An Application of ILP to Systems Biology

Samuel R Neaves

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

Abstract. We show how a logical aggregation method 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

This paper describes a method to use Inductive Logic Programming(ILP) techniques to mine structured biological data. The biological problem addressed in this paper is the task of understanding the system perturbations in different types of Lung Cancer; for details about the motivation of this task please see [**?**].

A basic pipe line for this kind of task is to identify differentially expressed genes, and then in subsequent analysis look for how these genes may be related and subject them to further research. Alternatively the task can be formalised as a classification task (of samples) or more generally as a ranking task. Remembering that classification is often not the true task -The classification performance is used as a measure of the quality of the identified differences. Classification might not be the true task as the class of a sample may already be known. (For example where the biopsy is taken from may indicate the class). When classification models are built it is common for this to be done using an attribute value representation where samples are vectors of gene expressions. The researchers are interested in the genes that are most important in the constructed model. Genes that are identified in this manner may be different to a simple list of differently expressed genes as the constructed model may have found dependencies on genes which are not highly differently expressed.

Whether researchers have identified a list of genes through differential expression analysis or by taking top features from a classification model, they have yet to take into account known relations between the genes. Therefore a common next step is for researchers to look for connections between these genes, for example by performing Gene Set Enrichment Analysis(GSEA) [**?**]. These gene sets define a relation between the genes, examples of gene sets include: functionally defined sets, cellular location defined sets, pathway defined sets, or network

defined topological sets (These can be identified by community detection in a gene-network).

Past ILP research 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. [**?**]. Other ways researchers have tried to use known relations include simple attempts to aggregate across the set of genes by taking an average expression or other measure such as the SVD, see [**?**] for a review of these methods. In most of these methods it is common to ignore the detailed relations between entities in a pathway, the pathway is treated as an unstructured collection of genes. Only do topologically defined sets take advantage of any known internal relations. But often these 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 because of the problem discussed in [**?**].

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 biological researchers 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 1. Consistency modelling [**?**] and 2. Network motif finding. [**?**]. In consistency modelling the pattern depicted in figure 1 is matched against gene regulatory networks to check for consistency, following this 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 [**?**].

In network motif finding, 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 (Pathway) giving in a effect a instantiated graph per sample, which allows us to search for common pathway activation patterns across samples taking into account the class of the sample which is the case in this paper.

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 samples expression patterns mapped on to the network. This work is most similar to [**?**] 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 Aim

The aim is to identify pathway activation patterns that differ between biological samples, in order to give a biologist more information than models built from simple gene features. This is achieved by using structural pathway information to build first order features that can be used to construct classification and ranking models, which will provide an appropriate level of abstraction for the interpretation of the model to be insightful for a biologist.

3 Method

3.1 GEO 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. [**?**].

3.2 Biological Pathway Data

Reactome [**?**] 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- either bipartite graphs or hypergraphs are required. We address this problem by using a directed reaction centric graph which we have extracted from Reactome. [**?**].

Boolean networks 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.

3.3 Barcode

The first step is to discretize the probe values into binary values. Barcode is a tool for applying learnt thresholds to microarray data, this gives a binary value for each probe in the microarray. Barcode makes it easy to compare gene expressions, both within a sample and between samples potentially measured by different arrays. This tool is implemented as a R package. [**?**].

3.4 Logical Aggregation

Once we have binary probe values, we can 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, A protein complex or protein set may itself include sub complexes or sub sets. This gives the input to a reaction a kind of tree structure. 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 and 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'.

3.5 Propositionalization

Best practise in ILP is to apply propositionalization methods when possible and so in this work we have experimented with two methods separately and then combining them. [**?**]

Warmr Warmr [**?**] 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.

Tree Liker Tree liker [**?**] 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 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 does not require a variable for the sample as this is done automatically. Treeliker unlike Warmr does not support explicit background knowledge and therefore all relevant relations need to be preprocessed using prolog.

Illustrative Tree Liker Template

```
set(output_type,single)
```

```
set(examples,'Pathway963.txt')
set(template,[reaction(-R1,#onoroff),link(+R1,-R2,!T1),
 reaction(+R2,#onoroff),link(+R2,-R3,!T2),reaction(+R3,#onorff),
 link(!RA,-R4,!T3),link(+R4,!RB,!T4),link(+R1,-R2,#T1),
 link(+R2,-R3,#T2),link(!RA,-R4,#T3),link(+R4,!RB,#T4)])
set(output, Pathway963.arff)
work(yes)
```

Example Tree Liker found feature

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

Combined Search For a combined search we first take the top features constructed by Treeliker and use these as the basis for the language bias input into Warmr. We then add appropriate background predicates that guide Warmr into searching for frequent queries that close the loops in the Treelike features. This results in longer features that contain loops than Warmr can find on its own.

3.6 Ranking and Classification

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 Ripper.

Appropriate Metrics Dream challenge uses the following metrics [**?**]. Flach et al recommend these metrics. [**?**].

4 Results

ROC curve Features Found and models built.

5 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 (by using pre-defined background knowledge predicate in Warmr such as onpath. The use of logical aggregation to a reaction graph simplifies the search for hypothesises to an extent where all pathways can be searched in reasonable time. The extension of Treeliker found features with Warmr to connect loops is believed to be novel. This allows for the finding of loops with additional properties in a reasonable amount of time.

