Modelling Inhibition in Metabolic Pathways Through Abduction and Induction

Alireza Tamaddoni-Nezhad[1], Antonis Kakas[2],
Stephen Muggleton[1], and Florencio Pazos[3]

[1] Department of Computing,Imperial College London
180 Queen's Gate, London SW7 2BZ, UK
{atn,shm}@doc.ic.ac.uk

[2] Dept. of Computer Science, University of Cyprus
antonis@ucy.ac.cy

[3] Dept. of Biological Sciences, Imperial College London
f.pazos@imperial.ac.uk

Abstract. In this paper, we study how a logical form of scientific modelling that integrates together abduction and induction can be used to understand the functional class of unknown enzymes or inhibitors. We show how we can model, within Abductive Logic Programming (ALP), inhibition in metabolic pathways and use abduction to generate facts about inhibition of enzymes by a particular toxin (e.g. Hydrazine) given the underlying metabolic pathway and observations about the concentration of metabolites. These ground facts, together with biochemical background information, can then be generalised by ILP to generate rules about the inhibition by Hydrazine thus enriching further our model. In particular, using Progol 5.0 where the processes of abduction and inductive generalization are integrated enables us to learn such general rules. Experimental results on modelling in this way the effect of Hydrazine in a real metabolic pathway are presented.

1 Introduction

The combination of abduction and induction has recently been explored from a number of angles [5]. Moreover, theoretical issues related to completeness of this form of reasoning have also been discussed by various authors [33,13,11]. Some efficient implemented systems have been developed for combining abduction and induction [19] and others have recently been proposed [23]. There have also recently been demonstrations of the application of abduction/induction systems in the area of Systems Biology [35,36,18] though in these cases the generated hypotheses were ground. The authors know of no published work to date which provides a real-world demonstration and assessment of abduction/induction in which hypotheses are non-ground rules, though this is arguably the more interesting case. The present paper provides such a study.

The research reported in this paper is being conducted as part of the MetaLog project [32], which aims to build causal models of the actions of toxins from

R. Camacho, R. King, A. Srinivasan (Eds.): ILP 2004, LNAI 3194, pp. 305–322, 2004.

empirical data in the form of Nuclear Magnetic Resonance (NMR) data, together with information on networks of known metabolic reactions from the KEGG database [30]. The NMR spectra provide information concerning the flux of metabolite concentrations before, during and after administration of a toxin.

In our case, examples extracted from the NMR data consist of metabolite concentrations (up-down regulation patterns extracted from NMR spectra of urine from rats dosed with the toxin hydrazine). Background knowledge (from KEGG) consists of known metabolic networks and enzymes known to be inhibited by hydrazine. This background knowledge, which represents the present state of understanding, is incomplete. In order to overcome this incompleteness hypotheses are entertained which consist of a mixture of specific inhibitions of enzymes (ground facts) together with general rules which predict classes of enzymes likely to be inhibited by hydrazine (non-ground). Hypotheses about inhibition are built using Progol5.0 [19] and predictive accuracy is assessed for both the ground and the non-ground cases. It is shown that even with the restriction to ground hypotheses, predictive accuracy increases with the number of training examples and in all cases exceeds the default (majority class). Experimental results suggest that when non-ground hypotheses are allowed the predictive accuracy increases.

The paper is organised as follows. Chapter 2 introduces the biological problem. Background to logical modelling of scientific theories using abduction and induction is given in Chapter 3. The experiments of learning ground and non-ground hypotheses are then described in Chapter 4. Lastly, Chapter 5 concludes the paper.

2 Inhibition in Metabolic Pathways

The processes which sustain living systems are based on chemical (biochemical) reactions. These reactions provide the requirements of mass and energy for the cellular processes to take place . The complex set of interconnected reactions taking place in a given organism constitute its *metabolic network* [14,22,2].

Most biochemical reactions would never occur spontaneously. They require the intervention of chemical agents called catalysers. Catalysers of biochemical reactions - enzymes - are proteins tuned by millions of years of evolution to catalyse reactions with high efficiency and specificity. One additional role of enzymes in biochemical reactions is that they add “control points” to the metabolic network since the absence or presence of the enzyme and its concentration (both controlled mainly by the transcription of the corresponding gene) determine whether the corresponding reaction takes place or not and to which extent.

The assembly of full metabolic networks, made possible by data accumulated through years of research, is now stored and organized on metabolic databases and allows their study from a network perspective [21,1]. Even with the help of this new Systems Biology approach to metabolism, we are still far apart from understanding many of its properties. One of the less understood phenomena, specially from a network perspective, is *inhibition*. Some chemical compounds

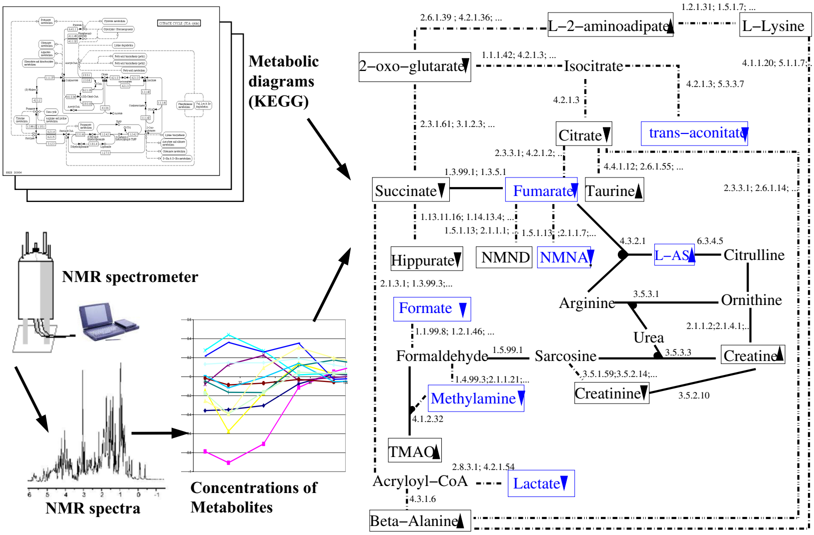

Fig. 1. A metabolic sub-network involving metabolites affected by hydrazine. Information on up/down changes in metabolite concentrations after hydrazine treatment is obtained from NMR spectra. This information is combined with KEGG metabolic diagrams, which contain information on the chemical reactions and associated enzymes.

can affect enzymes impeding them to carry out their functions, and hence affecting the normal flux in the metabolic network, which is in turn reflected in the accumulation or depletion of certain metabolites.

