Reducing the search-space and performance of Bayesian Networks

Richard Chen, Steven Chen

EN.600.438/638 Project Proposal

Problem

Network inference methods such as LASSO and Bayesian networks are widely used for gene regulatory network (GRN) reconstruction, as GRNs have huge potential in modeling the kinetics of cell state transitions, and discovering causal relationships between genes and proteins. However, there does not exist a single inference method that can scale well with large networks. Graphical LASSO, for example, has one of the fastest run-times of all inference methods, but suffers from low precision and recall. Bayesian networks on the other hand, are accurate in modeling small networks, but have exponential time complexity as the network size increases. In Marbach et al 2014, Marbach tested and validated over 30 interference methods on the DREAM4 challenge dataset, in which all methods performed poorly. In the landscape of GRN reconstruction methods, there exists a strong need for an inference method that preserves run-time and precision on large networks. In this project, we aim to analyze the pitfalls of current inference methods, and explore new techniques that will improve GRN reconstruction speed performance.

Approach

In Young et al 2015, Young found that Bayesian networks had the highest precision value and area under the curve (AUC) statistic amongst popular inference methods such as LASSO, ARACNE, and CLR on small networks, however, its performance dropped off on larger networks. To reduce the search space of calculating prior and posterior probabilities, we will first preprocess our expression data before hand using graphical community detection algorithms such as Fast Louvain, Stochastic Block Models and spectral clustering, and then use Bayesian networks to infer causal relationships amongst smaller communities. We hypothesize that given a large gene expression matrix, our community detection algorithm would cluster genes that share the same pathway, and our Bayesian network would draw causal relationships between genes in each pathway. We assume that our community detection algorithms would be able to detect overlaps for genes that exist in more than one pathway.

Other ways we can reduce the search space of Bayesian network reconstruction is looking at variable importance in Random Forests. Instead of calculating all possible priori probabilities, we can create a decision tree for each gene, and evaluate which set of genes best predict the expression value of each other gene. From these set of genes, we can draw causal relationships.

By the end of our project, we will have answered these questions:

1. Can we reduce the search space of Bayesian network reconstruction using other graphical machine learning algorithms?
2. Will an initial stage of preprocessing improve GRN reconstruction performance?
3. For preprocessing methods that did not increase performance, what are the pitfalls of each preprocessing method? What effects are we not modeling?

Data

Our data is taken from the DREAM4 simulated mRNA expression data used to assess various network inference methods. These networks are based on characteristics of two well-studied systems, E.coli and S.cerevisiae.

Tests/Metrics

In order to test our method, we would implement our algorithm on simulated data sets from DREAM4 with known gold standard networks from KEGG and GO, and also compare its accuracy to the accuracy of other network inference methods.

References

http://bmcsystbiol.biomedcentral.com/articles/10.1186/1752-0509-8-47
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0029348#references
http://www.ncbi.nlm.nih.gov/pubmed/22796662
http://arxiv.org/pdf/1110.5813v4.pdf
https://www.bioconductor.org/packages/3.3/data/experiment/vignettes/DREAM4/inst/doc/DREAM4.pdf