Abduction and induction for modelling inhibition in metabolic networks

Alireza Tamaddoni-Nezhad[1] and **Raphael Chaleil**[2] and **Antonis Kakas**[3] and **Stephen Muggleton**[1]
[1]Dept. of Computing, Imperial College London
Email: {atn,shm}@doc.ic.ac.uk
[2]Dept. of Biological Sciences, Imperial College London
Email: r.chaleil@imperial.ac.uk
[3]Dept. of Computer Science, University of Cyprus
Email: antonis@ucy.ac.cy

Abstract

This paper describes the use of a mixture of abduction and induction for the problem of identifying the effects of toxins in metabolic networks. Background knowledge is used which describes network topology and functional classes of enzymes. 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 the toxin (non-ground). The foreground examples were derived from in vivo experiments involving NMR analysis of time-varying metabolite concentrations in rat urine following injections of toxin. Hypotheses about inhibition are built using Progol5.0 and predictive accuracy is assessed for both the ground and the non-ground cases.

1 Introduction

The combination of abduction and induction has recently been explored from a number of angles [Flach and Kakas, 2000]. Moreover, theoretical issues related to completeness of this form of reasoning have also been discussed by various authors [Yamamoto, 1997; Ito and Yamamoto, 1998; Inoue, 2001]. Some efficient implemented systems have been developed for combining abduction and induction [Muggleton and Bryant, 2000] and others have recently been proposed [Ray *et al.*, 2003]. There have also recently been demonstrations of the application of abduction/induction systems in the area of Systems Biology [Zupan *et al.*, 2001; 2003; King *et al.*, 2004].

The research reported in this paper is being conducted as part of the MetaLog project [1], which aims to build causal models of the actions of toxins from empirical data in the form of Nuclear Magnetic Resonance (NMR) data, together with information on networks of known metabolic reactions from the KEGG database [2]. The NMR spectra provide information concerning the flux of metabolite concentrations before, during and after administration of a toxin.

In a previous paper [Tamaddoni-Nezhad *et al.*, 2004] describing the initial investigation in this topic we modelled the initial effects of a toxin (Hydrazine). The previous model ignored the temporal variance of metabolite concentrations. By contrast, in this paper we describe an extended study in which temporal variation is captured by the model.

In this study, 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) for different time periods. Background knowledge (from KEGG) consists of known metabolic networks and enzymes known to be inhibited by the toxin. 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 the toxin (non-ground). Hypotheses about inhibition are built using Progol5.0 [Muggleton and Bryant, 2000] and predictive accuracy is assessed for both the ground and the non-ground cases. Models performance is evaluated using a leave-one-out test procedure. It is shown that even with the restriction to ground hypotheses, predictive accuracy exceeds the default (majority class).

The paper is organised as follows. Section 2 introduces the biological problem. The logical modelling of inhibition is given in Section 3. The experiments of learning ground and non-ground hypotheses are then described in Section 4 and Section 5 concludes the paper.

2 Inhibition in metabolic pathways

Metabolism is made by the processes used by living organisms to transform elements from their environment into building bricks or as a source of energy and also to transform toxic compounds that may have penetrated in them into less toxic compounds to be excreted. All these processes form a very complex network of chemical reactions or *metabolic network* [Jeong *et al.*, 2000 Oct 5; Ravasz *et al.*, 2002; Alm and Arkin, 2003]. Not all reactions take place at the same time in this network, and need to be finely coordinated. Biochemical reactions are sped up by highly specialised proteins, the enzymes. Enzymes are the most efficient catalysers

[1]http://www.doc.ic.ac.uk/bioinformatics/metalog/
[2]http://www.genome.ad.jp/kegg/

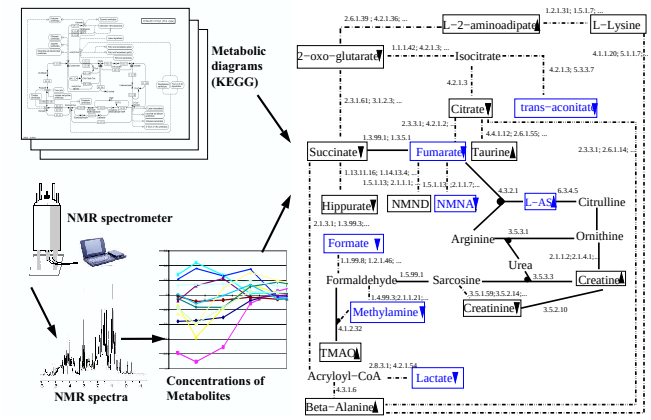

Figure 1: A metabolic sub-network involving metabolites affected by hydrazine. Information on up/down changes in metabolite concentrations after hydrazine treatment is obtained from NMR spectra. This information is combined with KEGG metabolic diagrams, which contain information on the chemical reactions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

known, and most of the reactions taking place in living organisms would be too slow without them to sustain life. Enzymes can be considered as the keys controlling the activation of different parts of the network, and are therefore the main element for coordination of the different parts of the metabolic network.

