Machine Learning for Systems Biology

S. H. Muggleton

Department of Computing,
Imperial College London.

Abstract. In this paper we survey work being conducted at Imperial College on the use of machine learning to build Systems Biology models of the effects of toxins on biochemical pathways. Several distinct, and complementary modelling techniques are being explored. Firstly, work is being conducted on applying Support-Vector ILP (SVILP) as an accurate means of screening high-toxicity molecules. Secondly, Bayes' networks have been machine-learned to provide causal maps of the effects of toxins on the network of metabolic reactions within cells. The data were derived from a study on the effects of hydrazine toxicity in rats. Although the resultant network can be partly explained in terms of existing KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway descriptions, several of the strong dependencies in the Bayes' network involve metabolite pairs with high separation in KEGG. Thirdly, in a complementary study KEGG pathways are being used as background knowledge for explaining the same data using a model constructed using Abductive ILP, a logic-based machine learning technique. With a binary prediction model (up/down regulation) cross validation results show that even with a restricted number of observed metabolites high predictive accuracy (80-90%) is achieved on unseen metabolite concentrations. Further increases in accuracy are achieved by allowing discovery of general rules from additional literature data on hydrazine inhibition. Ongoing work is aimed at formulating probabilistic logic models which combine the learned Bayes' network and ILP models.

1 Introduction

In the past experimental analysis of any single biological component, such as a gene or its protein product, was extremely time consuming. Consequently a single biology laboratory could study only a handful of such components at any one time. The recent revolution in high-throughput technologies offers an exciting opportunity to study such complex biological systems as an integrated whole. This new integrated approach to modelling of biological entities is known as Systems Biology. Systems Biologists use graph-based descriptions of bio-molecular interactions which describe cellular activities such as gene regulation, metabolism and transcription. Biologists build and maintain these network models based on the results of experiments in wild and mutated organisms. This paper will provide an overview of recent research in this area involving a consortium of computer scientists and biologists at Imperial College London. Some of the intrinsic interest in the area from a logic-based machine learning perspective include:

1. the availability of large-scale background knowledge on existing known biochemical networks from publicly available resources such as KEGG [2] (used in data sets such as those in [1, 11, 5]);
2. an abundance of training and test data from a variety of sources including microarray experiments (see for instance [3]) and metabolomic data [10] from NMR and mass spectroscopy experiments;
3. the inherent importance of the problem (see [6, 7]) owing to its application in biology and medicine;
4. the inherent relational structure in the form of spatial and temporal interactions of the molecules involved;

From a logical perspective the objects within this area include genes, proteins, metabolites, inhibitors and cofactors. The relations include biochemical reactions in which one set of metabolites is transformed to another in a biochemical reaction catalysed by an enzyme. One of the representational challenges is that within various databases the same object can be referred to in several ways.

A large part of the incentive for using machine learning techniques in this area comes from the incompleteness of detailed knowledge concerning the effects of inhibitors on known biochemical reactions. The requirement to infer such objects and relations indirectly from observational data necessitates the use of a mixture of abduction and induction within the ILP approaches to modelling in this problem.

Such models have wide potential application. For instance, in the new area of personalised medicines techniques which allow the construction of models of the toxic reactions of individuals to drug treatment would be of great benefit. Non-invasive testing, such as the NMR analysis of urine used in these studies, would be an appropriate basis for such modelling.

The paper is arranged as follows. Section 2 describes a novel approach to combining Support Vector Machines and ILP for directly predicting the effects of toxins on the basis of molecular features of the inhibitors. Section 3 describes the use of Bayes' network technology to estimate the structure and parameters of the causal network of interactions between metabolites whose up and down regulation patterns are observable within the NMR data. An ILP model built on the same data is described in Section 4. The model provides more detailed and testable predictions of the inhibited enzymes. Finally we conclude the paper in Section 5.

2 SVILP prediction of toxins

In [9] an accurate means of screening high-toxicity molecules is described. This approach uses a general method for constructing kernels for Support Vector Inductive Logic Programming (SVILP). The kernel not only captures the semantic and syntactic relational information contained in the data but also provides the flexibility of using arbitrary forms of structured and non-structured data coded in a relational way. While specialised kernels have been developed for strings, trees and graphs the approach uses declarative background knowledge to provide the learning bias. The use of explicitly encoded background knowledge distinguishes SVILP from existing relational kernels which in ILP-terms work purely at the atomic generalisation level.

