Int J Software Informatics, Vol.2, No.2, December 2008, pp. 215–231 E-mail: ijsi@iscas.ac.cn
International Journal of Software and Informatics, ISSN 1673-7288 http://www.ijsi.org
 Tel: +86-10-62661040

Mining Gene Expression Data using Domain Knowledge*

Nicolas Pasquier[1], Claude Pasquier[2], Laurent Brisson[3], and Martine Collard[1]

[1](Laboratoire I3S, Université de Nice Sophia-Antipolis/CNRS UMR-6070, 06903 Sophia Antipolis, France, {nicolas.pasquier,martine.collard}@unice.fr)

[2](Institute of Developmental Biology and Cancer, Université de Nice Sophia-Antipolis/CNRS UMR-6543, Parc Valrose, 06108 Nice, France, claude.pasquier@unice.fr)

[3](Département LUSSI, TELECOM Bretagne, CNRS FRE-3167, Technopôle Brest-Iroise, 29238 Brest, France, laurent.brisson@telecom-bretagne.eu)

Abstract Biology is now an information-intensive science and various research areas, like molecular biology, evolutionary biology or environmental biology, heavily depend on the availability and the efficient use of information. Data mining, that regroups several techniques for analyzing very large datasets, is used to solve problems in an increasing number of biological applications. This article focuses on the analysis of transcriptome, that reflects gene activity in a given cell population at a given time. We describe research themes in transcriptomics related to domain knowledge in biology. We are particularly interested in the way this knowledge can be efficiently combined and used during the various phases of a data mining process, in the most acknowledged applications in transcriptomics.

Key words: data mining; bioinformatics; clustering; association rules; gene expression data; gene annotations; knowledge integration

Pasquier N, Pasquier C, Brisson L, Collard M. Mining gene expression data using domain knowledge. *Int J Software Informatics*, **2008, 2(2): 215–231. http://www.ijsi.org/1673-7288/2/215.pdf**

1 Introduction

Over the last decade, a great number of genomes, from different organisms have been decoded. The knowledge of a genome sequence is an important step towards understanding it. However, the sequence itself provides little information about the role of genes contained within a genome. Old issues remain like: What are the functions of the different genes? In what cellular processes do they participate? How are genes regulated? In which cell types and depending on which conditions the genes become active? How various diseases or treatments influence the activity of genes? Or the reverse: How genes contribute to diseases?

Transcriptomics or global analysis of gene expression, also called genome-wide expression profiling, is a way to answer these questions. While the genome represents an inventory of every available gene in an organism, the transcriptome reflects the genes

* Corresponding author: Nicolas Pasquier, Email: nicolas.pasquier@unice.fr
Manuscript received 15 Oct., 2008; revised 4 Dec., 2008; accepted 16 Dec., 2008; published online 18 Dec., 2008.

that are being actively expressed at any given time. By studying the patterns of gene expression in different experimental conditions, researchers can get an understanding of genes and pathways involved in biological processes. A gene expression level is a numerical value assessing how this gene was over-expressed (intensively active) or under-expressed (weakly active) compared with his activity in normal conditions (apart from the experiment). Transcriptomics aims at discovering genes involved in similar biological processes using expression level measures. So called "in-silico" annotations are deduced from overall gene expression measurements in particular experimental contexts.

There are several methods to profile the expression of thousands of genes in parallel. These include hybridization-based technologies, such as DNA microarrays[1], and sequencing-based approaches like SAGE (Serial Analysis of Gene Expression)[2] and MPSS (Massively Parallel Signature Sequencing)[3]. Based on different principles, hybridization-based and sequencing-based technologies should be considered complementary to each other, rather than competitive alternatives for measuring gene expression, and currently, both are important tools for transcriptome profiling[4].

Several kinds of data mining techniques are currently used on biological data to extract knowledge on differentially expressed genes or co-expressed genes or other relevant patterns. To improve relevance and utility of extracted knowledge, most of these applications require to extend existing techniques to adapt them to biological data. The next challenge for this purpose is to integrate *biological knowledge* in all phases of the data mining process to optimize existing knowledge profit[5]. Biological knowledge refers to biological information describing known gene properties and relationships. This knowledge is spread over heterogeneous sources of information such as research papers, biological ontologies or regulation networks for instance. In a more general way, we call *domain knowledge* all information related to the domain studied and we refer to it as *background knowledge* or *apriori knowledge* when it is used in the data mining process.

Biological knowledge is widely available from public sources. Currently, most of the information is stored in plain text format into millions of biological research papers. However, a growing number of repositories make their data available in more structured formats, sometimes organized with ontologies. A widely used source of annotations is the Gene Ontology (GO)[6]. GO is a controlled vocabulary developed by a consortium of scientists that is used to describe ('annotate') a gene or a product of a gene in regard to its molecular functions (its activities or abilities, i.e. the catalysis of a biochemical reaction), cellular components (its localizations in the cell, i.e. in the plasma membrane) and biological processes (the processes in which it participates, i.e. the respiration). Other gene centric annotations include phenotypic annotations (the measurable characteristics of an organism controlled by genes), disease annotations (the propensity of genetic diseases associated with genes), tissue-expression patterns (the association of genes with the tissues in which they are preferentially expressed) and homology information (genes in different species that share a common ancestor). However, as no gene operates in an isolated way, it is also important to consider information about the complex molecular networks orchestrating the activity of cells. These networks and their constituents are depicted into compendia of pathways: Transcriptional, translational and regulatory pathways describe protein biosynthesis and

its regulation; Genetic and physical interactions describe the interactions occurring between genes and between proteins respectively; Metabolic pathways describe the series of chemical reactions occurring within a cell while signal transduction pathways describe the system of communication between cells.

