Variation of background knowledge in an industrial application of ILP

Stephen H. Muggleton*, Jianzhong Chen*, Hiroaki Watanabe*, Stuart J. Dunbar†, Charles Baxter†, Richard Currie†, José Domingo Salazar†, Jan Taubert+ and Michael J.E. Sternberg*

Imperial College London*
Syngenta Ltd†
BBSRC Rothamsted Research+

Abstract. In several recent papers ILP has been applied to Systems Biology problems, in which it has been used to fill gaps in the descriptions of biological networks. In the present paper we describe two new applications of this type in the area of plant biology. These applications are of particular interest to the agrochemical industry in which improvements in plant strains can have benefits for modelling crop development. The background knowledge in these applications is extensive and is derived from public databases in a Prolog format using a new system called Ondex (developers BBSRC Rothamsted). In this paper we explore the question of how much of this background knowledge it is beneficial to include, taking into account accuracy increases versus increases in learning time. The results indicate that relatively shallow background knowledge is needed to achieve maximum accuracy.

1 Introduction

Systems Biology is a rapidly evolving discipline that seeks to determine how complex biological systems function [6]. It works by integrating experimentally derived information with mathematical and computational models. Through an iterative process of experimentation and modelling, Systems Biology aims to understand how individual components interact to govern the functioning of the system as a whole. In several recent papers [11, 3], Inductive Logic Programing (ILP) has been applied to Systems Biology problems, in which it has been used to fill gaps in the descriptions of biological networks.

Two new industrial applications of this type in the area of plant biology are being explored within the Syngenta University Innovation Centre (see Section 2). This centre of excellence in Systems Biology aims to address biological research questions related to the improvement of plant strains involved in crops. The centre uses mathematical and computational modelling techniques developed at Imperial College London [2]. The background knowledge in these applications is generated by a system called Ondex [7]. Ondex is unique in its ability to generate large amounts of Prolog background knowledge on cell biochemistry by parsing, filtering and combining various publicly-available databases.

In this paper we use Ondex to explore the performance effects of varying the amount of background knowledge available to an ILP learning engine. This is done by generating variations of background knowledge using the Relation Neighbours Filter in Ondex together with a tehnique for sampling Relations. The effects of varying the background knowledge are measured on both learning time and predictive accuracy. The experimental results indicate that while learning time increases monotonically with the amount of background knowledge, relatively shallow degree of background knowledge is required to achieve maximum accuracy.

The paper is arranged as follows. In Section 2 we introduce the applications on which the experiments were conducted. We then describe the Ondex system for generating the background knowledge in Section 3. The experiments are then described in Section 4. Finally we conclude and describe further work in Section 5.

2 Application descriptions

2.1 University Innovation Centre (UIC) overview

Agricultural research relies on understanding interactions of genes and chemicals in a biological context. The search for a new biological trait to use in a conventional or genetic modification breeding programme is complex. It can take up to ten years and millions of dollars to bring such a development to market. The same is true for the search for new agrochemicals. Part of this development process is an assessment of the safety of the gene or chemical to the environment and its potential toxicity to both mammals and beneficial organisms. Systems Biology takes a new, integrated, approach to address these important challenges. Syngenta [1] is a leading Agrichemical company with a number one position in chemicals and is number three in high value seeds. Syngenta has established a "University Innovation Centre" (UIC) on Systems Biology at Imperial College London [2] to implement a "systems approach" to agricultural research. The centre has begun with two pioneer projects, tomato ripening and predictive toxicology.

2.2 Tomato application

The characteristics of the tomato fruit that reaches the consumer are defined by the combination of its biochemical and textural properties. Metabolic components (volatiles, pigments, sugars and amino acids) define the appearance and flavour whilst structural properties (cell adhesion, cell size, cuticle thickness, water content) define mouth-feel and texture perception. Together these components determine fruit quality and are crucial in influencing the success of commercial varieties.

