

LABORATOIRE



INFORMATIQUE, SIGNAUX ET SYSTÈMES  
DE SOPHIA ANTIPOLIS  
UMR 6070

# KTA: A FRAMEWORK FOR INTEGRATING EXPERT KNOWLEDGE AND EXPERIMENT MEMORY IN TRANSCRIPTOME ANALYSIS

*Laurent Brisson, Martine Collard, Kevin Le Brigant, Pascal Barbry*

*Projet EXECO*

Rapport de recherche  
ISRN I3S/RR-2004-27-FR

Octobre 2004

---

RÉSUMÉ :

MOTS CLÉS :

données génomiques, extraction de connaissances, ontologies

---

ABSTRACT:

This paper addresses the problem of the integration of expert knowledge in a data mining process. We present the KTA (integrating expert Knowledge in Transcriptome Analysis) framework which allows the mining process to be driven by prior knowledge on the application domain. KTA is embedded on the MEDIANTE project for evaluating and using DNA microarrays, the CORESE semantic search engine and the ANNOT module which annotates scientific publications.

KEY WORDS :

genomics data, data mining, ontologies

# KTA : A Framework for Integrating Expert Knowledge and Experiment Memory in Transcriptome Analysis

Laurent Brisson<sup>1</sup>, Martine Collard<sup>1</sup>, Kevin Le Brigand<sup>2</sup>, Pascal Barbry<sup>2</sup>

<sup>1</sup> Laboratoire I3S (CNRS UMR-6070), Université de Nice – Sophia-Antipolis,  
2000 route des lucioles, Les Algorithmes, 06903 Sophia-Antipolis, France;  
{brisson,mcollard}@i3s.unice.fr

<sup>2</sup> Laboratoire de Physiologie Génomique des Eucaryotes (CNRS UMR-6097), IPMC,  
Université de Nice – Sophia-Antipolis,  
660, Route des Lucioles, 06903 Sophia-Antipolis, France;  
{lebridand,barbry}@ipmc.cnrs.fr

**Abstract.** This paper addresses the problem of the integration of expert knowledge in a data mining process. We present the KTA (integrating expert Knowledge in Transcriptome Analysis) framework which allows the mining process to be driven by prior knowledge on the application domain. KTA is embedded in the MEDIANTE project for evaluating and using DNA microarrays, the CORESE semantic search engine and the ANNOT module which annotates scientific publications.

## 1 Introduction

The KTA framework allows to run data mining operations on transcriptome data and to integrate prior knowledge on the domain in the mining process in order to drive it.

Data Mining may be defined as the discovery of unexpected relationships by analyzing such large volumes of data that automated processes are necessary. The extracted knowledge is expressed as a model or a pattern like sets of rules, neural networks or clusters. For instance, in a data mining process, it is quite frequent to search for rule based models since they are easily understandable by end users and have been found to be useful concepts for learning interpretable knowledge from data. For a rule-like patterns  $A \rightarrow B$  where  $A$  and  $B$  are conjunctions of attribute-value terms, one essential issue is to measure the interestingness of the dependency link between the premises  $A$  and the consequent  $B$ . Standard algorithms currently use basic statistical measures for rule selection, but more specific indices were defined for addressing different facets of rule goodness. Furthermore, one can distinguish objective and subjective approaches. Subjective criteria generally are based on a comparison of learned rules against an *a priori* knowledge on data.

**Model Quality.** Numerous algorithms have been proposed for rule induction from data in the machine learning literature. Tree induction or separate-and-conquer rule learning for instance are used for prediction in data mining too. They provide classification rules which right-hand side attributes are predefined and represent the class. On the other hand, so called association rules are among the most popular representation for local patterns in data mining. They may be seen as probabilistic statements about the co-occurrence of events which satisfy statistical constraints on the database like minimum *Support* and minimum *Confidence*. Simple criteria for rule selection like *Accuracy* for entire model or *Support* and *Confidence* for individual rules are known to be insufficient

for extracting useful and interesting information.

Objective measures of interestingness take their origin from the proposition of G. Piatetsky-Shapiro [PIA91] who observed the weakness of the *Confidence* factor and defined the *rule interest*. Numerous other measures have been proposed for evaluating the quality of the extracted information: the *Lift* factor [IBM96], the *JMeasure* of Goodman and Smyth [GOO91], the measure defined by Sebag and Schoenauer [SEB88], the *Conviction* [BRI97]. It has been observed that some of them are quite identical since they rank rules in the same manner. But they may provide a complementary approach since their definition was motivated in order to fill the gap. However, the goodness cannot be specified in an absolute way, it depends on specific goals of the search process.

