Comparisons of Statistical Modeling for Constructing Gene Regulatory Networks

by

Xiaohui Chen

B.Sc., Zhejiang University, 2006

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

The Faculty of Graduate Studies

(Bioinformatics)

The University Of British Columbia

(Vancouver)

August, 2008

Abstract

Genetic regulatory networks are of great importance in terms of scientific interests and practical medical importance. Since a number of high-throughput measurement devices are available, such as microarrays and sequencing techniques, regulatory networks have been intensively studied over the last decade. Based on these high-throughput data sets, statistical interpretations of these billions of bits are crucial for biologist to extract meaningful results. In this thesis, we compare a variety of existing regression models and apply them to construct regulatory networks which span trancription factors and microRNAs. We also propose an extended algorithm to address the local optimum issue in finding the *Maximum A Posteriori* estimator. An *E.coli* mRNA expression microarray data set with known *bona fide* interactions is used to evaluate our models and we show that our regression networks with a properly chosen prior can perform comparably to the state-of-the-art regulatory network construction algorithm. Finally, we apply our models on a p53-related data set, NCI-60 data. By further incorporating available prior structural information from sequencing data, we identify several significantly enriched interactions with cell proliferation function. In both of the two data sets, we select specific examples to show that many regulatory interactions can be confirmed by previous studies or functional enrichment analysis. Through comparing statistical models, we conclude from the project that combining different models with over-representation analysis and prior structural information can improve the quality of prediction and facilitate biological interpretation.

Keywords: regulatory network, variable selection, penalized maximum likelihood estimation, optimization, functional enrichment analysis

Table of Contents

Appendices

List of Tables

List of Figures

Acknowledgements

First and foremost, I owe many thanks to my two supervisors: Prof. Raphael Gottardo and Prof. Kevin Murphy. This work would be impossible without their continuous efforts to keep my statistical ball rolling. Their many valuable and enlightened suggestions not only helped me from constantly studying new things, but also taught me how to be an independent researcher. Thanks to both of you for being great mentors, both professionally and personally.

I am also grateful to Dr.Mauricio Neira, who actually was my third unofficial supervisor. Thank you Mauricio for helping me so many times with valuable support and guidance on post-modeling bioinformatics analysis. Your never ending feedback on my analysis is highly appreciated. Besides, I would like to both thank you for helping prepare part of the data.

Last but not least, I have to say thanks to Prof. Arnaud Doucet and Dr. François Caron, both of whom instructed and helped me with the statistical modeling.

To my parents
Rui Chen and Yuping Ou

Chapter 1

Introduction and literature review

Gene regulatory networks have been intensively studied over the last decade because of their inherent scientific and medical importance (e.g., in the fight against cancer) as well as the development of high-throughput measurement devices. The data generated by these high-throughput experiments typically contain:

- **Gene microarray expression data:** Each microarray is used to capture the snapshot of the transcriptional status of cells, and one experiment usually includes several microarrays to reflect the change of status of a biological system over time (may contain repetitive measurements). Although mRNA is a precursor to its downstream protein product which directly executes biological functions, a large amount of previous regulatory network studies used mRNA microarray data as a proxy for the amount of protein products [24, 46, 64].

- **Sequence information:** The discovery and characterization of regulatory controlling sequences have been greatly facilitated by sequencing projects. Generally, identification of regulatory elements in the genome is mainly based on comparative sequence analysis and some Chromatin ImmunoPrecipitation (ChIP) binding experiment data [40]. There are many secondary curated databases which consist of reliable, verified regulations, but this approach is not scalable for quick identification of regulatory relationships *de novo*, such as comparative genomics methods and algorithms based on sequence complementarity.