At the genetic and biochemical level the regulation of fruit development and ripening remains poorly understood. In this project we are applying ILP to

deepen our understanding of the metabolic processes controlling tomato fruit development. By applying machine learning techniques to transcript and metabolite profiling data we are developing metabolic networks and building a predictive model of tomato ripening and fruit quality. Through the coordinated analysis of gene expression and metabolite changes across fruit development we aim to identify new genetic targets that play a role in controlling the ripening process. Such knowledge will allow us to focus on these genetic control points in breeding new tomato varieties, thus producing the most favourable combination of fruit quality characters in the ripe fruit.

2.3 Predictive Toxicology

An assessment of the potential to cause cancer is a key component of the risk assessment on a new Crop Protection Active Ingredient. The two year and 80 week bioassays in rats and mice, respectively, provide the Hazard Information to evaluate this risk. If tumours are observed in these trials, an assessment of their relevance for human risk may then be required. This typically makes use of the IPCS/HESI Human Relevance Framework, where the first step is the development of a mode-of-action case to describe the series of causal key events that lead to rodent tumours. The second step is to examine the plausibility of these key events occurring in humans and so guide an assessment of the relevance of the rodent findings.

This project aims to build a model that integrates the metabolic and gene expression regulatory networks that underlie initial key events in liver tumour promotion induced by model non-genotoxic carcinogens. It is envisaged that cycles of hypothesis generation informed by model building and experimental testing will allow the identification of those regulatory components that are key components in liver tumour promoting modes of action. Ultimately this will allow us to improve mechanistic understanding and so provide key data to explain the basis of the thresholds in dose and species specificity in response, thereby allowing more informed human health risk assessments.

3 Ondex: a Biological Background Knowledge Generator

Data integration in the life sciences still remains a significant challenge for bioinformatics [5]. Rather than developing a bespoke data integration solution for assembling the background knowledge for the machine learning task, the open source data integration framework Ondex [7] was selected as a general solution to bringing all the required data together. Ondex uses a graph-based approach with a data warehouse for integrating biological data. The nodes in the graph represent biological concepts, e.g. enzymes and metabolites. Edges in the graph represent relations between biological concepts, e.g. a set of enzymes catalyses a biochemical reaction. Both nodes and edges in the graph can have additional attributes, e.g. an enzyme name or an amino acid sequence. One of the reasons

for choosing the Ondex system as a background knowledge generator is the natural correspondence between its graph representation and the requirement of generating background knowledge as Prolog clauses for ILP learning.

Data from key biological pathway and gene function information resources including KEGG [10], LycoCyc [4] were transformed into a semantically consistent graph representation using Ondex. In order to create a non-redundant and coherent knowledge base, mapping methods were used to identify equivalent and similar entities among the different data sources. Once the databases were integrated, the resulting knowledge base was available for further analysis and visualization using the graph-based methods built in Ondex.

A key feature of the Ondex user client is that it allows the extraction of sub-graphs based on certain criteria. For example, it is simple to extract sub-graphs selected on by the class of biological concepts or relations, or where concepts posses a particular attribute, or from a graph-neighbourhood around particular nodes of interest. Such criteria can be combined in a workflow to manipulate the information to be included into the background knowledge. In order to support ILP learning, a general background knowledge generating utility was built that respected an agreed Prolog syntax, with Ondex concept class names and relation type names becoming predicate symbols and attributes becoming Prolog term structures. Every concept and relation was given a unique ID as the first argument in each predicate, which was used by other predicates to associate attributes with concepts and relations. The translation of attributes of concepts and relations were defined using the Ondex Prolog export utility. By following agreed conventions it was also possible to translate Prolog format back into an Ondex graph, thus enabling the results of the machine learning process to be imported back into Ondex where they could be visualised in the context of the original knowledge base.

4 Experiments

Two independent experiments[1] were conducted in the study to empirically investigate the *null hypothesis*: variations of background knowledge (BK) do not lead to increased predictive accuracy.

