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. This limitation is referred as single-clause learning (SCL) in this paper. However, according to the Blumer bound, incomplete learners such as Progol can have higher predictive accuracy while use 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 or its approximation is within the hypothesis space of the incomplete learner. This paper uses two real-world applications in systems biology to study whether it is necessary to have complete multi-clause learning (MCL) methods, which is computationally expensive but capable of deriving multi-clause hypotheses that is in the systems level. The experimental results show that in both applications there do exist datasets, in which MCL has significantly higher predictive accuracies than SCL. On the other hand, MCL does not outperform SCL all the time due to the existence of the target hypothesis or its approximations within the hypothesis space of SCL.

1 Introduction

Progol's inverse entailment [10] is incomplete, as first pointed out by Yamamoto[19]: Progol can only generalise a single example to a single clause, but not multiple clauses. This type of entailment-incompleteness can be characterised as single-clause learning (SCL). In contrast, entailment-complete methods are referred to 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 when increasing $|H|$ you also increase the bound on the size of required training set. 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

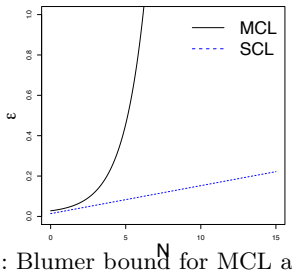


Fig. 1: Blumer bound for MCL and SCL

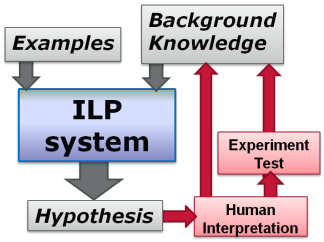


Fig. 2: Learning Cycle in ILP

error bounds than complete ones. Specifically, the error bounds for SCL and MCL are as given in equations (1) and (2), where N is the number of distinct atoms derivable from a hypothesis language. As shown in Fig. 1, MCL's error bound grows exponentially with the increasing N; while SCL's error bound is linear with respect to N. In terms of running time, MCL takes much longer than SCL due to a much larger $|H|$. Overall, the Blumer bound indicates that in the case that the target theory or it is approximations are within both hypothesis spaces, MCL is worse than SCL in terms of both running time and predictive error bounds for a randomly chosen training set.

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

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

However, the Blumer bound only holds if the target theory or its approximation is within the hypothesis space for both algorithms. In the case that both the target theory and its approximation are within the hypothesis space of the complete learner but not within the hypothesis space of the incomplete learner, 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 learning. However, this is not the case for a real-world dataset, so one of the motivations for this paper is to see whether completeness in learning does lead to higher accuracy in at least one real-world dataset.

1.2 Experimental Comparisons between SCL and MCL

Within ILP much effort has been put into designing methods that are complete for hypothesis finding. For example, ILP systems CF-Induction [7], XHAIL [15], TAL [4] and MC-TopLog [13] were designed to overcome Progol's entailment-incompleteness. Indeed, different from SCL that restricts its hypothesis spaces to single-clause hypotheses, MCL is able to suggest multi-clause hypotheses, which are more compressive. The difference between single-clause and multi-clause hypotheses can be analogous to that between reductionist and systems hypotheses, which is explained later in Section 3.4. However, it is unclear whether systems hypotheses are definitely better than reductionist hypotheses, especially in real-world applications, while no direct comparison has been done before using real-world datasets[1]. At the same time, Progol's entailment-incompleteness does

[1] Although [8] has compared CF-Induction to Progol, no predictive accuracies are provided, but only learned hypotheses ranked by a probability measure. Although

not stop it from 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 [16]. Therefore, another motivation for this paper is to use direct comparisons on the same datasets to demonstrate the necessity of having MCL, which is much more computationally expensive than SCL. We also analyse the cases when MCL does or does not improve the learning results of SCL.

