BIOINFORMATICS

Vol. 18 Suppl. 1 2002
Pages S241–S248

Modelling regulatory pathways in E. coli from time series expression profiles

Irene M. Ong [1], Jeremy D. Glasner [3] and David Page [1, 2]

[1]Department of Computer Sciences, [2]Department of Biostatistics & Medical Informatics and [3]Department of Genetics, University of Wisconsin, Madison, 53706, USA

Received on January 24, 2002; revised and accepted on April 1, 2002

ABSTRACT

Motivation: Cells continuously reprogram their gene expression network as they move through the cell cycle or sense changes in their environment. In order to understand the regulation of cells, time series expression profiles provide a more complete picture than single time point expression profiles. Few analysis techniques, however, are well suited to modelling such time series data.

Results: We describe an approach that naturally handles time series data with the capabilities of modelling causality, feedback loops, and environmental or hidden variables using a Dynamic Bayesian network. We also present a novel way of combining prior biological knowledge and current observations to improve the quality of analysis and to model interactions between sets of genes rather than individual genes. Our approach is evaluated on time series expression data measured in response to physiological changes that affect tryptophan metabolism in *E. coli*. Results indicate that this approach is capable of finding correlations between sets of related genes.

Contact: ong@cs.wisc.edu

Keywords: Dynamic Bayesian networks; regulatory pathways; time series gene expression; operon model.

INTRODUCTION

Living cells contain thousands of genes, each of which codes for one or more proteins. Many of these proteins in turn regulate expression of genes through complex regulatory pathways to accommodate changes in their environment or carry out the organism's developmental program. The key to understanding living processes is uncovering this genome-wide circuitry that underlies the regulation of cells.

Genome-wide DNA microarrays are a powerful tool, providing a glimpse of the signals and interactions within regulatory pathways of the cell. They enable the simultaneous measurement of mRNA abundance of most if not all identified genes in a genome under normal conditions or under various treatments or perturbations. A drawback of the current technology of DNA microarrays is that low mRNA expression levels can be very hard to detect. Additionally, steady state or single time point expression profiles do not allow us to discover sequences of regulatory events. The former problem can be controlled in some genes by the use of biological knowledge in the analysis. The latter problem can be alleviated by performing time series experiments using DNA microarrays, which will provide a better picture of the signals and interactions over time[†].

Friedman *et al.* (2000) were the first to address the task of determining properties of the transcriptional program of an organism (Baker's yeast) by using Bayesian networks to analyse gene expression data. Their method can represent the dependence between interacting genes, but it does not show how genes regulate each other over time in the complex workings of genetic networks. Analysis of time series data potentially allows us to determine regulatory pathways rather than just associating genes that are co-regulated together.

In certain organisms such as *Escherichia coli*, there are many sets of genes that are already known to be transcribed together and hence strongly co-regulated. These sequences of genes that are transcribed together into mRNA on their way to being expressed as proteins are known as operons. In this case, since we already know or can predict which genes are regulated together, it would be ideal to use this knowledge in the analysis technique rather than relearn it.

Using a Dynamic Bayesian network (DBN), a close relative of Bayesian network (BN), has several advantages. In addition to being well suited to handling time series data, this framework can handle missing data in a principled way as well as model stochasticity, prior knowledge and hidden variables (Murphy and Mian, 1999). To our knowledge Friedman *et al.* (1998) and Murphy and

[†]Time series experiments will provide a better understanding of the interactions within the cell provided the time steps are taken within intervals appropriate for capturing important molecular activity.

Mian (1999) are to be credited with first proposing the suitability of DBNs for modelling time series gene expression microarray data. The primary contribution of our paper is to test this DBN approach on real time series microarray data. A secondary contribution is the incorporation of the results of a previous application of Bayesian inference (naïve Bayes) as background knowledge for this new application. This naïve Bayes approach used a variety of evidence sources, including earlier microarray data from the Blattner Laboratory at the University of Wisconsin, to predict the operons in *E. coli*. The goal of that work was to produce an accurate operon map that could be used subsequently in the prediction of regulatory pathways in *E. coli*.

