Causal Bayesian network structure learning through effects of interventions to analyze Protein signaling pathways

Meghamala Sinha; Prasad Tadepalli; Stephen Ramsey

School of Electrical Engineering and Computer Science
Oregon State University

Motivation

Living cell molecules interact with each other in a coordinated and complicated fashion to carry out important biological functions. Building a rich network of these interactions can improve recognition of diseases by providing useful mechanistic interpretations of various phenotypes. A lot of discoveries in high-throughput technologies have given rise to numerous algorithms for reverse-engineering networks from molecular observations, as they provide an efficient and systematic way of analyzing the various molecular state and interaction of a number of genes. One class of such interaction networks that has recently generated much interest in recent years is transcriptional gene regulatory networks, which specify the set of genes that influence a given gene's expression level. This type of pattern can be naturally modeled in a causal graph or Bayesian network. In this work, we investigate the performance of our approach by analyzing experimental data from a pollychormatic flow cytometry experiment originally presented by Sachs, Perez, Pe´er, Lauffenburger and Nolan (2005). In this particular experiment, 11 well-studied proteins were selected from the MAPK pathways for fluorescent labeling, showed in Fig 1. This pathway was then perturbed by 9 different stimuli, each targeting a different protein in the selected pathway.

The original data set is continuous data that was processed by first eliminating outliers assuming a log-scale transform. The data were then discretized using an information preserving technique and adjusted such that perturbed vertices always reflect their perturbed values.

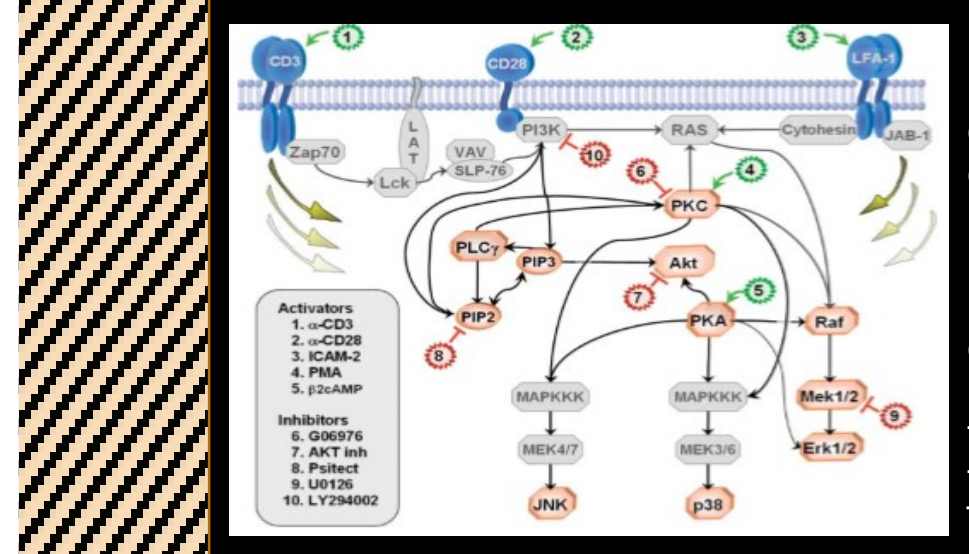
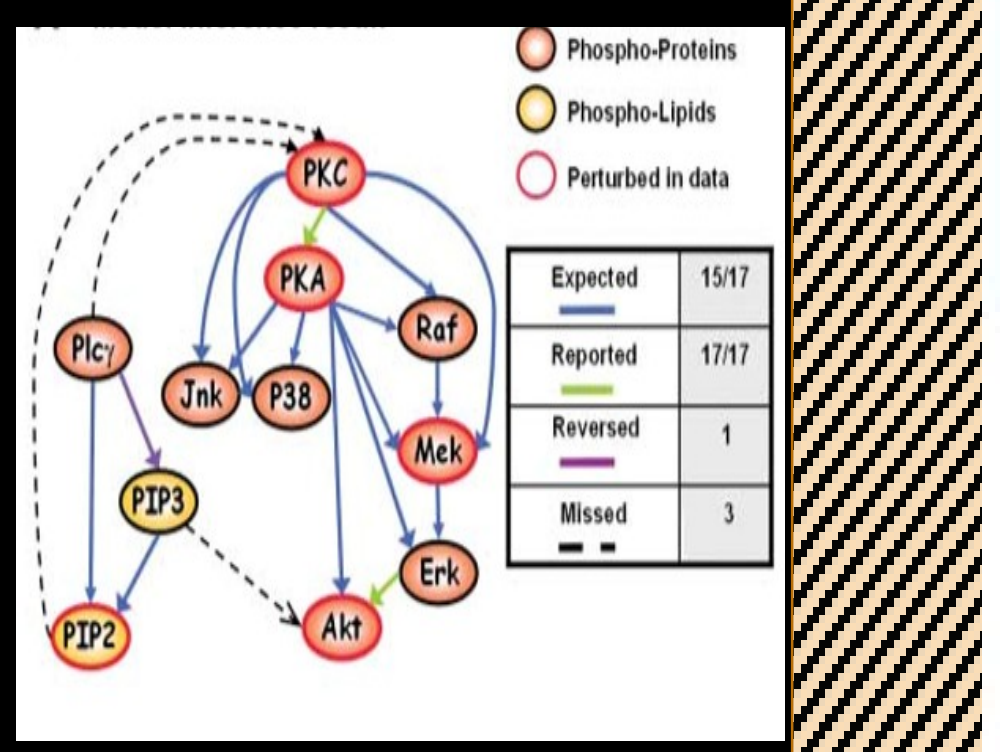


