Inferring large-scale gene regulatory networks using a low-order constraint-based algorithm

Mingyi Wang,[a] Vagner Augusto Benedito,[a,b] Patrick Xuechun Zhao[a] and Michael Udvardi[a] *

[a] *Plant Biology Division, The Samuel Roberts Noble Foundation, Inc., 2510 Sam Noble Parkway, Ardmore, Oklahoma, USA 73401. Fax: 580 224 6692; Tel: 580 224 6655; E-mail: mudvardi@noble.org*
[b] *Genetics & Developmental Biology Program, Plant & Soil Sciences Division, West Virginia University, 2090 Agricultural Sciences Building, Morgantown, WV 26506.*
* *Corresponding Author*

† Electronic Supplementary Information (ESI) available: The software and related datasets can be downloaded from http://bioinfo.noble.org/manuscript-support/lpc/.

Email addresses:

Mingyi Wang: mwang@noble.org

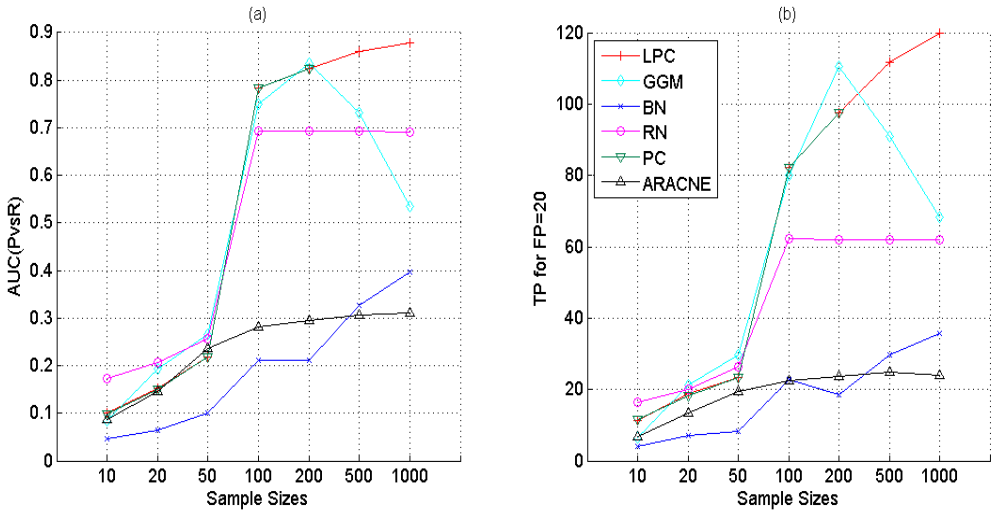
Michael Udvardi: mudvardi@noble.org

Supplementary Material - Tests over scale-free networks

We tested our methods over networks with scale-free topology [1], which have been used widely in literature as models of regulatory networks. Biological networks appear to exhibit such structure [2]. We generated 10 different scale-free networks with 100 genes each, using methods provided by [3]. All of these scale-free networks contained feedback loops. For each scale-free network, we used the same ODEs introduced in the "Evaluation" section of the paper and generated 7 synthetic data sets with different sample sizes (10, 20, 50, 100, 200, 500 and 1000). For all simulated datasets, we ran the five structure prediction methods used in the "Evaluation" section. In the same way that we tested DAGs, we adopted Precision versus Recall (PvsR) curves [4] and true positive (TP) number with fixed false positive (FP) numbers [5] to measure the quality of network reconstruction. The average AUC(PvsR) values for TP=20 for a given sample size were calculated and presented in Supplementary Fig. 1. For the PC-algorithm, only test results for sample sizes of 10, 20, 50, 100 and 200 are presented because the algorithm cannot complete the tests for samples sizes ≥500 in a reasonable time (seven days for each sample size). From the AUC(PvsR) evaluation (Supplementary Fig. 1a), the performances of LPC were better than other methods at sample sizes of 100, 500 and 1000, while GGM outperformed all other

algorithms at sample sizes of 50 and 200. RN outperformed all other algorithms when sample sizes were only 10, 20. From TP for fixed FP tests (Supplementary Fig. 1b), the LPC-algorithm was superior to RN, GGM, ARACNE and BN at sample sizes of 100, 500 and 1000, while GGM outperformed all other methods at sample sizes of 20, 50 and 200. ARACNE and BN still performed poorly in these tests. The results showed that the LPC-algorithm performance was comparable with GGM and outperformed other methods. Although different network topologies (scale-free networks with feedback loops) were used for evaluation tests, the similar conclusion can be drawn from the results as described in the "Evaluation of undirected graphs" section of the paper.

We did not test these methods for evaluation of directed graphs because the structures with feedback loops cannot be converted into the corresponding *complete partial directed acyclic graph* (CPDAG). For evaluation of directed graphs, we use the SHD metric (as described in the "Evaluation of directed graphs" section) to compare the structure of the learned and the original networks. For this comparison, we need to convert original DAGs into the corresponding CPDAG before calculating the SHDto avoid penalizing for structural differences that cannot be statistically distinguished even given an infinite number of observations (see details in the "Methods" section of the paper).



Supplementary Fig. 1 Performance comparisons between several methods using scale-free network structures. For (a) and (b), the average AUC(PvsR)s and true positives under the fixed 20 false positives were plotted for LPC, GGM, BN, RN, PC and ARACNE under 7 different sample sizes.

References

1. A. L. Barabasi and R. Albert, *Science*, 1999, **286**, 509-512.
2. M. E. J. Newman, *Siam Review*, 2003, **45**, 167-256.
3. N. Soranzo, G. Bianconi and C. Altafini, *Bioinformatics*, 2007, **23**, 1640-1647.
4. A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera and A. Califano, *BMC Bioinformatics*, 2006, **7 Suppl 1**, S7.
5. A. V. Werhli, M. Grzegorczyk and D. Husmeier, *Bioinformatics*, 2006, **22**, 2523-2531.