The present paper describes a next step in this direction. The focus of this paper addresses how one could approximately model the interactions of sets of genes automatically using prior biological knowledge and time series expression profiles. Does the use of prior knowledge or the use of time series expression profiles help determine broader correlations? Can we learn hierarchical connections between sets of co-regulated genes, and ultimately learn connections among multiple signal transduction pathways?

We introduce an approach to determining transcriptional regulatory pathways by applying the Dynamic Bayesian network to time series gene expression data from DNA microarray hybridization experiments. Our approach involves building an initial DBN structure that exploits biological knowledge of operons and their associated genes. We further use a domain expert's best guess to initialize the probabilities of how the state of an operon might affect the genes within that operon or that of another operon.

We evaluate our approach using a recent study by Khodursky *et al.* (2000), who performed time series experiments to analyse gene expression in response to physiological changes that affect tryptophan metabolism in *E. coli*.

MATERIALS: DATA AND SOFTWARE

To test our hypotheses, this paper reports the analysis of time series gene expression data from Khodursky *et al.* (2000). This data set is used because it is focused on tryptophan metabolism, a well studied regulatory process, making it an excellent check for the reverse engineering of a genetic network. Our eventual goal is to develop a tool for analysing larger collections of time series expression data on *E. coli*.

It should be noted that a common problem with current microarray expression data is a small number of data points and a large number of features. This is especially true of time series data. The present data set consists of 12 data points, from 4 time steps under tryptophan-rich conditions and 2 sets of 4 time steps under tryptophan-starved conditions. These data points consist of 169 genes that were selected based on their expression levels by Khodursky *et al.* (2000), and were the only data made available at the start of the present study. Nevertheless it is hoped that discretization and reasonable priors will partially offset noise and permit useful results to be obtained. All the data for both conditions are used (except where indicated) in each of the experiments below to learn how the different environmental conditions affect the regulatory pathway.

The Bayes Net Toolbox (Murphy, 2001) software package written by Kevin Murphy was used for the experiments in this paper because it already provided the necessary functionalities for building Bayesian networks, as well as an implementation of Expectation Maximization (EM) for learning the conditional probability tables. We constructed the initial BN structure and learned the parameters of the model using the methods provided in Bayes Net Toolbox. Within this framework we implemented the structure search described in the Section Structure learning.

Our operon map includes both known operons from Salgado *et al.* (1999) and predicted operons from Craven *et al.* (2000). The latter work maps every known and putative gene in the *E. coli* genome into its most probable operon. This map makes the simplifying assumption (rarely but occasionally violated in reality) that every gene appears in exactly one operon. The accuracy of the map (percentage of genes placed in the correct operon) is estimated at about 95% using 10-fold cross-validation.

We assume that this operon map is correct and use it to build our initial BN and DBN structures[‡]. Furthermore, the initial probabilities used in our BN and DBN structures are Dirichlet priors obtained from a domain expert. The initial probabilities for DBN could be dependent on the nature of the experiments and the amount of elapsed time between time points[§]. However, our only insight into these dependencies was gleaned largely from looking at the data; hence, we did not encode these insights into our DBN to avoid biasing our results.

The evidence variables in our BN and DBN are the discretized gene expression levels from the experiments with excess tryptophan and tryptophan starvation. We define up regulated (↑) or down regulated (↓) as the possible discrete values to avoid choosing arbitrary thresholds. In particular, we compare the expression levels between two consecutive time series measurements to determine whether there was an increase or decrease in expression levels. Note that we are determining the relative change in expression from one time step to another (even for the

[‡] The full operon map, with an interactive graphical interface, is available online at http://apps.biostat.wisc.edu/~ml-group/GenomeViewerButton.html.
[§] If time points are not equally spaced, we may want to initialize the DBN with different probabilities across time points.

non time series BN model) rather than absolute absent or present calls.