	MSE	R-squared
CHEM	1.04	0.48
PLS	1.03	0.47
TOPKAT	2.2	0.26
SVILP	**0.8**	**0.57**

Fig. 1. MSE and R-squared for CHEM, PLS, TOPKAT and SVILP.

The SVILP approach is a form of generalisation relative to background knowledge, though the final combining function for the ILP-learned clauses is an SVM rather than a logical conjunction. SVILP was evaluated empirically against related approaches, including an industry-standard toxin predictor called TOPKAT. Evaluation was conducted on a broad-ranging toxicity dataset DSSTox [12]. Figure 1 shows the cross-validated error of SVILP compared to a number of alternative predictors on the DSSTox dataset. The results demonstrate that the approach significantly outperforms other state-of-the-art approaches on the wide-ranging set of toxins represented.

Such toxin-substructure based techniques, much like the ILP approach to predicting mutagenesis in [4, 13], are appropriate for large-scale screening of potential toxic side-effects of drugs. By contrast, techniques for detailed analysis of the causes of toxic reaction are addressed in the next two sections.

3 Bayes' network model for metabolic pathways

Metabolism comprises the network of chemical reactions involved in the biological processes of cells. These reactions are typically catalysed by enzymes and are highly interconnected.

In [15] a modular approach for representing metabolic pathways using Bayes' networks is described. The authors examined different models for a single reaction metabolism and introduced a Bayes' network model for this purpose. The performance of the model was compared to a Stochastic Logic Program representation for learning the aromatic amino acid pathway of yeast.

In subsequent work the authors have used this approach to model the effects of the toxin hydrazine administered to rats. The data were derived from Nuclear Magnetic Resonance (NMR) studies conducted by the Consortium for Metabonomic Toxicology (COMET) [10]. The derived Bayes' network is shown in Figure 2. Although the resultant network can be partly explained in terms of existing KEGG pathway descriptions, several of the strong dependencies in the Bayes' network involve metabolite pairs which are distant in the KEGG network.

4 Abductive ILP models of toxicity

In [14] the hydrazine NMR toxicity data studied previously using Bayes' nets was reanalysed within an ILP framework using Progol5.0 [8]. Figure 3 shows the approach

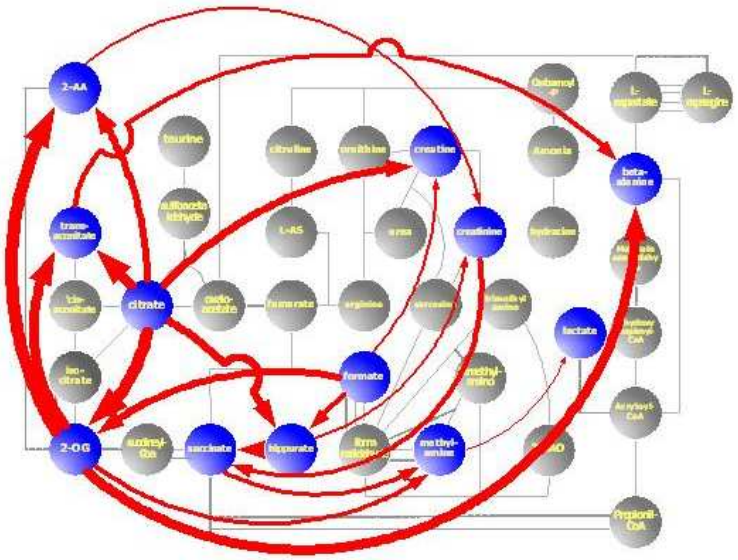

Fig. 2. Metabolic network showing the effects of hydrazine on rat metabolism. The thickness of arcs indicates the strength of dependencies between observable metabolites.

adopted. KEGG pathway descriptions were used as background knowledge. Only a limited subset (less than 10%) of the up/down regulation levels of metabolites in the KEGG model were directly observable within the NMR data. Progol5.0 was given this data as examples together with background knowledge consisting of the KEGG model and some general background rules concerning the transitive behaviour of the inhibitory effects of the toxin on various enzymes. From this it generated a set of ground hypotheses to explain the data in terms of inhibition of various enzymes. These ground hypotheses were then further generalised inductively together with known facts concerning the inhibition of various enzymes by hydrazine.

