Learning robust cell signalling models from high throughput proteomic data

Mitchell Koch

Department of Computer Science,
Rice University,
Houston, TX 77005, USA
E-mail: mkoch@rice.edu

Bradley M. Broom

Department of Bioinformatics and Computational Biology,
M.D. Anderson Cancer Center,
Houston, TX 77030, USA
E-mail: bmbroom@mdanderson.org

Devika Subramanian*

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

Abstract: We propose a framework for learning robust Bayesian network models of cell signalling from high-throughput proteomic data. We show that model averaging using Bayesian bootstrap resampling generates more robust structures than procedures that learn structures using all of the data. We also develop an algorithm for ranking the importance of network features using bootstrap resample data. We apply our algorithms to derive the T-cell signalling network from the flow cytometry data of Sachs et al. (2005). Our learning algorithm has identified, with high confidence, several new crosstalk mechanisms in the T-cell signalling network. Many of them have already been confirmed experimentally in the recent literature and six new crosstalk mechanisms await experimental validation.

Keywords: systems biology; Bayesian networks; T-cell signalling; crosstalk; flow cytometry; bioinformatics.

Reference to this paper should be made as follows: Koch, M., Broom, B.M. and Subramanian, D. (2009) 'Learning robust cell signalling models from high throughput proteomic data', *Int. J. Bioinformatics Research and Applications*, Vol. 5, No. 3, pp.241–253.

Biographical notes: Mitchell Koch is a sophomore (2nd year undergraduate student) in the Computer Science Department at Rice University.

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.

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 peer-reviewed papers.

1 Introduction

Cell signalling forms the core of the communication mechanism that coordinates both intra-cellular and inter-cellular activities. Extracellular stimuli promote the binding of signalling molecules to their receptors, initiating a series of intra-cellular reactions that regulate an individual cell's behaviour, including metabolism, movement, proliferation, survival and differentiation (Cooper, 2000). Signalling molecules that are expressed on the surface of one cell and that bind to receptors expressed by other cells, integrate and coordinate the functions of cells in a complex organism. Understanding the molecular mechanisms responsible for these pathways of cell signalling is a major area of current research. Many diseases, particularly cancers, arise as a result of a breakdown in the signalling pathways that control normal cell proliferation and survival. By mapping signalling pathways altered by disease, we can generate useful therapeutic targets to restore normal cell functioning. Traditional work in biology has focused on studying individual parts of cell signalling pathways. These studies (King et al., 1998) experimentally map out interactions between a small set of biomolecules in a pathway. Complete pathways are assembled by manually aggregating the interaction maps. In contrast, work in systems biology takes a more global view, and attempts to infer signalling networks from simultaneous measurements of a collection of biomolecules involved in the signalling process. The network view of signalling allows consideration of emergent properties, such as crosstalk and bistability, that cannot be modelled at the level of pathway components (Bhalla and Iyengar, 1999).

In this paper, we present an approach for computational mapping of cell signalling networks from high-throughput proteomic data (Sachs et al., 2005). For many cell signalling systems, there is not enough experimental data on the signalling proteins and the reactions in which they participate (e.g., rate constants) to construct detailed differential equation-based models. For such systems, an intermediate modelling methodology based on the probabilistic framework of Bayesian networks, is more appropriate. We use Bayesian networks to represent signalling systems, and we learn both the structure and the parameters of the networks from data. Bayesian networks represent probabilistic dependence relationships between multiple interacting components. Directed edges in a Bayesian network represent stochastic dependencies between biomolecules and their parents in the network. The network structure also encodes conditional independence between biomolecules using the

graph-theoretic notion of d-separation (Lauritzen, 1996). Standard algorithms (Friedman et al., 1999; Ellis and Wong, 2008; Eaton and Murphy, 2007) that infer the structure of Bayesian networks, heuristically maximise a scoring function based on the posterior probability of the learned structure with respect to the available data. When data is noisy, as is typical of proteomic data, it is not clear that the best models are networks that are most likely with respect to *all* of the available data. Here we demonstrate that model averaging using Bayesian bootstrap resampling generates more robust structures than procedures that learn structures using all of the data. We also develop an algorithm for ranking the importance of network features using bootstrap resample data. This algorithm helps us identify key edges in the learned network.