Inhibition is very important from the therapeutic point of view since many substances designed to be used as drugs against some diseases can eventually have an inhibitory side effect on other enzymes. Any system able to predict the inhibitory effect of substances on the metabolic network would be very useful in assessing the potential harmful side-effects of drugs.

In this work we use experimental data on the accumulation and depletion of metabolites to model the inhibitory effect of hydrazine (NH_2-NH_2) in the metabolic network of rats. Figure 1 shows the metabolic pathways sub-network of interest also indicating with “up” and “down” arrows, the observed effects of the hydrazine on the concentration of some of the metabolites involved.

This sub-network was manually built from the information contained in the KEGG metabolic database [30]. Starting from the set of chemical compounds for which there is information on up/down regulation after hydrazine treatment coming from the Nuclear Magnetic Resonance (NMR) experiments, we tried to construct the minimal network representing the biochemical links among them by taking the minimum pathway between each pair of compounds and collapsing all those pathways together through the shared chemical compounds. When there is more than one pathway of similar length (alternative pathways) all of them are included. Pathways involving "promiscuous" compounds (compounds involved in many chemical reactions) are excluded. KEGG contains a static representation of the metabolic network (reactions connecting metabolites). NMR data provides information on the concentrations of metabolites and their changes with time. These data represent the variation of the concentration of a number of chemical compounds during a period of time after hydrazine injection. The effect of hydrazine on the concentrations of chemical compounds is coded in a binary way. Only up/down changes (increasing/decreasing) in compound concentrations immediately after hydrazine injection are incorporated in the model. Quantitative information on absolute or relative concentrations, or fold changes are not used in the present model.

In this sub-network the relation between two compounds (edges in the network) can comprise a single chemical reaction (solid lines) or a linear pathway (dotted lines) of chemical reactions in the cases where the pathway between those compounds is composed by more than one reaction but not involving other compounds in the network (branching points). The directionality of the chemical reactions is not considered in this representation and in fact it is left deliberately open. Although metabolic reactions flow in a certain direction under normal conditions, this may not be the case in "unusual" conditions like the one we are modelling here (inhibition). Inhibition of a given reaction causes the substrates to accumulate what may cause an upstream enzyme to start working backwards in order to maintain its own substrate/product equilibrium.

The "one to many" relations (chemical reactions with more than one substrate or product) are indicated with a circle. The enzymes associated with the relations (single chemical reactions or linear pathways) are shown as a single enzyme or a list of enzymes.

3 Logical Modelling of Scientific Theories

Modelling a scientific domain is a continuous process of observing the phenomena, understanding these according to a currently chosen model and using this understanding, of an otherwise disperse collection of observations, to improve the current general model of the domain. In this process of development of a scientific model one starts with a relatively simple model which gets further improved and expanded as the process is iterated over. Any model of the phenomena at any stage of its development can be *incomplete* in its description. New information given to us by observations, O, can be used to complete this description. As proposed in [4,5], a logical approach to scientific modelling can then be set up by employing together the two *synthetic* forms of reasoning of *abduction* and *induction* in the process of assimilating the new information in the observations. Given the current model described by a theory, T, and the observations O both abduction and induction synthesize new knowledge, H, thus extending the model, T, to $T \cup H$, according to the same formal specification of: $T \cup H \models O$ and $T \cup H$ is consistent.

Abduction is typically applied on a model, T, in which we can separate two disjoint sets of predicates: the *observable* predicates and the *abducible* predicates. The basic assumption then is that our model T has reached a sufficient level of comprehension of the domain such that all the incompleteness of the model can be isolated (under some working hypotheses) in its abducible predicates. The observable predicates are assumed to be completely defined in T; any incompleteness in their representation comes from the incompleteness in the abducible predicates. In practice, observable predicates describe the scientific observations, and abducible predicates that describe underlying relations in our model that are not observable directly but can, through the model T, bring about observable information. We also have *background* predicates that are auxiliary relations that help us link observable and abducible information (e.g. they describe experimental conditions or known sub-processes of the phenomena).

Having isolated the incompleteness of our model in the abducible predicates, these will form the basis of *abductive explanations* for understanding, according to the model, the specific observations that we have of our scientific domain. Abduction generates in these explanations (typically) *extentional* knowledge that is specific to the particular state or scenario of the world pertaining to the observations explained. Adding an explanation to the theory then allows us to predict further observable information but again restricted essentially to the situation(s) of the given observations. On the other hand, inductive inference generates *intentional knowledge* in the form of general rules that are not restricted to the particular scenaria of the observations. The inductive hypothesis thus allows predictions to new, hitherto unseen, states of affairs or scenarios.

A *cycle of integration* of abduction and induction in the process of model development emerges. Abduction is first used to transform (and in some sense normalize) the observations to an extensional hypothesis on the abducible predicates. Then induction takes this as input (training data) and tries to generalize this extentional information to general rules for the abducible predicates. The cycle can then be repeated by adding the learned information on the abducibles back in the model as partial information now on these incomplete predicates.

