Learning Qualitative Metabolic Models.

G. M. Coghill [1] and **S. M. Garrett** [2] and **R. D. King** [2]

Abstract. The ability to learn a model of a system from observations of the system and background knowledge is central to intelligence, and the automation of the process is a key research goal of Artificial Intelligence. We present a model-learning system, developed for application to scientific discovery problems, where the models are scientific hypotheses and the observations are experiments. The learning system, Qoph learns the *structural* relationships between the observed variables, known to be a hard problem. Qoph has been shown capable of learning models with hidden (unmeasured) variables, under different levels of noise, and from qualitative or quantitative input data.

1 Introduction

The development of intelligent tools to aid in the process of Scientific Discovery, particularly in the construction of explanatory models, is an important goal of AI; and qualitative modelling provides an ideal representation. This is the ultimate in adaption, and a hybrid system merging Inductive Logic Programming and Qualitative Simulation is a suitable tool for acheiving it. Bioinformatics is an ideal domain for applying this technology: the data are sparse (making it unsuitable for numerical techniques), they are noisy and they require the construction of models which will inevitably include unobserved variables. Work on constructing models of systems in molecular biology is in the early stages of development and so, given the above stated challenges any useful results emerging will be of tremendous practical value.

The ultimate goal in this scientific quest is the production of quantitative models; however, the discovery of suitable structural models (qualitative differential equations) can be the means of directing the scientist as to which experiments to carry out next in the path towards this goal. In this paper we present Qoph a learning system which combines Inductive Logic Programming (ILP) with QSIM in order to construct qualitative models of physical and biological systems containing unmeasured variables.

2 Background

2.1 Qualitative Simulation

QSIM [8] is a constraint based qualitative simulation engine and utilises an equational representation which is an abstraction of *ordinary differential equations*. It is the most highly developed constraint based Qualitative Reasoning (QR) system available.

In QSIM, each model consists of a set of variables linked together via a set of *constraints*, called a *qualitative differential equation* (QDE). Each variable consists of a $\langle qmag, qdir \rangle$ pair. Here, *qmag* is the qualitative magnitude of the variable. It has a quantity space of varying resolution consisting of alternating points (called landmark values) and intervals; typically the quantity space is divided into the regions $[-\infty \ldots 0), [0], (0 \ldots \infty]$, where infinity is treated as a value. A *qdir* is the qualitative rate of change of the variable, which has a fixed, three valued resolution (the three quantities being *inc*, for increasing; *dec*, for decreasing; and *std*, for steady). Each constraint has only one operation and is defined between two or three variables.

There are several kinds of constraint which can appear in a QSIM model. There are predicates, implemented as relations, representing the usual algebraic operations of addition, multiplication, and sign inversion; plus a derivative predicate stating that one variable is the derivative of another.

One of the attractive features of QSIM is that it is designed to handle incompleteness in the knowledge of the model. The incompleteness here takes the form of a lack of knowledge concerning functional relations in the system. This situation is captured by the monotonic function constraints M^+ and M^- between two variables, which declares that one variable monotonically increases (+) or decreases (-) with respect to another variable, covering families of relations.

The conjunction of qualitative relations models the relationships between a set of measured variables, plus a number of unmeasured variables. There may be zero, one or more unmeasured variables, which we term the model's *intermediate variables.* Where there are sufficient intermediate variables, the method described here can discover *hidden relations* that relate only intermediate variables; this is a novel feature of the learning system presented here.

2.2 Inductive Logic Programming (ILP)

The general model learning problem can be represented deductively as follows: if we term the observations (*evidence*) E, the background knowledge B, and the hypothesis to be learnt H, then given that:

$$B \not\models E \tag{1}$$

find a hypothesis H so that

$$B \wedge H \models E \tag{2}$$

Many solutions to this problem are possible, e.g. the trivial solutions of E, or B → E. The problem is therefore how to

[1] Department of Computing Science, University of Aberdeen, Aberdeen AB24 3UE. email: gcoghill@csd.abdn.ac.uk

[2] Department of Computer Science, University of Wales, Aberystwyth SY23 3DB. emai: {smg, rdk}@aber.ac.uk

restrict solutions to suitable ones. In abduction [3] solutions are restricted to ground facts; in ILP more general solutions are allowed [11], although there are still typically syntactic restrictions on what form solutions can take. For most scientific discovery problems it is clear that ILP is advantageous, as we wish to learn general theories; and for similar reasons ILP is a sensible choice for learning QSIM models.