DYNAMIC BAYESIAN NETWORK

Modelling relationships among genes

The task of automatically discovering a model that represents relationships among genes from noisy expression data involves a significant amount of uncertainty. The entire experimental process allows for the introduction of uncountable variables as well as measurement errors. Also, the fact that we can only partially observe the happenings among a collection of cells makes it impossible to construct an accurate model from expression data. Therefore it is helpful to model this uncertainty. Instead of simply stating gene A and gene B are correlated, probability provides us with a way to express how certain we are about the correlation. If evidence strongly suggests that gene A and gene B are highly correlated, (i.e., the data show that when expression level of gene A is up regulated (↑), then gene B is also up regulated) then the probability that gene B is ↑ given that gene A is ↑ would be close to 1, otherwise it would be closer to 0. This probability assignment can be denoted as P(gene B =↑ |gene A =↑) = 0.95.

A visual, intuitive and compact way of representing relationships between genes is via the use of graphical structures. We let genes A and B be represented by nodes and an arc from gene A to gene B denote that gene A influences gene B. We associate small probability tables with the nodes to summarize how gene B is affected by gene A (its parent) and how gene A is not affected by anything since it has no parent. Evidence or data for genes A and B are assigned to gene A and B's nodes, respectively, and are used to adjust the values in the local probability tables. This example of a Bayesian network is shown in Figure 1a.

In order to find the relationships among genes in our dataset, we can use the BN model to represent all the genes in our dataset. Initial or prior probability settings for the local probability tables can be uniform ($\frac{1}{n}$ where n is the number of possible values) if no prior information is known. The local probability tables can then be updated automatically based on actual counts of the data. Now we can perform a search for the most likely graph given the data.

We construct such a network, called BN_{reg}, and perform the structure search described in the Section Structure learning. Below we compare the results of this network to that of a Bayesian network with an explicit model of operons. By performing this comparison we will be able to determine whether the latter model is better at learning useful correlations among genes than the straightforward approach of BN_{reg}.

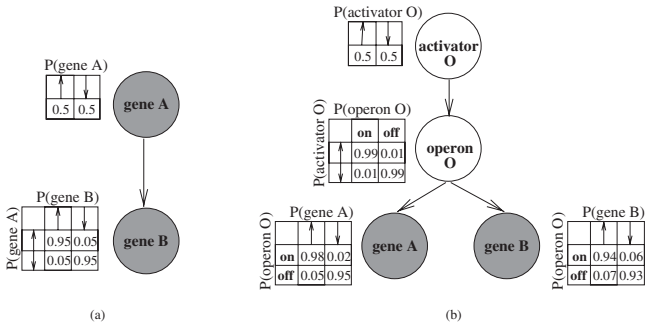

Fig. 1. (a) Example of a Bayesian network structure. ↑ represents up regulation and ↓ represents down regulation. (b) Nodes that are not shaded are hidden, i.e., nodes without observable data, and shaded nodes indicate nodes that have observable data. Hence the operon and activator nodes are hidden and the gene nodes are observed nodes.

Incorporating prior knowledge or environmental factors

There are three important reasons to incorporate explicit operon nodes into the BN model even though operon transcription levels are not observed. First, if we use nodes for genes only, and allow the learning algorithm to induce arcs between genes, it will induce many 'useless' arcs between genes in the same operon. For example, if gene A and gene B are both in operon O, then we would expect the expression level of gene A to be an excellent predictor of the expression level of gene B, but this would provide no new insight. Second, incorporation of operons in the model can help combat problems due to noise. For example, the operon that codes for tryptophan, *trpLEDCBA* (also known as *trp*), contains a leader, *trpL*, which is not detected and five genes: *trpA*, *trpB*, *trpC*, *trpD*, and *trpE*. Because of noise in microarray experiments, the measured expression level for *trpC* might be low. But the five different gene expression measurements give us essentially five independent indicators of *trp* transcription, reducing the effect of noise in the measurement of *trpC* expression. Third, by searching for interactions among operons rather than genes, we reduce the space of possible models.

Let us assume that we know gene A and gene B are co-transcribed genes in an operon, operon O, and that operon O is transcribed into mRNA when it is initiated by a molecule called an activator, activator O. The fact that gene A and gene B are in the same operon explains the high correlation between the two genes. Hence we can restructure the graph to represent this knowledge in Figure 1b. Because we cannot measure the operon or activator's expression level directly, we regard them as hidden nodes since we do not have any evidence for those nodes. If technology permitted us to obtain a measure of

the amount of activator molecule in a cell or if we knew the amount of activator molecule present in the environment, we would model the activator node as an observed node just like genes A and B. The values in the local probability tables for the hidden nodes are estimated from the graph structure, observed data and initial or prior probabilities associated with the hidden nodes.