1.1 MiroRNA and Transcription factors

At the molecular level, microRNA (miRNAs) and transcription factors (TFs) are two important categories of regulators that control thousands of mammalian genes. TFs are the proteins that can bind to specific parts of DNA (usually one gene(s)) using DNA binding domains and therefore help initiate/repress transcription, while miRNAs are small non-coding RNAs that regulate gene expression at the post-transcriptional level by affecting the stability and translational processes of gene transcripts. For TFs, they could either increase or decrease gene transcription, but generally for miRNAs, they mediate post-transcriptional gene silencing through degradation of mRNAs or inhibition of protein production [34]. In principle, miRNAs could help to explain discrepancies between mRNA and protein levels. For example, those discrepancies may seriously complicate the use of mRNA profiles to study chemoresistance [10]. Unfortunately, there still continues to be a question as to how well transcript levels predict translated protein levels; a graphical illustration of TFs and miRNAs coordinated gene expression processes is shown in Figure 1.1. Recently, [52] comprehensively studied the global architecture and network local motifs of human miRNA-TF regulatory networks. Based on the prediction datasets using sequence complementarity and conservation, they revealed that the combinatorial property of miRNA interactions and miRNA-TF cooperation for targeting genes is fundamental for the precise and complex nature of regulatory systems.

1.2 Previous work

Most of previous work on discovering regulator-target association is branched into two fields, corresponding to the two major types of data mentioned at the beginning of this chapter. The former field is based on mining correlated gene transcription patterns from mRNA microarray data. This includes clustering [41], Bayesian networks [24, 33, 46] and linear regression models [17, 64]. The latter one focuses on using sequence complementarity and phylogenetic conservation; for example, see [7, 35, 39, 58]. These

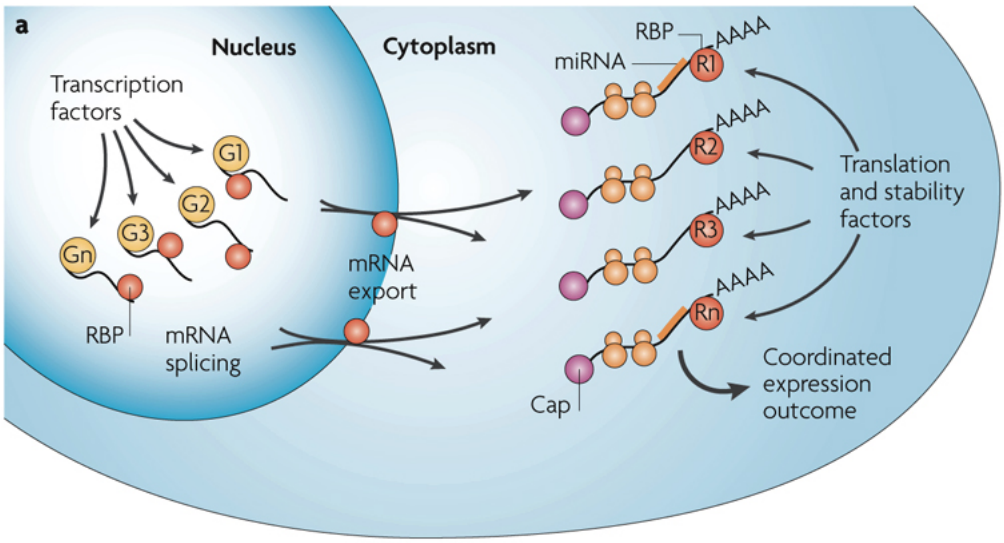


Figure 1.1: An illustration of coordinated gene expression processes. TFs activate a set of genes in nucleus (G1-Gn). After genes being transcribed, the resulting mRNAs are bound by RNA-binding proteins (RBP), spliced and subsequently exported to the cytoplasm. RBPs and miRNAs can affect the stability of transcripts, activate or repress their translations. Source: *RNA regulons: coordination of post-transcriptional events.* Keene, J. 2007 *Nature Reviews Genetics.* [34]

sequence based methods, however, have limited specificity due to imperfect matches between TF-TF Binding Sites (TFBS) and the miRNA-target. In the miRNA regulating transcripts domain, [31] proposed a Bayesian linear model to integrate the evidence from both sequence and expression data. The inferred regulatory network is a subset of the prior network given by a sequence-based target-finding program. To the same end, in the TF regulating gene domain, [17, 48] combined mRNA expression data with positive protein binding data from ChIP experiments to improve the predictive ability. Their method further classified ChIP positives into functional and non-functional TF targets using linear regression models. Currently, there is not much work on combining TFs and miRNAs to construct regulatory networks based on both microarray expression data and sequence-based structural information. Since more variety and volume of data sources are available, this joint modeling is expected to become popular in the near future.

