

# A Search-Based Approach for Bayesian Inference of the T-cell Signaling Network

Mitchell Koch  
Dept. of Computer Science  
Rice University  
Houston, TX 77005  
Email: mkoch@rice.edu

Bradley Broom  
Dept. of Bioinf. & Comp. Biology  
M.D. Anderson Cancer Center  
Houston, TX 77030  
Email: bmbroom@mdanderson.org

Devika Subramanian  
Department of Computer Science  
Rice University  
Houston, TX 77005  
Email: devika@rice.edu

**Abstract**—We apply a search-based technique for learning high-quality Bayesian networks from proteomic flow-cytometry data for a portion of the human T-cell signaling network. Although Bayesian network models have been learned from this data [1], [2], [3] using Markov Chain Monte Carlo (MCMC) methods that sample from a posterior distribution, we demonstrate that our more comprehensible search-based technique, which uses model averaging by Bayesian bootstrap replicates, provides comparable results. We also show that additional edges learned by our procedure, as well as that of Eaton and Murphy [3] are very unlikely to be artifacts of the network learning technique, and should be investigated further.

## I. INTRODUCTION

Learning models of biological networks from data is a key component of the systems biology approach. The learning process can be divided into two components: learning the structure of the network, and learning the parameters of the network. In this paper, we consider only the first part: learning the network structure. Bayesian networks are a widely used approach for learning network structure from data. In a Bayesian network, every edge represents a statistical dependency between two nodes that cannot be explained by any other edges in the network. Conversely, the absence of an edge indicates that any statistical dependency that exists between the nodes can be explained by the edges that are present. We have previously described a strategy [4] for learning a Bayesian network model from data. This strategy combines Bayesian bagging, the Dirichlet Prior Scoring Metric (DPSM), a greedy hill-climbing search algorithm with randomized restarts, and Friedman’s sparse candidate algorithm [5]. For large datasets, our strategy is as effective as other reported strategies, and for small datasets it is significantly more effective.

In this paper, we apply our Bayesian network learning strategy to a biological dataset first analyzed by Sachs et al. [1]. The dataset was gathered using multicolor

flow-cytometry to measure proteins in a portion of the human T-cell signaling network. It contains information about 11 proteins using 8 interventions and 2 general perturbation measurements. Although there is a recognized “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 an a Markov Chain Monte Carlo (MCMC) approach based on multiple-restart simulated annealing in the space of networks to construct a computational model of the T-cell signaling network. In addition to the accepted model at the time, their computational analysis identified two additional edges that they then confirmed experimentally. Besides Sachs et al., this data was also previously analyzed using Bayesian inference methods that sample from the posterior distribution; specifically, by Eaton & Murphy [3] using structure-MCMC, as well as by Ellis & Wong [2] using order-sampling techniques.

## II. APPROACH

The most fundamental difference between our method and those used to analyze the dataset previously is that we use a method that does not sample from the posterior distribution. Instead, at its core, our technique is based around a greedy hill-climbing search algorithm in the space of structures. We use the well-established scoring function, DPSM with low values of  $\lambda$ . Although we used  $\lambda = 0.1$  in analyzing datasets with fewer than 1000 examples, as the Sachs’ discretized dataset contains 5400 examples, we used  $\lambda = 1$ . We use model averaging by Bayesian bootstrap aggregation [6], which allows learning from a more diverse population of structures by averaging over the single high-scoring Bayesian network from each resample.

Together, these additional enhancements form a method sophisticated enough to compete with methods that sample from the posterior distribution, while also

providing several theoretical benefits over those methods. Our method also has different parameters than those in MCMC methods, parameters whose effects on the end result can be readily understood such as number of restarts, maximum number of parents  $k$ , the  $\lambda$  value of DPSM, and the number of Bayesian bootstrap resamples over which to average.

We developed a Bayesian network inference package, `gbayesnet`, to implement this method and added the capability to deal with interventions according to the perfect intervention model proposed by Eaton & Murphy [3] in order to apply it to this dataset. It is available from the authors upon request.

### III. EXPERIMENTAL RESULTS

In this paper, we report the results of two related experiments. In the first experiment, we applied our method to data obtained by Sachs et al. We used the data as already discretized by Sachs et al., leaving the Bayesian inference method as the only difference between our experiment and those of Sachs et al., Eaton & Murphy, and Ellis & Wong. In the second experiment, we applied our method to synthetic data generated from a consensus model. For both experiments we applied the method previously described, using the DPSM scoring metric with  $\lambda = 1$ , averaged over 2500 Bayesian bootstrap replicates, each of which used the sparse candidate algorithm with  $k = 6$  and 100 restarts to search for the highest scoring network(s).

For the average model obtained from the first experiment, we report results for two thresholds: an arbitrarily chosen high threshold of 0.99, which results in a learned network containing 22 edges, and a low threshold of 0.6 which was selected by applying the permutation-based threshold selection method described in [4], and which results in a learned network containing 30 edges.