As an example consider the integration of abduction and induction for modelling inhibition as shown in Figure 2. The purpose of the abduction process is to generate hypotheses about inhibited enzymes from the NMR observations of metabolite concentration. For this purpose we need a logic program which models how the concentration of metabolites (e.g. up-down regulations) is related to inhibition of enzymes (see Section 3.2 for such a model). The purpose of the induction process is to learn from the abduced facts, general rules about inhibition of enzymes in terms of chemical properties of the inhibitor, functional class of enzymes etc. Part of the information about inhibition required by the induction process can be obtained from databases such as BRENDA [29]. However, for many inhibitors the available data may not be enough to generate any general rule. The results of abduction, from the previous stage, then act as invaluable training examples for the induction process.

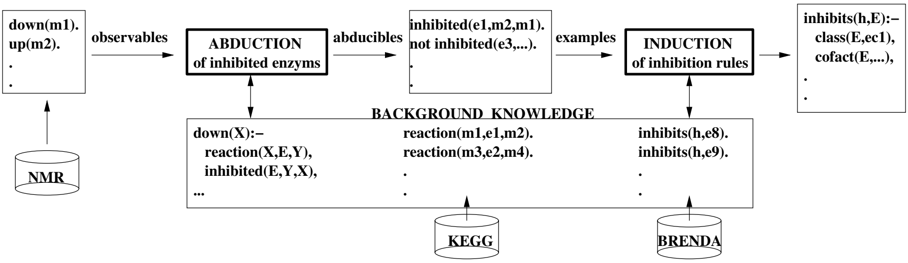

Fig. 2. An Abductive/Inductive framework for modelling inhibition.

In general, the integration of abduction and induction enhances the model development. Moreover, it provides a better opportunity to test the correctness of the generated hypotheses as this can increase the scope of testing. In a *tight integration* of abduction and induction the choice of an explanation in the first abductive phase of the cycle is linked to the second phase of how well the explanation generalizes through induction. Such frameworks of tight integration already exist, e.g. Progol 5.0 [19], ACL [17], SOLDR [34], CF-Induction [12], HAIL [23]. We will use Progol 5.0 to carry out the experiments in our study in this paper.

3.1 Modelling in Abductive Logic Programming

A framework that allows declarative representations of incomplete theories is that of Abductive Logic Programming (ALP) [16,15]. In this framework a model or a theory, T, is described in terms of a triple (P, A, IC) consisting of a logic program, P, a set of abducible predicates, A, and a set of classical logic formulas IC, called the *integrity constraints* of the theory. The program P contains *definitional knowledge* representing the general laws about our problem domain through a complete definition of a set of *observable predicates* in terms of each other, background predicates (which are again assumed to be completely specified in P) and a set of abducible predicates that are open. Abducible predicates appear only in the conditions of the program rules with no definition in P. The integrity constraints, IC, represent *assertional knowledge* that we may have about our domain, augmenting the model in P, but without defining any predicates.

Given such an ALP theory the inference of abduction (i.e. of abductive explanation) is then specialized accordingly in the following way:

Definition 1. *Given an abductive logic theory (P, A, IC), an abductive explanation for an observation O, is a set, Δ, of ground abducible atoms on the predicates A such that:*

- $P \cup \Delta \models_{LP} O$
- $P \cup \Delta \models_{LP} IC$.

where $\models_{LP}$ *denotes the logical entailment relation in Logic Programming*[1]*.*

The abductive explanation Δ represents a hypothesis which when taken together with the model described in the theory T explains how a nonempty experimental observable O could hold. An abductive explanation partially completes the model as described in the theory T. The important role of the integrity constraints IC, is to impose *validity* requirements on the abducible hypotheses Δ. They are modularly stated in the theory, separately from the basic model captured in P, and they are used to augment this with any partial information that we may have on the abducible predicates or other particular requirements that we may want the abductively generated explanations of our observations to have. In most practical cases the integrity constraints are of the form of clausal rules: $B_1 \wedge ... \wedge B_n \rightarrow A_1 \vee ... \vee A_k$ where $A_1, ..., A_k$ and $B_1, ..., B_n$ are positive literals. In these constraints, k can be possibly zero (we will then write the conclusion as **false**) in which case the constraint is a denial prohibiting any set of abducibles that would imply the conjunction $B_1, ..., B_n$.

3.2 Modelling Inhibition in ALP

We will develop a model for analyzing (understanding and subsequently predicting) the effect of toxin substances on the concentration of metabolites. The ontology of our representation will use as *observable* predicates the single predicate:

$$concentration(Metabolite, Level)$$

where *Level* can take (in the simplest case) the two values, *down* or *up*. In general, this would contain a third argument, namely the name of the toxin that we are examining but we will assume here for simplicity that we are studying only one toxin at a time and hence we can factor this out. *Background* predicates such as:

$$reactionnode(Metabolites1, Enzymes, Metabolites2)$$

describe the topology of the network of the metabolic pathways as depicted in figure1. For example, the statement

$$reactionnode('l-2-aminoadipate', '2.6.1.39', '2-oxo-glutarate')$$

expresses the fact that there is a direct path (reaction) between the metabolites $l-2-aminoadipate$ and $2-oxo-glutarate$ catalyzed by the enzyme 2.6.1.39. More generally, we can have a set of metabolites on each side of the reaction and a set of different enzymes that can catalyze the reaction.

