A Bayesian Model for Metabolic Pathways

Alireza Tamaddoni-Nezhad and **Stephen Muggleton** and **Jung-Wook Bang**
Department of Computing, Imperial College London
180 Queen's Gate, London SW7 2AZ, UK
email: {atn,shm,jwb3}@doc.ic.ac.uk

Abstract

In this paper we introduce a modular approach for modelling metabolic pathways using Bayesian networks. We examine different models for a single reaction metabolism and introduce a Bayesian model for this purpose and then demonstrate this approach by developing a Bayesian model for the aromatic amino acid pathway of yeast. We compare the performance of this model with the performance of a probabilistic model of the same pathway which is based on Stochastic Logic Programs (SLPs). Preliminary results suggest that in parameter estimation from data, the Bayesian model for the yeast metabolic pathway outperforms the SLP model. These results also show that unlike the SLP model, introducing an additional pathway within the Bayesian model does not result in a significant quantitative difference.

1 Introduction

Genomic data is now being obtained on an industrial scale. The complete genomes of at least a dozen microorganisms have been sequenced (e.g. *E. coli* and bakers yeast *S. cerevisiae*). The genomes of about another 50 organisms are in the process of being sequenced and the first complete drafts of the human genome were published in 2001 [Consortium, 2001; Venter *et al.*, 2001]. The focus of genome research is moving to the problem of identifying the biological functions of genes. This is known as *Functional Genomics*. This problem is important because nothing is known about the function of between 30-60% of all new genes identified from sequencing [Oliver, 1996]. Functional Genomics is recognised as central to a deeper understanding of biology, and the future exploitation of biology in medicine, agriculture, and biotechnology in general.

The analysis of the Genomic data needs to become as industrialised as the methods for obtaining it. Within functional genomics, probabilistic approaches such as hidden Markov models, stochastic grammars and Bayesian networks have been proved to be useful [Durbin *et al.*, 1998]. Bayesian networks have been recently used for modelling gene expression data [Friedman *et al.*, 2000]. The benefit of Bayesian networks in this domain has been justified by the comprehensive graphical representation of gene expression data with the possible explanation of causal relations among gene variables (levels). This could be explained with the causality modelling originally proposed by Pearl and Verma [Pearl and Verma, 1991]. In this approach each arc can be interpreted as a causal connection between a prior gene variable and a posterior gene variable at a given time. This approach has been used to extract biologically plausible conclusions from real expression data [Spellman *et al.*, 1998] by deploying search algorithms and statistical confidence measurements [Friedman *et al.*, 2000]. An extension of this approach has been also suggested for dealing with temporal reasoning using dynamic Bayesian networks [Friedman *et al.*, 1998]. In a different attempt [Imoto *et al.*, 2002] in gene expression domain, non-parametric regression and Bayesian networks have been used for constructing genetic networks from gene expression data and to deal with continuous variables.

In addition to gene expression, a crucial body of genomic data is the information on functional pathways in a given cell or tissue, representing processes such as metabolism. This information is available via *metabolic pathways*. Online databases such as KEGG[1], WIT[2] and BRENDA[3] describe relationships between tens of thousands of enzymes and metabolites.

In this paper we introduce a modular approach for modelling metabolic pathways using Bayesian networks. We examine different models for a single reaction metabolism and introduce a Bayesian model for this purpose and then demonstrate this approach by developing a Bayesian model for the aromatic amino acid pathway of yeast. We compare the performance of this model with the performance of a probabilistic model [Angelopoulos and Muggleton, 2002] of the same pathway which is based on Stochastic Logic Programs (SLPs) [Muggleton, 1999].

This paper is organised as follows. Section 2 de-

[1] http://www.genome.ad.jp/kegg/
[2] http://wit.mcs.anl.gov/WIT2/
[3] http://www.brenda.uni-koeln.de/

scribes metabolic pathways and in particular the aromatic amino acid pathway of yeast. Section 3 examines different models for a single reaction metabolite and introduces a Bayesian model for this purpose. In section 4, we evaluate the models discussed in section 3 and also we develop and examine a Bayesian model for the yeast pathway. Finally, section 5 concludes the paper and suggests directions for further research.

