Predicting altered pathways using extendable scaffolds

B.M. Broom

Department of Biostatistics and Applied Mathematics,
MD Anderson Cancer Center,
1515 Holcombe Blvd,
Houston TX 77030, USA
E-mail: broom@mdanderson.org

T.J. McDonnell

Department of Molecular Pathology
and General Medical Oncology,
MD Anderson Cancer Center,
1515 Holcombe Blvd, Houston TX 77030, USA
E-mail: tmcdonne@mdanderson.org

D. Subramanian*

Department of Computer Science,
Rice University, 6100 Main St,
Houston TX 77005, USA
E-mail: devika@rice.edu
*Corresponding author

Abstract: Many diseases, especially solid tumors, involve the disruption or deregulation of cellular processes. Most current work using gene expression and other high-throughput data, simply list a set of differentially expressed genes. We propose a new method, PAPES (predicting altered pathways using extendable scaffolds), to computationally reverse-engineer models of biological systems. We use sets of genes that occur in a known biological pathway to construct component process models. We then compose these models to build larger scale networks that capture interactions among pathways. We show that we can learn process modifications in two coupled metabolic pathways in prostate cancer cells.

Keywords: Bayesian networks; gene expression data analysis; prostate cancer; glutathione pathway; urea cycle.

Reference to this paper should be made as follows: Broom, B.M., McDonnell, T.J. and Subramanian, D. (2006) ‘Predicting altered pathways using extendable scaffolds’, *Int. J. Bioinformatics Research and Applications*, Vol. 2, No. 1, pp.3–18.

Biographical notes: Bradley M. Broom received his PhD in Computer Science at the University of Queensland. He is an Associate Professor at the Department of Biostatistics and Applied Mathematics, University of Texas MD Anderson Cancer Center. His research interests include high-performance computing, computer networking, programming languages and models and computational methods for machine learning. He is the author of 23 peer-reviewed papers and two book chapters.

Timothy J. McDonnell, MD, PhD, is a Professor of Molecular Pathology and Genitourinary Medical Oncology at the University of Texas MD Anderson Cancer Center. His research interests include cell death regulation in cancer progression and therapeutic resistance. He is the author of 125 peer-reviewed publications, 18 invited reviews and 13 book chapters.

Devika Subramanian is a Professor of Computer Science at Rice University. She received her PhD in Computer Science from Stanford University. Her research interests include machine learning and artificial intelligence and their applications in computational biology. She is the author of 80 peer-reviewed papers.

1 Introduction

The introduction of microarray technology has made it possible to measure the expression levels of thousands of genes in cells in different metabolic or physiological states. Such measurements are the basis of numerous studies that identify hundreds of genes that are differentially expressed in diseased cells (Eder et al., 2004; Henrique and Jeronimo, 2004; Rhodes et al., 2002; Singh et al., 2002). However, they are yet to yield detailed understanding of the underlying genetic and/or regulatory events that lead to the initiation and progression of complex diseases such as cancer. The central question is: how do we make sense of the vast list of differentially expressed genes and gene products; what roles do they play in the runaway metabolic processes in the cell? To answer this question, we propose a new approach to computationally reverse-engineer models of biological systems. Our goal is to construct models for normal and for diseased cells from data, so that we can explain changes in gene expression levels as a function of changes in the underlying biological processes.

By a model of a biological process, we mean a quantitative representation of the functional and regulatory relationships between expressions levels of genes themselves, as well as between genes and gene products in a cell. The role of such models is to identify the root causes of changes in expression levels, so that we can limit our attention to a very small set of genes that explain differences in function between normal and diseased cells.

Richer models could consider interactions between cell constituents and its environment. However, gene-expression microarray data, upon which we primarily rely in this study, measure only one aspect of the cell's activity – and not any of the myriad biological conditions that directly influence that expression. This is why we consider models that only capture dependencies between genes and gene products. Inferring dependencies between expression levels of genes and gene products is further complicated by the fact that typical case/control studies only provide data from a few hundred patients. Learning process models *ab initio* for complex diseases from such small samples is challenging; the data is simply not sufficient to uniquely identify a model. Adding to the challenge is the fact that gene expression data itself is noisy. There is tremendous variability caused by tissue heterogeneity, particularly for cells from solid tumour samples. Further, mRNA levels measured are approximate surrogates for gene expression levels – there is no way of directly assessing transcriptional, translational and post-translational effects. All of these problems make the task of inferring good predictive process models involving genes and gene products in mammalian cells very difficult.

Our method, predicting altered pathways using extendable scaffolds (PAPES), compensates for the small sample size problem by exploiting available knowledge of genetic and metabolic processes. There are two key ideas in this paper. First, instead of working with individual genes, we use sets of genes that occur in a *known* biological pathway to construct component process models. We derive the structure of each component model for normal cells from pathway databases. We do not restrict the genes in these models to those that are differentially expressed. We represent each component process model as a Bayesian network with nodes that include not just the observed gene expression levels as is standard in the literature (Friedman et al., 2000), but unobserved metabolites in the pathway as well. We use information from known pathways to constrain the number and placement of the unobserved nodes in the network. We could use expectation maximisation to learn the parameters of the stochastic process represented by the network. The network parameters are chosen to maximise the likelihood of observed data. We demonstrate that including hidden metabolite nodes improves the classification accuracy of the models inferred from data. We are also able to make testable hypotheses regarding the levels of these metabolites in normal as well as in diseased cells.