After an overview of the field of transcriptomics presented in Section 2, the approaches for driving expression data analysis with background knowledge are presented in three main trends. These three trends correspond to different methods of representation and use of background knowledge. Section 3 presents methodologies using background knowledge as an a priori assumption on data to supervise the mining process. In these methodologies, background knowledge is used to drive the dataset preparation, by identifying and selecting relevant genes or experiments for instance, or the search space traversal, by pruning irrelevant patterns for instance, during the mining process. Section 4 shows how background knowledge can be used during the post-processing phase to improve the relevance of extracted patterns and simplify the end-user's task. In these methods, extracted patterns are re-formated, compared, evaluated or visualized according to background knowledge. These methods require that background knowledge is represented in a format enabling comparisons with extracted patterns. Section 5 is devoted to data mining methods that take into account expression data and background knowledge simultaneously to extract patterns showing relationships between both. These methods integrate in a single framework expression data and biological knowledge from several heterogeneous sources such as bibliographic databases, research papers and texts, bio-ontologies or semantic networks for instance.

2 Bioinformatics and Transcriptomics

High throughput techniques used in biological research are routinely producing an extraordinary amount of data. These data need to be stored, analyzed and interpreted to serve the progress of knowledge. Applying data management techniques to handle biological data is challenging because data are spread over the web, hosted in many independent, heterogeneous and highly focused data repositories. In addition, biological data are of diverse types, including experimental measures, digital images, 3D structures or sequences. In conjunction with these raw data, produced by biological experiments, researchers have access to a domain knowledge which is also widely available from public sources. Most of the information is stored in millions of biological research papers. Other sources of information include repositories of gene's functions, activities, similarities, interactions, mutations, homologies or implication in diseases.

2.1 Biological data

To answer these questions, one needs to look at the activity of genes. Indeed, a genome represents an inventory of every available gene in an organism but few of them are active at a given time. When a gene is active, the gene's information is copied in a ribonucleic acid (RNA) in a process called transcription. Some RNA fragments, called messenger RNA (mRNA), are translated in proteins, which are one of the active compounds of cells. Other kinds of RNA, called non-coding RNA, can be involved in the translation process, in the regulation of genes' activity or may have

a catalytic activity. Therefore, the activity of a gene can be assessed by measuring the abundance of the matching RNA.

There are several methods to profile the expression of thousands of genes in parallel. These include hybridization-based technologies, such as DNA microarrays, and sequencing-based approaches like SAGE (Serial Analysis of Gene Expression) and MPSS (Massively Parallel Signature Sequencing)[4].

The most commonly used technology is DNA microarrays[1]. A DNA microarray works by using the ability of a given DNA molecule to bind specifically to, or hybridize to, its original DNA coding sequence. Many DNA segments (also known as probes or reporters), each one matching a specific RNA, are arrayed on a solid surface. The expression levels of hundreds or thousands of genes within a cell are obtained by measuring the amount of RNA bound to each site on the array.

The older sequencing based approach, which was developed in parallel with the sequencing of the genome, consist in sequencing all cDNAs obtained from all RNAs expressed in a tissue[7]. This technique produces small fragments (called Expressed Sequence Tags or ESTs) of all expressed genes. ESTs are counted and mapped to the DNA from which they are derived, allowing to estimate transcripts abundance. The SAGE technique[2], which is derived from cDNA sequencing, is based on the fact that, most of the time, a small part of a cDNA is sufficient to unambiguously identify it. In this technique, small fragments (also called tags), are cut from the cDNA sequences derived from RNA. These tags are linked together to form long serial molecules that are cloned and sequenced. The abundance of transcripts is then estimated by counting the number of sequenced tags. SAGE allows the sampling of 12 to 20 transcripts per sequencing reaction, compared to one EST with cDNA sequencing. MPSS technology[3] is another sequencing based approach. In this technique, cDNA sequences are placed on microbeads (with each microbead containing only one cDNA) and sequenced in parallel. The abundance of transcript is evaluated by counting the number of beads matching a specific sequence. With MPSS the number of sequences obtained is much more larger than with SAGE as more than 1 million cDNA can be sequenced simultaneously[8].

2.2 Transcriptomics' background knowledge

One crucial source of information that must be considered when dealing with gene expression data is information about the way gene expression measures were obtained. Information about the technology used and the way experiments were performed is of the highest importance because it influences the pre-processing of the data. Information about experimental contexts is also needed for analyzing gene expression measurements. For example, comparing the activity of genes in a healthy and cancerous tissue may give some hints about the genes that are involved in cancer. However, this approach is very limited because many of the genes serve multiple functions and changes in gene expression can be due to factors not directly connected to the experiment under study[9]. Deeper and more accurate analyzes require the use of other sources of information.

A widely used source of annotations is the Gene Ontology (GO)[6]. GO is a controlled vocabulary developed by a consortium of scientists. It is used to describe ('nnotate') a gene or a product of a gene in regard to its molecular functions (its activ-

ities or abilities, i.e. the catalysis of a biochemical reaction), cellular components (its localizations in the cell, i.e. in the plasma membrane) and biological processes (the processes in which it participates, i.e. the respiration). Other gene centric annotations include phenotypic annotations (the measurable characteristics of an organism controlled by genes), disease annotations (the propensity of genetic diseases associated with genes), tissue-expression patterns (the association of genes with the tissues in which they are preferentially expressed) and homology information (genes in different species that share a common ancestor).

However, no gene operates in an isolated way. The activity of a cell is orchestrated by complex molecular networks consisting of entities such as proteins or RNAs connected by different kind of interactions. Information about this network and its constituents includes compendia of pathways describing different aspects of genes interactions. Protein biosynthesis and its regulation are depicted in transcriptional, translational and regulatory pathways. Genetic and physical interactions describe the interactions occurring between genes and between proteins respectively. At the cellular level, metabolic pathways describe the series of chemical reactions occurring within the cell while signal transduction pathways describe the system of communication between cells.

3 Process Supervisation using Background Knowledge

3.1 Biological data preparation, filtering and selection

Background knowledge based approaches use domain knowledge to direct the mining process during the dataset preparation phase, to select data that are relevant to the mining task, or during the computation phase, to prune irrelevant results. These approaches use gene expression measures to discover co-regulated genes, but the task of interpreting these links from a biological viewpoint is left to the expert as a post-processing phase[10, 11].

