Critique 1 – Integrative Network Reconstruction

Problem Overview

Addressed in these papers is the task of integrating various sources and types of data within a network inference algorithm. Network inference is an important task, and the authors of these papers make note of its relevance in diverse areas such as understanding gene expression, cell function, disease, and cell and organismal responses to new environmental conditions. Despite its importance, network inference remains a technically challenging task.

The Methods

1. *Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks (Greenfield et al.)* – In this paper, the authors make note that most previous network reconstruction methods only make use of gene expression data and occasionally other proteomic measurements, but cite an increase in the availability of other diverse datasets related to gene regulation. They develop two closely related methods that incorporate these diverse datasets, namely Modified Elastic Net (MEN) and Bayesian Best Subset Regression (BBSR). At the core of both of these methods is an ODE model that uses time-course data, where the expression of a target gene is modeled as a function of the time-lagged expression of its regulators. The MEN approach uses the l_1 and l_2 penalties to solve an optimization problem and learn the relationships between each target and its regulators in the form of a sparse network, while additionally incorporating prior belief in an interaction into their optimization formulation. BBSR is an alternative approach to network inference that assumes a target genes expression can be modeled using a Gaussian distribution. Here, Bayesian Information Criterion (BIC) is used for model selection.
2. *Physical Module Networks: an integrative approach for reconstructing transcription regulation (Novershtern et al.)* – The approach described by Novershtern et al aims to learn regulatory networks at the module level, where a module is defined as a set of co-expressed genes, each sharing a set of regulators. In learning the module network the algorithm looks for a "path" between a regulator and a target, consisting of the regulator's protein product, optional protein-protein interactions, and ultimately ending in a TF known to bind the promoter of the target gene. These paths can be variable in length, but shorter paths whose edges are supported by prior datasets are preferred, resulting in a parsimonious solution. The set of selected paths are referred to as a "Physical Interaction Graph" and help guide the learning process towards regulatory edges that are supported by additional data beyond expression.
3. *Learning Regulatory Programs That Accurately Predict Differential Expression with MEDUSA (Kundaje et al.)* – MEDUSA learns context-specific regulatory networks by incorporating promoter sequences, transcription factor occupancy data, and gene expression levels. Instead of using expression data in a traditional manner such as clustering, MEDUSA uses discretized expression data along with sequence properties to learn for each sample whether the gene is up regulated or down regulated. To accomplish this task, MEDUSA uses an alternating decision tree (ADT). Additionally, MEDUSA has the ability to learn sequence-specific motifs as part of the learning process.

Evaluation

1. Greenfield et al evaluate their methods using a published *Bacillus subtilis* network, as well as two networks from the DREAM consortium: one of *Escherichia coli* and one created *in silico*. Before testing the accuracy of their methods against that of other methods, the authors do several experiments to test the robustness of their methods to parameter choices, the relationship between the correlation of regulator-target pair datasets and the confidence in a prediction, and the impact of prior knowledge. They then compare both of their methods to several other methods that do not use prior knowledge, and show that both of their methods outperform these competitor methods even when the prior knowledge given to them as input includes a significant amount of false positives.
2. Norvershtern et al test their physical module network approach against an earlier module level approach that did not make use of a "Physical Interaction Graph", using a synthetic network consisting of 312 genes and 7 modules regulated by 10 of these genes. They show that their method outperforms a traditional module network approach in terms of precision when tested with a varying number of modules, and state (but do not show actual values) that "both models have good recall" and recall "ranges between 80% and 100%." They also experiment with smoothing of input expression data, and show that their approach is more robust to the smoothing parameter. They also tested their approach using several yeast expression datasets, assessing the ability of their algorithm to accurately identify biologically relevant pathways, and testing the full extent of the model against a yeast cell-cycle network.
3. MEDUSA was evaluated using three yeast expression datasets under diverse conditions. With an ESR dataset and 10-fold cross-validation, MEDUSA achieved an error rate of just 13.4%, better than the error rate when MEDUSA was given a database of motifs instead of inferring the motifs itself. They compared this to a baseline k-nearest neighbors approach, and showed that the error achieved with MEDUSA was significantly lower. They also evaluate MEDUSA against a more diverse DNA damage dataset and achieve an error rate of 20.7%, which the authors claim is good considering the diversity of the dataset. A third test was done using a Hypoxia dataset, where MEDUSA had

an error rate of 8.0%. The authors note that this was an easier dataset to make predictions with given the presence of replicates in the data. To show that MEDUSA was still effective in this setting, they repeat the tests without using promoter sequences and get an error rate of 26.0%. Additionally, they grouped replicates into the same folds using cross-validation so that replicates were never found in both a train and test set, and got an error rate of 23.9%.

