# Two-Stage Bayesian Networks for Metabolic Network Prediction

**Jung-Wook Bang[1]* Raphael Chaleil[2]* and Jon Williamson[3]***

[1]Computational Bioinformatics Group, Imperial College London, UK
[2]Structural Bioinformatics Group, Imperial College London, UK
[3]Dept. of Philosophy, King's College London, UK

## Abstract

Metabolism is a set of chemical reactions, used by living organisms to process chemical compounds in order to take energy and eliminate toxic compounds, for example. Its processes are referred as metabolic pathways. Understanding metabolism is imperative to biology, toxicology and medicine, but the number and complexity of metabolic pathways makes this a difficult task. In our paper, we investigate the use of causal Bayesian networks to model the pathways of yeast *saccharomyces cerevisiae* metabolism: such a network can be used to draw predictions about the levels of metabolites and enzymes in a particular specimen. We, propose a *two-stage methodology* for causal networks, as follows. First construct a causal network from the network of metabolic pathways. The viability of this causal network depends on the validity of the causal Markov condition. If this condition fails, however, the principle of the common cause motivates the addition of a new causal arrow or a new `hidden' common cause to the network (stage 2 of the model formation process). Algorithms for adding arrows or hidden nodes have been developed separately in a number of papers, and in this paper we combine them, showing how the resulting procedure can be applied to the metabolic pathway problem. Our general approach was tested on neural cell morphology data and demonstrated noticeable improvements in both prediction and network accuracy.

## 1 Introduction

Functional genomics is the search for understanding of the functionality of specific genes, their relations to diseases, their associated proteins and their roles in biological processes. In functional genomics, a cell can be seen as a biochemical machine that consumes simple molecules to generate more complex ones by chaining together biochemical reactions into long sequences; its processes are referred to as *metabolic pathways*[4]. Genes play an essential role in these networks by providing the information to synthesize the enzymes that catalyze biochemical reactions. Understanding metabolism is an important problem for biology, pharmacology (in particular toxicology) and medicine but the size, complexity and uncertainty of the network of pathways has made this task difficult.

Lately, Friedman *et al.* used Bayesian nets to model gene expression data [9,10] and justified their use by their rich graphical and probabilistic representation of gene expression data and their ability to explain relations among gene variables. They reported many biologically plausible conclusions from real expression data of Spellman *et al.* by deploying heuristic search algorithms and statistical confidence measurements [17]. They proposed adopting continuous variables to capture precise local probabilities and improving the heuristic search algorithm as topics for further study. Imoto *et al.* applied a non-parametric regression model in Bayesian networks for constructing genetic networks from gene expression data [12]. They claimed to have success in microarray gene expression data and generalized method to deal with more general cases in the future.

In this paper, we demonstrate how causal networks can be used to model and predict yeast metabolism whose pathways essentially form a causal graph, one component of a causal net. Causal nets depend on the causal Markov condition as a primitive assumption and we propose a *two-stage methodology* to deal with any failure of the condition. First construct a causal net from the net of metabolic pathways; second alter that net to ensure the causal Markov condition is satisfied by adding new causal arrows or new `hidden' common causes. In section 2, we illustrate issues in metabolic pathway in yeast described in KEGG. In section 3, we review causal Bayesian networks and present a causal Bayesian network modeling an aromatic amino acid pathway of yeast *saccharomyces cerevisiae.* In section 4 and 5, we discuss the case where the causal Markov condition fails and propose our two-stage method to deal with the problem. In section 6, we illustrate the effectiveness of adding hidden nodes in a real biological domain. In section 7 and 8, we discuss our approach and issues to be studied in the near future.

* Co-authors in alphabetical order

## 2 Metabolic Networks

Metabolism is a set of chemical reactions, used by living organisms to process chemical compounds in order to take energy, extract building blocks and eliminate toxic compounds. Most of these reactions would not be executed without specialized proteins called enzymes, whose function is to catalyze these chemical reactions.