2 Metabolic Pathways and the yeast metabolism

In computational genomic, a cell can be viewed as a biochemical machine that consumes simple molecules to generate more complex ones by chaining together biochemical reactions into long sequences referred to as *metabolic pathways*. Genes play an essential role in these networks by providing the information to synthesise the enzymes that catalyse biochemical reactions. Figure 1 illustrates an abstract, highly simplified, model of a living cell. The cell imports molecules from the growth medium (Amino acids, Purines, etc.). These are converted via a pathway of chemical reactions to essential molecules for growth (Proteins, Nucleic Acids, etc.). Each chemical reaction is catalyzed by an enzyme. Some of these enzymes are known and others are not.

Figure 2 shows the aromatic amino acid pathway of yeast [Bryant *et al.*, 2001]. The aromatic amino acids are essential amino acids i.e. amino acids which humans must obtain from dietary sources. However microorganisms, such as yeast, and plants can synthesise them using the aromatic amino acid pathway which is never found in animals. The pathway is therefore an important target for herbicides, antibiotics and live vaccines.

In figure 2 circles represent metabolites and the pathway pertains to the biosynthesis of the aromatic amino acids phenylalanine, tyrosine and tryptophan which are shown in red. The main reaction paths are shown in black. Molecules which are involved in the reactions but which do not lie on the main pathway are shown in blue. Each rectangle, together with those metabolites which are linked to it, represents a chemical reaction. Reactants are shown entering on one side of each rectangle and products leaving on the other. Each rectangle is labelled by the class of the enzyme which catalyses the reaction.

The model uses unique identifiers from the literature to refer to metabolites and enzymes. Metabolites are referred to using their accession numbers in the LIGAND database in the Kyoto Encyclopedia of Genes and Genomes (KEGG) [Goto *et al.*, 2000] and enzymes by their Enzyme Commission classification numbers [IUBMB, 1992].

3 Models for a single reaction metabolism

In this section we discuss different models for a single reaction metabolism and introduce a Bayesian model for this purpose. A single reaction metabolism consists of an enzyme and a set of metabolites each can be either a substrate or a product in a biochemical reaction. In theory, a chemical reaction can occur in both directions and a metabolite, therefore, can act as a substrate as well as a product in the same reaction. However, in this paper we consider reactions which only occur in one direction (i.e. irreversible reactions). Figure 3.a shows a single reaction metabolism with two substrates and two products. Each metabolism is regulated by an enzyme which can be 'active' or 'inactive'. An enzyme is inactive when enzyme inhibition occurs. Inhibition is a reduction in the rate of a catalysed reaction by substances called inhibitors. In biochemical reactions, in which the catalysts are enzymes, if the inhibitor molecules resemble the reactants they may bind to the active site of the enzyme, so preventing normal enzymatic activity.

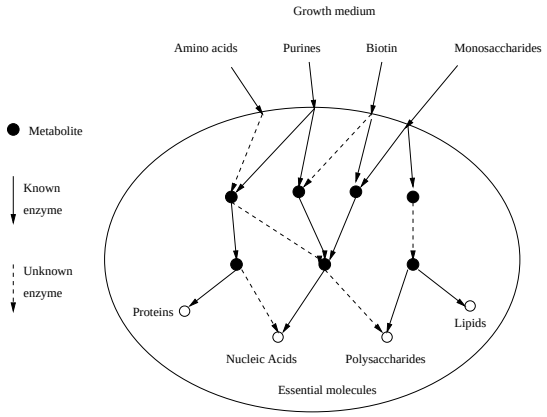


Figure 1: Illustration of cell with metabolic network involved in converting growth media into molecules essential for life.

A single reaction metabolism can be viewed as a logic circuit as shown in Figure 3.b. According to this logic view, products of the reaction are generated if all substrates are present and none of inhibitors are present. A similar assumption was used in the previous studies [Bryant *et al.*, 2001; Reiser *et al.*, 2001] where a metabolic pathway was modelled as a logic program. In this approach substrates and products are represented by arguments in predicates which code for enzyme reactions. The metabolic network is then modelled by the calling diagram of different predicates in the logic program. In [Bryant *et al.*, 2001] it was demonstrated that active learning in a logic programming setting could reduce the expected cost of experimentation for discovering the function of genes.