Note that the local probability tables for gene A and gene B have changed from Figure 1a. Gene A and gene B now depend on their parent, operon O. While gene A and gene B are correlated, once we know the value of operon O (that it is activated), gene A becomes independent of gene B. Additionally, genes A and B are independent of activator O given its parent, operon O. This is known as conditional independence, which is a key property of Bayesian networks. The lack of arcs implies conditional independence, i.e., a node is independent of all non-descendent nodes in the graph given its parents. A fundamental property of BNs is that given an acyclic graph and the set of local probabilities associated with it (also known as conditional probability tables since the probabilities are conditional on the parent's values) we can determine the joint probability distribution uniquely.

We build our initial BN structure with the model of operons as described above for all the genes in our dataset from our operon map. Since an operon's transcription level affects the expression levels of the genes in that operon, we show this causality with arcs from the operon to its associated genes. Uniform priors are used for the operons, and a domain expert's best guess is used to set the informative priors for how the effect on an operon would affect the genes within that operon. The latter initial probability values, the same as those in our initial DBN model, are shown abstractly in Figure 4. The search results for this BN model with operons, BN_{op}, are compared with those of BN_{reg} in the Section Results.

Two other experiments using the BN_{op} structure are performed to determine whether separating the data based on the treatment of the cells would allow us to learn a finer structure. BN_{op_excess} uses the data under tryptophan-rich conditions, whereas BN_{op_starve} uses data under tryptophan-starved conditions.

Modelling the concept of time

Time series expression data can provide insight into causality[¶] and the regulation of cells as they change over time. Dynamic Bayesian networks gracefully scale up BNs to handle the analysis of time series data. In addition, DBNs can also model feedback loops, which are not possible for BNs due to the acyclicity constraint. To see why this is an important feature for modelling regulatory pathways, see Figure 2, which shows how some operons rely on a feedback mechanism to regulate transcription.

[¶] Causality can also be inferred by using the method proposed by Pe'er *et al.* (2001) if cells with deletions or mutations of specific genes are available.

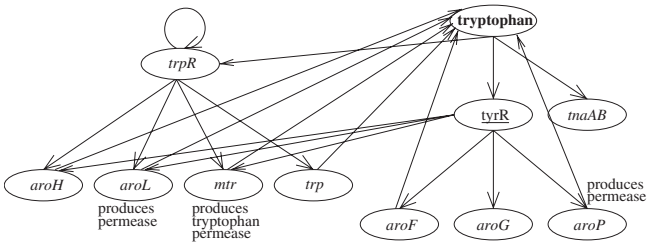

Fig. 2. A model (not a BNgraph) of how the 9 key operons (italicized) in the tryptophan regulon, groups of operons that are co-regulated, influence each other. This structure was constructed by the authors based on Khodursky *et al.* (2000). Operon names are abbreviated. The **tryptophan** node represents the molecule. tyrR is not part of the tryptophan regulon but it influences key operons within the regulon. The **tryptophan** and tyrR nodes serve to connect the interactions between the 9 key operons.

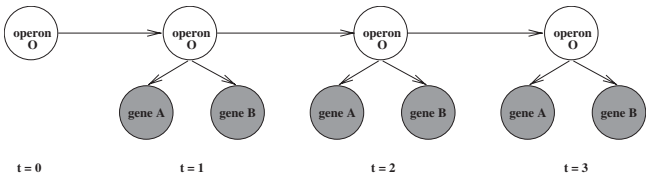

Fig. 3. An example of a Dynamic Bayesian network structure. Time slices are represented by $t = 0, t = 1, \ldots, t = n$, where n is the number of time step experiments for which there are observable data.

