A New algorithm for Inferring Gene Regulatory Networks Based On Combination of Bayesian Approach and MIT Score.

Rosa Aghdam

Faculty of Mathematical Sciences, Department of Statistics, Shahid Beheshti University, G.C., Tehran, Iran.

Institute for Research in Fundamental Sciences (IPM), Iran. n-aghdam@sbu.ac.ir

Changiz Eslahchi

Faculty of Mathematical Sciences, Department of Computer Science, Shahid Beheshti University, G.C., Tehran,Iran. ch-eslahchi@sbu.ac.ir

Mojtaba Ganjali

Faculty of Mathematical Sciences, Department of Statistics, Shahid Beheshti University, G.C., Tehran, Iran.

m-ganjali@sbu.ac.ir

Abstract

Gene regulatory networks explain how cells control the expression of genes, which, together with some additional regulation downstream, determines the production of proteins essential for cellular function. Bayesian networks (BNs) are practical tools which have been successfully implemented in learning gene networks based on microarray gene expression data. All existing methods for inferring Gene Regulatory Networks (GRNs) from gene expression data sets have some strengths and weaknesses. There is still a large space for current approaches to be improved. In the Bayesian network the dependency of two variables needs to be determined. Conditional mutual information (CMI) is a suitable tool for detecting the joint conditional linear and nonlinear dependency between genes, which in accordance with the complexity of biology instead of linear assumption.

In this work, we introduce an iterative algorithm for inferring GRNs from gene expression data to improve the prediction accuracy of the PC Algorithm based on conditional mutual information test (PCA-CMI).

We applied an iterative strategy to identify the directed acyclic graph. First, score searching method is applied to direct the edges of S_i (the skeleton of order i). Second, some scores

values are defined for $\mathrm{ADJ}(\mathrm{X}) \cap \mathrm{ADJ}(\mathrm{Y})$(let ADJ(X) denotes the set of variables in the graph which are adjacent to X) and the nodes of separator set belongs to the set of nodes with high score values. Finally, to construct S_{i+1} conditional independence relationship between two genes given separator set is determined in $\mathrm{G_i}$ (directed acyclic graph of order i). This iterative procedure is repeated until a stopping condition is met.

In this work a mutual information test is applied in Max-Min Hill Climbing algorithm to direct the edges of skeleton. Only the local changes related to reversed edges between nodes are considered in the algorithm to determine suitable directed network. We run the algorithm on 50 different starting graphs which are chosen randomly then one with the maximum score value is selected.

The achieved improvement of our algorithm in comparison with PCA-CMI (Zhang et al., 2011)is derived from reduction of statistical errors in the process of learning the skeleton of gene network.

We use Red.Pen (java package for MIT score) to direct the edges of skeleton which can reduce running time and the required memory in comparison with Elvira system. The merits of the new algorithm are evaluated by applying this algorithm on the Dream3 challenge and real data set such as SOS DNA repair network with experiment data set in Escherichia coli. The results indicate that applying the proposed algorithm improves the precision of learning the structure of the GRNs.

Keywords: Gene regulatory networks; Gene Expression Bayesian networks; Conditional mutual information; MIT score ; Max-Min Hill Climbing algorithm.

Acknowledgement

The authors would like to thank Departments of Research Affairs of Shahid Beheshti university. The research presented in this paper was carried out on the High Performance Computing Cluster supported by the computer science department of Institute for Research in Fundamental Sciences (IPM). We are also grateful to Luis M. de Campos and Xiujun Zhang for their excellent comments on several parts of this work.

References:

1. Acid S. and de Campos, L. M. (2001) **A hybrid methodology for learning belief networks: Benedict,** *International Journal of Approximate Reasoning, 27:235-262.*

2. Akaike, H. (1974)**A new look at the statistical model identification.** *IEEE Transactions on Automatic Control, 19:716-723.*

3. Cheng, J., Bell, D. and Liu, W. (1998) **Learning Bayesian Networks from Data: An Efficient Approach based on Information Theory.**

4. Chickering DM, Geiger D, Heckerman D (1995) **Learning Bayesian networks: search methods and experimental results.** *In: Proceedings of the fifth international workshop on artificial intelligence and statistics, pp 112128.*

5. Cooper, G. F. and Herskovits, E. A. (1992) **Bayesian method for the induction of probabilistic networks from data**. *Machine Learning, 9:309348.*

6. De Campos, L.M. **(**2006**). A Scoring Function for Learning Bayesian Networks based on Mutual Information and Conditional Independence Tests**. *Journal of Machine Learning Research 7: 2149.2187.*

7. Faulkner, E. (2007). **K2GA: heuristically guided evolution of Bayesian network structures from data.** *In Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining (CIDM 2007), IEEE, 1825. doi:10.1109/CIDM.2007.368847.*

8. Friedman, N., Nachman, I. and Peer, D. (1999).**Learning Bayesian network structure**

from massive datasets: the .Sparse Candidate. algorithm. *In Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI-99), Prade, H. and Laskey, K. (eds). Morgan Kaufmann, 206215.*