Note also that these reactions are in general reversible, i.e. they can occur in either direction and indeed the presence of a toxin could result in some reactions

[1] For example, when the program P contains no negation as failure then this entailment is given by the minimal Herbrand model of the program and the truth of formulae in this model.

changing their direction in an attempt to compensate (re-balance) the effects of the toxin. The incompleteness of our model resides in the lack of knowledge of which metabolic reactions are adversely affected in the presence of the toxin. This is captured through the declaration of the *abducible* predicate:

$$inhibited(Enzyme, Metabolites1, Metabolites2)$$

capturing the hypothesis that the toxin inhibits the reaction from $Metabolites1$ to $Metabolites2$ through an adverse effect on the enzyme, $Enzyme$, that normally catalyzes this reaction. For example,

$$inhibited('2.6.1.39', 'l-2-aminoadipate', '2-oxo-glutarate')$$

expresses the abducible hypothesis that the toxin inhibits the reaction from $l-2-aminoadipate$ to $2-oxo-glutarate$ via the enzyme 2.6.1.39.

Hence the set of abducibles, A, in our ALP theory (P, A, IC), contains the only predicate *inhibited*/3. Completing this would complete the given model. The experimental observations of increased or reduced metabolite concentration will be accounted for in terms of hypotheses on the underlying and non-observable inhibitory effect of the toxin represented by this abducible predicate.

Given this ontology for our theory (P, A, IC), we now need to provide the program rules in P and the integrity constraints IC of our model representation. The rules in P describe an underlying mechanics of the effect of inhibition of a toxin by defining the observable *concentration*/2 predicate. This model is simple in the sense that it only describes at an appropriate high-level the possible inhibition effects of the toxin, abstracting away from the details of the complex biochemical reactions that occur. It sets out simple general laws under which the effect of the toxin can increase or reduce their concentration, Examples of these rules in P are:

```
concentration(X,down):-
    reactionnode(X,Enz,Y),
    inhibited(Enz,Y,X).
concentration(X,down):-
    reactionnode(X,Enz,Y),
    not inhibited(Enz,Y,X),
    concentration(Y,down).
```

The first rule expresses the fact that if the toxin inhibits a reaction producing metabolite X then this will cause down concentration of this metabolite. The second rule accounts for changes in the concentration through indirect effects where a metabolite X can have down concentration due to the fact that some other substrate metabolite, Y, that produces X was caused to have low concentration. Increased concentration is modelled analogously with rules for "up" concentration. For example we have

```
concentration(X,up):-
    reactionnode(Y,Enz,X),
    inhibited(Enz,X,Y).
```

where the inhibition of the reaction from metabolite X to Y causes the concentration of X to go up as X is not consumed due to this inhibition.

Note that for a representation that does not involve negation as failure, as we would need when using the Progol 5.0 system, we could use instead the abducible predicate $inhibited(Enz, TruthValue, Y, X)$ where $TruthValue$ would take the two values $true$ and $false$. The underlying and simplifying working hypotheses of our model are:

(1) the primary effect of the toxin can be *localized* on the individual reactions of the metabolic pathways;
(2) the underlying network of the metabolic pathways is correct and complete;
(3) all the reactions of the metabolic pathways are a-priori equally likely to be affected by the toxin;
(4) inhibition in one reaction is sufficient to cause change in the concentration of the metabolites.

The above rules and working hypotheses give a relatively simple model but this is sufficient as a starting point. In a more elaborate model we could relax the fourth underlying hypothesis of the model and allow, for example, the possibility that the down concentration effect on a metabolite, due to the inhibition of one reaction leading to it, to be compensated by some increased flow of another reaction that also leads to it. We would then have more elaborated program P rules that express this. For example, the first rule above would be replaced by:

```
concentration(X,down):-
    reactionnode(X,Enz,Y),
    inhibited(Enz,Y,X),
    not compensated(X,Enz).
compensated(X,Enz):-
    reactionnode(X,Enz1,Y),
    different(Enz1,Enz),
    increased(Enz1,Y,X).
```

where now the set of abducible predicates A includes also the predicate $increased(Enzyme, Metabolites1, Metabolites2)$ that captures the assumption that the flow of the reaction from $Metabolites1$ to $Metabolites2$ has increased as a secondary effect of the presence of the toxin.

Validity requirements of the model. The abducible information of *inhibited*/3 is required to satisfy several *validity requirements* captured in the integrity constraints IC of the model. These are stated modularly in IC separately from the program P and can be changed without affecting the need to reconsider the underlying model of P. They typically involve general self-consistency requirements of the model such as:

$$concentration(X, down), concentration(X, up) \rightarrow \boldsymbol{false}$$

expressing the facts that the model should not entail that the concentration of any metabolite is at the same time down and up.

Example Explanations. Let us illustrate the use of our model and its possible development with an example. Given the pathways network in figure 1 and the experimental observation that:

$$concentration('2-oxo-glutarate', down)$$

the following are some of its possible explanations

$$E_1 = \{inhibited(2.3.1.61,' succinate',' 2-oxo-glutarate')\}$$
$$E_2 = \{inhibited(2.6.1.39,' l-2-aminoadipate',' 2-oxo-glutarate')\}$$
$$E_3 = \{inhibited(1.1.1.42,' isocitrate',' 2-oxo-glutarate')\}$$

Combining this observation with the additional observation that

$$concentration('isocitrate', down)$$

makes the third explanation E_3 inconsistent, as this would imply that the concentration of isocitrate is up. Now if we further suppose that we have observed

$$concentration('l-2-aminoadipate', up)$$

then the above explanation E_2 is able to account for all three observations with no added hypotheses needed. An alternative explanation would be

$$E_2^{'} = \{inhibited(2.6.1.39,' l-2-aminoadipate',' 2-oxo-glutarate'),$$
$$inhibited(1.2.1.31,' l-2-aminoadipate',' l-lysine')\}$$

Applying a principle of *minimality* of explanations or more generally of *maximal compression* we would prefer the explanation E_2 over $E_2^{'}$.