For modelling diseased cells, we adopt a two-stage approach. We first conservatively assume that the structures of the pathways remain intact, and that the parameters defining the component processes are altered by disease. Thus, we use the same network structure as for normal cells. The expectation maximisation learning procedure instantiates the parameters of the Bayesian network model from data derived from diseased cells. If the likelihood of the data, given the model, is low, we attribute it to poor choice of network structure. Thus, we let the data dictate the need for new structural models. We then use structure-learning methods to determine the most biologically plausible perturbations of the nominal pathways that are consistent with the data. By learning separate component models for normal and diseased states, we can perform structural and/or parameter-level comparisons between them. The comparisons can yield testable hypotheses about changes in the underlying biological processes.

Pathways in mammalian cells are often coupled. Modelling complex diseases requires that we learn multiple pathways jointly. Small sample sizes force us to consider an indirect approach to this problem. This leads to the second key idea in this paper – a new method for composing component models learned from data. We merge two component Bayesian network models by taking the union of their nodes and edges. Additional hidden metabolite nodes are added to link nodes between the component models as dictated by pathway databases. We only reestimate the conditional probability tables for nodes that are common to the two networks, and use the parameters learned during component modelling for the rest of the nodes. The data requirements for this local reestimation are far smaller than for *ab initio* learning of the merged network.

The composition yields larger scale networks that capture interactions among pathways. Differences in gene and protein expression levels can then be explained as functions of differences, if any, in the composite genetic and metabolic processes inferred from data. Our approach is generally applicable to a variety of complex diseases. In this paper, we illustrate our approach in the context of prostate cancer using publicly available expression data on the disease (Singh et al., 2002). We show that we can learn process modifications in two coupled metabolic pathways (glutathione metabolism and the urea cycle) in prostate cancer cells. Since our technique makes predictions about metabolite levels in normal and diseased cells, we can validate our method by checking these predictions in the laboratory.

Our paper is organised as follows. Section 2 describes related work in applying machine learning to expression data. Section 3 describes how we construct component models and how we compose them to construct process models that characterise normal cells and diseased cells. Section 4 presents initial results obtained by applying our approach to two pathways that have been implicated in prostate cancer, and makes testable predictions about changes in the underlying process models in diseased cells. Section 5 summarises the contributions of this paper and outlines our plans for subsequent research.

2 Related work

2.1 Representing metabolic and genetic networks

Pathways can be represented at several levels of abstraction ranging from network models which emphasise the fundamental components (genes and metabolic products) and connections between them (the L1 models as defined in Ideker and Lauffenberger, 2003), to detailed differential equation models of the kinetics of specific reactions (the L2 models) (de Jong, 2002; Ideker and Lauffenberger, 2003). The choice of abstraction level is generally a function of the biological problem being addressed and the type and quantity of data available. For instance, models based on differential equations have been used for detailed modelling of specific molecular interactions, when time series data for the concentrations of the various molecular components involved is available. Boolean networks approximate gene expression by binary variables that are either on or off, and gene interactions as Boolean functions. Approaches based on Bayesian networks (Friedman et al., 2000) and their generalisations allow representations of continuous quantities. They have a solid formal foundation and naturally support reasoning about incomplete and noisy data.

Bayesian networks are directed acyclic graphs that can be viewed as factored representations of the joint probability distribution on the values (or levels) of all the nodes in the network. They have been used in a wide variety of models generated from gene expression data including (Friedman et al., 2000; Bar-Joseph et al., 2003). Unlike purely qualitative models, Bayesian networks represent quantitative information in the form of conditional probabilities of nodes given their parent nodes in the network. Bayesian networks naturally represent the modularity found in metabolic processes. The directionality of parent-child relationships can be used to encode causality. Feedback loops in pathways can be represented by unrolling the loop over time using a temporal extension called dynamic Bayesian networks.

Given a Bayesian network and its parameters, the network can be queried to obtain the probability distribution of unobserved nodes conditional on the values of the observed nodes. For most machine learning applications, the goal is to learn the network parameters and often the network structure itself from the data. In this paper, we use Bayesian networks to represent pathways implicated in prostate cancer.

2.2 Learning interaction networks from high-throughput data

Learning Bayesian network models from gene expression microarrays raises many computational and representational challenges. First, there is the issue of choosing nodes to be included in the network; which genes and gene products should be considered in a 'network' explanation for a given data set? Second, how do we evaluate a given structure with respect to available data? Third, how can we efficiently find structures that optimise the chosen evaluation criterion? Fourth, how do we determine the biological validity of the computed networks?

The number of possible Bayesian networks on n nodes is super-exponential in n, so the question of which nodes to include in the network construction process is very important. In *ab initio* construction of gene regulatory networks, a starter set of differentially expressed genes obtained from a preprocessing phase (such as by clustering or correlational analyses followed by thresholding on p-values) is used. Another approach for limiting the number of genes is to consider overlaps between differentially expressed genes and those that occur on specific pathways of interest, as in Draghici et al. (2003). Tools such as Cytoscape make this overlap computation easy to perform.

All learning algorithms in the literature use a scoring function that evaluates the probability of a given network G with respect to the data G.

$$P(G \mid D) = \alpha \times P(D \mid G) \times P(G)$$

where $P(G)$ is a prior on the network structure and $P(D|G)$ is the likelihood of the data given the network; that is, how well the data is explained by the network. An optimal network maximises this scoring function.

