Predicting Altered Pathways using Extendable Scaffolds

Bradley M. Broom
Dept. of Biostat. and App. Math.
M.D. Anderson Cancer Center
E-mail: bmbroom@mdanderson.org

Timothy J. McDonnell
Dept. of Mol. Path. and Gen. Med. Oncology
M.D. Anderson Cancer Center
E-mail: tmcdonne@mdanderson.org

Devika Subramanian
Dept. of Computer Science
Rice University
E-mail: devika@rice.edu

Abstract

We propose a new approach to computationally reverse-engineer models of biological systems from data. Our goal is to construct models for normal cells and for diseased cells, so that we can explain changes in gene expression levels as a function of changes in the underlying biological processes. Limited data availability makes it challenging to learn accurate process models *ab initio* on a genome-wide scale. Our method — Predicting Altered Pathways using Extendable Scaffolds (PAPES) — compensates for the lack of data 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—not restricted to those that are differentially expressed—to construct component process models. Each component process model is represented as a Bayesian network with nodes for both the observed gene expression levels and the unobserved metabolites in the pathway. The structure of each component model is derived from pathway databases. Second, we compose these models in a novel manner to construct process representations for both normal and diseased cells. This composition yields larger scale networks that capture interactions among pathways in complex diseases. Using publicly available gene expression data on prostate cancer, we show that the method can learn process modifications in two coupled metabolic pathways (glutathione and urea) in prostate cancer cells.

1 Introduction

There is increasing evidence that disease progression in complex diseases, especially solid tumors, does not arise from an individual molecule or gene, but from complex interactions between a cell's numerous constituents and its environment.While numerous studies identify genes and proteins differentially expressed in diseased cells [4, 6, 8], they are yet to yield detailed understanding of the underlying genetic and/or regulatory events that lead to the initiation and progression of diseases such as cancer. Our goal is to use computational learning methods to elucidate from high-throughput gene expression data and additional proteomic or metabolic data, the nature of the altered interactions that characterized diseased cells.

A significant obstacle to the application of computational learning methods for solving this problem is the extremely small amount of data—a few hundred samples in a typical study—which effectively eliminates *ab initio* methods. Our approach, therefore, combines known information about the genetic and metabolic pathways affected by a disease with the available gene expression and proteomic data, and attempts to reverse-engineer the most plausible modifications to these pathways.

This paper is organized as follows. Section 2 describes related work in applying machine learning to expression data. Section 3 describes our iterative method for incrementally constructing a disease model incorporating gene expression data and biological pathway information. Our approach is generally applicable to a variety of complex genetic diseases. Section 4 presents results obtained by applying our approach to two pathways that have been implicated in prostate cancer, and makes testable predictions about the levels of a number of significant metabolites involved. With our clinical collaborator, we plan to validate the structures of the learned networks. Section 5 summarizes the contributions of this paper and outlines our plans for subsequent research.

2 Related Work

2.1 Representing networks

Pathways can be represented at several levels of abstraction ranging from network models which emphasize the fundamental components (genes and metabolic products) and connections between them (the L1 models as defined in [5]), to detailed differential equation models of the kinetics of specific reactions (the L2 models). The choice of abstraction level is generally a function of the biological problem being addressed and the type and quantity of data available.

Bayesian networks are directed acyclic graphs which 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. Unlike purely qualitative models, Bayesian networks represent quantitative information in the form of conditional probabilities of nodes given their parent nodes in the network.

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.

2.2 Learning interaction networks

Algorithms for learning the structure of a Bayesian network from data use a scoring function that evaluates the probability of a given network G with respect to the data D: $P(G|D) = \alpha * P(D|G) * P(G)$, 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 maximizes this scoring function.

Learning Bayesian network models from gene expression microarrays raises many computational and representational challenges. The number of possible Bayesian networks on n nodes is super-exponential in n, and learning an optimal Bayesian network from data is NP-hard. Considerable research is devoted to finding good approximations to optimal networks, although many of these approximation methods are themselves computationally infeasible for problems involving several thousand genes. The *sparse candidate algorithm* [3] is suitable for problems involving up to a few hundred nodes.