Metabolism was previously seen as a combination of distinct pathways, such as glycolysis, citrate cycle, urea cycle, amino acid biosynthesis and many others. All these pathways are connected to each other. In recent years, metabolism has started to be studied in a network approach. Information about the structure of metabolic networks can now be partially extracted and represented in graphical form using KEGG, WIT and MetaCyc. For example, the data of the Kyoto Encyclopaedia of gene and Genomes (KEGG) consists of information on interacting molecular and gene pathways. Related to KEGG are the Biochemical Pathways (BP) index of Boehringer Mannheim and the Encyclopaedia of E. Coli Genes and Metabolism (EcoCyc).

Enzymes are proteins encoded by genes and these genes can be expressed at will. Therefore some subgraphs of the network or pathways can be activated or inactivated. In prokaryotic cells, a whole subgraph can be activated or inactivated at once, this is the notion of operons; and in eukaryotic cells which do not have operons, genes can be controlled individually allowing an even thinner regulation. The whole network is very dynamic, responding to the environment and the cell's needs, and some parts of the network can be activated in mutually exclusive or inclusive way. If we look at the level of a single biochemical reaction in the network, its activation depends on the presence of the reaction substrates, therefore depends on the previous step. It also depends on whether the gene coding for the enzyme is activated, and of course what activates the gene, which can be one of the substrates or some external stimuli.

## 3 Constructing Causal Networks

A Bayesian network is a tool for representing a probability function. It is defined over a finite domain V of variables, each of which may be discrete or continuous - for ease of exposition we will restrict attention to discrete variables which take a finite number of values. A Bayesian network consists of a DAG $G$ whose nodes are the variables in $V$; a probability specification $S$ which contains the probability distribution $p(V_i|Par_i)$ of each variable $V_i \in V$ conditional on its parents $Par_i$ in $G$; and an assumption, called the *Markov condition*, which states that each variable $V_i \in V$ is probabilistically independent of its non-descendants, $ND_i$, conditional on its parents, written $V_i \perp ND_i \mid Par_i$.

A *causally interpreted Bayesian network*, or *causal network*, is a Bayesian network in which the graph $G$ represents the causal relations amongst the variables in $V$, with an arrow from $V_i$ to $V_j$ if $V_i$ is a direct cause of $V_j$ [14]. There are arguments to the effect that, if a Bayesian network is causally interpreted then the Markov condition - now called the *causal Markov condition* - is a valid assumption: *i.e.* that any variable is probabilistically independent of its non-effects, conditional on its direct causes [15, 20]. Thus in cases where causal relations are known, the graphical structure $G$ in a Bayesian network can just be taken to be the causal graph. This reduces the problem of constructing a Bayesian network to that of determining the probability specification $S$, and this is normally done by taking the corresponding sample frequencies from a database of past case data, or by eliciting degrees of belief from experts.

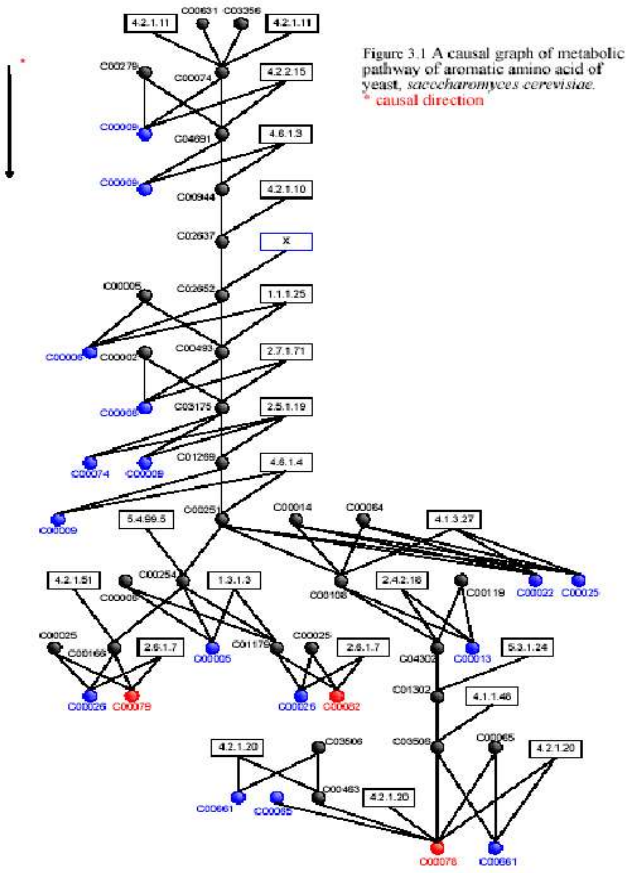


Figure 3.1 A causal graph of metabolic pathway of aromatic amino acid of yeast, *saccecharomyces cerevisiae*.
* causal direction