Fig 1. Basic structure of MAPK pathways

Reagent	Effects
Anti-CD3/CD28	General Perturbation
ICAM-2	General Perturbation
b2cAMP	Activates PKA
AKT inhibitor	Inhibits AKT
U0126	Inhibits Mek1/2
PMA	Activates PKC
G06976	Inhibits PKC
Psitectorigenin	Inhibits PIP2
LY294002	Activates AKT

Fig 2. Summary of the 9 experimental stimuli and their effect on the proteins

This is necessary because some of the perturbation methods, particularly the inhibition, only affect the activity of the protein and not their phosphorolation so while the activity of the protein appears to vary, its activity is in fact controlled. After pre-processing, 500 data sets are generated. Each data set consists of 600 cells sampled from each of the 9 experiments. Simulated annealing was used to obtain an optimal DAG from each of the data sets. Thus, after 500 simulated annealing runs, they now have a set of 500 DAGs each with an associated score. To estimate the marginal feature probabilities, Bootstrap samples are generated from this data set according to their scores.



Expected	15/17
Reported	17/17
Reversed	1
Missed	3

Fig 3. Results from Sachs et al experiment.

Introduction

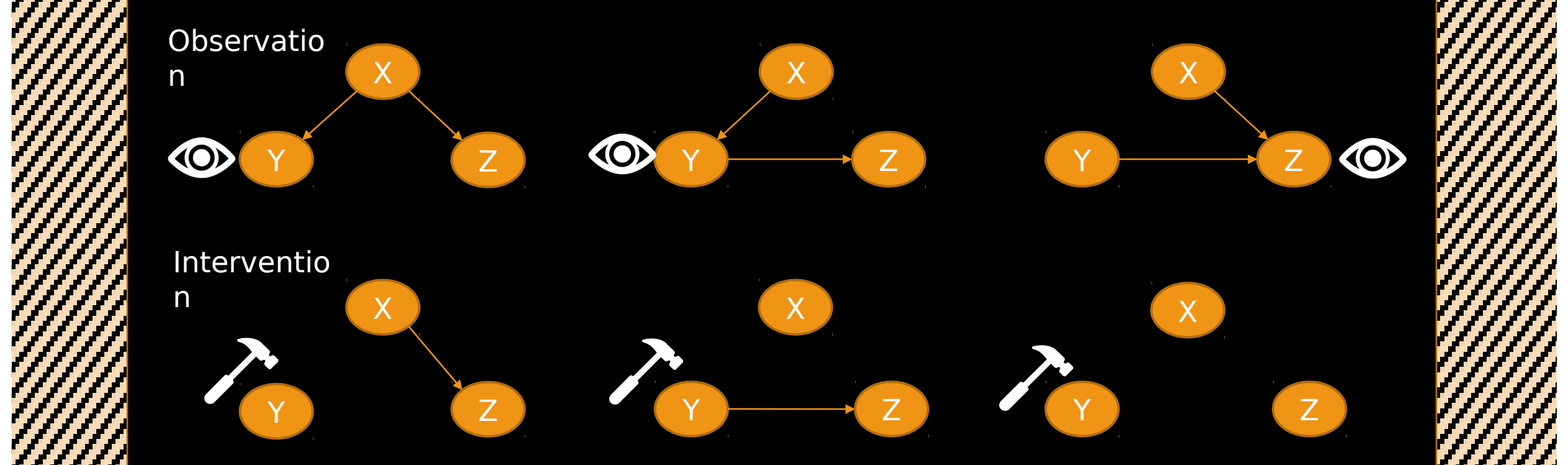


Fig 4. Observation vs Intervention

Observation of Y:
$P(X, Y=1, Z) = P(Z|Y=1) * P(Y=1|X) * P(X)$

Intervention on Y:
$P(X, do(Y=1), Z) = P(Z|Y=1) * P(X)$

Significance of interventions for causal discovery in biological networks:

- Given a directed acyclic graph over a set of variables , an edge *X* → *Y* encodes a causal influence of *X* on *Y*
- For data containing only passive observations of the underlying system, the causal structure is only identifiable up to Markov equivalence classes, it can be claimed as an association
- To overcome this limitation, intervention experiments, in which some variables are controlled to take specific values, can be used to guarantee full identifiability, so given intervention on X, X → Y implies X to be a causal parent of Y
- Living cells and molecules of organisms incorporates a lot of such relationship, which remain intractable

We apply Bayesian causal reconstruction methods over Sachs et al data to analyze the various aspect of intervention in causality.

