Using ILP to Identify Pathway Activation Patterns in Systems Biology

Samuel R Neaves, Sophia Tsoka

Department of Informatics King's College London,
Strand, London, UK
{samuel.neaves,sophia.tsoka}@kcl.ac.uk
http://www.kcl.ac.uk

Abstract. We show a logical aggregation method that, combined with propositionalization methods, can construct novel structured biological features from gene expression data. We do this to gain understanding of pathway mechanisms, for instance those associated with a particular disease.

Keywords: ILP, Biological pathways, Warmr, Treeliker, Reactome, Barcode, Logical aggregation, RDF.

1 Introduction and Background

In the field of Systems Biology researchers are often interested in identifying perturbations within a biological system that are different across experimental conditions. In this paper we use the example of identifying differences in perturbations between two types of Lung Cancer.

A typical pipeline for this kind of task has three distinct stages. The first stage is to use a technology such as a microarray or RNAseq to measure gene expression across the genome in a number of samples from each of the experimental conditions. The second stage is to identify a subset of genes whose expression values differ across conditions. This stage is commonly achieved by performing differential expression analysis and ranking genes by a statistic such as fold change values. A statistical test is then used to identify the relevant set to take forward to stage 3. Alternatively for stage 2 researchers may train a model using machine learning to classify samples into experimental conditions, often using an attribute value representation where the features are a vector of gene expression values. This approach has the advantage that the constructed model may have found dependencies between genes which would not have been identified otherwise. Researchers will use the 'top' features from the model to identify the set of genes to take on to stage 3.

In stage 3 researchers look for connections between these genes, for example by performing Gene Set Enrichment Analysis (GSEA) [1]. Here the set of genes identified in stage 2 are compared with predefined sets of genes. Each predefined set of genes indicate a known relation. For example having a related function, existing in the same location in the cell or taking part in the same pathway.

To bring background knowledge of relations into the model building process, past ILP research [2] integrated stage 2: finding differentially expressed genes and stage 3: GSEA, into a single step. This was achieved using Relational Subgroup Discovery, which has the advantage of being able to construct novel sets by sharing variables across predicates that define the sets. For example a set could be defined as the genes that have been annotated with two Gene Ontology terms.

Other ways researchers have tried to integrate the use of known relations is by adapting the classification approach. New features are built by aggregating across a predefined set of genes - for example by taking an average expression value for a pathway, see [3] for a review of these methods. A major limitation of current classification approaches is that the models are constructed from either genes or crude aggregates of sets of genes, and so ignore the detailed relations between entities in a pathway. In order to incorporate more complex relations a network representation is needed. It is important that this representation is appropriate such that biological relations are adequately represented. For example a simple directed network of genes and proteins does not adequately represent the complexities of biochemical pathways such as the dependencies of biochemical reactions. To do this bipartite graphs or hypergraphs can be used (see [4] for more details).

One way to incorporate more complex relations is by creating topologically defined sets, for example by performing community detection in a gene regulatory network. However, this approach can create crude clusters of genes, that do not account for important known biological concepts. Biologists are also interested in complex biological interactions rather than just sets of genes, as we now describe.

Network motif and frequent subgroup mining [5] are methods that can look for structured patterns in biological networks. However, in these approaches the patterns are often described in a language which is not as expressive as first order logic. This means they are unable to find patterns with uninstantiated variables, or with relational concepts such as paths or loops. For example, in Figure 1a we show a toy example of four instantiations of the same graph, three in class A and 1 in class B. A frequent pattern distinguishing between the classes may be a chain of three on reactions. This may be represented in Prolog as *on(a),on(X),on(c),link(a,X),link(X,c)*. However, network motif finding can only work with patterns such as $a_{on} \rightarrow b_{on} \rightarrow c_{on}$, and hence would not be able to find the repeated pattern.

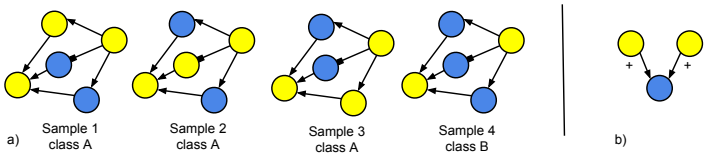

Fig. 1. Example graphs. Node color indicates expression activity. Yellow: on; blue: off.

To our knowledge only one previous work has looked to use ILP for this task [6]. Here the authors propose identifying features consisting of the longest possible chain of vertices in which non-zero vertex activation implies a certain (non-zero) activation in its successors, which they call a Fully Coupled Flux.

The aim of this paper is to identify pathway activation patterns that differ between biological samples of different classes, in order to give a biologist different information than models built from simple gene features.