1.3 Project motivations and goals

As discussed in previous sections, the regulatory systems seldom function through only a few mechanisms. Hence, building a comprehensive regulatory network spanning TFs and miRNAs at different molecular levels and integrating both microarray expression data and prior structural information from sequence based predictions are essential. Moreover, due to the sparseness property of regulatory networks, we formulate the problem as a variable selection problem. The goals of the project are summarized as below:

1. Compare various statistical models and algorithms without integrating structural information.

2. By incorporating available prior structural information, compare the results of different models. Comparisons are also made between combining two information sources and merely using expression data to construct networks.

3. Combine functional enrichment analysis to interpret learned regulatory interactions.

The rest of the thesis is organized as following:

- Chapter 2 briefly presents the models we are using and the existing and our proposed algorithms to solve the models. Details of the modeling issues are put into the appendix.

- Chapter 3 presents the results of a simulation study and a real data example (from *E.coli*) where the true regulatory network is known. The purpose of this chapter is to compare various models and algorithms without prior structural information, from a statistical point of view. Model performance is measured by a precision-recall curve. Further, several *hub* genes are selected with common predictions from high performance algorithms and we show that at the functional level, the learned networks agree with current literature.

- Chapter 4 presents the results on a p53-centered data set (NCI60) and prior structural information is used to infer a network. The results are presented in several specific examples, which are of our biological interests, i.e. p53 tumor related genes/miRNAs. We combine the variable selection and functional over-representation analysis to detect *bona fide* interactions.

- Chapter 5 concludes the thesis with a discussion.

Chapter 2

Models and algorithms

2.1 Model: Linear regression

The model we are assuming is a multivariate linear regression model.

$$y = X\beta + e \tag{2.1}$$

where

- $X \equiv X_{n\times p}$ is the design matrix consisting of expression values for all known candidate regulators, including TFs and microRNAs
- y is the response vector of target gene expression values
- β is the regulatory strength: $\beta > 0$ means up-regulation, $\beta < 0$ means down-regulation, and $\beta = 0$ represents no regulation
- e is an error term assumed to be Gaussian with constant variance σ^2, i.e. $N(0, \sigma^2 I)$.

To capture the dependency between the regulators and target gene, we consecutively regress the expression levels of every target gene to X, while keeping X fixed for all.

2.2 Variable selection: penalized likelihood

In many situations, the assumed linear model may be redundant in the sense that not all of its predictors have significant effects on the response. For example, there may be hundreds or even thousands candidate regulators for one gene. Thus the full linear model is usually over-parameterized.

Intuitively, including more β's, the prediction error would decrease and the fit would improve. However, the improvement is very marginal compared with the cost of estimating a much more complex model. Furthermore, the interpretation becomes unclear for separating the true predictors from the irrelevant variables. This is because the estimations yielded from minimizing a least squares (LS) objective function typically do not equal zeros. Hence, selecting a proper subset of predictors from the full model is crucial for high-dimensional statistical modeling. The basic idea is to penalize more complex models while preferring simpler models. In terms of mathematical formulation, many variable selection problems aim to maximize a penalized likelihood function, which is equivalent to minimizing a penalized negative-log-likelihood $\mathcal{O}(\beta, \lambda|X, y)$ function

$$\mathcal{O}(\beta, \lambda|X, y) = -\ell(\beta) + \sum_{j=1}^{p} p_{\lambda_j}(|\beta_j|) \tag{2.2}$$

where ℓ is the log-likelihood and $p_{\lambda_j}(|\beta_j|)$ is the penalty term for each large β absolute value. Generally speaking, large penalties tend to shrink the β coefficients to zeros. Here in the dimensionality reduction context and the rest of the thesis, we use variable selection and model selection interchangeably.

In the context of linear regression, i.e. when $p_{\lambda_j} = 0$, the maximum likelihood (ML) and least squares approaches provide the same estimator $\hat{\beta}_{MLE}$, which does not depend on σ^2. Thus σ^2 can be absorbed in the penalty terms. That is to say given λ_j, we can omit σ^2 and minimize the penalized LS objective function over β and λ_j

$$\mathcal{O}(\beta, \lambda|X, y) = \frac{1}{2}||y - X\beta||_2^2 + \sum_{j=1}^{p} p_{\lambda_j}(|\beta_j|) \tag{2.3}$$

where $||\cdot||_2$ is the Euclidean norm of a vector, and without loss of generality, we assume p_{λ_j} takes the same functional form for all $j = 1 : p$, i.e. $p_{\lambda_j} = p_\lambda$ in Eq.(2.3). Specifically, choosing $p_\lambda = \lambda|\beta|^p$ yields the $\mathcal{L}^p$ penalty [22]. By avoiding the over-penalization of β with large absolute values, Fan and Li (2001) proposed a modified $\mathcal{L}^1$ penalty [19], i.e the Smoothly Clipped

Absolute Deviation (SCAD) penalty.