**Prior knowledge.** From a user point of view, we have to find patterns from databases that are useful or interesting. It's quite a difficult task since people are interested in different things. Furthermore given a set of patterns different users may be interested in different subsets of patterns and the interest of a same user may also vary over time. That's why subjective measures of interestingness may be more relevant than statistical criteria. They measure the subjective interestingness of a pattern to a user and are defined as follows [LIU99]:

- Unexpectedness: Patterns are interesting if they are unexpected or previously unknown to the user
- Actionability: Patterns are interesting if the user can do something with them to his advantage

In the KTA approach, interesting models are selected according the expert knowledge stored in ontologies and scientific publications. The paper is organized as follows: section 2 gives an overview of genomics and DNA microarrays and section 3 presents our approach KTA.

## 2 Mining genomics data

**Genomic data and DNA Microarrays** In recent years there has been an explosion in the acquisition of biomedical data. Advances in molecular genetics technologies such as DNA arrays allow us to obtain one (or several) global description of a living cell. The microarray technology makes it now possible to rapidly measure, through the process of hybridization, the levels of virtually all the genes expressed in a biological sample. Microarrays allow to create data sets of molecular information to represent many systems of biological interest. The gene expression patterns in microarray data have already provided some valuable insights in a variety of problems, and it is expected that knowledge gleaned from microarray data will contribute significantly to advances in fundamental questions in biology as well as in clinical medicine. Gene expression profiles can be used as inputs to large-scale data analysis for identifying previously unknown relationships between genes or understanding the way genes are involved in specific physiological or pathological events.

Machine learning and statistical techniques applied to gene expression data may be used for identifying clusters of genes which have similar behaviours, predicting treatment outcome or drug response.

A joint effort between the "Réseau National des Genopoles" (<http://rng.cnrg.fr>) and the HGMP (MRC, Hinxton, UK) has led to the creation of an experimental resource

allowing the production of pangenomic human and mouse microarrays. The objective was to launch a procedure open to all academic laboratories, in order to select and validate oligonucleotide collections, allowing potential users (especially specialists in precise fields) to participate in improving the panels of selected oligonucleotide probes. A second objective was to standardize experimental methods. This not only requires the involvement of French and English platforms, but also calls for international cooperation. This process should help to form the basis for a European transcriptome standard. 143 274 oligonucleotides 50-52 mers long have been calculated against 28 074 distinct human transcripts, and 118 307 oligos were calculated against 25 173 murine transcripts. A preselection procedure was used to determine the "best" long oligo available in the collection for each transcript. In a few cases, 2 or 3 oligos were finally selected for a single transcript (especially for some splice variants). The selection procedure for the "best" oligo integrated several data, including the specificity of each oligo for transcripts contained in the ENSEMBL (<http://www.ensembl.org>), NCBI, RefSeq databases (<http://www.ncbi.nlm.nih.gov/RefSeq/>), the number of ESTs specifically recognized by each oligo, ... The chosen oligo was the one with the best possible score.

**Why integrating expert knowledge ?** In the context of genomics data, expert knowledge resides into ontologies on genes and diseases, scientific publications and experiment results. Most current methods for transcriptome data analysis, result interpretation and explanation are not automatic. They do not take advantage of volume, heterogeneity and complexity of knowledge stored in all information sources.

### 3 The KTA Approach

KTA is embedded in the MEDIANTE (<http://www.microarray.fr>) project for evaluating and using DNA microarrays, the CORESE [COR02] semantic search engine and the ANNOT module [KHE04] which annotates scientific publications.

#### 3.1 Framework

The interface set up for the MEDIANTE project allows remote users to evaluate the appropriateness of all available probes for their genes of interest. The user can then examine whether the desired microarray pattern is already present on the pan-genomic microarray, or in a microarray being developed by one of the Network platforms. In either case, *ad hoc* transfer tools allow the online user to download all the information needed to independently pursue a transcriptome project (oligo sequences, annotations, etc.). A search tool is provided in order to create a panel of sequences of interest for a personal project managed within the application, based on references (Ensembl, RefSeq, LocusLink, Unigene, ...), keywords (sequence descriptors), exact Gene Ontology (GO) terms, or the chromosomal position. Blast comparison of sequences of interest with all the oligonucleotides in the MEDIANTE database, or all selected transcripts, or all ENSEMBL transcripts, or all NCBI REFSEQ transcripts also allows new sequences to be compared with sequences already stored in MEDIANTE. The transcripts and relevant oligos can be visualized on-screen. A new oligo can then be added to the shopping basket, and a microarray specific to a particular task can be created. If no sequence is recognized, the requested sequence appears as an unknown sequence in the transcript database. A new function will allow the oligos for these "orphan" sequences to be recalculated (preliminary tests have shown that 95% of the analyzed sequences were probed by at least