Even with the help of the 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 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.

In this work, toxicity experimental data were used, to look at the normalised variations of different metabolites concentrations in rat urine, measured by Nuclear Magnetic Resonance (NMR). The data was obtained after injecting Hydrazine (NH_2-NH_2) to rats, and following the concentration of metabolites over time using NMR. The NMR data represent variations of concentration of the metabolites relative to their concentration before injection of hydrazine, at 8 hours, 24 hours, 48 hours, 72 hours, 96 hours, etc.

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 [2]. Starting from the set of chemical compounds for which there is information on up/down regulation after toxin treatment coming from the 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.

3 Abduction for modelling inhibition

We will develop a model for analyzing (understanding and subsequently predicting) the effect of toxin substances on the concentration of metabolites. We use as the set of *observables* the single predicate:

$$concentration(Metabolite, Level, Time)$$

expressing the fact that at some time, $Time$, a metabolite, $Metabolite$, has a certain level of concentration, $Level$ which in the simplest case can take the two values, $down$ or up. In general, the concentration predication would contain a fourth 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('l2aminoadipate','2.6.1.39','2oxoglutarate')

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 changing their direction in an attempt to compensate (re-balance) the effects of the toxin. The model also involves background biochemical data on enzymes and metabolites that would be used in the process of inductive generalization of the abduced hypotheses.

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:

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

capturing the hypothesis that at the time $Time$ the reaction from $Metabolites1$ to $Metabolites2$ is inhibited by the toxin through an adverse effect on the enzyme, $Enzyme$, that normally catalyzes this reaction. For example,

$$inhibited('2.6.1.39',' l2aminoadipate',' 2oxoglutarate', 8)$$

expresses the abducible hypothesis that at time 8 the the reaction from $l-2-aminoadipate$ to $2-oxo-glutarate$ via the enzyme $2.6.1.39$ is inhibited by the toxin.

Hence the set of abducibles contains the only predicate $inhibited/4$ and 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.

We now need to provide the rules and the integrity constraints of our model representation. The rules describe an underlying mechanics of the effect of inhibition of a toxin by defining the observable $concentration/3$ 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 are:

concentration(X,down,T):-
reactionnode(X,Enz,Y),
inhibited(Enz,Y,X,T).

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

The first rule expresses the fact that if a reaction producing metabolite X is inhibited at time T then this will cause down concentration of this metabolite at this time. 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 (even when the reaction is not currently inhibited). Increased concentration is modelled analogously with rules for "up" concentration. For example we have

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

where the inhibition of the reaction from metabolite X to Y causes the concentration of X to go up as X is not (currently) 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 $inhibited(Enz, TruthValue, Y, X, T)$ 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 rules that express this. For example, the first rule above would be replaced by:

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

compensated(X,Enz,T):-
reactionnode(X,Enz1,Y),
different(Enz1,Enz),
increased(Enz1,Y,X,T).

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

The abducible information of $inhibited/4$ is required to satisfy several *validity requirements* captured in the integrity constraints of the model. These are stated modularly and separately from the program rules and can be changed without affecting the need to reconsider the underlying model. They typically involve general self-consistency requirements of the model such as:

$: -concentration(X, down, T), concentration(X, up, T)$

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

In addition, specific partial information that we may have on the abducible predicates $inhibited/4$ (such as that a certain reaction cannot be inhibited by the toxin that we are examining) can be captured as a validity requirement.

Other such constraints can help us restrict further the form of the abductive explanations that we are looking for, essentially adding in this way extra working hypotheses to our model. We could, for example, be interested only in explanations whose inhibition effects are separated apart on the pathways network. This would be captured by an integrity constraint of the form:

:-inhibited(Enz,X,Y,T),inhibited(Enz1,Y1,Z,T), close(Y,Y1)

where the auxiliary background predicate $close(Y, Y1)$ holds true iff the shortest distance between the two metabolite nodes Y and $Y1$ is smaller than a given minimum distance.

4 Empirical evaluation

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

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

Abduction and induction could be combined to generate general (non-ground) rules about inhibition of enzymes. In this experiment we also examine if we can achieve this by further generalising the ground hypotheses. In particular, we test the following null hypothesis:

Null hypothesis: Generalising ground hypotheses, which are generated from the abductive model for inhibition, does not lead to increased predictive accuracy.