The Dynamic Bayesian network relies on the same properties as that of Bayesian networks with the addition of modelling genes or operons as they evolve over time. Using a simplified version of the model from Figure 1b as an example, we show in Figure 3 how the concept of time can be modelled by simply replicating the structure for each time step. The structure at time $t = 0$ is not the same as the other time slices because the purpose of operon O at $t = 0$ is to model the effect it has on operon O at $t = 1$ (to start off the chain).

Since each time slice in a DBN is identical in structure to the next, we can just replicate the structure of BN_{op} four times (our data has four time steps) to build our initial DBN structure modelling operons. This leaves undetermined the arcs from the hidden variables of one time step to the hidden variables of the next time step. Since an operon's expression level from one time step typically reflects its expression level at the next time step, we add these arcs as shown in the detailed DBN structure in Figure 5.

Note that the operons at time $t + 1$ (operon$(i)_{t+1}$, where

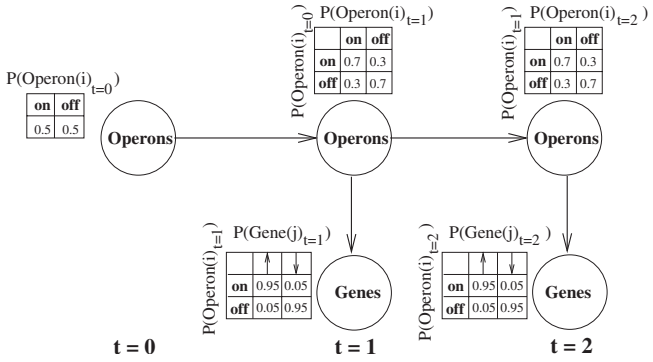

Fig. 4. Abstract Dynamic Bayesian network structure with conditional probability tables (CPTs) for each arc in the model. Time slices 3 and 4 (not shown) are identical to 1 and 2.

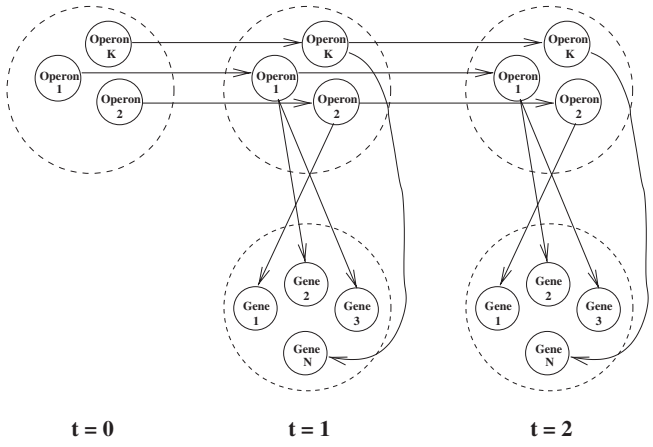

Fig. 5. Detailed view of our initial Dynamic Bayesian network structure showing intra and inter time slice connections.

i indicates the ith operon) is independent of operon$(i)_{t-1}$ given operon$(i)_t$ because of the conditional independence assumptions. This states that the future is independent of the past given the present. We can work around this assumption by adding an additional arc from operon$_{t-2}$ to operon$_t$ to indicate that the present also depends on the events from two time steps ago. However, the computational costs increase exponentially.

As before, the domain expert's best guess is used to set the initial probability values. The abstract structure of our initial DBN model, DBN$_{op}$, along with the prior probabilities are shown in Figure 4. Any additional arc among hidden nodes, as well as all posterior CPT probabilities, must be inferred from time series microarray data for *E. coli*.

Structure learning

To perform structure learning for our BN models, we focused on key genes or operons known to be affected by the absence or presence of tryptophan in the environment. For BN$_{reg}$, there are 15 such key genes. Similarly, for BN$_{op}$ we focused on the corresponding 9 key operons. For each of these genes (operons) in BN$_{reg}$ (BN$_{op}$), we consider all the other genes (operons) as possible parents. At each step, we use the Expectation Maximization (EM) algorithm to update all CPTs in the model to give a (local) maximum log likelihood. EM will infer values for the hidden variables as well as for missing observations. The log likelihood score from the previous step is used as the scoring measure to select the 20 most probable parents. This is done because EM is not guaranteed to find the maximally probable parent and because a large number of structures have the same or very close scores.