We demonstrate the effectiveness of our algorithm in the context of deriving the T-cell signalling network from the flow cytometry data of Sachs et al. (2005). While individual pathways in T-cell signalling are relatively well understood, our knowledge of how they interact with each other is incomplete. Our learning algorithm has identified, with high confidence, several new crosstalk mechanisms in the T-cell signalling network. Many of them have already been confirmed experimentally in the recent literature. Six new crosstalk mechanisms proposed by our algorithm, await experimental validation. Our algorithm allows robust assembly of cell signalling networks from high-throughput data, and also serves as an effective hypothesis generator for experimental investigation of crosstalk mechanisms in signalling networks.

Our paper is organised as follows. Section 2 describes our structure learning algorithm based on Bayesian model averaging, as well as our procedure for ranking edges in the learned network. In Section 3 we present our computational results on the T-cell signalling network, as well as results validating our learned model. We conclude in Section 4 with several new computationally hypothesised crosstalk mechanisms for the T-cell signalling network.

2 Computational approach

Bayesian networks offer flexible and modular representations of complex multivariate distributions. They have been widely used in the literature to model gene regulatory networks (Friedman et al., 2000; Yu et al., 2004), metabolic networks (Broom et al., 2006), as well as cell signalling networks (Sachs et al., 2005). They can computationally simulate the modelled networks, and these simulations can be used to generate experimentally testable hypotheses about the functioning of the network. Because of their stochastic nature, they are particularly well suited for learning networks from noisy, incomplete data.

2.1 Modeling signalling networks using Bayesian networks

A Bayesian network models a cell signalling network as a joint probability distribution over variables denoting the expression levels of all the biomolecules in that network. It is simply a compact, graphical representation of that full joint distribution. The graph structure reflects conditional independence relationships between the biomolecules. A Bayesian network for a set $\mathbf{X} = \{X_1, \ldots, X_n\}$ of n discrete random variables is a pair $\mathcal{B} = (\mathcal{G}, \Theta)$ where $\mathcal{G}$ is a *directed acyclic* graph whose vertices represent

the random variables $X_1, \ldots, X_n$, and whose edges represent direct dependencies between these variables. Θ represents the set of conditional probability distributions of the form $\Theta_{X_i \mid \mathbf{U}_i} = P(X_i \mid \mathbf{U}_i), 1 \leq i \leq n$, where $\mathbf{U}_i$ denotes the parents of variable X_i in the graph $\mathcal{G}$. The joint probability distribution $\mathbf{P}(X_1, \ldots, X_n)$ encoded by $\mathcal{B}$ can be reconstructed as the product of the individual conditional probability distributions in Θ:

$$\mathbf{P}(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \mathbf{U}_i).$$

2.2 Learning network structure from data

The problem of learning a network from data is normally posed as an optimisation problem: Given a set $\mathcal{D} = \{\mathbf{x}^{(j)} \mid 1 \leq j \leq N\}$ of instances drawn from a multivariate joint probability distribution $\mathbf{P}(\mathbf{X})$, find a network $\mathcal{B}^* = (\mathcal{G}^*, \Theta^*)$ which maximises the posterior probability of the network given the data:

$$\begin{aligned} \mathcal{B}^* = \text{argmax}_{\mathcal{B}} P(\mathcal{B} \mid \mathcal{D}) &= \text{argmax}_{\mathcal{B}}\ P(\mathcal{D} \mid \mathcal{B}) P(\mathcal{B}), \\ &= \text{argmax}_{\mathcal{G},\Theta}\ P(\mathcal{D} \mid \mathcal{G}, \Theta) P(\mathcal{G}, \Theta), \\ &= \text{argmax}_{\mathcal{G},\Theta}\ P(\mathcal{D} \mid \mathcal{G}, \Theta) P(\mathcal{G}) P(\Theta \mid \mathcal{G}). \end{aligned}$$

The first term above is the likelihood of the data given the network structure $\mathcal{G}$ and its parameters Θ. The second and third terms are priors: a discrete probability distribution $P(\mathcal{G})$ over graph structures $\mathcal{G}$, and for each possible graph, a density measure $P(\Theta \mid \mathcal{G})$ over possible values of the parameter Θ. The most popular choice for the prior $P(\mathcal{G})$ over graphs is the uniform distribution, and for $P(\Theta \mid \mathcal{G})$ is the Dirichlet distribution. The parameter priors for each variable are considered independent of one another, an assumption called global parameter independence in Friedman and Koller (2003),

$$P(\Theta \mid \mathcal{G}) = \prod_{i=1}^{n} P(\Theta_{X_i \mid \mathbf{U}_i}).$$