The problem becomes finding $\hat{\beta}$ such that

$$(\hat{\beta}, \hat{\lambda}) = \arg\min_{\beta,\lambda} \mathcal{O}(\beta, \lambda | X, y) \tag{2.4}$$

λ can be determined on a linear grid by cross-validation (CV) giving the smallest prediction errors.

2.3 Optimization: finding penalized ML estimator

It turns out the minimization of the objective function, $\mathcal{O}(\cdot)$, is not trivial. For $p_\lambda(\cdot) = |\cdot|$ (a.k.a. LASSO) [56], it can be efficiently solved by Osborne's [44] and the Least Angle Regression (LARS) algorithm [16]. But these algorithms are not generalized to other penalties (see Appendix A Methods). Caron and Doucet [12] proposed an Expectation-Maximization (EM) algorithm to find the *Maximum A Posteriori* (MAP) estimator for general penalty functions. Further, for $\mathcal{L}^1$ and SCAD penalties (see Appendix A Methods), Hunter and Li [32] proposed a Majorize-Minimize (MM) algorithm to optimize the same objective function and avoid the local optimum problem with the EM algorithm. The MM can be seen as a generalization of the EM algorithm. In this thesis, we extended the idea of the MM algorithm and propose a new MM algorithm to handle the general penalizations. The details will be addressed in the following paragraphs and also Appendix A Methods section.

2.3.1 EM algorithm

[20], [27] and [12] proposed an EM algorithm [15] to find the posterior modes of Laplacian (L1), Normal-Gamma (NG), Normal-Jeffreys (NJ), and Normal-Inv-Gaussian (NIG) models. In general, the marginal prior distribution of regression coefficients in this family can be factored as a scale

mixture of Gaussian densities:

$$\pi(\beta) = \int \mathcal{N}(\beta;\, 0,\, \tau^2)\, p(\tau^2)\, d\tau^2 \tag{2.5}$$

Choosing the prior of τ_j^2 to be Laplace, Gamma, Jeffreys and Inverse Gaussian corresponds to the L1, NG, NJ and NIG models, respectively.

In the EM framework, β is the parameter vector of interest, and τ can be seen as missing data. Then, the posterior mode can be iteratively found by maximizing conditioned on the last estimate $\hat{\beta}^{(k)}$:

$$\hat{\beta}^{(k+1)} = \arg\max_{\beta} Q(\beta, \hat{\beta}^{(k)}) \tag{2.6}$$

where $Q(\beta, \hat{\beta}^{(k)})$ is defined as the expected complete-data log-likelihood:

$$\int \log(p(\beta|X, y, \tau)) p(\tau|\hat{\beta}^{(k)}, X, y)\; d\tau$$

Then, in the next iteration, $\hat{\beta}^{(k+1)}$ is given by maximizing the $Q(\beta, \hat{\beta}^{(k)})$ function over β.

For linear regression models, the update expression of $\hat{\beta}^{(k+1)}$ can be given in closed form:

$$\hat{\beta}^{(k+1)} = (X^T X + U^{(k)})^{-1} X^T y \tag{2.7}$$

where

$$U^{(k)} = \begin{pmatrix} u_1^{(k)} & 0 & \dots & 0 \\ 0 & u_2^{(k)} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & u_p^{(k)} \end{pmatrix}$$

with

$$u_j^{(k)} = \frac{p^{'}(|\beta_j|)}{|\beta_j|}.$$

Here $p^{'}(x)$ denotes the derivative of penalty function $p(.)$ evaluated at x.

