result stems from the property of specific patterns, the δ-free patterns which are made of attributes without frequency relations between them (Boulicaut et al., 2003). Experiments (Crémilleux & Boulicaut, 2002) show that the number of characterization rules is at least an order of magnitude lower than the number of classification rules. Unfortunately, the method given in (Crémilleux & Boulicaut, 2002) does not run on large datasets such as the SAGE data. For that reason we have proposed a new method (Hébert et al., 2005) based on the extension of patterns (the extension of a pattern X is the maximal set of the objects containing X), because the extension has few objects in large databases. We give now a formal definition of these characterization rules (X and Y are patterns and C_i is an item referring to a class value).

Definition 1 (characterization rules): *Let minfreq be a frequency threshold, δ be an integer, a rule* $X \rightarrow C_i$ *is a characterization rule if there is no rule* $Y \rightarrow C_i$ *with* $Y \subset X$ *and a confidence greater than or equal to 1-(δ/minfreq).*

Given a frequency threshold *minfreq*, this definition means that we consider only the minimum sets of attributes (i.e., the minimal premises) to end up C_i, the uncertainty being controlled by δ. For instance, in our running example (Table 1), with δ = 1 and *minfreq* = 2, $C \rightarrow C_2$ is a characterization rule (there is one exception), but $CD \rightarrow C_2$ is not a characterization rule (it is covered by the previous rule). We argue that this property of minimal premise is a fundamental issue for classification. Not only it prevents from over-fitting but also it makes the characterization of an example easier to explain. It provides a feedback on the application domain expertise that can be reused for further analysis.

The value of δ is fundamental to discover relevant rules. With δ = 0, every rule must have a confidence value of 1 (i.e., *exact* rule). In many practical applications, such as the SAGE data, there are generally very few exact rules due to the non-determinism of the phenomena. We have to relax the condition on δ to accept exceptions (the more δ raises, the more the confidence decreases).

We developed the FTCminer prototype which extracts the sound and complete collection of frequent characterization rules (Hébert et al., 2005). FTCminer follows the outline of a level-wise algorithm (Mannila & Toivonen, 1997). Its originality is the use of the extension of patterns and that there is no generation phase of all the candidates at a given level since the candidates are generating one at a time. Thanks to these techniques, we are able to mine characterization rules even in large data sets whereas it was impossible before (Becquet at al., 2002; Hébert et al., 2005). Main results on SAGE data are given in the section on experiments.

INTEGRATING INFORMATION SOURCES SYNTHESIZING BACKGROUND KNOWLEDGE

This section sketches our approach to integrate background knowledge (BK) in the mining process to focus on the most plausible patterns consistent with pieces of existing knowledge. For instance, biologists are interested in constraints both on synexpression groups and common characteristics of the descriptions of the genes and/or biological situations under consideration. BK is available in relational and literature databases, ontological trees and other sources. Nevertheless, mining in a heterogeneous environment allowing a large set of descriptions at various levels of detail is highly non-trivial. There are various ways to interconnect the heterogeneous data sources and express the mutual relations among

the entities they address. We tackle this issue with the constraint paradigm. We think it is a promising way for such a work, the constraints can effectively link different datasets and knowledge sources (Soulet et al., 2007).

Our approach is based on the primitive-based constraints (Soulet & Crémilleux, 2005). There are no formal properties required on the final constraints and they are freely built of a large set of primitives. The primitives have to satisfy solely a property of monotonicity according to their variables (when the others remain constant). We showed that the whole set of primitive-based constraints constitutes a super-class of monotone, anti-monotone, succinct and convertible constraints (Soulet & Crémilleux, 2008). Consequently, the proposed framework provides a flexible and rich constraint (query) language. For instance, the product of two primitives *count*(X) x *freq*(X) may address the patterns having a certain minimum length (i.e., containing a minimum number of genes) and frequency (i.e., covering a minimum number of situations). We referred to it as *area*(X) above.

Furthermore, this framework naturally enables to integrate primitives addressing external data. Let us consider the transcriptomic mining context given in Figure 1. The involved data include a transcriptome dataset also called internal data as in our running example. External data - a similarity matrix and textual resources - summarize BK that contains various information on genes. Each field of the triangular matrix $s_{ij} \in [0,1]$ gives a similarity measure between the genes *i* and *j*. The textual dataset provides a description of genes. Details on the processing of textual resources within this approach and primitives tackling external data are given in another chapter of this book (Klema & Zelezny). The mined patterns are composed of genes of the internal data, the corresponding objects are usually also noted (and possibly analyzed). 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. The user can iteratively develop complex constraints integrating various knowledge types.