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 shelf life, flavour and nutritional quality is a major part of Syngenta's breed selection and seed development programme. The aim of applying an ILP approach in this programme 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 liver tumour promoter (phenobarbital) in the rat. In both applications, the predictive models derived would potentially influence the experimental design, 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 disprove 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

Progol's hypothesis is only ranked at 13th, it does not mean it has lower predictive accuracy than the one ranked at the top.

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. ILP techniques make such a 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 a reaction state that is not observable and could be simply a ground fact. 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 [8], 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 [17]. 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 take into account the effects 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 we chose these two applications to study the question in the title is that they could potentially benefit significantly from multi-clause learning. Firstly, the background knowledge is highly incomplete, since none of the reaction states are known beforehand in the two applications. Secondly, the explanations for each example inevitably involve multiple reaction states, which will be explained later 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 the more effective usage of transcriptomic data.

[2] There are billions of candidate hypotheses, which exceeds the capacity of a Binary Decision Diagram (BDD), thus the algorithm in [8] is practically inapplicable here.

2 ILP Models

2.1 Examples

The aim in both applications is to hypothesise the changes in reaction states, which reflect the genetic control of reactions. Although reaction states are not observable, they affect the flux through reactions, which leads to changes 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 that 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 does 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 by the information theory as follows. There are fewer reactions regulated by genes directly than indirectly, therefore the more frequent h_s is encoded using shorter description length than h_e to achieve minimum description length.

2.3 Background Knowledge

Regulation Rules Fig. 3 lists the seven regulation rules suggested by biologists. 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 rule about substrate limiting (e.g. b_7) is 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 dominating effect that is controlling the flux through a reaction, rather than knowing exactly what happens 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 dominates and leads to the final observed change. Therefore, the rules in Fig. 3 are sufficient for 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(Metabolite,Reaction),
reaction_state(Reaction,enzymeLimiting,cataIncreased,Time).
b_3: $concentration(Metabolite, down, Time)$ ← produced_by(Metabolite,Reaction),
reaction_state(Reaction,enzymeLimiting,cataDecreased,Time).
b_4: $concentration(Metabolite, up, Time)$ ← consumed_by(Metabolite,Reaction),
reaction_state(Reaction,enzymeLimiting,cataDecreased,Time).
b_5: $concentration(Metabolite, no\_change, Time)$ ← produced_by(Metabolite,Reaction),
reaction_state(Reaction,enzymeLimiting,cataNoChange,Time).
b_6: $concentration(Metabolite, no\_change, Time)$ ← consumed_by(Metabolite,Reaction),
reaction_state(Reaction,enzymeLimiting,cataNoChange,Time).
b_7: $concentration(Metabolite1, Change, Time)$ ←
produced_by(Metabolite1,Reaction), reaction_state(Reaction, substrateLimiting,_,Time),
consumed_by(Metabolite2,Reaction), concentration(Metabolite2,Change,Time).

Fig. 3: Regulation Rules

Metabolic Networks For the tomato application, the metabolic network is derived from the LycoCyc database [9], which contains 1841 reactions, 1840 metabolites and 8726 enzymes. For the predictive toxicology application, it is obtained from the rat specific network in the 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 (i.e. *substrate limiting*, *catalytically increased*, *catalytically decreased* and *catalytically no-change*).

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.

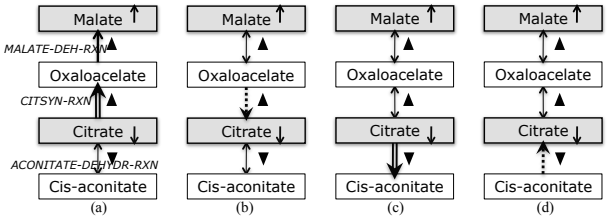


Fig. 4: Candidate Hypotheses for the decreased Citrate (Tomato Application). A reaction arrow is in double direction if its state is not hypothesised, otherwise it is not just in one direction, but also changed in the line style. The reaction states of substrate limiting, catalytically decreased and increased are respectively represented by thicker, dashed and double lines. Measured metabolites are highlighted in grey, and their corresponding values are annotated in their upper right corner. Gene expression levels are represented by the small triangles next to the reaction arrows. The upward and downward triangles mean increased and decreased.