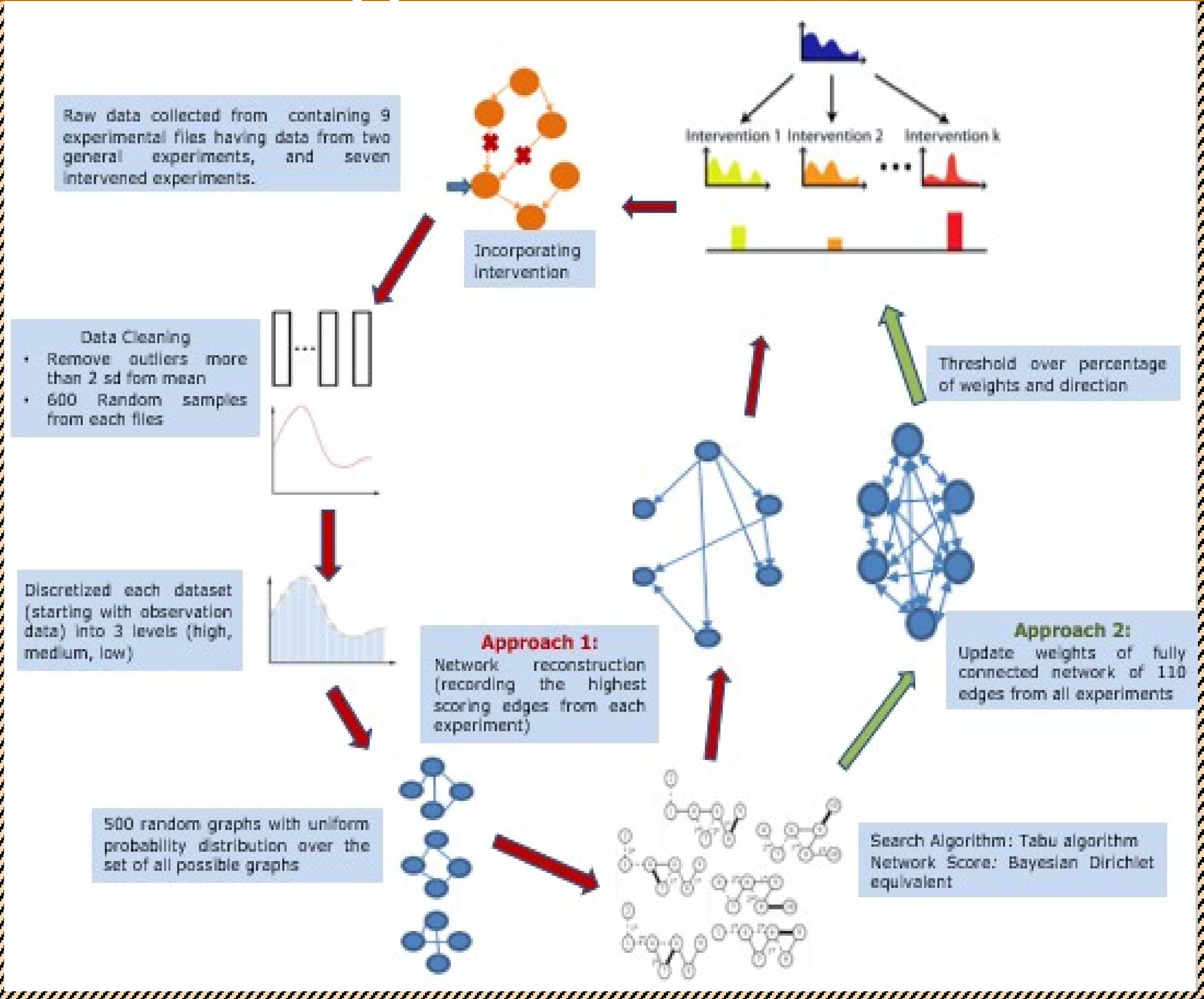
```
Approach 1: Combine most probable arcs learnt from each experiment

Input: $X = \{X_n, n=1,\dots, k\}$, $k^{th}$ interventional experiments dataset
Output: $G' = (X, E', V')$, final reconstructed network

for n=1 to k do
    • generate 500 random equivalent class graphs (given set of nodes) using Melancon's Digraph algorithm
    • Use Tabu search by scoring each edge changes using bde score (excluding the score for arcs directing toward an intervened node) $\{E_n'\} = E_n \setminus E_{V->V_n}$
    • Use Bayesian Model Averaging and record the highest scoring edges to learn a directed graph $G'_n = (X_n, E_n', V_n')$

Combine $\{G'_n, n=1,..,k\}$ to form a merged graph $G' = (X, E', V')$ with $V' = \mathbf{U}_{n=1 \text{ to } k} (V_k)$ and $E' = \mathbf{U}_{n=1 \text{ to } k} (E_k)$

Return G′
```