4.1 Materials and methods

An initial ground background knowledge base was derived from *LycoCyc* database [4, 8] and exported as Prolog format using the Ondex system. The knowledge base depicts the relational structure of tomato biological network (shown in Fig. 1), including the fundamental components, e.g. *compounds, reactions, enzymes*, and their relations, e.g. *consumed_by, produced_by, catalysed_by*, etc. Two types of raw

[1] The same methods were applied in both the tomato biology and the predictive toxicology applications with similar results. Due to space limitations, only the tomato biology application is reported.

data were provided by the domain experts in the experiments using gene mutants to study altered tomato ripening - concentration changes of metabolites and gene transcripts for four genotypes during 13 time slices. The data were expressed in terms of binary (up/down-regulation) for the purposes of applying ILP. In order to deal with the many-to-one relationships among gene transcripts, enzymes and reactions, a set of relevant transcriptomic data were further 'compressed' into one value using *SUM aggregate* for a reaction. Three datasets were chosen for modeling *tomato aspartate and the connected subnetwork*. Each contains the concentration changes of 10 metabolites as learning examples and 16 transcripts as observed facts on a particular time point[2].

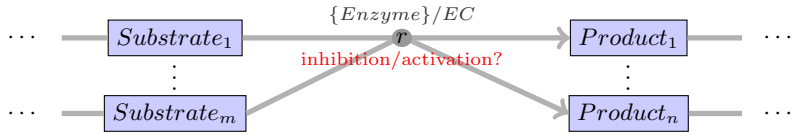

Fig. 1. Illustration of the relational background knowledge, where a set of compounds $\{Substrate\}_1^m$ are consumed by a reaction r, which produce a set of compounds $\{Product\}_1^n$; r is catalysed by a set of enzymes with an EC number $\{Enzyme\}/EC$ which is aggregated from a set of gene transcripts; the ILP learning is to abduce inhibition/activation occured in r.

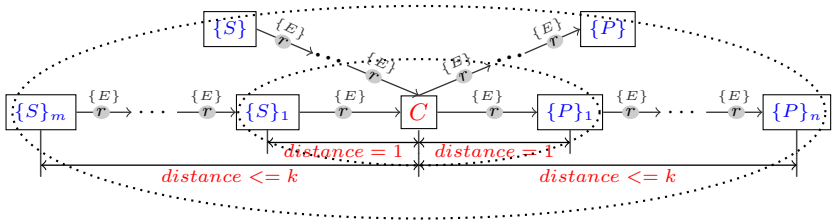

Fig. 2. Illustration of biological network structure and k-compound-neighbours (k-cn, $k \geq 1$) of a centroid compound C, where $\{S\}, \{P\}, \{E\}, r$ stand for a set of substrate compounds, product compounds, enzymes and a reaction, respectively

Variations of the background knowledge were generated using the Ondex *relation neighbours filter* (RNF), which extracts a subset from an original network given a set of centroid nodes and some distance k. In our experiments, the subnetwork consists of k-compound-neighbours (k-cn, as illustrated in Fig. 2) , which

[2] They respectively represent three time points - 15 days post anthesis, the breaker point when the fruit starts to change colour, and 7 days after the breaker stage.

is defined based on a corresponding k-reaction-neighbours (k-rn) as follows.

$$0\text{-cn} = \{\text{centroid compounds}\}$$

$$k\text{-rn} = \bigcup_{c \in (k-1)\text{-cn}} (r \mid \text{consumed\_by}(c, r)) + \bigcup_{c \in (k-1)\text{-cn}} (r \mid \text{produced\_by}(c, r))$$

$$k\text{-cn} = \bigcup_{r \in k\text{-rn}} (c \mid \text{consumed\_by}(c, r)) + \bigcup_{r \in k\text{-rn}} (c \mid \text{produced\_by}(c, r))$$

Furthermore, a *sampling relation neighbours filter* (SRNF) is developed in order to generate a series of evenly varied variations, containing sampled subsets of k-rn in which the size of k-cn can be controlled by a given sampling rate. The algorithm of SRNF is shown in Table 1.