The resulting set of predicted inhibitions are shown in Figure 4. Owing to the spareseness of the known and documented inhibitory effects of hydrazine, all but one of the predicted inhibitory effects is novel. With the help of biological experts the model has been compared in detail with the Bayes' net model shown in Figure 2. In general the ILP model gives more detailed suggestions for the location of the inhibitory effects of the toxin. This level of detail allows for the possibility of laboratory testing of the inhibitory effects suggested by the ILP model.

The model was tested by randomly leaving out subsets of the examples and testing the predictions on the remaining observations. The resulting learning curves for abduction on its own versus the combination of abduction and subsequent induction are shown in Figure 5.

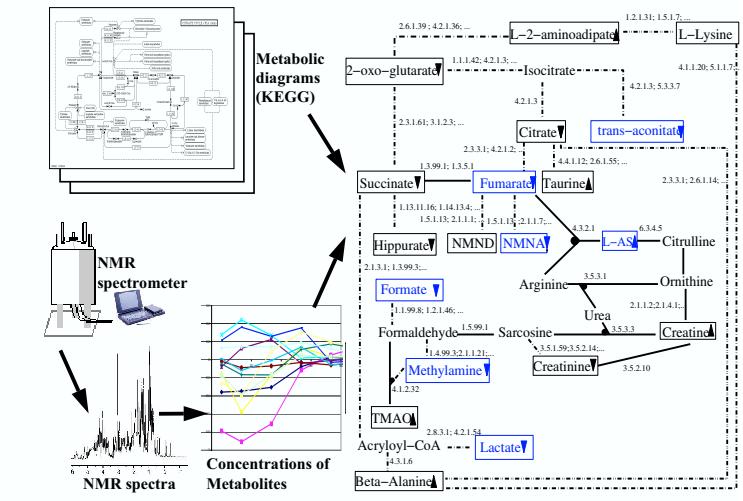

Fig. 3. Abductive ILP modelling of hydrazine toxicity

5 Conclusions and further work

In this paper we have described three distinct machine learning approaches to modelling the toxic effects of molecules. These three approaches should not be considered as being in competition, but rather as complementary approaches to the problem which can be used in a series of analytical phases. During drug development it is usual to consider a large set of potential candidates in the early stages of development. The SVILP toxicity predictor described in Section 2 has been proved to be at least as good as alternative state-of-the-art toxicity predictors in such a setting.

Having selected a particular candidate for further investigation it would be advantageous to apply more detailed analysis to the compound in question to understand its toxic effects better. NMR analyses of urine could then be conducted. Such an analysis incurs experimental costs, but could potentially be applied non-invasively in both animal and human testing. The results could then be modelled in a broad-brush fashion using the Bayes' network technology described in Section 3. An advantage of this approach is that no additional background knowledge of metabolism need be considered.

A more detailed analysis of inhibitory effects of the compound could be produced using the ILP modelling approach. It makes sense to apply ILP modelling last, since both the development of appropriate background knowledge and the computational costs of running ILP models incur more costs than the other two modelling approaches.

Many open questions are still to be addressed in this work. In particular, the treatment of time in both the Bayes' network and ILP models is presently inadequate. The NMR data are available as a time series measured at intervals of several hours over a 72 hour period. Within the Bayes net setting it may seem attractive to build dynamic Bayes' nets to deal with this temporal data. However, such models do not adequately account for the underlying causality of the domain. Metabolic reactions and fluxes take