Computing Explanations by ALP and ILP systems. There are several systems (e.g. [28,27]) for computing abductive explanations in ALP. Also some ILP systems, such as Progol 5, can compute abductive explanations as well as generalizations of these. Most ALP systems, unlike ILP systems, do not employ an automatic way of comparing different explanations at generation/search time and selecting from these those explanations that satisfy some criterium of compression or simplicity. On the other hand, ALP systems can operate on a richer representation language, e.g. that includes negation as failure. Hence although Progol 5 can provide compact and minimal explanations ALP systems can provide explanations that have a more complete form.

In particular, Progol 5 explanations are known to be restrictive [33,23], in that for a single observation/example they can not contain more than one abducible clause. Despite this in many domains where this single clause restriction is acceptable, as is the case in our present study of inhibition in metabolic networks, ground explanations of Progol 5 are closely related to (minimal) ALP explanations. ALP explanations may contain extra hypotheses that are generated from ensuring that the integrity constraints are satisfied. Such hypotheses are left implicit in Progol 5 explanations. This means that Progol 5 and ALP explanations have corresponding predictions, modulo any differences in their vocabularies of representation. For example, referring again to Figure 1, a Progol 5 explanation for the two observations for metabolites $l-2-aminoadipate$ and *succinate* would be:

$$E_{ILP} = \{inhibited(2.6.1.39, true, 'l-2-aminoadipate', '2-oxo-glutarate'),\\ inhibited(1.2.7.3, false, '2-oxo-glutarate', 'succinate')\}$$

This explanation does not carry any information on the rest of the network that is not directly connected with the observations and the abducible hypotheses that it contains. The corresponding ALP explanation(s) have the form:

$$E_{ALP} = \{inhibited(2.6.1.39, 'l-2-aminoadipate', '2-oxo-glutarate'),\\ not\ inhibited(1.2.7.3, '2-oxo-glutarate', 'succinate')\} \cup E_{Rest}$$

where E_{Rest} makes explicit further assumptions required for the satisfaction of the integrity constraints. In this example, if we are interested in the metabolite *isocitrate* then we could have two possibilities:

$$E^1_{Rest} = \{not\ inhibited(1.1.1.42., '2-oxo-glutarate', 'isocitrate'),\\ not\ inhibited(1.1.1.42., 'isocitrate', '2-oxo-glutarate')$$

$$E^2_{Rest} = \{not\ inhibited(1.1.1.42., '2-oxo-glutarate', 'isocitrate'),\\ inhibited(1.1.1.42., 'isocitrate', '2-oxo-glutarate')$$

These extra assumptions are left implicit in the ILP explanations as they have their emphasis on maximal compression. But the predictions that we get from the two types of ALP and ILP explanations are the same. Both types of explanations predict $concentration('2-oxo-glutarate', down)$. For *isocitrate* the first ALP explanation predicts this to have down concentration whereas the second one predicts this to have up concentration. The non-committal corresponding ILP explanation will also give these two possibilities of prediction depending on how we further assume the flow of the reaction between $2-oxo-glutarate$ and *isocitrate*. In our experiments, reported in the following section, we could examine a-posteriori the possible ALP explanations and confirm this link between ground Progol 5 explanations with minimal ALP explanations.

4 Experiments

The purpose of the experiments in this section is to empirically evaluate the inhibition model, described in the previous section, on a real metabolic pathway and real NMR data.

4.1 Experiment 1: Learning Ground Hypotheses

In this experiment we evaluate ground hypotheses which are generated using the inhibition model given observations about concentration of some metabolites.

Materials. Progol 5.0[2] is used to generate ground hypotheses from observations and background knowledge. As a part of background knowledge, we use the relational representation of biochemical reactions involved in a metabolic pathway which is affected by hydrazine. The observable data is up-down regulation of metabolites obtained from NMR spectra. These background knowledge and observable data were explained in Section 2 and illustrated in Figure 1.

[2] Available from: http://www.doc.ic.ac.uk/~shm/Software/progol5.0/

```
for $i$=1 to 10 do
  $Ts_i = m$ test example randomly sampled from $E$
  $Tr_i = E - Ts_i$
  for $j$ in (2,4,6,8,10) do
    $Tr_{ij} = j$ training example randomly sampled from $Tr_i$
  end
end
for $i$=1 to 10 do
  for $j$ in (2,4,6,8,10) do
    $H_{ij}$ = learned hypotheses using the training set $Tr_{ij}$
    $A_{ij}$ = predictive accuracy of $H_{ij}$ on the test set $Ts_{ij}$
  end
end
for $j$ in (2,4,6,8,10) do
    Plot average and error bars of $A_{ij}$ versus $j$ ($i \in [1..10]$)
```

Fig. 3. Experimental method used for Experiment 1. E is the set of all examples and in this experiment $m = 7$.

Methods. In the first attempt to evaluate the model we tried to predict the concentration of a set of metabolites which became available later during the Metalog project. Hence, we have used the previously available observations (shown in black arrows in Figure 1) as training data and the new observations (shown in blue arrows in Figure 1) as test data. According to our model, there are many possible hypotheses which can explain the up-regulation and down-regulation of the observed metabolites. However, Progol's search attempts to find the most compressive hypotheses. The following are examples of hypotheses returned by Progol:

```
inhibited('2.6.1.39',true,'l-2-aminoadipate','2-oxo-glutarate').
inhibited('2.3.1.61',false,'2-oxo-glutarate','succinate').
inhibited('1.13.11.16',false,'succinate','hippurate').
inhibited('2.6.1.-',true,'taurine','citrate').
inhibited('3.5.2.10',true,'creatine','creatinine').
inhibited('4.1.2.32',true,'tmao','formaldehyde').
inhibited('4.3.1.6',true,'beta-alanine','acryloyl-coA').
```

Using these ground hypotheses, the model can correctly predict the concentration of six out of the seven new metabolites. In order to evaluate the predictive accuracy of the model in a similar setting, we generate random test sets (with size equal to seven) and use the remaining examples for training. Figure 3 summarises the experimental method used for this purpose.