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

The term 'single-clause learning'(SCL) comes from the entailment-incompleteness of Progol. As first pointed out by Yamamoto [19], 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 SCL and MCL 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 [19] to demonstrate Progol's entailment-incompleteness. This example involves mutual recursion, so that the target clause h needs to be applied several times in a refutation proof for the 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 issue of incomplete background knowledge, such as the two applications studied in this paper.

3.2 MCL $\neq$ Global Optimisation

The term 'learning multiple clauses'(LMC) is used to describe 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. LMC is related to the issue of selecting hypotheses globally rather than locally. The hypotheses from which it selects can be derived either by 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 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 locally most compressed.

3.3 Difference in Hypothesis Space

SCL's hypothesis space is a subset to that of MCL, and their difference is not insignificant. Specifically, the upper bound on the hypothesis space of SCL is $O(2^N)$, where N is the number of distinct atoms derivable from a hypothesis language. In contrast, it is $O(2^{2^N})$ for MCL, because it does not ignore the hypotheses with dependent clauses. Such a large hypothesis space makes MCL not PAC-learnable (Probably approximately correct learnable [18]). Because the number of examples m grows exponentially with increasing N, rather than polynomial as that in SCL, which can be seen by rewriting SCL and MCL's Blumer bounds as $m \geq \frac{1}{\epsilon}(Nln2+ln\frac{1}{\delta})$ and $m \geq \frac{1}{\epsilon}(2^N ln2+ln\frac{1}{\delta})$, respectively. Even when N is small and fixed for a particular learning problem, MCL's hypothesis space is still much larger than that of SCL. That is why it is particularly important for a multi-clause learner to bound its search space like that in MC-TopLog.

3.4 Reductionist vs. Systems Hypothesis

SCL can only generalise an example to a single clause, thus its hypotheses are in the style of 'H_1 causes O_1, ... H_n causes O_n', where O_i represents an observation and each H_i is not necessarily related to the others. This kind of hypotheses can be referred to as reductionist hypotheses. In contrast, MCL is able to generalise an example to mutiple clauses so that its hypotheses are rich enough to be in the systems-level, and they are in the style of 'H_1, H_2...H_j together cause O_1, O_2 ... O_i'. Table 1 summarises the differences between SCL and MCL.

Entailment-**Incomplete**	Entailment-**Complete**
Single clause per example	**Multiple clauses** per example
Constrained hypothesis space	**Less** constrained hypothesis space
Reductionist	**Systems**
H_1 causes O_1 ... H_n causes O_n	H_1, H_2...H_m together cause O_1, O_2 ... O_n

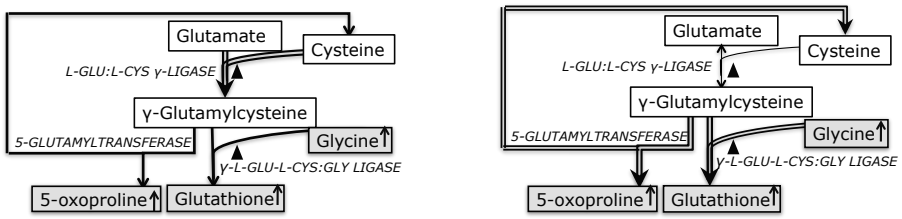
Table 1: Single-clause Learning vs. Multi-clause Learning

3.5 SCL and MCL in the Context of the Two Applications

This subsection uses 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, it is two different reactions whose activation that results in the increased glutathione and 5-oxoproline, as suggested by the two double line arrows in Fig. 5(b).

