WASHINGTON UNIVERSITY IN ST. LOUIS SCHOOL OF ENGINEERING AND APPLIED SCIENCE

CSE543T Algorithms for Nonlinear Optimization

Determination of Transcription Factor Control Strength and Activity Values from Transcription Factor Deletion Expression Data

John Gibson and Patrick Di Rita

December 7, 2018

1 Abstract

Of current interest in computational and systems biology are transcription factor networks. A transcription factor network is a system of regulatory elements, denoted enhancers and repressors, that interact with genes to promote or suppress transcription, respectively. We propose a simple nonlinear model to approximate the effect of enhancers and repressors on a gene network.

2 Introduction

2.1 Biological Review

In a biological context, the gene is the basic unit of heredity. From a molecular perspective, a gene is a DNA sequence that carries the information for a single RNA (ribonucleic acid) molecule, which is sometimes translated into a polypeptide or protein. A gene includes a coding region, promotor, any regulatory elements involved with its expression, and a transcription termination site. Transcription is the process of building an

RNA molecule from a corresponding DNA sequence. In transcription, the DNA template strand is used to direct synthesis of the RNA molecule. This results in an RNA transcript identical to the non-template (coding) strand, except that all T's in the DNA sequence are replaced with U's. This process is catalyzed by RNA polymerase, which unwinds the DNA strands and uses the template sequence to incorporate complementary free ribose-based nucleotides into the growing RNA strand. Transcription is a highly regulated process; each cell in an organism contains the same DNA, but differences in gene expression caused by the regulation of transcription allow the differentiation of different cell types and allow the individual cells to react to their environment. Regulation of transcription can be either positive, in which transcription is promoted, or negative, in which transcription is repressed. Positive regulation involves gene expression being turned on by activators, especially transcription factors. Transcription factors are proteins which bind to DNA or other proteins and affect transcription rates of various genes. Negative regulation involves the blocking of transcription by repressor proteins binding to operator sequences. Basal expression, or basal transcription, occurs in the absence of both activator and repressor at very low levels. This is also called "leaky" transcription. Repressor proteins bind to operator sequences in order to prevent transcription by physically blocking RNA polymerase, while activators bind to activator binding sites generally called enhancers and help induce transcription by changing the conformation of DNA to allow RNA polymerase to access the promoter or by assisting in the recruitment and assembly of the RNA polymerase complex. RNA transcripts are then used as instructions for the assembly of proteins in a process called translation, directed by the ribosome.
The ribosome is an RNA-protein complex that directs elongation of polypeptides at a rate of 3-5 amino acids added per second. Proteins are the basic functional unit of life, and perform a vast array of roles. Proteins are responsible for catalysis of metabolic reactions, DNA replication, responding to stimuli, providing structure to cells and organisms, and transporting molecules from one location to another. In prokaryotes, transcription and translation are tightly coupled and occur concurrently in the nucleolus, and thus RNA level approximates protein expression level well in prokaryotes. In eukaryotes, transcribed RNA molecules are subject to additional regulation via nuclear transport. A eukaryotic RNA molecule must be transported from the nucleus into the cytoplasm for translation and can be degraded, modified, or otherwise regulated before it is loaded onto a ribosome. Therefore RNA level is only a rough approximation of protein expression in eukaryotes; however current methods for measuring protein concentration are inexact and expensive, and thus measurements of RNA levels are more commonly used to infer gene expression.

2.2 Transcription Factor Networks

The rise of next-generation sequencing technology has resulted in the ability of researchers to easily and cheaply measure RNA expression levels in a variety of different conditions. In addition, continuous development of genome editing technologies such as CRISPR has allowed for easy manipulation of the genetic code. Together, these techniques have led to a method of detailing the effects of a certain transcription factor on gene expression.