The main shortcoming of this logic-based approach is that it cannot represent the degree of uncertainty which is involved in each reaction. It also cannot distinguish between situations that might be different in reality. For example, in this model there is no difference between situations where all inhibitors are present and when only one of them is present. This model also fails to account

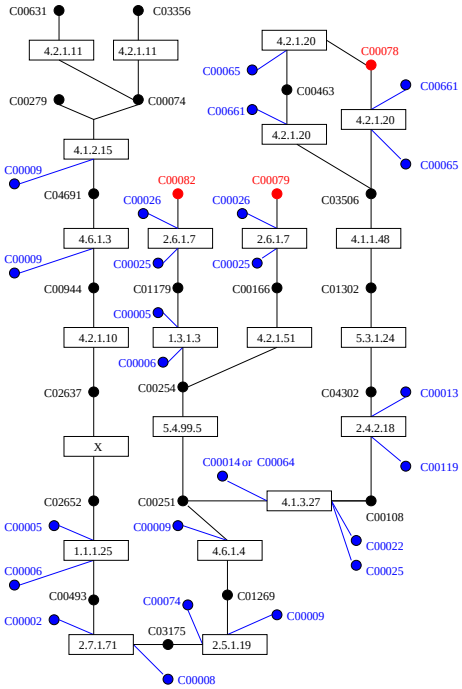


Figure 2: Aromatic amino acid pathway of yeast.

for relative rates of reactions. To overcome these shortcomings, a natural way is to use probabilistic logic representations rather than pure logic programming. Angelopoulos and Muggleton showed that for modelling metabolic networks, a probabilistic representation is required [Angelopoulos and Muggleton, 2002]. They used Stochastic Logic Programming (SLP) [Muggleton, 1999] to model a metabolic pathway and a parameter estimation algorithm [Cussens, 2001] to estimate the parameters of the metabolic pathway from artificial data. Unlike the logic-based models, a model using SLP is able to capture the relative rate of reactions. Figure 4 shows an SLP representation of a single reaction metabolism. In addition to its ability for representing probabilities, SLP is especially useful for representing relational background knowledge about biochemical reactions. However, the learning techniques for SLP are still under development [Muggleton, 2002].

In this paper we use a graphical probabilistic representation for modelling metabolic pathways. This probabilistic representation is based on Bayesian Belief Networks. Bayesian belief networks (or briefly Bayesian networks) are Directed Acyclic Graphs (DAGs) where each node represents a random variable. The intuitive meaning of an arrow from a parent node to a child node is that the parent node directly influences the child node. These influences are quantified by Conditional Probability Tables (CPTs). In Bayesian networks we assume that each node is conditionally independent of all of its non-descendants given its parents. Bayesian networks are compact methods for representing joint probability

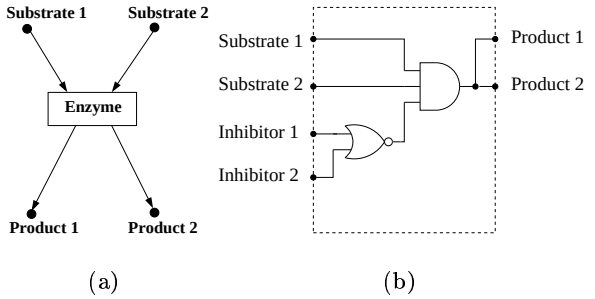


Figure 3: a) A single reaction metabolism b) A logic view of a single reaction metabolism

```
enzyme(enzyme1,reaction1,[substrate1,substrate2],[product1,product2]).
0.80 :: reaction1(yes,yes,yes,yes).
0.20 :: reaction1(yes,yes,no,no).
```

Figure 4: An SLP model for the single reaction metabolism.

distributions and the marginal probabilities can be computed very efficiently. Bayesian networks have been used for causal modelling [Pearl, 2000]. For this purpose a causal Bayesian network is defined as a Bayesian network in which each arc is interpreted as a direct causal influence between a parent and a child node, relative to the other nodes in the network.

Bayesian networks have some properties which are of special interest in learning metabolic pathways. Firstly, there are well-developed methods for learning parameters as well as structure of a Bayesian network [Heckerman, 1995], secondly there are techniques for introducing missing or unobservable nodes, and thirdly Bayesian networks support incremental learning.

In this section we introduce a Bayesian model for a single reaction metabolism. Table 1 shows a general mapping between Bayesian networks and metabolic networks. In this mapping the existence of each metabolite is represented by a propositional variable (i.e. a node in the Bayesian network). If we assume that each metabolite in the pathway has only two possible states, present and absent, then each metabolite can be represented by a binary random variable. The relationship between substrates and products in a reaction can be represented by parent-child relations. Finally, different probabilities involved, including the probability of detectable products, are represented by Conditional Probability Tables (CPTs).