Fig 5. Our algorithm for Approach 1

```
Approach 2: Threshold over edge-weights from each experiment

Input: $X = \{X_n, n=1,\dots, k\}$, $k^{th}$ interventional experiments dataset
Output: $G' = (X, E', V')$, final reconstructed network

for n=1 to k do
    • generate 500 random equivalent class graphs (given set of nodes) using Melancon's Digraph algorithm
    • Use Tabu search by scoring each edge changes using bde score (excluding the score for arcs directing toward an intervened node) $\{E_n'\} = E_n \setminus E_{V->V_n}$
    • Use Bayesian Model Averaging and add all the edge weights such that $E' = \sum\{E_n'\}$ and $V' = \mathbf{U}_{n=1 \text{ to } k} (V_k)$

Select a threshold T (for example, $\sum E_n'/n$) such that $E_T' | E_n' > T$
 to form a graph $G' = (X, E_T', V')$

Return G′
```

Fig 6. Our algorithm for Approach 2

Approach and Method



Raw data collected from containing 9 experimental files having data from two general experiments, and seven intervened experiments.

Incorporating intervention

Data Cleaning
- Remove outliers more than 2 sd fom mean
- 600 Random samples from each files

Threshold over percentage of weights and direction

Discretized each dataset (starting with observation data) into 3 levels (high, medium, low)

Approach 1: Network reconstruction (recording the highest scoring edges from each experiment)

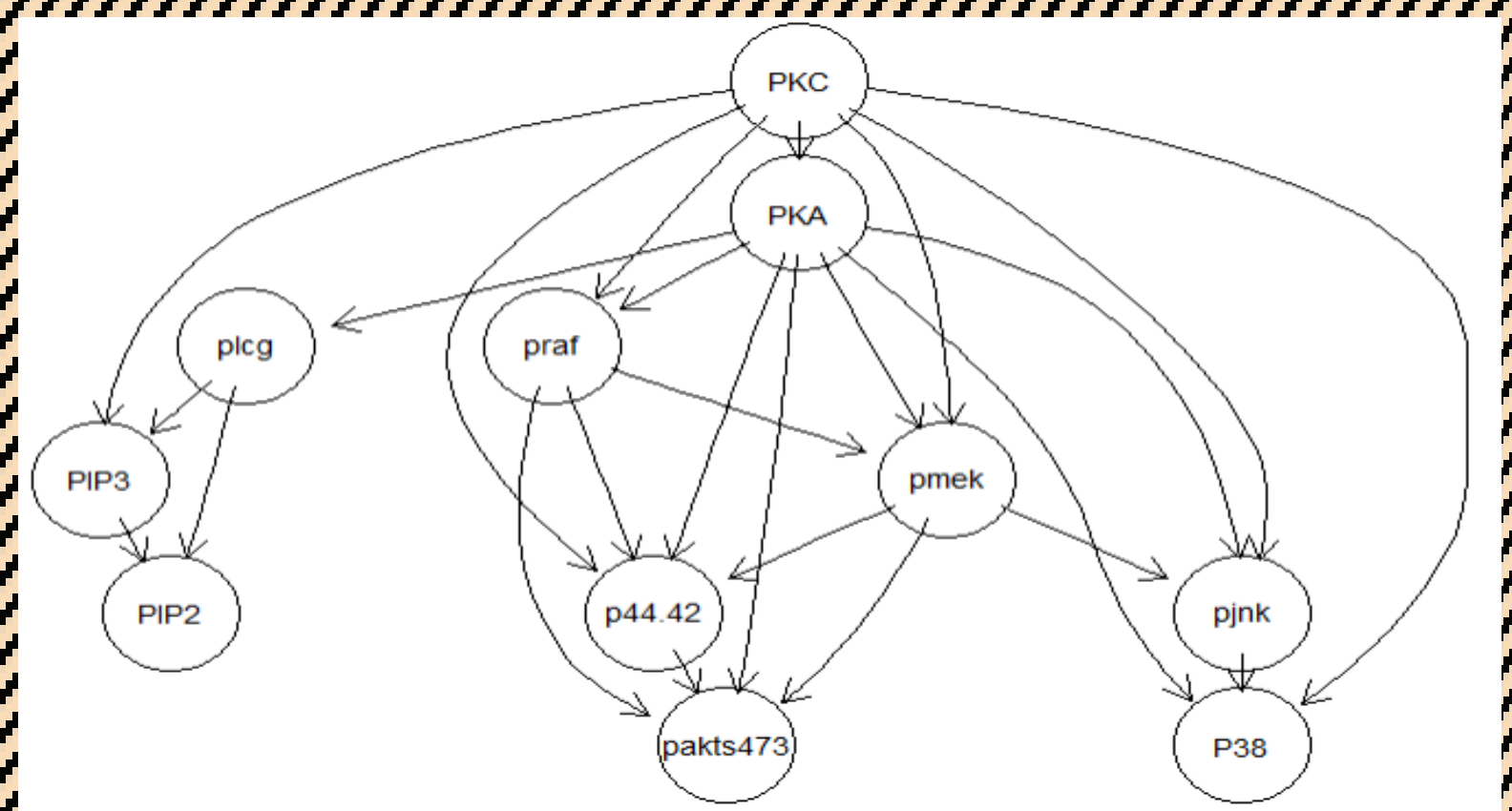
Approach 2: Update weights of fully connected network of 110 edges from all experiments