In the case of yeast *saccharomyces cerevisiae* metabolism the network of aromatic amino acid pathways takes the form of a causal graph (Figure 3.1). The direction of causation (*i.e.* the direction of the reaction) is from the top to the bottom of the diagram. The rectangular nodes are enzymes and the circular nodes are metabolites. The red circular nodes are the aromatic amino acids - phenylalanine (C00079), tyrosine (C00082) and tryptophan (C00078) whose values we are interested in predicting. The values that the variables take are their concentrations (by mass). C00079 is produced by C00166 and C00025 under enzyme 2.6.1.7. There is a period of flux before each reaction settles down to an equilibrium during which the reaction often takes place in both directions. There is inevitably a single overall direction to the reaction however, which is determined after

equilibrium is reached. To complete the causal network all that remains is to add probability specifiers, i.e. the probability distribution of each node conditional on its direct causes. In KEGG, the metabolic pathways represent all known pathways in a given organism. However, when, in some case, enzymes could not be located, a simple deductive rule was used to uncover alternative reaction paths from an initial substrate and a final product [11]. However, Goto describes difficulties in path computation from a given list of enzymes - there still exists a number of unknown pathways for secondary metabolisms and metabolisms that are revealed under stressful conditions.

## 4 Two-Stage Methodology

As mentioned above, taking the graph of a Bayesian network to be the causal graph can simplify the problem of network construction. However there is a potential difficulty with this strategy: the causal Markov condition, which is required to hold if the causal network is to coincide with physical probability (frequency, propensity, chance), may in fact fail. There are a number of ways in which the causal Markov condition may fail [18]: 1. Causal information may be missing: some causal relationships amongst the variables may simply not be known; some common causes of variables in $V$ may be omitted from $V$. 2. Probability specifiers may be poor estimates of physical probabilities: if specifiers were determined from a database of past case data there may be too little data to determine the required probabilities accurately, or the database may represent a biased sample from the population at large; if specifiers be elicited from experts, the experts' degrees of belief may poorly reflect physical probabilities. 3. Probabilistic dependencies which contradict the causal Markov condition may be induced by non-causal relationships amongst the variables: variables may have overlapping meaning, they may be logically or mathematically related, they may be related by non-causal physical laws or by problem constraints, or they may be subject to accidental correlations. Although the causal Markov condition may fail, it remains a good default assumption, in the following sense. If an agent's background knowledge consists just of the two components of a causal network, a causal graph and the associated probability specifiers, then the agent's personal probabilities (her degrees of belief) ought to satisfy the causal Markov condition [20, 21]. Thus the causal network is the best model available given just causal knowledge and knowledge of the conditional probability distribution of each variable conditional on its parents.

This suggests a *two-stage methodology* for employing Bayesian networks: Stage One: Construct a causal network from causal knowledge and corresponding probability specifiers. This is a good default model. Stage Two: If the network fails to perform well (this is indicative of failure of the causal Markov condition), modify the network so that it better approximates physical probability.

## 5 Network Modification

If a Bayesian network is to be modified to better represent physical probability, one can either change its graphical structure, or its probability specifiers or both. Changing probability specifiers to better approximate the corresponding physical probabilities is a statistical problem. We shall assume here that this problem is solvable - i.e. a mechanism is available for determining physical probabilities - and focus our attention on the graphical problem. Two routes are available for changing the graph in a Bayesian network: one can change the nodes in the graph (add, delete or combine nodes) or change the arrows in the graph (add, delete or re-orient arrows) or both. We shall present an example of each strategy: adding arrows and adding hidden nodes.