An example of when such a pattern is useful is in consistency modelling [7]. Here a pattern of inconsistency, an example of which is given in Figure 1b, is matched against gene regulatory networks. Then for each match Answer Set Programming techniques are used to amend the pathways and remove the inconsistency. Similar work has been carried out in ILP where biological pathways have been constructed and amended [8]. The patterns identified with our approach could be used in a similar task to further understand the system level perturbations between classes.

2 Methods

We use structural pathway information to build first order features that are used to construct classification models that discriminate between two lung cancer types. We collate raw data from online sources, process this data to create a Prolog knowledge base, and then learn first order models using this knowledge base.

2.1 Raw Data

We obtained from the GEO a two class Lung Cancer data set containing 37 SCC examples and 33 AC examples; the accession number is GSE2109.The classes correspond to different types of Lung Cancer. For details of this task please see [9]. We use the Reactome database to provide the background knowledge about pathways. Reactome [10] is a collection of manually curated peer reviewed pathways. Reactome is made available as an RDF XML Biopax level 3 file file, which allows for simple passing using SWI-Prolog's semantic web libraries.

2.2 Data Processing

Reactome uses the bipartite network representation of entities and reactions. We extract and process this to create a reaction centric graph, where nodes are reactions and directed edges are labelled either as 'activation' ,'inhibition ' or 'follows' corresponding to how reactions are connected. Boolean networks [11] are a common abstraction in biological research, but these are normally applied at the gene or protein level not at the reaction level. In order to use a boolean network abstraction on a reaction network, we apply a logical aggregation method that aggregrates measured probe values in the microarray into reactions.

The first step of this aggregation is the discretization of the probe values into binary values. We do this using Barcode [12], a tool for converting the continuous probe values to binary variables by applying previously learnt thresholds to microarray data. This makes it possible to compare gene expressions, both within a sample and between samples potentially measured by different arrays.

Once we have binary probe values, we use the structure provided by the Reactome RDF graph and key biological concepts to build reaction level features. Each reaction has a set of inputs that are required for a particular reaction. In addition a reaction may be controlled (activated or inhibited) by particular entities. Entities in Reactome include protein complexes and protein sets, which can themselves comprise of other complexes or sets. We interpret each reaction input as a logical circuit. The relationship between probes and proteins is treated as an OR gate, protein complexes as an AND gate, and protein sets also as an OR gate. A final step takes into account the activation or inhibition of the reaction by any controlling entities. In this way we can say that a reaction 'can proceed' if and only if the input circuit evaluates to true. If a reaction 'can proceed' then we say that it is 'on' otherwise we say that it is 'off'.

2.3 Searching for Pathway Activation Patterns

We experimented with two propositionalization methods, Warmr and Treeliker, separately and then combining them. We search for features independently in each pathway and with a language bias that suitably constrains the search in order to achieve a manageable number of structures.

The first method, Warmr [13], is the first order equivalent of item set and association rule mining. It can be used as a propositionalization method by independently searching for frequent queries in the two classes. An advantage of Warmr is that it is possible to define background predicates for relevant concepts. For example a path or loop of all 'on' reactions. As Warmr does not prune by relevance to classification tasks it can however quickly build to an intractable search with many irrelevant or similar queries/features built.

The second method, TreeLiker [14], is a modern ILP tool that implements a number of algorithms. It has been shown to produce long features by building features bottom up in a blockwise manner. This is desirable for our task as longer features will provide more mechanistic insight to a biologist. A limitation is that the features are 'tree like ' which means there can be no cycles in the variables. Unlike Warmr, TreeLiker does not support explicit background knowledge and therefore all relevant relations need to be preprocessed using Prolog.

Our combined method takes a top feature constructed by TreeLiker and uses this as the basis for the language bias input into Warmr. We then add language bias constraints that guide Warmr to add cycles to the tree like feature. This results in long cyclical features that Warmr would not be able to find on its own.

To evaluate the generated features we use them to build classifiers, to estimate their ability to discriminate between the two lung cancer types. We use the J48 and JRIP tree and rule building algorithms of the Weka package to do this, chosen because these produce interpretable models.

3 Preliminary Results

We give the following example features found by 1) Warmr, 2) TreeLiker and 3) our combined approach:

```
1:array(A),reaction(A,B,1),reaction(A,C,0),link(C,B,D),link(B,C,E).
```

```
2:reaction(A,0), link(A,B,follows), reaction(B,1), link(B,C,_),
reaction(C,0), link(A,D,activation), reaction(D,0).
```

```
3:array(A),reaction(A,B,0),link(B,C,follows),reaction(A,C,1),
link(C,D,E),reaction(A,D,0),link(B,F,activation),reaction(A,F,1),
link(F,D,E),link(D,G,E),reaction(A,G,0)
```

Feature 1 is a simple cyclical feature found by Warmr, the variable A matches one sample. Feature 2 is a longer tree like feature found by TreeLiker. Notice TreeLiker does not require a variable for the sample. Feature 3 is found by our combined method, it is both long and contains a cycle.

