Learning responsive cellular networks by integrating genomics and proteomics data

Elma Akand[1], Michael Bain[1], and Mark Temple[2]

[1] School of Computer Science and Engineering, University of New South Wales, Sydney, Australia 2052
[2] School of Biomedical and Health Sciences, University of Western Sydney, Locked Bag 1797, Penrith South, DC NSW 1797, Australia

Abstract. Systems biology is an important application area for ILP. We investigate learning networks of activity in response to stress in yeast cells. Diverse heterogeneous empirical data is used rather than curated knowledge sources. We find that although this representation is restricted and predictive accuracy is low this can still lead to useful outputs.

1 Introduction

The number of organisms for which the complete genome (DNA sequence) is available continues to grow, and there have been major advances in laboratory techniques to analyse complex cellular processes. This has led to *systems biology*, in which responsive phenotypes, the measurable characteristics of the organism in response to environmental or genetic perturbations, can be investigated *genome-wide*, i.e., by collecting data on the activity of all the organism's genes simultaneously [10].

Thus we can investigate the *cellular network response* of the genes that give rise to an observed phenotype as the downstream effect of an external stimulus through signal transduction. For example, when cells adapt to sudden changes in the environment, cellular network responses include the action of sets of transcription factors (proteins) to activate sets of genes involved in biochemical pathways.

The baker's and brewer's yeast *Saccharomyces cerevisiae* is a key model organism for systems biology, due to the ease with which genetic manipulation can be carried out. Importantly, many fundamental processes in yeast are conserved through to humans [11]. However, although more than a decade has passed since the sequencing of the complete yeast genome, around 25% of yeast genes still do not have a known molecular function [12].

Data sets describing yeast cellular network responses derived from new high-throughput genome-wide experimental techniques are increasingly available [6], and are often inherently relational. Therefore the aim of this work is to apply ILP to uncover significant logical relationships that govern cellular network responses, such as those involved in the onset of oxidative stress-related phenotypic responses that are important in many human diseases, through the integration of genome-wide data sets.

Presented at the 19th International Conference on Inductive Logic Programming (ILP'09).

In this paper we investigate two problems of modelling responsive phenotypes in yeast using ILP: protein expression under oxidative stress, and sensitivity of gene deletion mutants to multiple stresses. The basic setting is described in Section 2, initial results are in Section 3, and discussion in Section 4.

2 Learning the logic of responsive cellular networks

A responsive sub-network of a cell is referred to as the *genetic regulatory network* (GRN) [5]. The protein products of co-expressed genes in a GRN combine to form interacting *molecular machines* that produce a responsive cellular phenotype. This responsive sub-network is partly described by the *protein-protein interaction* (PPI) network. In turn, proteins act to regulate cellular *metabolism* in pathways of biochemical reactions and, by subtle feedback mechanisms, their own GRNs and PPI networks.

We do not expect to be able to learn an entire cellular response network from data on its behaviour. That would be pointless since, in some sense, it is implicit in the empirical data on the GRNs and PPI networks, although this is typically *incomplete* and *incorrect*. Instead, we aim to learn theories on network components that may be predictive *and* explanatory of an observed cellular response. These may be used, for example, in visualization or further learning.

We assume a logical language $\mathscr{L}_{\text{Net}}$ to represent cellular networks, as follows. In this paper we use as constants only gene symbols (genes represent proteins in certain contexts). Function symbols are not currently used. Predicate symbols express properties or relations, such as gene expression or protein interactions. This is similar to representations used in previous work (e.g., [2, 14, 7]).

We assume a supervised learning framework, but adopt a simpler setting than typical in ILP. The task of learning a *logical network* will be to discover a theory T defined in $\mathscr{L}_{\text{Net}}$ which is *over-represented* with respect to a data set E and background knowledge B. We assume there is a function $f_{E,B}(T)$ to evaluate candidate theories, and that some threshold can be set by the biologist on this function to decide if the network may be of interest, i.e., is over-represented. Note that in the work reported here, theories are constructed by learning individual clauses separately, and the evaluation is therefore applied per clause.