Even so, learning a Bayesian network for an entire microarray experiment is computationally infeasible, 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 pre-processing phase (such as by clustering or correlational analyses followed by thresholding on p-values) is used [3].

A significantly greater problem with current data sets is that 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. This 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 is to extract common features (such as edges between nodes) in all high scoring networks as suggested in [2]. Even common features, however, may simply be artifacts of noisy data and its preprocessing (such as discretization), since there is so little data.

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. [7] have combined gene expression data and promoter sequence data to identify transcriptional modules in *Saccharomyces cerevisiae.*

3 Predicting Altered Pathways using Extendable Scaffolds

Can we begin with data on gene expression and possibly protein levels from both normal and cancerous tissue of various grades of prostate cancer, and derive network models that explain the differences in the observed data? The most straightforward approach would be to gather huge amounts of gene expression and proteomic data on large samples of tissues and cells, and infer discriminative Bayesian network models directly. This is not feasible because we simply do not have enough data on normal and tumor samples, and *ab initio* learning of pathway 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 to which they belong. Component networks are generated from portions of the pathways in which the differentially expressed genes occur, and

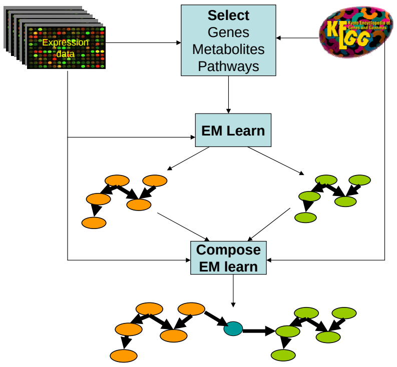

Figure 1: The extendable scaffold method for predicting altered pathways. It begins with an initial Bayesian network representing a pathway. The network parameters are optimized with respect to the available data. This network serves as a scaffold onto which additional genes and gene products are added, guided by the set of differentially expressed genes and the pathways in which they participate.

contain not just differentially expressed genes, but also other genes and gene products with which they interact. The component networks are the pieces in the network scaffold, and each is represented as a Bayesian network. The pieces will be composed by adding genes and gene products that link pathways.

Our Bayesian network models differ from standard models 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 selection of these hidden (i.e., unobserved) metabolite nodes is constrained by our knowledge of the pathway.

The structure of the Bayesian networks for normal cells is derived from pathway databases, and the parameters of the networks are learned by expectation maximization (EM). For modeling diseased cells, we adopt a two-stage approach. We initially use the same network structure for diseased cells as for normal 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 optimizing 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. We compare the parameters learned for normal and diseased networks and use the networks to predict metabolite levels. Predicted differences in metabolite levels between normal and diseased cells are hypotheses that can be verified in the laboratory.

To explain complex diseases we may need to consider interactions between pathways. We merge two component Bayesian network models by taking the union of their nodes and edges, as in Figure 1. 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 re-estimate the conditional probability tables for newly added nodes and nodes common to the two networks. We re-use the parameters learned during component modeling for the rest of the nodes. The data requirements for this local re-estimation 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.

4 An application of PAPES: prostate cancer pathways

In this section, we apply PAPES to the prostate gene expression data of Singh et al. [8]. This data set comprises gene expression measurements for 12,625 genes across 102 prostate samples, 50 normal and 52 cancer. We identified the top 50 differentially expressed genes using Fisher scores and mutual information, and mapped these genes to over 20 known metabolic and signaling pathways in the KEGG database. To illustrate how PAPES works, we selected two of these pathways that interact with one another—the glutathione (GSH) metabolism and the urea cycle—and which are known to be implicated in prostate cancer. The differentially expressed gene GSTP1 in the GSH mechanism is believed to be epigenetically silenced in prostate cancer. The polyamines ornithine and putrescene in the urea cycle are overexpressed

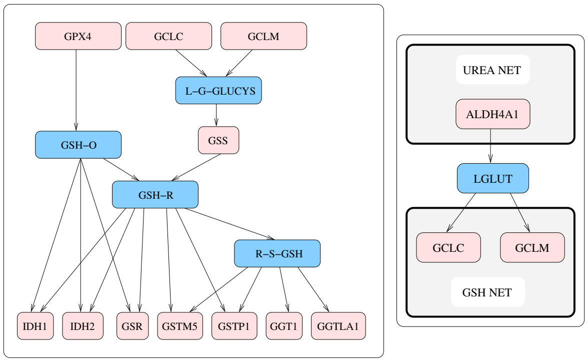

Figure 2: Component network representing a portion of the GSH metabolism (left), and its composition with the urea cycle via a hidden metabolite node (right). The blue nodes in the networks represent metabolite fluxes; they are hidden nodes, because we do not have observed metabolite data.