The use of analysis techniques for processing gene expression data must cope with noise, or random variations, inherent to living systems[12]. Replication is the key to produces more consistent and reliable findings despite such noise. Shortly after the advent of microarrays, it was suggested that at least three replicates be used in designing experiments[13]. Several techniques can be used to extract the genes whose expression level varies significantly, from all replicates. One of the most commonly used method, known as the "fold change method", is straightforward: It consists in selecting genes whose expression level varies by a predetermined threshold (usually by a factor of 2). A method that relies on more statistically motivated criteria is SAM (Significance Analysis of Microarrays). It uses the conventional t tests to estimate the false discovery rate (FDR), which is the expected proportion of false positive among all tests. This method allows to adjust the threshold in order to correspond to an acceptable FDR. Some researchers use the analysis of variance (ANOVA) to discriminate between the variability explained by experimental factors and the variability due to random noise.

3.2 Gene expression data clustering

Gene-based clustering is the process of grouping genes into a set of classes (or

clusters) according to their expression in given experimental conditions (samples or time points). Each cluster intends to contain *co-expressed genes* that exhibit a common expression profile. Gene-based clustering was investigated to understand biological processes since genes grouped in the same cluster are expected to be involved in common biological processes. Most popular gene-based clustering algorithms are partitional, self-organized maps (SOM) and hierarchical. Although clustering techniques proved to be useful for identifying co-expressed genes, interpreting gene co-expression without ambiguity has remained a challenge since it depends on other sources of knowledge such as expert knowledge of biologists[14]. While the seminal paper by Eisen *et al.*[15] showed clusters which significance was demonstrated by common functional categorization, other works concluded that further statistical analyzes were required. Indeed, a microarray dataset contains numerous groups of co-expressed genes. Then, a typical strategy for a biologist is to start from genes which are known to be closely related to a biological function and to browse a preliminary rough clustering result, to focus on a small subset of those genes which are supposed to play a role. Thus, currently, biologists follow exploratory strategies by manually selecting potential groups of genes according to their knowledge.

A first attempt to provide more automatic solutions for a relevant clustering was *background knowledge based approaches* which integrate knowledge in a preliminary stage of the clustering process. Prior knowledge of biologists may be brought into gene-based clustering either by introducing assumptions or constraints, like in semi-supervised analysis and bi-clustering, or by initiating the clustering with complementary data sources like ontological annotations. *Semi-supervised clustering*[16, 17] uses existing domain knowledge to guide the clustering process either by constraints or by specific distances. For instance, in Ref.[16] *must-link* and *cannot-link* constraints are defined with associated costs of violation. A unified model for semi-supervised clustering with constraints proposed by Segal *et al.*[18] combines a binary Markov network derived from constraints on pairwise protein interaction data and a Naive Bayes Markov network modeling gene expression data.

Bi-clustering techniques also referred as *subspace clustering* for microarray data[19−22] enhance simple gene-based clustering by supplying knowledge for selecting clusters as sub-matrices of the initial datasets. A bi-cluster is defined as a subset of genes that exhibit compatible expression patterns over a subset of conditions (samples or time points). Bi-clustering reveals groups of genes that are co-regulated only under specific conditions and are independent under other conditions. The underlying assumption that genes are active only over some, but not all, conditions has been demonstrated as quite relevant for different organisms and datasets[23]. More generally, subspaces of the gene expression dataset may be defined as submatrices satisfying some constraints. In Ref.[24], the fluctuation and trend constraints require that for all genes in a cluster the differences of expression levels between two conditions are similar and the expression levels of two genes are correlated. *Annotation-based clustering* build co-annotated gene groups sharing common genomic and biomedical annotations in a first stage. Afterwards, they integrate the gene expression profiles into co-annotated groups and highlight groups of co-expressed genes. Finally, the statistical significance of co-annotated and co-expressed gene groups is tested. Annotations generally come from public available knowledge bases. These approaches[25−28]

depend mainly on the availability and completeness of the annotation bases.

3.3 Pattern extraction from gene expression data

Pattern extraction techniques aim at discovering correlations and links between data that are represented as association rules and frequent patterns. For applicability and performance reasons, most of these techniques work on categorical data and require that numerical values are discretized during a pre-processing phase. Hence, biostatistical methods are generally used to discretize numerical gene expression measures into gene expression levels indicating if a gene was *under-expressed*, *unchanged* or *over-expressed* in the corresponding biological condition. In the resulting N×M data matrix, each of the N lines corresponds to a gene and contains M expression levels corresponding to the M experimental conditions.

Association rules are conditional rules expressing correlations between sets of attribute values, called *items*, in data lines. An association rule $A \rightarrow C$ states that a significant proportion of data lines containing items in A also contain items in C. The *frequent itemsets* framework for association rule discovery (ARD) was introduced with the seminal Apriori algorithm. Frequent itemsets are sets of items contained in a significant proportion of data lines and association rules are straightforwardly generated from them. Several extensions of Apriori, using optimized dataset representations, data structures, search space traversals and redundant rules filtering methods, have been proposed[29].

ARD was first applied to gene expression data for generating association rules between gene under- and over-expressions such as: *gene1*[↑] → *gene2*[↑], *gene3*[↑]. This rule states that in a significant number of biological conditions, when *gene1* is over-expressed it is likely to observe an over-expression of *gene2* and *gene3*. Applications of the Apriori algorithm[30, 10] and its extensions[31, 32] for such global gene expression profiling pointed out co-regulated genes supported by recent biological literature. These applications showed that methods for filtering and selecting rules are required when using frequent itemset based approaches as a huge number of rules, containing many redundant rules, are generated when data is dense. Different solutions were investigated to address this problem: The use of post-processing techniques to filter and explore extracted rules[33], the selection of the top-k most interesting rules according to a statistical criterium[34, 35], the fusion of rules according to gene co-regulation significance[36] and the evaluation of rules significance using biostatistical measures[37]. The *frequent closed itemsets* framework was introduced for ARD from dense data. Frequent closed itemsets are a minimal representation for frequent itemsets and consequently allow to reduce the search space of ARD. See[38] for an extensive review of frequent closed itemset based approaches. Several of these approaches were applied for mining association rules[39, 40], condensed representation of association rules[30] and maximal sets of co-regulated genes[41] from gene expression data. These applications showed that frequent closed itemset based approaches both improve extraction efficiency and exclude redundant patterns when mining gene expression data as they are dense[35, 40].