In this method, the gene for a transcription factor is deleted and the RNA levels of all other genes expressed by the cell are measured. Repeating this process for many transcription factors while keeping the environment constant allows for the characterization of the effects of each transcription factor on gene regulation. Together, a system of genes and the transcription factors that regulate them are called a gene regulatory network or a transcription factor network (TFN). Transcription factor networks are of great interest in systems biology, and have been studied in great detail (Wilkinson et al. 2017). Many computational and empirical methods of determining TFN structure exist, including *in silico* sequence motif analysis and binding site prediction as well as experimental methods such as chromatin immunoprecipitation sequencing (ChIP-seq) experiments. However, in general these methods only attempt to reconstruct the structure of the transcription factor network and do not provide much information about the strength of the interactions. In general, these networks only detail the system of activation and repression interactions and can be represented in many ways (Schlitt et al. 2007). The classical representation of a TFN is the so-called "topological model," where genes (including transcription factors) are represented by nodes in a graph and different types of interactions are represented by different types of edges. Control system dynamics can be extracted from these graphs, allowing prediction of upregulated and downregulated genes from the active transcription factors in the cell. In addition, models based on differential equations have been developed. The basic form of these models is:

$$
\begin{aligned}
g_1(t+\Delta t) &= (w_{11}g_i(t) + ... + w_{1n}g_n(t))\,\Delta t \\
g_2(t+\Delta t) &= (w_{21}g_i(t) + ... + w_{2n}g_n(t))\,\Delta t \\
&\vdots \\
g_n(t+\Delta t) &= (w_{n1}g_i(t) + ... + w_{nn}g_n(t))\,\Delta t
\end{aligned}
$$

Notice that this model assumes a linear relationship between activities of genes. Additional terms can be added to the model to account for additional parameters. Although this linear model well-approximates certain networks found in nature, it requires a large number of additional terms to be accurate, and the weights w_{ij} for each gene pair are difficult to determine experimentally. In this work, we examine how to determine gene-specific and transcription factor-specific parameters from easily determined expression data.

3 Related Work

As it is easy to optimize and implement, much work has been done on the related linear model of TFN gene expression (Oron et al. 2008). In this model, gene expression is modeled by the weight of each transcription factor control strength on the transcription factor activity, as follows:

$$E^g_{\text{total}} = \beta^g + \sum_{i=1}^{n} \mathit{tfa}_i \cdot cs^g_i$$

Where β^g is the baseline (i.e. basal transcription rate) of gene g, tfa_i is the activity of transcription factor i, and cs^g_i is the control strength of transcription factor i on gene g. However, this model is quite limited; it does not express the concept of saturation of a specific gene or nonlinear relationships between transcription factor activity and gene expression. However, it does make intuitive sense and is known to approximate gene expression in some scenarios, most notably in early embryo developments of *Drosophila*(von Dassow et al. 2000).

4 Technical Details

We define the following:

1. A **control strength** value is a real-valued parameter denoted by cs^g_i that gives the effect of transcription factor i on gene g.

2. A **transcription factor activity** value is a positive real-valued parameter denoted by tfa_i that gives the activity (i.e. a measure of how much transcription factor is present and if it is active) of transcription factor i.

3. A **baseline** value is a real-valued parameter denoted by β^g that describes the baseline (i.e. basal) transcription rate of gene g.

4. We define a **scaling factor** α^g for each gene to account for differences in gene product production, transcription/translation rate, and other biological and stochastic differences that affect the production of a gene product.

5. We define the **influence** I^g_{TF}, a measure of how transcription factor activity and control strength affect a gene product, separately for enhancers and repressors.
   - **Enhancers:** $I^g_{E_i} = \frac{\mathit{tfa}_i}{\mathit{tfa}_i + cs_i}$
   - **Repressors:** $I^g_{R_i} = \frac{1}{\mathit{tfa}_i + cs_i}$

Our two definitions of influence allow for different behaviors of enhancers and repressors as the transcription factor activity value is changed.

Transcription Factor Activity Limit	Enhancer	Repressor
$\mathit{tfa} \to \infty$	$I \to 1$	$I \to 0$
$\mathit{tfa} \to 0$	$I \to 0$	$I \to \frac{1}{cs}$

Naturally, we can then define the total influence of either activators or repressors on a gene. Note, however, that the presence of an enhancer for the gene will mask the presence of a repressor; thus this model's main limitation is the inability to accurately model the interactions between enhancers and repressors.