Although gene expression microarrays measure the expression levels of many thousands of genes simultaneously, a typical study includes at most a few hundred different samples, which is far too few to reliably reconstruct a unique network model. In fact, it is not unusual for an exponential number of different networks on a given set of nodes to have the same high score! To circumvent this fundamental limitation on the amount of data needed to learn network structures with high confidence, two approaches have been considered.

One approach is to extract common features (such as edges between nodes) in all high scoring networks as suggested in Friedman et al. (2000). If the set of genes in the network are highly correlated, nearly every pair of genes is identified as a significant interaction. It therefore makes it difficult to learn pathways by this method, unless we consider methods that add hidden nodes to the network.

A different approach for overcoming the problems created by the small number of samples is to incorporate known biological information and incrementally add additional genes into the network using existing knowledge about gene interactions. Segal et al. (2003) have combined gene expression data and promoter sequence data to identify transcriptional modules in *Saccharomyces cerevisiae*. Bar-Joseph et al. (2003) combined genome-wide location data with the gene-expression data to obtain insights into regulatory networks for the same organism.

2.3 Perturbation analysis and pathway databases

Our approach is inspired by the work of Ideker et al. (2001) who performed an integrative study of the yeast galactose-utilisation pathway. They first identified all known genes, proteins and small molecules involved in this pathway and systematically considered genetic (overexpression and gene deletion) and environmental perturbations of the pathway. They used the resulting microarray and proteomic data to refine the existing galactose pathway model and to uncover new interactions between it and other pathways in the yeast cell. This paper also introduced physical interaction networks in which genes are nodes and there are edges between genes that regulate others through protein-DNA binding, and edges between coding genes whose proteins interact with one another. Physical interaction networks represent metabolic activity in a cell together with the regulatory components. While an exact replication of the Ideker yeast study is infeasible at present in the context of a complex disease such as prostate cancer, we adopt two of their key ideas in this work: the concept of physical interaction networks, and the types of pathway perturbations to consider.

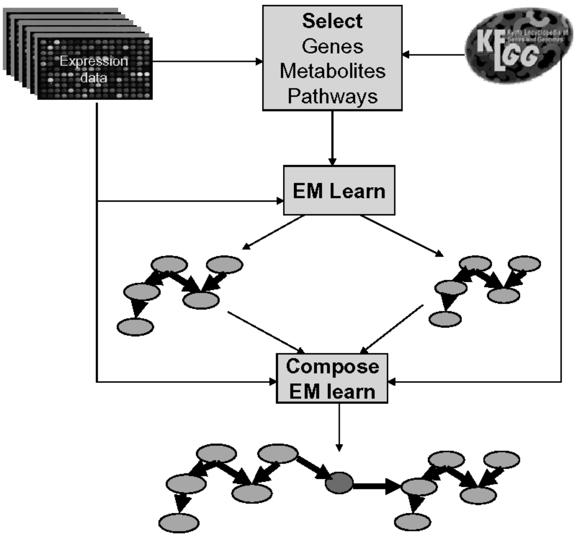
A variety of primary resources are available for constructing nominal networks of genes and their metabolic products. One is the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Ogata et al., 1999) pathway database, which contains networks covering metabolic processes, signal transcription and transduction as well as cellular process such as the cell cycle and apoptosis. This database identifies key enzymatic reactions and the genes that code for the enzymes concerned. The key information missing from the KEGG networks, which is important from the point of view of constructing process-level explanations of disease, are the regulatory and signalling genes. These ‘missing’ genes can be added to KEGG networks by merging information contained in three different databases. They include the DIP/BIND database, the proprietary Ingenuity knowledge base and the Transfac (Matys et al., 2003) database of transcriptional regulation, of genes. The Ingenuity database includes over one million manually curated relationships between genes, cells, diseases, drugs and other biological entities extracted from the literature, and includes information about transcription factors as well as protein–protein interactions. The TransFac database (Matys et al., 2003) can be used as a supplementary resource for relating transcriptional factors to genes in the interaction networks that we construct, while the HumanCyc database links genes with their metabolic products.

3 Predicting altered pathways using extendable scaffolds

The question that motivates our work is the following: can we begin with data on gene expression from both normal and cancerous tissue of various grades of prostate cancer, and derive process models that explain the differences in the observed data? The most straightforward approach would be to gather huge amounts of gene expression data on large samples of tissues and cells, and infer discriminative Bayesian network models directly. As explained in the introduction, this is not possible, because we simply do not have enough data on normal and tumour samples, and *ab initio* learning of Bayesian networks with thousands of genes is computationally infeasible.

Figure 1 shows our approach – predicting altered pathways using extendable scaffolds (PAPES) – for finding disease-affected cellular pathways. PAPES begins with a set of differentially expressed genes and the pathways that they participate in. Component networks are generated from portions of the pathways in which the differentially expressed genes occur. The component networks are pieces of the network scaffold. The pieces will be composed by adding genes and gene products that link pathways together. A scaffold piece contains not just the genes that are differentially expressed, but also other genes and gene products that they interact with. Each scaffold piece is represented as a Bayesian network. Bayesian network specifications have two parts. The first part is the structure, represented as a directed acyclic graph $G = (V, E)$, whose nodes (V) represent expression levels of genes or gene products, and whose edges ($E \subseteq V \times V$) denote dependence between expression levels. The structure of a Bayesian network exposes conditional independencies between expression levels. The second part, θ, is the set of conditional probability distributions for each node in the network. In this paper, all conditional probability distributions, $P(v|Parents(v))$ where $v \in V$ and $Parents(v) = \{u|(u,v) \in E\}$, are discrete multinomials. We discretise all expression levels into three discrete categories: low, medium and high. It is also possible and sometimes more accurate to model these distribution as continuous mixture distributions.