In addition, we make the parameter modularity assumption (Heckerman et al., 1995). For two graph structures $\mathcal{G}$ and $\mathcal{G}'$ such that the parents of node X_i are the same in both graphs, we have

$$P(\Theta_{X_i \mid \mathbf{U}_i} \mid \mathcal{G}) = P(\Theta_{X_i \mid \mathbf{U}_i} \mid \mathcal{G}').$$

The graph structure prior $P(\mathcal{G})$ is assumed to satisfy structural modularity; that is, it can be factored as a product of distributions over the possible parent sets of node X_i in the graph. The posterior probability of the graph structure with these assumptions reduces to

$$P(\mathcal{G} \mid \mathcal{D}) = P(\mathcal{G}) \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(N'_{ij})}{\Gamma(N'_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(N'_{ijk} + N_{ijk})}{\Gamma(N'_{ijk})},$$

where q_i is the number of values for the parents of node i, r_i is the number of values for node i itself, N'_{ijk} is the Dirichlet distribution order for variable i with value k

and parent value j, and $N'_{ij} = \sum_{k=1}^{r_i} N'_{ijk}$. N_{ijk} is the number of instances in the data set $\mathcal{D}$ where variable i with value k has parents with value j and $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$. In this paper, we assume that the Dirichlet orders for all sets of parameters (N'_{ijk}) is a constant $\lambda \geq 1$ as in Yang and Chang (2002). The choice of N'_{ijk} is critical, particularly for small data sets. If it is large, the N'_{ijk} values dominate the N_{ijk} values, making the available data have less influence in determining the space of structures explored. The posterior probability $P(\mathcal{G} \mid \mathcal{D})$ together with a penalty adjustment to account for the complexity of $\mathcal{G}$, is used as a scoring function by structure learning algorithms. The scoring function can be viewed as a measure of how well the network explains the data.

The inference of network structure from data is the most interesting aspect of Bayesian network learning. It allows for the identification of real dependencies between the measured biomolecules, as opposed to simple correlations (Needham et al., 2007). The problem of learning a network which maximises a given scoring function is a combinatorial optimisation problem. The number of model structures is super-exponential in the number of variables, making enumeration-based approaches impractical. In fact, the problem of learning a network which maximises a scoring function is known to be NP-complete (Chickering, 1996). There are two approaches to finding approximate solutions: direct search for the structure guided by a scoring function, and sampling structures from the posterior distribution $P(\mathcal{G} \mid \mathcal{D})$. We use direct search in structure space in this paper. The local search algorithm used in our experiments is a greedy hill climber with randomised restarts. It starts with a specially generated family of initial networks (detailed below), then iteratively adds, deletes or reverses an edge, scores the resulting network at each stage, continuing until a local maximum is found. Every network explored by the algorithm during the search is recorded, so no network is ever considered twice.

We generate an initial set of networks based on Friedman's sparse candidate algorithm (Friedman et al., 1999) with $k = 6$ (maximum number of parents per node). Initially, the k most likely parents for each node, the candidates, are selected by scoring all networks with a single edge between that node and another. We denote by K, the set of all networks in which the parents of every node belong to the set of k candidate parents for that node. A list of starting networks containing all networks in K with up to two edges is then generated. A starting network is picked at random from this initial list. From this starting network, all neighbouring networks in K that have not been considered before and which differ by an additional edge, one less edge, or a reversed edge are evaluated. The highest scoring neighbouring network, if its score is higher, replaces the current network. The search is continued until no new networks are generated or all generated networks score less than the current network. New sets of k candidate parents are then generated following Friedman's algorithm, and the search is continued. New candidate parents sets are picked until a previously seen set of candidate parents is revisited, or ten different candidate parent sets have been considered. Such searches starting from a randomly picked member of the initial network list are performed a total of 25 times. Another 25 such searches are performed starting from a network chosen randomly from all of those seen during the first 25 searches.

Repeated randomised restarts of the structure space search algorithm yield a collection of 50 networks which are local maxima with respect to the scoring function. The highest scoring networks in this collection are reported. When networks are scored

against noisy data, the highest scoring networks tend to be overfitted, and it is unclear that they are the best models. To insulate ourselves against overfitting, we use bootstrap resampling and learn networks using each resampled data set.