For n transcription factors and k genes, our terms become:

$$I^g_{E_{\text{total}}} = \beta^g + \alpha^g \sum_{i=1}^{n} \frac{tfa_i}{tfa_i + cs^g_i}$$

$$I^g_{R_{\text{total}}} = \beta^g + \alpha^g \sum_{i=1}^{n} \frac{1}{tfa_i + cs^g_i}$$

and,

$$I^g_{\text{total}} = \beta^g + \alpha^g \sum_{i=1}^{n} \left(\gamma_g \frac{tfa_i}{tfa_i + cs^g_i} + (1 - \gamma_g) \frac{1}{tfa_i + cs^g_i} \right)$$

for $g \in 1..k$, where γ_g is 1 if transcription factor i is an activator of gene g and 0 otherwise. The influence value for each gene is a measure of gene expression. We wish to find parameters for *tfa*, *cs*, α, and β that give the optimal values for I. In order to determine these values, we collect data from **transcription factor deletion experiments**. In these experiments, a transcription factor is knocked out of an organism, allowing us to set the activity value, and thus the influence, of that gene to zero. Repeating this experiment for each transcription factor in the set of factors we wish to study gives us sufficient data to solve for each parameter.

We can define a loss function as follows: Given a vector β of baselines, a vector α of scaling factors, and n by k *tfa* and *cs* matrices, and a vector of actual gene expression data E, then we have our loss function:

$$L(E) = \sum_{g \in G} \left(E^g - I^g_{\text{total}} \right)^2$$

Minimizing $L(E)$ across all parameters gives the optimal set of parameters. The values for E generally come from RNAseq or microarray experiments, which measure a quantity proportional to gene expression (number of fragments mapped per kilobase of gene and fluorescence, respectively). We will utilize randomly-generated test datasets in order to determine the quality of our solutions.

As the process is much the same for repressors, we focus our attention on determining parameters for simulated activator data. In particular, we do the following:

1. Define k values of α to denote the scaling factors of each gene. In our test model these are all set to 1.0.

2. Define k values of β to denote the baseline expression levels of each gene. In our test model these are all set to 0.0.

3. Generate a $k + 1$ by n matrix. The last column of the matrix (the $(k + 1)$st) represents the *tfa* values.

4. Using that matrix, for each transcription factor i:
    - Compute the gene expression value for all genes with transcription factor i deleted (i.e. its activity value set to zero)
    - Create a k by n matrix M where the vector in column j is the expression vector where transcription factor j is deleted. Let M_i be the column i of matrix M.
5. Using the computed gene expression matrix with deletions as input, along with the α and β values, to minimize the following:

$$L(M) = \sum_{i=1}^{n} \sum_{g=1}^{k} \left(M_i^g - I_{\text{total}}^g\right)^2$$

In addition, we add a regularization term to the above error term using the LASSO constraint (Tibshirani 1994):

$$L(M) = \left(\sum_{i=1}^{n} \sum_{g=1}^{k} \left(M_i^g - I_{[i]}^g\right)^2 \right) + \sum_{i=1}^{n} \sum_{j=1}^{k} \|M_{j,i}\|_1$$

where $M_{j,i}$ is the entry of M at row j and column i, and $I_{[\mathrm{i}]}^g$ is the total influence with transcription factor i deleted (e.g. activity = 0).

In order to verify our results, we can check our determined parameters by performing double deletions as follows: perform a deletion, as above, of two transcription factors, i and j. Then, the validation statistic is as follows:

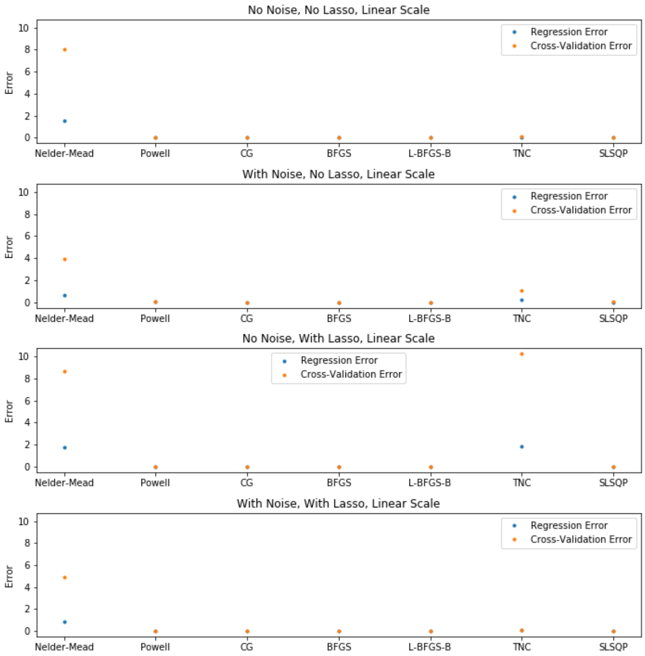
$$E(M) = \sum_{i \neq j}^{n} E_{[i,j]}^g - I_{[i,j]}^g$$

where, similarly to before, $E_{[i,j]}^g$ is the true expression value with transcription factors i and j deleted, and $I_{[i,j]}^g$ is the total calculated influence with activities of i and j set to 0.