The explanation depicted in Fig. 5(a) can be encoded by a logic program $H_{mc} = \{h_1, h_2, h_3\}$, where h_i is in Fig. 6(a). Similarly, the explanation in Fig. 5(b) can be encoded as $H_{sc} = \{h_4, h_5\}$. Although both H_{mc} and H_{sc} consist of multiple clauses, H_{sc} is aggregated from two single-clause hypotheses: $H_{sc1} = \{h_5\}$ and $H_{sc2} = \{h_4\}$, which are respectively generalised from e_1 and e_2. In other words, each clause in H_{sc} is derived independently from different examples, and each alone is sufficient to explain an example. In contrast, H_{mc} comes from two multi-clause hypotheses: $H_{mc1} = \{h_1, h_3\}$ and $H_{mc2} = \{h_1, h_2\}$, which are also generalised from e_1 and e_2, respectively. However, none of the clauses in H_{mc} is able to explain any examples alone 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. For example, the two double-line arrows in Fig. 5(b) are connected directly to glutathione and 5-oxoproline, whose abundances are measurable. In contrast, a multi-clause learner is able to explore any possible regulatory reactions that are several reactions away from the observed metabolites. For example, the reaction arrow with double-line in Fig. 5(a) is not directly connected to either glutathione or 5-oxoproline. However, the regulatory effect of this reaction is passed through the metabolite γ-glutamylcysteine, which is a common substrate of the two substrate limiting reactions ('γ-L-GLU-L-CYS:GLY LIGASE' and '5-GLUTAMYLTRANSFERASE'). The hypothesis H_{mc} in Fig. 5(a) agrees with the one suggested by biologists [5], but it is not derivable by SCL.



(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

In terms of compression, H_{mc} is more compressive than H_{sc}, according to the description length defined in the previous section. Intuitively, H_{mc} is more compact since it suggests a single control point for two observed metabolites, while H_{sc} involves two control points for the same number of observations. On the other hand, H_{sc} is a reductionist hypothesis while H_{mc} is in the systems level. Because H_{sc} suggests that h_4 causes e_1 and h_5 causes e_2. In contrast, H_{mc} says it is the combination of h_1, h_2 and h_3 that leads to e_1 and e_2. The higher compression of H_{mc} can also be explained by the fact that it is a systems-level description, which is more compact than the non-systems one.

3.6 Reducing MCL to 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. 4(a) gives such an example. H_{4a} consists of two clauses h_6 and h_7, which are in Fig. 6(b). 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 suggests 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 metabolite abundances are observable, then we can simply apply SCL to reconstruct each reaction state. However, not all metabolite abundances are measurable due to technological limitations.

4 Experiments

The two null hypotheses to be tested are: (1) MCL does not have higher predictive accuracies than SCL for any real-world datasets; (2) MCL always has higher predictive accuracies than SCL for all 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 [5]. 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. Leave-one-out cross validation was used to compute the predictive accuracies. The closed world assumption applied during the testing phase was that "a reaction state is substrate limiting if it is not hypothesised". For the comparison of running time, we compared the number of search nodes instead. Because Progol5 and MC-TopLog's running time are not comparable. Specifically, Progol5 was implemented in C, while MC-TopLog used 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 despite of a much larger search space. For example, in the experiments, MC-TopLog takes maximum 10 mins for each run, while Progol 5 can take up to 3 hours.

4.3 Predictive Accuracies

As shown in the tables below, there are two datasets (i.e. 'NOR_Mid' and 'NOR_Late') in the tomato application and one dataset (i.e. 'Day 3') in the predictive toxicology application, where MC-TopLog's accuracies are significantly higher than that of Progol5 at the 95% confidence level (i.e. p-value$\leq$ 0.05). While for the rest of the datasets, the two systems have the same or similar accuracies. Therefore both our null hypotheses are rejected by the accuracy results: (1) there is at least one dataset in both applications where MCL has significantly higher accuracy than SCL; (2) MCL does not outperform SCL all the time in terms of predictive accuracies. The explanation for such results will be given later after seeing a concrete example of the hypotheses derived by the two systems.

