Using Prior Biological Knowledge when Constructing Regulatory Networks

Sara Mostafavi SARAM@CS.STANFORD.EDU
Stanford University, Palo Alto, CA

Daphne Koller KOLLER@CS.STANFORD.EDU
Stanford University, Palo Alto, CA

Abstract

We present a model for constructing regulatory networks from gene expression datasets while accounting for prior biological knowledge in terms of known gene function and pathway annotations. We will show that our new approach results in improved performance in terms of three different evaluation metrics compared to an equivalent model that does not consider such prior biological knowledge.

1. Introduction

Complex diseases such as cancer involve genetic variants that are associated with dysfunction of regulatory factors, effects of which propagate through networks of molecular interactions and affect a large number of genes or proteins. Reconstructing such networks of molecular interactions between potential regulatory genes and their targets can improve our understanding of cellular processes that lead to cancer and its progression. The importance of this problem is reflected in the large number of papers, published over the last decade, that have proposed various methods for reconstructing such networks from high-throughput genomics datasets (see (Kim et al., 2009) for a review).

Typically, regulatory networks are constructed from gene expression datasets that measure the expression levels of tens of thousands of genes in hundreds of samples. In its simplified form, the task at hand is to estimate a bipartite network where the weighted edges represent regulatory interactions between a set of potential regulators and their inferred targets. The main challenge is the abundance of spurious correlations between gene expression profiles, exacerbated by the low signal-to-noise ratio and high-dimensionality of the high-throughput datasets, which leads to high false positive rates among the inferred interactions. Currently, there are two types of approaches for addressing this issue: (1) dimensionality reduction in the gene-space, or (2) use of prior biological knowledge about which regulatory interactions are more plausible to somehow prone the inferred edges. Though each approach has its own strengths, combining both of these approaches can result in improved accuracy and interpretability of the inferred regulatory interactions (*e.g.,* (Lee et al., 2009; Novershtern et al., 2011; Lee et al., 2010)).

Here, building on previous work (Lee et al., 2007; 2009), we propose a model that extends the dimensionality reduction approach by incorporating prior biological knowledge to construct more accurate and biologically coherent regulatory networks. In particular, we adopt the *module network* approach (Segal et al., 2003) to dimensionality reduction, by inferring regulatory interactions for groups of co-expressed (*modules*), and hence possibly co-regulated, genes, simultaneously. However, within the same framework, we take the additional step of using prior biological knowledge, in the form of pathway and gene set annotations, to infer which regulatory interactions are more likely. The main challenge here is that a gene can be involved in tens to hundreds of biological functions or pathways and it is not *a priori* known which pathways or gene set annotations are relevant in a given context (*e.g.,* a particular module). Our model offers a principled and coherent approach for determining the relevance of prior knowledge in a context-specific manner, and thus it supports its inferences of regulatory interactions based on correlations in a given dataset and specific prior annotations.

We evaluate our model, and compare its performance to a baseline version that doesn't consider prior knowledge, on a glioblastoma cancer (GBM) dataset from The Cancer Genome Atlas (TCGA). Our model results

in improved performance on three different evaluation metrics: (1) prediction performance in terms of percent explained variance on test data, (2) prediction of known cancer and GBM genes as regulatory hubs, (3) coherence of inferred modules in terms of enrichments of shared up-stream sequence motifs.

2. Methods

In this section, we will first review the module network approach for constructing regulatory networks from gene expression data. We will then describe our extension that incorporates prior biological knowledge when learning such networks.

2.1. Review: Module Networks for Constructing Regulatory Networks

In its simplified form, constructing a regulatory network involves learning *regulatory interactions* between potential regulators and target genes. This problem can be cast as multiple independent sparse regression problems. In particular, given a set of r potential regulators and their expression profiles in n subjects, as summarized in an $n \times r$ matrix X, and the expression profiles of g genes in n subjects summarized in the matrix $Y_{n\times g}$, we solve for regulatory interactions by attempting to predict Y from X (*e.g.,* (Lee et al., 2009)).