1.0-cn={learning examples};
2.for each distance $k = 1, \ldots, K$, where K is a distance threshold corresponding to the original background knowledge base
2.1.k-rn={}, for each $c \in (k-1)$-cn k-rn = k-rn $\cup\{r \mid \text{consumed\_by}(c, r) \vee \text{produced\_by}(c, r)\}$;
2.2.ks-rn = sample(k-rn,sr), where sr is a manually set sampling rate;
2.3.k-cn={}, for each $r \in ks$-rn k-cn = k-cn $\cup\{c \mid \text{consumed\_by}(c, r) \vee \text{produced\_by}(c, r)\}$;
2.4.output k-cn.

Table 1. Algorithm of sampling relation neighbours filter (SRNF)

The datasets and variations of ground background knowledge were given to the abductive ILP [11] system Progol5.0 [9] together with a set of non-ground rules, which describe the underlying transitive behaviour of concentrations of metabolites and enzymes. Progol5.0 was then required to derive *inhibition* on reactions. Predictive performance was tested and evaluated against variations of the background knowledge (k-cn).

4.2 Results and discussion

Leave-one-out cross validation was used to evaluate the experiments, in which predictive accuracy and running time were computed for variation of the background knowledge. Fig. 3(a) shows that the sizes of k-cn generated by RNF increased sharply when $k \leq 2$ but only have minor changes when $k > 2$; whereas the sizes of k-cn generated by SRNF increase gradually and evenly with increasing values of k. Fig. 3(b) indicates that the running time increases linearly with the size of the background knowledge. Fig. 4 shows that (1) all the experiments get at least default predictive accuracy; (2) maximum accuracy could be achieved with relatively shallow bcakground knowledge (i.e. smaller k); (3) the sizes of background knowledge generated by SRNF at which the predictive

accuracy reaches its maximum value s smaller than those generated by RNF for two datasets, and are almost level in the third dataset.

In addition, we define a k-model to be the learning result (inhibition/activation abduced) using k-cn. A k-model will be referred to as *stable* if it is equivalent to a K-model (see step 2 of Table 1) with maximum accuracy. The vertical lines in Fig. 3 show that SRNF generates a smaller size background knowledge with less running time to reach the least k values of which k-model starts to be stable than RNF for all the three datasets ($k \geq 3$ for RNF and $k \geq 7$ for SRNF). In summary, it is possible to find more stable, shallow and fine (or smaller) background knowledge that achieves maximum predictive accuracy with less running time by using SRNF rather than RNF. SRNF also enables us to investigate the finer changes between variation of BK in a controllable way. The null hypothesis set for the experiments has been rejected by these results.

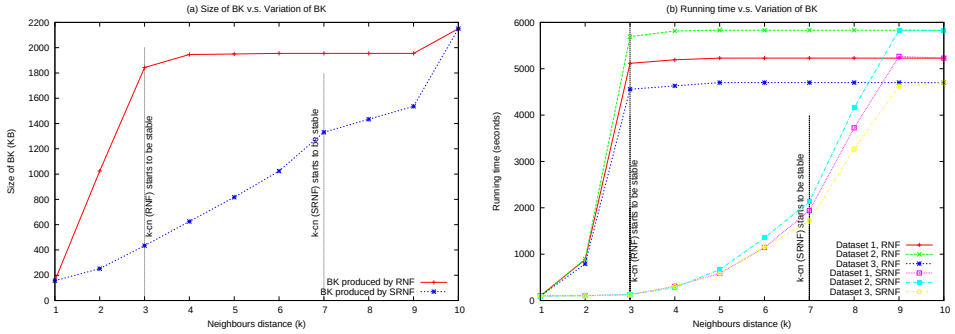

Fig. 3. (a) Size of variation of BK (b) Running time v.s. variation of BK (k-cn) produced by *RNF* and *SRNF* for the three datasets

