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 generalise a single example to multiple clauses. However, according to the Blumer bound, incomplete learners such as Progol, can have higher predictive accuracy using less search than more complete learners. This issue is particularly relevant in real-world problems, in which it is unclear whether the unknown target theory is within the hypothesis space of the incomplete learner. This paper uses two real-world applications in systems biology to study whether there exist datasets where a complete multi-clause learning (MCL) method can significantly outperform a single-clause learning (SCL) method. The experimental results show that in both applications there do exist datasets, in which hypotheses derived by MCL have significantly higher predictive accuracies. On the other hand, for most of the datasets in the two applications, there are good approximations of the target within the hypothesis space of SCL, so that MCL does not outperform SCL.

1 Introduction

Progol's inverse entailment [10] is incomplete, as first pointed out by Yamamoto [21]: Progol can only generalise a single example to a single clause, but not multiple clauses. This type of entailment-incompleteness can be characterised by single-clause learning (SCL). In contrast, entailment-complete methods are referred as multi-clause learning (MCL) in this paper.

1.1 Relationship between Completeness and Accuracy

It might be imagined that by achieving completeness of search, a learning algorithm necessarily increases the accuracy of prediction on unseen examples. However, the Blumer bound [2] indicates this is not necessarily the case.

$$\textbf{Blumer bound } \; m \geq \frac{1}{\epsilon}(ln|H| + ln\frac{1}{\delta})$$

In the above m stands for the number of training examples, ϵ is the bound on the error, $|H|$ is the cardinality of the hypothesis space and $(1-\delta)$ is the bound on the probability with which the inequality holds for a randomly chosen set of training examples. Note that by increasing $|H|$ you increase the bound on the required training set size. Given a fixed training set for which the bound holds as an equality, the increase in $|H|$ would need to be balanced by an increase in ϵ, i.e. a larger bound on predictive error. Therefore on the face of it, the Blumer bound indicates that incomplete learning algorithms have lower bounds on error

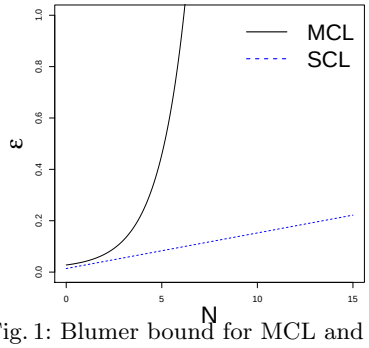


Fig. 1: Blumer bound for MCL and SCL

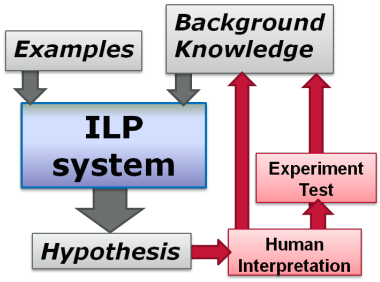


Fig. 2: Learning Cycle in ILP

than complete ones. And this difference in error bounds is not insignificant, as shown in Fig. 1. $|H|$ is 2^N for SCL, where N is the number of distinct atoms derivable from a hypothesis language, while it is 2^{2^N} for MCL. Therefore the Blumer bound for SCL and MCL are as below. In terms of running time, MCL also takes much longer than SCL, as its $|H|$ is much bigger. Overall, the Blumer bound indicates that in the case that the target theory is within both hypothesis spaces, MCL is worse than SCL, both in running time and in terms of predictive error bounds for randomly chosen training sets.

$$\textbf{SCL's Blumer bound } \epsilon \geq \frac{1}{m}(Nln2 + ln\frac{1}{\delta}),\ |H| = 2^N$$

$$\textbf{MCL's Blumer bound } \epsilon \geq \frac{1}{m}(2^N ln2 + ln\frac{1}{\delta}),\ |H| = 2^{2^N}$$

However, the Blumer bound only holds if the target theory is within the hypothesis space for both algorithms. In the case that target theory is within the hypothesis space of the complete learner but not within the hypothesis space of the incomplete learner then the complete learner will have a lower error bound. For an artificial dataset, it is possible to decide whether the target theory is within the hypothesis space before the learning is carried out. However, this is not the case for a real-world dataset. So the motivation for this paper was to see whether completeness in learning does lead to higher accuracy in at least one real-world dataset.

1.2 Complete and Incomplete ILP systems

Within ILP much effort has been put into designing methods that are complete for hypothesis finding. Multi-clause learning (MCL) systems like CF-Induction [6], XHAIL [15], TAL [4] and MC-TopLog [13] were designed to overcome Progol's entailment-incompleteness. However, as discussed above, it is not clear whether completeness is important in real-world applications. Although some of the MCL systems have been applied to real-world domains [9,16,22], no direct comparison to a single-clause learning (SCL) method (e.g. Progol) has been made using experiments[1]. On the other hand, Progol's entailment-incompleteness does not

[1] Although [7] has compared CF-Induction to Progol, no predictive accuracies are provided, but only learned hypotheses ranked by a probability measure. Although Progol's hypothesis is only ranked at 13th, it does not mean it has lower predictive accuracy than the one ranked at top.

stop it being applied to real-world applications, because in certain cases, it is possible to construct a multi-clause hypothesis by sequentially adding single clauses. For example, a network of food webs, whose logical description consists of multiple clauses, can be constructed from scratch using Progol5 as shown in [17]. Therefore, it is still unclear, for applications such as those studied in this paper, whether it is necessary to use MCL, which is much more computationally expensive than SCL.