These applications showed several ARD features suggesting that association rules can reveal patterns that might not have been revealed using clustering and vice versa. First, the data pre-processing phase allows to address independently problems of noise

in the data[42, 10] that is inherent to biological systems[12]. Second, a gene expression can appear in any number of association rules and thus all its links with different sets of genes can be enlighten during the same experiment[10, 41]. Third, association rules are directed relationships and thus provide deeper insights into specific relationships than clustering[33]. Fourth, association rules can integrate both numerical data such as gene expression measures and categorical data such as gene annotations or biological condition information[42, 10]. These features suggest that ARD and clustering are complementary methods which results should be combined to achieve biological knowledge[30,10,11].

4 Background Knowledge Based Post-Processing

Since modeling step can extract lots of patterns (rules, clusters, decision trees), a post-processing step is necessary to filter, to re-format and to evaluate them. During this step, domain experts have to assess how extracted knowledge meets business objectives and success criteria. However, although some statistical indexes can measure accuracy or precision they cannot measure the real interestingness of discovered knowledge for the domain expert. Consequently, the use of domain knowledge representation is necessary to evaluate and to validate extracted knowledge.

4.1 Extracted pattern management

A main issue with pattern management is to deal with heterogeneous pattern representation. Since extracted patterns could be trivial or irrelevant according to domain knowledge, analyzing and accessing patterns is a laborious task and it is necessary to store, query, compare and combine various patterns in a unified way.

4.1.1 Pattern representation

Several approaches define logical models for pattern representation, for example the Predictive Model Markup Language (PMML), an XML representation language for data mining and statistical model sharing, and the Common Warehouse Model for Data Mining (CWM-DM), a specification for data mining metadata exchanges. Although these approaches are well-suited for data model sharing they seem inadequate to represent and handle different classes of patterns in a flexible, effective and coherent way because predefined pattern types are considered. Rizzi *et al.* introduced Pattern Base Management Systems (PBMS) to provide a new logical model for patterns describing a model structure, a measure evaluating pattern interestingness, a raw data source and a formula to map raw data space to model space[43]. The main interest of PBMS is to provide flexibility to incorporate novel pattern types and mechanisms for constructing composite patterns. Although this approach improves heterogeneous pattern management, it is necessary to use advanced semantic information in pattern representation[44].

4.1.2 Management systems for querying and indexing patterns

From an architectural point of view, the next step after defining a model for pattern representation is to consider storage, indexation and query aspects. There are some standardization projects providing end-user tools: SQL/MM DM, a standard which defines an Application Programming Interface (API) to access data and Java Data Mining (JDM) API, offering a standard way to handle data and metadata

supporting data mining models. However these standards are more centered on the data mining process and they lack pattern management solutions.

Another approach is the use of Inductive Databases (IDB). In such databases data and patterns are handled in the same way with an inductive query language. This language is an extension of a database query language that allows us to select, manipulate and query data or patterns satisfying some user-specified constraints.

Rizzi *et al.* developed an approach slightly different from IDB[43]. They argue that a logical separation between database and pattern-base is needed to ensure efficient handling of both raw data and patterns through dedicated management systems and they defined PBMS. PBMS provides methods for representing and storing patterns, but also for processing queries and for efficiently retrieving patterns. Kotsifakos *et al.* improved PBMS architecture by enabling the support of domain ontologies[45]. They propose the integration of PBMS and ontologies as a solution to the need of many scientific fields for efficient extraction of useful information from large databases. Their ontology-enhanced PBMS is independent of any data mining engine and uses XML to store patterns in the pattern base. Their PBMS provides pattern filtering functionalities using ontologies to automate the pattern evaluation step. Thus, pattern querying functionalities are greatly improved with the use of domain specific knowledge. However, it is necessary to define semantic similarity between objects in an ontology to validate the extracted patterns.

4.2 Extracted pattern validation

Issues in evaluating and interpreting results of the mining process are currently major research challenges. Detailed studies have been devoted to interestingness measures. A consensus among researchers is now established to consider objective interestingness and subjective interestingness. Objective interestingness is traditionally evaluated by various of statistical indexes[46] while subjective interestingness is generally evaluated by comparing discovered patterns to user knowledge or *a priori* convictions of domain experts. A way to improve subjective interestingness measures is to deeply explore expert knowledge and source data in order to formalize them in conceptual structures and exploit these structures for flexible model interpretation.

4.2.1 Subjective interestingness measures

Numerous works focus on indexes that measure the interestingness of a mined pattern[47] and propose unexpectedness and actionability as subjective criteria. According to the actionability criteria, a model is interesting if the user can start some action depending on it[48]. On the other hand, unexpected models are considered interesting since they contradict user expectations which depend on his beliefs. User expectations is a method developed by Liu[47]. The user has to specify a set of patterns according to his previous knowledge and intuitive feelings, then the system matches them against each discovered patterns using a fuzzy matching technique. Silberschatz and Tuzhilin propose a method to define unexpectedness via belief systems[48]. A pattern is said to be interesting relatively to some belief system if it "affects" this system, and the more it "affects" it, the more interesting it is.

4.2.2 Ontology-Based validation methods

Subjective interestingness measures were developed in order to give an insight

into real human interest. However, these measures lack semantic formalization, and force the user to express all of his expectations. Consequently, the extracted pattern validation process must involve not only the study of patterns and domain experts' expectations but also the use of a domain ontology. Thus, rules expressed to filter out noisy patterns or to select the most interesting ones will be relevant.