A real example of a constraint $q(X)$ is given in Figure 1. The first part (a) of q addresses the internal data and means that the biologist is interested in patterns satisfying a minimal area. The other parts deal with the external data: (b) is used to discard ribosomal patterns (one gene exception per pattern is allowed), (c) avoids patterns with prevailing items of an unknown function and (d) is to ensure a minimal

Figure 1. Example of a toy (transcriptomic) mining context and a constraint

Internal data

Boolean matrix $\mathcal{D}$

	Gene expressions					
Situations	*A*	*B*	*C*	*D*	*E*	*F*
o_1			1			
o_2	1	1		1		1
o_3	1			1	1	
o_4	1	1		1		

External data

Similarity matrix

	A	*B*	*C*	*D*	*E*	*F*
A		.07	?	?	.2	0
B			.06	?	?	0
C				.07	.05	.04
D					.03	.1
E						?

Textual data

A	'metal ion binding' 'transcription factor'
B	'serine-type peptidase activity' 'proteolysis'
C	'DNA binding' 'metal ion binding'
D	'ATP binding' 'nucleotide binding'
E	'proteolysis'
F	'ATP binding' 'metal ion binding'

freq, count,... ***sumsim, svmsim,...*** ***regexp***

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

average gene similarity. The usefulness of such a constraint is shown in the section on experiments.

We have proposed a general prototype Music-dfs which discovers soundly and completely all the patterns satisfying the specified set of constraints (Soulet et al., 2007). Its efficiency lies in its depth-first search strategy and a safe pruning of the pattern space by pushing the constraints. Extractions in large data sets such as the SAGE data are feasible. Section on experiments demonstrates that our procedure leads to a very effective reduction of the number of patterns, together with an "interpretation" of the patterns.

RECURSIVE PATTERN MINING

This section outlines the recursive pattern mining framework and the discovery of the recursive emerging patterns (Soulet, 2007). The key idea is to repeat the pattern mining process on output to reduce it until few and relevant patterns are obtained. The final recursive patterns bring forward information coming from each mining step.

As often in mining constraint-based local patterns, the so-called collections of frequent emerging patterns (EPs) are huge and this hinders their uses. Several works address methods to reduce these collections by focusing on the most expressive ones (Bailey et al., 2002) (which are only present in one class) or by mining a lossless condensed representation (Li et al., 2007; Soulet et al., 2004). Nevertheless, these approaches do not reduce enough the number of mined patterns. Moreover, setting thresholds (i.e., *minsupp* or *mingr*) is often too subtle. Both the quantity and the quality of desired patterns are unpredictable. For instance, a too high threshold may generate no answer, a small one may generate thousands of patterns. Increasing thresholds to diminish the number of output patterns may be counterproductive (see the example with the area constraint in the section on experiments). Mining recursive patterns aims at solving these pitfalls.

In this work, we deal with frequent emerging patterns. Recursive emerging patterns (REPs) are the EPs which frequently occur within the outputted EPs according to the classes. The assumption is that these EPs are significant because the recursive mining process enables to synthesize and give prominence to the most meaningful contrasts of a dataset. A recursive emerging pattern *k*-summary (a REP *k*-summary, see Definition 2) provides a short description of the dataset constituted at most *k* REPs summarizing the contrasts according to the classes. It is produced by the generic recursive pattern mining framework: for each step, the previous mined patterns constitute the new transactional dataset. A first step mines all the frequent emerging patterns, as usual in the constraint-based pattern mining framework. Then the outputted EPs are joined to form a new dataset $D^2 = D^2_{\text{cancer}} \cup D^2_{\text{no cancer}}$. The EPs concluding on the class cancer (or no cancer) constitute the new sub-dataset D^2_{cancer} (or $D^2_{\text{no cancer}}$) and the process is repeated. This recursive process is ended as soon as the result becomes stable. At the end, we get at most *k* patterns brought forward information coming from each mining step. They summarize the main contrasts repeated through the outputs. From an abstract point of view, REPs can be seen as generalizations of emerging patterns. Main features on the method, (e.g., the theoretical convergence of recursive mining, number of steps) are given in (Soulet, 2007) and are not developed here because they are not crucial in practice.

