ILP Identifies Pathway Activation Patterns

Samuel R Neaves, Dr 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 for classification and ranking tasks in Systems Biology.

Keywords: Biological Pathways, Reactome, RDF, Barcode, Logical Aggregation, Warmr, Treeliker, ILP.

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 these perturbations in two types of Lung Cancer (for details see [1]). To clarify we are looking to identify differences rather than building a diagnostic tool for 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 correlate with the conditions. Stage 2 is commonly achieved by performing differential expression analysis and ranking genes by their fold change values or other statistic. A statistical test is then used as a cut off value 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 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) [2]. Here the set of identified genes are compared with predefined sets of genes. The predefined sets of genes indicate a known relation. For example having a related function, existing in the same location in the cell or taking part in a pathway. A further way a set could be defined is a topological set, this takes into account finer grained internal relations to define a set. For example a topological set could be identified by community detection in a gene regulatory network.

In contrast to having two stages for analysis, analysis can be performed in one step. Past ILP research [3] has integrated the two steps of finding differently expressed genes and GSEA by using relational subgroup discovery. These have used the hierarchical Gene Ontology to relate genes and 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 GO terms. Other ways researchers have tried to use 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 [4] for a review of these methods.

In most of these methods it is common to ignore the detailed relations between entities in a pathway, whereby the pathway is treated as an unstructured collection of genes. Only topologically defined sets take advantage of any known internal relations. On the other hand these often create crude clusters of genes, and do not follow biological intuition of information flow through the pathway.In addition the chosen network representation is often inappropriate for pathway analysis as described later. Therefore a major limitation of the current classification approaches is that the models constructed are limited to being constructed from either genes or crude aggregates of sets of genes. However biologists are not only interested in these kind of models. More general pathway activation patterns are also of interest, including paths and loops where some of the entities could possibly be represented by uninstantiated variables. Two examples are i). Consistency modelling and ii). Network motif finding. In consistency modelling [5] a pattern of inconsistency is matched against gene regulatory networks and for each match Answer Set Programming techniques are used to amend the pathways to remove the inconsistency. Similar work has been done in ILP where biological pathways have been constructed and amended [6]. In network motif finding [7], again the biological intuition of information flow is often ignored and it is uncommon to either take into account the class of the network sample or to include uninstantiated variables in the search for motif patterns. Motifs are most commonly searched for across a network constructed from an agglomerative of the experimental data rather than by mapping expression values to a standard graph of the pathway which would in effect give an instantiated graph for each sample. This approach would allow for the searching of common pathway activation patterns across samples - taking into account the class of each sample.

In contrast to the work on consistency modelling, in this work we assume the pathways are in some sense correct and we are looking at individuals sample expression patterns mapped on to the pathway. This work is most similar to [8] in which the authors propose using Fully Coupled Fluxes as features. An FCF is the longest possible chain of vertices in which non-zero vertex activation implies a certain (non-zero) activation in its successors.

2 Method

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. We use structural pathway information to build first order features that are used to construct classification and ranking models, which provide an appropriate level of abstraction for the interpretation of the model to be insightful for a biologist.

2.1 Raw Data

The GEO data set we use for this study is GSE2109. This is a two class data set with the classes corresponding to different types of lung cancer(SCC and AC). There are 38 SCC samples and 33 AC samples.

Reactome [9] is a collection of manually curated peer reviewed pathways. Reactome is made available as an RDF XML Biopax level 3 file. This allows for simple passing using SWI-Prolog's semantic web libraries. In this work we use a high level logical abstraction of the idea of a biological pathway. In biological research it is common to see many types of biological networks represented by mathematical graphs. When considering biological pathways simple directed networks of genes and proteins are not appropriate, because this formalism does not adequately model the dependencies of activation patterns i.e. bipartite graphs or hypergraphs are required. See [10] for more details on this. We address this problem by using a directed reaction centric graph which we have extracted from Reactome.

2.2 Data Processing

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 from the measured probes in the microarray to reactions as defined in Reactome.

We first discretize the probe values into binary values using Barcode [12] which is a tool for applying learnt thresholds to microarray data, Barcode 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 graphs and our biological understanding to build Reaction level features. Reactome defines reactions with entities divided into substrates, enzymes and products. Each reaction has a set of inputs and set of outputs, in addition a reaction may be controlled (activated or inhibited) by an entity. Entities in Reactome include protein complexes and protein sets, which in turn can themselves comprise of other complexes or sets. For the logical aggregation step we interpret this structure as a simple 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. With a final step taking 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

In this work we have experimented with two propositionalization methods, 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 propositionalization method by independently searching for frequent queries in the two classes. An advantage of using Warmr is that it is easy 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 a desirable trait for this task as longer features will give more information to a biologist. A limitation is that the features are 'tree like ' which means there can be no cycles in the variables. TreeLiker unlike Warmr does not support explicit background knowledge and therefore all relevant relations need to be preprocessed using prolog.

For a combined method we first take a top feature constructed by TreeLiker and use this as the basis for the language bias input into Warmr. We then add further language bias constraints that guide Warmr into searching for frequent queries that create cycles in the tree like features. This results in longer features that contain cycles than Warmr can find on its own.

3 Results

As we are interested in producing comprehensible models, we limit our experiments to rule and tree building algorithms.We use Wekas implementation of classifications trees (J48) and rule learning (JRip). The best classifiers achieve 81.29% accuracy (std 13%) in 10 fold cross validation.

An Example Warmr found feature is:

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

This is a simple cyclical feature, the variable A matches one sample.

An Example Tree Liker found feature is:

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

This is a longer tree like feature -notice TreeLiker does not require a variable for the sample.

An Example combined method found feature is:

```
microarray(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)
```

This final feature is both long and contains a cycle. With the structured features found, we were able to achieve 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

This work has shown the appropriateness of using ILP methods to mine 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 to a reaction graph and how this simplifies the search for hypothesises to an extent where searching all pathways is tractable. The extension of TreeLiker found features with Warmr to connect cycles is believed to be novel. This enables the finding of long cyclical features which are of interest to biologists.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

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

[6] 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.

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

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

[9] 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.

[10] 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.

[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.