Kotsifakos *et al.* present an approach based on PBMS where a subgraph of the ontology that contains the attributes under consideration is constructed[45]. Then, if some of the attributes are not in the subgraph, rules containing them are marked as "noisy". Brisson and Collard propose the KEOPS methodology[49]. They suggest comparing extracted rules with expert's knowledge. By comparing coverage the most informative rule is deduced, i.e. the rule predicting the largest consequence from the smallest condition. Finally, the IMAK interestingness measure, both objective and subjective, is applied.

An important issue in ontology-based validation methods is the definition of semantic similarity measures between ontology concepts.There are two kinds of methods in order to measure semantic similarity within an ontology: *edge counting* methods and *information-theoretic* based methods. *Edge counting* methods involve calculating the distance between concepts in the ontology, similarity decreasing while distance increasing. Leacock and Chodorow measure semantic similarity by finding the shortest method distance between two concepts and then scale the distance by the maximum distance in the "is-a" hierarchy[50]. Zhong *et al.* define weights for the links according to their position in the taxonomy[51]. Resnik introduces *information-theoretic* measures[52] based on the information content of the lower common ancestor of two concepts and demonstrates that such methods are less sensitive, and in some cases not sensitive, to the problem of link density variability[52]. Lin improves Resnik's measure considering how close the concepts are to their lower common ancestor[53]. Jiang presents a combined approach that inherits the edge counting based approach, which is enhanced by the node-based approach of the information content calculation[54]. Lord compares Resnik's, Lin's and Jiang's measures in order to use them to explore GO. His results suggest that all three measures show a strong correlation between sequence similarity and molecular function semantic similarity[55]. He concluded that none of the three measures has a clear advantage over the others, although each has strengths and weakness. Schlicker *et al.* introduce a new measure of similarity between GO terms that is based on Resnik's and Lin's definitions[56]. This measure takes into account how close these terms are to their lower common ancestor and a uses a score allowing one to identify functionally related gene products from different species that have no significant sequence similarity.

5 Expression Data and Biological Knowledge Integration

Integrating biological knowledge and expression data in a single framework is a major challenge to improve relevance of mined patterns and simplify their interpretation by the biologists. This section reviews mining applications to datasets integrating both expression data and biological knowledge from various knowledge bases such as bio-ontologies, descriptions of regulation pathways and literature.

5.1 Heterogeneous data integration

Although a large amount of information is accessible to researchers, it is often difficult to use it because this information is spread over different sources, represented under different formats and, most of the time, generated with different techniques that make it hardly comparable. This is the case for gene expression measurements. Special care is needed when gene expression measures are generated by different technologies because they cannot be easily merged. Hybridization-based technologies measure the ratios of expression changes while sequencing-based approaches produce an estimation of the number of transcripts. Several studies, comparing microarrays with SAGE or MPSS conclude in a poor overlapping in the set of expressed genes revealed by these technologies[57]. Hybridization-based technologies show greater consistency across platforms than sequencing-based approaches. To ease the comparison among different microarray experiments, public databases currently require that microarray data be encoded in the MIAME format[58]. MIAME includes information about the design of the experiment and microarray layout, the preparation of the biological samples, the protocol used to hybridize the sample, the way intensities are quantified and the method used to normalize data. Merging measurements based on abundance is much more difficult because experimental conditions are not strictly described and technology differences can have more impact on the results than biological differences between samples and tissues[57].

Integrating domain knowledge is even more challenging. Classical approaches to data integration[59] face difficulties which are amplified because of some specificities of biological knowledge (see Ref.[60] for a detailed description). Because of these specificities, biological knowledge integration is done, most of the time, on a case by case basis. Excepted in the rare cases where all necessary information is gathered in a dedicated database, hands-off data integration in life science is still impracticable.

Data integration approaches range from light solutions like link integration and Web 2.0 mashups to heavyweight mechanisms like data warehousing and view integration. See Ref.[61] for a review of popular approaches to data integration for bioinformatics. It is envisioned that Semantic Web technologies, which provide a common framework allowing data to be shared and reused between applications, might be well suited for managing disseminated biological data. The Semantic Web Health Care and Life Sciences Interest Group (HCLSIG) was launched to explore the application of these technologies in various areas[62].

5.2 Co-Clustering techniques

As previously argued in this article, gene expression data are noisy and gene-based clustering gains more reliability if other sources of knowledge are associated to the process. *Co-clustering* techniques are solutions for clustering datasets combining gene expression data and other kinds of data modeling knowledge. Generally, integrated knowledge data are annotations from biological knowledge bases and informative data on metabolic pathways or protein interactions. Standard clustering algorithms are then applied to combined datasets and the effort focuses on distance functions and cluster quality measures.

The application of a standard hierarchical clustering method to integrated metabolic network knowledge and gene expression data is reported in Ref.[63]. A graph

distance function was defined for metabolic networks and a correlation-based distance was used for expression data. Clusters were evaluated according to three distances: a distance on expression data only, a distance on networks only and a combined distance. The last one was able to yield clusters with low internal distances according to both expressions and metabolic networks. In Ref.[64], clustering was applied to different datasets to assess the value of gene annotation integration in the clustering process. The first dataset contained only expression data while other datasets integrated expression data with enzyme classification annotations. Several SOM clusterings were performed and clusters were evaluated by measuring how genes in clusters were correlated regarding functional and metabolic annotations. Experiments showed that clusters mined from integrated expression data and annotations were better, that is the correlation between genes was higher in clusters extracted from combined datasets. Clustering was also applied to integrated gene expression data and protein interaction data[65, 66]. In Ref.[65], an EM algorithm was used to extract a unified probabilistic model for identifying pathways. Expression data were clustered in a first step and each cluster, considered as defining a pathway, was the input of the probabilistic algorithm for protein interactions. The biological coherence of clusters was evaluated according to GO functional annotations. In Ref.[66], a graph-based approach taking into account noise in data was used. Experimental results showed interesting relations between co-expressed genes and interacting proteins.