Figure 1 The extendable scaffold method for predicting altered pathways



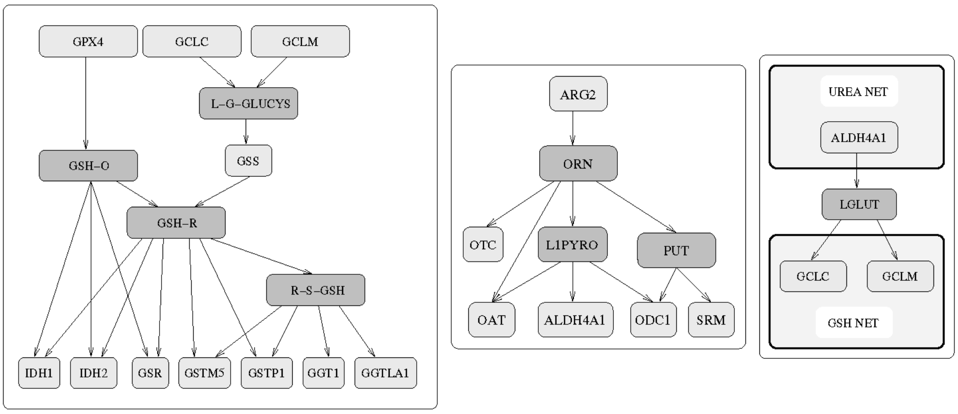
Our Bayesian network models differ from standard models such as (Friedman et al., 2000) in two key ways. Our goal is to use expression data from normal and diseased cells to discover structural and/or parametric perturbations in pathways that are attributable to disease. We therefore work with sets of genes, including those that are differentially expressed, requiring them to be part of a single pathway. Second, we enrich our network with nodes that represent the levels of metabolites that occur in the pathway. The placement and the selection of these hidden (i.e., unobserved) metabolite nodes are constrained by our knowledge of the pathway. As an example, consider the urea cycle in which the enzyme ODC (coded for by the gene ODC) converts the metabolite ornithine to putrescene. Putrescene is converted to spermidine by the enzyme coded for by the gene SRM. In a network that only includes observed gene expression levels, we can build a two-node network connecting ODC to SRM to model this portion of the urea cycle. The addition of hidden nodes for putrescene and ornithine constrains the interaction between ODC and SRM in a more biologically meaningful way. As a side effect, we can learn conditional probability models for ornithine (resp. putrescene) of the form $P(ornithine|Parents(ornithine))$, from gene expression data. Our stochastic process models are in fact the conditional probability tables associated with these metabolite nodes.

The structure of the Bayesian networks for normal cells is obtained from pathway databases, and the parameters of the networks are learned by expectation maximisation (EM). For modelling diseased cells, we adopt a two-stage approach. We first conservatively assume that the structure of the pathways remains intact, and that the parameters defining the component processes are altered by disease. Thus, we use the same network structure for diseased cells as for normal cells. The EM learning procedure instantiates the parameters of the Bayesian network model using data from diseased cells. A low likelihood for the data given the structure tells us that the network choice is incorrect. Thus, we let the data dictate the need for new structural models. We then use structure-learning methods to determine the most biologically plausible perturbations of the nominal pathways that are consistent with the data.

By optimising the component networks separately for both the normal and diseased cells, we aim to identify whether the differential gene expression is simply the pathway's response to the diseased state, or whether the pathway has been disrupted by disease. In the latter case, we can search for alterations to the network structure that best explains the changed response. We compare the parameters learned for normal and diseased networks and use the networks to predict levels of metabolites. Predicted differences in metabolite levels between normal and diseased cells are hypotheses that can be directly verified in the laboratory.

To explain complex diseases we may need to consider interactions between pathways. In Figure 1, we schematically show that we compose component network models to form larger scale process models. We merge two component Bayesian network models by taking the union of their nodes and edges. Additional hidden metabolite nodes are added to link nodes between the component models in a biologically consistent manner, as dictated by pathway databases. We only reestimate the conditional probability tables for newly added nodes and nodes common to the two networks. We reuse the parameters learned during component modelling for the rest of the nodes. The data requirements for this local reestimation are far smaller than *ab initio* learning of the merged network. Our approach to composition mitigates the problem of small sample sizes and yields robust larger scale networks that capture interactions among pathways. As an example, portions of the urea cycle interact with the glutathione metabolism in cells via a metabolite called L-glutamate. The Bayesian component networks corresponding to the urea cycle and the glutathione metabolism are merged by the introduction of a new hidden node as shown in Figure 2. In Section 4, we show the improvement in classification accuracy that results from the composition.