For example, Table 2 depicts the mining of REPs from D (cf. Table 1) with *minsupp*=0.1 and *mingr*=2. Obviously, the datasets D^2_1 and D^2_2 are exactly the EPs in $D = D^1$ with *minsupp*=0.1 and *mingr*=2. At the

Table 2. REPs mined from D with minsupp = 0.1 and mingr = 2

$\mathcal{D}_1^2$

A			E		
A		D	E		
A		D			F
A	B	D			F
A		D			
A	B	D			
A	B				
		D	E		
		D			F
	B	D			F
	B	D			

$\mathcal{D}_2^2$

A		C			
A	B	C			
A	B	C			F
A		C			F
		C			
		C		E	
		C			F
	B	C			F
		C	D		
	B	C	D		
	B	C			

$\mathcal{D}_1^3$

A		D			
			E		
		D			F
	B	D			
		D			

$\mathcal{D}_2^3$

A		C			
		C			F
	B	C			
		C			

Table 3. REP 10-summary of D with mingr = 2

REPs of $\mathcal{D}_1$

REP	*supp*	gr_1
AD	0.5	3
E	0.25	1
DF	0.125	∞
BD	0.375	2
D	0.625	1.5

REPs of $\mathcal{D}_2$

REP	*supp*	gr_2
AC	0.125	∞
CF	0.125	∞
BC	0.25	∞
C	0.5	3

next mining step, the number of REPs (i.e., union of D^3_1 and D^3_2 : 9 patterns) is lower than the number of EPs (i.e., union of D^2_1 and D^2_2 : 22 patterns). In this example, EPs in D^4 are exactly the patterns of D^3 and then the collection of frequent REPs is stable: final REPs come from D^3 . We define below a REP *k*-summary which straightforwardly stems from REPs:

Definition 2 (REP *k*-summary): *A REP k-summary (according to mingr) is the whole collection of REPs obtained with minsupp=1/k and mingr.*

We argue that a REP *k*-summary is a compact collection of EPs having a good trade-off between significance and representativity. We proved (Soulet, 2007) that the size of a REP *k*-summary is bounded according to *minsupp*: to get at most *k* patterns in a REP *k*-summary, it is enough to fix *minsupp* = 1/*k*. For instance, the 10-summary in Table 3 contains 9 patterns (we have $9 \leq 10$ with *k*=10=1 /*minsupp* = 1/0.1). Moreover, we claim that it is easier for a user to fix a maximal value for the number of patterns than the support threshold.

Besides a REP *k*-summary covers a large part of the dataset D: most objects support at least one EP of the summary. This is due to REPs are frequent patterns in the dataset of each step. Thus, they are representative of the original dataset D, but also of all the emerging patterns from D. Table 3 recalls the REP 10-summary with *mingr*=2 from our running example. Supports (column *supp*) and growth rates (column gr_i) in the initial dataset D are added. As *minsupp*=1/10, this summary is exactly the REPs given in Table 2. Interestingly, we note that the growth rates of the REPs may be lower than *mingr* (e.g., $gr_1(D,\mathrm{D})=1.5$ whereas *mingr* = 2). This avoids the crisp effect of a threshold where a promising pattern is deleted only because its value for the measure is just under the threshold. The power of the recursive mining approach relies on the summarization: most of the REPs have a significative growth rate and all the objects (except o_1 and o_5) are covered by a REP concluding to their class values. Clearly, o_1 is closer to the objects of D_2 than objects of D_1, this explains why o_1 is not characterized by a REP. A similar reasoning can be done with o_5.

The tunable concision of REPs favours users' interpretation. Each REP can be individually interpreted as usual EPs, providing a qualitative and quantitative information. Appropriately, the small collection of REPs offers a global and complementary description of the whole dataset.

LESSONS FROM MINING SAGE DATA

The section outlines the main results achieved on SAGE data thanks to the previous data mining methods. Then, we synthesize the major lessons both from the data mining and the biological points of view.

A Sketch of Biological Results

To fully understand the results of experiments, we have to precise that each attribute of a SAGE data set is a *tag*. The identification of genes is closely related to the tags and biologists are able to associate genes and tags. In the case of the 207x11082 data set, each tag is unambiguously identified. This property is very useful to link together the information coming from several sources of BK.