5.3 Pattern extraction from heterogeneous data

Pattern extraction from integrated gene expression data and biological knowledge (annotations or class information) aims at mining association rules or classification rules describing relations between biological functions and co-regulated genes.

The application of a frequent itemset based algorithm, combined with a statistical test of significance for pruning redundant rules, to mine association rule with the form *annotation* $\rightarrow$ *gene expressions* from integrated gene expression data and annotations was reported in Ref.[42]. The application of a frequent closed itemset based approach to extract non-redundant association rules, not constrained in their form, from integrated gene expression data and annotations was studied in Ref.[11]. In both applications, extracted association rules showed biologically meaningful relations between gene expressions and annotations supported by recent biological literature. They also confirmed that tackling the problem of redundant association rules is necessary for better result's quality and interpretability.

In classification, a model, called classifier, is built from a training dataset with given class labels and then used to classify data of unknown class label. Several classification techniques, such as support vector machines, neural networks and decision trees, were applied to gene expression data[67]. Recently, pattern based classifiers constructed from association rules or frequent patterns were also applied. The use of a frequent itemset based algorithm to generate classification rules with the form *gene expressions* $\rightarrow$ *class* from gene expression data of cancerous and healthy tissues (cancer data) was studied in Ref.[68]. Another study addressed the application of a frequent closed itemset based algorithm to extract non-redundant association rules with the same form and generate a rule based classifier from them. This approach was extended in Ref.[34] by selecting the top-k first rules with maximal confidence to

construct the classifier. Experiments conducted on cancer data showed that pattern based classifiers are particularly accurate for such high-dimensional data as they do not suffer from the single coverage constraint and the fragmentation problem[68].

Emerging patterns (EP) are patterns whose support varies significantly between two sets of data. EP have been proposed to capture significant differences between two classes of high dimensional data and construct classifiers from them. The application of a frequent itemset based approach to cancer data for discovering itemsets that are frequents in one tissue and not in the other is reported in Ref.[69]. Classifiers constructed from these EP correctly predicted 57 of the 62 tissues tested. In Ref.[70], a new approach combining EP extraction and a statistical test to evaluate the significance of each EP is proposed. Application to gene expression data classification showed the efficiency of the EP approach for classifiying high-dimensional data. The application of EP classification to analyze a dataset integrating gene expression data and phylogenetic profiles, reflecting whether a gene has a close homologue in the corresponding genome, was studied in Ref.[71]. These experiments showed that EP are efficient for multi-source data classification and that they are, as classification and association rules, easily understandable, a property that is especially important in bioinformatics application problems[34,68−71].

6 Conclusion

After more than one decade of researches in data mining, efficient and scalable techniques for mining relevant patterns from large datasets are available. In the meantime, the development of high-throughput methods for quantitative monitoring of gene expression has generated vast amounts of data about the activity of genes. These gene expression data contain implicit knowledge on the biological role of genes and data mining techniques are well-suited to extract this knowledge from such high-dimensional data. To improve the relevance of extracted patterns, biological prior knowledge about the genes (gene annotations, regulation pathways, biological condition and homology information, protein interactions) should be integrated in the mining process. However, integrating this knowledge is not an easy task since different types of information are represented in various data formats (research papers, digital images, raw or semantically rich data) and stored in heterogeneous data structures (bio-ontologies, knowledge bases, relational databases and bibliographic repositories). The value of combining prior knowledge and gene expression data is quite obvious in each phase of the mining process (data preparation, mining and post-processing) either to select and evaluate data and extracted patterns, or to enhance gene expression datasets to form rich mining contexts and extract more predictive patterns.

In this paper, data mining techniques for gene expression data analysis are reviewed from the viewpoint of their usage of biological knowledge. In background knowledge based approaches, prior knowledge is useful to prepare data to be mined or to drive the mining process. In knowledge based post-processing techniques, this knowledge is used to select the most interesting patterns and provide support to explore and evaluate them. In background knowledge and data integrated co-analysis techniques, prior knowledge is integrated in the data to be mined to extend extracted patterns and simplify their interpretation by the experts.

Several research issues remain to extract patterns that are more adequate to

the biologists' needs and to combine results from different analyzes. An important issue is the use of text mining to extract more valuable knowledge from biological literature. Since most of biological knowledge is represented in such biobliographic sources, automatically extracting this knowledge to integrate it in the mining process could substantially improve biological value of extracted patterns.

References

[1] Adams MD, Kelley J, Gocayne J, Dubnick M, Polymeropoulos M, Xiao H, Merril C, Wu A, Olde B, Moreno R, Kerlavage A, McCombie W, Venter J. Complementary dna sequencing: Expressed sequence tags and human genome project. Science, 1991, 252(5013): 1651–6.

[2] Ashburner M, Ball C, Blake J, Botstein D, Butler H, Cherry J, Davis A, Dolinski K, Dwight S, Eppig J, Harris M, Hill D, Issel-Tarver L, Kasarskis A, Lewis S, Matese J, Richardson J, Ringwald M, Rubin G, Sherlock G. Gene ontology: Tool for the unification of biology. the gene ontology consortium. Nat Genet, 2000, 25(1).

[3] Asyali MH, Colak D, Demirkaya O, Inan MS. Gene expression profile classification: A review. Current Bioinformatics, January 2006, 1(1): 55–73.

[4] Becquet C, Blachon S, Jeudy B, Boulicaut JF, Gandrillon O. Strong-association-rule mining for large-scale gene-expression data analysis: A case study on human sage data. Genome Biol, 2002, 3(12).

[5] BenYahia S, Hamrouni T, Mephu Nguifo E. Frequent closed itemset based algorithms: A thorough structural and analytical survey. SIGKDD Explorations, 2006, 8(1): 93–104.

[6] Bilenko M, Basu S, Mooney RJ. Integrating constraints and metric learning in semi-supervised clustering. In: *Proc. ICML conference*, 2004.

[7] Boulesteix AL, Tutz G, Strimmer K. A cart-based approach to discover emerging patterns in microarray data. Bioinformatics, 2003, 19(18): 2465–2472.