To assess the closeness of our results (see Figure 1), we compare them with the network that Sachs et al. introduce in Figure 3A of [1]. This network consists of edges from the accepted model, as well as five additional edges with varying support in the literature, for a total of 20 edges. We call this the *consensus model*.

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,  $\text{Plc-}\gamma$  to  $\text{PIP}_3$ ). 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 “incorrect.” When we lower the threshold to 0.6, we add

one additional edge from the consensus model (reversed) and 7 additional “incorrect” edges. Table I summarizes these results alongside those of Ellis & Wong’s “average graph constructed from the average of each subset in the cross validation experiment” [2], as well as Eaton and Murphy’s result using the perfect intervention model. Table II lists, with their frequencies of occurrence, all of the “incorrect” edges that our analysis found more frequently than the 0.6 threshold.

TABLE I  
NUMBER OF EDGES ACCORDING TO THE SACHS CONSENSUS NETWORK

	Ellis & Wong	Eaton & Murphy	gbayesnet $t = 0.99$	gbayesnet $t = 0.6$
Correct	8	16	16	16
Reversed	4	2	1	2
Incorrect	8	9	5	12
Missing	8	2	3	2

TABLE II  
“INCORRECT” EDGES FOUND BY OUR SEARCH PROCEDURE AND THEIR FREQUENCY OF OCCURRENCE.

From	To	Frequency	From	To	Frequency
Raf	Jnk	0.648	PKC	Erk	0.987
Raf	Erk	0.819	Mek	Jnk	0.995
Plcg	P38	0.963	Mek	Plcg	0.999
PKA	Plcg	0.976	Jnk	P38	1.00
PKC	PIP3	0.983	Mek	Akt	1.00
Jnk	PIP3	0.984	Raf	Akt	1.00

Comparing our results to these other analyses, using Sachs’ consensus model as “truth”, we find the same number of correct edges as Eaton & Murphy, with the two networks having 15 edges in common. Eaton & Murphy find one additional (reversed) edge from the consensus model that occurs relatively infrequently in our network. Our high-threshold network contains three fewer incorrect edges than Ellis & Wong’s network and four fewer than Eaton & Murphy’s. When we lower our threshold to 0.6, we find all the “incorrect” edges reported by Eaton & Murphy plus three more.

To summarize the results of this first experiment, our Bayesian network learning procedure constructs network models that compare favorably to the best previously reported analyses of this data set. Most of the “incorrect” edges that we found were also found by Eaton & Murphy, using a very different learning procedure, suggesting that these edges warrant further investigation. In fact, all three models find the Mek to Jnk edge, leading us to conclude that it is an edge especially worthy of experimental investigation.