Gene expressions are quantitative values and we must identify a specific gene expression property to get binary value and run the data mining methods depicted above. In principle, several properties per gene could be encoded, e.g. over-expression and under-expression. In our studies, we decided to focus on over-expression (over-expression has been introduced in the beginning of the paper). Several ways exist for identifying gene over-expression (Becquet et al., 2002). Results given in this paper are performed by using the mid-range method: the threshold is fixed w.r.t. the maximal value (*max*) observed for each tag. All the values which are greater than (100 - X%) of *max* are assigned to 1, 0 for the others (here, $X = 25$). For the 90x27679 data set, the values of tags vary from 0 to 26021. The percentage of tags which values are different from 0 is 19.86% and the arithmetic mean is around 4. As already said, the biological situations are divided into two classes (cancer and no cancer). 59 situations are labelled by cancer and 31 by no cancer (i.e., normal).

Characterization rules. We give the mean features on our work on mining characterization rules on SAGE data (more experiments and details are provided in (Hébert et al., 2005). In this paper, we only deal with the classes cancer and no cancer. More fruitful further biological investigations will require to use sub-groups of these classes, such sub-groups being defined according to biological criteria (e.g., a cancer type).

Table 4 presents a selection of rules with at least two tags in their body and a rather high confidence and frequency with *minfreq* and δ=1. Table 5 provides the description of tags (identification number, sequence and description) only for the tags which appear the most frequently in our results. Some tags are identified by several genes: their identifications are separated by ";".

Few tags (e.g., 4602, 8255, 11115, 22129) clearly arise in many rules concluding on cancer. They may have an influence on the development of this disease. It is interesting to note that the frequencies of these tags strongly varies from one class to another. For example, the tag 11115 appears 28.7 times more in rules characterizing cancer than no cancer. The tag 11115 is identified as GPX1. The expression of GPX1 has been found in various studies to be correlated with cancerous situations (Korotkina et al., 2002; Nasr et al., 2004). On the contrary, the tag 22129 appears 22 times more in rules concluding on no cancer than concluding on cancer. It might mean that this tag is related to normal development. We will come back on this tag below, with regard to the interestingness of biological results.

Table 4. Examples of potential relevant rules with minfreq = 4 and δ= 1

Premise	Conclusion	Exceptions	Frequency	Confidence
11115 19811	cancer	1	13	0.92
5961 11115	cancer	0	12	1
8279 23600	cancer	1	12	0.92
10960 11115	cancer	1	12	0.92
11115 20766	cancer	1	12	0.92
4602 7259 18882	cancer	1	10	0.9
4602 7259 24686	cancer	1	10	0.9
8255 11115 19811	cancer	1	10	0.9
4602 7259 20461	cancer	1	9	0.89
4602 7259 25202	cancer	1	9	0.89
4602 18882 24686	cancer	1	9	0.89
4287 4602 7818	cancer	1	8	0.88
4287 4602 19811	cancer	1	8	0.88
4602 7259 19734	cancer	1	8	0.88
4602 24686 25202	cancer	1	8	0.88
4602 25128 25202	cancer	1	8	0.88
7259 12667 16807	cancer	1	8	0.88
8255 11115 13642	cancer	0	8	1
8255 11115 26846	cancer	1	8	0.88
8255 19811 26846	cancer	1	8	0.88
22619 25202 26846 27358	cancer	1	5	0.8
16786 26715	no cancer	1	7	0.86
22129 25356	no cancer	1	7	0.86
22129 27414	no cancer	1	7	0.86
22647 25356	no cancer	1	7	0.86
1722 25202 26715	no cancer	1	6	0.83

Table 5. Characteristics of potential relevant tags

Number	Sequence	Description
4287	AGCTCTCCCT	RPL17 ribosomal protein L17
4602	AGGCTACGGA	Similar to ribosomal protein L13a, 60S ribosomal protein L13a, 23 kD highly basic protein
8255	CATCCAAAAC	HNRPH1 Heterogeneous nuclear ribonucleoprotein H1 (H)
11115	CTCTTCGAGA	GPX1 Glutathione peroxidase 1
19811	GTTGCTGCCC	NIFIE14 Seven transmembrane domain protein
22129	TCAGAGAATA	SLC25A22 Solute carrier family 25 (mitochondrial carrier: glutamate), member 22; IRS2 Insulin receptor substrate 2
25202	TGTGCTAAAT	RPL34 Ribosomal protein L34

Integrating BK. A highly valuable biological knowledge comes from the patterns that concern genes with interesting common features (e.g., process, function, location, disease) whose synexpression is observed in a homogeneous biological context (i.e., in a number of analogous biological situations). We give now an example of such a context with the set of medulloblastoma SAGE libraries discovered from constrained patterns taking into account the BK. We use the 207x11082 data set because each tag is unambiguously identified. This property is very useful to link together the information coming from several sources of BK.