2.3.2 MM algorithm

Depending on the optimization context, the MM algorithm refers to the majorize-minimize or minorize-maximize algorithm, which is a generalization of the famous EM algorithm. Because the MM algorithm is a very general optimization method (for example iteratively reweighted least squares algorithm is also a special MM algorithm which uses tangent lines to surrogate the original objective function), we adopt the MM algorithm to solve penalized models with various penalties. In the problem where no missing data can be naturally introduced, MM requires the construction of a surrogate function for the objective function, which is equivalent to the E-step in the EM framework. The M-step is subsequently implemented to optimize the transfered surrogate function. These two steps are iteratively performed in turn until convergence. The MM algorithm has the monotonicity property of the EM if the majorize/surrogate function constructed satisfies the following two properties:

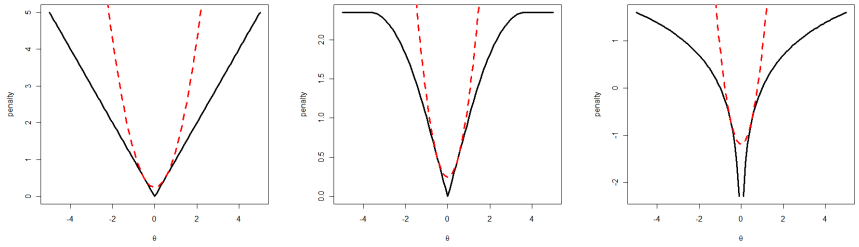
$$\Phi(\theta, \theta_0) \geq p(\theta), \forall \theta \in \Theta \tag{2.8}$$

$$\Phi(\theta_0, \theta_0) = p(\theta_0) \tag{2.9}$$

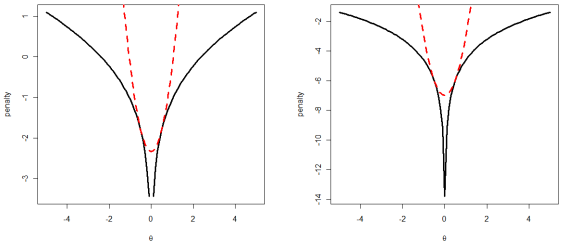
where, $\Phi(\theta, \theta_0)$ is the majorize function of the objective function $p(\theta)$ at θ_0. Specifically, in the context of penalized linear regression, $p(\cdot)$ is the penalty. Then, it is obvious that by subtracting Eq.(2.9) from Eq.(2.8), we get $\Phi(\theta, \theta_0) - \Phi(\theta_0, \theta_0) \geq p(\theta) - p(\theta_0)$, which in turn guarantees the monotonically non-increasing property of original objective function:

$$\Phi(\theta, \theta_0) \leq \Phi(\theta_0, \theta_0) \implies p(\theta) \leq p(\theta_0) \tag{2.10}$$

Thus by choosing a more smooth surrogate function, instead of directly minimizing $p(\theta)$, the optimization is transfered to minimizing the differentiable function $\Phi(\theta, \theta_0)$ over θ. The penalty functions and corresponding upper bound surrogate functions are shown in Figure 2.1.



(a) L1 (Lipschitz continuous) (b) SCAD (Lipschitz continuous) (c) NJ (Infinite derivative)



(d) NG (Infinite derivative) (e) NIG (Infinite derivative)

Figure 2.1: Penalty functions and their corresponding quadratic surrogate functions around origin. Solid lines are original penalty and the red dashed lines are the upper bound surrogate functions of solid lines with $\theta_0 = 0.5$ and $\epsilon = 10^{-8}$.

MM1 algorithm

In the context of the linear regression, one can define the quadratic function

$$\Phi_1(\theta, \theta_0) = p_\lambda(|\theta_0|) + \frac{(\theta^2 - \theta_0^2)p_\lambda^{'}(|\theta_0|+)}{2|\theta_0|} \tag{2.11}$$

it can be shown that $\Phi_1(\theta, \theta_0)$ majorizes $p_\lambda(|\theta|)$ at $\pm|\theta_0|$ (Proposition 3.1 [32]). Here, $p_\lambda^{'}(|\theta|+)$ denotes the limit of $p_\lambda^{'}(x)$ as $x \to |\theta|$ from above. By iteratively minimizing $\Phi_1(\theta, \theta_0)$, the solution to this algorithm is identical to the solution of the M-step in the EM algorithm. Fan and Li called this algorithm Local Quadratic Approximation (LAQ) [19], but here we call it MM1 for simplicity.