Figure 5.a shows a Bayesian model for a single reaction metabolism. This model combines the logical view in Figure 3.b with the general mapping scheme in Table 1. In this model a metabolic pathway is described as a directed graph, where the vertices (nodes) are metabolites (substrates and products). Each metabolite node corresponds to a binary random variable which determines whether the metabolite is 'present' or 'absent'. Enzymes

Table 1: A mapping between Bayesian networks and Metabolic networks

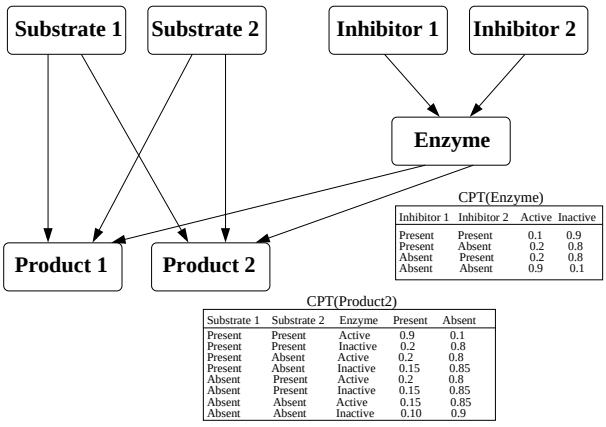
Metabolic network	Bayesian network
Existence of metabolites	Propositional variables
Reaction - Substrates & Products	Parent-child relation
Probability of detectable product	Conditional Probability Table

regulating reactions and their inhibitors are also represented by binary random variables. An edge between two node represents that the parent directly influences the child. These influences are quantified by Conditional Probability Tables (CPTs) for each node. Figure 5.a also shows examples of CPT for nodes 'Enzyme' and 'Product 2'. In this example, 'Product 2' is present with probability 90% if 'Substrate 1' and 'Substrate 2' are present and 'Enzyme' is active. However, this probability is dramatically reduced if any of substrates are absent or the enzyme is inactive and the probability varies between 10%,15% and 20% for different configurations of substrates and enzyme.

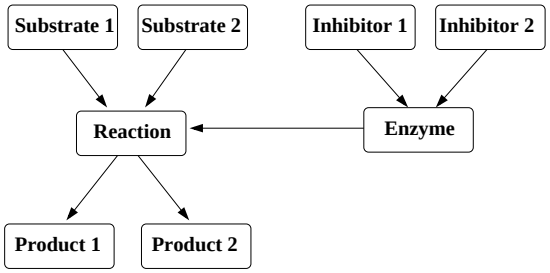
In this model each product is directly connected to the enzyme and all substrates in the reaction. The number of parameters in the CPT for a product node increases exponentially with the number of substrates. For example, if we have three substrates (instead of two), then the number of parameters increases from 2×8 to 2×16 and 2×32 for four substrates and so on. This rapid increase in the number of parameters could result in difficulties when learning parameters from data, mainly because more training examples are required when we have more parameters to estimate.

According to the theoretical results on the sample complexity of learning Bayesian networks [Friedman and Yakhini, 1996], one measure of the complexity is the number of parameters in B^*, where B^* denote the (minimal) Bayesian network which describe the underlying distribution. As shown in [Friedman and Yakhini, 1996], the sample complexity of learning is sub-linear in the number of parameters of B^*.

To improve Model 1 in terms of number of parameters, we propose Model 2 in which product nodes are connected to enzyme and substrates through an intermediate node called reaction. Even though in this model we have an additional unobservable node (i.e. reaction), the total number of parameters in this model is less than Model 1. This is because the product nodes in Model 2 only have one parent and a 2×2 CPT while each product node in Model 1 has three parental nodes and a 2×8 CPT. The difference between total number of parameters increases when we have more substrates and products for each reaction. For this reason it is expected that Model 2 outperform Model 1 especially when number of substrates and products are increased. In section 4 we present an experiment for testing this conjecture. Another advantage of Model 2 is that in this model it is easier to define and re-use reactions and enzymes in a modular way.



(a) Model 1



(b) Model 2

Figure 5: Two Bayesian models for a single reaction metabolic pathway.

