Analysis of Efficacy and Runtime in Bayesian Networks

—

Steven Chen
Richard Chen

Background

- Gene Regulatory Networks (GRNs) are molecular networks representing connections between genes that regulate mRNA/protein expression levels downstream.
- Network inference methods (Graphical LASSO, ARACNE, CLR, Bayesian Networks) are used to predict these networks from expression data.
- We explore the pros and cons of one of these methods, the Bayesian Network inference

Bayesian Networks

- Bayesian Networks - Represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG)
- Score-based learning
    a. Define scoring function that measures how well a certain structure fits the observed data
    b. Start with random network
    c. Score all possible changes
    d. Search for structure with highest score, and apply changes
    e. Repeat c-d until no further modification improves score
- In genomics, Bayesian Networks predict causal relationships between genes.

Problem

When Bayesian Network inference is scaled to larger network sizes, runtime increases exponentially

Approach

- 2 steps
  - Cluster genes
  - Perform Bayesian network inference on the individual clusters
- Hope is to reduce runtime by reducing the search space for Bayesian Network inference while producing similar results to inferring the entire network

Clustering

- Evaluate optimal k using BIC values.
- Gaussian Mixture Models/Expectation Maximization
  - R: mclust, self-written code
  - MATLAB: default function, self-written code
- K-Means
- Spectral Clustering
- Fast Louvain
- Stochastic Block Modeling

Data

- DREAM5 challenge data sets (from Marbach et. al.)
- 3 networks from well-studied or simulated systems
  - E.Coli - 4511 genes, 805 samples
  - S.cerevisiae - 5950 genes, 536 samples
  - in silico - 1643 genes, 805 samples
- Gold Standards for the systems used for network inference evaluation

Issues

- Though we are trying to address the problem of runtime, it still presents a problem during analysis because of computational power
- Clustering algorithms for $k > 100$ take a long time
- Bayesian network inference on smaller cluster sizes still takes a long time
- Learning curve of understanding:
  - MATLAB, Python, and R packages
  - Gene Set Enrichment Analysis

Preliminary Results