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 2: Predictive accuracies with standard errors in Tomato Application

Timepoint	**default(no change)**,%	**Progol**,%	**MC-TopLog**,%	p-value
Day 1	55.77	63.46±6.68	73.08±6.15	0.058
Day 3	30.77	**44.23**±6.89	**59.62**±6.80	**0.010**
Day 7	40.38	53.85±6.91	59.62±6.80	0.182
Day 14	48.08	61.54±6.75	63.46 ±6.67	0.569

Table 3: Predictive accuracies with standard errors in Predictive Toxicology Application

4.4 Hypothesis Interpretation

This subsection exemplifies the different hypotheses suggested by Progol5 and MC-TopLog. The dataset used here is the abundances of six metabolites (Citrate, Malate, GABA, Alanine, Serine and Threonine) measured in the mutant 'CNR_Late' of the tomato application. MC-TopLog suggests a single control point to co-regulate the six metabolites. As can be seen in Fig. 7(a), there is only one ground fact with enzyme limiting, while the rest are about substrate limiting, which are also indispensable in explaining the six observations together

rs(reversed-'GLYCINE-AMINOTRANSFERASE-RXN',enzymeLimiting,cataDecreased,'CNR_L').
rs('MALSYN-RXN',substrateLimiting,_,'CNR_L').
rs(reversed-'ALANINE–GLYOXYLATE-AMINOTRANSFERASE-RXN',substrateLimiting,_,'CNR_L').
rs(reversed-'GLYOHMETRANS-RXN',substrateLimiting,_,'CNR_L').
rs(reversed-'THREONINE-ALDOLASE-RXN',substrateLimiting,_,'CNR_L').
rs('GABATRANSAM-RXN',substrateLimiting,_,'CNR_L').
rs(reversed-'RXN-6902',substrateLimiting,_,'CNR_L').

(a) MC-TopLog's Hypothesis

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

(b) Progol's Hypothesis

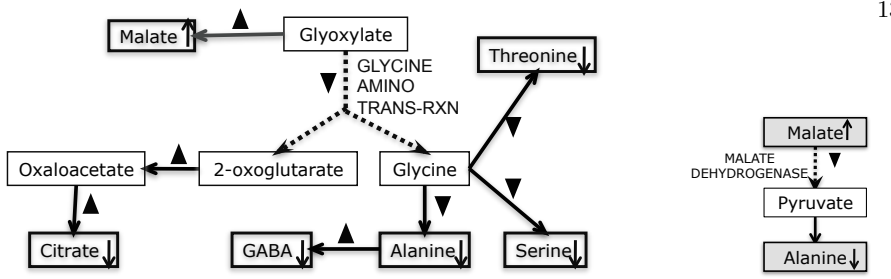
Fig. 7: Hypotheses Comparison. The predicate 'rs' is short for 'reaction_state'

with the suggested control point. For the same set of observations, Progol suggests a reductionist hypothesis with six control points, since it hypothesises one control point for each metabolite. As can be seen in Fig. 7(b), all the ground facts there are about enzyme limiting.

Biological Significance Fig. 8(a) visualises the hypothesis in Fig. 7(a) suggested by MC-TopLog. It is the reaction 'GLYCINE-AMINOTRANS-RXN' that is suggested to be the control point for the six observations. This hypothesis is particularly interesting to biologists. Firstly, it is suggested in [6] that the abundance of organic acids is controlled via TCA-Cycle, while this hypothesis indicates that the flux through the Malate can also be regulated by Glyoxylate shunt, independently of TCA cycle. Secondly, this hypothesis involves three intricately connected pathways (TCA-Cycle, Glyoxylate Shunt and GABA Shunt pathway), which is difficult for human beings to come up with. Different from the multi-clause hypothesis depicted in Fig. 5(a) which has been confirmed by biologists [5], no previous study is available to confirm the one in Fig. 8(a), thus new biological experiments will be designed to test this hypothesis. Thirdly, this hypothesis could be of industrial interest since higher organic acid content in particular Malate is a commercially important quality trait for tomatoes [3].

