Using Bio-Pathways in Relational Learning

Matěj Holec[1], Filip Železný[1], Jiří Kléma[1], Jiří Svoboda[1] and Jakub Tolar[2]

[1]Czech Technical University, Prague
e-mail: {holecm1, zelezny, klema, svoboj1}@fel.cvut.cz
[2]University of Minnesota, Minneapolis
e-mail: tolar003@umn.edu

Abstract. In the present work we compile expression and pathway data related to a specific biological classification problem, into a relational database in Prolog, providing benchmarking material for ILP experimentation. We also review the principal pitfalls arising in the attempts to connect the two sources of knowledge through the relational formalism.

1 Introduction

In the last ten years, gene expression data measured by high-throughput technologies such as DNA microarrays [2] have become an important challenge for machine learning. The typical approaches include the construction of classifiers for tissue or disease class from the expression data (the paper [3] representing the first significant success story), or using clustering algorithms for discovering previously unknown classes characterized by distinguished expression profiles.

To boost the explanatory power of gene expression based classifiers, relevant biological background knowledge must be integrated into the learning process. Inductive logic programming (ILP) here thus offers itself as a suitable tool for data analysis. The review paper [5] indicated several ways in which relevant genomic background knowledge, such as primary gene structure (the order of DNA bases in the gene), could be exploited. The paper [7] presented an application of ILP for inducing relational descriptions of groups of co-expressed genes. The descriptions were first-order conjunctions addressing genes' joint relationships to functions, processes or cellular locations formalized by the gene ontology [1]. The descriptions also pertained to mutual gene relationships derived from another piece of background knowledge – the database of known gene interactions.

In extensive discussions of the results of the study [7] with biologists, we perpetually received the feedback that Gene Ontology terms, acting as the primary vocabulary for expressing classifiers, are overly general to allow for any serious biological interpretation, let alone experimental validation. The similar held for rules conditioned on gene-interaction assumptions. Offering instead to use a language addressing the properties of gene primary structure, we seemed to have jumped to the opposite extreme, in that this level of information would be too specific to allow interpretation in terms of systems biology.

In this work we try to exploit as background knowledge the descriptions of known biochemical pathways (metabolic, gene regulatory, signalling) which appear

[1] http://www.geneontology.org/

to possess just the right level of generality for a biologist. In their simplest form abstracting from reaction stochiometry and kinetics, pathways may be seen as directed graphs with labeled nodes representing compounds (enzymes, metabolites, etc.) and labeled edges standing for various relationships among them (inhibition, activation, reaction product, etc.). Previous work in ILP [6] exists, where metabolic networks were constructed or completed from empirical data. We are however not aware of previous ILP research exploiting pathway descriptions as background knowledge incorporated for sakes of gene-expression based classification.

In the general area of bioinformatics, pathways are already recognized as units exploitable for prediction tasks in gene expression mining. However, the technical means actually provided for their exploitation in current gene expression analysis systems[2] are surprisingly stone-age. In particular, given a sample of expression values and a pathway, the systems calculate the "pathway expression". This is the average of the expressions of genes which code for enzymes present in the pathway. The researcher then looks for the pathways most over- or under-expressed for a given set of samples, compared to a control sample group.

This approach offends biological intuition in several ways, out of which the most important is that a pathway rarely activates as a whole. Usually, it is proper subgraphs of the pathway graphs whose expression may be specific for a given biological condition. In fact, the very division of the cell processes into separate units called pathways is rather arbitrary and has been guided by human convenience rather than any systematic method of network clustering. The recent paper [4] demonstrated a systematic way to extract pathway subgraphs called *fully coupled fluxes* (FCF). It was shown that FCF's comply better than pathways to the intuitive notion of "working units" as the correlation of gene expression in FCF's is larger than in pathways.

ILP systems have in principle all the necessary means to identify the pathway fragments relevant for a given classification task, without much human intervention. With this paper we thus wish to stimulate ILP researchers to explore ways in which this can be done, i.e. how to best exploit pathway information as background knowledge for gene-expression based classification. Note that this task is more ambitious than ordinary search for over(under)-expressed pathway subgraphs. For example, recursion may prove as a suitable syntactic instrument to express that *paths* of certain properties in pathway graphs are typically activated in a given biological condition.

2 Synthesis of the ILP input data

The biological problem motivating the ILP task is to distinguish between two types of cell tissues produced in bone marrow important in blood forming, so called *stromal* cells and hematopoietic *cells*. The full biological background of this problem is out of the scope of this short paper. In the machine learning perspective, the problem can be formalized as classification or knowledge extraction. The former seeks to classify the tissue samples into two distinct classes/cell types. The latter

[2] the Expression Profiler available from the European Bioinformatics Institute, or the DAVID system provided by the US National Center for Biotechnology Information

aims to identify and interpret emerging molecular patterns, i.e. gene sets whose expression and/or coregulation differentiate between the cell types. In this paper we confine ourselves to classification.