Novel Insights

1. The main conclusion to be drawn from Greenfield et al is that the use of regulatory information as a prior on the graph structure can lead to networks that are overall more in line with the ground truth. However, since the evaluation was done using DREAM networks, the authors do not make mention of any novel biological predictions.
2. Norvershtern et al apply their approach to real yeast data and identify pathways enriched for various biological processes. Using their inferred networks and interaction graphs, the authors are able to make several biological hypotheses consistent with existing literature. These include a role for STBF, "a zinc-finger with an unknown function," in "induction of early meiosis genes," and novel stress response pathways associated with "reduction of cell growth" starting "with the knockout of GRC1, a transcriptional activator of glycolysis genes." The authors also test their method using a dataset that measured expression over a time-course during human-flu infection, and identify putative paths "that lead from the viral proteins to changes in expression." This method was able to follow up on the original paper that identified a role for the "viral polymerase subunits NP and PB2" in perturbing host signal, by "identifying a pathway that includes apoptic proteins TRAF1, API1, and p53."
3. The analysis done with MEDUSA is rich in biological predictions, including the association of several regulators including Hap4 and Wtm1, among others, with DNA damage response. The authors also compare and contrast inferred regulators between DNA damage response and general ESR, and make special note of Msn4, noting that "Msn4 was predicted to be an important regulator primarily through its binding site," whereas its expression profile "was only weakly predictive." They are furthermore able to identify particular TF binding sites associated with DNA damage response regulation. They applied MEDUSA to yeast Hypoxia data, and were able to discover not only regulators associated with this condition, but also distinguish between them and general stress regulators.

Strengths and Weaknesses

1. An obvious strength of the method by Greenfield et al is its ability to improve predictive accuracy by incorporating prior regulatory information from diverse datasets. One weakness pertaining to the paper is that while the authors show that the method outperforms methods that do not incorporate prior knowledge, they do not compare their method to other methods that include prior knowledge, nor do they compare it to methods that learn module-based networks. It is possible that the method is in fact superior to these other methods, but the authors don't make an attempt to show this.
2. Not only does the "Physical Module Network" approach presented by Novershtern et al yield networks with improved accuracy, but the approach outputs a "Physical Interaction Graph" that can be used by a user to understand not only which genes are regulated by which regulators, but also how the regulation takes place. This could potentially guide biologists in performing experiments. Additionally, a module network can make up for limited expression data by pooling the data of those genes within the same module. However, in the case of Physical Module Networks, additional data is needed such as protein-protein interactions and TF occupancy data, so this particular module network approach is still limited to specific scenarios where various datasets are available.
3. In addition to its ability to accurately predict up/down regulation of genes, a primary strength of the MEDUSA approach is its ability to learn TF binding motifs *de novo*. This may make the method effective with species that have not been heavily studied, for which motif binding sites have not been documented. Another strength lies in that is able to identify context specific. One weakness of the method is the discretization of gene expression data. As discussed by the authors, this results in loss of information regarding more subtle variation in gene expression.

Extensions

As is briefly discussed by Kundaje et al (and in class), I would be interested in modifying MEDUSA to use real-valued expression levels of regulators rather than discretized up/down/normal values. The authors cite good reasons for discretizing the data such as reduction in noise, and straightforward application of current machine learning classification algorithms. However, although this is outside of my area of expertise, it seems as if since 2007 (the publication year of this paper) improved technologies and an increase in available expression data may have led to generally less noisy data, or perhaps a more rigorous quantification of the noise, making it easier to model within a learning framework. Additionally, as computing power continues to increase, it may be more computationally feasible to employ sophisticated statistical techniques to predict real-valued expression rather than discrete labels.