2.3 Model averaging with Bayesian bootstrap

Bootstrap aggregating or bagging (Breiman, 1996), which is a model averaging procedure, has been used to reduce variance and improve robustness of learned networks (Friedman et al., 1999). Suppose we have a data set $\mathcal{D} = \{\mathbf{x}^{(j)} \mid 1 \leq j \leq N\}$, where each $\mathbf{x}^{(j)}$ is a vector of size n drawn from the cross product of the domains of variables $X_1, \ldots, X_n$. The basic idea of the standard non-parametric bootstrap is to randomly draw datasets with replacement from $\mathcal{D}$, with each sample the same size as the original set, that is N. This is done B times, producing B bootstrap replicates. We learn Bayesian networks from each bootstrap resample. We then average the adjacency matrices of the networks generated over the B resamples, producing estimates of posterior probabilities $P(e \mid \mathcal{D})$ for all edge features. Given a threshold $0 \leq t \leq 1$, we generate a thresholded graph by including only those edges whose posterior probabilities exceed t ($P(e \mid \mathcal{D}) > t$). We have experimentally determined that it takes 2500 resamples for estimates of these posterior probabilities to stabilise.

Each instance in $\mathcal{D}$ is represented between 0 and B times among the bootstrap resamples. Thus one can think of the standard bootstrap procedure as assigning each example in $\mathcal{D}$ an integer weight drawn from a multinomial distribution, representing its number of occurrences in the B resamples. The probability of not including a specific instance in a resample is about $1/e \approx 37\%$. Since an instance contributes to the count N_{ijk} in the scoring function; dropping instances biases the counts, as well as the structures that are learned from them.

Therefore, we use the Bayesian bootstrap which is a continuous analog of the discrete bootstrap (Rubin, 1981). Bayesian bootstrap is a resampling procedure that is operationally similar to the standard non-parametric bootstrap. In the Bayesian bootstrap, examples are assigned continuously varying weights drawn from a Dirichlet distribution. The Bayesian bootstrap procedure has a Bayesian interpretation. Assume that instances are drawn from some unknown multivariate distribution $P(\mathbf{X})$, and that we have no specific priors on that distribution. The uninformative prior on P combined with the multinomial sample likelihood yields, via Bayes Theorem, a Dirichlet posterior distribution on the fraction of the original population that each sampled instance represents. The ensemble of Bayesian bootstrap resamples, and the distribution of statistics derived from them, can be viewed as samples from a Bayesian posterior distribution. The continuously varying weights of the Bayesian bootstrap ensure that there is a vanishingly small chance of assigning a zero weight to any instance in a resample. Thus, all of the inter-relationships between examples are preserved in the resample, but in reweighted form. Statistics derived from these resamples do not embody the bias introduced by the discreteness of the standard bootstrap.

2.4 Testing the robustness of the model

The best locally optimal network learned from all of the data may not necessarily be the best model for the underlying biological system, especially when the data gathered from the system is noisy and sparse. Bayesian bootstrap aggregation alleviates some of the overfitting that arises from noise in the data. To further test the robustness

of the learned network, we delete elements of the network in turn to determine their importance to the model. We focus on edge elements because direct edges are suggestive of underlying causal relationships. The deletion score of an edge e in a network G is defined as the difference between the score of the network G without e, denoted as G_{-e}, and the score of the original network G.

$$\text{deletionScore}(G, e) = \text{score}(G_{-e}) - \text{score}(G).$$

The score differences for an edge are computed with respect to every bootstrap resample, and overall mean and variance of the deletion score over all the resampled data sets are reported. Important edges have high positive mean and a low variance in the deletion score. We thus use the deletion score to rank the importance of edges in the network.