The usual way of classifying gene expression data falls in attribute-value learning as tissue samples can be characterized by an invariable probe/gene set. However, in this experiment we use samples from four different species. In particular, from human, macaque, mouse and rat. The instant reason is that any single organism does not provide representative samples in both of the classes. More importantly, learning and generalizing over genomic properties of different species is of fundamental importance in the study of biological and evolutionary principles [1]. In any case, the tissue expression vectors cannot be directly matched as they are measured by different arrays using diverse probe (and thus gene/attribute) sets.

In this paper we propose alternative "working units" whose expression can be figured out and matched in different species – fully coupled fluxes. The various gene vector spaces are transformed into a uniform FCF space in which the classification is carried out. This approach not only allows to generalize beyond species, but also introduces a vocabulary of terms more robust than the original probes, respectively genes. The following subsections give a detailed description of the original data and their Prolog counterparts as well as construction of the abstract FCF attributes.

2.1 Gene Expression Data

We searched the Gene Expression Omnibus (GEO)[3], a public gene expression data library, for expression samples from various experiments involving different organisms. Irrespectively of the objectives of these different experiments, we selected those which included one of the mentioned tissue types among their measured samples. We were using only the measurements acquired by different Affymetrix DNA microarrays (chips).

We obtained 268 biological samples measured by 8 different arrays. 150 samples represent stromal cells while 118 samples correspond to hematopoietic cells. 163 samples were human, 11 of macaque, 8 of rat and 97 murine. Each GEO sample has a XML annotation that gives basic information about it. Among others, the annotation gives CellType which in our case can be hematopoietic or stromal. Further, several allied samples can be acquired within the same experiment, GDSno carries its identification. SampleID characterizes the sample, TissueState distinguishes normal and treated tissues, Organism determines the species and MArrayID gives the identification of the used chip. The annotations were parsed and stored into the Prolog predicate *array*. The predicate as well as a particular fact are shown below:

array(CellType,GDSno,SampleID,TissueState,Organism,MArrayID,Comment).

```
array(hematopoietic,'GDS2718','GSM169465','normal','Mus musculus',
'GPL1261','GCOS 1.4 software (Affymetrix)').
```

Measurements were available as plain text files where rows contain microarray probe identifiers and corresponding expression values. They can easily be transformed into a list of Prolog facts, where SampleID characterizes the sample, AffyID

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

identifies the probe and ExpressionValue gives the expression measured on the given probe in the given sample:

e(SampleID, AffyID, ExpressionValue).

```
e('GSM101111', 'AFFX-BioB-M_at', 1016.3).
```

2.2 Pathway Data

Kyoto Encyclopedia of Genes and Genomes (KEGG)[4] is a collection of manually drawn pathway maps representing common knowledge on the molecular interaction and reaction networks. KEGG stores pathways as XML files with a strictly defined structure. We transformed the four species specific KEGG XML files to Prolog facts, the transformations preserved as many graph features as possible (in fact, only visual representation and position of the elements were neglected). The predicate *entry* represents vertices, the predicate *relation* corresponds to edges. The argument Organism gives the species ('hsa' stands for homo sapiens). PathwayID identifies the reference pathway – a unique pathway of the given function shared by various organisms, KeggNodeID determines its vertex. ListOfEntrezIDs provides a list of genes that map on the given vertex within the specific organism. The genes are given by identifiers that are used by Entrez[5] – the integrated, search and retrieval system developed and maintained by National Center for Biotechnology Information (NCBI)[5]. NodeType specifies the type of vertex (e.g. gene product, group of gene products, compound or map). Interaction or relation is basically an oriented edge among nodes given by BeginNodeID and EndNodeID. A more detailed description can be found in KEGG Markup Language[4].

entry(Organism, PathwayID, KeggNodeID, ListOfEntrezIDs, NodeType).

```
entry('hsa', 04520,1, [hsa:4089], 'gene').
```

relation(Organism, PathwayID, BeginNodeID, EndNodeID, TypeOfRelation, SubTypeName, SubTypeValue).

```
relation('hsa', 04520, 14, 16, 'pprel', ['activation','phosphorylation'],
['-->','+p']).
```

2.3 Data Processing – Fully Coupled Fluxes

The above-mentioned representation enables to merge the species dedicated pathway data along the enzymes exhibiting the same behavior. In other words, the orthologous genes involved in the same vertex and having a similar function in the pathway can be mapped across all of the species under consideration. However, to improve robustness we use linear pathway subgraphs instead of vertices – FCF's. FCF is the longest possible chain of vertices in which non-zero vertex activation implies a certain (non-zero) activation in its successors. FCF's make the abstract attributes which substitute the original probes/genes.