4 Experimental evaluation

In this section we examine the performance of the models discussed in section 3 and also we develop and evaluate a Bayesian model for the yeast pathway. For this purpose we conduct learning experiments for estimating parameters of each model from artificial data. In Experiment 1 we compare the performance of Model 1 and Model 2 as described in section 3. In Experiment 2 we conduct a similar experiment as in [Angelopoulos and Muggleton, 2002] to learn parameters of the metabolic pathway of yeast. The purpose of the later experiment is to compare parameter learning of a pathway modelled by SLP with parameter learning of the same pathway modelled by Bayesian network. In this experiment we also investigate the effect of introducing an additional pathway within the network and compare the result with the result reported in [Angelopoulos and Muggleton, 2002]. We use the theoretical results on the sample complexity of learning Bayesian networks [Friedman and Yakhini, 1996] to evaluate different Bayesian networks. In both Experiment 1 and Experiment 2 we compare the number of training examples which are required by different models (with different number of parameters) to achieve a certain level of accuracy.

4.1 Experiment 1: Model 1 vs. Model 2

In this experiment we compare the required number of training examples for Model 1 and Model 2 (see Figure 5) to achieve a certain level of accuracy. As discussed before, in Model 1 products are directly influenced by the enzyme and substrates, whereas in Model 2 an intermediate node, namely reaction, influences the products. Even though the number of nodes is increased, the total number of parameters in this model is less than Model 1. For this reason it is expected that Model 2 require less training examples than Model 1 to achieve the same level of accuracy. The purpose of Experiment 1 is to test this conjecture. In the following, we explain the material and methods used in the experiment and then discuss the results.

Material and methods

In this experiment we use Netica which is a commercial Bayesian network software and can be used for inference and learning in Bayesian belief networks [4]. Figure 6 shows the experimental method used in this experiment. This experimental method has been used to measure the predictive accuracy of both Model 1 and Model 2. The main scenario is as follows. First we set up each network using Netica and assign the Conditional Probability Tables (CPTs) with some fictional values which will be used as 'true ' probabilities. Then we run the network to generate a given number of random samples for all variables in the model. These samples will be used for training and testing purposes. The 'true' CPT values are then replaced with uniformly distributed probabilities and the learning facility of Netica is used to estimate the network parameters from 'training' cases. The accuracy of the learned network can then be tested by measuring the error rate of particular nodes (i.e. product nodes, in this experiment). Finally, the average and standard error of the error rate for the mentioned nodes for 10 different runs are plotted against the number of training examples. This procedure is repeated for both Model 1 and Model 2.

[4] A free version and documentation are available from: www.norsys.com

```
for i = 1 to 10 do
    for j in (10, 100, 1000) do
      Set up Conditional Probability Tables (CPTs) with 'true' values
      Simulate network to generate j random 'training' cases
      Simulate network to generate 10000 random 'test' cases
      Replace CPT values with uniformly distributed probabilities
      Learn CPTs from 'training' cases (parameter estimation)
      E_ij =average error rate of the output nodes on 'test' cases
    end
end
for j in (10, 100, 1000) do
    Plot average and standard error of E_ij versus j (i ∈ [1..10])
```

Figure 6: Experimental method used to examine Bayesian models.

The error rate measured in this experiment is predictive error of the model on 10000 random test cases (generated in step 5). Equation 1 shows how predictive error is calculated. In the contingency table shown, P stands for predicted by the model and $\overline{P}$ not predicted, A stands for actual positive example and $\overline{A}$ actual negative examples. In equation 1, a and d are the number of positive and negative test examples which are correctly predicted by the model and b and c are the number of negative and positive test examples which are not correctly predicted by the model.

$$Error\% = \frac{b+c}{a+b+c+d} \qquad (1)$$

	A	$\overline{A}$
P	a	b
$\overline{P}$	c	d

Results and discussion

The results of the experiment are shown in Figure 7. This graph suggests that in general the error rate of Model 2 is less than Model 1 and therefore Model 2 requires less training examples than Model 1 to achieve the same level of accuracy. The total number of parameters in Model 2 is 20 compared to 24 in Model 1. These results are consistent with the theoretical results on the sample complexity of learning Bayesian networks. According to [Friedman and Yakhini, 1996], the sample complexity of learning is sub-linear in the number of parameters of B^*, where B^* denote the (minimal) Bayesian network which describe the underlying distribution.