As an over-representation measure $f_{E,B}(C)$ to evaluate a clause C we used the cumulative probability from the hypergeometric test [3] to obtain P-values:

$$P(r, s, m, n) = 1 - \sum_{i=0}^{r-1} \frac{\binom{m}{i}\binom{n-m}{s-i}}{\binom{n}{s}} \tag{1}$$

where n is the total number of genes, m is the number of genes in the positive class, s is the number of genes covered by the clause and r is the number of genes in the positive class covered by the clause. This does not apply correction for multiple testing, which would reduce the P-values obtained [3]. However, biologists are familiar with its usage, and apply conservative thresholds. It also has the property of favouring higher-coverage clauses; given two clauses C_1, C_2 with equal accuracy, if C_1 covers more examples, it will have a lower P-value.

3 Experimental results

To develop and test our approach we selected two problems of learning elements of the cellular response network in yeast. The first was to learn a protein expression network from integrated data sets. The second was a more typical laboratory problem, where multiple stresses were applied to yeast cells which were then screened for changes in growth.

3.1 Learning to predict an intra-cellular response phenotype

In this experiment proteomics and genomics data were integrated to learn to predict protein expression in yeast in response to the environmental addition of hydrogen peroxide (H2O2), a condition known to produce "oxidative stress". The response was taken from a proteomics experiment by Godon et al. [8]. There were 56 proteins whose synthesis was stimulated and 36 that were repressed under oxidative stress. As background knowledge, we took two independently generated data sets.

Microarray data from the study by Causton et al. [4] was discretized from time-courses into the values "up" or "down" for a number of conditions. Transcription factor binding (ChIP-chip) data from the study by Harbison et al. [9] provided a set of potential links between genomics and proteomics. Lastly, a set of protein-protein interactions was downloaded from the BioGRID repository [3]. We used Aleph to learn clauses to predict whether genes in the Godon et al. data had their protein expression induced or repressed. Since the majority of genes in the positive class were not generalised the 10-fold cross-validation mean accuracy (sample std. dev.) was 62.5($\pm$17.7)%.

The set of ground instantiations of the learned theory was translated into 'canned' natural language and one of us (M.T.) manually assembled the network diagram of Figure 1. In Figure 1 the horizontal bold line represents the intragenic (promoter) region of the gene named on the left-hand side of the diagram. The filled circle on each promoter links vertically (dotted line) to the transcribed mRNA indicated by the right-pointing arrowhead — a bold vertical line indicates that the transcript is up-regulated in the microarray data [4].

The circle around each arrowhead indicates the response to hydrogen peroxide — a grey filled circle indicates that the protein is induced [8]. Proteins connected by a curved arrow indicate that a protein-protein interaction is in BioGRID. The downward pointing triangle represents a transcription factor (protein) bound to the gene promoter (DNA sequence). The identity of this factor is indicated by the bold vertical line attached to the labelled circle below. The condition under which the transcription factor is bound is indicated by the box to the left of the labelled factor [9].

The clause `(induced(A) :- h2o2(B,A), ppi(A,C), acid(D,C))` denotes that a gene `A` has its protein induced under H2O2 addition since two transcription factors, `B` and `D`, bind the promoter regions of the genes, `A` and `C`, and there is a protein-protein interaction between `A` and `C`.