The experiments in this paper, where direct comparisons between MC-TopLog and Progol5 were made using the same datasets, demonstrate that there indeed exist datasets where a MCL method achieves significantly higher accuracies than a SCL method. On the other hand, this is not always the case, as there are also datasets where a good approximation to the target theory exists in the hypothesis space of a SCL method, so that SCL can have lower error than MCL, as suggested by the Blumer bound.

1.3 Two Biological Applications

The two biological applications studied in this work are of commercial interest to Syngenta [1], which is a leading agribusiness company providing crop protection and genetic solutions to growers. Developing tomato varieties optimised for the shelf life, flavour and nutritional quality is a major part of Syngenta's breed selection and seed development program. The aim of applying ILP approach in this program is to identify genetic control points regulating metabolic changes that occur during tomato fruit ripening. The other application about predictive toxicology is important to Syngentas crop protection initiatives. The objective is to identify control points for metabolic pathway perturbations caused by a model liver tumour promoter (phenobarbital) in the rat liver. In both applications, the respective predictive models would potentially influence the experimental design by generating not only testable hypotheses but providing explanations as well, thus saving time, experimental cost and labour involved with cycles of trial runs.

Why ILP? For centuries scientists have used telescopes and microscopes to enhance their natural abilities to perceive the world. In an analogous way ILP can be used to magnify the abilities of scientists to reason about complex datasets. The biological applications to which ILP systems are applied in this work are typical of situations in which biologists have limited comprehension of the impact of perturbing a cellular pathway. The scale of the metabolic network and the interconnections among various pathways add another challenge to overcome. For example, during the tomato ripening, the genes that control the texture may also indirectly affect the flavour. It would be undesirable to sacrifice the taste of tomato to its firmness, although the firmness improves the shelf life. Therefore, all pathways related to flavour, texture and colour have to be considered together, which is difficult for biologists to conceptualise. Biologists therefore need a testable hypothesis suggested by an ILP system in order to carry out their studies. This is where ILP comes to their aid.

ILP has the advantage of suggesting readily comprehensible hypotheses, due to the use of logic programs as a uniform representation for B, E and H. Biologists can then examine the hypotheses using their existing knowledge. Those

plausible hypotheses that are impossible to be disproved can be considered for further experimental validation, while a biologically non-meaningful hypothesis may indicate that insufficient background knowledge has been provided. Being a knowledge discovery task it is often difficult to know a priori, the depth of the knowledge required to circumvent such non-meaningful hypotheses. For example, in the predictive toxicology application, there are candidate hypotheses that explain the decrease of glucose and fructose from the reactions that produce them. However, in the given environment a decrease in glucose and fructose can only be explained by the reactions consuming them. Therefore, we updated the background knowledge with this knowledge as an integrity constraint to filter non-meaningful hypotheses. No matter whether a suggested hypothesis is disproved by biologists' existing knowledge or further tested by experiments, the background knowledge needs to be updated. It is ILP techniques that make such learning cycle feasible in a controlled manner. The diagram in Fig. 2 not only shows such a learning cycle, but also highlights the fact that an ILP system does need scientists' help in providing/updating its input and interpreting its output. This supports our analogy of ILP technique as tools, which enhance scientists' capacity, rather than making scientists redundant.

Why not Technologies other than ILP? In the two applications considered, the learning target is the reaction state that is not observable and could be simply ground facts. Therefore, using abduction alone seems sufficient. However, an abductive system suggests all of the candidate hypotheses instead of the most promising ones. Although, an algorithm for ranking abductive hypotheses has already been proposed [7], it is not applicable to the current study due to the sheer number of candidate hypotheses generated[2]. Hence, in this study compression is used to select the most promising candidate hypotheses for further interpretation by biologists and/or experimental validation.

Difference from Previous ILP Applications The use of transcriptomics as well as metabolomics data in the modelling distinguishes the two applications from the previous biological application of ILP, e.g. MetaLog project [18]. This integrative omics approach is also different from the traditional approach used by biologists, where only transcriptomic data from treated groups and the control group is compared to find differentially expressed genes (control points). The integration of the metabolic data could potentially complement the affects due to the post-translational modification and protein-protein interactions that would otherwise not be captured by the differential gene expression alone.

1.4 Why These Two Applications?

The reason to use these two applications to study the question in the title is that they could potentialy benefit significantly from multi-clause learning. First, the background knowledge is highly incomplete, since none of the reaction states are known beforehand in the two applications. Second, the explanations for each example inevitably involve multiple reaction states, which will be explained later

[2] There are billions of candidate hypotheses, which exceeds the capacity of a Binary Decision Diagram (BDD), thus the algorithm in [7] is practically inapplicable here.

in section 3. The same applications were also used in [12] to study how varying the background knowledge affects the accuracy, but the modelling has been extended by better way of using transcriptomic data.