MM2 algorithm

Note that in Eq.(2.11), $\Phi_1(\theta, \theta_0)$ is undefined at $|\theta_0| = 0$, and hence when a parameter is estimated to be zero or close enough to zero, then it is excluded from the subsequent sub-model and never re-enters the model again. To ease this problem, [32] proposed a modified version of LQA/MM1 with perturbed penalty function $p_{\lambda 2}$

$$p_{\lambda 2}(|\theta|) = p_\lambda(|\theta|) - \epsilon \int_0^{|\theta|} \frac{p_\lambda^{'}(t)}{\epsilon + t} dt \tag{2.12}$$

paired with its majorize function

$$\Phi_2(\theta, \theta_0) = p_{\lambda 2}(|\theta_0|) + \frac{(\theta^2 - \theta_0^2)p_\lambda^{'}(|\theta_0|+)}{2(\epsilon + |\theta_0|)} \tag{2.13}$$

Under certain regularity conditions (c.f. Proposition 3.2 [32]) and provided the space of θ is compact, it can be shown as $\epsilon \downarrow 0$, $|p_{\lambda 2}(\theta) - p_\lambda(\theta)| \to 0$ uniformly over θ. The iterative ridge-type solution can be given in closed-form expression

$$\hat{\beta}^{(k+1)} = (X^T X + V^{(k)})^{-1} X^T y \tag{2.14}$$

where

$$V^{(k)} = \begin{pmatrix} v_1^{(k)} & 0 & \dots & 0 \\ 0 & v_2^{(k)} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & v_p^{(k)} \end{pmatrix}$$

with

$$v_j^{(k)} = \frac{p_\lambda^{'}(|\hat{\beta}^{(k)}|+)}{|\hat{\beta}^{(k)}| + \epsilon}.$$

MM3 algorithm

For NG, NJ and NIG models, the regularity conditions stated in [32] are not met. For example, in the NJ model, the first derivative of the penalty function is $\frac{1}{|\theta|}$ (see Table A.1), which is unbounded as $|\theta| \downarrow 0$. Also the penalty itself is unbounded below when $|\theta| \downarrow 0$ (Figure 2.2(c)). Even worse, when $\frac{p_\lambda^{'}(t)}{\epsilon+t}$ is not integrable around the origin, it is not possible to use this functional form to derive the new penalty function. To handle this issue, we propose a perturbed penalty function, fix $\epsilon > 0$:

$$p_{\lambda 3}(|\theta|) = p_\lambda(|\theta| + \epsilon) - \epsilon \int_0^{|\theta|} \frac{p_\lambda^{'}(\epsilon + t)}{\epsilon + t} dt \tag{2.15}$$

Define:

$$\Phi_3(\theta, \theta_0) = p_{\lambda 3}(|\theta_0|) + \frac{(\theta^2 - \theta_0^2) p_\lambda^{'}((|\theta_0| + \epsilon)+)}{2(|\theta_0| + \epsilon)} \tag{2.16}$$

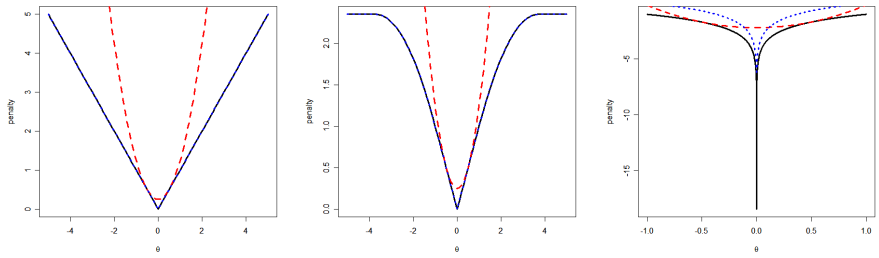
Similarly, by the above construction, we could prove the following theorem.