Because of limited data, we consider only simple DBN structural models in which each operon has at most two incoming arcs, from (1) the same operon at the previous time step, and (2) one other operon from the previous time step. Each operon begins with one parent—the same operon at the previous time step. In our full algorithm, for each operon we consider adding a different operon from the previous time step as a second parent. Each potential parent is considered. For each such potential second parent, the EM algorithm is employed. If any choice of second parent increases the log likelihood, then the choice that provides the highest log likelihood is selected.

In general, the preceding cycle through all the operons may need to be repeated several times for convergence to a locally optimal structure. Despite our structural restrictions the run time would take over 9 months using the junction tree algorithm as implemented in BN Toolbox. This is because of the large number of nodes (142 operons, 169 genes, for four time steps) in our DBN structure. For the long-term, we are experimenting with approximate approaches to speed up the computations. For the short-term, we focus the algorithm on the 9 key operons; the algorithm cycles once through only these, but *all* the other 141 operons are considered as potential parents. For each operon we record the best 20 choices for the second parent.

RESULTS

The results from the BN$_{reg}$ and BN$_{op}$ experiments indicate that modelling operons into the BN structure provide a more comprehensive view of the tryptophan regulon, groups of operons that are co-regulated. Without the operon structure, BN$_{reg}$ found some correlations between genes in the same operon but missed many correlations between genes in different operons.

The BN_{op} structure was able to identify correlations with direct siblings (other children of a node's parent), parents, or children for 4 of the 9 operons. *trp*'s parent, *trpR*, and siblings, *aroH* and *mtr*, were among *trp*'s 20 most probable parents. Similarly, each of *trpR*, *mtr* and *aroH* identified correlations with each other and *trpR*. The other 3 operons showed correlations with 2 of the 9 key operons.

The BN_{reg} results showed that all 15 key genes correlated with a subset of these six genes: *trpR*, *trp*, *aroH*, *mtr*, *aroF* and *aroP*. BN_{reg} found correlations between some genes but missed correlations between other genes within the same operon. All five genes in the *trp* operon were found to be probable second parents when any of the 5 were present. However, *tnaA* and *aroF* were not always found to be correlated with *tnaB* and *tyrA*, respectively.

It was interesting that the results for BN_{op_excess} showed that the operon *hisGDCBHAFI* influenced 5 of the 9 key operons. These histidine genes might be correlated to operons in the tryptophan regulon or, alternatively, the cells may have consumed all the histidine in the media resulting in histidine biosynthesis. BN_{op_excess} also showed that many of the 9 key operons were influenced by the excess tryptophan condition as all of the key operons (except *trpR*) have between 2 to 6 (which includes *trpR*) of the 9 operons as a possible 2nd parent.

We were surprised by the results from BN_{op_starve}. BN_{op_starve} showed that known operons, *hybGFEDCBA*, *artPIQMJ*, *rplK-rplA*, *fecIRABCDE*, and predicted operons, *yciGFE*, *yafDE*, *yi21-yi22*, were influenced by all of the 9 key operons. Khodursky *et al.* (2000) note that in their cluster analysis genes *yciF* and *yciG* form a tight cluster with *trpR* and related operons. They also noticed that arginine biosynthetic operons were sensitive to tryptophan changes. The *artPIQMJ* operon codes for proteins involved in the arginine transport system.

Are the other correlations meaningful? It turns out that the *rplK* ribosomal protein is involved in regulating the response to starvation for amino acids (Yang *et al.*, 2001). The *fecIRABCDE* operon is actually 2 distinct operons *fecABCDE* and *fecIR* (probably an error in the database of Salgado *et al.* (1999)). *fecABCDE* is induced under an iron limiting condition or in the presence of ferric citrate and is under the control of the regulator FUR. We are not sure why it would be involved in the tryptophan starvation response, but *aroH* is known to be more active in the presence of iron (Ray *et al.*, 1991). The *hybGFEDCBA* operon encodes hydrogenase-2, usually used under conditions of anaerobiosis (low oxygen). We do not know of a reason why the culture conditions would lead to low oxygen levels, although constant vigorous shaking of the culture during growth is important for maintaining good aerobic growth conditions. The results could be related to unknown factors in the experimental methodology of Khodursky *et al.* (2000) such as oxygen levels. *yci21* and *yci22* proteins are encoded by IS2, an insertion sequence in *E. coli*. These insertion sequences are mobile DNA elements that are often induced by growth of cells under stressful conditions. The role of the *yafD* and *yafE* proteins is not clear as they are hypothetical proteins.