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

2.1 Examples

The aim in both applications is to hypothesise the change in reaction states, which reflects the genetic control of a reaction. Although reaction states are not observable, they affect the flux through reactions, which leads to the change in metabolic abundance. Therefore, we can hypothesise the changes in reactions states through the changes in metabolic abundances that are observable. Accordingly, changes in metabolic abundance are used as examples E for learning. By comparing the treated group to the control group, three possible changes (i.e. *up*, *down* and *no-change*) in metabolic abundance can be observed. In the tomato application, the treated groups are obtained by knocking out specific genes related to the tomato ripening process, which results in 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.

2.2 Hypothesis Space

The hypotheses are ground facts about reaction states. A reaction state can be substrate limiting or enzyme limiting. Substrate limiting means the flux through a reaction is determined by the abundance of its substrates; while enzyme limiting implies that the flux through a reaction is controlled by the activity of its catalysing enzymes. Depending on the activity of catalysing enzymes, enzyme limiting can be further divided into three states: *catalytically increased*, *catalytically decreased* and *catalytically no-change*. These three states refer to the relative changes in the treated group against the control group, therefore they are not exactly the same as being activated or inhibited. For example, a relatively decreased reaction state doe not necessarily mean inhibited.

An enzyme limiting reaction is assumed to be under genetic regulation, while a substrate limiting reaction is not, and its flux is affected by the nearby enzyme limiting reactions. Therefore, a hypothesis h_e about enzyme limiting contains more information than a hypothesis h_s about substrate limiting. Thus the description length for different hypotheses is different. Specifically, if h_s is encoded by L bit, then $k * L$ bits are required for h_e, where $k > 1$. Considering each metabolite's abundance is controlled by one regulatory reaction, each example is also encoded by L bits to make compression achievable. The difference in the description length can also be explained using the frequency in information theory. There are much smaller number of reactions regulated by genes directly than indirectly. To achieve minimum description length, the more frequent h_s is encoded using shorter description length than h_e.

2.3 Background Knowledge

Regulation Rules Fig. 3 lists the eight regulation rules suggested by biologiests. These rules tell how changes in reaction states affect metabolic abundances. For example, if a reaction is catalytically increased, which means the flux through that reaction increases, then the concentration of its product goes up, while its substrate's concentration goes down because of the quicker consumption. These are encoded as b_1 and b_2 in Fig. 3. The rules b_1 to b_6 are all about enzyme limiting, and they are non-recursive, because the change in the substrate concentration will not affect the flux through the reaction but the enzyme activity itself. In contrast, the rules about substrate limiting (b_7 and b_8) are both recursive, because the substrate concentration would determine the flux through the reaction therefore affect the abundance of the product. These recursive rules essentially model the *indirect* effect of gene regulation.

These regulation rules seem to consider only one aspect, either enzyme limiting or substrate limiting, while in reality, both substrate abundances and enzyme activities may act together. However, it is unnecessary to consider the rules about the cumulative effect in our models, because the aim is to identify the dominated effect that is controlling the flux through a reaction, rather than knowing exactly what happen for each reaction. Similarly, as a node in a well-connected network, a metabolite's concentration is not just affected by one reaction's flux, but all reactions that consume or produce it. It seems the regulation rules should also capture this and consider how the fluxes from different reactions are balanced. However, no matter how fluxes from different branches are balanced, there is one branch whose effect is dominated and leads to the final observed change. Therefore, the rules in Fig. 3 are sufficient to our models.

b_1: $concentration(Metabolite, up, Time)$ ← produced_by(Metabolite,Reaction),
reaction_state(Reaction,enzymeLimiting,cataIncreased,Time).
b_2: $concentration(Metabolite, down, Time)$ ← consumed_by(Metabolite1,Reaction),
reaction_state(Reaction,enzymeLimiting,cataIncreased,Time).
b_3: $concentration(Metabolite, down, Time)$ ← produced_by(Metabolite1,Reaction),
reaction_state(Reaction,enzymeLimiting,cataDecreased,Time).
b_4: $concentration(Metabolite, up, Time)$ ← consumed_by(Metabolite1,Reaction),
reaction_state(Reaction,enzymeLimiting,cataDecreased,Time).
b_5: $concentration(Metabolite, no\_change, Time)$ ← produced_by(Metabolite1,Reaction),
reaction_state(Reaction,enzymeLimiting,cataNoChange,Time).
b_6: $concentration(Metabolite, no\_change, Time)$ ← consumed_by(Metabolite1,Reaction),
reaction_state(Reaction,enzymeLimiting,cataNoChange,Time).
b_7: $concentration(Metabolite1, up, Time)$ ←
produced_by(Metabolite1,Reaction), reaction_state(Reaction, substrateLimiting,_,Time),
consumed_by(Metabolite2,Reaction), concentration(Metabolite2,up,Time).
b_8: $concentration(Metabolite1, down, Time)$ ←
produced_by(Metabolite1,Reaction), reaction_state(Reaction, substrateLimiting,_,Time),
consumed_by(Metabolite2,Reaction), concentration(Metabolite2,down,Time).