The area constraint is the most meaningful constraint on the internal data for the search of such synexpression groups. On the one hand, it products large patterns (the more genes they contain, the better ; the higher the frequency is, the better). On the other hand, it enables exceptions on genes and/or biological situations contrary to the maximal patterns (Rioult et al., 2003; Blachon et al., 2007) (i.e.,

formal concepts) which require that all the connected genes are over-expressed. In domains such as gene expressions where the non-determinism is intrinsic, this lead to a fragmentation of the information embedded in the data and a huge number of patterns covering very few genes or biological situations.

We fix the area threshold thanks to statistical analysis of random datasets having the same properties as the original SAGE data. We obtain a value of 20 as an optimal area threshold to distinguish between spurious (i.e., occurring randomly) and meaningful patterns (first spurious patterns start to appear for this threshold area). Unfortunately, we get too many (several thousands) candidate patterns. Increasing the threshold of the area constraint to get a reasonable number of patterns is rather counterproductive. The constraint $area \geq 75$ led to a small but uniform set of 56 patterns that was flooded by the ribosomal proteins which generally represent the most frequent genes in the dataset. Biologists rated these patterns as valid but useless.

The most valuable synexpression groups expected by biologists have non-trivial size containing genes and situations whose characteristics can be generalized, connected, interpreted and thus transformed into knowledge. To get such patterns, constraints based on the external data have to be added to the minimal area constraint just like in the constraint *q* given in the section on integration of information sources synthetizing BK. It joins the minimal area constraint with background constraints coming from the NCBI (cf. http://www.ncbi.nlm.nih.gov) textual resources (gene summaries and adjoined PubMed abstracts). There are 46671 patterns satisfying the minimal area constraint (the part (a) of the constraint *q*), but only 9 satisfy *q*. This shows the efficiency of reduction of patterns brought by the BK. One of these patterns is of biological interest (Klema et al.). It consists of 4 genes (KHDRBS1 NONO TOP2B FMR1) over-expressed in 6 biological situations (BM_P019 BM_P494 BM_P608 BM_P301 BM_H275 BM_H876), BM stands for brain medulloblastoma. A cross-fertilization with other external data was obviously attractive. So, we define a constraint *q'* which is similar to *q*, except that the functional Gene Ontology (cf. http://www.geneontology.org/) is used instead of NCBI textual resources. Only 2 patterns satisfy *q'*. Interestingly, the previous pattern that was identified by the expert as one of the "nuggets" provided by *q'* is also selected by *q'*. The constraints *q* and *q'* demonstrate two different ways to reach a compact and meaningful output that can be easily human surveyed.

REP summaries. Following our work to study the relationships between the genes and the type of biological situations according to cancer and no cancer, we computed REP summaries from the SAGE data. We use the same binary data set as in the characterization rules task.

Table 6 depicts the REP 4-summary with *mingr*=2. We observe that all patterns describe the class cancer. Using other values for the parameters *k* and *mingr* also leads to only characterize cancer. Interestingly, the 3 extracted genes characterize 40% of biological situations and even 61% of cancerous situations. We will see below that this REP summary confirms the results obtained with characterization rules. Nevertheless, a great interest of the approach based on the summarization is to directly isolate genes without requiring a manual inspection of rules.

A Breakthrough on Mining and Using Local Pattern Methods

A first challenge in discovery knowledge from local patterns in SAGE data is to perform the local pattern extractions. Recalling that few years ago it was impossible to mine such patterns in large datasets and only association rules with rather a high frequency threshold were used (Becquet et al., 2002). Relying on recent progress in constraint-based paradigm, we have proposed efficient data mining methods to mine local patterns solving the problem due to the size of the search space. Key ideas are the use of

Table 6. REP 4-summary of SAGE data with mingr = 2

Sequence (tag)	Description (gene)	*supp*	*gr*
	`cancer`		
CATCCAAAAC	HNRPH1 Heterogeneous nuclear ribonucleoprotein H1 (H)	0.28	2.10
CTCTTCGAGA	GPX1 Glutathione peroxidase 1	0.32	3.28
GTTGCTGCCC	NIFIE14 Seven transmembrane domain protein	0.26	2.50

Class coverage : 40% / Running-time: 1.37s

the extension of patterns and depth-first search. Thanks to the constraint-based mining approach, the user can handle a wide spectrum of constraints expressing a viable notion of interestingness. We deal with characterization rules, emerging patterns, minimal area (which is the translation in the constraint paradigm of a synexpression group), but many other possibilities are offered to the user.

