Does Multi-clause Learning Help in Real-world Applications?

Dianhuan Lin*, Jianzhong Chen*, Hiroaki Watanabe*, Stephen H. Muggleton*,
Pooja Jain*, Michael J.E. Sternberg*, Charles Baxter†, Richard A. Currie†,
Stuart J. Dunbar†, Mark Earll†, José Domingo Salazar†

Imperial College London*
Syngenta Ltd†

Abstract. The ILP system Progol is incomplete in not being able to derive a multi-clause hypothesis from an example. However, due to the assumption that a multi-clause hypothesis can be built by sequentially adding single clauses, Progol's incompleteness does not stop it being applied to real-world applications. This paper uses two real-world applications in systems biology to study whether a complete multi-clause learning method MC-TopLog can make a difference to learning results compared to the single-clause learning method Progol5. The experimental results show that in both applications there exist data sets, in which hypotheses derived by MC-TopLog have higher predictive accuracies, as well as better biological significance than those of Progol5.

1 Introduction

Yamamoto [12] first pointed out that Progol's inverse entailment [3] is incomplete, which means Progol can only derive hypotheses with a single clause, but not with multiple clauses from an example. The learning of the concept of odd-numbers was used as a counter example by Yamamoto to demonstrate Progol's incompleteness. However, his example contains a number of particular facets, such as mutual recursion. In this paper we investigate whether the form of incompleteness studied by Yamamoto significantly affects learning performance in two real-world applications. Neither application involves mutual recursion, and although the background knowledge is recursive in both cases, the hypothesis space consists of ground facts. In such a case one might imagine that it should be possible to always build up the hypothesis sequentially by adding single facts to explain individual examples. For example, a network of food webs, whose logical description consists of multiple clauses, can be constructed from scratch using Progol5 [4], as shown in [10].

Although other multi-clause learning methods have been applied to real-world domains [8] [13], where there are no mutual recursion, no direct comparison to an single-clause learning method has been made using experiments. Therefore, it is still unclear whether the assumption that single-clause learners can perform as good as multi-clause learners by sequentially constructing each clause in a multi-clause hypothesis is still valid in applications like those studied in [8] and [13]. The experiments in this paper, where direct comparisons between

MC-TopLog and Progol5 were made using the same data sets, demonstrate that a complete multi-clause learning method can significantly improve learning results of certain real-world problems, compared to a single-clause learning method.

The two real-world applications studied in this paper are tomato and predictive toxicology applications. Both come from Syngenta [1], which is a world leading agribusiness supplying crop protection and genetic solutions to growers. These two applications were also used in [5] to study how the variation of the background knowledge affects learning results.

Developing new varieties of tomato is a major part of Syngenta‘s vegetable seed business. The tomato application aims to identify new molecular-genetic targets that play a role in controlling tomato ripening and fruit quality. The output of this project aims to increase the efficiency of breeding selection processes, resulting in new tomato varieties optimised for shelf life and quality.

The predictive toxicology application is important to the sector of crop protection at Syngenta, since an assessment of the potential to cause cancer is a key component in the risk assessment of a new crop protection active ingredient. The objective of the predictive toxicology application is to devise a predictive model describing xenobiotic-induced alterations in metabolism in the rat that underlie tumour promoting activity. This model will direct experimental design and choices to reduce the cost and number of experiments needed to predict toxicological end points from a range of chemistries.

The rest of the paper will describe ILP models of the two applications first, and then explain the definition of multi-clause learning in the context of the applications. Finally, the experimental results are presented.

2 ILP Models for Tomato and Predictive Toxicology Applications

The abstract models of tomato and predictive toxicology applications are similar, although their underlying biological processes are different. In both applications, changes in the metabolite abundances are observed in the treated group compared to the control group. In the tomato application, the treated groups are ripening mutants, such as colourless non-ripening (CNR), ripening-inhibitor (RIN) and non-ripening (NOR); in the predictive toxicology application, the treated groups are Fischer F344 rats treated with different doses of phenobarbital (a non-genotoxic liver carcinogen). The observed changes in the metabolite abundances are classified into *up*, *down* and *no-change*, and used as examples E.