one oligo of the Mediante database). At the end of the session, the "virtual microarray" can be saved to disk, or a summary e-mail can be requested, with all the data necessary to manage the associated information in-house (oligo sequence, references of recognized sequences, ...). Further bio-informatics development will focus on the archiving of the data, and subsequent datamining. Platforms in Evry, Nice/Sophia Antipolis, Strasbourg (FR), and Hinxton (UK) are in charge of producing pangenomic arrays for British and French public-sector laboratories.

### 3.2 Conceptual Resource Search Engine and Annotations

KTA aims at extending the current version of MEDIANTE with data mining tools in order to analyze experiments, identify experiment profiles and compare a new experiment to previously identified profiles. Once an experiment is achieved, the biologist proceeds a first step preprocessing data, correcting bad data points and then runs statistical studies on a curated set of data. Statistical results about altered genes, or group of genes need then to be explained. Thus the biologist generally looks for related knowledge on modulated genes from diverse information sources like ontologies or scientific publications. The interpretation of results leads to identify previously unknown relationships or to confirm established information. It implies looking for information about genes in the wide volume of publications and ontologies. One of the first objectives of the KTA approach is to provide assistance to the biologist and guide him for selecting genes or choosing a mining algorithm. In the KTA approach, the Miner tool is supported by the search engine CORESE and the ANNOT module. CORESE stands for Conceptual Resource Search Engine. It is an RDF engine based on Conceptual Graphs (CG). It enables the processing of RDF Scheme and RDF statements within the CG formalism. The ANNOT Module produces annotations on scientific publications about the transcriptome provided by biologists. Indeed ANNOT provides semantic ontology-based annotations for each document. These annotations give information on genes described by the publication and relationships among these genes, biological functions or cellular components.

### 3.3 KTA Miner

The KTA Miner module provides user data mining tools which are able to integrate information from an expert knowledge base. The selection of genes of interest will be done according different sources: their expression intensity, a list of genes submitted by the biologist, a list of genes found into annotations according the experiment description. The search for gene clusters will be executed by using either standard algorithms from existing platforms (BioConductor, GeneSpring, ...) or a specific clustering algorithm which will group genes according their annotations in ANNOT. The experiment base will store both expression data and analysis results on each validated experiment. A tool will allow to query the database and characterize a new experiment.

**Description** The KTA approach is based on the concept of scenario. It is useful to improve accessibility of data mining tools and to allow a better understanding of knowledge extraction mechanisms. Our goal is to obtain, to validate and to interpret experimental results and afterwards to build an experiment repository. Consequently, scenarios we use combine various data mining tools suitable for gene expression data analysis. Currently, we focus on three main tasks : relevant gene selection, model interpretation and experiment knowledge repository enhancement.

**Concept of scenario.** A scenario is a specific sequence of operations. Among them there are statistical measures to analyze data, selection techniques, data mining algorithms (classification, clustering, association search, ...), annotation mechanisms and knowledge integration methods. These scenarios help experts to analyze the data, so they can find new interesting relationships between proteins, gene functions, regulatory networks and metabolic pathways for example. Scenarios we propose, can be driven in two ways:

- Interactively driven: Each step of a scenario can lead to different choices for next steps. Choices depend on results, comparison between several methods, biologist goals and criteria than can evolve according time. Consequently it is important to know how matching a data mining task to a specific goal.
- Driven by experiment description: For repetitive tasks, our approach give a choice of well fitted scenarios. Biologists without data mining knowledge will be able to use a set of tools to analyze their data.

**Gene selection.** Current manual techniques focus on few gene selection methods considering for example gene expression intensity. Data mining can increase number of techniques involved in this process, integrate knowledge from heterogeneous sources (UMLS, GO, UNIGENE, LocusLink, GeneBank, Protein, ...) or publications databases (Medline, Books, OMIM, ...) and compare at a large extent results of these various selection methods. In a first approach we consider four different selection methods.