3 Experimental results

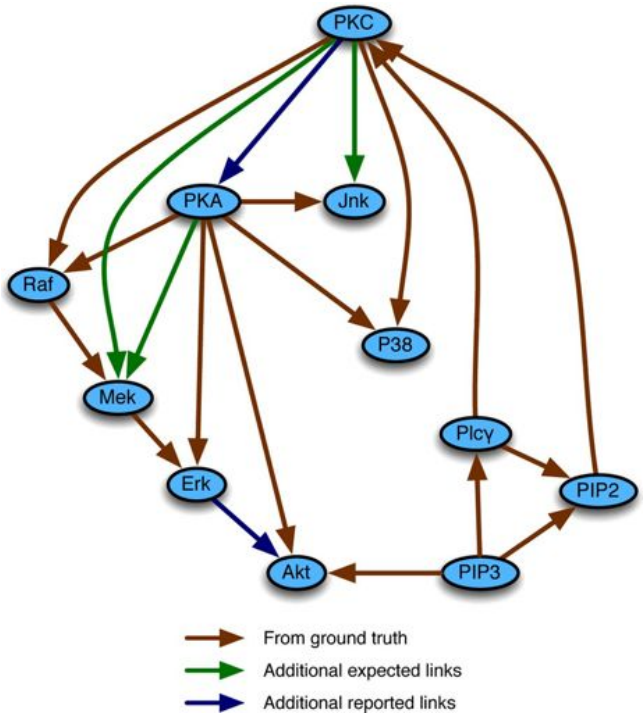
In this paper, we apply our Bayesian network learning strategy to a biological dataset first analysed by Sachs et al. (2005). The dataset was gathered using multicolour flow cytometry to measure proteins in a portion of the human T-cell signalling network. It contains 5400 simultaneous measurements of 11 proteins under 9 interventions, including two general conditions. Although there is a recognised *ground truth* network, it is rather sparse, and there is no general agreement on the interactions between the proteins beyond that, which makes computational analysis a helpful tool for this problem.

Sachs et al. used a Markov Chain Monte Carlo (MCMC) search approach in the space of networks, based on multiple-restart simulated annealing, to construct a computational model. This computational model closely approximates their selected benchmark, shown in Figure 1, which consists of edges from the accepted *ground truth* model, as well as five additional edges selected by Sachs et al. from varying support in the literature, for a total of 20 edges. We call their benchmark the *consensus model*, and compare our results against this model.

Besides Sachs et al. the flow cytometry data was also previously analysed using Bayesian inference methods that sample from the posterior distribution $P(\mathcal{G} \mid \mathcal{D})$; specifically, by Eaton and Murphy (2007) using structure-MCMC, as well as by Ellis and Wong (2008) using order-sampling techniques.

In this section, we report on the results of two experiments. In the first experiment, we apply our learning method to the discretised flow cytometry data of Sachs et al. (2005), leaving the Bayesian inference method as the only difference between us and Sachs et al. (Eaton and Murphy, 2007; Ellis and Wong, 2008). This experiment allows us to directly and fairly compare the networks learned by our method against those generated by the other methods. From the Bayesian bootstrap averaged networks, we generate two thresholded graphs with thresholds 0.6 and 0.99. These networks are shown in Figure 2. The 0.6 threshold is derived from our permutation experiments in Broom et al. (2007), while the 0.99 threshold is chosen as a high threshold bar. Our result graph thresholded at 0.99 contains 22 edges, while the same graph thresholded at 0.6 contains 30 edges. We also compute the deletion scores for all edges in our 30 edge network, and order them according to those scores (see Table 2).

Figure 1 Benchmark T-cell signalling network used by Sachs et al. (2005) to assess the validity of their results. This Bayesian network has 20 edges. The brown edges form the conventionally accepted *ground truth* model of T-cell signalling with respect to the observed biomolecules. The green edges have many references in the literature and are referred to as *expected* by Sachs et al. The blue edges have at least one reference in the literature and are referred to as *reported* by Sachs et al. We refer to this network as the *consensus model* to which we compare our results (see online version for colours)



Seventeen of the 20 edges in the consensus model are in the network we obtain using the high threshold of 0.99. One edge is reversed from an edge in the consensus model (it is actually the same edge that was reversed in Sachs et al.'s own results, Plcγ to PIP3). The three edges in the consensus model that are missing from our network are the same three edges missing from Sachs et al.'s reconstruction. We find five edges that do not appear in the consensus model, and we label them as 'new'. When we lower the threshold to 0.6, we add one additional edge from the consensus model (reversed) and seven additional 'new' edges. Table 1 summarises these results alongside those of Ellis and Wong (2008), as well as Eaton and Murphy's result using the perfect intervention model (Eaton and Murphy, 2007).