Fig. 3: Regulation Rules

Metabolic Networks For tomato application, it is derived from the LycoCyc database [8], which contains 1841 reactions, 1840 metabolites and 8726 enzymes. For the predictive toxicology application, it is obtained from the rno KEGG database [14], which consists of 2334 reactions, 1366 metabolites and 1397 enzymes. In both applications, each reaction is considered as reversible. Therefore, the actual number of reactions N_r are doubled in the models. Since a subset of

these reactions' states have to be hypothesised in order to explain the observed changes, the size of hypothesis spaces for the two applications are 2^{4N_r}, where the number 4 corresponds to the four possible reaction states.

Transcript Profiles Transcript profiles represent expression data for the genes encoding the enzymes. However, gene expression alone is not always indicative of the reaction states. This is due to the other cellular processes, such as post-translational modification that could change the activity of the enzyme. Therefore, instead of using transcription profiles as training examples, they were used as an integrity constraint in our model to filter hypotheses. Any hypotheses about enzyme limiting have to be consistent with their gene expression data. Specifically, if a reaction state is hypothesised to be catalytically increased, its expression data, if available, should be increased and vise-versa. For example, without considering gene expression data, the four hypotheses shown in Fig. 4 are all candidates. However, the hypotheses (b) and (c) have inconsistent reaction states (arrow color) with the change in the expression (colored squares), hence these two hypotheses will be filtered after applying the integrity constraint about gene expression.

Integrity Constraint Apart from the integrity constraint about gene expression, there is another constraint about reaction states: a reaction can not be in different states at the same time. Please note that, there is no constraint that a metabolite's concentration cannot be both up and down at the same time. Because as explained earlier, the model is about the dominated branch that leads to the final observation, while it is possible that different branches to the same metabolite have different contributions of fluxes.

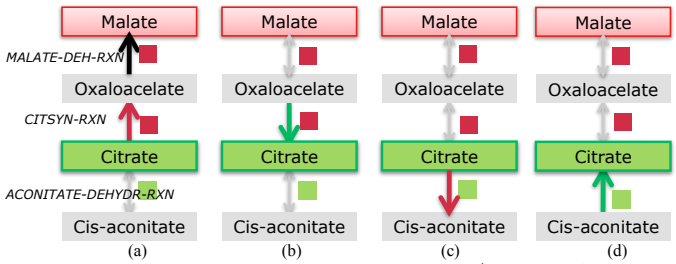


Fig. 4: Candidate Hypotheses for the decreased Citrate (Tomato Application). A reaction arrow is grey if it is not hypothesised, otherwise, it is coloured red, green or black to represent catalytically increased, decreased and substrate limiting reaction states, respectively. A metabolite is also in grey if not measured, otherwise, it is coloured red, green and blue to represent up, down and no-change, respectively. Gene expression levels are represented by the small squares beside the reaction arrows, and are applied the same colour scheme as that of metabolites.

3 Multi-clause Learning *vs* Single-clause Learning

The term 'single-clause learning'(SCL) comes from the entailment-incompleteness of Progol. As first pointed out by Yamamoto [21], the inverse entailment operator in Progol can only derive hypotheses that subsume an example e relative to

B in Plotkin's sense. This entailment-incompleteness restricts its derivable hypothesis to be a single clause, and that clause is used only once in the refutation proof of the example e. Thus we define SCL and MCL as follows. More details about MCL and SCL can be found in [13].

Definition 1. *Let c_i be a clause, which is either from background knowledge B or hypothesis H. Suppose $R = \langle c_1, c_2, ..., c_n \rangle$ is a refutation sequence that explains a positive example e. Let N be the number of c_i in R that is from H. It is single-clause learning (SCL) if $N = 1$; while it is multi-clause learning (MCL) if $N \geq 1$.*

3.1 Examples of MCL

An example of learning odd-numbers was used by Yamamoto [21] to demonstrate Progol's entailment-incompleteness. This example involves mutual recursion, so that the target clause h needs be applied several times in a refutation proof for an example $odd(s(s(s(0))))$. According to the definition above, this learning task is MCL, even though there is only one target clause to be learned. Progol's entailment-incompleteness is not only to do with mutual recursion, but also related to the issues of incomplete background knowledge. When B is incomplete, the missing clauses in B need to be hypothesised together with the clause about the observable predicate. In this case, a clause alone will not be able to explain an example, and have to rely on hypothesising other clauses in B in order to complete a refutation-proof of that example.

3.2 MCL $\neq$ Global Optimisation

The term 'learning multiple clauses'(LMC) is used in the description of a global-optimisation approach, in which multiple clauses that compressed from the *whole* set of examples are refined together, as opposed to a local-optimisation approach like the covering algorithm, where clauses compressed from a *subset* of examples are added to the final H iteratively. However, learning multiple clauses (LMC) referred in the global-optimisation approach and the mutli-clause learning (MCL) defined in this paper are related to different issues: hypothesis selection and hypothesis generation. LMC is related to the issue of selecting hypotheses globally, rather than locally. The hypotheses from which it selects can be derived either from MCL or SCL. Even if a learning algorithm's hypothesis space consists of single clauses derived by SCL, its final hypothesis may still have multiple clauses, which are aggregated from single clauses that generalised from different examples. In contrast, MCL is to do with generalising an example to multiple clauses, rather than a single clause. It can be combined with a selection method that is either global or local. Specifically, after deriving all candidate hypotheses using a MCL method, the covering algorithm is still applicable to greedily choosing a hypothesis which is a locally most compressed.