In this experiment Progol 5.0 [3] 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 the toxin. 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. Background knowledge required for non-ground hypotheses can be obtained from databases such as BRENDA [4] and LIGAND [5]. This background information can include information about enzyme classes, co-factors etc.

In this experiment we use NMR observations for 8hrs to 96hrs as training/test examples and apply a leave-one-out test strategy (leaving out one example as the test data and using the rest as training data).

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:

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

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

concentration1(X,down,T).

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 ground hypotheses returned by Progol for the inhibitory effect of Hydrazine at time 8hrs:

```
inhibited('2.6.1.39',true,'l2aminoadipate','2oxoglutarate',8).
inhibited('2.3.1.61',false,'2oxoglutarate','succinate',8).
inhibited('1.13.11.16',false,'succinate','hippurate',8).
inhibited('2.6.1.-',true,'taurine','citrate',8).
inhibited('3.5.2.10',false,'creatine','creatinine',8).
inhibited('4.1.2.32',true,'methylamine','tmao',8).
inhibited('4.3.1.6',true,'beta-alanine','acryloyl-coA',8).
```

In this experiment Progol is also used to generate general rules for inhibition by generalizing the ground facts in the abductive explanations. An example of such a rule is:

inhibited(Enz, true, M1, M2, T) :-
reactionnode(M2,Enz,M1),
class(Enz,'aminotransferase').

expressing the information that reactions that are catalysed by enzymes in the enzymatic class 'aminotransferase' are inhibited by the toxin.

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

[4] http://www.brenda.uni-koeln.de/

[5] http://www.genome.ad.jp/ligand/

Model	Predictive accuracy
Ground hypotheses	62%
Ground hypotheses + cwa	83%
Non-ground hypotheses + cwa	86%
Default class	54%

Table 1: Overall predictive accuracies of ground (abduction only) and non-ground (abduction + induction) hypotheses with the closed world assumption (cwa) from a leave-one-out test procedure. The 'Default class' model is one that simply guesses the majority class.

The overall predictive accuracies of ground and non-ground hypotheses are summarised in Table 1. According to this table, in all cases the overall accuracies are above the default accuracy of 54% (a model that simply guesses the majority class). This table also suggests that a model which uses non-ground hypotheses has a bettter performance than the one which only uses ground hypotheses.

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.

The abduction technique which is used in this paper can be compared with the one in the robot scientist project [King *et al.*, 2004] where ASE-Progol was used to generate ground hypotheses about the function of genes. Abduction has been also used within a system, called GenePath [Zupan *et al.*, 2001; 2003], 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 [Friedman *et al.*, 1998; 2000; Imoto *et al.*, 2002]. On the MetaLog project Bayes' nets have also been used to model metabolic networks [Tamaddoni-Nezhad *et al.*, 2003]. 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.

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 [Muggleton *et al.*, 2003].

References

[Alm and Arkin, 2003] E. Alm and A.P. Arkin. Biological networks. *Curr. Opin. Struct. Biol.*, 13(2):193–202, 2003.

[Flach and Kakas, 2000] P. A. Flach and A. C. Kakas, editors. *Abductive and Inductive Reasoning*. Pure and Applied Logic. Kluwer, 2000.

[Friedman *et al.*, 1998] 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.

[Friedman *et al.*, 2000] 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.

[Imoto *et al.*, 2002] 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.

[Inoue, 2001] 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.

[Ito and Yamamoto, 1998] 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.

[Jeong *et al.*, 2000 Oct 5] 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.

[King *et al.*, 2004] 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.

[Muggleton and Bryant, 2000] 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.

[Muggleton *et al.*, 2003] 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.

[Ravasz *et al.*, 2002] 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.

[Ray *et al.*, 2003] 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.

[Tamaddoni-Nezhad *et al.*, 2003] 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.

[Tamaddoni-Nezhad *et al.*, 2004] A. Tamaddoni-Nezhad, A. Kakas, S.H. Muggleton, and F. Pazos. Modelling inhibition in metabolic pathways through abduction and induction. In *Proceedings of the 14th International Conference on Inductive Logic Programming*, pages 305–322. Springer-Verlag, 2004.

[Yamamoto, 1997] 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.

[Zupan *et al.*, 2001] B. Zupan, I. Bratko, J. Demsar, J. R. Beck, A. Kuspa, and G. Shaulsky. Abductive inference of genetic networks. *AIME*, pages 304–313, 2001.

[Zupan *et al.*, 2003] 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.