[3] http://www.biogrid.org

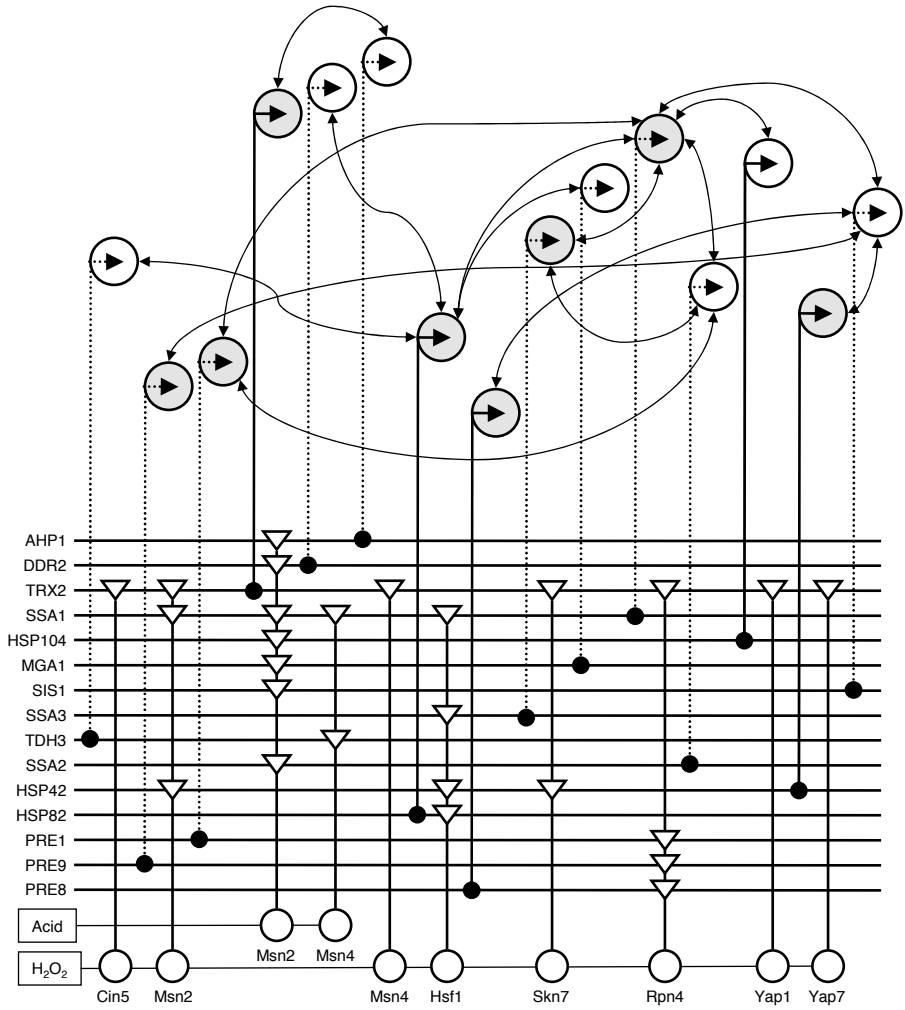

Fig. 1. A network that describes interactions between diverse heterogeneous data leading to protein induction or repression in response to H2O2 treatment.

By way of example, in one of the instantiations of this clause, the gene Trx2 is bound by the transcription factor Msn2/Msn4 (and others as indicated in Figure 1) under hydrogen peroxide treatment conditions, the Trx2 mRNA is up regulated under similar treatment conditions and the protein itself is induced upon peroxide treatment. In addition the Trx2 protein exhibits a protein-protein interaction to Ahp1 (as shown in Figure 1). In turn, further learned clauses describe the AHP1 gene's relationships to others the network.

3.2 Learning to predict an extra-cellular multiple response phenotype

The *Saccharomyces* Genome Deletion Project [15] is a set of yeast strains in each of which exactly one gene from the genome has been systematically removed. Biologists have carried out many "screens" of the deletant set — selecting a subset of genes and subjecting each of the corresponding deletants to that stress searching for a "sensitive" phenotype (e.g., abnormal growth) that would suggest a role for the deleted gene in the cellular response to that stress. This is known to present a hard problem in functional genomics, since there is very little correlation of these screens to microarray data [13].

We assembled 26 screens on 1016 genes from various different laboratories. Of these, 409 deletants were sensitive to three or more screens. These may have a general role in cellular stress response, whereas the remaining 607 are impicated in specific responses. The classification problem was then to learn to discriminate these "general" response genes from those sensitive to only one or two screens (since many stresses had two screens, this is roughly equivalent to being sensitive to one stress). The background data was essentially the same as used in Section 3.1. Although the screen data is known to be noisy, we have so far found a number of clauses at a P-value below the standard cutoff of 0.01. Work is continuing to evaluate their biological plausibility.