3.3 Difference in Hypothesis Space

The hypothesis space of SCL is a subset of that of MCL. Apart from single-clause hypotheses, it also considers hypothesising multiple clause together to handle the incomplete background knowledge. Since the clauses within a multi-clause hypothesis depend on each other, the increase of search space is dramatic.

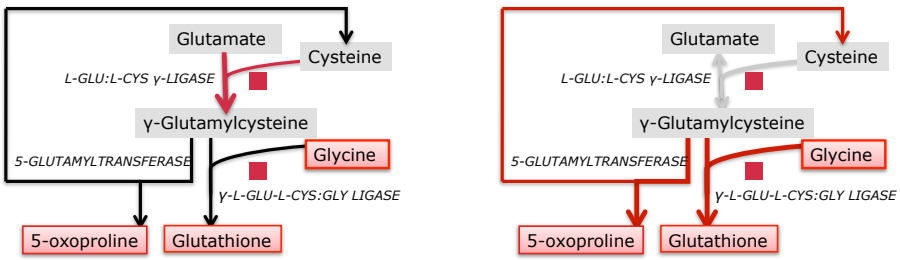
The upper bounds on the hypothesis spaces of SCL and MCL are $O(2^N)$ vesus $O(2^{2^N})$, where N is the number of distinct atoms derivable from a hypothesis language. As discussed in the introduction, the Blumer Bound of MCL is $\epsilon \geq \frac{1}{m}(2^N ln2 + ln\frac{1}{\delta})$, where the error bound grows exponentially with increasing N, thus is not PAC-learnable. Although MCL's error bound becomes polynomial when N is fixed, its hypothesis space is still much larger than that of SCL. Therefore, strong declarative biases are particularly important for MCL to constraint the hypothesis space. For example, in the two applications, the hypotheses space is constraint to the reactions whose distance to the observed metabolite is within six reactions, according to the small-world assumption in biology[19] .

3.4 MCL and SCL in the Context of the Two Applications

This subsection will use specific examples from the two applications to exemplify what has been discussed so far in this section. The two figures in Fig. 5 are from the predictive toxicology application. They show two possible explanations for the increase in the abundances of glutathione and 5-oxoproline. Fig. 5(a) says it is the reaction 'L-GLU:L-CYS γ-LIGASE' that is catalytically increased, which *indirectly* leads to the increase of glutathione and 5-oxoproline. In contrast, Fig. 5(b) suggests, it is the catalytical increase of the reaction 'γ-L-GLU-L-CYS:GLY LIGASE' that *directly* leads to the increased glutathione, while it is a different reaction '5-GLUTAMYLTRANSFERASE' whose catalytical increase results in the increased 5-oxoproline.

The explanation depicted in Fig. 5(a) can be encoded by a logic program as $H_1 = \{h_1, h_2, h_3\}$, where h_i is in Fig. 6(a)). Similarly, $H_2 = \{h_4, h_5\}$ encodes the explanation in Fig. 5(b). H_1 comes from two multi-clause hypotheses: $H_{mc1} = \{h_1, h_3\}$ and $H_{mc2} = \{h_1, h_2\}$, which are generalised from e_1 and e_2, respectively. While H_2 are aggregated from two single-clause hypotheses: $H_{sc1} = \{h_5\}$ and $H_{sc2} = \{h_4\}$, which are also generalised from e_1 and e_2, respectively. Although H_2 does consist of two clauses, each of them is derived independently from different examples, and each alone is sufficient to explain an example. In contrast, none of the clauses in H_{mc_i} is able to explain any examples without other clauses.

In the context of the two applications, single-clause learning means hypothesising a single reaction state for an example. This limitation restricts its derivable explanations to the reactions that directly connect to the observed metabolites, e.g. the two reactions coloured in red in Fig. 5(b). In contrast, a multi-clause



(a) Multi-clause hypotheses H_1 (b) Single-clause hypotheses H_2

Fig. 5: Explanations for the increase of Glutathione and 5-oxoproline

h_1: reaction_state('γ-L-GLU-L-CYS:GLY LIGASE', substrateLimiting, _ , day14).
h_2: reaction_state('5-GLUTAMYLTRANSFERASE', substrateLimiting, _ , day14).
h_3: reaction_state('L-GLU:L-CYS γ-LIGASE', enzymeLimiting, cataIncreased, day14).
h_4: reaction_state('5-GLUTAMYLTRANSFERASE', enzymeLimiting, cataIncreased, day14).
h_5: reaction_state('L-GLU:L-CYS γ-LIGASE', enzymeLimiting, cataIncreased, 'dat14').

(a) Predictive Toxicology Application

h_6: reaction_state('CITSYN-RXN', enzymeLimiting, cataIncreased, 'NOR_Late').
h_7: reaction_state('MALATE-DEH-RXN', substrateLimiting, _ , 'NOR_Late').
h_8: reaction_state('ACONITATE-DEHYDR-RXN', enzymeLimiting, cataDecreased, 'NOR_Late').