[8] Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M. Minimum information about a microarray experiment (miame)-toward standards for microarray data. Nat Genet, 2001, 29(4): 365–71.

[9] Breitling R, Amtmann A, Herzyk P. Iterative group analysis (iga): A simple tool to enhance sensitivity and facilitate interpretation of microarray experiments. BMC Bioinformatics, 2004, 5.

[10] Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M, Ewan M, Roth R, George D, Eletr S, Albrecht G, Vermaas E, Williams SR, Moon K, Burcham T, Pallas M, DuBridge RB, Kirchner J, Fearon K, Mao J, Corcoran K. Gene expression analysis by massively parallel signature sequencing (mpss) on microbead arrays. Nat Biotechnol, 2000, 18(6): 630–4.

[11] Brisson L, Collard M. An ontology driven data mining process. In: Proc. ICEIS conference, 2008.

[12] Carmona-Saez P, Chagoyen M, Rodríguez A, Trelles O, Carazo J, Pascual-Montano A. Integrated analysis of gene expression by association rules discovery. BMC Bioinformatics, 2006, 7(1): 54–69.

[13] Ceglar A, Roddick JF. Association mining. ACM Computing Surveys, 2006, 38(2).

[14] Chen J, Agrawal V, Rattray M, West MA, St Clair DA, Michelmore RW, Coughlan SJ, Meyers BC. A comparison of microarray and mpss technology platforms for expression analysis of arabidopsis. BMC Genomics, 2007, 8:414.

[15] Chung S, Deng Z, Shu C, Hu D. Clustering analysis of gene expression data based on semi-supervised visual clustering algorithm. Soft Comput, 2006, 10(11): 981–993.

[16] Cong G, Tan KL, Tung AK, Xu X. Mining top-k covering rule groups for gene expression data. In: Proc. ACM SIGMOD Conference, 2005. 670–681.

[17] Creighton C, Hanash S. Mining gene expression databases for association rules. Bioinformatics, 2003, 19(1): 79–86.

[18] Eisen M, Spellman P, Brown PO, Botstein D. Cluster analysis and display of genome wide expression patterns. Proc. Natl. Acad. Sci. USA, 1998, 95(25): 14863–14868.

[19] Fan H, Zhai C, Liu L, Yang J. Subspace clustering for microarray data analysis: Multiple criteria and significance assessment. In: Proc. IEEE CSB Conference, 2004. 582–583.

[20] Goble CA, Stevens R. State of the nation in data integration for bioinformatics. J. Biomed. Inform., 2008.

[21] Gyenesei A, Wagner U, Barkow-Oesterreicher S, Stolte E, Schlapbach R. Mining co-regulated gene profiles for the detection of functional associations in gene expression data. Bioinformatics, May 2007.

[22] Hanisch D, Zien A, Zimmer R, Lengauer T. Co-Clustering of biological networks and gene expression data. Bioinformatics, 2002, 18(Suppl 1): 145–154.

[23] Hene L, Sreenu VB, Vuong MT, Abidi SH, Sutton JK, Rowland-Jones SL, Davis SJ, Evans EJ. Deep analysis of cellular transcriptomes - longsage versus classic mpss. BMC Genomics, 2007, 8: 333.

[24] Huang Z, Li J, Su H, Watts GS, Chen H. Large-scale regulatory network analysis from microarray data: modified bayesian network learning and association rule mining. Decision Support Systems, 2007, 43(4): 1207–1225.

[25] Icev A, Ruiz C, Ryder EF. Distance-enhanced association rules for gene expression. In: Proc. BioKDD Conference, 2003. 34–40.

[26] Jiang D, Pei J, Ramanathan M, Lin C, Tang C, Zhang A. Mining gene-sample-time microarray data: A coherent gene cluster discovery approach. Knowl. Inf. Syst., 2007, 13(3): 305–335.

[27] Jiang J, Conrath DW. Semantic similarity based on corpus statistics and lexical taxonomy. CoRR, cmp-lg/9709008, September 1997.

[28] Jiang XR, Gruenwald L. Microarray gene expression data association rules mining based on jg-tree. In: Proc. DEXA Workshop, 2003. 27.

[29] Jonsson P, Laurio K, Lubovac Z, Olsson B, Andersson ML. Using functional annotation to improve clusterings of gene expression patterns. Information Sciences, September 2002, 145(3-4): 183–194.

[30] Jun J, Chung S, McLeod D. Subspace clustering of microarray data based on domain transformation. Lecture Notes in Computer Science, 2006, 4316: 14–28.

[31] Kim SY, Volsky DJ. Page: Parametric analysis of geneset enrichment. BMC Bioinformatics, 2005. 8.

[32] Kotala P, Zhou P, Mudivarthy S, Perrizo W, Deckard E. Gene expression profiling of dna microarray data using peano count trees. In: Proc. VCGB Conference, 2001.

[33] Kotsifakos E, Marketos G, Theodoridis Y. A framework for integrating ontologies and patternbases, chapter 12. Information Science Reference, July 2007.

[34] Leacock C, Chodorow M. Combining local context with wordnet similarity for word sense identification. In: Fellbaum C, ed. WordNet: A Lexical Reference System and its Application. MIT Press, Cambridge, MA, 1998.

[35] Lee ML, Kuo FC, Whitmore GA, Sklar J. Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cdna hybridizations. Proc. Natl. Acad. Sci. USA, 2000, 97(18): 9834–9.

[36] Lenca P, Meyer P, Vaillant B, Lallich S. On selecting interestingness measures for association rules: User oriented description and multiple criteria decision aid. European Journal of Operational Research, 2008, 184(2): 610–626.

[37] Li J, Liu H, Ng SK, Wong L. Discovery of significant rules for classifying cancer diagnosis data. Bioinformatics, 2003, 19(Suppl 2): 93–102.

[38] Li J, Wong L. Identifying good diagnostic gene groups from gene expression profiles using the concept of emerging patterns. Bioinformatics, 2002, 18(5): 725–734.