To date we have found promising features for the Apoptosis pathway, which, using the Jrip model, achieved mean 81.29% accuracy (std 13%) using 10 fold cross validation. This is a comparable accuracy to that of a model built with raw expression values, but now we have identified pathway activation perturbations rather than just gene expression perturbations.

4 Discussion

The Pathway Activation Patterns we found using this approach are in clinically relevant pathways. These patterns may give diagnostic and clinical insights that biologists can develop into new hypotheses for further investigation.

This work has shown the potential of ILP methods for mining the abundance of highly structured biological data. Using this method we have identified differences in Pathway Activation Patterns that go beyond the standard analysis of differentially expressed genes, enrichment analysis, gene feature ranking and pattern mining for common network motifs. We have also demonstrated the use of logical aggregation with a reaction graph and how this simplifies the search for hypotheses to an extent where searching all pathways is tractable. We have introduced a novel approach that uses Warmr to extend features initially identified with TreeLiker. This makes it possible to search for long cyclical features.

References

[1] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, October 2005.

[2] Dragan Gamberger, Nada Lavra, Filip elezn, and Jakub Tolar. Induction of comprehensible models for gene expression datasets by subgroup discovery methodology. *Journal of Biomedical Informatics*, 37(4):269–284, August 2004.

[3] Matj Holec, Ji Klma, Filip elezn, and Jakub Tolar. Comparative evaluation of set-level techniques in predictive classification of gene expression samples. *BMC Bioinformatics*, 13(Suppl 10):S15, June 2012.

[4] Ken Whelan, Oliver Ray, and Ross D King. Representation, simulation, and hypothesis generation in graph and logical models of biological networks. In *Yeast Systems Biology*, pages 465–482. Springer, 2011.

[5] Wooyoung Kim, Min Li, Jianxin Wang, and Yi Pan. Biological network motif detection and evaluation. *BMC Systems Biology*, 5(Suppl 3):S5, December 2011.

[6] Matej Holec, Filip Zeleznỳ, Jirı Kléma, Jirı Svoboda, and Jakub Tolar. Using bio-pathways in relational learning. *Inductive Logic Programming*, page 50, 2008.

[7] Carito Guziolowski. *Analysis of Large-Scale Biological Networks with Constraint-Based Approaches over Static Models*. PhD thesis, Université Rennes 1, 2010.

[8] Oliver Ray, Ken Whelan, and Ross King. Automatic revision of metabolic networks through logical analysis of experimental data. In *Inductive Logic Programming*, pages 194–201. Springer, 2010.

[9] Kahn Rhrissorrakrai, J. Jeremy Rice, Stephanie Boue, Marja Talikka, Erhan Bilal, Florian Martin, Pablo Meyer, Raquel Norel, Yang Xiang, Gustavo Stolovitzky, Julia Hoeng, and Manuel C. Peitsch. sbv IMPROVER Diagnostic Signature Challenge: Design and results. *Systems Biomedicine*, 1(4):3–14, September 2013.

[10] D. Croft, A. F. Mundo, R. Haw, M. Milacic, J. Weiser, G. Wu, M. Caudy, P. Garapati, M. Gillespie, M. R. Kamdar, B. Jassal, S. Jupe, L. Matthews, B. May, S. Palatnik, K. Rothfels, V. Shamovsky, H. Song, M. Williams, E. Birney, H. Hermjakob, L. Stein, and P. D'Eustachio. The Reactome pathway knowledgebase. *Nucleic Acids Research*, 42(D1):D472–D477, November 2013.

[11] Rui-Sheng Wang, Assieh Saadatpour, and Rka Albert. Boolean modeling in systems biology: an overview of methodology and applications. *Physical Biology*, 9(5):055001, October 2012.

[12] M. N. McCall, H. A. Jaffee, S. J. Zelisko, N. Sinha, G. Hooiveld, R. A. Irizarry, and M. J. Zilliox. The Gene Expression Barcode 3.0: improved data processing and mining tools. *Nucleic Acids Research*, 42(D1):D938–D943, January 2014.

[13] Luc Dehaspe and Luc De Raedt. Mining association rules in multiple relations. In Nada Lavra and Sao Deroski, editors, *Inductive Logic Programming*, number 1297 in Lecture Notes in Computer Science, pages 125–132. Springer Berlin Heidelberg, January 1997.

[14] Doc Ing Filip Železnỳ. *Fast Construction of Relational Features for Machine Learning*. PhD thesis, Czech Technical University in Prague, 2013.