[4] http://www.genome.jp/kegg/, http://www.genome.jp/kegg/docs/xml/
[5] http://www.ncbi.nlm.nih.gov/, http://www.ncbi.nlm.nih.gov/sites/gquery

The transformation proceeds in the following steps. First, the probes are assigned EntrezIDs, i.e. the probes are matched with genes. Prevailingly, there is multiple probes mapped to single EntrezID. The conversion predicate *affy2entrez* extracted from corresponding BioConductor libraries[6] maps AffyIDs and EntrezIDs introduced earlier:

affy2entrez(MArrayID,AffyID,EntrezID).
`affy2entrez('GPL1261','1452692_a_at',72900).`

Second, the activity of enzymes can be inferred from the predicates *affy2entrez*, *e* and *entry*. The activity is considered in terms of expression of the underlying genes, respectively probes. Obviously, this step involves one of the major difficulties. There is many-to-many relationship between array probes and pathway vertices, respectively enzymes. Moreover, the relationship varies across chips. It is also advisable to consider dependencies among contextual enzymes – the same enzyme can be produced by different genes in different contexts. That is why, this enumeration is delayed until FCF's are constructed. Third, the FCF's are found on basis of *relation* – the predicate *flux* is introduced:

flux(FCFID,PathwayID,ListOfKeggNodeIDs).
`flux(flux9,04520,[6,7,8,9,10,11,12,13,14,15,16,17,18,19]).`

Last, the activity of FCF's can be enumerated in every sample and used to classify them. This step is decomposed into two substeps. Initially, flux interpretations have to be found. These interpretations aim to find the most plausible definition of FCF's in all chips. For each flux, enzyme and chip, the corresponding interpretation picks the "optimal" probe. Correlation of expression values underlies the process of selection – the selected probes show the highest mean correlation in the given FCF within all samples measured by the given chip. The predicate *fi* gives the Prolog definition of flux interpretation in the given chip:

fi(MArrayID,FCFID,ListOfAffyIDs,CorrelationOfAffyIDs)
`fi('GPL1261',flux0,['1451002_at','1450048_a_at'],0.86).`

Having the flux interpretations, it is straightforward to calculate FCF activity in every sample. It is given by mean expression of the probes taken from the appropriate interpretation. The overview of entities, predicates and their relations is given in Figure 1. For the sake of lucidity, we do not explicitly mention other necessary and implemented features – missing value treatment or normalization of raw expression data.

3 Conclusions

The proposed ILP task introduces a new way to mine heterogeneous genomic data. It allows to generalize beyond genes as well as species. There are manifold direct implications to gene expression mining. Considering biological viewpoint, no targeted assay has yet been conducted to measure the expression profiles of the two types of tissues in a single experimental setup. The general methodology provides means to increase robustness and explanatory power of molecular classifiers. We

[6] http://bioconductor.org/packages/1.9/data/annotation/

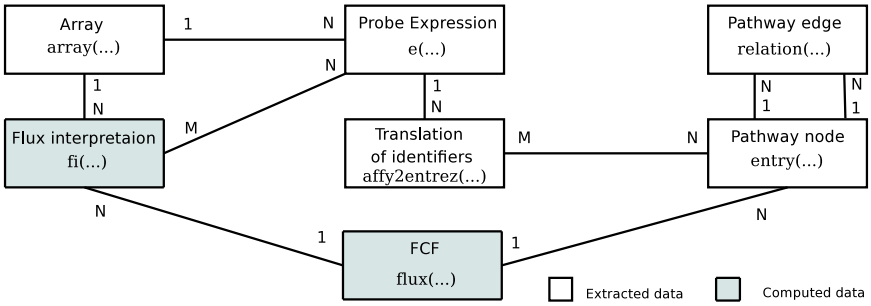

Fig. 1. Overview of entities, predicates and their relations.

have already obtained basic mining results enabled by the developed representation, although not yet with an ILP system. There were found FCF's with plausible biological interpretation that exhibit a statistically significant fold-change (ratio between the mean activity in both classes). Principal component analysis done in FCF's feature space confirmed that fluxes capture the difference between stromal and hematopoietic cells. The Prolog knowledge base is available on request.

Acknowledgements

This work was supported by the grant 1ET101210513 "Relational Machine Learning for Analysis of Biomedical Data" funded by the Czech Academy of Sciences and by the Czech Technical University grant CTU0814413. The Czech-USA travels were covered by Czech Ministry of Education through the project ME910.

References

1. S. Bergmann, J. Ihmels, and N. Barkai. Similarities and Differences in Genome-Wide Expression Data of Six Organisms. *PLoS Comput Biol*, 2(1):e9, 2003.
2. W. Dubitzky, M. Granzow, and Berrar D.P. *Fundamentals of Data Mining in Genomics and Proteomics*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.
3. T. Golub, D. Slonim, P. Tamayo, and et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, October 1999.
4. R.A. Notebaart, B. Teusink, R.J. Siezen, and B. Papp. Co-Regulation of Metabolic Genes Is Better Explained by Flux Than Network Distance. *PLoS Comput Biol*, 4(1):e26, 2008.
5. D. Page and M. Craven. Biological applications of multi-relational data mining. *SIGKDD Explor. Newsl.*, 5(1):69–79, 2003.
6. A. Tamaddoni-Nezhad, R. Chaleil, A. Kakas, and S. Muggleton. Abduction and induction for learning models of inhibition in metabolic networks. In *ICMLA '05: Proc of the Fourth Int Conf on Machine Learning and Applications*, pages 233–239, Washington, DC, USA, 2005.
7. I. Trajkovski, F. Železný, N. Lavrač, and J. Tolar. Learning relational descriptions of differentially expressed gene groups. *IEEE Trans. Sys Man Cyb C*, 38(1):16–25, 2008.