BN_{op_starve} also showed that *trpR* was identified to be correlated with almost all of the 9 key operons under the tryptophan starvation condition. *mtr* and *trp* were also among the 20 most probable parents for 2 of the 9 operons. No possible correlation was found for any of the other 9 operons.

A summary of the results from the search of the 20 most probable parents in the DBN structure are listed in Table 1. Seven of the operons that were found among the 20 best parents would correctly model causality if they were selected as the second best parent to be added. The probability of at least 1 of the 9 operons plus *tyrR* being in the top 20 best parents for each of the 9 operons is quite low, at 0.059[‖]. While further experimentation is required, these results provide some initial evidence supporting the use of time series data to learn causality.

DISCUSSION AND FUTURE DIRECTIONS

We have reported an initial experiment in learning Dynamic Bayesian networks as a means of modelling time series gene expression microarray data, with the aim of gaining insights into regulatory pathways. The prior structure and prior CPTs of our BN and DBN encode background knowledge about gene expression in the organism being modelled, *E. coli*. The experiments provide evidence that BN and DBN, together with an explicit model of operons, are capable of identifying operons in *E. coli* that are in a common regulatory pathway.

There are several directions for further research. First, a larger data set may improve the performance of the approach and allow us to determine whether causality can be determined from time series data. Some additional time series data recently have been made available by the Blattner Laboratory at the University of Wisconsin, under a different set of conditions, and we anticipate the availability of further time series data on *E. coli* in the year ahead. Nevertheless, potentially offsetting any such gain is the need to include additional genes (observed variables) and operons (hidden variables) in the analysis. The present data set used only 169 genes appearing in 142 operons. But the full *E. coli* genome has over 4000 genes, and

[‖] The probability of picking an operon that is not one of the 9 key operons or *tyrR* (if all genes are equally likely) is $\frac{132}{141}$. The probability of picking all 20 operons that are not one of the 10 is $(\frac{132}{141})^{20} = 0.27$. Thus, the probability of getting at least one of the 9 key operons is $1 - 0.27 = 0.73$. However, the probability of doing this for all 9 operons is $(0.73)^9 = 0.059$.

Table 1. Summary of the results of the DBN model. The operons listed on the right are one of the 9 key operons or *tyrR* that appeared as one of the 20 most probable parents of the operons on the left.

Key operons in tryptophan regulatory pathway	Operons known to be involved in the tryptophan regulatory pathway that appeared as one of the 20 most probable parents of the operon on the left
aroF-tyrA	**tyrR**, **trpR**
aroG	**aroF-tyrA**, **aroP**, **aroH**, **tyrR**
aroH	**tyrR**
aroL-yaiA-aroM	**trpR**, **tyrR**
aroP	**tyrR**
mtr	**tyrR**
tnaLAB	**aroF-tyrA**
trpLEDCBA	**aroH**, **tyrR**
trpR	**aroG**

the predicted operon map has well over 1000 multigene operons.

A second, and perhaps more important, shortcoming of the present work is that computation time did not permit a more encompassing search to be employed. Our full search algorithm modifies incoming arcs to every hidden node. As the arcs coming into one hidden node are modified, and the CPTs updated, this node may become a better parent for another node. A cascade of such improvements could improve the fit of the model dramatically and hence, potentially, the match of the model with the actual regulatory structure of the organism. Therefore, a crucial direction for further work is to decrease the computation time of the learning algorithm. An approximate approach to updating CPTs and calculating scores based on local Markov blanket (the parents, children, and children's parents of a node) of operons, where the structure changed, might speed up the learning process. Alternatively, increasing the efficiency of the implementation of the learning algorithm by parallel execution on a Condor pool (Litzkow *et al.*, 1988) can allow the full algorithm to be tested. In addition, the faster implementation will facilitate more extensive experimentation, including cross-validation to estimate the accuracies of expression levels that the model predicts for various genes at various time steps.