5 Experimental Results

Using SciPy's Optimize module, we obtained an optimal expression matrix and regression error from each of the following algorithms: Nelder-Mead, Powell, conjugate gradient (CG), BFGS, L-BFGS-B, truncated Newton (TNC), and sequential least squares programming (SLSQP). We then calculated the cross-validation error for each solver's output, compared it to that solver's regression error, and plotted the two values on the same graph. Afterwords, we implemented two modifications: a slight amount of Gaussian noise (μ=0, σ=0.01) added to the expression matrix, which helps to more closely replicate a true biological system, and a regularization term using a LASSO constraint. The optimization, error-calculation, and plotting procedure was then re-run using no noise with a LASSO constraint, added noise without a LASSO constraint, and added noise with a LASSO constraint. The results to all of which can be seen in figures 1 and 2, as well as table 1.

As an interesting side-note, although all solvers did eventually converge to an optimal value, we observed that the conjugate gradient solver took substantially longer to do so. This author believes that the drastic increase in run-time relative to all other algorithms was due to the objective matrix's very irregular and difficult to calculate gradient, an inherent byproduct of randomly generating the initial guess.



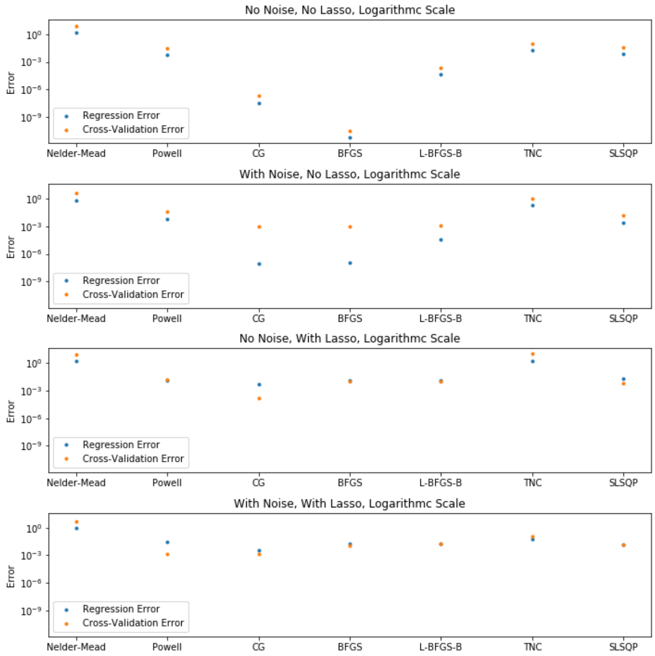


Figure 1: (left, top-to-bottom) comparison between optimization algorithms' regression and cross-validation errors under varying conditions, with linear Y-axis scaling

Figure 2: (right, top-to-bottom) the same data as in figure 1, with logarithmic Y-axis scaling for visible scaling between data points

	Nelder-Mead	Powell	CG	BFGS	L-BFGS-B	TNC	SLSQP
Regression Error, No Noise, No Lasso	1.570399	0.005297	3.442840e-08	5.509661e-12	0.000041	0.016734	0.006539
Cross-Validation Error, No Noise, No Lasso	8.016431	0.031481	2.045252e-07	3.152610e-11	0.000243	0.088541	0.038802
Regression Error, With Noise, No Lasso	0.663720	0.005797	8.458132e-08	9.971839e-08	0.000038	0.203556	0.002327
Cross-Validation Error, With Noise, No Lasso	3.939525	0.036057	1.065899e-03	1.066311e-03	0.001363	1.050634	0.014795
Regression Error, No Noise, With Lasso	1.733664	0.012525	4.951642e-03	1.338924e-02	0.014795	1.809919	0.022748
Cross-Validation Error, No Noise, With Lasso	8.702683	0.018723	1.582944e-04	1.028889e-02	0.009835	10.244930	0.007273
Regression Error, With Noise, With Lasso	0.887041	0.029083	3.474896e-03	1.614842e-02	0.019201	0.053610	0.014834
Cross-Validation Error, With Noise, With Lasso	4.921173	0.001264	1.409979e-03	1.214882e-02	0.016386	0.128326	0.014611

