Critique 1: Expression-based Network Inference Methods

Problem Overview

The three papers reviewed here discuss methods which aim to recover gene regulatory networks from expression data. This high-level goal is accompanied with two major logistical problems: the number of genes/variables/features far outstripping the number of available sample data to learn from (p >>> n) and the challenge of efficiently searching through a huge space when learning the network.

Method Overview

1. How does the learned network structure represent the underlying regulatory network?

Using Bayesian Networks to Analyze Expression Data (Friedman et al.) employs a Bayesian Network, a directed acyclic graph (DAG). The nodes represent a random variable (gene/regulatory factor). The edges correspond to a conditional probability distribution (CPD) given the parent node(s) – for discrete variables, the CPD is in a table format; for continuous variables, in Gaussian distribution. *Learning Module Networks* (Segal et al.) groups the nodes in a Bayesian Network into modules and derives the relationship between the modules instead of individual variables/nodes; such relationship is represented in a regression tree. In *Inferring Regulatory Networks from Expression Data Using Tree-Based Methods* (Huynh-Thu et al.), GENIE3 algorithm treats genes/regulatory factors as a fixed-length feature vector to predict the expression level of a particular gene; a directed, possibly cyclic network is implicitly derived from this framework, where a given node is connected from a set of nodes/features that ranked highest in their predictive power.

2. How does each method go about addressing the issue of p >>> n?

Friedman et al. acknowledge that "when learning models with many variables, small data sets are not sufficiently informative to significantly determine that a single model is the "right" one." Their proposed solution is to instead identify key features of the network (e.g. Markov relations, order relations) and measure the confidence on those features. Paring down p to a "tractable" level was the fundamental motivation behind modularizing approach to Bayesian Network taken by Segal et al. Huynh-Thu et al. "decompose the problem of inferring a network of size p into p different feature selection problems," each with sample size of n.

3. What is the search algorithm? What are the strategies to prune the search space?

In deriving a Bayesian Network from expression data, Friedman et al. start the search with a "blank" network with all variables independent from one another and locally searches for a graph that maximizes a Bayesian score. The sparse candidate algorithm is employed to initialize a small number of candidate parents; those that improve the score are added to the network iteratively until the score plateaus. The search problem for Segal et al. is two-fold: "searching through the space of structures of the graph and the space of module assignments." The algorithm starts with initial assignment of variables to modules (clustering method can be employed); then in a local search, the module dependency structure is optimized to a Bayesian score; with the given structure, the module assignment is then optimized by sequentially reassigning a variable from one module to another. For each module, a step-wise split is used to derive the regression tree. The steps are repeated until convergence. Both Friedman et al.'s algorithm and Segal et al.'s can employ simulated annealing instead of greedy hill climbing to 'escape' from a local maximum. Huynh-Thu et al. turn the expression data into output (target gene) and feature vector (all other genes) for each of the p regression problems, builds an ensemble of decision trees (Random Forests or Extra Trees), ranks the gene interactions on an error-reduction score, then finally global ranks the interactions.

Method Evaluation

1. What types of data sets are used for testing the methods?

Friedman et al. employ a randomly and independently permutated data set from an actual expression data of S. cerevisiae, along with the original mRNA expression level data of S. cerevisiae. Segal et al. utilize synthetic data generated by a known module network and mRNA expression level data of yeast, as well as stock market data. Huynh-Thu et al. similarly use synthetic data generated from E-coli and S. cerevisiae, and actual expression data from E. coli.

2. What are the quantitative metrics used?

Friedman et al. mainly use the higher number of features (order relations, Markov relations) distributed along higher confidence levels against the feature-confidence distribution on randomized data to show the statistical significance of their results. Segal et al. use relatively high % of recovered structure from the synthetic data from a known module network, along

with higher data log likelihood than a traditional Bayesian network, as a statistical analysis of the efficacy of the module network framework. Huynh-Thu et al. use various metrics related to precision-recall curve and receiver operating characteristics curve to demonstrate the tree-based method's superiority to other existing algorithms (CLR, ARACNE, MRNET, GGM).

3. Was there any qualitative analysis of the results?

Friedman et al. perform a qualitative review of their output network by extrapolating from the identified features using known gene functionality. Segal et al. attempt to deduce within-in module shared functionality and between-module regulation using known gene function/regulation information. They are able to quantify this analysis using annotation enrichment metric and show stronger enrichment than in a clustering approach.

Novel Insights

1. Were they able to predict any biological function or relationship from the network?

Inspecting the list of dominant genes in the ordering relation feature, Friedman et al. identify genes that are known to be essential in cell function (many whose null mutant is inviable or lethal). In the list of top Markov relations, they assert that the best Markov relations reveal a functional relationship between the genes; Markov relations are also able to identify potentially hierarchical multi-gene relationship that would simply be buried into a single cluster in previous clustering approaches. Segal et al. use the module network model to predict the previously unknown functionality of certain genes, and performed in vitro experiments to confirm the three novel predictions.

Strengths & Weaknesses

1. How did the method improve upon previous methods?

At the time Friedman et al. published, "Most of the analysis tools currently used are based on clustering algorithms." By implementing a Bayesian network learning onto expression data, they are able to take advantage of "probabilistic semantics [which] betters fits the stochastic nature of both the biological processes and noisy experiments." Their approach to extracting features allows the important trees to stick out from the forest, so to speak. Module Networks improve upon the Bayesian network approach by tackling the issue of a domain with large number of variables, much larger than the sample size. Segal et al. postulate that the gains made upon non-modular Bayesian Network derive from parameter sharing in a module, wherein each parameter can be estimated on a larger sample, which allows learning dependencies too weak to be noticed with a single variable. Apart from empirical data supporting improvement in learning regulatory networks compared to relevance networks and Gaussian graphical models, tree-based inference methods have a manageable computational complexity by converting Bayesian Networks' intractable search problem (which can only be tamed with heuristics) into distinctive feature selection problems.

2. What are the unmet needs?

The key shortcoming of a Bayesian network is made obvious by the existence of module networks (too large of a p). Module networks' limitations come from each variable being assigned to only one module (which may not be a realistic representation), and the search arriving at a single network and missing out on comparable or equivalent structures. Furthermore, both types of networks cannot have cycles, which may result in excluding key cyclic relationship information. The most alarming part of the tree-base inference method was its diminishing ability to identify regulatory factors as the number of regulatory factors increased, and its no-better-than-random predictive capability against a real E. coli expression data without initial selection of transcription factors, which taints its scalability claim. Although not necessarily a shortcoming, the issues of key parameter selection (e.g. # of parents, # of modules, # of features) and a mechanism for a priori knowledge incorporation (e.g. candidate parent selection, initial 'skeleton' construction, initial module assignment, potential feature/transcription factor identification) were a common thread throughout the different methods in terms of technical challenges to be improved upon.