GPX	GSH-O (normal)		
	low	med	high
low	0.67±0.25	0.23±0.24	0.10±0.24
med	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

GPX	GSH-O (tumor)		
	low	med	high
low	0.74±0.35	0.11±0.16	0.14±0.32
med	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

Figure 3: Stochastic process models for the production of oxidized and reduced GSH in normal cells (top), and tumor cells (bottom). Oxidized GSH is catalyzed by GPX. When GPX levels are medium, the probability distribution of GSHO is skewed to the left in tumor cells. These stochastic models account for individual variation as well as incompleteness of our understanding of the process.

in prostate cancer, due to the overexpression of the enzyme ODC which regulates the conversion of ornithine to putrescene [1]. The two pathway segments interact through the metabolite L-glutamate as shown in Figure 2.

4.1 GSH component network

In conjunction with GSH S-transferases (GST*), GSH 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 GSH pathway include the oxidized and reduced forms of GSH, R-S-glutathione and L-γ-glutamylcysteine. Since metabolite levels are not observed, they are hidden nodes in the Bayesian network representation of this pathway component. The other nodes in the network shown in Figure 1 correspond to expression levels of genes that code for the named enzymes.

The structure of the Bayesian network model of the pathway component departs where necessary from a causal model in favor of one that can be learnt from the limited data available. Specifically, we desire a model in which each node has at most two (or possibly three) parents, all hidden nodes have at least one parent and one child, and no hidden node has two or more hidden nodes as parents. Subject to these constraints, 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. Metabolites such as GSHO which are generated from metabolites not included in the analysis have as parents the catalyzing enzymes that generate them. Therefore, there is an edge from GPX4 to GSHO. Enzymes such as GSTP1 which convert one metabolite to another have as parents the metabolites they consume and the ones they produce. Thus, GSTP1 has GSHR and R-S-GSH as parents. GSS is a parent of GSH-R to avoid a model 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 discretized levels of these nodes. Enzymes such as GGT1 which consume a modeled metabolite, and produce an unmodeled one, have as parents the input metabolites they convert. Thus we have an edge from R-S-GSH to GGT1.

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 discretize gene expression levels and metabolite levels, into three categories: low, medium and high. For each gene, the discretization points were chosen by exhaustively searching for the two values that maximized the weighted average of the sample's self-information (I_s) and the mutual information (I_m) between the sample and its type (tumor or normal) using the following formula: $0.425 I_s + I_m$. We found that by including the weighted self-information, the tendency to select very narrow discretization 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 maximize the likelihood of the given data. Since not all nodes are observable, we use the expectation maximization (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 likelihoods, 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 tumor samples. The learned parameters for oxidized GSH for normal and tumor cells are shown in Figure 3. The models are very similar except for a skew to the left in the tumor distributions for medium GPX levels.

Having determined the structure and parameters of the nominal GSH network, we can calculate using standard Bayesian network inference the probabilities of metabolite levels induced by a specific configuration of observed gene expression levels. For every normal sample in our gene expression data, we calculate the probability distributions over the metabolite nodes. As shown in Figure 4, we predict that tumor cells have lower levels of reduced GSH than normal cells, and that a higher proportion of tumor samples will have high oxidative stress (the ratio of reduced to oxidized GSH).

4.3 Urea-cycle component network

The portion of the urea cycle of interest to us is the conversion of ornithine into putrescene catalyzed by the enzyme ODC, which is known to be overexpressed in prostate cancer. Using the same network design principles as in the GSH metabolism, we derived a Bayesian network for this portion of the urea cycle. The network parameters were learnt for normal and tumor cells, using the same algorithm as the GSH metabolism.