Figure 2 Component networks representing portions of the glutathione metabolism (left); the urea cycle (middle); and their composition via a hidden metabolite node (right)



4 An application of PAPES: glutathione pathways in prostate cancer

We use gene expression data from Singh et al. (2002) with 50 normal and 52 prostate cancer cases. Measurements of 12,625 genes for the 102 samples are in this data set. We identified the top 50 differentially expressed genes using Fisher scores and mutual information. These 50 genes map to over 20 known metabolic and signalling pathways in the KEGG database. To illustrate how PAPES works, we selected two of these pathways that interact with one another, and which are known to be implicated in prostate cancer. One of these is the glutathione metabolism. The differentially expressed gene GSTP1 in the glutathione mechanism is believed to be epigenetically silenced in prostate cancer (Eder et al., 2004). The other related pathway is the urea cycle containing polyamines ornithine and putrescene, which are overexpressed in prostate cancer. The overexpression of these metabolites is attributed to the overexpression of the enzyme ODC that regulates the conversion of ornithine to putrescene (Dhanasekaran et al., 2001). We generate Bayesian network component models for portions of these pathways in which the genes GSTP1 and ODC participate. The two pathway segments interact through the metabolite L-glutamate as shown in Figure 1. Our goal is to determine if these pathway disruption explanations in the literature are derivable from the Singh et al. data.

4.1 Modelling the glutathione component network

Glutathione reduces substances, such as peroxides or free radicals, which accumulate in cells under oxidising conditions. By maintaining an intracellular reducing environment, glutathione prevents intracellular protein thiols from oxidising to disulfides. In conjunction with glutathione S-transferases (GST*), it participates in detoxification of organic halides, fatty acid peroxides and products derived from radiation-damaged DNA. When the GST enzymes are underexpressed, as has been observed in prostate cancer cells, the detoxification process is disrupted.

The metabolites in this portion of the glutathione (GSH) pathway include the oxidised and reduced forms of glutathione, R-S-glutathione (a conjugate of reduced glutathione) and L-γ-glutamylcysteine. Since metabolite levels are not observed, they are hidden nodes in the Bayesian network representation of this segment of the glutathione pathway. The other nodes in the network shown in Figure 1 correspond to expression levels of genes that code for the named enzymes. The network structure reflects the following process. The glutathione per-oxidase enzymes (GPX*) catalyse the reduction of peroxides, producing oxidised glutathione (GSHO). Glutathione reductase (GSR) recovers reduced glutathione (GSHR) from oxidised glutathione. There is an alternate synthesis pathway for reduced glutathione from L-γ-glutamylcysteine (L-G-GLUCYS) fueled by the enzyme glutathione synthase (GSS). L-γ-glutamylcysteine is produced from cysteine by enzymes GCLC and GCLM. The glutathione-S-transferase enzymes (GSTM1 and GSTP1) catalyse the conjugation of reduced glutathione into R-S-glutathione (R-S-GSH). Enzymes GGT1 and GGTLA1 consume R-S-glutathione and generate the metabolite R-S-Alanylglycine that is not modelled in our network.

In the Bayesian network modelling of the glutathione metabolism, a node representing a metabolite has the raw material metabolites needed for its synthesis, as parents. Thus, there is an edge from GSHO to GSHR, and from GSHR to R-S-GSH. Metabolites such as GSHO and L-G-GLUCYS, which are generated from metabolites not included in the analysis, have as parents the catalysing enzymes that generate them. Therefore, there is an edge from GPX4 to GSHO and from GCLC and GCLM to L-G-GLUCYS. Enzymes such as IDH1/2, GSR, GSTP1 and GSTM1, which convert one metabolite to another have as parents the metabolites they consume and the ones they produce. Thus, GSR as well as IDH1/2 has GSHO and GSHR as parents, while GSTP1 and GSTM5 have GSHR and R-S-GSH as parents.[1] Enzymes such as GGT1 and GGTLA1, which consume a modelled metabolite and produce an unmodelled one, have as parents the input metabolite they convert. Thus, we have an edge from R-S-GSH to GGT1 and to GGTLA1. Each edge in the network denotes a dependence between the expression levels of genes and gene products.

While the KEGG pathway constrains the structure of our Bayesian network, we still need to represent the quantitative part of the model. These parameters are probability distributions of each node as a function of its parents in the network. For nodes with no parents, such as GPX4, GCLC and GCLM, we learn unconditional probability distributions of the form $P(node = value)$, over the range of values that these nodes take. To simplify the specification of these distributions, we discretise gene expression levels and metabolite levels, into three categories: low, medium and high. For each gene, the discretisation points were chosen by exhaustively searching for the two values that maximised the weighted average of the sample's self-information (I_s) and the mutual information (I_m) between the sample and its type (tumour or normal) using the following formula: $0.425I_s + I_m$. We found that by including the weighted self-information, the tendency to select very narrow discretisation levels was reduced. For all other nodes, we learn conditional probability distributions of the form $P(node = value|Parents(node) = value\_vector)$.

4.2 Learning network parameters

The algorithms that learn network parameters find values of the probability distributions of the nodes to maximise the likelihood of the given data. Since not all nodes are observable, we use the expectation maximisation (EM) method, which uses the distributions of values of the hidden nodes computed by standard Bayesian inference. When EM is applied multiple times to the same network, there is significant variation in the resulting network likelihood, which we ascribe to the EM procedure finding local maxima, perhaps because of the comparatively high proportion of hidden nodes and the small number of data points. To reduce the resulting variability, we repeat the EM procedure multiple times (30) and average the top few (6) sets of network parameters, where the best networks are those that best discriminate between normal and tumour samples. The learned parameters for oxidised glutathione for normal and tumour cells are shown in Table 1, while those for reduced glutathione are shown in Table 2. The models are very similar except for a skew to the left in the tumour distributions when the catalysing enzyme levels are medium.