4.5 Explanations for the Accuracy Results

The higher predictive accuracies by MC-TopLog in the three datasets can be explained by the fact that in those datasets neither target hypotheses nor their approximations are within the hypothesis space of Progol. Although the target hypotheses are unknown for the two real-world applications, the hypotheses searched by Progol are less likely to be the targets. Because as mentioned before, Progol's hypotheses are not just reductionist, but also restricted to the reactions directly connected to the observed metabolites, so that they are usually specific to the example that they are generalised from. Such specific hypotheses may not be generalisable to the test data, thus they fail to predict the test data. In constrast, the multi-clause hypotheses suggested by MC-TopLog are not just in the systems-level, but also more compressive. For example, the multi-clause hypothesis in Fig. 8(a) generalises six examples. When any of the six examples are left-out as test data, they can always be predicted by the hypothesis generalised from the remaining five examples. That is why MC-TopLog achieves higher accuracy for the three datasets.



(a) (b)

Fig. 8: (a) Three organic acids (Citrate, Malate, GABA) and three amino acids (Alanine, Serine and Threonine) are hypothesised to be controlled by the reaction 'GLYCINE-AMINOTRANS-RXN'. The decrease in the flux through this reaction (represented by the dashed line) 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 the 'Malate Synthase' 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 suggested to be controlled by the reaction catalysed by malate dehydrogenase.

On the other hand, it turns out that the systems hypotheses suggested by MC-TopLog does not always have higher predictive accuracies than the reductionist hypotheses suggested by Progol. That is because there do exist good approximations to the targets within the hypothesis space of Progol. Fig 8(b) shows such a good approximation, where a pair of metabolites are suggested to be co-regulated by Malate Dehydrogenase. This systems hypothesis is essentially derived by aggregating two reductionist hypotheses. Specifically, in Fig 8(b), the dash line denoting catalytically decrease is hypothesised from the increased Malate, while the solid line representing substrate limiting is derived from the decreased Alanine. Although the number of co-regulated metabolites in Fig 8(b) is not as large as the one in Fig. 8(a), it manages to predict one of the co-regulated metabolites when it is left-out as test data. There are other similar small co-regulated modules in Progol's hypothesis space, so that they together approximate the large module (Fig. 8(a)) suggested by MC-TopLog. That is why in the dataset like 'CNR_Late' MC-TopLog does not outperform Progol5. In fact, the hypotheses with small co-regulated modules are not disprovable by the existing knowledge. Additionally, there is no evidence that a control point regulating more metabolites is definitely better. Nevertheless, biologists tend to follow Occam's razor and prefer a more compressive hypothesis with fewer control points.

There is even one dataset 'CNR_Early' where Progol has a slightly higher accuracy than MC-TopLog. This is consistent with the Blumer bound argument, where it indicates that MC-TopLog is in the risk of overfitting when it searches within a much larger hypothesis space to find a high-compression hypothesis. In the context of the two applications, the high-compression hypotheses correspond to the control points that co-regulates as many metabolites as possible.

4.6 Search Space and Compression

Table 4 shows that MC-TopLog always has a larger search space than Progol5. This is consistent with the theoretical analysis discussed earlier. The larger search space make it possible for MC-TopLog to find hypotheses with higher compression than Progol5. Indeed as shown in Table 4, hypotheses suggested by MC-TopLog always has higher compression than those suggested by Progol. In that table, the compression of a hypothesis H is defined as $N_p - N_n - DL$, where N_p and N_n are respectively the number of positive and negative examples covered by H, while DL is short for description length. As explained in Section 2.2, the DL of a hypothesis about substrate limiting and the one about enzyme limiting are respectively L and $k * L$. Here we choose $k = 10$ and $L = 1$, therefore a compression value of 10 in the Table 4 means only one example is compressed by H. Note that more compressive hypotheses does not necessarily correspond to higher accuracies, as you can see when lining up Table 4 with Table 2, This implies that a more complete search to find a more compressive hypothesis does not necessarily gain higher accuracies, which is consistent with the Blumer bound argument. However this does not mean that compression is not a good heuristic for search, but is related to other problems like overfitting.