The model which has been used for evaluating the hypotheses generated by Progol explicates the Closed World Assumption (CWA). In other words, we are working under the assumption that a reaction is not inhibited unless we have a fact which says otherwise:

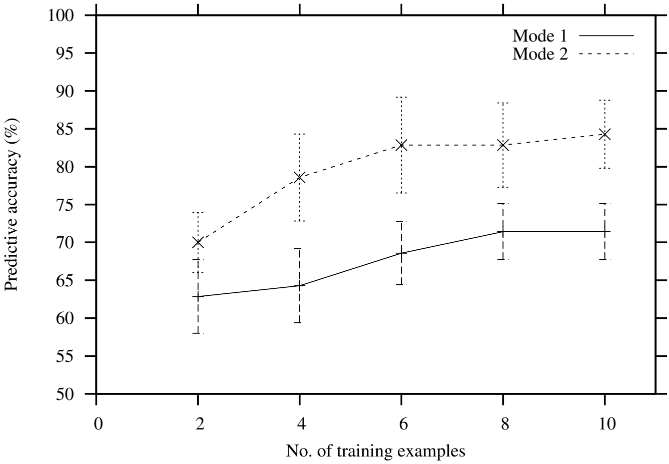

Fig. 4. Performance of the hypotheses generated by Progol in Experiment 1.

inhibited(Enz,false,X,Y):-
    reactionnode(Y,Enz,X),
    not(inhibited(Enz,true,_,_)).

When we include this we will call this evaluation, *mode 2*, and without it we will call the evaluation *mode 1*.

The predictor which we have used in our experiments converts the three class problem which we have ('up', 'down' and 'unknown') to a two class prediction with 'down' as the default class. For this purpose we use the following test predicate:

concentration1(X,up):-
    concentration(X,up),
    not(concentration(X,down)).
concentration1(X,down).

Results and discussion. The results of the experiments are shown in Figure 4. In this graph, the vertical axis shows the predictive accuracy and the horizontal axis shows the number of training examples. According to this graph, we have a better predictive accuracy when we use the closed world assumption (*Mode 2*) compared to the accuracy when we do not use this assumption (*Mode 1*). The reason for this is that the closed world assumption allows the rules of the model (as represented in Progol) have apply in more cases than without the assumption. According to the number of up and down regulations in the examples, the default accuracy is 64.7%. For both *Mode 1* and *Mode 2*, the overall accuracy is above the default accuracy and inreases with the number of training examples.

```
for $i$ in (1,4,8,16) do
    for $j$=1 to $n$ do
        $Ts_{ij} = i$ test examples randomly sampled from $E$
        $Tr_{ij} = E - Ts_{ij}$
    end
end
for $i$ in (1,4,8,16) do
    for $j$=1 to $n$ do
        $H_{ij}$ = learned hypotheses using the training set $Tr_{ij}$
        $A_{ij}$ = predictive accuracy of $H_{ij}$ on the test set $Ts_{ij}$
    end
end
for $i$ in (1,4,8,16) do
        Plot average of $A_{ij}$ versus $j$ ($j \in [1..n]$)
```

Fig. 5. Experimental method used for Experiment 2. E is the set of all examples and in this experiment $n = 17$.

4.2 Experiment 2: Learning Non-ground Hypotheses

As mentioned in the previous sections, abduction and induction can be combined to generate general rules about inhibition of enzymes. In this experiment we attempt to do this by further generalising the kind of ground hypotheses which were learned in Experiment 1.

Materials and Methods. Background knowledge required for this experiment can be obtained from databases such as BRENDA [29] and LIGAND [31]. This background information can include information about enzyme classes, cofactors etc. For example, information on the described inhibition by hydrazine and/or presence of the pyridoxal 5'-phosphate (PLP) group can be extracted from the BRENDA database when such information exists. In our experiments for learning non-ground hypotheses we include the possibility that a given chemical compound can be inhibiting a whole enzymatic class, since this situation is possible in non-competitive inhibition. For example, a very strong reducer or oxidant affecting many oxidoreductases (1.-.-.-). In our case, since the mechanism (competitive/non-competitive) of inhibition of hydrazine is unknown, we leave this possibility open. In this experiment we use all available observations and we apply a leave-out test strategy (randomly leave out 1, 4, 8 and 16 test examples and use the rest as training data). The experimental method is detailed in Figure 5.

Results and discussion. In this experiment Progol attempted to generate general rules for inhibition effectively trying to generalize from the ground facts in the abductive explanations. Among the rules that it had considered were:

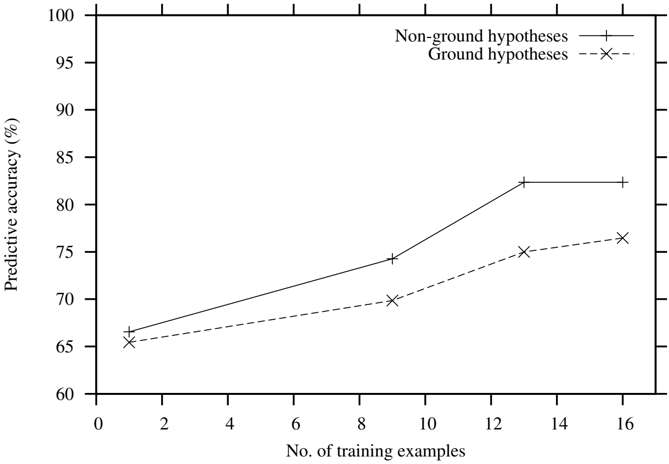

Fig. 6. Performance of ground and non-ground hypotheses generated by Progol using a leave-out test strategy as detailed in Figure 5.