Table 1 Stochastic process models for the production of oxidised and reduced glutathione in normal cells (top), and tumor cells (bottom)

GPX	*Low*	*Medium*	*High*
		GSH-O (normal)	
Low	0.67 ± 0.25	0.23 ± 0.24	0.10 ± 0.24
Medium	0.33 ± 0.40	0.65 ± 0.40	0.00 ± 0.01
High	0.04 ± 0.07	0.13 ± 0.10	0.83 ± 0.09
		GSH-O (tumour)	
Low	0.74 ± 0.35	0.11 ± 0.16	0.14 ± 0.32
Medium	0.68 ± 0.34	0.09 ± 0.13	0.23 ± 0.27
High	0.02 ± 0.02	0.02 ± 0.02	0.96 ± 0.02

Table 2 Stochastic process models for the production of reduced glutathione in normal cells (top), and tumor cells (bottom)

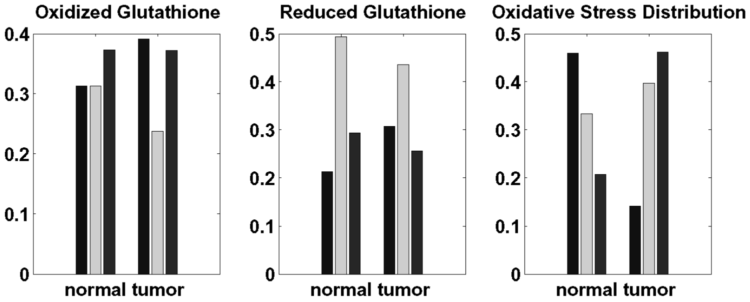
GSS	*GSHO*	*Low*	*Medium*	*High*
			GSHR (normal)	
Low	Low	0.60 ± 0.30	0.12 ± 0.13	0.27 ± 0.36
Medium		0.50 ± 0.49	0.33 ± 0.50	0.17 ± 0.34
High		0.60 ± 0.38	0.26 ± 0.42	0.14 ± 0.22
Low	Medium	0.23 ± 0.37	0.30 ± 0.44	0.47 ± 0.50
Medium		0.01 ± 0.01	0.34 ± 0.49	0.65 ± 0.49
High		0.10 ± 0.14	0.45 ± 0.36	0.44 ± 0.41
Low	High	0.02 ± 0.02	0.17 ± 0.39	0.81 ± 0.39
Medium		0.01 ± 0.00	0.01 ± 0.01	0.97 ± 0.01
High		0.01 ± 0.00	0.04 ± 0.07	0.96 ± 0.07
			GSHR (tumour)	
Low	Low	0.97 ± 0.02	0.01 ± 0.01	0.01 ± 0.01
Medium		0.79 ± 0.27	0.15 ± 0.23	0.06 ± 0.06
High		0.52 ± 0.52	0.20 ± 0.39	0.28 ± 0.44
Low	Medium	0.49 ± 0.41	0.19 ± 0.39	0.32 ± 0.36
Medium		0.18 ± 0.37	0.34 ± 0.48	0.48 ± 0.49
High		0.39 ± 0.36	0.39 ± 0.43	0.32 ± 0.37
Low	High	0.27 ± 0.32	0.01 ± 0.01	0.71 ± 0.31
Medium		0.20 ± 0.39	0.01 ± 0.02	0.79 ± 0.38
High		0.20 ± 0.39	0.00 ± 0.00	0.81 ± 0.39

With the structure and parameters of the nominal glutathione network set up, we can calculate the probabilities of metabolite levels induced by a specific configuration of observed gene expression levels. These computations are performed by standard Bayesian network inference. For every normal sample in our gene expression data, we calculate the probability distributions over the metabolite nodes. We predict that tumour cells have lower levels of reduced glutathione than normal cells, and that a higher proportion of tumour samples will have high oxidative stress. Oxidative stress is the ratio of reduced to oxidised glutathione. Our predictions are borne out in Figure 3.

4.3 Modelling the urea cycle component network

The urea cycle is a metabolic process in which ammonia, produced during amino acid degradation, is converted to urea in the liver, through a series of reactions. The portion of the urea cycle of interest to us is the conversion of ornithine into putrescene catalysed the enzyme ODC, which is known to be overexpressed in prostate cancer. Ornithine (ORN) is produced from the metabolite arginine by the enzyme ARG. Ornithine is consumed by three different processes. One is the conversion to putrescene (PUT) by ODC. Putrescene is converted to spermidine and spermine downstream by enzyme SRM. Ornithine is also used for the production of metabolite citrulline by enzyme OTC. The metabolite L-Glutamate 5-semialdehyde and L1Pyrroline 5 carboxylate (we abbreviate it as L1PYRO) is made from ornithine by the enzyme OAT. Finally, L1PYRO itself is consumed by an enzyme ALDH4A1 to metabolite glutamate.