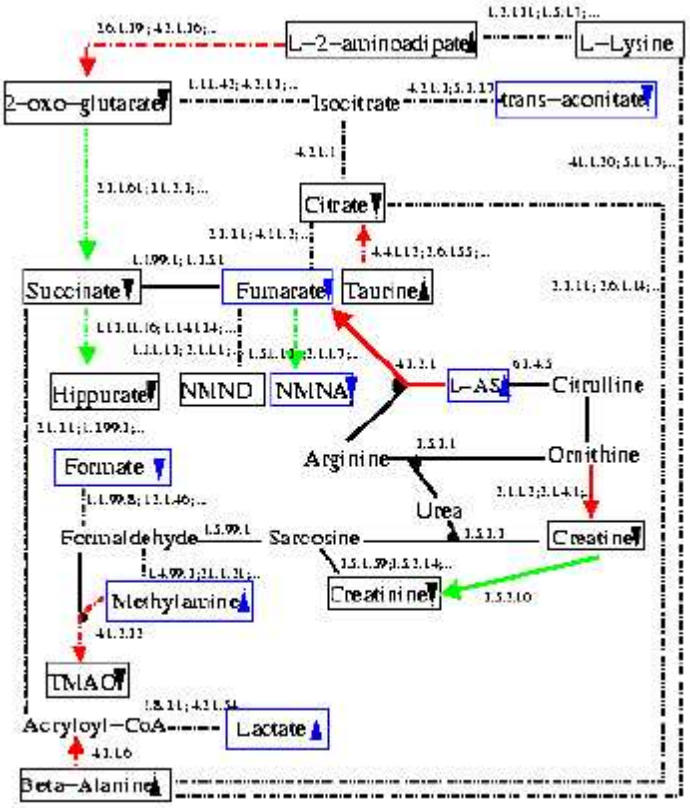

Fig. 4. Predicted inhibitions within the network. Modes represent metabolites with up and down regulation indicated by arrows when observable. Solid arcs represent single reactions, labelled by the catalysing enzyme's classification number. Dotted lines indicate a reaction sequence, with the list of associated enzymes. Red/green arrows indicate an inhibited/uninhibited (respectively) reaction (or reaction sequence). The arrow head shows the direction of inhibition.

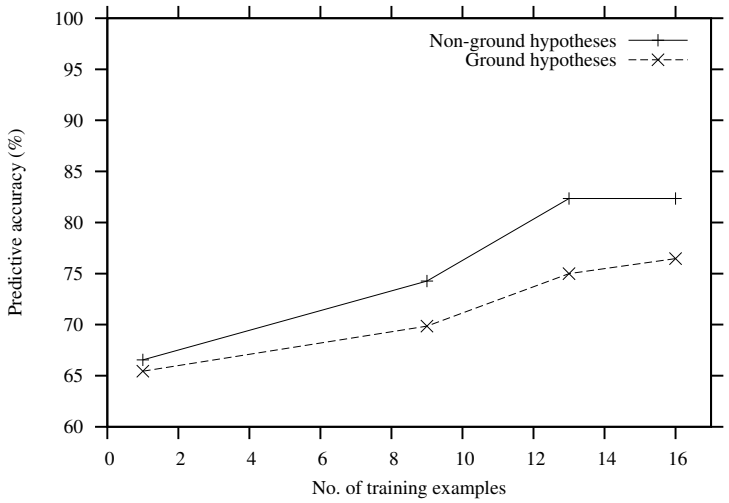

Fig. 5. Learning curves for abduction on its own versus abduction and subsequent induction. The X-axis indicates the number of examples in the training test after removal of a randomly chosen test set of varying sizes. The Y-axis gives predictive accuracy on the test set for results averaged over multiple trials.

place within a period of under 1 millisecond, while the intervals between readings are at least 8 hours. Persistence of inhibitory effects are due to the toxin remaining in the blood stream over an extended period of time. While it is possible to model such persistence axiomatically within a logical model, it is unclear how this could be achieved in the case of the Bayes' model.

Work is presently progressing on including aspects of both the Bayes' and ILP models within a Probabilistic Logic Programming model. Such integration holds the promise of allowing uncertainty to be expressed explicitly within ILP-generated models.

Finally, modelling within Systems Biology is a key application area for Machine Learning in general. The studies described in this paper indicate that ILP has the potential to be a key technology in an area which is now drawing major scientific interest internationally.