Although our method differs significantly from those of previous analyses, our results are in close agreement with those of the structure-MCMC approach used by Eaton and Murphy and surpass those of the order-sampling approach used by Ellis and Wong.

In the second experiment, our goal is to confirm that the additional edges learned by our algorithm, over and above the 20 edge consensus model are legitimate, and not mere artifacts of our computation. We assume that the structure of the consensus

model shown in Figure 1 is correct, and learn the parameters of that network from the available data. From this quantitative network, we generate 60 synthetic data sets each containing 5400 data points. For each synthetic data set, we apply the network learning procedure in exactly the same way as we did for the original data set. The network models generated in this way each contain 17 or 18 edges from the consensus model, and are missing two or three edges. In no case are additional edges obtained at threshold values of 0.6 or above. Consequently, we are confident that the additional edges have real support in the data from which they were learned.

Figure 2 The network on the left is our result graph thresholded at 0.99, while the network on the right is the same graph thresholded at 0.6. The graph on the left has 22 edges, while the graph on the right has 30 edges. The edges are annotated to show how they relate to the 20-edge consensus model of T-cell signalling (see online version for colours)

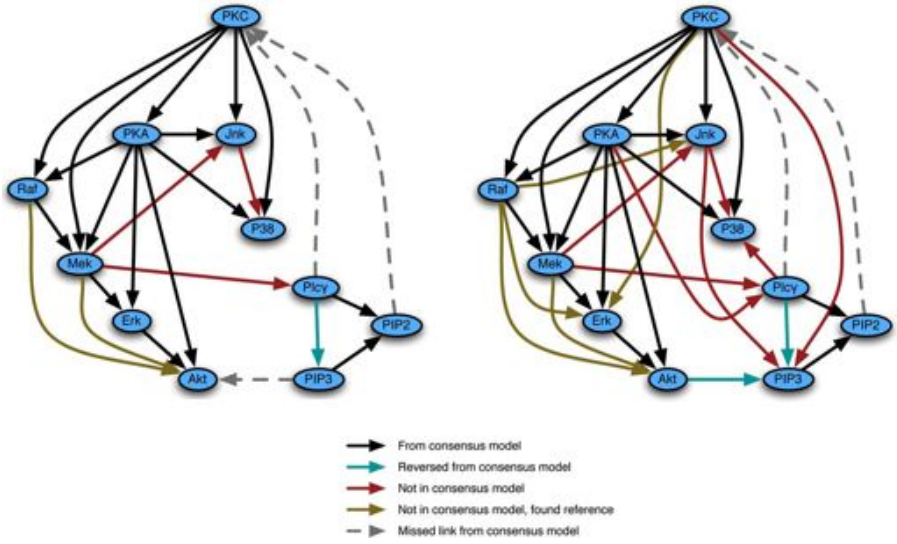


Table 1 Edge comparisons of networks learned by different methods from the same flow cytometry data. The edge labels (correct, reversed, new, missing) are relative to the consensus model shown in Figure 1. t is the threshold for edge features in our bootstrap averaged model

	Ellis and Wong	*Eaton and Murphy*	*Our method* $t = 0.99$	*Our method* $t = 0.6$
Correct	8	16	16	16
Reversed	4	2	1	2
New	8	9	5	12
Missing	8	2	3	2

It is possible that these additional edges are the result of noise in the original flow cytometry data or flaws in the discretisation of the data. However, since we have a relatively large number of data points (5400), as well as a large number of bootstrap resamples (2500), we consider this alternate explanation to be somewhat unlikely.

In future work, we plan to investigate the robustness of our results with respect to perturbations of the discretisation technique.

Table 2 shows the importance of each of the 30 edges in our result graph thresholded at 0.6, as measured by their deletion scores. It is interesting to note that eight of the top ten edges in the network are already in the consensus model. There are 11 edges in our network that are not present in the generally accepted model. We have found confirmation for five of these edges in the recent literature. Especially noteworthy are the transitive edges `Raf` to `Akt`, `Raf` to `Erk`, and `Mek` to `Akt`, all of which are confirmed in Dougherty et al. (2005) and Kolch (2000) as playing critical feedback roles in the well known `Raf` $\rightarrow$ `Mek` $\rightarrow$ `Erk` $\rightarrow$ `Akt` pathway.