500 random graphs with uniform probability distribution over the set of all possible graphs

Search Algorithm: Tabu algorithm
Network Score: Bayesian Dirichlet equivalent

Fig 7. Flowchart of our work

Experiments

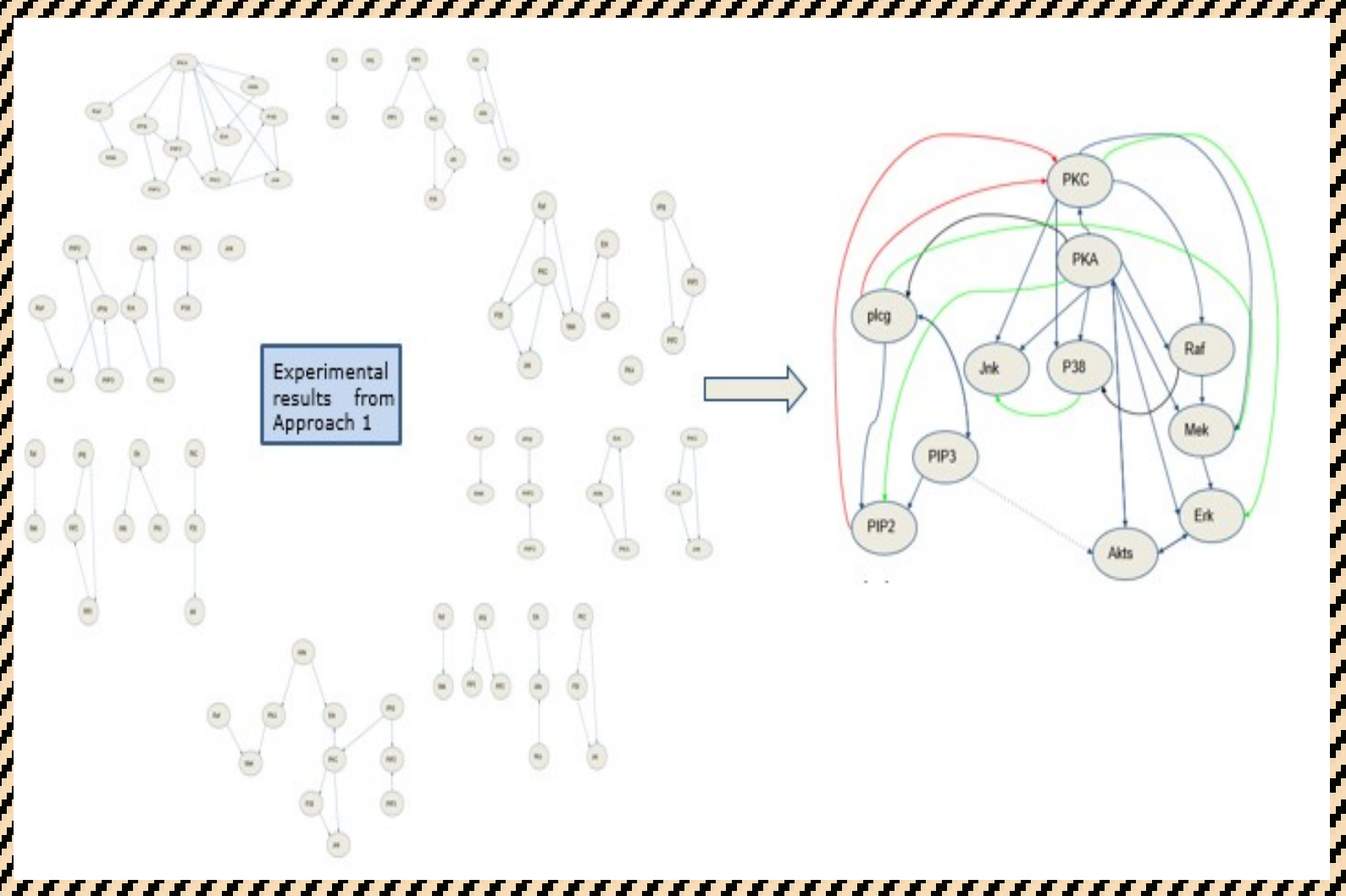