Acknowledgements

Many thanks are due to my wife, Thirza and daughter Clare for the support and happiness they give me. This work was supported by the DTI Beacon project "Metalog - Integrated Machine Learning of Metabolic Networks Applied to Predictive Toxicology", Grant Reference QCBB/C/012/00003, the ESPRIT IST project "Application of Probabilistic Inductive Logic Programming II (APRIL II)", Grant Reference FP-508861 and BBSRC Bio-informatics and E-Science Programme, "Studying Biochemical networks using probabilistic knowledge discovery", Grant Reference 28/BEP17011.

References

1. C.H. Bryant, S.H. Muggleton, S.G. Oliver, D.B. Kell, P. Reiser, and R.D. King. Combining inductive logic programming, active learning and robotics to discover the function of genes. *Electronic Transactions in Artificial Intelligence*, 5-B1(012):1–36, November 2001.
2. S. Goto, Y. Okuno, M. Hattori, T. Nishioka, , and M. Kanehisa. Ligand: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Research*, 30:402–404, 2002.
3. T.R. Hughes, M.J. Marton, A.R. Jones, C.J. Roberts, R. Stoughton, C.D. Armour, H.A.Bennett, E. Coffey, H. Dai, Y.D. He, M.J. Kidd, A.M. King, M.R. Meyer, D. Slade, P.Y. Lum, S.B. Stepaniants, D.D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, and S.H. Friend. Functional discovery via a compendium of expression profiles. *Cell*, 102(1):109–126, 2000.
4. R.D. King, S.H. Muggleton, A. Srinivasan, and M. Sternberg. Structure-activity relationships derived by machine learning: the use of atoms and their bond connectives to predict mutagenicity by inductive logic programming. *Proceedings of the National Academy of Sciences*, 93:438–442, 1996.
5. R.D. King, K.E. Whelan, F.M. Jones, P.K.G. Reiser, C.H. Bryant, S.H. Muggleton, D.B. Kell, and S.G. Oliver. Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature*, 427:247–252, 2004.
6. H. Kitano. Computational systems biology. *Nature*, 420:206–210, 2002.
7. H. Kitano. Systems biology: a brief overview. *Science*, 295:1662–1664, 2002.
8. S.H. Muggleton and C.H. Bryant. Theory completion using inverse entailment. In *Proc. of the 10th International Workshop on Inductive Logic Programming (ILP-00)*, pages 130–146, Berlin, 2000. Springer-Verlag.
9. S.H. Muggleton, H. Lodhi, A. Amini, and M.J.E. Sternberg. Support Vector Inductive Logic Programming. In D. Holmes and L.C. Jain, editors, *Recent Advances in Machine Learning*. Springer-Verlag, 2005. To appear.
10. J.K. Nicholson, J. Connelly, J.C. Lindon, and E. Holmes. Metabonomics: a platform for studying drug toxicity and gene function. *Nature Drug Discovery*, 1:153–161, 2002.
11. P.G.K. Reiser, R.D. King, D.B. Kell, S.H. Muggleton, C.H. Bryant, and S.G. Oliver. Developing a logical model of yeast metabolism. *Electronic Transactions in Artificial Intelligence*, 5-B2(024):223–244, November 2001.
12. A.M. Richard and C.R. Williams. Distributed structure-searchable toxicity (DSSTox) public database network: A proposal. *Mutation Research*, 499:27–52, 2000.
13. A. Srinivasan, S.H. Muggleton, R. King, and M. Sternberg. Theories for mutagenicity: a study of first-order and feature based induction. *Artificial Intelligence*, 85(1,2):277–299, 1996.
14. A. Tamaddoni-Nezhad, A. Kakas, S.H. Muggleton, and F. Pazos. Modelling inhibition in metabolic pathways through abduction and induction. In *Proceedings of the 14th International Conference on Inductive Logic Programming*. Springer-Verlag, 2004.
15. A. Tamaddoni-Nezhad, S. Muggleton, and J. Bang. A Bayesian model for metabolic pathways. In *International Joint Conference on Artificial Intelligence (IJCAI03) Workshop on Learning Statistical Models from Relational Data*, pages 50–57. IJCAI, 2003.