Third, we could include a variety of other variables, hidden or observed. Observed variables might include environmental factors such as glucose, tryptophan or temperature. Unobserved variables might include transcription factors and other protein products. In the present work we omitted these because we wanted to see the extent to which changes in the expression levels of a gene could be predicted solely on the basis of changes to other genes, regardless of the environmental causes of those changes.

This paper has presented the first application (to our knowledge) of Dynamic Bayesian networks to time series gene expression microarray data. It also has shown how background knowledge about an organism's genome (in this case, an operon map) can be used to construct the initial, core structure of the DBN. This background knowledge can be taken from the scientific literature or can itself be the output of another modelling system. In this case, the operon map consisted partially of each type of knowledge. This paper has provided some evidence that the results of such an application of DBNs provide additional insights into the organism's regulatory network and that such an application has the potential to reveal hierarchical connections among signal transduction pathways. This paper also has demonstrated that our DBN approach has the potential for inducing direct causal links, that is, direct arcs in the regulatory network. The approach proposed by Pe'er *et al.* (2001) can be used in conjunction with our approach if knockout experiments are available. Further experiments will provide insight into whether causality can really be learned from time series data.

ACKNOWLEDGEMENTS

The authors are grateful to Mark Rich and Ian Alderman for comments on drafts of this paper. The first author was supported by NIH Biotechnology Training Grant NIH 5 T32 GM08349. The third author was supported by NSF Grant 9987841.

REFERENCES

Craven,M., Page,D., Shavlik,J., Bockhorst,J. and Glasner,J. (2000) A probabilistic learning approach to whole-genome operon prediction. *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, San Diego, CA, pp. 116–127.

Friedman,N., Linial,M., Nachman,I. and Pe'er,D. (2000) Using Bayesian networks to analyze expression data. *J. Comput. Biol.*, **7**, 601–620.

Friedman,N., Murphy,K. and Russell,S. (1998) Learning the structure of dynamic probabilistic networks. *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*. Cooper,G.F. and Moral,S. (eds), Morgan Kaufmann, Madison, WI, pp. 139–147.

Khodursky,A., Peter,B.J., Cozzarelli,N.R., Botstein,D., Brown,P.O. and Yanofsky,C. (2000) DNA microarray analysis of gene expression in response to physiological and genetic changes that affect tryptophan metabolism in *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **97**, 12170–12175.

Litzkow,M.J., Livny,M. and Mutka,M.W. (1988) Condor—a hunter of idle workstations. *Proceedings of the 8th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, San José, CA, pp. 104–111.

Murphy,K. (2001) The Bayes net toolbox for matlab. *Computing Science and Statistics: Proceedings of Interface*. **33**.

Murphy,K. and Mian,S. (1999) Modelling Gene Expression Data using Dynamic Bayesian Networks. Technical Report, Computer Science Division, University ofCalifornia, Berkeley, CA.

Pe'er,D., Regev,A., Elidan,G. and Friedman,N. (2001) Inferring subnetworks from perturbed expression profiles. *Proceedings of the 9th International Conference on Intelligent Systems for Molecular Biology*, S215–S224.

Ray,J.M. and Bauerle,R. (1991) Purification and properties of tryptophan-sensitive 3-deoxy-D-arabino-heptulosonate-7-phosphate synthase from *Escherichia coli*. *J. Bacteriol.*, **173**, 1894–1901.

Salgado,H., Santos,A., Garza-Ramos,U., van Helden,J., Diaz,E. and Collado-Vides,J. (1999) RegulonDB (version 2.0): a database on transcriptional regulation in *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **27**, 59–60.

Yang,X. and Ishiguro,E.E. (2001) Involvement of the N terminus of ribosomal protein L11 in regulation of the RelA protein of *Escherichia coli*. *J. Bacteriol.*, **183**, 6532–6537.