The background knowledge B consists of: (1)The regulation rules about how changes in reaction states affect metabolite abundances. (2) Metabolic network. For tomato application, it is derived from LycoCyc database [2], while for the predictive toxicology application, it is obtained from KEGG database [7]. (3) Gene expression (transcript profiles). A longer version paper will provide details about how to use the transcript information for constructing hypotheses. Fig. 1 shows a part of the background knowledge.

```
Background Knowledge B:
(1) Regulation Rules (9 rules in total):
concentration(Metabolite1, up, Time) ←
                    produced_by(Metabolite1,Reaction),
                    reactionState(Reaction, enzymeLimiting, cataIncreased,Time).
concentration(Metabolite1, down, Time) ←
                    consumed_by(Metabolite1,Reaction),
                    reactionState(Reaction, enzymeLimiting, cataIncreased,Time).
concentration(Metabolite1, up, Time) ←
                    produced_by(Metabolite1,Reaction),
                    reactionState(Reaction, substrateLimiting, _,Time),
                    consumed_by(Metabolite2,Reaction),
                    concentration(Metabolite2,up,Time).
(2) Metabolic Network:
consumed_by(glutamate,‘L-GLU:L-CYS γ-LIGASE’).
produced_by(γ–glutamylcysteine,‘L-GLU:L-CYS γ-LIGASE’).
catalyzed_by(‘L-GLU:L-CYS γ-LIGASE’, ‘glutamate–cysteine ligase’).
part_of_catalysing_class(‘glutamate–cysteine ligase’,‘E.C.6.3.2.2’).
(3) Gene Expression Data:
concentration_e(‘E.C.6.3.2.2’,up,day14).      concentration_e(‘E.C.6.3.2.3’,up,day14).
Examples E:
e1: concentration(glutathione,up,day14).      e2: concentration(5–oxoproline,up,day14).
Candidate Hypothesis Clauses:
h1: reaction_state(‘γ-L-GLU-L-CYS:GLY LIGASE’, substrateLimiting, _ , day14 ).
h2: reaction_state(‘5-GLUTAMYLTRANSFERASE’, substrateLimiting, _ , day14 ).
h3: reaction_state(‘L-GLU:L-CYS γ-LIGASE’, enzymeLimiting, cataIncreased, day14 ).
```

(a) Predictive Toxicology Application

```
Examples: e4: concentration(citrate,down,‘NOR_Late’). e5: concentration(malate,up,‘NOR_Late’).
Candidate Hypothesis Clauses:
h4: reaction_state(‘CITSYN-RXN’, enzymeLimiting, cataIncreased, ‘NOR_Late’).
h5: reaction_state(‘MALATE-DEH-RXN’, substrateLimiting, _ , ‘NOR_Late’).
h6: reaction_state(‘ACONITATE-DEHYDR-RXN’, enzymeLimiting, cataDecreased, ‘NOR_Late’).
```

(b) Tomato Application

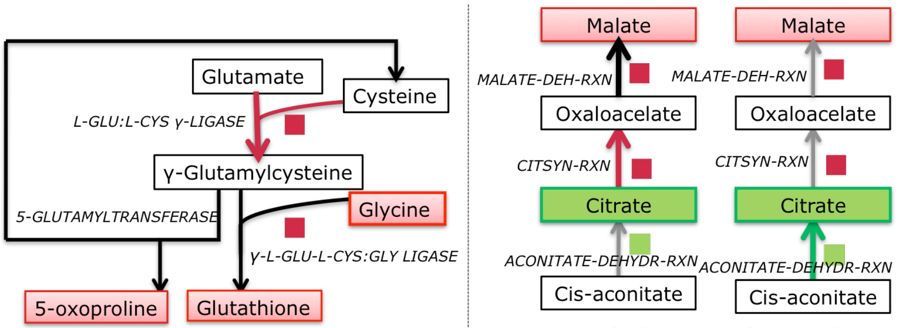
Fig. 1: ILP Models

To explain examples E, we need to hypothesise changes in reaction states, which are not observable. Reaction states can be classified as enzyme limiting or substrate limiting. Enzyme limiting implies that the flux through the reaction is controlled by the activity of the catalysing enzyme that can either be catalytically increased, decreased or no-change. Similarly, substrate limiting means the flux through the reaction is determined by the abundance of substrates that can be observed either as up, down or no-change.

3 Multi-clause Learning

Progol's incompleteness can be characterised by single-clause learning, because Progol can only derive a single-clause hypothesis that subsumes an example e relative to B in Plotkin's sense. More details of multi-clause learning *vs.* single-clause learning can be found in [6].