As previously mentioned, the challenges in this setting are the high-dimensionality of the data (*i.e.,* large g) and the abundance of spurious correlations between expression profiles. To help address this issue, the module network approach makes a biologically intuitive assumption that both reduces the dimensionality of the problem and improves the interpretability of the learnt networks. In particular, in the module network approach, target genes are organized into co-regulated *modules* that have the same regulatory interactions. More formally, the learning task can be summarized as follows:

$$\underset{W}{\operatorname{argmin}} \sum_{m} \sum_{g\in \text{module}_m} ||\vec{y}_g - X\vec{w}_m||_2^2 + \lambda \sum_m |\vec{w}_m| \quad (1)$$

where m is the number of modules (note that $m << g$), $\vec{y}_i$ and $\vec{x}_i$ are the i^{th} column of X and Y, respectively, W is a matrix of regression weights with columns $\vec{w}_m$ (an r-vector) that represent the *interaction weights* between the regulators and all target genes in module m (referred to as the regulatory program), and $|\vec{w}_m|$ is the L^1-norm of $\vec{w}_m$.

So in the above formulation, target genes are grouped into modules, and all genes that are a member of the same module have the same regulatory program (*i.e.,* $\vec{w}_m$). We can either use a fixed assignment of genes into module (for example by clustering target genes based on their expression profiles), or treat module membership as hidden variables and infer them, along with $\vec{w}_m$'s, in an iterative fashion (as done in (Lee et al., 2009)). In our experiments, we use the second approach.

2.2. Transfer Learning and Incorporating Prior Knowledge

We will use the probabilistic interpretation of the above model to describe how we incorporate prior biological knowledge when learning a regulatory network. In particular, we can represent sparse regression using a probabilistic formulation by assuming a gaussian likelihood and a Laplace prior:

$$\begin{aligned} p(W,Y,X) &\propto \prod_m \prod_{g\in\text{module}_m} p(\vec{y}_g|\vec{w}_m, X)p(\vec{w}_m) \\ &= \prod_m \prod_{g\in\text{module}_m} N(\vec{y}_g|X\vec{w}_m, I) \\ &\times \prod_r Lap(w_{r,m}|0,\lambda^{-1}). \end{aligned}$$

Obtaining the MAP estimates of $\vec{w}_m$'s, by directly optimizing the logarithm of the above joint likelihood, results in the same solution as the one obtained by solving Equation (1).

We will now describe how we use prior biological knowledge in the context of a given module by using *transfer learning*. In particular, we assume that we have a matrix $F_{r\times f}$ that represents our prior knowledge. Here, we are working with prior knowledge about biological functions and pathway memberships (which will refer to as gene annotations), so we have $f_{r,f} = 1$ if regulator r has the f^{th} annotation, and $f_{r,f} = 0$, otherwise. We obtain such annotations from the GSEA database (see "Results"). Intuitively, we would like to assign a prior probability to each regulator-module interaction based on our prior biological knowledge. Relying on the probabilistic interpretation of sparse regression, we can achieve our goal by assigning variable variance parameters $\lambda_{r,m}$ for each regulatory-module pair. In this way, we set an initial prior on the likelihood of each interaction. However, since we often have multiple annotations for each regulator, we would like to *learn* how to combine these annotations to assign such prior probabilities.

Formally, to learn the relevance of this prior knowledge for each module, we parameterize $\lambda_{r,m}$ as a function of $F_{r,:}$ (the r^{th} row of F): $\lambda_{r,m} = h(F_{r,:}\vec{b}_m)$ where $\vec{b}_m$'s

(vectors of size $f \times 1$) are unknown hyper-parameters, themselves drawn from a Gaussian distribution that encourages hyper-parameter sharing across modules, where $h(z) = \frac{1}{1+e^{-z}}$ is the sigmoid function. Put together, we have the following joint likelihood:

$$\begin{aligned} & p(Y, W, B, X, F) \\ \propto & \prod_m \prod_{g \in \text{module}_m} p(\vec{y}_g|\vec{w}_m, X) p(\vec{w}_m|\vec{b}_m) p(\vec{b}_m) \\ = & \prod_m \prod_{g \in \text{module}_m} N(\vec{y}_g|X\vec{w}_m, I) \\ \times & \prod_r Lap(w_{r,m}|0, \lambda_{r,m}^{-1}) \prod_r N(B_{r,:}^{\mathsf{T}}|0, \alpha K^{-1}) \\ \text{s.t.} & \quad \lambda_{r,m} = \frac{1}{1+e^{-F_{r,:}\vec{b}_m}} \end{aligned}$$

where $B_{r\times m}$ is the matrix of hyper-parameters, with $\vec{b}_m$ as its m^{th} column, and $B_{r,:}$ denoting its r^{th} row. K is the precision matrix of the Gaussian distribution (its derivation is described below), and α is a scalar that can be set to adjust the extend of hyper-parameter sharing across modules. optimizing the negative logarithm of the above joint likelihood with respect to W and B results in the following objective:

$$\begin{aligned} \underset{W,B}{\text{argmin}} & \sum_m \sum_{g \in module_m} ||\vec{y}_g - X\vec{w}_g||_2^2 \\ + & \sum_{r,m} \lambda_{r,m}|w_{r,m}| + 2\sum_{r,m} \log(\lambda_{r,m}) \quad (2) \\ + & \sum_m \text{trace}(\frac{1}{2}\alpha BKB^{\mathsf{T}}) \\ \text{s.t.} & \quad \lambda_{r,m} = \frac{1}{1+e^{-F_{r,:}\vec{b}_m}} \end{aligned}$$

Note that the logarithm term in Equation (2) corresponds to the logarithm of the normalization term for the Laplace distribution over $\vec{w}_m$: since we are optimizing $\lambda_{r,m}$'s, this term is not a constant anymore (as in standard sparse regression) and thus must be optimized for. Following the terminology used by (Lee et al., 2009), we call F the matrix of meta-features, $\vec{b}_m$'s are meta-weights, and $1 - \vec{\lambda}_{r,m}$'s are meta-priors.

Although in the above formulation each module has its own set of meta-priors and meta-weights, we encourage parameter sharing across modules by assuming a multivariate Gaussian prior on rows of B (*i.e.,* $B_{r,:} \sim N(0, \alpha K^{-1})$) where the precision matrix $K_{m\times m}$ encourages *similar* modules to have similar meta-weights. We construct K as a graph Laplacian where co-expression values between modules indicate module-module similarities: that is, $K_{i,j} = -c_{i,j}$ if $c_{i,j} \geq 0$, where $c_{i,j}$ is the correlation coefficient between mean expression levels of genes in module i and j, we set $K_{i,j} = 0$ if $c_{i,j} < 0$ and $K_{i,i} = 1, \forall i$.

The objective function in Equation (2) is not jointly convex in W and B, but it is convex in each given a fixed setting of the other. Therefore, we solve for W and B by coordinate descent, where we estimate W and B iteratively.

There are two key differences between what we have proposed here and the work of (Lee et al., 2009). First, we allow each module to have its own meta-priors, but allow sharing of parameters by specifying a multivariate Gaussian prior on rows of B. Second, by optimizing the joint likelihood (as opposed to ignoring the Laplace normalization constant), we have a well-formed probabilistic model.

3. Results

3.1. Datasets and Pathway/Gene Set Annotations

We evaluate our model on a glioblastoma cancer (GBM) dataset consisting of 477 samples and 8000 genes. This dataset was downloaded from the cancer genome project's (TCGA) data portal (Level 3 Agilent data). We then filter the genes based on standard deviation, keeping the top 8000 most variable genes. We denote genes as regulators if they have a "regulatory role" (including transcription regulators, chromatin modifiers, and signaling molecules) according to the Gene Ontology (GO), and treat all other genes as potential target genes.

We use gene sets and pathway annotations provided by Broad's GSEA website (called C2) as prior knowledge. These gene sets were collected from various sources such as online pathway databases, publications in PubMed, and knowledge of domain experts. In our analysis, we only consider gene sets that contain more than 5 and less than 300 genes.

3.2. Evaluation on GBM dataset

We trained our model, and a baseline version where we don't use prior pathway annotations (Equation (1)), on the GBM data and used BIC to select the parameter settings for both models. We set the initial module memberships using affinity propagation clustering (Frey & Dueck, 2007). For our pathway model, learning consists of (a) estimating the regulatory interactions, $\vec{w}_m$'s, for each module m, (b) estimating

the meta-weights, $\vec{b}_m$'s, and the meta-priors $\lambda_{r,m}$, that allow for preferential selection of regulatory interactions and (c) re-estimating the module membership for each target gene using hard EM. For the baseline model, we (a) estimate regulatory interactions and (b) re-estimate the module memberships. As we will describe below, we used three different evaluation metrics: (1) prediction performance on test data in terms of percent explained variance (PEV), (2) prediction of important GBM or cancer genes as "top" regulators (*e.g.,* regulatory hubs in the network), (3) concordance of the final gene modules in terms of sequence motif enrichment.