Table 2 The means and standard deviations of the deletion scores of the 30 edges in the thresholded graph obtained from the bootstrapped averaged graph thresholded at 0.6. The means and variances of the deletion scores are computed over the bootstrap resamples used in the learning process

From	*To*	Δ*Score*	*Status*
Plcγ	PIP2	1143 ± 44	consensus
PKC	PKA	885 ± 40	consensus
Erk	Akt	858 ± 42	consensus
PKA	Raf	660 ± 34	consensus
Mek	Plcγ	541 ± 24	**unknown**
Raf	Mek	433 ± 33	consensus
Mek	Jnk	373 ± 27	**unknown**
Mek	Erk	350 ± 26	consensus
PKC	Jnk	323 ± 32	consensus
PKC	Raf	305 ± 24	consensus
PIP3	PIP2	196 ± 19	consensus
PKA	Jnk	192 ± 23	consensus
PKA	Erk	173 ± 30	consensus
Jnk	P38	137 ± 23	**unknown**
PKA	Mek	134 ± 27	consensus
PKA	Plcγ	134 ± 20	**unknown**
PKC	Mek	115 ± 23	consensus
Raf	Akt	111 ± 20	Kolch (2000) and Jun et al. (1999)
PKA	Akt	86 ± 24	consensus
PKA	P38	57 ± 24	consensus
PKC	P38	57 ± 20	consensus
Plcγ	PIP3	47 ± 25	consensus
Mek	Akt	44 ± 19	Kolch (2000)
Plcγ	PKC	6 ± 15	consensus
Raf	Jnk	-33 ± 16	Adler et al. (2005)
PKC	Erk	-67 ± 18	Besson et al. (2001)
Raf	Erk	-74 ± 18	Dougherty et al. (2005)
PKC	PIP3	-93 ± 16	**unknown**
Jnk	PIP3	-111 ± 15	**unknown**
Akt	PIP3	-132 ± 13	consensus

In fact, the key motif in our network is the transitive triangle formed by three nodes `A`, `B`, and `C`, with edges `A` $\rightarrow$ `B`, `B` $\rightarrow$ `C`, and `A` $\rightarrow$ `C`. The transitive link `A` $\rightarrow$ `C` is generally

a negative regulator of the A $\rightarrow$ B $\rightarrow$ C process. The confirmed new edges all belong to this family. An important transitive triangle that awaits confirmation is the PKA, Jnk and P38 triplet. This triangle suggests crosstalk between the Jnk and P38 pathways that is independent of the upstream PKA biomolecule. We are presently working with our biological collaborators to confirm the existence of this mechanism.

4 Discussion and conclusion

We have demonstrated that our method learns a more complete and robust model of T-cell signalling than previous approaches to the problem on the same flow cytometry data. Our method identifies additional new edges, that are not present in the consensus network. Five of them have experimental support in the literature. Further investigation of these edges is likely to refine our understanding of the T-cell signalling network. Using synthetic data generated from a model of the consensus network, we showed that it is exceptionally unlikely that these additional edges are artifacts of the learning process, and that they all have significant support in the flow cytometry data from which they were learned. The major goal of our future research, is to investigate these new edges by conducting interventional experiments on T-cells in a lab with our biological collaborators.

Our computational results point to incompleteness in our understanding of the theory of T-cell signalling. Transitive edges (Raf to Akt, Mek to Akt) are new regulatory edges discovered by our algorithm which have been confirmed by recent literature. There appears to be significant crosstalk between major cellular signalling pathways, the Jnk to P38 connection is an especially interesting one to validate. The real value of our methods will be established when we are able to experimentally confirm at least one of these predicted links.

Acknowledgement

This work was supported by a grant from the Doerr Foundation.

References

Adler, V., Qu, W., Smith, S.J., Iztova, L., Pestka, S., Kung, H.F., Lin, M., Friedman, F.K., Chie, L., Chung, D., Boutjdir, M. and Pincus, M.R. (2005) 'Functional interactions of Raf and MEK with JNK result in a positive feedback loop on the oncogenic Ras signaling pathway', *Biochemistry*, Vol. 44, No. 32, pp.10784–10795.

Besson, A., Davy, A., Robbins, S.M. and Yong, V.W. (2001) 'Differential activation of ERKs to focal adhesions by PKC epsilon is required for PMA-induced adhesion and migration of human glioma cells', *Oncogene*, Vol. 20, No. 50, pp.7398–7407.