In the context of the two applications studied in this paper, deriving a multi-clause hypothesis means hypothesising multiple reaction states. Fig. 2(1) gives an example of a multi-clause hypothesis H_1 from the predictive toxicology application. H_1 consists of the three clauses h_1, h_2 and h_3 in Fig. 1(a), and it explains e_1 as well as e_2. Specifically, both reactions ‘γ-L-GLU-L-CYS:GLY LIGASE’ and ‘5-GLUTAMYLTRANSFERASE’ are hypothesised to be substrate limiting, thus flux through them depends on the abundance of their common substrate γ-glutamylcysteine. While the reaction ‘L-GLU:L-CYS γ-LIGASE’ that



(1) A Multi-clause hypothesis in Glutathione Pathway (Predictive Toxicology Application)

(a) Multi-clause (b) Single-clause

(2) Hypotheses in TCA cycle (Tomato Application)

Fig. 2: Hypothesis Visualization. A reaction arrow is grey if it is not hypothesised, otherwise, it is coloured red, green or black to represent catalytically increased, decreased and substrate limiting reaction states, respectively. Metabolite abundance and gene expression (small squares beside the reaction arrows) are coloured red, green and black to represent up, down and no-change, respectively.

produces γ-glutamylcysteine is hypothesised as enzyme limiting and catalytically increased, which means the enzyme catalysing 'L-GLU:L-CYS γ-LIGASE' regulates the abundance of glutathione and 5-oxoproline both. This hypothesis is indeed consistent with that in [11]. However, such a multi-clause hypothesis can not be derived by a single-clause learner like Progol5, unless the abundance of γ-glutamylcysteine is available[1], while that is practically immeasurable due to technological limitations. Another multi-clause hypothesis $H_{2a}=\{h_4, h_5\}$ (where h_4 and h_5 are in Fig. 1(b)) is shown in Fig. 2(2)(a). This hypothesis is possible to be derived by a single-clause learner. Specifically, the single clause h_4 can be derived from the example e_4. After h_4 is added to the background knowledge, another clause h_5 can be derived from the example e_5. Despite the fact that H_{2a} can be sequentially constructed using Progol5, Progol5 does not necessarily suggest this hypothesis, but instead hypothesises $H_{2b}=\{h_6\}$ shown in Fig. 2(2)(b).

4 Experiments

Two independent experiments were conducted to empirically investigate the null hypothesis: multi-clause learning does not outperform single-clause learning.

Materials In the tomato application, transcript and metabolite profiles for three developmental stages (Early, Mid and Late) were obtained for wild type and three mutants (CNR, RIN, NOR) from Syngenta. This gave nine data sets in total (3 stages x 3 mutants). In the cancer application, transcript and metabolite profiles were obtained for 1, 3, 7 and 14 days post treatment, which were from a published study [11]. All the materials used in the experiments can be found at http://ilp.doc.ic.ac.uk/mcTopLog.

[1] Suppose $concentration(\gamma\text{–}glutamylcysteine, up, day14)$ exists as an example e_3, then a single-clause learner can sequentially derive each clause in H respectively from e_1, e_2 and e_3.

Methods Progol5 and MC-TopLog [6] were used to represent single-clause learner and multi-clause learner, respectively. In the tomato application, leave-one-out cross validation was used to compute the predictive accuracies due to the availability of a limited set of abundance data (22 metabolites). However, in the predictive toxicology application 10-fold cross validation was employed as a larger set of metabolite abundance data (52 metabolites) was available. The closed world assumption was applied during the testing phase to define an un-hypothesised reaction state as substrate limiting.

Results The following two tables show the predictive accuracies of two applications using Progol5 and MC-TopLog. The accuracies of both Progol5 and MC-TopLog are higher than default, which shows learning is effective. Compared to Progol5, MC-TopLog suggested hypotheses with higher accuracies for certain data sets. In the tomato application, there are five (CNR_Mid, CNR_Late, NOR_Mid, NOR_Late, RIN_Late) out of nine data sets, in which MC-TopLog's accuracies are significantly higher than that of Progol5 at the 95% confidence level. While in the predictive toxicology application, only for the data at day 14 that MC-TopLog performs significantly better. Overall our null hypothesis is rejected by the accuracy results: there exist cases where multi-clause learning significantly outperforms single-clause learning.

