Scalable learning of large networks

Sushmita Roy[1], Sergey Plis[1], Margaret Werner-Washburne[2], and Terran Lane[1]

***Short Abstract* — Structure inference of cellular networks from microarray experiments provides important insight about the network of interactions in cells under different conditions. Unfortunately, many structure inference algorithms do not scale to whole-genome data that have several thousands of variables. We propose a two-step approach for learning the structure of large networks. We first pre-cluster the network nodes and learn networks per cluster. We then revisit the cluster assignment of selected variables with poor neighborhoods. Preliminary results suggest that our approach performs at par to approaches that learn the complete network without pre-clustering.**

***Keywords* — Cellular networks, network structure inference**

I. Introduction

Cellular adaptations essential for survival under changing environmental conditions are driven by a complex, but coordinated, set of interactions among genes, proteins and metabolites. Identification of these interactions using whole-genome microarray data is crucial for understanding the functional aspects of these networks and, therefore, how cells respond to changing environmental conditions.

Probabilistic graphical models are well-known frameworks for modeling cellular networks. Unfortunately, when the structure is not known, *de-novo* reconstruction of general probabilistic graphs becomes intractable. To address this issue, researchers use more scalable models such as the ARACNE algorithm [1], which estimate only pair-wise dependencies. As biological networks are likely to have both pair-wise and higher-order dependencies [2], accurate estimation of higher-order dependencies is important to further our understanding of cellular networks.

We present a tractable approach to learning the structure of undirected graphical models. Our main idea is to pre-cluster the nodes into smaller, possibly overlapping groups and learn separate networks per cluster. Because the initial clustering may not be perfect, we use an iterative procedure, which reassigns nodes to clusters based on the quality of the current node neighborhoods, repeating the procedure until convergence.

We compare our pre-clustering approach with and without cluster reassignment, against an approach with no pre-clustering. Our results indicate that, although pre-clustering without reassignment gives significant speed improvements, the cluster reassignment step provides further performance improvement.

II. Experiments

We used data generated from a network of known topology with n=200 nodes. This network is sufficiently large to require our pre-clustering approach, yet small enough to enable structure learning via an approach without pre-clustering. The k-means algorithm was used to pre-cluster the data. The no pre-clustering approach estimates the “best possible” performance that would be obtained if we had enough time.

We evaluated the performance by comparing how well sub-graphs in the true networks matched sub-graphs in the inferred networks. Sub-graphs rather than individual edges were used because connected sub-graphs represent higher-order dependencies. The sub-graphs considered were: (a) all node pairs shortest paths (SPN), (b) sub-graphs per vertex and its 1-step neighbors (1-N), (c) sub-graphs per vertex and its 1 and 2-step neighbors (2-N).

Our preliminary results (Table 1) indicate that revisiting clusters gives better results than does fixed clustering. Although neither pre-clustering approaches perform as well as the no pre-clustering approach, it should be noted that such estimates are intractable to compute for genome-scale networks.

Algorithm	Runtime	SPN	1-N	2-N
No pre-clustering	16m 17s	0.552	0.595	0.484
Fixed clustering	8m 46s	0.505	0.540	0.442
Cluster revisiting	14m 58s	0.519	0.555	0.453

Table 1 F-scores of sub-graphs match between true and inferred networks

III. Conclusion

We present here a tractable approach for learning large networks from genome-scale expression data. Our preliminary results indicate that the pre-clustering approach gives speed improvements without incurring substantial performance loss. We are currently experimenting with different clustering approaches, and comparing our approach against scalable algorithms such as ARACNE.

References

[1] Margolin A, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera RD, Califano A (2005) ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics.

[2] Q Yuan, G Hui (2005) Modularity and dynamics of cellular networks. PLOS Computational Biology.

Acknowledgements: This work was funded by NIMH grant 1R01MH076282-01 and NSF grant MCB-0645854.

[1] Department of Computer Science, University of New Mexico,

[2] Department of Biology, University of New Mexico