A second challenge is to deal with the (huge) collections of local patterns. We claim that we propose fruitful methods to eliminate redundancy between patterns and highlighting the most promising ones or summarizing them. Integrating BK in a data mining process is a usual work for the biologist, but he did it manually. To the best of our knowledge, there is no other constraint-based method to efficiently discover patterns from large data under a broad set of constraints linking BK distributed in various knowledge sources. Recursive mining is a new and promising way which ensures to produce very few patterns summarizing the data. These summaries can easily be inspected by the user.

Interestingness of Biological Results

A first result is that most of the extracted patterns were harboring (or even composed only of) genes encoding ribosomal proteins, and proteins involved in the translation process. This is for example the case for the vast majority of the characterization rules concluding on cancer (see Tables 4 and 5). Such an overexpression has been documented in various contexts ranging from prostate cancer (Vaarala et al., 1998) to v-erbA oncogene over-expression (Bresson et al., 2007). The biological meaning of such an over-expression is an open question which is currently investigated in the CGMC lab.

As a second lesson, we demonstrated that mining local patterns discovers promising biological knowledge. Let us come back on the pattern highlighted by the BK (see above). This pattern can be verbally characterized as follows: it consists of 4 genes that are over-expressed in 6 biological situations, it contains at most one ribosomal gene, the genes share a lot of common terms in their descriptions as well as they functionally overlap, at least 3 of the genes are known (have a non-empty record) and all of the biological situations are medulloblastomas which are very aggressive brain tumors in children. This pattern led to an interesting hypothesis regarding the role of RNA-binding activities in the generation and/or maintenance of medulloblastomas (Klema et al.).

Finally, our data mining approaches enable a cross-fertilization of results, indicating that a relatively small number of genes keeps popping up throughout various analysis. This is typically the case of the GPX1 gene highlighted both on characterization rules and REP summaries to have an influence on the development of cancer. This gene encodes a cytosolic glutathione peroxidase acting as an antioxidant by detoxifying hydroperoxides (Brigelius-Flohe, 2006). It is known that exposition to an oxidative stress is a factor that favors development of different types of tumors (Halliwell, 2007). It is therefore

reasonable to suggest that this gene is over-expressed to respond to an oxidative stress to which cells have been exposed. It would be of interest to verify its expression level by RT-PCR in normal versus cancerous samples in human.

CONCLUSION

There are now a few methods to mine local patterns under various sets of constraints even in large data sets such as gene expression data. Nevertheless, dealing with the huge number of extracted local patterns is still a challenge due to the difficult location of the most interesting patterns. In this chapter, we have presented several methods to reduce and summarize local patterns. We have shown the potential impact of these methods on the large SAGE data. By highlighting few patterns, these approaches are precious in domains such as genomics where a manual inspection of patterns is highly time consuming. Our methods provide qualitative information (e.g., biological situations associated to genes, text resources) but also quantitative information (e.g., growth rate or other measures). Such characteristics are major features in a lot of domains with noisy data and non-deterministic phenomena for knowledge discovery. We think that our results on SAGE data illustrate the power of local patterns to highlight gene expression patterns appearing through very different conditions, and that such patterns would not be captured by global tools such as hierarchical clustering.

A future issue is the combination of these methods: how to ensure to build non redundant optimal recursive patterns? how to integrate BK in recursive mining? Another way is to design new kinds of constraints to directly mine global patterns as sets of local patterns or produce models.

ACKNOWLEDGMENT

This work has mainly been done within the Bingo project framework (http://www.info.unicaen.fr/~bruno/bingo). The work of Jiri Klema was funded by the Czech Ministry of Education in terms of the research programme Transdisciplinary Research in the Area of Biomedical Engineering II, MSM 6840770012. The authors thank all members of the

Bingo project and especially Sylvain Blachon for generating the SAGE gene expression matrices and Jean-François Boulicaut for fruitful discussions. This work is partly supported by the ANR (French Research National Agency) funded project Bingo2 ANR-07-MDCO-014 (http://bingo2.greyc.fr/), which is a follow-up of the first Bingo project and the Czech-French PHC Barrande project “Heterogeneous Data Fusion for Genomic and Proteomic Knowledge Discovery”.