Timepoint	**Compression**		**Number of Search Nodes**	
	Progol	**MC-TopLog**	**Progol**	**MC-TopLog**
CNR_Early	0	49	352	1240
CNR_Mid	0	33	350	11890
CNR_Late	10	75	322	3654
NOR_Early	10	30	318	411
NOR_Mid	0	34	352	10851
NOR_Late	0	13	354	14032
RIN_Early	20	40	312	350
RIN_Mid	20	40	312	793
RIN_Late	0	14	354	14584

Table 4: Comparing Compression and Search nodes (Tomato Application)

5 Conclusions and Future Work

The use of ILP in the two real-world problems supported efficient analysis of the biological data. Additionally, interesting hypotheses were produced that are different from what the biologists had stated prior to the machine learning. In both applications, MC-TopLog's hypotheses were also compared against human hypotheses provided by the Syngenta project leaders. It was noted that the human hypotheses were closer in form to the reductionist hypotheses generated by Progol. In several cases the MC-TopLog were both more complex and more accurate than the human ones and indicated quite distinct control points within the relevant sub-networks. The plausible hypotheses that do not have support from existing studies will be tested experimentally in future.

As shown by our experiments, there do exist datasets in which systems hypotheses derived by MCL have significantly higher predictive accuracies than the reductionist ones derived by SCL. On the other hand, MCL does not outperform SCL all the time due to the existence of good approximations to the target hypothesis within SCL's hypothesis space. In this case, it seems not worth to apply MCL considering that MCL is much more computationally expensive than SCL. However, for real-world applications whose target theories are unknown,

it is worth trying MCL, as there are datasets where neither the target theory nor its approximations exist within the hypothesis space of SCL, thus MCL can improve the learning results of SCL.

Acknowledgements

The authors would like to acknowledge the support from Syngenta Ltd for funding the University innovations Centre at Imperial College.

References

1. Syngenta Ltd. http://www.syngenta.com/en/index.html.
2. A. Blumer, A. Ehrenfeucht, D. Haussler, and M.K. Warmuth. Occam's razor. *Information Processing Letters*, 24(6):377–380, 1987.
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. C.L. Waterman et al. An integrated functional genomic study of acute phenobarbital exposure in the rat. *BMC Genomics*, 11(1):9, 2010.
6. 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.
7. K Inoue. Induction as consequence finding. *Machine Learning*, 55:109–135, 2004.
8. K. Inoue, T. Sato, M. Ishihata, and et al. Evaluating abductive hypotheses using an EM algorithm on BDDs. In *IJCAI-09*, pages 810–815, 2009.
9. LycoCyc. Solanum lycopersicum database. http://solcyc.solgenomics.net//LYCO/.
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. In *ILP2010*. Springer-Verlag, 2010.
13. S.H. Muggleton, D. Lin, and A. Tamaddoni-Nezhad. MC-TopLog: complete multi-clause learning guided by a top theory. In *ILP11*. Springer-Verlag, 2012.
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. O. Ray. Nonmonotonic abductive inductive learning. *Journal of Applied Logic*, 7(3):329–340, 2009.
16. A. Tamaddoni-Nezhad, D. Bohan, A. Raybould, and S.H. Muggleton. Machine learning a probabilistic network of ecological interactions. In *ILP11*. Springer-Verlag, 2011.
17. 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.
18. L. Valiant. A theory of the learnable. *Journal of the ACM*, 27(11):1134–1142, 1984.
19. A. Yamamoto. Which hypotheses can be found with inverse entailment? In N. Lavrač and S. Džeroski, editors, *ILP97*, pages 296–308. Springer-Verlag, 1997.