Bhalla, U. and Iyengar, R. (1999) 'Emergent properties of networks of biological signaling pathways', *Science*, Vol. 283, pp.381–387.

Breiman, L. (1996) 'Bagging predictors', *Machine Learning*, Vol. 24, pp.123–140.

Broom, B.M., Do, K.A. and Subramanian, D. (2007) *Model Averaging Strategies for Structure Learning with Limited Data using Bayesian Networks*, Technical Report, Department of Computer Science, Rice University, December.

Broom, B.M., McDonnell, T.J. and Subramanian, D. (2006) 'Predicting altered pathways using extendable scaffolds', *International Journal of Bioinformatics Research and Applications*, Vol. 2, No. 1, pp.3–18.

Chickering, D.M. (1996) 'Learning Bayesian networks is NP-complete', *Learning from data: Artificial Intelligence and Statistics V*, Springer Verlag, pp.121–130.

Cooper, G.M. (2000) *Cooper. Cell: A Molecular Approach*, 2nd ed., ASM Press, Sinauer Associates, Inc., Massachusetts.

Dougherty, M.K., Miller, J., Ritt, D.A., Zhou, M., Zhou, X.Z., Copeland, T.D., Conrads, T.P., Veenstra, T.D., Lu, K.P. and Morrison, D.K. (2005) 'Regulation of Raf-1 by direct feedback phosphorylation', *Molecular Cell*, Vol. 17, No. 2, pp.215–224.

Eaton, D. and Murphy, K. (2007) *Exact Bayesian Structure Learning from Uncertain Interventions*, AI & Statistics.

Ellis, B. and Wong, W.H. (2008) 'Learning causal Bayesian network structures from experimental data', *Journal of the American Statistical Association*, Vol. 103, No. 482, June, pp.778–789(12).

Friedman, N. and Koller, D. (2003) 'Being Bayesian about network structure: a Bayesian approach to structure discovery in Bayesian networks', *Machine Learning*, Vol. 50, pp.95–126.

Friedman, N., Goldszmidt, M. and Wyner, A. (1999) 'Data analysis with Bayesian networks: a bootstrap approach', *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp.196–205.

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

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

Heckerman, D., Geiger, D. and Chickering, D.M. (1995) 'Learning Bayesian networks: the combination of knowledge and data', *Machine Learning*, Vol. 20, No. 3, pp.197–243.

Jun, T., Gjoerup, O. and Roberts, T.M. (1999) 'Tangled webs: evidence of crosstalk between C-Raf-1 and Akt', *Science STKE*, Vol. 1999, No. 13, p.PE1, http://stke.sciencemag.org/cgi/content/abstract/sigtrans;1999/13/pe1

King, A.J., Sun, H., Diaz, B., Barnard, D., Miao, W., Bagrdoia, S. and Marshall, M.S. (1998) 'The protein kinase pak3 positively regulates Raf-1 activity through phosphorylation of serine 338', *Nature*, Vol. 396, pp.180–183.

Kolch, W. (2000) 'Meaningful relationships: the regulation of the Ras/Raf/MEK/ERK pathway by protein interactions', *Biochem. J.*, Vol. 351, pp.289–305.

Lauritzen, S.L. (1996) *Graphical Models*, Clarendon Press, Oxford University Press, Oxford.

Needham, C.J., Bradford, J.R., Bulpitt, A.J. and Westhead, D.R. (2007) 'A primer on learning in Bayesian networks for computational biology', *PLoS Computational Biology*, Vol. 3, No. 8, pp.1409–1416.

Rubin, D.B. (1981) 'The Bayesian bootstrap', *The Annals of Statistics*, Vol. 9, No. 1, pp.130–134.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D.A. and Nolan, G.P. (2005) 'Causal protein-signaling networks derived from multiparameter single-cell data', *Science*, Vol. 308, No. 5721, pp.523–529.

Yang, S. and Chang, K-C. (2002) 'Comparison of score metrics for Bayesian network learning', *IEEE Transactions on Systems, Man and Cybernetics: Part A: Systems and Humans*, Vol. 32, No. 3, pp.419–428.

Yu, J., Smith, V., Wang, P., Hartemink, A. and Jarvis, E. (2004) 'Advances to Bayesian network inference for generating causal networks from observational biological data', *Bioinformatics*, Vol. 20, pp.3594–3603.