ILP is distinguished from other machine learning techniques by using first-order predicate logic (specifically logic programs) to represent background knowledge, observations, and hypotheses [10]; and we have previously applied machine learning and ILP to many scientific problems with success(e.g. [6]).

The learning of qualitative models from examples is a great challenge for current machine learning methods since the search space is very large. The problem is also interesting because the data are *positive only*, i.e. when identifying a system, nature only provides positive examples of states of the system, not examples the system can *not* be in. This hinders machine learning as there are no negative examples to restrict over-generalisation.

2.3 Related Work

Automated model construction is an important and growing area of research which has as a central aim the provision of appropriate models for scientific and industrial tasks. The ideal situation would be for a learning system to be supplied with only positive data for some of the variables of the system of interest plus some background knowledge and then produce a model which explains the data in a physically meaningfully manner, identify any hidden (unmeasured) variables and not be overconstrained. This is the hard to achieve target at which researchers are aiming. Previous work in the area has tended to either require that all variables be measured (e.g. [5]), required negative data (e.g. [1]), generate models that were overconstrained (e.g. [7, 12]) or models that were logically but not physically equivalent to the plant being modelled (e.g. [1]). In addition there has been no comprehensive testing of the conditions under which learning of qualitative models is possible. For further details see Garrett *et al* [4].

3 Model Learning Methodology - The Qoph Method

The Aleph ILP system [13] was used as a wrapper for the Qoph implementation, which was written separately. As with [1], we used a subset of QSIM, implemented in Prolog, as background knowledge for ILP. The task of the model learning method is to induce a model given example values for a known set of qualitative variables (a set of qualitative states), and the model language of qualitative relations that can be applied to those variables.

ILP, like much of learning, can be considered to be a search through a space of possible solutions. In the case of learning QSIM models, this space is the set of all possible QSIM models, partially ordered by generality. The relation-variable lattice is traversed by best-first search, and the search of this space can be constrained by the use of various heuristics. These heuristics can be generated from a number of sources: for example systems theory or the domain knowledge of the areas under investigation. In the former case the heuristics consist of general principles from systems theory such as: models must be parsimonious, operate under integral causality, and contain no algebraic loops (although these latter are preferences rather than absolute rules - since for some systems it is not always possible to achieve them). Also, for example, if one is working in the biological area some of the domain knowledge may consist of a set of rules regarding legal chemical reactions that may take place.

3.1 Testing the QOPH method

The Qoph system was developed as a tool to aid in the construction of structural models of systems in molecular biology. This is a domain in which data are sparse and inherently noisy; therefore it was important that Qoph be thoroughly tested under these conditions in order to ascertain its potential as such a tool; and the following set of experiments were devised for this purpose.

1. Starting with a complete envisionment (containing N states) every combination of $N - K$ states from the envisionment (for $K = 0 \ldots N$) was created (giving an experiment space of $2^N - 1$ experiments) and the ability of Qoph to learn the target model from each set of states was tested. This set of experiments measures the sensitivity of Qoph to sparcity of data alone.
2. For the complete envisionment of N states, experiments were run (termed *inverse noise*) in which the total number of states used to by Qoph to learn from was kept constant (at N) with the number of real states being progressively replaced by a number of qualitative noisy states; from 0 (no noise) to N (only noise). A noisy qualitative state is defined as a state that is not part of the complete envisionment but is of the same form, containing the same number and type of variables. This tests the supposed effect of noise introduced in the quantitative to qualitative conversion process.
3. For a selection of the experiments used in (1) a random number of qualitative noisy states were added to the real ones and the effect on learning measured. This was done to simulate the effect of converting noisy signals.
4. Finally experiments were run to test the whole process (from data acquisition and interpretation to model construction) for both clean and noisy quantitative data.