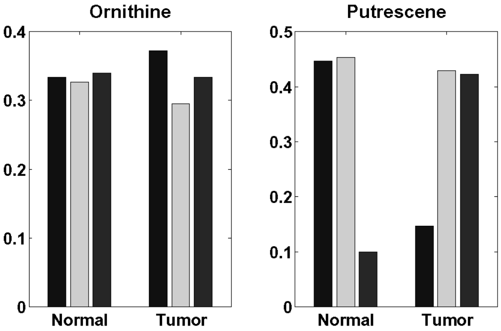
Figure 3 The predictions made by the learned normal and tumor glutathione networks



We use the same principles for modelling the above process as a Bayesian network, as we did for the glutathione metabolism. The modelled metabolites are ornithine, putrescene and LIPYRO. There are edges from ORN to PUT and ORN to L1PYRO because ornithine is converted to these other metabolites in this cycle. Since ARG produces ORN, we have an edge from ARG to ORN. ODC and OAT catalyse the conversions from ornithine to putrescene and ornithine to L1PYRO, respectively. Thus, ODC has parents ORN and PUT, and OAT has parents ORN and L1PYRO. Finally, edges from PUT to SRM and L1PYRO to ALDH4A1 denote processes that consume these metabolites.

We learn the parameters of the urea network for normal and tumour cells, using the same algorithm as the glutathione metabolism. We use the learned networks to make predictions about the levels of ornithine and putrescene. We are able to derive the fact that putrescene levels are elevated in tumour cells as shown in Figure 4.

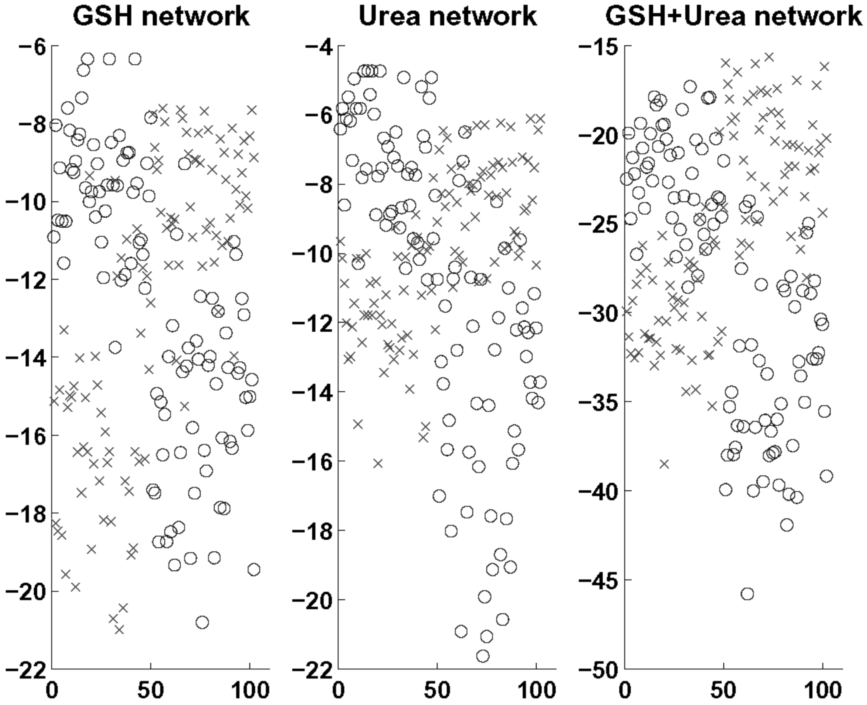
Figure 4 The predictions made by the learned normal and tumor urea networks



4.4 Composing the component networks

To build the combined network, we fix the parameters of all nodes that also occur in a component network, and have the same parents, to those learned for the component network. The EM procedure applied to the combined network therefore only learns the parameters for a small number of nodes. In our example, the parameters for the hidden metabolite node glutamate, which unites the glutathione and urea networks, is learned. A comparison of the classification performance of the combined network and the component networks is shown in Figure 5.

Figure 5 These figures show the log likelihood of normal samples (o's) and tumor samples (x's). The normal samples are the first 50 elements on the x axis, the tumor samples are the rest



4.5 Robustness and sensitivity analysis

To ensure that our EM network learning procedure is not overfitting the data, we checked the method's classification accuracy using leave-one-out cross validation. Using the discretised data, for each sample we learned the parameters of each network using all but that data point, and then used those parameters to classify the excluded data point. Table 3 presents the results that are in broad accord with the classification accuracy obtained using all samples, giving us confidence in the robustness of the EM learning process.

Table 3 Results of leave-one-out cross-validation

		Actual					
		Glutathione network		*Urea network*		*Combined network*	
		Normal	*Tumour*	*Normal*	*Tumour*	*Normal*	*Tumour*
Predicted	Normal	41	8	42	13	45	7
	Tumour	9	44	8	39	5	45

5 Conclusions and further research

We introduced a new approach to computationally reverse-engineer models of biological systems from data. Our method – Predicting Altered Pathways using Extendable Scaffolds (PAPES) – built on available pathway knowledge to construct component networks is based on subsets of differentially expressed genes. We represented each component process model as a Bayesian network with nodes that include not just the observed gene expression levels, but unobserved metabolites in the pathway as well. We then used expectation maximisation to learn the parameters of the stochastic process represented by the network. We obtained models that predicted metabolite levels in normal and tumour cells. These predictions are directly testable in the laboratory. We then proposed a method for composing component network to build larger scale process models. We demonstrated that the classification accuracy and explanatory power of the composed network was better than the individual component networks. The incremental construction of models that explain differences between metabolic process in normal and tumour cells with gene expression data alone, forms the first step in elucidating the molecular basis of complex diseases.