5.1 Figures

For LaTeX users, we recommend using the *graphics* or *graphicx* package and the `\includegraphics` command.

Please check that the lines in line drawings are not interrupted and are of a constant width. Grids and details within the figures must be clearly legible and may not be written one on top of the other. Line drawings should have a resolution of at least 800 dpi (preferably 1200 dpi). The lettering in figures should have a height of 2 mm (10-point type). Figures should be numbered and should have a caption which should always be positioned *under* the figures, in contrast to the caption belonging to a table, which should always appear *above* the table; this is simply achieved as matter of sequence in your source.

Please center the figures or your tabular material by using the `\centering` declaration. Short captions are centered by default between the margins and typeset in 9-point type (Fig. 1 shows an example). The distance between text and figure is preset to be about 8 mm, the distance between figure and caption about 6 mm.

To ensure that the reproduction of your illustrations is of a reasonable quality, we advise against the use of shading. The contrast should be as pronounced as possible.

If screenshots are necessary, please make sure that you are happy with the print quality before you send the files.

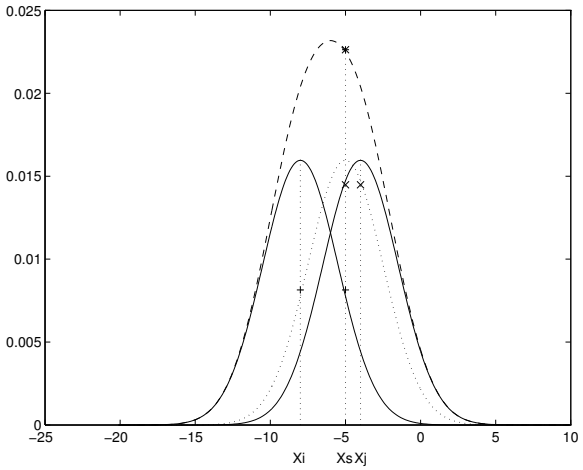

Fig. 1. One kernel at x_s (*dotted kernel*) or two kernels at x_i and x_j (*left and right*) lead to the same summed estimate at x_s. This shows a figure consisting of different types of lines. Elements of the figure described in the caption should be set in italics, in parentheses, as shown in this sample caption.

Please define figures (and tables) as floating objects. Please avoid using optional location parameters like "`[h]`" for "here".

Remark 1. In the printed volumes, illustrations are generally black and white (halftones), and only in exceptional cases, and if the author is prepared to cover the extra cost for color reproduction, are colored pictures accepted. Colored pictures are welcome in the electronic version free of charge. If you send colored figures that are to be printed in black and white, please make sure that they really are legible in black and white. Some colors as well as the contrast of converted colors show up very poorly when printed in black and white.

5.2 Citations

For citations in the text please use square brackets and consecutive numbers: [1], [2], [4] – provided automatically by LaTeX's `\cite ... \bibitem` mechanism.

Acknowledgments. The heading should be treated as a subsubsection heading and should not be assigned a number.

6 The References Section

In order to permit cross referencing within LNCS-Online, and eventually between different publishers and their online databases, LNCS will, from now on, be standardizing the format of the references. This new feature will increase the visibility of publications and facilitate academic research considerably. Please base your references on the examples below. References that don't adhere to this style will be reformatted by Springer. You should therefore check your references thoroughly when you receive the final pdf of your paper. The reference section must be complete. You may not omit references. Instructions as to where to find a fuller version of the references are not permissible.

We only accept references written using the latin alphabet. If the title of the book you are referring to is in Russian or Chinese, then please write (in Russian) or (in Chinese) at the end of the transcript or translation of the title.

The following section shows a sample reference list with entries for journal articles [1], an LNCS chapter [2], a book [3], proceedings without editors [4] and [5], as well as a URL [6]. Please note that proceedings published in LNCS are not cited with their full titles, but with their acronyms!

References

1. Smith, T.F., Waterman, M.S.: Identification of Common Molecular Subsequences. J. Mol. Biol. 147, 195–197 (1981)
2. May, P., Ehrlich, H.C., Steinke, T.: ZIB Structure Prediction Pipeline: Composing a Complex Biological Workflow through Web Services. In: Nagel, W.E., Walter, W.V., Lehner, W. (eds.) Euro-Par 2006. LNCS, vol. 4128, pp. 1148–1158. Springer, Heidelberg (2006)

3. Foster, I., Kesselman, C.: The Grid: Blueprint for a New Computing Infrastructure. Morgan Kaufmann, San Francisco (1999)
4. Czajkowski, K., Fitzgerald, S., Foster, I., Kesselman, C.: Grid Information Services for Distributed Resource Sharing. In: 10th IEEE International Symposium on High Performance Distributed Computing, pp. 181–184. IEEE Press, New York (2001)
5. Foster, I., Kesselman, C., Nick, J., Tuecke, S.: The Physiology of the Grid: an Open Grid Services Architecture for Distributed Systems Integration. Technical report, Global Grid Forum (2002)
6. National Center for Biotechnology Information, `http://www.ncbi.nlm.nih.gov`

7 Checklist of Items to be Sent to Volume Editors

Here is a checklist of everything the volume editor requires from you:

☐ The final LaTeX source files

☐ A final PDF file

☐ A copyright form, signed by one author on behalf of all of the authors of the paper.

☐ A read me giving the name and email address of the corresponding author.