In order to illustrate the approach used and the results obtained we will utilise a coupled two compartment model, since compartmental models are oftem used to represent metabolic systems. Details of the full set of tests and results can be found in [**?**]. In this system the input, $inflow_1$, is the input to compartment 1 and the output, $outflow_2$, is the elimination to the environemnt from compartment 2 (see Fig. 1). The model of this system is:

DERIV($conc_1$, $netflow_1$),
DERIV($conc_2$, $netflow_2$),
ADD($conc_2$, $concDiff$, $conc_1$),
M^+($concDiff$, $flow_{1-2}$),
M^+($conc_2$, $outflow_2$),
ADD($netflow_2$, $outflow_2$, $flow_{1-2}$),
ADD($flow_{1-2}$, $netflow_1$, $inflow_1$).

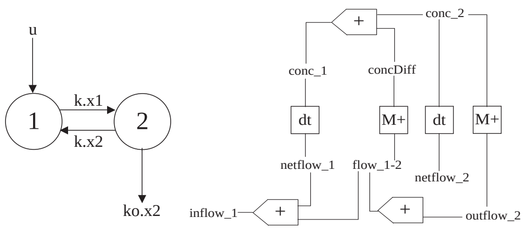

Figure 1. The coupled system (a) compartments; (b) QSIM

Here there are three unmeasured variables: '$netflow_1$', '$netflow_2$', and '$concDiff$'. For the system to be correctly learned these variables will have to be induced. Variable '$inflow_1$' (the input) is exogenous to the model and so appears only once.

4 Results

Since any given experiment will induce its models from a finite number of states, it is possible to plot the *average* reliability for all the experiments for a particular number of states, from one state up to the number of states in the complete envisionment. This 'Average reliability' is given in the range [0 1]. For the noise experiments, the noise dimension is projected on the comparative 2-D plot (this assumes an average noise for each point on the state dimension) to allow comparison with clean data experiments, but a 3-D plot is also presented for the noise experiments that includes the noise dimension.

The plots of the number of states used against average reliability for the coupled tanks are shown in Fig. 2.

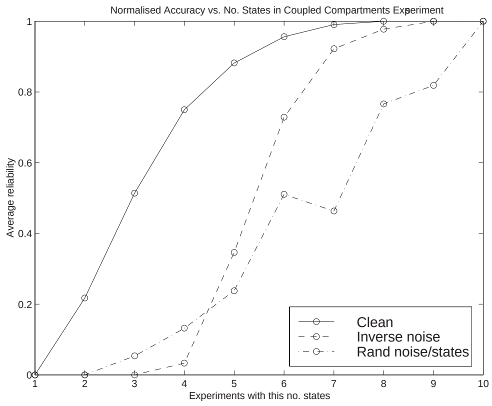

Figure 2. Coupled compartments reliability graphs

We analysed the performance of subsets of the complete envisionment to test whether certain subsets helped QOPH to learn the correct model more reliably than others. If this were the case then there would a a number of *minimal subsets* that contained the lowest number of states that reliably lead to the correct model being found. Subset analysis of the clean data experiments for the coupled tanks give the following states as the minimal subsets.

[1,6], [6,8], [6,9] (state 6 with 1, 8 or 9)
[2,8], ([6,8]), [7,8] (state 8 with 2, 6 or 7)
[1,2,3], [1,2,4], [1,2,5] (states 1 and 2 with 3, 4 or 5)
[1,3,7], [1,4,7], [1,5,7] (states 1 and 7 with 3, 4 or 5)
[3,7,9], [4,7,9], [5,7,9] (states 2 and 9 with 3, 4 or 5)
[2,3,9], [2,4,9], [2,5,9] (states 7 and 9 with 3, 4 or 5)

Fig. 3 shows the relationship of these states in the envisionment graph. A comparison with Table 1 reveals two key features: a selection of states from different behaviours and the use of the critical points of the system are the key to inducing the correct model reliably (see Discussion section below).