REFERENCES

Bailey, J., Manoukian, T., & Ramamohanarao, K. (2002). Fast algorithms for mining emerging patterns. *Proceedings of the Sixth European Conference on Principles Data Mining and Knowledge Discovery (PKDD’02)* (pp. 39-50). Helsinki, Finland: Springer.

Bayardo, R. J. (2005). The hows, whys, and whens of constraints in itemset and rule discovery. *Proceedings of the workshop on Inductive Databases and Constraint Based Mining* (pp. 1-13) Springer.

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

Blachon, S., Pensa, R. G., Besson, J., Robardet, C., Boulicaut, J.-F., & Gandrillon, O. (2007). Clustering formal concepts to discover biologically relevant knowledge from gene expression data. *Silico Biology, 7.*

Bonchi, F., & Lucchese, C. (2006). On condensed representations of constrained frequent patterns. *Knowledge and Information Systems, 9*, 180-201.

Boulicaut, J.-F., Bykowski, A., & Rigotti, C. (2003). Free-sets: A condensed representation of boolean data for the approximation of frequency queries. *Data Mining and Knowledge Discovery journal, 7*, 5-22. Kluwer Academics Publishers.

Bresson, C., Keime, C., Faure, C., Letrillard, Y., Barbado, M., Sanfilippo, S., Benhra, N., Gandrillon, O., & Gonin-Giraud, S. (2007). Large-scale analysis by sage reveals new mechanisms of v-erba oncogene action. *BMC Genomics, 8.*

Brigelius-Flohe, R. (2006). Glutathione peroxidases and redox-regulated transcription factors. *Biol Chem, 387*, 1329-1335.

Calders, T., Rigotti, C., & Boulicaut, J.-F. (2005). A survey on condensed representations for frequent sets. *Constraint-Based Mining and Inductive Databases* (pp. 64-80). Springer.

Crémilleux, B., & Boulicaut, J.-F. (2002). Simplest rules characterizing classes generated by delta-free sets. *Proceedings 22nd Int. Conf. on Knowledge Based Systems and Applied Artificial Intelligence* (pp. 33-46). Cambridge, UK.

De Raedt, L., Jäger, M., Lee, S. D., & Mannila, H. (2002). A theory of inductive query answering. *Proceedings of the IEEE Conference on Data Mining* (ICDM'02) (pp. 123-130). Maebashi, Japan.

De Raedt, L., & Zimmermann, A. (2007). Constraint-based pattern set mining. *Proceedings of the Seventh SIAM International Conference on Data Mining.* Minneapolis, Minnesota, USA: SIAM.

DeRisi, J., Iyer, V., & Brown, P. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science, 278*, 680-686.

Dong, G., & Li, J. (1999). Efficient mining of emerging patterns: discovering trends and differences. *Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD'99)* (pp. 43-52). San Diego, CA: ACM Press.

Halliwell, B. (2007). Biochemistry of oxidative stress. *Biochem Soc Trans, 35*, 1147-1150.

Hand, D. J. (2002*). ESF exploratory workshop on pattern detection and discovery in data mining, 2447 of Lecture Notes in Computer Science*. Chapter Pattern detection and discovery, 1-12. Springer.

Hébert, C., Blachon, S., & Crémilleux, B. (2005). Mining delta-strong characterization rules in large sage data. *ECML/PKDD'05 Discovery Challenge on gene expression data co-located with the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'05)* (pp. 90-101). Porto, Portugal.

Klema, J., Blachon, S., Soulet, A., Crémilleux, B., & Gandrillon, O. *Constraint-based knowledge discovery from sage data*. Submitted.

Klema, J., & Zelezny, F. In P. Berka, J. Rauch and D. J. Zighed (Eds.), *Data mining and medical knowledge management: Cases and applications, chapter Gene Expression Data Mining Guided by Genomic Background Knowledge*. IGI Global.

Knobbe, A., & Ho, E. (2006). Pattern teams. *Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'06)* (pp. 577-584). Berlin, Germany: Springer-Verlag.

Korotkina, R. N., Matskevich, G. N., Devlikanova, A. S., Vishnevskii, A. A., Kunitsyn, A. G., & Karelin, A. A. (2002). Activity of glutathione-metabolizing and antioxidant enzymes in malignant and benign tumors of human lungs. *Bulletin of Experimental Biology and Medicine, 133*, 606-608.