$$inhibited(Enz, true, M1, M2)\colon\!\!-reactionnode(M2, Enz, M1), class(Enz, 2.6.1)$$
$$inhibited(Enz, true, M1, M2)\colon\!\!-reactionnode(M2, Enz, M1), class(Enz, 4.1.2)$$

expressing the information that reactions that are catalysed by enzymes in either of the two classes '2.6.1' and '4.1.2' are inhibited by Hydrazine. These rules had to be eventually rejected by the system as they are inconsistent with the given model. This is because they imply that these reactions are inhibited in both directions while the model assumes that any reaction at any particular time only flows in one direction and hence can only be inhibited in that direction. In fact, the available data is not sufficient for the learning method to distinguish the direction in which the reactions of the network flow. Moreover, it is not appropriate to learn such a relation as we know that metabolic pathways reactions are reversible and so depending on the circumstances they can flow in either direction (see Section 2). The problem therefore is a problem of representation where we simply want to express that these reactions are inhibited in the one direction that they flow whatever this direction might be.

Nevertheless it was instructive to accept these (seemingly overgeneral) rules into our model by adopting a default direction of the reactions of the network involved (i.e. whose enzymes fall in these two classes) and examine the effect of this generalization on the predictive accuracy of our new model compared with the case where the ground abductive explanations are added to the model. This comparison is shown in Figure 6 indicating that the predictive accuracy improves after generalization.

5 Conclusions

We have studied how to use abduction and induction in scientific modelling concentrating on the problem of inhibition of metabolic pathways. Our work has demonstrated the feasibility of a process of scientific model development through an integrated use of abduction and induction. This is to our knowledge the first time that abduction and induction are used together in an enhancing way on a real-life domain.

The abduction technique which is used in this paper can be compared with the one in the robot scientist project [18] where ASE-Progol was used to generate ground hypotheses about the function of genes. Abduction has been also used within a system, called GenePath [35,36], to find relations from experimental genetic data in order to facilitate the analysis of genetic networks. Bayesian networks are among the most successful techniques which have been used for modelling biological networks. In particular, gene expression data has been widely modelled using Bayes' net techniques [7,6,10]. On the MetaLog project Bayes' nets have also been used to model metabolic networks [24]. A key advantage of the logical modelling approach in the present paper compared with the Bayes' net approach is the ability to incorporate background knowledge of existing known biochemical pathways, together with information on enzyme classes and reaction chemistry. The logical modelling approach also produces explicit hypotheses concerning the inhibitory effects of toxins.

A number of classical mathematical approaches to metabolic pathway analysis and simulation exist. These can be divided into three main groups based around Biochemical Systems Theory (BST), Metabolic Control Analysis (MCA) and Flux Balance Analysis (FBA). BST and MCA are oriented toward dynamic simulation of cellular processes based on physicochemical laws [8,9,25]. However, progress towards the ultimate goal of complete simulation of cellular systems [25] has been impeded by the lack of kinetic information and attention in the last decade has been diverted to analysing the relative importance of metabolic events. FBA [26,3] unlike BST and MCA, does not require exact kinetic information to analyse the operative modes of metabolic systems. FBA, which includes the techniques of Elementary Flux Mode Analysis and Extreme Pathway Analysis, only requires stochiometric parameters (the quantitative relationship between reactants and products in a chemical reaction). However, by contrast with the approach taken in the present paper, BST, MCA and FBA are not machine learning approaches, and most importantly do not incorporate techniques for extending the structure of the model based on empirical data.

In the present study we used simple background knowledge concerning the class of enzymes to allow the construction of non-ground hypotheses. Despite this limited use of background knowledge we achieved an increase in predictive accuracy over the case in which hypothesis were restricted to be ground. In future work we hope to extend the representation to include structural descriptions of the reactions involved in a style similar to that described in [20].

Acknowledgements. We would like to thank the anonymous reviewers for their comments. We also thank M. Sternberg, R. Chaleil and A. Amini for their useful discussions and advice, T. Ebbels and D. Crockford for their help on preparing the NMR data and O. Ray for his help with the ALP system. This work was supported by the DTI project "MetaLog - Integrated Machine Learning of Metabolic Networks applied to Predictive Toxicology". The second author is grateful to the Dept. of Computing, Imperial College London for hosting him during the 2003/4 academic year.