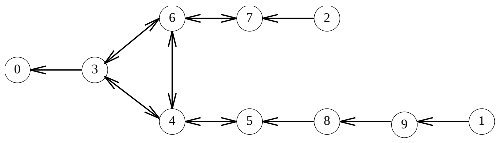

Figure 3. The envisionment graph for the coupled two compartment model

The results from the numerical data experiments are presented in Fig. 4. The legend in the top right corner associates initial values of the state variables (given as two concatenated digits) to a plot; 'all' is the case where the union of states from all initial conditions were used in learning. These results show that it is possible to learn models from clean and noisy numerical data. As discussed above, the qualitative states generated from the clean numerical data contain a number of unavoidable data transformation errors, and the resulting qualitative states form at most a single behaviour of the system under investigation. The set of states gleaned from quantitative to qualitative conversion did not form a full behaviour for the coupled two compartment model, which makes the ability to learn a model from them even more impressive.

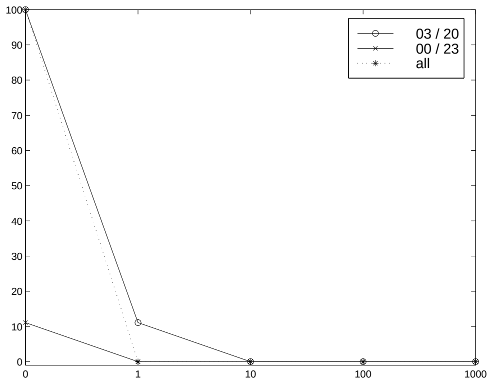

Figure 4. Reliability of learning the correct model from numerical data vs. 1000ths of full Gaussian noise for coupled tanks

5 Application to Biological Systems

As well as exploring the effects of sparcity of data and adding noise it was important to test the *scalablilty* of the QOPH

State	$level_A$	$level_B$	$crossflow_{AB}$	$outflow_B$
0	$< 0, std >$	$< 0, std >$	$< 0, std >$	$< 0, std >$
1	$< 0, inc >$	$< (0,\infty), dec >$	$< (-\infty,0), inc >$	$< (0,\infty), dec >$
2	$< (0,\infty), dec >$	$< 0, inc >$	$< (0,\infty), dec >$	$< 0, inc >$
3	$< (0,\infty), dec >$	$< (0,\infty), dec >$	$< (0,\infty), dec >$	$< (0,\infty), dec >$
4	$< (0,\infty), dec >$	$< (0,\infty), dec >$	$< (0,\infty), std >$	$< (0,\infty), dec >$
5	$< (0,\infty), dec >$	$< (0,\infty), dec >$	$< (0,\infty), inc >$	$< (0,\infty), dec >$
6	$< (0,\infty), dec >$	$< (0,\infty), std >$	$< (0,\infty), dec >$	$< (0,\infty), std >$
7	$< (0,\infty), dec >$	$< (0,\infty), inc >$	$< (0,\infty), dec >$	$< (0,\infty), inc >$
8	$< (0,\infty), std >$	$< (0,\infty), dec >$	$< 0, inc >$	$< (0,\infty), dec >$
9	$< (0,\infty), inc >$	$< (0,\infty), dec >$	$< (-\infty,0), inc >$	$< (0,\infty), dec >$

Table 1. The envisionment states for the coupled compartmental system.

learning method. So far we have only described models constructed from the basic QSIM primitives; to improve scalibility it was useful to be able to use the well-established AI principle of *chunking* [9].

Metabolic pathways essentially contain only two types of molecule: metabolites and enzymes, we therefore designed two *Metabolic Components*, built from standard QSIM relations, to model *metabolites* and *enzymes*. Concentrations of metabolites vary over time as they are synthesised or utilised by enzymatically catalysed reactions. This means that their concentration at time t is a function of their concentration at time $t-1$, and the amount that they are used or created by various enzyme reactions. This can be expressed as a simple summation in QSIM. The qualitative equation for the metabolite components is therefore:

$$\frac{dM}{dt} = \sum_{i=0}^{n}(enzm\_flow_i). \tag{3}$$

The other form of high-level metabolic component in a metabolic pathway are enzymes. Each enzyme is assumed to have one or two inputs and one or two outputs. If there are two inputs or outputs these are considered to form an input or output complex, such that the amount of the complex is proportional to the amount of the inputs or outputs multiplied together. The input complex is converted into the output complex which then disassociates into the output metabolites, and vice versa. The overall flow through the enzyme is the amount of input complex formed minus the amount of output complex formed. The qualitative equation for the enzyme components is therefore[3]:

$$flow = \underbrace{\mathbf{M}^{+}(\prod_{i=1}^{n} M_i)}_{inputcomplex} - \underbrace{\mathbf{M}^{+}(\prod_{j=1}^{m} M_j)}_{outputcomplex}. \tag{4}$$