(b) Tomato Application

Fig. 6: Candidate Hypothesis Clauses

learner is able to explore any possible regulatory reactions that are several reactions away from the observed metabolites. For example, the reaction coloured in red in Fig. 5(a) is not directly connected to glutathione and 5-oxoproline. However, the regulatory effect of this reaction is passed through γ-glutamylcysteine, which is a common substrate of the two substrate limiting reactions, 'γ-L-GLU-L-CYS:GLY LIGASE' and '5-GLUTAMYLTRANSFERASE, producing the two observed metabolites. The multi-clause explanation (Fig. 5(a)) is also hypothesised by the biologists [20], while it is not derivable by SCL.

As mentioned earlier in the introduction, it is possible to construct a multi-clause hypothesis by sequentially adding single-clauses. The hypothesis H_{4a} drawn in Fig. 6(b)(a) gives such an example. H_{4a} consists of two clauses h_6 and h_7, which are in Fig. **??**. The single clause h_6 can be derived from the example of decreased Citrate. After h_6 is added to the background knowledge, another clause h_7 can be derived from the example of increased Malate. Despite the fact that H_{4a} can be sequentially constructed using Progol5, Progol5 does not necessarily suggest this hypothesis, but instead hypothesise H_{4d}=$\{h_8\}$ shown in Fig. 4(d). Whether a MCL problem can be reduced to a SCL problem depends on the degree of incompleteness in the background knowledge and the distributions of given examples. For the two applications studied in this paper, imagine an extreme case where all metabolites' abundances are observable, then we can simply apply SCL to reconstruct each reaction state. However, not all metabolites' abundances are practically measurable due to technological limitations.

A multi-clause learner is likely to find a hypothesis with higher compression than a single-clause learner because of a more complete hypothesis space. According to the description length defined in the previous section, H_1 shown in Fig. 5(a) is more compressed than H_2 shown in Fig. 5(b). Intuitively, H_1 suggests a single control point for two observed metabolites, while H_2 involves two control points for the same number of observations.

4 Experiments

Two independent experiments were conducted to empirically investigate the null hypothesis: MCL does not have higher predictive accuracies than SCL for any real-world datasets.

4.1 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 datasets in total (3 stages*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 [20]. All the materials used in the experiments can be found at http://ilp.doc.ic.ac.uk/mcTopLog.

4.2 Methods

Progol5 [11] and MC-TopLog [13] were used to represent SCL and MCL, 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 that "a reaction state is substrate limiting if it is not hypothesised" was applied during the testing phase. The running time for Progol5 and MC-TopLog is not comparable, because Progol5 has been implemented in C, while MC-TopLog uses Prolog and was executed using YAP. Since YAP is optimised towards efficiency, it is much faster, thus MC-TopLog's running time is even shorter than Progol5 and each run takes less then 5 mins. Therefore, it is the number of generated hypotheses that is compared.

4.3 Predictive Accuracies

The predictive accuracies of Progol5 and MC-TopLog are given below. In the tomato application there are two datasets i.e. NOR_MID and NOR_LATE in which MC-TopLog's accuracies are significantly higher than that of Progol5 at the 95% confidence level. There is even one dataset CNR_EARLY that Progol5 has lower error. In the predictive toxicology application, there also exist one dataset, i.e. day 14, where MC-TopLog has a significantly higher accuracy than Progol5. Overall our null hypothesis is rejected by the accuracy results: There is at least one dataset in both applications where MCL has significantly lower error than SCL.

Timepoint	**default(no change)**,%	**Progol**,%	**MC-TopLog**,%	p-value
CNR_Early	63.64	86.36±7.32	81.82±8.22	0.576
CNR_Mid	36.36	86.36±7.32	86.36±7.32	1.000
CNR_Late	40.90	90.91±6.13	90.91±6.13	1.000
NOR_Early	86.36	86.36±7.32	86.36±7.32	1.000
NOR_Mid	50.00	68.18±9.93	86.86±7.32	0.042
NOR_Late	31.82	68.18±9.93	86.36±7.32	0.042
RIN_Early	100.00	100±0.00	100±0.00	1.000
RIN_Mid	90.91	90.91±6.13	90.91±6.13	1.000
RIN_Late	36.36	77.27±8.93	77.27±8.93	1.000

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 Predictive Toxicology Application

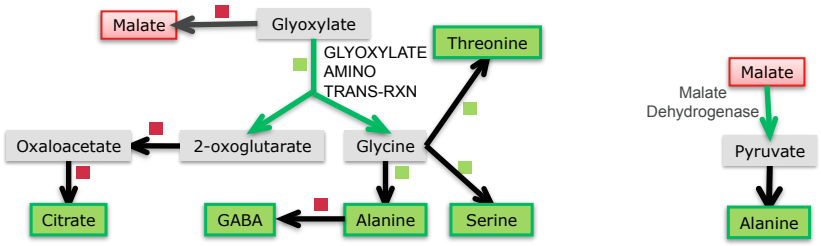
4.4 Hypothesis Interpretation

The differences between the hypotheses suggested by Progol5 and MC-TopLog are explained in this subsection using a concrete example. The dataset used here is CNR_Late from the tomato application. As shown in Fig 8(a), there is only one ground fact with enzyme limiting, which means only one control point is hypothesised to explain six metabolites' concentration changes. In contrast, all the ground facts suggested by Progol5 are enzyme limiting, which means hypothesising six control points to explain the same set of observations.