References

1. R. Alves, R.A. Chaleil, and M.J. Sternberg. Evolution of enzymes in metabolism: a network perspective. *Mol. Biol.*, 320(4):751–70, 2002 Jul 19.
2. E. Alm E and A.P. Arkin. Biological networks. *Curr. Opin. Struct. Biol.*, 13(2):193–202, 2003 April.
3. J. S. Edwards, R. Ramakrishna, C. H. Schilling, and B. O. Palsson. Metabolic flux balance analysis. In S. Y. Lee and E. T. Papoutsakis, editors, *Metabolic Engineering.* Marcel Deker, 1999.
4. P. Flach and A.C. Kakas. Abductive and inductive reasoning: Background and issues. In P. A. Flach and A. C. Kakas, editors, *Abductive and Inductive Reasoning*, Pure and Applied Logic. Kluwer, 2000.
5. P. A. Flach and A. C. Kakas, editors. *Abductive and Inductive Reasoning.* Pure and Applied Logic. Kluwer, 200.
6. Nir Friedman, Michal Linial, Iftach Nachman, and Dana Pe'er. Using bayesian networks to analyze expression data. *J. of Comp. Bio.*, 7:601–620, 2000.
7. Nir Friedman, Kevin Murphy, and Stuart Russell. Learning the structure of dynamic probabilistic networks. In *Uncertainty in Artificial Intelligence: Proceedings of the Fourteenth Conference (UAI-1998)*, pages 139–147, San Francisco, CA, 1998. Morgan Kaufmann Publishers.
8. B.C. Goodwin. *Oscillatory organization in cells, a dynamic theory of cellular control processes.* Academic Press, New York, 1963.
9. B. Hess and A. Boiteux. Oscillatory organization in cells, a dynamic theory of cellular control processes. *Hoppe-Seylers Zeitschrift fur Physiologische Chemie*, 349:1567 – 1574, 1968.
10. S. Imoto, T. Goto, and S. Miyano. Estimation of genetic networks and functional structures between genes by using bayesian networks and nonparametric regression. In *Proceeding of Pacific Symposium on Biocomputing*, pages 175–186, 2002.
11. K. Inoue. Induction, abduction and consequence-finding. In C. Rouveirol and M. Sebag, editors, *Proceedings of the International Workshop on Inductive Logic Programming (ILP01)*, pages 65–79, Berlin, 2001. Springer-Verlag. LNAI 2157.
12. K. Inoue. Inverse entailment for full clausal theories. In *LICS-2001 Workshop on Logic and Learning*, 2001.
13. K. Ito and A. Yamamoto. Finding hypotheses from examples by computing the least generlisation of bottom clauses. In *Proceedings of Discovery Science '98*, pages 303–314. Springer, Berlin, 1998. LNAI 1532.
14. H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, and A.L. Barabasi. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, 2000 Oct 5.

15. A. C. Kakas and M. Denecker. Abduction in logic programming. In A. C. Kakas and F. Sadri, editors, *Computational Logic: Logic Programming and Beyond. Part I*, number 2407, pages 402–436, 2002.
16. A. C. Kakas, R. A. Kowalski, and F. Toni. Abductive Logic Programming. *Journal of Logic and Computation*, 2(6):719–770, 1993.
17. A.C. Kakas and F. Riguzzi. Abductive concept learning. *New Generation Computing*, 18:243–294, 2000.
18. R.D. King, K.E. Whelan, F.M. Jones, P.K.G. Reiser, C.H. Bryant, S.H. Muggleton, D.B. Kell, and S.G. Oliver. Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature*, 427:247–252, 2004.
19. S.H. Muggleton and C.H. Bryant. Theory completion using inverse entailment. In *Proc. of the 10th International Workshop on Inductive Logic Programming (ILP-00)*, pages 130–146, Berlin, 2000. Springer-Verlag.
20. S.H. Muggleton, A. Tamaddoni-Nezhad, and H. Watanabe. Induction of enzyme classes from biological databases. In *Proceedings of the 13th International Conference on Inductive Logic Programming*, pages 269–280. Springer-Verlag, 2003.
21. J.A. Papin, N.D. Price, S.J. Wiback, D.A. Fell, and B.O. Palsson. Metabolic pathways in the post-genome era. *Trends Biochem. Sci.*, 28(5):250–8, 2003 May.
22. E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai, and A.L. Barabasi. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–5, 2002.
23. O. Ray, K. Broda, and A. Russo. Hybrid Abductive Inductive Learning: a Generalisation of Progol. In *13th International Conference on Inductive Logic Programming*, volume 2835 of *LNAI*, pages 311–328. Springer Verlag, 2003.
24. A. Tamaddoni-Nezhad, S. Muggleton, and J. Bang. A bayesian model for metabolic pathways. In *International Joint Conference on Artificial Intelligence (IJCAI03) Workshop on Learning Statistical Models from Relational Data*, pages 50–57. IJCAI, 2003.
25. J.J. Tyson and H. G. Othmer. The dynamics of feedback control circuits in biochemical pathways. *Progress in Theoretical Biology*, 5:1–62, 1978.
26. A. Varma and B. O. Palsson. Metabolic flux balancing: Basic concepts, scientific and practical use. *Bio/Technology*, 12:994–998, 1994.
27. A-System Webpage:. http://www.cs.kuleuven.ac.be/~bertv/asystem/.
28. ALP-Systems Webpage:. http://www.doc.ic.ac.uk/ ~ or/abduction/alp.pl and http://www.cs.ucy.ac.cy/aclp/.
29. BRENDA Webpage:. http://www.brenda.uni-koeln.de/.
30. KEGG Webpage:. http://www.genome.ad.jp/kegg/.
31. LIGAND Webpage:. http://www.genome.ad.jp/ligand/.
32. MetaLog Webpage:. http://www.doc.ic.ac.uk/bioinformatics/metalog/.
33. A. Yamamoto. Which hypotheses can be found with inverse entailment? In *Proceedings of the Seventh International Workshop on Inductive Logic Programming*, pages 296–308. Berlin, 1997. LNAI 1297.
34. A. Yamamoto and B. Fronhöfer. Finding hypotheses by generalizing residues hypotheses. pages 107–118, September 2001.
35. B. Zupan, I. Bratko, J. Demsar, J. R. Beck, A. Kuspa, and G. Shaulsky. Abductive inference of genetic networks. *AIME*, pages 304–313, 2001.
36. B. Zupan, I. Bratko, J. Demsar, P. Juvan, J.A Halter, A. Kuspa, and G. Shaulsky. Genepath: a system for automated construction of genetic networks from mutant data. *Bioinformatics*, 19(3):383–389, 2003.