Li, J., Liu, G., & Wong, L. (2007). Mining statistically important equivalence classes and delta-discriminative emerging patterns. *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'07)* (pp. 430-439). New York, NY, USA: ACM.

Madeira, S. C., & Oliveira, A. L. (2004). Biclustering algorithms for biological data analysis: A survey. IEEE/ACM Trans. Comput. *Biology Bioinform., 1*, 24-45.

Mannila, H., & Toivonen, H. (1997). Levelwise search and borders of theories in knowledge discovery. *Data Mining and Knowledge Discovery, 1*, 241-258.

Morik, K., Boulicaut, J.-F., & (eds.), A. S. (Eds.). (2005). *Local pattern detection, 3539* of *LNAI*. Springer-Verlag.

Nasr, M., Fedele, M., Esser, K., & A, D. (2004). GPx-1 modulates akt and p70s6k phosphorylation and gadd45 levels in mcf-7 cells. *Free Radical Biology and Medicine, 37*, 187-195.

Ng, R. T., Lakshmanan, V. S., Han, J., & Pang, A. (1998). Exploratory mining and pruning optimizations of constrained associations rules. *Proceedings of ACM SIGMOD'98* (pp. 13-24). ACM Press.

Pan, F., Cong, G., Tung, A. K. H., Yang, Y., & Zaki, M. J. (2003). CARPENTER: finding closed patterns in long biological datasets. *Proceedings of the 9th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'03)* (pp. 637-642). Washington, DC, USA: ACM Press.

Pensa, R., Robardet, C., & Boulicaut, J.-F. (2005). A bi-clustering framework for categorical data. *Proceedings of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'05)* (pp. 643-650). Porto, Portugal.

Rioult, F., Boulicaut, J.-F., Crémilleux, B., & J., B. (2003). Using transposition for pattern discovery from microarray data. *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD'03)* (pp. 73-79). San Diego, CA.

Siebes, A., Vreeken, J., & Van Leeuwen, M. (2006). Item sets that compress. *Proceedings of the Sixth SIAM International Conference on Data Mining*. Bethesda, MD, USA: SIAM.

Soulet, A. (2007). Résumer les contrastes par l'extraction récursive de motifs. Conférence *sur l'Apprentissage Automatique (CAp'07)* (pp. 339-354). Grenoble, France: Cépaduès Edition.

Soulet, A., & Crémilleux, B. (2005). An efficient framework for mining flexible constraints *Proceedings 9th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'05)* (pp. 661-671). Hanoi, Vietnam: Springer.

Soulet, A., & Crémilleux, B. (2008). Soulet A., Crémilleux B. Mining constraint-based patterns using automatic relaxation. *Intelligent Data Analysis, 13*(1). IOS Press. To appear.

Soulet, A., Crémilleux, B., & Rioult, F. (2004). Condensed representation of emerging patterns. *Proceedings 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'04)* (pp. 127-132). Sydney, Australia: Springer-Verlag.

Soulet, A., Klema, J., & Crémilleux, B. (2007). *Post-proceedings of the 5th international workshop on knowledge discovery in inductive databases in conjunction with ECML/PKDD 2006 (KDID'06), 4747 of Lecture Notes in Computer Science, chapter Efficient Mining under Rich Constraints Derived from Various Datasets*, 223-239. Springer.

Vaarala, M. H., Porvari, K. S., Kyllonen, A. P., Mustonen, M. V., Lukkarinen, O., & Vihko, P. T. (1998). Several genes encoding ribosomal proteins are over-expressed in prostate-cancer cell lines: confirmation of l7a and l37 over-expression in prostate-cancer tissue samples. *Int. J. Cancer, 78*, 27-32.

Velculescu, V., Zhang, L., Vogelstein, B., & Kinzler, K. (1995). Serial analysis of gene expression. *Science, 270*, 484-487.

KEY TERMS

Background Knowledge: Information sources or knowledge available on the domain (e.g., relational and literature databases, biological ontologies).

Constraint: Pattern restriction defining the focus of search. It expresses a potential interest given by a user.

Functional Genomics: Functional Genomics hints at understanding the function of genes and other parts of the genome.

Local Patterns: Regularities that hold for a particular part of the data. It is often required that local patterns are also characterized by high deviations from a global model (Hand, 2002).

Recursive Mining: Repeating the pattern mining process on output to reduce it until few and relevant patterns are obtained.

SAGE Method: SAGE produces a digital version of the thanscriptome that is made from small sequences derived from genes called "tags" together with their frequency in a given biological situation.