We illustrated our approach using portions of the glutathione metabolism and the urea pathway with which it interacts. We computationally reconstructed the parameters of the glutathione metabolism for normal and tumour cells by the EM procedure on Bayesian networks. We used the models to show that levels of reduced glutathione are lower in tumour cells than in normal cells and that a much larger proportion of tumour cells have high oxidative stress compared to normal cells (Eder et al., 2004). We also reconstructed a portion of the urea cycle involving the metabolites, ornithine and putrescene. Our computational reconstruction allowed us to to infer that while ornithine levels are similar for normal and tumour cells; the levels of putrescene in tumour cells are markedly higher. This prediction is borne out in the literature (Dhanasekaran et al., 2001). We composed these two component networks into a single network and used it to classify the samples in a leave-one-out setting. We showed that the combined network has higher classification accuracy than either component network. Our composition method could be extended to cover metabolic processes on a genome-wide scale. As metabolic data becomes more readily available, we can extend our methods to learn new network structures to better explain differences in functioning between normal and tumour cells.

References

Bar-Joseph, Z., Gerber, G.K., Lee, T.I., Rinaldi, N.J., Yoo, J.Y., Robert, F., Gordon, D.B., Fraenkel, E., Jaakkola, T.S., Young, R.A. and Gifford, D.K. (2003) 'Computational discovery of gene modules and regulatory networks', *Nature Biotechnology*, Vol. 21, No. 11, pp.1337–1342.

de Jong, H. (2002) 'Modeling and simulation of genetic regulatory systems: a literature review', *Journal of Computational Biology*, Vol. 9, No. 1, pp.67–103.

Dhanasekaran, S.M., Ghosh, D., Barrette, R.R., Shah, R., Varambally, S., Kurachi, K., Pienta, K.J. and Chinnaiyan, A.M. (2001) 'Delineation of prognostic biomarkers in prostate cancer', *Nature*, Vol. 412, August, pp.822–826.

Draghici, S., Khatri, P., Martins, R.P., Ostermeier, G.C. and Krawetz, S.A. (2003) 'Global functional profiling of gene expression', *Genomics*, Vol. 81, pp.98–104.

Eder, L.E., Bektic, J., Baartsch, G. and Klocker, H. (2004) 'Genes differentially expressed in prostate cancer', *BJU International*, Vol. 93, pp.1151–1155.

Friedman, M., Linial, I., Nachman and Pe'er, D. (2000) 'Using Bayesian networks to analyze expression data', *Journal of Computational Biology*, Vol. 7, Nos. 3–4, pp.601–620.

Henrique, R. and Jeronimo, C. (2004) 'Molecular detection of prostate cancer: a role for GSTP1 hypermethylation', *Eur. Urol.*, Vol. 46, No. 5, November, pp.660–669.

Ideker, T. and Lauffenberger, D. (2003) 'Building with a scaffold: emerging strategies for high to low-level cellular modeling', *Trends in Biotechnology*, Vol. 21, No. 6, pp.255–262.

Ideker, T., Thorsson, V., Ranish, J.A., Christmas, R., Buhler, J., Eng, J.K., Bumgarner, R., Goodlett, D.R., Aebersold, R. and Hood, L. (2001) 'Integrated genomic and proteomic analyses of a systematically perturbed metabolic network', *Science*, Vol. 292, No. 5518, pp.929–934.

Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubroack, M., Hehl, R., Hornisher, K., Karas, K., Kel, D., Kel-Margoulis, A.E., Kloos, O.V., Land, D.U., Lewicki-Potapov, S., Michael, H., Munch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S. and Wingender, E. (2003) 'Transfac: transcriptional regulation: from patterns to profiles', *Nucleic Acids Research*, Vol. 31, No. 1, pp.374–378.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. *et al.* (1999) 'Kegg: Kyoto encyclopedia of genes and genomes', *Nucleic Acids Research*, Vol. 29, pp.29–34.

Rhodes, D.R., Barrette, T.R., Rubin, M.A., Ghosh, D. and Chinnaiyan, A.M. (2002) 'Metaanalysis of miroarrays: interstudy valication of gene expression profiles reveals pathway dysregulation in prostate cancer', *Cancer Research*, Vol. 62, August, pp.4427–4433.

Rosser, C.J., Tanaka, M., Pisters, L.L., Tanaka, N., Levy, L.B., Hoover, D.C., Grossman, H.B., McDonnell, T.J., Kuban, D.A. and Meyn, R.E. (2004) 'Adenoviral-mediated pten transgene expression sensitizes bcl-2-expressing prostate cancer cells to radiation', *Cancer Gene Ther.*, Vol. 11, No. 4, pp.273–279.

Segal, E., Yelensky, R. and Koller, D. (2003) ‘Geneome-wide discovery of transcriptional modules from DNA sequence and gene expression’, *Bioinformatics*, Vol. 19, Suppl. 1, pp.i273–i282.

Singh, D., Febbo, P.G., Riss, K., Jackson, D.G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A.A., D’Amico, A.V., Richie, J.P., Lander, E.S., Loda, M., Kantoff, P.W., Golub, T.R. and Sellers, W.R. (2002) ‘Gene expression correlates of clinical prostate cancer behavior’, *Cancer Cell*, Vol. 1, March, pp.203–209.

Note

[1] The reason for this choice is that we avoid models in which a hidden node has more than one hidden node as a parent. The interpretation of the conditional probability tables of such nodes becomes difficult, because we have to assign semantics to the discretised levels of these nodes.