4.2 Experiment 2: A Bayesian model for the yeast pathway

The purpose of this experiment is to compare the performance of the Bayesian model for the yeast pathway (see figure 2) with the SLP model for the same pathway in learning parameters from data. To be able to compare

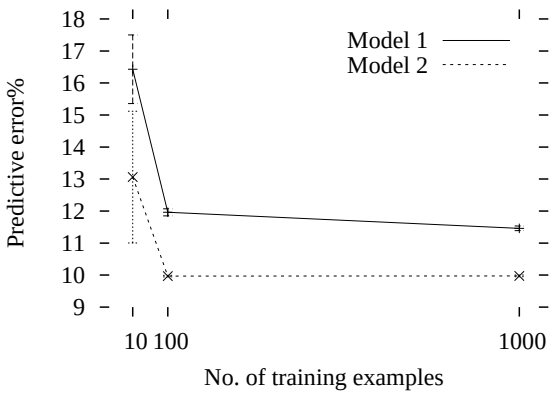


Figure 7: Predictive error for the Bayesian models 'Model 1' and 'Model 2' shown in Figure 5.

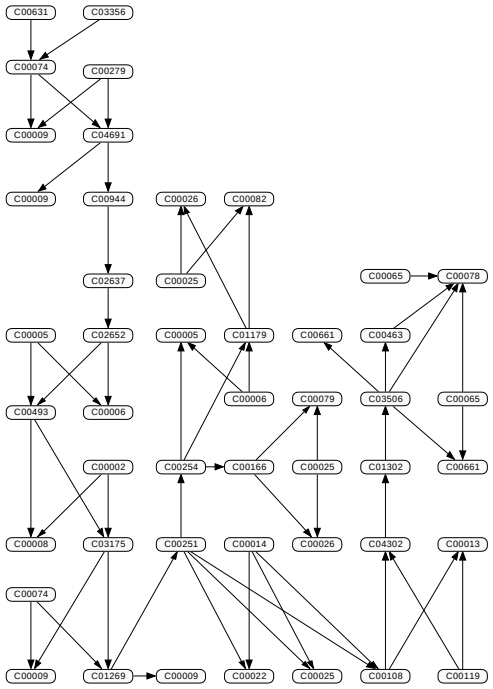
the results of this experiment with the results for the SLP model reported in [Angelopoulos and Muggleton, 2002], we need to consider the following assumptions: 1) we use the mapping scheme used in Model 1 rather than Model 2 to avoid intermediate nodes (i.e. reaction nodes) so that the nodes in the network correspond to the metabolites in the same pathway which were considered in the experiments in the SLP model 2) we assume that all enzymes are active (i.e. none of inhibitors are present), so we do not need to represent enzyme and inhibitors in the model. In addition to these assumption we consider the following simplifying assumptions which were also considered in the SLP model: a) reactions deplete their substrates b) each reaction is only considered once.

Material and methods

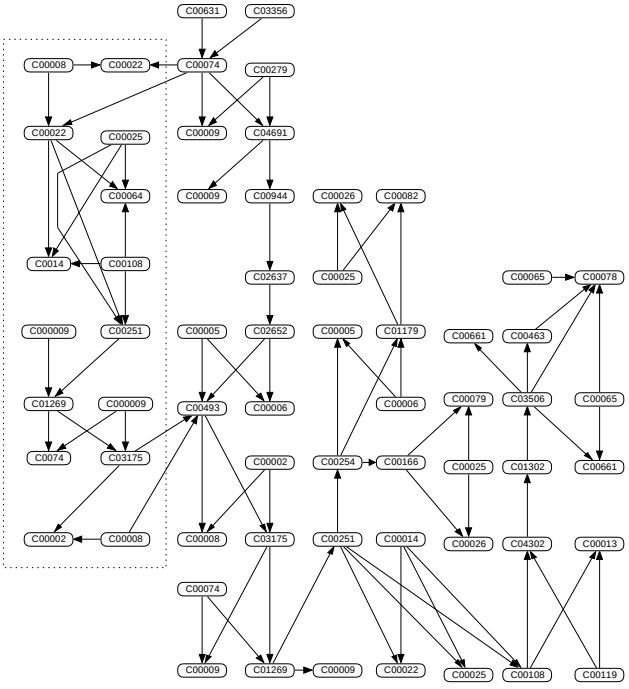
Figure 8.a shows the Bayesian model for the aromatic amino acid pathway of yeast used in this experiment. In this experiment we also examine the effect of introducing an additional branch to the pathway (Figure 8.b). This fictional pathway is the same pathway introduced as 'branching pathway' in [Angelopoulos and Muggleton, 2002]. To examine the performance of the models shown in Figures 8.a and 8.b, we use the same experimental method described in Experiment 1. In this experiment we also measure Root Mean Square (RMS) error in addition to predictive error defined before. The reason for using this measure is that this was used to evaluate the SLP model, and again we intend to compare the results. Root Mean Square (RMS) error is defined in Equation 2.