[39] Lin D. An information-theoretic definition of similarity. In: Proc. ICML Conference, 1998.

[40] Liu B, Hsu W, Mun LF, Lee HY. Finding interesting patterns using user expectations. Knowledge and Data Engineering, 1999, 11(6): 817–832.

[41] Liu F, Jenssen TK, Trimarchi J, Punzo C, Cepko CL, Ohno-Machado L, Hovig E, Kuo WP. Comparison of hybridization-based and sequencing-based gene expression technologies on bio-

logical replicates. BMC Genomics, June 2007, 8(153).

[42] Lord PW, Stevens R, Brass A, Goble CA. Semantic similarity measures as tools for exploring the gene ontology. In: Proc. PSB Conference, 2003.

[43] Maddalena A. Pattern based management: Data models and architectural aspects. Lecture Notes in Computer Science, 2005, 3268: 54–65.

[44] Madeira SC, Oliveira AL. Biclustering algorithms for biological data analysis: A survey. IEEE/ACM Trans. Comput. Biology Bioinform., 2004, 1(1): 24–45.

[45] Martinez R, Collard M. Extracted knowledge interpretation in mining biological data: A survey. International Journal of Computer Science & Applications, August 2007, IV(2).

[46] Martinez R, Pasquier C, Pasquier N. Genminer: Mining informative association rules from genomic data. In: Proc. IEEE BIBM Conference, 2007. 15–22.

[47] Martinez R, Pasquier N, Pasquier C, Collard M, Lopez-Perez L. Co-expressed gene groups analysis (cgga): An automatic tool for the interpretation of microarray experiments. Journal of Integrative Bioinformatics, August 2006, 3(2): 1–12.

[48] McIntosh T, Chawla S. High confidence rule mining for microarray analysis. IEEE/ACM Trans. Comput. Biology Bioinform., 2007, 4(4): 611–623.

[49] Mootha VK, Lindgren C, Eriksson K, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrrale M, Laurila E, Houstis N, Daly M, Patterson N, Mesirov J, Golub T, Tamayo P, Spiegelman B, Lander E, Hirschhorn J, Altshuler D, Groop L. Pgc-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet, July 2003, 34(3): 267–273.

[50] Pan F, Cong G, Tung AK, Yang J, Zaki MJ. Carpenter: Finding closed patterns in long biological datasets. In: Proc. ACM SIGKDD Conference, 2003. 637–642.

[51] Parsons L, Haque E, Liu H. Subspace clustering for high dimensional data: A review. SIGKDD Explorations, 2004, 6(1): 90–105.

[52] Pasquier C. Biological data integration using semantic web technologies. Biochimie, 2008.

[53] Pei J, Jiang D, Zhang A. Mining cross-graph quasi-cliques in gene expression and protein interaction data. ICDE Conference, 2005. 353–354.

[54] Pfaltz J, Taylor C. Closed set mining of biological data. In: Proc. BioKDD Conference, 2002.

[55] Prelic A, Bleuler S, Zimmermann P, Wille A, Bühlmann P, Gruissem L, Thiele L, Zitzler E. A systematic comparison and evaluation of biclustering methods for gene expression data. Bioinformatics, 2006, 22(9):1122–1129.

[56] Resnik P. Using information content to evaluate semantic similarity in a taxonomy. In: Proc. IJCAI Conference, 1995. 448–453.

[57] Rizzi S, Bertino E, Catania B, Golfarelli M, Halkidi M, Terrovitis M, Vassiliadis P, Vazirgiannis M, Vrachnos E. Towards a logical model for patterns. In: Proc. ER Conference, 2003. 77–90.

[58] Ruttenberg A, Clark T, Bug W, Samwald M, Bodenreider O, Chen H, Doherty D, Forsberg K, Gao Y, Kashyap V, Kinoshita J, Luciano J, Marshall MS, Ogbuji C, Rees J, Stephens S, Wong G, Wu E, Zaccagnini D, Hongsermeier T, Neumann E, Herman I, Cheung KH. Advancing translational research with the semantic web. BMC Bioinformatics, 2007, 8(Suppl 3(S2)).

[59] Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary dna microarray. Science, 1995, 270(5235): 467–70.

[60] Schlicker A, Domingues F, Rahnenfuhrer J, Lengauer T. A new measure for functional similarity of gene products based on gene ontology. BMC Bioinformatics, 2006, 7(1): 302.

[61] Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N. Module networks: Identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet, 2003, 34(2): 166–176.

[62] Segal E, Wang H, Koller D. Discovering molecular pathways from protein interaction and gene expression data. Bioinformatics, 2003, 19(Suppl 1): 264–272.

[63] Silberschatz A, Tuzhilin A. What makes patterns interesting in knowledge discovery systems. IEEE Trans. Knowl. Data Eng., 1996, 8: 970–974.

[64] Tanay A, Sharan R, Shamir R. Discovering statistically significant biclusters in gene expression data. Bioinformatics, 2002, 18(Suppl 1): 136–144.

[65] Thattai M, Van Oudenaarden A. Intrinsic noise in gene regulatory networks. Proc. Natl. Acad.

Sci. USA, 2001, 98(15): 8614–9.

[66] Tuzhilin A, Adomavicius G. Handling very large numbers of association rules in the analysis of microarray data. In: Proc. SIGKDD Conference, 2002. 396–404.

[67] Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. Science, 1995, 270(5235): 484–7.

[68] Werner T. Bioinformatics applications for pathway analysis of microarray data. Curr. Opin. Biotechnol., 2008, 19(1): 50–4.

[69] Yoon HS, Lee SH, Kim JH. Application of emerging patterns for multi-source bio-data classification and analysis. Lecture Notes in Computer Science, 2005, 3610.

[70] Zhong J, Zhu H, Li J, Yu Y. Conceptual graph matching for semantic search. In: Proc. ICCS Conference, 2002. 92–196.

[71] Ziegler P, Dittrich KR. Three decades of data integration - all problems solved? In: Proc. IFIP Congess Topical Sessions, 2004. 3–12.