Result from our closest implementation of the Sachs et al method as described in their supplementary material

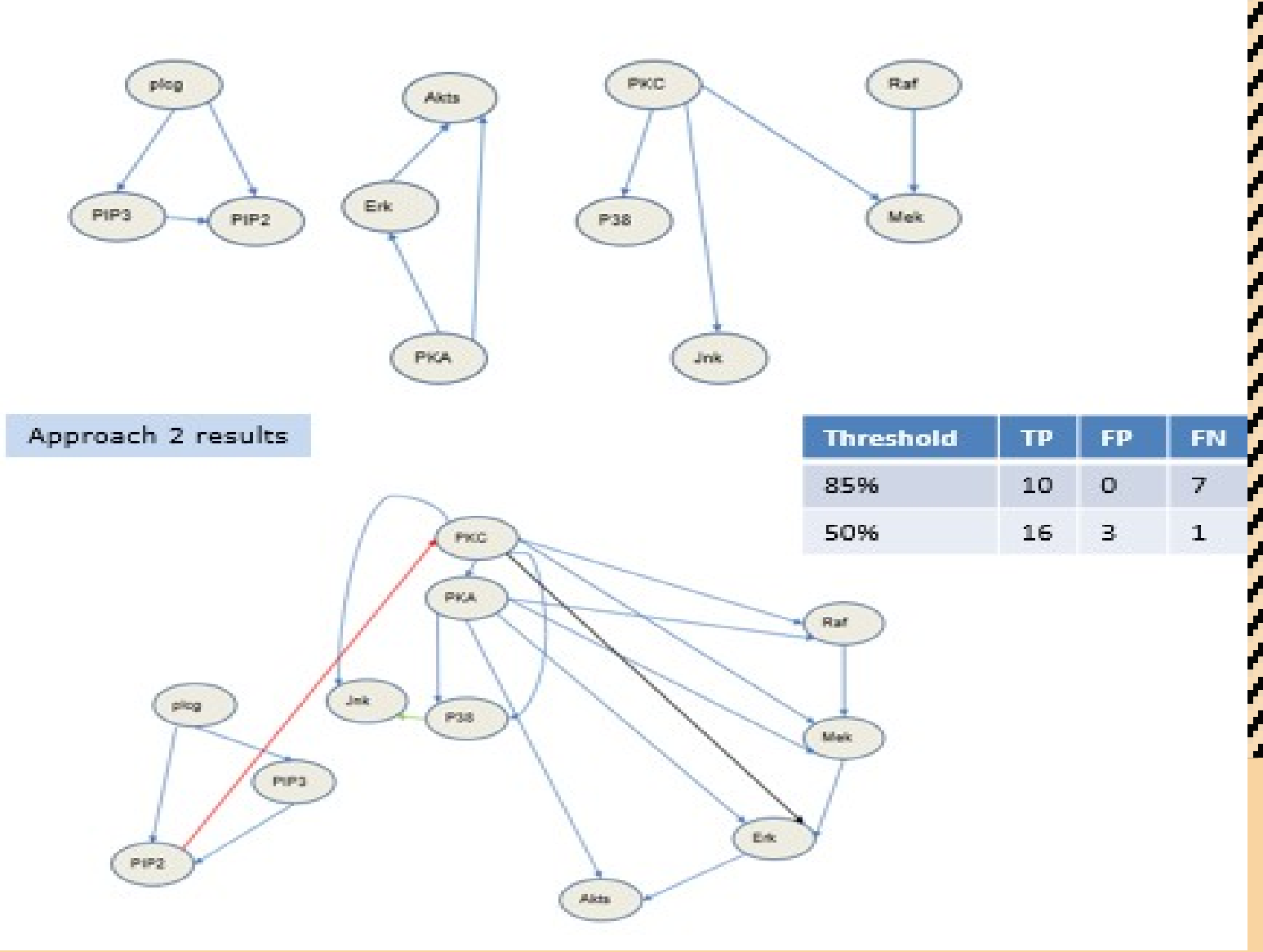
True Positive	17
False Positive	8
False Negative	0

A Significant Improvement !!!