First, we evaluated our model based on its test set prediction. To do so, we trained our model on 2/3 of the data, and predicted the expression levels of all test genes. We measured the performance in terms of percent explained variance (defined as PEV(gene g)$=1-\frac{MSE}{\text{var}(\vec{y}_g)}$). As shown in Figure 1, including prior information drastically improves the prediction performance on the test data compared to the equivalent baseline version that doesn't consider such information.

Next, we evaluated our model in terms of its prediction of known GBM or general cancer genes. To do so, we ranked each regulator based on its weighted interactions with all modules, and then compared our ranked list of regulators to two different sets: (1) a set of known somatically mutated GBM genes consisting of 13 genes that were expressed in the GBM dataset, and (2) a set of all known cancer genes with known somatic mutation ($n = 233$). We obtained both of these lists from The Cancer Gene Census[1]. We measured the performance by computing the area under the ROC curve using either the GBM genes or the general cancer genes as the true-positive set. As shown in Figure 2(a), our pathway models results in a significant performance improvement in this evaluation measure.

Finally, we evaluated the inferred module memberships by computing the enrichment of genes in a given module for the same sequence motif (using lists of genes with the same up-stream motifs from the GSEA database). As shown in Figure 2(b), the module memberships that are estimated by our pathway model are more consistent with prior knowledge about co-regulation.

[1]Available from http://www.sanger.ac.uk/genetics/CGP/Census/.

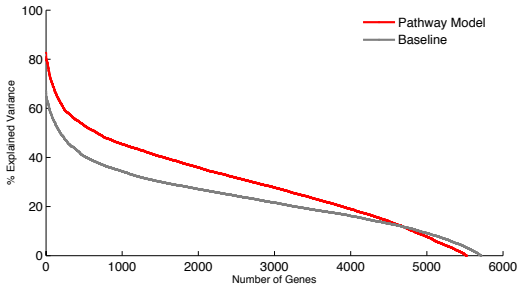
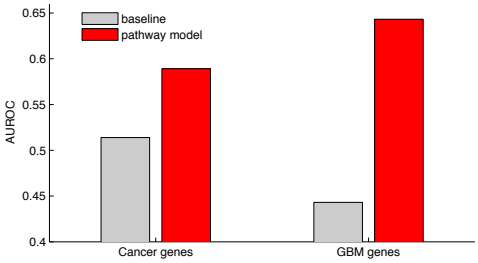
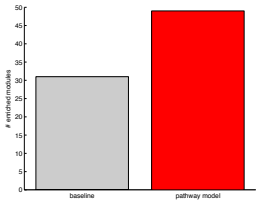


Figure 1. Plot shows the cumulative distribution of percent explained variance (PEV) of target genes (on test data) when using the pathway model or the baseline model. As shown, incorporating pathway information results in improved PEV.



(a)



(b)

Figure 2. (a) Plot shows the area under the ROC curve (AUROC) in predicting known cancer genes (bars on the left) and GBM gene (bars on the right). (b) Plot shows the number of modules in which members are enriched for the same up-stream motif.

4. Conclusion

We have presented a model for inferring regulatory networks while accounting for prior biological knowledge. Our model offers a principled and coherent approach for determining the relevance of prior knowledge in a context-specific manner, and using such information to infer more informative regulatory interactions. We have shown that, in addition to improved test set accuracy, using such information leads to more biologically coherent results.

References

Frey, Brendan J. and Dueck, Delbert. Clustering by passing messages between data points. *Science*, 315: 972–976, 2007.

Kim, H., Shay, T., O'Shea, E., and Regev, A. Transcriptional regulatory circuits: predicting numbers from alphabets. science. *Science*, 325:429432, 2009.

Langley, P. Crafting papers on machine learning. In Langley, Pat (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Lee, S., Zhu, J., and Xing, E. P. Adaptive multi-task Lasso: with application to eQTL detection. In *NIPS*, 2010.

Lee, S.I., Chatalbashev, V., Vickrey, D., and Koller, D. Learning a meta-level prior for feature relevance from multiple related tasks. In *ICML*, 2007.

Lee, S.I., Dudley, A., Drubin, D., Silver, P., Krogan, N., Pe'er, D., and Koller, D. Learning a prior on regulatory potential from eqtl data. *PLoS Genetics*, 5, 2009.

Novershtern, N., Regev, A., and Friedman, N. Physical module networks: an integrative approach for reconstructing transcription regulation. *PLoS Genetics*, 27(13):i177–i185, 2011.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., and Friedman, N. Minreg: A scalable algorithm for learning parsimonious regulatory networks in yeast and mammals. *Nature Genetics*, 34(2):166–176, 2003.