

Title: Learning Causal Protein-Signaling Network From Experimental Data

Ping He

Zhi Geng

Wei Yan

Zhihai Liu

School of Mathematical Sciences,

Peking University

Beijing 100871, China

SUNHP@PKU.EDU.CN

ZGENG@MATH.PKU.EDU.CN

YANWEI1982@PKU.EDU.CN

DEMETRIO@PKU.EDU.CN

Editor:

Task solved: CYTO

Reference:

1. A. J. Hartemink. Principled Computational Methods for the Validation of and Discovery of Genetic Regulatory Networks. Unpublished doctoral thesis, Massachusetts Institute of Technology, 2001.
2. G. F. Cooper and C. Yoo. Causal discovery from a mixture of experimental and observational data. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 116-125, 1999.
3. K. Sachs, O. Perez, D. Per, D. A. Lauffenburger, and G. P. Nolan. Causal Protein-Signaling Networks Derived From Multiparameter Single- Cell Data. *Science*, 308, 523C529, 2005.
4. Y. He and Z. Geng. Active learning of causal networks with intervention experiments and optimal designs. To appear in *J. Machine Learning Research*, 9, 2008.
5. Y. He, Z. Geng and X. Liang Learning causal structures based on Markov equivalence class. LNAI 3734, 92-106, ALT 2005, S. Jain, H. U. Simon and E. Tomita Eds. Springer-Verlag, Berlin, 2005.

Method:

At the preprocessing step, the original continuous data are discretized into 3 levels by using the information-preserving technique (Hartemink, 2001).

We propose an approach for discovering causal networks from multiple data bases with external interventions. In our approach, we first find a skeleton or a Markov equivalence class of networks, in which there are undirected and directed edges. Then we orient undirected edges in terms of information on causality from data sets with external interventions. Intuitively intervening a cause affects its effects, but intervening an effect does not affect its causes. For each undirected edge, we determine its orientation using data sets with external interventions to its nodes.

Results:

The path matrix of the causal network obtained by using our approach for the CYTO data set is shown in Table 1, where ‘0’ in cell (i, j) denotes no edge between nodes i and j , ‘1’ denotes a directed edge $i \rightarrow j$, ‘-1’ denotes a directed edge $i \leftarrow j$, and ‘2’ denotes an unoriented edge. Our causal network includes 11 directed edges and 5 undirected edges. The network depicted in Figure 1 shows a comparison between our network G with the classic network G' mentioned by Sachs et al. (2005), where the black edges are consistent in both G and G' , the red edges are those with reversed orientation in both networks, the blue edges are those in G but not in G' , and the dashed ones are those not in G but in G' . The network after drawing the dashed edges is our result G . Our network G is quite different to G' . It may be because our G is constructed as a whole network.

	praf	pmek	plcg	PIP2	PIP3	Erk	Akt	PKA	PKC	P38	pjnk
praf	0	2	0	0	0	0	0	-1	0	0	0
pmek	2	0	0	0	1	0	0	0	0	0	0
plcg	0	0	0	-1	2	0	0	0	0	0	0
PIP2	0	0	1	0	1	0	0	0	0	0	0
PIP3	0	-1	2	2	0	2	-1	-1	0	2	2
Erk	0	0	0	0	2	0	1	-1	0	0	0
Akt	0	0	0	0	1	-1	0	1	0	0	0
PKA	1	0	0	0	1	1	-1	0	0	0	0
PKC	0	0	0	0	0	0	0	0	0	1	0
P38	0	0	0	0	2	0	0	0	-1	0	2
pjnk	0	0	0	0	2	0	0	0	0	2	0

Table 1: The path matrix of the network obtained by our approach for the CYTO data set

Advantages of our approach: We propose an approach for structural learning from multiple data bases with external interventions. Comparing with the Bayesian approach via MCMC proposed by Sachs et al. (2005), our approach have higher computational efficiency. The Bayesian approach is a score-based method, and our approach is a constraint-based method. Since it is difficult to find posteriors for all possible causal networks in Bayesian approach, it uses MCMC to find posteriors of edges and then combines those edges with high posteriors together to construct a network, but such a network may not have the maximum posterior. Different to this, our approach determines orientation of undirected edges selecting data sets with external interventions, and thus this approach can also be used for intervention design. Our approach is based on conditional independence test, which can be easily executed when the number of variables is small. According to our simulation, we find that our approach has high accuracy when the sample size is small.

Keywords: Active learning, Causal Network, Structural learning

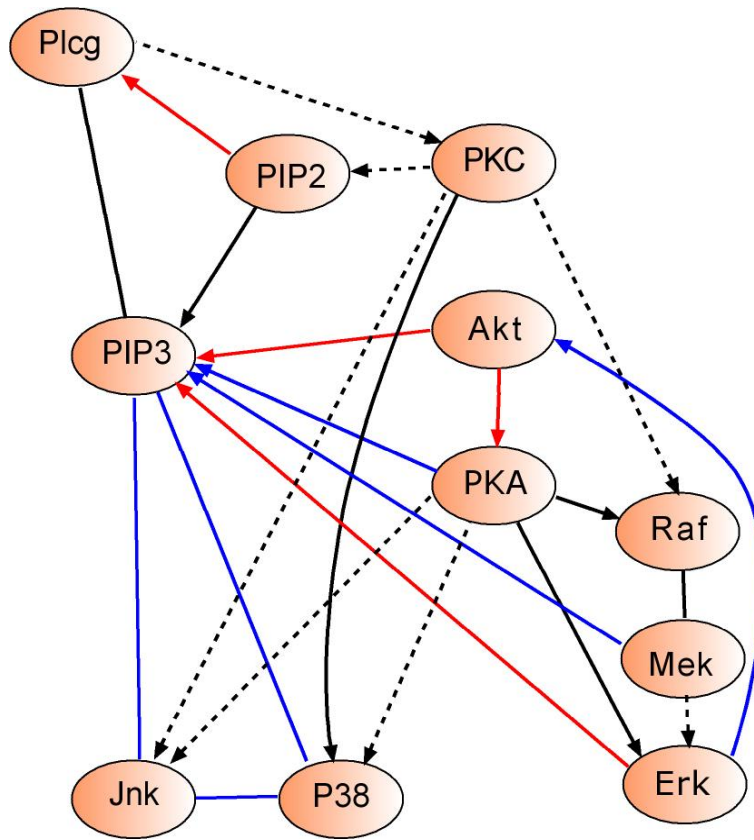


Figure 1: Results of our approach on CYTO data.