New Arcs found :
P38 -> pjnk (reported in PCViz) ☺
Plcg -> pmek (reported in PCViz) ☺
P44.4 -> PKC (reported in PCViz) ☺
Praf -> P38 ☹
PKA-> plcg ☹
Pip2->PKA (reported in PubMed) ☺

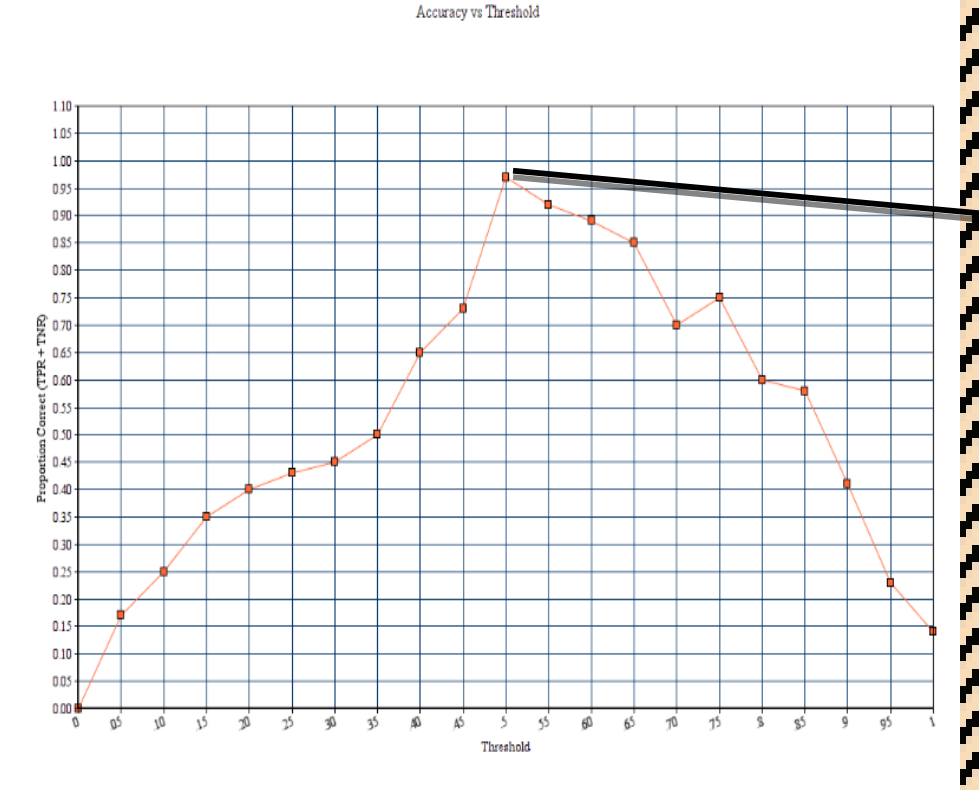


Expected	17/17
New	2
Reported	4/6
Missed	1



Approach 2 results

Threshold	TP	FP	FN
85%	10	0	7
50%	16	3	1



Select an appropriate threshold for sensitivity vs specificity tradeoff !!!

best threshold found

Discussion

The main aim of this work is a deep understanding of the classical work done by Sachs et al in their ground-breaking work. We have thoroughly analyzed the data set used by the authors and recreated their method.

We tried to separate each experiments instead of combining them, in order to make more extensive use of the intervention nodes and hoped to extract more information, since the perturbed nodes were intervened simultaneously changing the states of all the other nodes as well. Hence an intra-experimental method (Fig 7) made more since than inter-experimental way.

The results showed a significant improvement than the classical paper. All the expected nodes found by Sachs et al were also found by our method (Approach 1) including an additional discovery of 2 expected missed node. Most of the newly found arcs were further discovered to be true by literature survey and labelled as reported.

Approach 2 gave a lower accuracy than Approach 1, but the computation time was higher in the later. The accuracy rate for all the methods have been described in Fig 8.