Biological Significance Fig. 7(a) visualises the hypothesis suggested by MC-TopLog. It is the reaction catalysed by glyoxylate amino transferase that is suggested to be the control point for three organic acids (Citrate, Malate, GABA) and three amino acids (Alanine, Serine and Threonine). This hypothesis is particularly interesting to biologists. First, biologists used to believe that the abundance of organic acids is controlled via TCA-Cycle [5], while this hypothesis indicates that the flux through the Malate can also be regulated by Glyoxylate shunt, independently of TCA cycle. Second, this hypothesis involves three intricately connected pathways (TCA-Cycle, Glyoxylate Shunt and GABA Shunt pathway), which is difficult for a human being to come up with. Different from the multi-clause hypothesis depicted in Fig. 5(a) which is relatively simple and has been confirmed by biologists [20], no previous study is available to confirm the one in Fig. 7(a), thus new biological experiments will be designed to test this hypothesis. Thirdly, this hypothesis could be of industrial interest as the higher organic acid content in particular Malate is a commercially important quality trait for tomatoes [3], therefore this plausible hypothesis is subjected to the experimental investigations.

Why Different Accuracy? Compared to Progol5's hypothesis, where there are six control points, the one suggested by MC-TopLog involves a single control point co-regulating six metabolites. It is not just much simpler, but also potentially has higher predictive accuracy. During the leave-one-out cross validation, when one of co-regulated metabolites (e.g. Alanine) is left out as test data, a hypothesis with the same control point (e.g. glyoxylate amino transferase) can be reconstructed using the remaining co-regulated metabolites in the training data. With the closed world assumption, the hypothesis derived from the training data will explain the test data since it is co-regulated with other metabolites in the training data. In contrast, a control point suggested by Progol5 only regulates a single metabolite in the training set, therefore is less likely to be co-regulated with the test data, thus may not be able to explain that test data.

However, that is not always the case, as shown by the accuracy results. It turns out there exist good approximations of such module of co-regulated metabolites in the hypothesis space of Progol5. That is why MC-TopLog did not show higher accuracies in those cases. Fig 7(b) shows such a good approximation, where a pair of metabolites are suggested to be co-regulated by Malate Dehydrogenase. Although the number of co-regulated metabolites is not as large as the one in Fig. 7(a), it manages to predict the decrease of Alanine when it is



(a) (b)

Fig. 7: (a) Three organic acids (Citrate, Malate, GABA) and three amino acids (Alanine, Serine and Threonine) are hypothesised to be controlled by the reaction catalysed by Glyoxylate Amino Transferase. The decrease in the flux through this reaction would increase the abundance of the reactants (Glyoxylate) and would decrease the abundance of the products (Glycine and 2-oxoglutarate). This would subsequently affect the flux through the Glyoxylate shunt and GABA shunt pathways and a part of the TCA cycle involved with the synthesis of organic acids. Specifically, decrease in the flux would lead to the accumulation of glyoxylate and a reversed flux to Malate via malate synthase (2.3.3.9) reaction would lead to an accumulation of Malate. On the other hand, glycine's production would be hampered and is reflected in the decreased abundance of the three amino acids that are being synthesized by glycine in different condensation reactions. (b) Malate and Alanine are hypothesised to be controlled by the reaction catalysed by Malate Dehydrogenase.

left-out as test-data. There are other similar small co-regulated modules in Progol's hypothesis, so that they together approximate the large module (Fig. 7(a)) suggested by MC-TopLog. The hypothesis with small co-regulated modules is not disprovable according to biologists' existing knowledge, and it does not have existing studies to support either, therefore further experimental test will be carried out. There is no existing evidence that a control point that regulate more metabolites is always better. It is just Occam's razor that make biologists prefer simpler hypothesis with smaller number of control points. Therefore, it is difficult to conclude that a multi-clause hypothesis (a large co-regulated module) is better than a single-clause hypothesis (a small co-regulated module), even though the previous is more compressive.

4.5 Compression and $|H|$

The following table 3 shows that MC-TopLog always derives hypothesis with higher compression, while the trade-of is a much larger hypothesis space, as can be seen from their search nodes. For the datasets NOR_EARLY and RIN_EARLY, the difference in search nodes is relatively small. That is because at early time points, there are few changes in metabolic abundances, while the rule about *no_change* is non-recursive.