This is an abstraction of standard kinetic equations [2] and is an expression of the collision probabilities of the metabolites and enzyme. We assume for simplicity that enzymes are taken to exist in constant amounts; although this is clearly a simplification this assumption is also used in ODE modelling. These metabolic components are shown in Fig. 5.

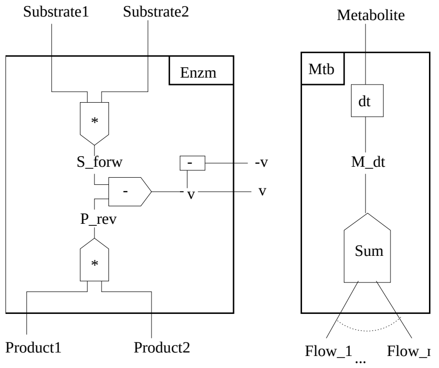

Figure 5. Metabolic components for metabolic system modelling

A model of glycolysis in *Trypanasoma brucei* constructed from these Metabolic Components is shown in Fig. 6. The qualitative model is easier to understand than an ODE since it extracts out detail and allows a complete envisionment of the states.

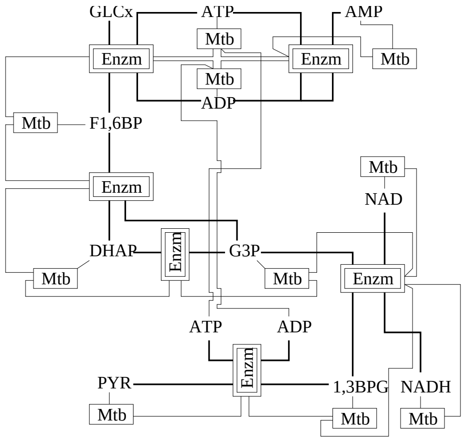

Figure 6. The glycolysis metabolic pathway, built from metabolic components

Using the states from the model, the metabolic components, described above, and general backgound knowledge about biochemical interactions between molecules, it was possible to

[3] Note the distinction between M and $\mathbf{M}^{+}$, the amount of a metabolite and the monotonically increasing relation respectively.

learn a model of the first half of the glycolysis pathway in only a few hours.

Learning a model of this sort represents a major step forward because the learned system is equivalent to a QSIM model consisting of 36 relations, It was calculated that introducing high level components has more than doubled the complexity of the models that can be learned, as well as making the resulting models easier to read.

6 Discussion and Conclusions

The first general point to note is that for all the experiments the number of measured variables from which learning took place remained constant and was less than the total number of variables in the target model. Thus in all circumstances the learning system had to find the intermediate variables and their relationships to the other variables of the model.

Analysis of the clean data experiments showed that given the complete envisionment of a system the correct model was always reliably found. As one would expect there was a gradual deterioration in the reliability as the number of states presented as data was reduced. However, a closer analysis of the results in conjunction with the envisionment graphs for the target models reveals that there is a strong relationship between the reliability of the learning process and the number, and type, of states used in an experiment.

An interesting result from this analysis is the observation that models can be reliably learnt from a minimal number of qualitative states (two in the case of the coupled two compartment system) if the states come from different branches in the envisionment graph. So we can hypothesise first of all that in order to reliably learn a system the data used should come from experiments yielding qualitatively different behaviours (that is behaviours which would appear as distinct branches in an envisionment graph).