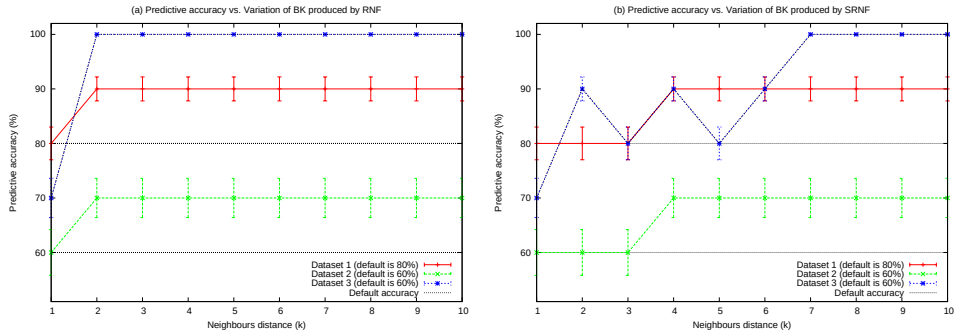

Fig. 4. Predictive accuracy v.s. variation of BK (k-cn) produced by (a) *RNF* and (b) *SRNF* for the three datasets

5 Conclusions and further work

This paper explores the application of ILP to Systems Biology. These applications involve modelling interations between components of biological systems using abductive inductive logic programing. We also introduce a powerful new system called Ondex which can be used to generate Prolog background knowledge iby parsing and filtering public dataases on cell biochemistry. Two industrial applications are described which are being studied in the Syngenta University Innovation Centre. With the extensive background knowledge generated, we explore the question of how variations of background knowledge affect learning time and predictive accuracy of the same ILP learning system.

Through two independent experiments in tomato biology and predictive toxicology, we conclude that relatively shallow background knowledge can be used to achieve maximum accuracy. In addition the experiments indicate that use of neighbourhod further reduces the learning time required to achieve maximum accuracy.

In further work we aim to improve the non-ground background knowledge rules used in the experiments. We also intend to extend the results using datasets of all time points, and investigate the biological significance of the learned theories.

Acknowledgments

The authors would like to acknowledge the support of Syngenta in its fundig of the University innovations Centre at Imperial College. The first author would also like to thank the Royal Academy of Engineering and Microsoft Research for their support of his research chair.

References

1. Syngenta Ltd. http://www.syngenta.com/en/index.html.
2. Syngenta University Innovation Centre. http://www3.imperial.ac.uk/syngenta-uic.
3. J. Chen, S.H. Muggleton, and J. Santos. Learning probabilistic logic models from probabilistic examples. *Machine Learning*, 73(1):55–85, 2008. 10.1007/s10994-008-5076-4.
4. L.A. Mueller et al. The SOL Genomics Network. A Comparative Resource for Solanaceae Biology and Beyond. *Plant Physiology*, 138(3):1310–1317, 2005.
5. C. Goble and R. Stevens. State of the nation in data integration for bioinformatics. *Journal of Biomedical Informatics*, 41(5):687–693, 2008.
6. H. Kitano. Computational systems biology. *Nature*, 420:206–210, 2002.
7. J. Kohler, J. Baumbach, J. Taubert, M. Specht, A. Skusa, A. Ruegg, C. Rawlings, P. Verrier, and S. Philippi. Graph-based analysis and visualization of experimental results with ONDEX. *Bioinformatics*, 22(11):1383–1390, 2006.
8. LycoCyc. Solanum lycopersicum database. http://solcyc.solgenomics.net//LYCO/.

9. 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.
10. H Ogata, S Goto, K Sato, W Fujibuchi, H Bono, and M Kanehisa. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucl. Acids Res.*, 27(1):29–34, 1999.
11. A. Tamaddoni-Nezhad, R. Chaleil, A. Kakas, and S.H. Muggleton. Application of abductive ILP to learning metabolic network inhibition from temporal data. *Machine Learning*, 64:209–230, 2006. DOI: 10.1007/s10994-006-8988-x.