9. Friedman, N., Linial, M., Nachman, I. and Peer, D. (2000). **Using Bayesian networks to analyze expression data**. *Journal of Computational Biology 7(3/4), 601620.*

10. Friedman, N. (2004). **Inferring cellular networks using probabilistic graphical models**. *Science 303(5679), 799805.*

11. Heckerman, D., Geiger, D. and Chickering, D. M. (1995) **Learning Bayesian networks: The combination of knowledge and statistical data**. *Machine Learning, 20:197243.*

12. Imoto, S., Goto, T. and Miyano, S. (2002) **Estimation of genetic networks and functional structures between genes by using Bayesian networks and nonparametric regression Pac**. *Symp. Biocomput ,7, 175-186.*

13. Kalisch,M. and Buhlmann,P. (2007) **Estimating high-dimensional directed acyclic graphs with the PC-algorithm.** *J. Mach. Learn. Res., 8, 613-636.*

14. Larraaga P, Poza M, Yurramendi Y, Murga RH, Kuijpers CMH (1996) **Structure learning of Bayesian networks by genetic algorithms: a performance analysis of control parameters.** *IEEE Trans Pattern Anal Mach Intell 18(9):912926.*

15. Marbach, D., Prill, R.J., Schaffter, T., Mattiussi, C., Florea no, D. and Stolovitzky, G.(2010) **Revealing strengths and weaknesses of methods for gene network inference** *Proc. Natl Acad. Sci. USA, 107, 6286-6291.*

16. Pearl, J. (1988) **Probabilistic Reasoning in Intelligent Systems: in Intelligent Systems** *Pac. Symp. Biocomput, 9, 557 567.*

17. Pearl, J. (2000) **Causality. Models, Reasoning and Inference**: *Cambridge University Press*

18. Pena, J. M., Bj orkegren, J. and Tegnr, J. (2005): **Growin Bayesian network models of gene networks from seed genes.** *Bioinformatics, 25, 224229.*

19. Ronen,M., Rosenberg, R., Shraiman, B. and Alon, U. (2002) **Assigning numbers to the arrows: Parameterizing a gene regulation network by using accurate expression kinetics.** *Proc. Natl. Acad. Sci. USA, 99, 10555-10560.*

20. Saito,S., Zhou, X., Bae, T., Kim, S., Horimoto,K. (2011) **A procedure for identifying master regulators in conjunction with network screening and inference.**. *IEEE Int. Conf. Bioinf. Biomed., 296- 301.*

21.Shen-Orr SS, Milo R, Mangan S, Alon U. (2002) **Network motifs in the transcriptional regulation network of Escherichia coli.** *Nat Genet 31:64-68.*

22. Smet,R.D. and Marchal,K. (2010) **Advantages and limitations of current network inference methods** *Nat. Rev. Microbiol., 8, 717-729.*

23. Spirtes, P., Glymour, C., Scheines, R. (1993). **Causation, Prediction and Search**: *Springer-Verlag.*

24. Spirtes P., Meek C., Richardson T. (2000) **An algorithm for causal inference in the presence of latent variables and selection bias, in Computation, Causation and Discovery. C. Glymour and G. Cooper.** *MIT Press*

25. Tibshirani,R. (1996) **Regression shrinkage and selection via the Lasso.** *J. R. Stat. Soc. B, 58, 267-288.*

26. Tsamardinos I, Brown LE, Aliferis CF (2006). **The max- min hill-climbing bayesian network structure learning algorithm**. *Machine Learning 65(1):3178*

27. Villanueva, E. and Maciel., C.D. (2012) **Optimized algorithm for learning Bayesian network superstructures.** *In Proceedings of the 2012 International Conference on Pattern Recognition Applications and Methods, ICPRAM12.*

28. Wallace, C.S., Korb, K.B., Dai, H. (1996) **Causal discovery via MML.** *Proceedings of the Thirteenth International Conference of Machine Learning (ICML'96), Morgan Kaufmann, San Francisco CA USA, pp. 516-524.*

29. Wang,Y., Joshi, T., Zhang, X., Xu3, D. and Chenm, L. (2006) **Inferring gene regulatory networks from multiple microarray datasets.** *Bioinformatics, 22, 2413-2420.*

30. Wang, K., Saito, M., Bisikirska, B.C., Alvarez, M.J., Lim, W.K., Rajbhandari, P., Shen, Q., Nemenman, I., Basso, K., Margolin, A.A.,(2009) **Genome-wide identification of post-translational modulators of transcription factor activity in human B cells.**. *Nat. biotechnol., 27, 829-839.*

31. Zhang, Y. and Ji, Q. (2005) **Active and dynamic information fusion for facial expression understanding from image sequence**. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27, 699714.*

32. Zhang, X., Zhao, X., He, K., L, L., Cao, Y., Liu. J.,Hao, J., Liu, Z and Chen, L. (2011)**Inferring gene regulatory networks from gene expression data by PC algorithm** based on conditional mutual information. *Bioinformatics*.