However, this hypothesis only provides a necessary, but not a sufficient condition for learning. It was noted during the analysis that in each case where the model was reliably learnt with a minimal number of states, at least one of the states is a critical point of the first derivative of at least one of the state variables: indicating the importance of these critical point states to the definition of a system. What this means is that if an experiment were set up in which all the state variables were exactly at their critical points then the experiment could be run for a very short time and the correct model structure identified. Of course, it is impossible to set up such an experiment, especially in the situation where the structure of the system is completely unknown. Another alternative is to set up multiple experiments with the state variables set to their extremæ: from these initial conditions all the states of the envisionment will eventually be passed through. The downside of this is that the experiments may be difficult to set up and could take an very long time to complete. These two scenarios form the ends of a spectrum within which the optimal experimental setting will lie; the identification of the the best strategies is an important area of research arising from the results of the present work, but it is beyond the scope of this paper.

The main results from testing QOPH on benchmark problems, as illustrated here by the coupled compartmental system, can be summarised as:

- The benchmark models could be induced from their complete envisionments.
- As the number of states chosen from the complete envisionment increases so does the frequency and reliability of finding the correct model.
- The correct model can always be reliably found given a relatively small subset of the total envisionment. There is a set of these subsets such that other state subsets are either supersets of one member of this set, or do not reliably give rise to the correct model.
- Even though subsets containing *very few* states can reliably give rise to the correct model, it is possible to select subsets containing *almost all* the states that do *not* reliably lead to the correct model.
- Qualitative models can be leart from data containing noisy qualitative states, though the overall reliability is reduced.
- Models can be learned from noisy simulated real data for the benchmark systems.

In addition to the results presented here we have also used QOPH to learn a qualitative model representing the complex biological process of glycolysis [4].

Acknowledgement

This work was partly supported by BBSRC/EPSRC grant BIO10479. The authors would like to thank Ashwin Srinivasan, Stephen Oliver and Douglas Kell for their helpful discussions on various aspects of this paper.

REFERENCES

[1] I. Bratko and S. Muggleton, 'Learning qualitative models of dynamic systems', in *Inductive Logic Programming*, ed., S. Muggleton, 437–452, Academic Press, (1992).

[2] W. W. Cleland, 'The kinetics of enzyme-catalysed reactions with two or more substrates and products: 1. nomenclature and rate equations', *Biochim. Biophys. Acta*, **67**, 104–137, (1963).

[3] P. A. Flach and A. C. Kakas, *Abduction and Induction: Essays on their relation and integration*, Kluwer Academic Publishers, 2000.

[4] S. M. Garrett, G. M. Coghill, R. D. King, and A. Srinivasan, 'On learning qualitative models of qualitative and real-valued data', Technical Report UWA-DCS-01-037, University of Wales, Aberystwyth, (August 2001).

[5] D. T. Hau and E. W. Coiera, 'Learning qualitative models of dynamic systems', *Machine Learning*, **26**, 177–211, (1993).

[6] R. D. King and A. Srinivasan, 'The discovery of indicator variables for qsar using inductive logic programming', *Journal of Computer-Aided Molecular Design*, **11**, 571–580, (1997).

[7] I. C. Kraan, B. L. Richards, and B. J. Kuipers, 'Automatic abduction of qualitative models', in *Proceedings of Qualitative Reasoning 1991 (QR'91)*, (1991).

[8] B. Kuipers, *Qualitative Reasoning*, MIT Press, 1994.

[9] J. E. Laird, P.S. Rosenbloom, and A. Newell, 'Chunking in soar: The anatomy of a general learning mechanism', *Machine Learning*, **1**, 11–46, (1986).

[10] S. Muggleton, *Inductive Logic Programming*, Academic Press, London, 1992.

[11] S. Muggleton and C. Feng, 'Efficient induction of logic programs', in *Proc. of the First Conf. on Algorithmic Learning Theory*, OHMSHA, Tokyo, (1990).

[12] A. C. C. Say and S. Kuru, 'Qualitative system indentification: deriving structure from behavior', *Artificial Intelligence*, **83**, 75–141, (1996).

[13] A. Srinivasan, *Aleph web site :*, web.comlab.ox.ac.uk/oucl/research/areas/machlearn/Aleph/aleph_toc.html, 2000.