$$RMS = \sqrt{\frac{\sum_{i=1}^{N}(p_i - \hat{p}_i)^2}{N}} \qquad (2)$$

In this equation p_i are true parameters and $\hat{p}_i$ are estimated or learned parameters. In SLP these parameters correspond to the labels of stochastic clauses (see figure 4). In a Bayesian model, however, these parameters correspond to the probabilities in CPTs. Thus, the number of parameters in the Bayesian model is more than the number of parameters in the SLP model, though we have



(a)



(b)

Figure 8: a) A Bayesian model for the aromatic amino acid pathway of yeast b) a Bayesian model for the yeast pathway with an extra branch

the same pathway and the same number of metabolites. In models in figures 8.a and 8.b there are 137 and 203 parameters respectively, compared to 21 and 26 in the corresponding SLP models.

Results and discussion

The results of the experiment are shown in figure 9. Graphs 9.a and 9.b compare the predictive error and RMS error respectively. According to graph 9.b the RMS error for the Bayesian model varies from around 0.25 for 10 training examples to 0.14 for 1000 training examples. This graph suggests that the RMS error of the Bayesian model is in general less than the RMS of the SLP model which varies from around 0.25 for 100 training examples to 0.18 for 1000 training examples [Angelopoulos and Muggleton, 2002]. These graphs also suggest that there is not a substantial decrease in efficiency for the Bayesian model with an additional branch, whereas the results from the SLP model show a significant quantitative difference between branching and non-branching models (for the branching model, RMS varies between around 0.29 for 100 training examples to 0.22 for 1000 training examples). A better performance of the Bayesian learning algorithm, especially in the branching model, could be related to the way these algorithms deal with multi-branch situations. Learning techniques for Bayesian networks which are a kind of graphical representation are naturally suitable to cope with multiple branches, whereas in SLP this corresponds to non-determinism which is not perfectly addressed in the learning algorithms for SLP. This preliminary conjecture, however, requires more investigation.

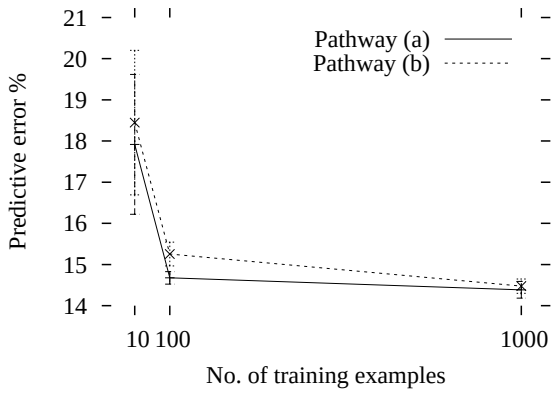
5 Conclusions and further work

In this paper we introduced a framework for modelling metabolic pathways using Bayesian belief networks. This framework was used to develop a Bayesian model of the aromatic amino acid pathway of yeast.

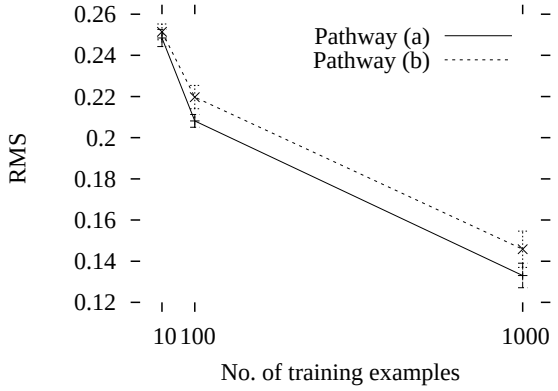
Preliminary results suggest that in parameter estimation from data, the Bayesian model for the yeast metabolic pathway outperforms an SLP model for the same pathway. These results also show that unlike the SLP model, introducing an additional pathway within the Bayesian model does not result in a significant quantitative difference.