4 Discussion

Badea [2] was the first to learn theories of gene expression using ILP. Fröhler and Kramer [7] included genomic and proteomic data in a similar task. These approaches are learning to predict an *intra-cellular* "phenotype". Trajkovski et al. [14] applied ILP and propositionalization to learn an *extra-cellular* phenotype (cancer type), also from integrated data sets. These approaches, as ours, depend on intra-cellular measurement data. However, in our case, the extra-cellular phenotype is known to be more difficult to predict from such data [13].

We have shown elsewhere [1] that it is possible to obtain reasonable accuracy on the tasks in this paper with propositional learning, given non-discretised microarray data (task in Section 3.1) or Gene Ontology data (task in Section 3.2). The key differences here are (1) that we are trying to learn a representation of the underlying networks with ILP, and (2) we are using experimental data only.

5 Conclusions

Reviewing progress [12] in the years since the sequencing of the yeast genome, the role of "human inference and domain knowledge" in functional genomics was emphasised. We believe our ILP approach to learning networks contributes to this goal, providing a path from large multi-relational data sets to comprehensible diagrams or other biologist-oriented applications of learned theories.

We note that it appears harder to learn to predict phenotype at the cellular level (such as sensitivity to environmental stresses) than a quantitative intracellular measure such as protein or gene expression. For further work we will continue to investigate this difference.

References

1. E. Akand, M. Bain, and M. Temple. Learning with Gene Ontology Annotation using Feature Selection and Construction. *(submitted)*, 2009.
2. L. Badea. Functional discrimination of gene expression patterns in terms of the gene ontology. In *Pacific Symposium on Biocomputing (PSB 2003)*, pages 565–576, 2003.
3. E. Boyle, S. Weng, J. Gollub, H. Jin, D. Botstein, J. M. Cherry, and G. Sherlock. GO::TermFinder – open source software for accessing Gene Ontology information and finding signficantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, 20(18):3710–3715, 2004.
4. Causton, H., Ren, B., Koh, S. et al. Remodeling of Yeast Genome Expression in Response to Environmental Changes. *Molecular Biology of the Cell*, 12:323–337, 2001.
5. E. Davidson. A view from the genome: spatial control of transcription in sea urchin development. *Curr. Opinion in Genetics and Development*, 9:530–541, 1999.
6. K. Dolinksi and D. Botstein. *Genome Research*, 15(12):1611–1619, 2005.
7. S. Fröhler and S. Kramer. Inductive logic programming for gene regulation prediction. *Machine Learning*, 70:225–240, 2008.
8. Godon, C., Lagniel, G., Lee, J. et al. The H202 Stimulon in Saccharomyces cerevisiae. *Journal of Biological Chemistry*, 273(34):22480–22489, 1998.
9. Harbison, C., Gordon, D., Lee, T. et al. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431:99–104, 2004.
10. T. Ideker, T. Galitski, and L. Hood. A new approach to decoding life: systems biology. *Ann. Review of Genomics and Human Genetics*, 2:343–372, 2001.
11. L. Steinmetz et al. *Nature Genetics*, 31(4):400–404, 2002.
12. L. Pĕna-Castillo and T. Hughes. Why Are There Still Over 1000 Uncharacterized Yeast Genes? *Genetics*, 176:7–14, 2007.
13. G. Thorpe, C. Fong, N. Alic, V. Higgins, and I. Dawes. Cells have distinct mechanisms to maintain protection against different reactive oxygen species: oxidative-stress-response genes. *Proc. Natl. Acad. Sci. USA*, 101(17):6564–9, 2004.
14. I. Trajkovski, F. Zelezny, N. Lavrač, and J. Tolar. Learning Relational Descriptions of Differentially Expressed Gene Groups. *IEEE Trans. Systems, Man and Cybernetics*, pages 1–10, 2007.
15. Winzeler, E., Shoemaker, D., Astromoff, A. et al. Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science*, 285:901–906, 1999.