### 5.1 Adding Arrows

The adding-arrows approach is conceptually very simple. If one adds an arrow from $V_i$ to $V_j$ in a Bayesian network (and change the corresponding probability specifiers accordingly) then the new network will be no worse an approximation to physical probability than the old network, and will be a closer approximation if and only if $V_i$ and $V_j$ are probabilistically dependent conditional on the other direct causes of $V_j$[18]. So a simple strategy for changing a network to better approximate physical probability is to add arrows corresponding to conditional dependencies. If at each stage one adds the arrow corresponding to strongest conditional dependence then one achieves the closest approximation at each stage. Moreover this simple greedy algorithm finds networks that are close to the global best approximation [18,21,22].

### 5.2 Symmetric Hidden Node Method

The hidden node approach was originally proposed by Pearl and Verma [13]. Whenever two nodes $B$ & $C$ with no arrow between them are probabilistically dependent conditional on a common parent $A$ (a violation of the causal Markov condition), then a 'hidden node' $H$ is added as a new parent of $B$ and $C$, with the arrows from $A$ to $B$ and $C$ redirected through $H$. Then probability specifiers for $H$, $B$ and $C$ will be learned from data using the symmetric propagation algorithm called *Symmetric Hidden Node Method* (SHNM) [1]. In neural cell morphology, SHNM improved the prediction accuracy in Bayesian networks up to 42% (from 59% to 84%) [2,3]. Comparative analysis on other machine learning techniques also showed the strength of SHNM. These included neural networks and C4.5. The neural networks had one hidden layer (with up to 5 hidden nodes). The number of learning cycles was in the range 10,000 to 500,000, compared to 800 cycles for learning neural networks and Bayesian networks, respectively. The C4.5 weights were set in the range 2 to 4. It showed that the C4.5 method gave a comparable performance to the naive Bayesian network, but neural networks were considerably worse in this case. This paper also details how to systematically identify the place to add a hidden node,

using a conditional dependency measure to test for violations of the Markov condition.

### 5.3 A Combined Approach

In this paper we advocate a combination of these two strategies for network modification. According to the principle of the common cause a probabilistic dependency which violates the causal Markov condition indicates that either a direct causal relation between the dependent nodes, or a common cause of the two nodes, is missing from the causal graph [18]. Thus to generate a causal network that satisfies the causal Markov condition we need the flexibility to add either a new arrow or a new common cause (a hidden node). The new graph can be treated as a new causal hypothesis, and can motivate closer scrutiny to verify the new posited causal connections [20].

In deciding whether to add an arrow or add a hidden node to modify a network, there are two key considerations to take into account. First the new network should be plausible when construed as a hypothesis about causal relations. Thus if it is implausible that two dependent variables are directly causally related, one ought not add an arrow between them - one ought to add a hidden node (interpreted as a common cause) to account for their dependency. Second, (Occam's razor) one ought to pursue the option that, other things being equal, increases the complexity of the network least. The complexity of a network can be measured in terms of the number of probability specifiers required in the network. In most situations adding an arrow will increase complexity least, but in cases where two or more nodes share a large number of parents, adding a hidden node can even decrease complexity.

## 6 Application To Yeast Metabolism

Our methodology for network modification has two objectives: Firstly, to significantly improve the prediction accuracy of the causal network. Secondly, to suggest new common causes (chemical reactions) and causal relations (reaction pathways between substrates and products) in the network. Our algorithm is as follows

1. For each pair of variables (substrates and products) $B$ and $C$ in the network, check their probabilistic dependence conditional on the parents $A$ of $C$, via the mutual information formula

$$\mathrm{MI}_c(\mathrm{B},\mathrm{C}\mid \mathrm{A}) = \sum_{\mathrm{B},\mathrm{C}}\left[\mathrm{P}(\mathrm{b},\mathrm{c}\mid \mathrm{a})\log\frac{\mathrm{P}(\mathrm{b},\mathrm{c}\mid \mathrm{a})}{\mathrm{P}(\mathrm{b}\mid \mathrm{a})\mathrm{P}(\mathrm{c}\mid \mathrm{a})}\right]$$

2. If there are any such dependencies then the causal Markov condition has failed. Choose the maximal dependency and generate two new causal hypotheses: one by adding a hidden node and the other by adding an arrow from $B$ to $C$.