5 Conclusions and Future Work

Applying ILP technique to these two real-world problems not only improve the efficiency of the whole studies, but also suggest interesting hypotheses that are

```
reaction_state(reversed-'GLYCINE-AMINOTRANSFERASE-RXN',enzymeLimiting,cataDecreased,'CNR_L').
reaction_state('MALSYN-RXN',substrateLimiting,sUp,'CNR_L').
reaction_state(reversed-'ALANINE–GLYOXYLATE-AMINOTRANSFERASE-RXN',substrateLimiting,sDown,'CNR_L').
reaction_state(reversed-'GLYOHMETRANS-RXN',substrateLimiting,sDown,'CNR_L').
reaction_state(reversed-'THREONINE-ALDOLASE-RXN',substrateLimiting,sDown,'CNR_L').
reaction_state('GABATRANSAM-RXN',substrateLimiting,sDown,'CNR_L').
reaction_state(reversed-'RXN-6902',substrateLimiting,sDown,'CNR_L').
```

(a) MC-TopLog's Hypothesis

```
reaction_state('2.6.1.18-RXN',enzymeLimiting,cataIncreased,'CNR_L').
reaction_state(reversed-'5.1.1.18-RXN',enzymeLimiting,cataDecreased,'CNR_L').
reaction_state('THREDEHYD-RXN',enzymeLimiting,cataIncreased,'CNR_L').
reaction_state(reversed-'ACONITATEDEHYDR-RXN',enzymeLimiting,cataDecreased,'CNR_L').
reaction_state('GABATRANSAM-RXN',enzymeLimiting,cataIncreased,'CNR_L').
reaction_state('1.1.1.39-RXN',enzymeLimiting,cataDecreased,'CNR_L').
```

(b) Progol's Hypothesis

Fig. 8: Comparing Induced Hypotheses

Timepoint	**Compression**		**Search Nodes**	
	Progol	**MC-TopLog**	**Progol**	**MC-TopLog**
CNR_Early	10	845	352	1240
CNR_Mid	12265	315095	322	11890
CNR_Late	8358	150342	318	3654
NOR_Early	5382	64	352	411
NOR_Mid	12555	117592	354	11890
NOR_Late	11362	129813	312	14032
RIN_Early	2940	40	312	350
RIN_Mid	2552	404	793	10851
RIN_Late	11475	132320	354	14584

Table 3: Comparing Search nodes and Compression

different from what biologists used to think. Those plausible hypotheses without support from existing studies will be test by biological experiments in future studies.

As shown by experiments, there do exist datasets, in which hypotheses derived by MCL have significantly higher predictive accuracy than SCL. On the other hand, for most of the datasets in the two applications, there are good approximation of the target within the hypothesis space of a single-clause learner, so that MCL does not necessarily have higher accuracy than SCL.

MCL has a much larger hypothesis space than SCL, and it tends to find a more compressive hypothesis. However, this does not mean it will gain higher accuracy.

Acknowledgements

The authors would like to acknowledge the support of Syngenta in its fundig of the University innovations Centre at Imperial College. The fourth author would like to thank the Royal Academy of Engineering and Microsoft for funding his present 5 year Research Chair.

References

1. Syngenta Ltd. http://www.syngenta.com/en/index.html.
2. A. Blumer, A. Ehrenfeucht, D. Haussler, and M.K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, 1989.
3. D.C. Centeno and S.Osorio et al. Malate plays a crucial role in starch metabolism, ripening, and soluble solid content of tomato fruit and affects postharvest softening. *Plant Cell*, 23:162–184, 2011.

4. D. Corapi, A. Russo, and E. Lupu. Inductive logic programming as abductive search. In *ICLP2010 Technical Communications*, Berlin, 2010. Springer-Verlag.
5. A.R. Fernie, F. Carrari, and L.J. Sweetlove. Respiratory metabolism: glycolysis, the tca cycle and mitochondrial electron transport. *Current Opinion in Plant Biology*, 7:254–261, 2004.
6. K Inoue. Induction as consequence finding. *Machine Learning*, 55:109–135, 2004.
7. K. Inoue, T. Sato, M. Ishihata, Y. Kameya, and H. Nabeshima. Evaluating abductive hypotheses using an EM algorithm on BDDs. In *IJCAI-09*, pages 810–815. AAAI Press, 2009.
8. LycoCyc. Solanum lycopersicum database. http://solcyc.solgenomics.net//LYCO/.
9. A. Markitanis, D. Corapi, A. Russo, and E. Lupu. Learning user behaviours in real mobile domains. Submitted to ILP2011.
10. S.H. Muggleton. Inverse entailment and Progol. *New Generation Computing*, 13:245–286, 1995.
11. S.H. Muggleton and C.H. Bryant. Theory completion using inverse entailment. In *ILP-00*, pages 130–146. Springer-Verlag, 2000.
12. 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. ILP2010.
13. S.H. Muggleton, D. Lin, and A. Tamaddoni-Nezhad. MC-TopLog: complete multi-clause learning guided by a top theory. 2011. Submitted to ILP2011.
14. 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.
15. Oliver Ray. Nonmonotonic abductive inductive learning. *Journal of Applied Logic*, 7(3):329 – 340, 2009.
16. Oliver Ray, Ken Whelan, and Ross King. Automatic revision of metabolic networks through logical analysis of experimental data. In *ILP2009*, pages 194–201, 2009.
17. A. Tamaddoni-Nezhad, D. Bohan, A. Raybould, and S.H. Muggleton. Machine learning a probabilistic network of ecological interactions. Accepted to ILP2011.
18. 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.
19. Andreas Wagner and David A. Fell. The small world inside large metabolic networks. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268(1478):1803–1810, September 2001.
20. 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.
21. A. Yamamoto. Which hypotheses can be found with inverse entailment? In N. Lavrač and S. Džeroski, editors, *ILP97*, pages 296–308. Springer-Verlag, 1997.
22. 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.