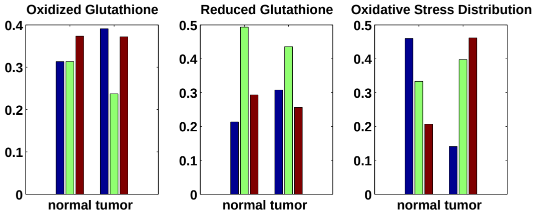

Figure 4: The predictions made by the learned normal and tumor GSH networks. We predict that the levels of reduced GSH will be reduced in tumor cells. The oxidative stress, measured as the ratio of oxidized GSH to reduced GSH, is also significantly greater in tumor cells than normal cells.

4.4 Component Composition

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 GSH and urea networks, is learned.

4.5 Robustness and sensitivity

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. The results, see Figure 5, are in broad accord with the classification accuracy obtained using all samples, giving us confidence in the robustness of the EM learning process.

5 Conclusions

We introduced a knowledge-based approach for inferring pathway modifications that explain differences in gene expression data gathered from normal and tumor samples. Our method — Predicting Altered Pathways using Extendable Scaffolds (PAPES) —

	GSH Network		Urea Network		Combined Network	
	Actual		Actual		Actual	
Predicted	N	T	N	T	N	T
N	41	8	42	13	45	7
T	9	44	8	39	5	45

Figure 5: Results for leave-one-out cross validation. The results are broadly in accordance with the classification accuracy obtained using all samples, giving us confidence in the robustness of the EM learning process.

builds on available pathway knowledge to build component networks based on subsets of differentially exprssed genes. We explicitly model metabolite fluxes in our network scaffold which is represented as a Bayesian network. We use expectation maximization to learn optimized parameters for the network from available data. We obtained models that predict differences in metabolite levels in normal and tumor cells. Such differences are directly testable in the laboratory. We propose to extend such biologically validated networks with genes known to interact with those in the scaffold. This incremental construction of models that explain differences between metabolic process in normal and tumor cells with limited gene expression data forms the first step in elucidating the molecular basis of complex diseases.

We illustrated our approach using portions of the GSH metabolism and the urea cycle with which it interacts. We computationally reconstructed the parameters of the GSH metabolism for normal and tumor cells by the EM procedure on Bayesian networks. We used the models to show that levels of reduced GSH are lower in tumor than in normal cells and that a much larger proportion of tumor cells have high oxidative stress compared to normal cells. 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 tumor cells; the levels of putrescene in tumor cells are markedly higher. This prediction is borne out in the literature [1]. 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 tumor cells.

References

[1] S. M. Dhanasekaran, D. G. R. R. Barrette, R. Shah, S. Varambally, K. Kurachi, K. Pienta, and A. M. Chinnaiyan, *Delineation of prognostic biomarkers in prostate cancer*, Nature, 412 (2001), pp. 822–826.

[2] N. Friedman and D. Koller, *Being bayesian about network structure: A bayesian approach to structure discovery in bayesian networks*, Machine Learning, 50 (2003), pp. 95–126.

[3] N. Friedman, I. Nachman, and D. Pe'er, *Learning Bayesian network structure from massive datasets: the "sparse candidate" algorithm*, in Proc. Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI '99), H. Dubios and K. Laskey, eds., Morgan Kaufmann, 1999, pp. 206–215.

[4] R. Henrique and C. Jeronimo, *Molecular detection of prostate cancer: A role for GSTP1 hypermethylation*, Eur Urol., 46 (2004), pp. 660–669.

[5] T. Ideker and D. Lauffenberger, *Building with a scaffold: emerging strategies for high to low-level cellular modeling*, Trends in Biotechnology, 21 (2003).

[6] D. R. Rhodes, T. R. Barrette, M. A. Rubin, D. Ghosh, and A. M. Chinnaiyan, *Meta-analysis of microarrays: interstudy valication of gene expression profiles reveals pathway dysregulation in prostate cancer*, Cancer Research, 62 (2002), pp. 4427–4433.

[7] E. Segal, R. Yelensky, and D. Koller, *Geneome-wide discovery of transcriptional modules from DNA sequence and gene expression*, Bioinformatics, 19 (2003), pp. i273–i282.

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