3. The corresponding probability specifiers need to be learned. In the case of the model with added arrow these can be equated with the corresponding frequencies in the database. In the case of the model with hidden node, execute a learning process in that particular local structure [1].

4. Try to find the referent of a hidden node $H$ as follows. Regenerate a data set for $H$ relevant to the original data set. Using a pattern matching techniques with correlation calculation, try to locate a variable with unknown functionality that scores the highest mark.

5. Try to verify a new arrow by checking that it corresponds to probability raising of the effect by the cause, conditional on the effects other direct causes. If so, check that intervening to fix the value of the cause fixes the values of the effect, controlling for the effect's other direct causes.

6. Eliminate a hypothesis if it has no plausible causal interpretation. If both hypotheses remain, then eliminate the most complex hypothesis. Proceed to step 1.

Consider the following example. Suppose enzyme $e$ catalyzes a chemical reaction with substrate $m_i$ and product $m_j$ where $i = 3$ and $j = 3$. Figure 6.1a shows an example of a highly connected causal graph in a single chemical reaction. We start by examining conditional dependency between the substrates and products to identify the location by deploying systematic search (step 1). If a dependency is found, we spawn new hypotheses, test their causal interpretation and eliminate one (steps 2-6). We see how adding a hidden node (Figure 6.1c) can in some cases offer a hypothesis of lower complexity than that generated by adding arrows (Figure 6.1b). A mixed approach (Figure 6.1d) is likely to result however.

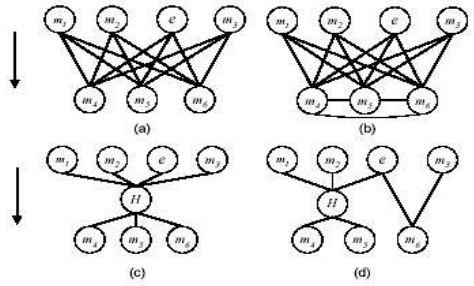


Figure 6.1 A possible network scheme between substrates, enzyme and products. (a) Original network (b) Adding arrows (c) Adding a hidden node (d) Adding a limited hidden node * causal direction (top to button)

## 7 Conclusion

In this paper we have shown how causal networks can be used to model and predict yeast metabolism. A network of metabolic pathways is essentially a causal graph. By augmenting this causal graph with probability specifiers (the probability distribution of each variable conditional on its direct causes) we construct a causal network. The viability of a causal network depends on the validity of the causal Markov condition. If this condition fails, the principle of the common cause motivates the addition of a new causal arrow or a new `hidden' common cause to the network. Algorithms for adding hidden nodes and adding arrows have been developed separately in a number of papers, and in this paper we combine them, showing how the resulting procedure can be applied to the metabolic pathway problem.

The next step in this line of research is clearly to test the resulting methodology on real yeast metabolism data. We plan

to use data from Biochemistry group (Prof. Jeremy K. Nicholson) at Imperial College. Having developed one or more causal networks which model the data well, we intend to examine their plausibility as causal hypotheses. i.e. we intend to see whether new common causes and causal connections that have been posited in these models really do correspond to causes and causal connections. This will be done by collecting further observational and experimental data, to see whether each new posited cause raises the probability of its effects and whether intervening to change the value of causes changes the values of their effects.

In sum, our two-stage methodology for causal networks has a double goal: to model probabilistic relations amongst the variables involved in the metabolic pathways and to model the causal connections amongst these variables. In this paper we present the biological problem and modeling methodology; in future papers we intend to assess our proposed solution.

## Acknowledgments

The authors are grateful to Mike Sternberg and Jeremy K. Nicholson at Imperial College. This paper was partially supported by DTI project (Metalog Project), UK.

## References

[1] J-W Bang and D. Gillies. Estimating Hidden nodes in Bayesian Networks. *Proceedings of Int'l Conference on Machine Learning and Applications*, Las Vegas, USA, 2002.