Table 1: Numerical data for the plots shown in figures 1 and 2

6 Discussion

We obtained satisfactory results in regression error and cross-validation across all categories. The addition of Gaussian noise, intended to mimic noisy biological datasets, generally resulted in an increase in regression error. However, this increase was generally small and corresponded to a simple increase in regression error due to the added differences in the expected and actual expression values and was not indicative of loss of convergence. In addition, the LASSO constraint increased regression error but decreased validation error, as expected, as the regularization was intended to produce greater fidelity to the original parameter matrix and thus would produce lower values for the validation error statistic. Overall, although the matrices obtained from the minimization were not unique, our method generally produced good approximations to the true expression values in the validation step.

7 Recommendations for Future Work

Further work can be done to test the effects of repressor presence on the minimization. In addition, other regularizations such as the Ridge constraint and Elastic Net constraint can be added. In addition, adding in other biologically inspired constraints, parameters, and estimates based on enzyme activity, RNA degradation, ribosome activity, and RNA transport can be added to the model to increase predictive ability and applicability to real biological situations. Furthermore, estimates of basal activity level (β) and scaling factors (α) could be incorporated into the regression, and additional experiments that change these values for each gene could be performed. This work could also be extended to a more sophisticated model, such as the thermodynamic model of gene expression, which models interactions between individual components of the model. Furthermore, a more sophisticated solver could be designed that takes advantage of inherent constraints in the system.

8 Contributions

JG contributed the Introduction, Related Work, and Technical Details sections. PDR contributed the Experimental Results section, including figures and tables. Both participated in development of the model and programs for determining parameters from expression data, as well as writing the Discussion and Recommendations for Future Work sections of the paper.

References

[1] Tibshirani, R. (1994). Regression Shrinkage and Selection Via the Lasso. JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B, 58, 267–288. Retrieved from http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.35.7574

[2] von Dassow, G., Meir, E., Munro, E. M., & Odell, G. M. (2000). The segment polarity network is a robust developmental module. Nature, 406(6792), 188âĂŞ192. http://doi.org/10.1038/35018085

[3] Dong, Z., Song, T., & Yuan, C. (2013). Inference of Gene Regulatory Networks from Genetic Perturbations with Linear Regression Model. PLoS ONE, 8(12), e83263. http://doi.org/10.1371/journal.pone.0083263

[4] Li, Y., Liang, M., & Zhang, Z. (2014). Regression Analysis of Combined Gene Expression Regulation in Acute Myeloid Leukemia. PLoS Computational Biology, 10(10), e1003908. http://doi.org/10.1371/journal.pcbi.1003908

[5] Oron, A. P., Jiang, Z., & Gentleman, R. (2008). Gene set enrichment analysis using linear models and diagnostics. Bioinformatics (Oxford, England), 24(22), 2586âĂŞ91. http://doi.org/10.1093/bioinformatics/btn465

[6] Kang, Y., Liow, H.-H., Maier, E. J., & Brent, M. R. (2018). NetProphet 2.0: mapping transcription factor networks by exploiting scalable data resources. Bioinformatics, 34(2), 249âĂŞ257. http://doi.org/10.1093/bioinformatics/btx563

[7] Schlitt, T., & Brazma, A. (2007). Current approaches to gene regulatory network modelling. BMC Bioinformatics, 8(Suppl 6), S9. http://doi.org/10.1186/1471-2105-8-S6-S9

[8] Becker, M. G., Walker, P. L., Pulgar-Vidal, N. C., & Belmonte, M. F. (2017). SeqEnrich: A tool to predict transcription factor networks from co-expressed Arabidopsis and Brassica napus gene sets. PLOS ONE, 12(6), e0178256. http://doi.org/10.1371/journal.pone.0178256

[9] Wilkinson, A. C., Nakauchi, H., & Gottgens, B. (2017). Mammalian Transcription Factor Networks: Recent Advances in Interrogating Biological Complexity. Cell Systems, 5(4), 319âĂŞ331. http://doi.org/10.1016/j.cels.2017.07.004