Even though Bayesian networks have shown a better performance for capturing probabilistic information, they are weaker than logic-based methods (e.g. SLP) in representing structural information involved in metabolic reactions. For example logic-based representations allow detailed encoding of physical and structural properties as well as chemical reactions associated with the metabolites together with an encoding of the metabolic network. These models can be machine learned from online databases of metabolic networks (e.g. KEGG) which are now publicly available.

In future work we intend to develop a hybrid framework for learning metabolic pathways which involves



(a) Predictive Error%



(b) Root Mean Square (RMS) error

Figure 9: (a) Predictive Error and (b) RMS for the Bayesian models of Figure 8.

Bayesian parameter learning and a logic-based representation and inference.

Acknowledgement

We would like to thank R.D. King and C.H. Bryant for the information on metabolic pathways and yeast metabolism. We would also like to thank N. Angelopoulos and J. Cussens for useful discussions on SLPs and Bayesian networks. This work was supported by the DTI project "MetaLog - Integrated Machine Learning of Metabolic Networks" (grant reference: QCBB/C/012/00003)

References

[Angelopoulos and Muggleton, 2002] Nicos Angelopoulos and Stephen Muggleton. Machine learning metabolic pathway descriptions using a probabilistic relational representation. In *Machine Intelligence 19*, September 18-20 2002. Presented. Also submitted to ETAI for referred post-proceedings publication.

[Bryant *et al.*, 2001] C.H. Bryant, S.H. Muggleton, S.G. Oliver, D.B. Kell, P. Reiser, and R.D. King. Combining inductive logic programming, active learning and robotics to discover the function of genes. *Electronic Transactions in Artificial Intelligence*, 5-B1(012):1–36, November 2001.

[Consortium, 2001] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 15th February 2001.

[Cussens, 2001] J. Cussens. Parameter estimation in stochastic logic programs. *Machine Learning*, 44(3):245–271, 2001.

[Durbin *et al.*, 1998] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.

[Friedman and Yakhini, 1996] Nir Friedman and Zohar Yakhini. On the sample complexity of learning bayesian networks. In *Uncertainty in Artificial Intelligence: Proceedings of the Twelfeth Conference (UAI-1996)*, pages 274–282, San Francisco, CA, 1996. Morgan Kaufmann Publishers.

[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. *Journal of Computational Biology*, 7:601–620, 2000.

[Goto *et al.*, 2000] S. Goto, T. Nishioka, and M. Kanehisa. Ligand: chemical database of enzyme reactions. *Nucleic Acids Research*, 28:380–382, 2000.

[Heckerman, 1995] D. Heckerman. A tutorial on learning with bayesian networks. Technical report, Microsoft Research, Redmond, Washington, 1995.

[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*, volume 7, pages 175–186, 2002.

[IUBMB, 1992] IUBMB. *Enzyme Nomenclature: Recommendations (1992) of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology*. Academic Press, New York, 1992.

[Muggleton, 1999] S. Muggleton. Stochastic logic programs. *Journal of Logic Programming*, 1999.

[Muggleton, 2002] S. Muggleton. Learning structure and parameters of stochastic logic programs. *Electronic Transactions in Artificial Intelligence*, 6, 2002.

[Oliver, 1996] S.G. Oliver. From DNA sequence to biological function. *Nature*, 379:597–600, 1996.

[Pearl and Verma, 1991] Judea Pearl and Tom S. Verma. A theory of inferred causation. In James F. Allen, Richard Fikes, and Erik Sandewall, editors, *KR'91: Principles of Knowledge Representation and Reasoning*, pages 441–452, San Mateo, California, 1991. Morgan Kaufmann.

[Pearl, 2000] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

[Reiser *et al.*, 2001] P.G.K. Reiser, R.D. King, D.B. Kell, S.H. Muggleton, C.H. Bryant, and S.G. Oliver. Developing a logical model of yeast metabolism. *Electronic Transactions in Artificial Intelligence*, 5-B2(024):223–244, November 2001.

[Spellman *et al.*, 1998] P.T. Spellman, G. Sherlock, MQ. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, 9:3273–3297, 1998.

[Venter *et al.*, 2001] J.C. Venter, M.D. Adams, and E.W. Myers et al. The sequence of the human genome. *Science*, 291(5507):1304–1351, 16th February 2001.