[2] J-W Bang and D. Gillies. Using Bayesian Networks with Hidden Nodes to Recognize Neural cell Morphology. *In Proceedings of the Seventh Pacific Rim Int'l Conference in Artificial Intelligence*, LNAI, Springer-Verlag, Tokyo, Japan, 2002.

[3] J-W Bang, Alexandros Pappas and Duncan Gillies, "Interpretation of Hidden Node Methodology with Network Accuracy." Technical Report: ISSN 1469-4166, Dept. of Computing, Imperial College, London, UK, 2003.

[4] C. Bryant, S. Muggleton, S.G. Oliver, D.B. Kell, P. Reiser, and R.D. King. *ElectronicTransactions on Artificial Intelligence*, 5(B1):1-36, 2001.

[5] D. Chickering, D. Geiger & D. Heckerman: 'Learning Bayesian networks is NP-hard', Technical Report MSR-TR-94-17, Microsoft Research, November 1994.

[6] G.F. Cooper: 'The computational complexity of probabilistic inference using Bayesian belief networks', Artificial Intelligence 42, pages 393-405, 1990.

[7] David Corfield & Jon Williamson(eds.): 'Foundations of Bayesianism', Kluwer Applied Logic Series, Dordrecht: Kluwer Academic Publishers, 2001.

[8] A. Flook. The use of Dilation Logic on the Quantimet to Achieve Fractal Dimension Characterization of Textured and Structured Profiles. Powder Technology, 21, 195-198, 1978.

[9] N. Friedman, K. Murphy and S. Russell. Learning the structure of dynamic probabilistic networks. In Proc. Fourteenth Conference on Uncertainty in Artificial Intelligence, 1998.

[10] N. Friedman, M. Linial, I. Nachman, D. Pe'er, 'Using Bayesian networks to analyze expression data, *Journal of Computational Biology*, 7:601—620, 2000.

[11] S. Goto, H. Bono, H. Ogata, W. Fujibuchi, T. Nishioka, & K. Sato: 'Organizing and Computing Metabolic Pathway Data in Terms of Binary Relations', Pacific Symposium Biocomputing, 2, 175-186, 1996.

[12] S. Imoto, T. Goto and S. Miyano. Estimation of genetic networks and functional structures between genes by using Bayesian network and nonparametric regression', Proc. Pacific Symposium on Biocomputing. 7, 175-186, 2002.

[13] J. Pearl and T. Verma. 'Statistical Semantics for Causation, *Statistics and Computing*, 2, Chapman and Hall, 1993.

[14] J. Pearl: 'Probabilistic reasoning in intelligent systems: networks of plausible inference', San Mateo, CA: Morgan Kaufmann, 1988.

[15] J. Pearl: 'Causality: models, reasoning, and inference', Cambridge University Press, 2000.

[16] D. Sholl. Dendritic Organization in the Neurons of the Visual and Motor cortices of the Cat. *Journal of Anatomy*, 87, 387-406, 1953.

[17] Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., Brown, P., Botstein, D. & Futcher, B. `Comprehensive identification of cell cycle-regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization', *Molecular Biology of the Cell,* 9, 3273–3297, 1998.

[18] Jon Williamson: 'Foundations for Bayesian networks', in [7], pages 75-115, 2001.

[19] Jon Williamson: 'Maximizing Entropy Efficiently', Electronic Transactions in Artificial Intelligence 6, www.e-taij.org, 2002.

[20] Jon Williamson: 'Learning causal relationships', Technical Report 02/02, LSE Centre for Natural and Social Sciences, www.lse.ac.uk/Depts/cpnss/proj_causality.htm, 2002.

[21] Jon Williamson: `Approximating discrete probability distributions with Bayesian networks', in Proceedings of the International Conference on Artificial Intelligence in Science and Technology, Hobart Tasmania, 16-20 December 2000, pp. 106-114.

[22] Jon Williamson: `A probabilistic approach to diagnosis', Proceedings of the Eleventh International Workshop on Principles of Diagnosis (DX-00), Morelia, Michoacen, Mexico, June 8-11 2000.