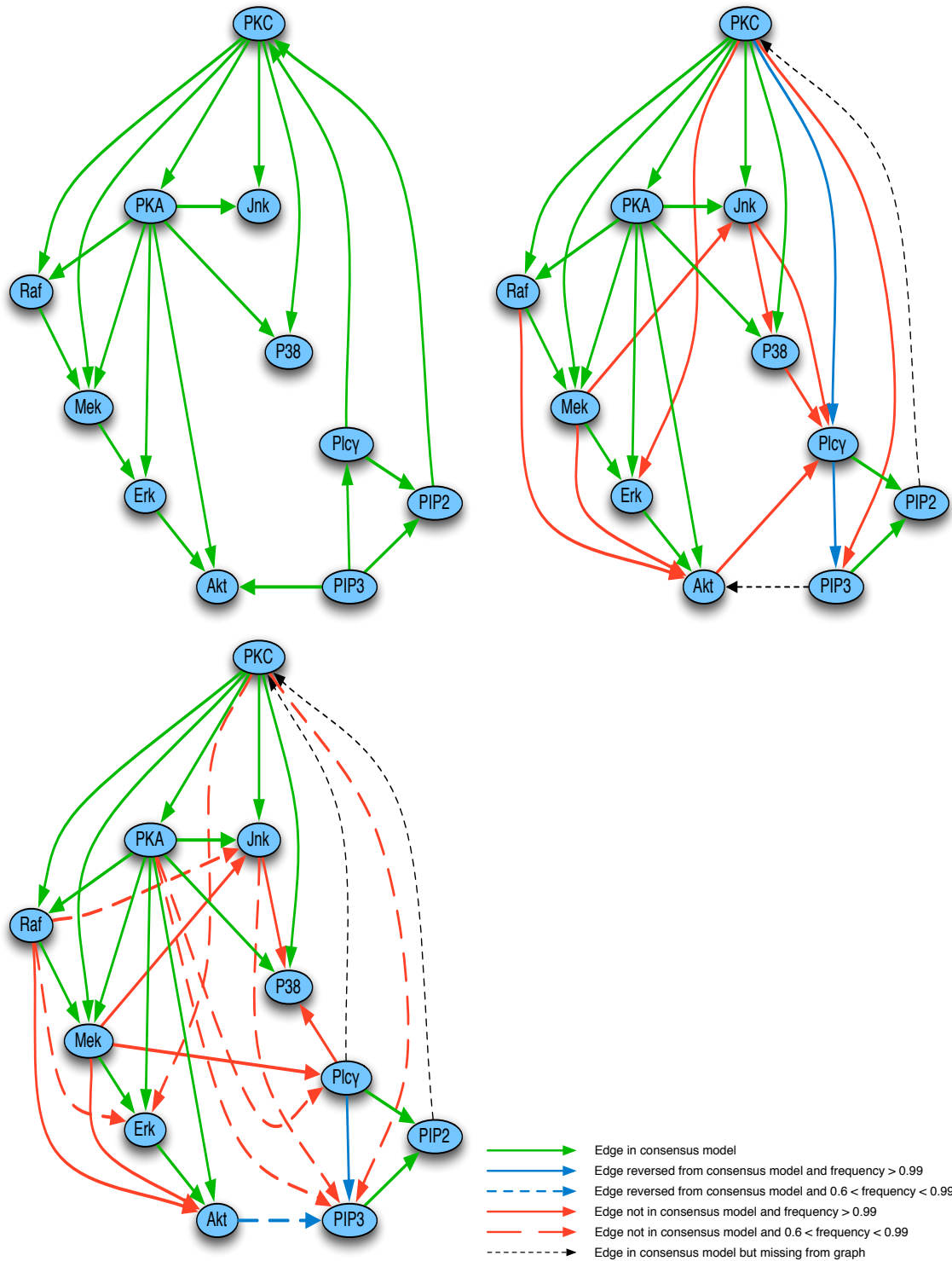


Fig. 1. Network on top left is the consensus model of T-cell signaling derived from Sachs' et. al [1]. Network on top right is derived by Murphy and Eaton using MCMC sampling. Network on bottom left is derived by our method with an edge threshold of 0.6 selected by our permutation test. Green edges match those in the consensus model, blue edges are reversed from those in the consensus model, red edges do not appear in the consensus model, and black dashed edges appear in the consensus model, but not in our result or Eaton and Murphy's.

To determine whether these additional “incorrect” edges are an artifact of the learning procedure, we conducted a second experiment in which we learned network models from synthetic data generated from the consensus model. We assumed that the consensus model was correct, and learned the network’s conditional probability tables (CPTs) from the actual discretized data used above. From that quantitative model, we then generated 60 synthetic data sets each containing 5400 data points. For each synthetic data set, we then applied 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 incorrect edges obtained at a frequency near those that occur in the network model learned from the actual data. Consequently, we are confident that the “incorrect” edges are not simply artifacts of the learning procedure, but have some real support in the (discretized) data from which they were learned.

#### IV. CONCLUSIONS

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

Our method identifies additional edges, five with particularly high confidence, that are not present in the consensus network. Further investigation of these edges is likely to refine our understanding of a well-studied signaling 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 discretized flow cytometry data from which they were learned.

It could still be that there is a non-biological explanation for these additional edges, based on some peculiarities of the discretization of the original flow cytometry data. Given the relatively large number of data points (5400), such an explanation would be somewhat surprising. Nevertheless, we plan to investigate the robustness of our results with respect to perturbations of the discretization schedule.

The major goal of our future research, however, is to investigate those additional edges in which we have the highest confidence, by conducting interventional experiments on T-cells in a lab with our biological collaborator. If these edges are confirmed, our understanding of the T-cell signaling network will be much

improved. Should they not be confirmed, we will have nevertheless obtained a more concrete foundation from which to understand network learning methods and to devise better ones.

#### REFERENCES

- [1] K. Sachs, O. Perez, D. Pe’er, D. A. Lauffenburger, and G. P. Nolan, “Causal protein-signaling networks derived from multiparameter single-cell data,” *Science*, vol. 308, no. 5721, pp. 523–529, 2005.
- [2] B. Ellis and W. H. Wong, “Learning causal Bayesian network structures from experimental data,” 2006. [Online]. Available: <http://www.stanford.edu/group/wonglab/docs/EllisWong-061025.pdf>
- [3] D. Eaton and K. Murphy, “Exact Bayesian structure learning from uncertain interventions,” *AI & Statistics*, 2007.
- [4] B. Broom, K.-A. Do, and D. Subramanian, “Model averaging strategies for structure learning in Bayesian networks with limited data,” *submitted*, 2007.
- [5] 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.
- [6] M. A. Clyde and H. K. H. Lee, “Bagging and the Bayesian bootstrap,” 2000. [Online]. Available: [citeseer.ist.psu.edu/401334.html](http://citeseer.ist.psu.edu/401334.html)