Gene selection according expression intensity: Once we have applied fitting, smoothing and normalization algorithms to data, we select all of the genes with distinct expression intensity. Lots of methods are available (SAM: Significance Analysis Microarray, EBAM: Empirical Bayes Analysis for Microarray, VSN: Variance Stabilization, ...), either based on thresholds and statistics or taking into account missing values.

Gene selection according experiment analogy: we select genes which were expressed (or not expressed) in a similar experiment. It allows to identify genes which have similar behaviors (ie. are expressed in every experiment at the same level) or opposite ones (ie. are expressed in every experiment at very different levels). It could be useful that biologists define their interestingness notion in order to select the best comparison criterion for the current experiment.

Gene selection according experiment description: we select genes which were related to keywords into ontologies and publications. In this step the ANNOT tool gives to the expert a whole set of annotations to discriminate genes of interest.

Gene selection according expert conjectures: genes of interest and relationships among them are submitted by biologists.

A gene selection step provides guidelines for further mining algorithms. For instance, if we select a group of genes according their intensity and their relationship with a disease, the next mining step could be a clustering operation to describe its internal structure. However if we select few genes which are not expressed during an experiment, the next mining step could be to search for characterizing rules. At this time, we focus on finding all the best algorithms which match specific tasks, and how to integrate them in scenarios.

**Model interpretation.** Clusters and decision trees are able to show similarities and differences among gene biological functions or metabolic pathways. Rules are expressing

relationships between genes taking into account high level concepts such as diseases or protein interactions. Instead of rules, we can also use frequent closed itemsets to reduce the size of extracted knowledge or to point out some interesting equivalence classes. Eventually, in some specific cases temporal patterns are helpful in order to model biological processes.

**Knowledge repository enhancement.** For each experiment, the repository contains a description, data and results. The description provides experiment conditions, biologist objectives and references towards related experiments. Each experiment validated by a biologist enhances the repository which may be queried like ontologies and scientific publications from CORESE interface.

## 4 Conclusion

In this paper, we have presented KTA, a new approach we are currently developing for integrating prior knowledge in a data mining process. This proposition follows two main objectives: providing a semantic description of extracted models and improving their interestingness. KTA is based upon the concept of data mining scenarios which consists in sequences of atomic mining operations. We have applied our ideas in the context of transcriptome analysis. In this context, we have proposed an innovative solution which takes advantage of previous DNA microarray experiments, scientific publications and gene and medical ontologies. Data mining operations for gene selection and model extraction are driven by this knowledge base.

## Acknowledgments

This work was performed thanks to supports from the Réseau National Genopole<sup>®</sup>, the Association Vaincre la Mucoviscidose, the GIP Aventis, the ARC, the CNRS, and the French Ministry of Industry (réseau GenHomme).

## References

- [LIU99] Bing Liu, Wynne Hsu, Lai-Fun Mun, and Hing-Yan Lee. Finding interesting patterns using user expectations. *Knowledge and Data Engineering*, 11(6) :817-832, 1999.
- [TAN03] Tan et al Evaluation of gene expression measurements from commercial microarray platforms *Nucleic Acid Research*, 2003, 31:5676-5684.
- [PIA91] G. Piatetsky-Shapiro Discovery, analysis and presentation of strong rules *Knowledge Discovery in Databases*, 1991, MIT Press, G. Piatetsky-Shapiro and W. J. Frawley, editors
- [IBM96] International Business Machines IBM intelligent Miner, User's guide 1996
- [GOO91] R. M. Goodman and P. Smyth Rule induction using information theory *Knowledge Discovery in Databases* KDD-1991, MIT Press
- [SEB88] M. Sebag and M. Schoenauer Generation of rules with certainty and confidence factors from incomplete and incoherent learning bases, *Proceedings of the European Knowledge Acquisition Workshop*, EKAW-1988
- [BRI97] S. Brin, R. Motwani, J.D. Ullman and S. Tsur Dynamic Itemset Counting and Implication Rules for Market Basket Data *Proceedings ACM SIGMOD International Conference on Management of Data*, SIGMOD-1997
- [KHE04] R. Khelif, R. Dieng Annotations sémantiques pour le domaine des biopuces *IC'04*, 2004
- [COR02] O. Corby, C. Faron-sZucker Corese: A corporate semantic web engine *Workshop on real world RDF and semantic web applications*, Hawaii, 2002