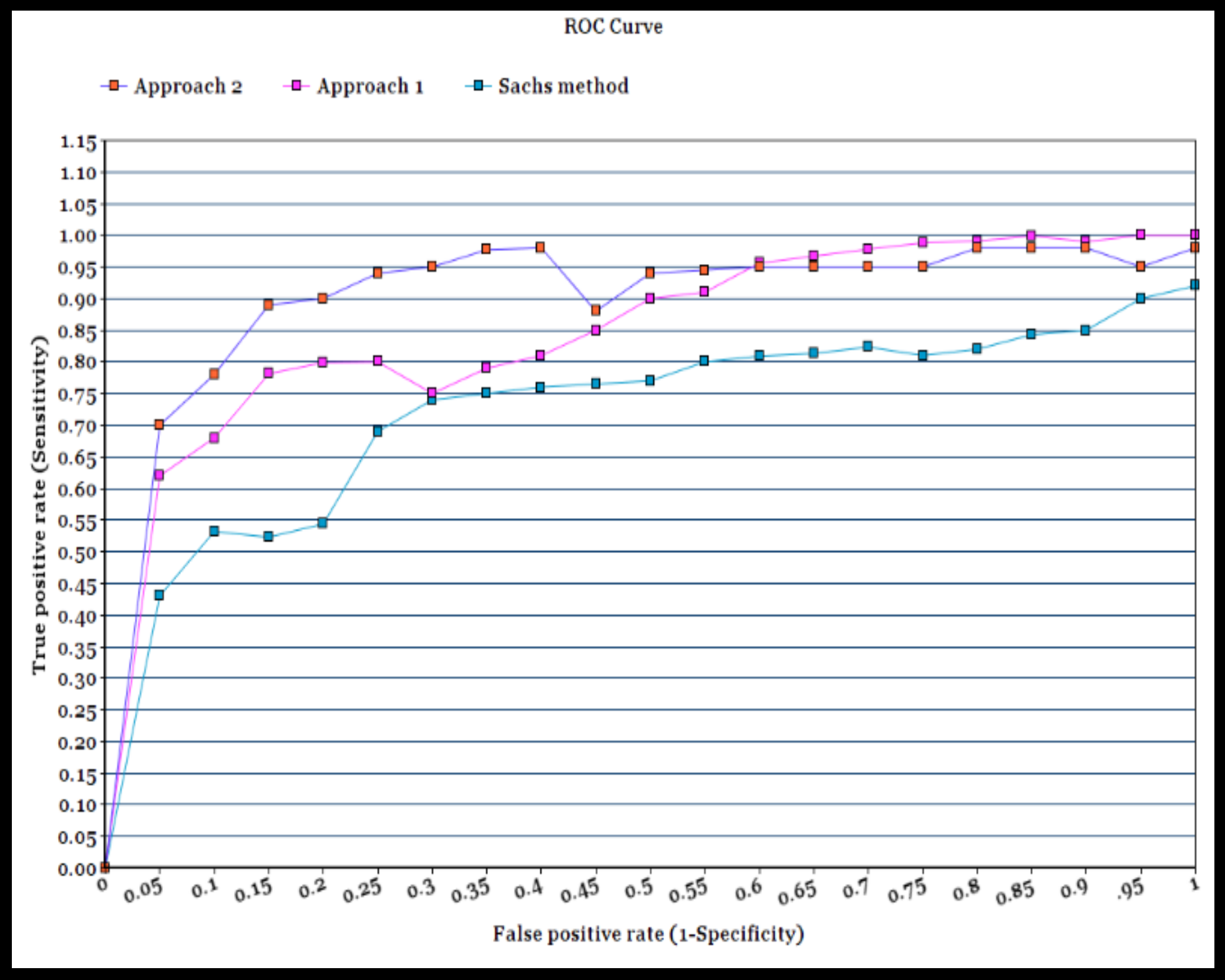


Fig 8. Performance analysis of all the methods

References

- Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., & Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, *308*(5721), 523-529.
- Cho, H., Berger, B., & Peng, J. (2016). Reconstructing causal biological networks through active learning. *PloS one*, *11*(3), e0150611
- He, Y., & Geng, Z. (2016). Causal Network Learning from Multiple Interventions of Unknown Manipulated Targets. *arXiv preprint arXiv:1610.08611*
- Hagmayer, Y., Sloman, S. A., Lagnado, D. A., & Waldmann, M. R. (2007). Causal reasoning through intervention. *Causal learning: Psychology, philosophy, and computation*, 86-100
- Ellis, B., & Wong, W. H. (2008). Learning causal Bayesian network structures from experimental data. *Journal of the American Statistical Association*, *103*(482), 778-789.
- Druzdzel, M. J., & Díez, F. J. (2003). Combining knowledge from different sources in causal probabilistic models. *Journal of Machine Learning Research*, *4*(Jul), 295-316.
- Pearl, J. (2009). *Causality*. Cambridge university press.
- Scutari, M., & Denis, J. B. (2014). *Bayesian networks: with examples in R*. CRC press.

❖ Pubmed : https://www.ncbi.nlm.nih.gov/pubmed/
❖ PCViz: http://www.pathwaycommons.org/pcviz/

Contact:
Meghamala Sinha
Graduate Student (CS)
Email: sinham@oregonstate.edu
1148 Kelley Engineering Center, 2500 NW Monroe Ave, Corvallis, OR 97331