Timepoint	**default(no change)**,%	**Progol**,%	**MC-TopLog**,%	p-value
CNR_Early	63.64	72.73±9.49	81.82±8.22	0.162
CNR_Mid	36.36	36.36±10.26	77.27±8.93	0.001
CNR_Late	40.90	54.55±10.62	86.36±7.32	0.005
NOR_Early	86.36	86.36±7.32	86.36±7.32	1.000
NOR_Mid	50.00	59.09±10.48	77.27±8.93	0.043
NOR_Late	31.82	40.91±10.48	81.82±8.22	0.001
RIN_Early	100.00	100±0.00	100.00±0.00	1.000
RIN_Mid	90.91	90.91±6.13	90.91±6.13	1.000
RIN_Late	36.36	45.45±10.62	81.82±8.22	0.002

Table 1: Predictive accuracies with standard errors in Tomato Application

Timepoint	**default(no change)**,%	**Progol**,%	**MC-TopLog**,%	p-value
Day 1	55.0	75.00±6.06	78.0±5.74	0.7304
Day 3	30.6	56.66±6.87	59.00±6.82	0.5554
Day 7	40.6	60.33±6.78	66.00±6.57	0.4250
Day 14	48.0	50.33±6.93	68.00 ±6.47	0.0039

Table 2: Predictive accuracies with standard errors in Cancer Application

Hypothesis Interpretation Here we compare the hypotheses generated by MC-TopLog and Progol5 by examining their biological significance from published data. The example used here is a hypothesis around organic acids citrate and malate, both of which are targets for improving fruit quality. Fig 2(2)(a) depicts a multi-clause hypothesis derived by MC-TopLog. It suggests that the enzyme catalysing the reaction 'CITSYN-RXN' (citrate synthase) regulates the abundance of citrate as well as malate. Different from MC-TopLog, Progol5 hypothesises the enzyme aconitate hydratase to be the regulator of citrate abundance, as shown in Fig 2(2)(b). From a biological perspective, both hypotheses are plausible, and the role of both enzymes have not been functionally tested in tomato fruit. However, the MC-TopLog hypothesis is supported by indirect evidence [9], in which down regulated expression of citrate synthase in tomato leaves influences the abundance of citrate and malate. Those plausible hypotheses without current literature validation will be experimentally tested and validated by biologists in a future study.

5 Conclusion

The two real-world applications studied in this paper do not involve recursive target hypotheses and mutually dependent predicates like Yamamoto's odd–numbers example. However, multi-clause learning still outperforms single-clause learning in certain data sets of the two applications, as shown by the direct comparisons. Therefore, we conclude that there exist general classes of problems in the real-world that benefit from complete multi-clause learning methods.

References

1. Syngenta Ltd. http://www.syngenta.com/en/index.html.
2. LycoCyc. Solanum lycopersicum database. http://solcyc.solgenomics.net//LYCO/.
3. S.H. Muggleton. Inverse entailment and Progol. *New Generation Computing*, 13:245–286, 1995.
4. S.H. Muggleton and C.H. Bryant. Theory completion using inverse entailment. In *ILP-00*, pages 130–146. Springer-Verlag, 2000.
5. S.H. Muggleton, J. Chen, H. Watanabe, S. Dunbar, C. Baxter, R. Currie, J.D. Salazar, J. Taubert, and M.J.E. Sternberg. Variation of background knowledge in an industrial application of ILP. 2010. To appear in ILP2010.
6. S.H. Muggleton, D. Lin, and A. Tamaddoni-Nezhad. MC-TopLog: complete multi-clause learning guided by a top theory. 2011. Submitted to ILP2011.
7. 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.
8. Oliver Ray, Ken Whelan, and Ross King. Automatic revision of metabolic networks through logical analysis of experimental data. In *ILP2009*, pages 194–201, 2009.
9. A. Sienkiewicz-Porzucek and A. Nunes-Nesi et al. Mild reductions in mitochondrial citrate synthase activity result in a compromised nitrate assimilation and reduced leaf pigmentation but have no effect on photosynthetic performance or growth. plant physiolog. *Plant Physiology*, 147:115–127, 2008.
10. A. Tamaddoni-Nezhad, D. Bohan, A. Raybould, and S.H. Muggleton. Machine learning a probabilistic network of ecological interactions. Submitted to ILP2011.
11. C.L. Waterman and R.A. Currie et al. An integrated functional genomic study of acute phenobarbital exposure in the rat. *BMC Genomics*, 11(1):9, 2010.
12. A. Yamamoto. Which hypotheses can be found with inverse entailment? In N. Lavrač and S. Džeroski, editors, *ILP97*, pages 296–308. Springer-Verlag, 1997.
13. Y. Yamamoto, K. Inoue, and A. Doncescu. Integrating abduction and induction in biological inference using cf-induction. In Huma Lodhi and Stephen Muggleton, editors, *Elements of Computational Systems Biology*, pages 213–234. 2010.