Theorem 2.3.1. *Let $p_\lambda(\cdot)$ be a function defined on $[0, \infty)$. Suppose $p_\lambda(\cdot)$ satisfies the regularity conditions:*

1. *differentiable, nondecreasing and concave on $(0, \infty)$*
2. *continuous at the origin*

$\forall \epsilon > 0$, define $p_{\lambda 3}(|\theta|)$ and $\Phi_3(\theta, \theta_0)$ as above. Then

1. *$\Phi_3(\theta, \theta_0)$ majorizes $p_{\lambda 3}(|\theta|)$ at $\pm|\theta_0|$, $\forall \theta_0 \in \mathbb{R}$.*



(a) L1 (Lipschitz continuous) (b) SCAD (Lipschitz continuous) (c) NJ (Infinite derivative)

Figure 2.2: Penalty functions and their corresponding surrogate functions around origin. Dotted lines (in blue) are original penalty, solid lines are the proposed perturbed penalty function, and dashed lines (in red) are the surrogate functions of solid lines with $\theta_0 = 0.5$ and $\epsilon = 10^{-8}$.

2. $\forall \mathcal{C} \subset \Theta$*, the space of* θ*, where* $\mathcal{C}$ *is compact. As* $\epsilon \downarrow 0$*,*

$$\sup_{\theta \in \mathcal{C}} |p_{\lambda 3}(\theta) - p_\lambda(\theta)| \to 0 \tag{2.17}$$

In particular, if $p_\lambda(\cdot)$ *is Lipschitz continuous on* $[0, \infty)$*, then* $\sup_{\theta \in \Theta} |p_{\lambda 3}(\theta) - p_\lambda(\theta)| \to 0$.

The proof of this theorem is given in Appendix B. Theorem 2.3.1 tells us that under the new surrogate function, the condition $p_\lambda(0+) < \infty$ can even be abandoned. As we have seen, this is essential for many kinds of priors that have been used in variable selection, for example NJ and NG etc. Then iteratively update expression for the parameters β can also be analytically given by

$$\hat{\beta}^{(k+1)} = (X^T X + W^{(k)})^{-1} X^T y \tag{2.18}$$

where

$$W^{(k)} = \begin{pmatrix} w_1^{(k)} & 0 & \dots & 0 \\ 0 & w_2^{(k)} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & w_p^{(k)} \end{pmatrix}$$

with

$$w_j^{(k)} = \frac{p_\lambda^{'}((|\hat{\beta}^{(k)}| + \epsilon)+)}{|\hat{\beta}^{(k)}| + \epsilon}.$$

The convergence criterion adopted here is different from the one used in [32]. We adjust ϵ, hence also the surrogate function (if necessary), to ensure that, for $\beta_j \neq 0$, $|\frac{\partial \mathcal{O}(\hat{\beta})}{\partial \beta_j}| < \tau$ at the convergence, where τ is a predefined *effective-zero* level. Algorithm 1 summarizes MM3 algorithm steps:

Algorithm 1 MM3 algorithm

Require: Design matrix X, response y, penalty name

1: Initialize ϵ to be moderately small
2: **repeat**
3: Run MM3 algorithm
4: **until** $\mathcal{O}(\hat{\beta}^{(k)}) - \mathcal{O}(\hat{\beta}^{(k+1)}) < \frac{\tau}{2}$
5: **if** $|\hat{\beta}_j^{(k+1)}| < \tau$ **then**
6: Set $\hat{\beta}_j^{(k+1)} = 0$
7: **end if**
8: Compute

$$\delta = \max_{j:\beta_j \neq 0} \left(\left| \frac{\partial p_{\lambda 3}(\beta)}{\partial \beta_j} - \frac{\partial p_\lambda(\beta)}{\partial \beta_j} \right|_{\hat{\beta}^{(k+1)}} \right)$$

9: **if** $\delta < \frac{\tau}{2}$ **then**
10: Finish and exit
11: **else**
12: Set $\epsilon \leftarrow \frac{\epsilon}{2}$
13: Go to line 2
14: **end if**
15: **return** Estimated coefficients/Selected variables

Remark: we do not adopt the procedure of presuming $\hat{\beta}_j^{(k+1)} = 0$ if its sub-differential is greater than τ, i.e. $\left|\frac{\partial \mathcal{O}(\hat{\beta})}{\partial \beta_j}\right|_{\hat{\beta}_j^{(k+1)}} > \tau$ as in [32]. Since any coefficient close to zero has been removed in the last step, we argue that the remaining coefficients after the algorithm converges are all non-zero. In fact, these *effective-zero* coefficients are due to the perturbation introduced, so they should be exactly zero if $\epsilon = 0$. Also from the simulation studies,

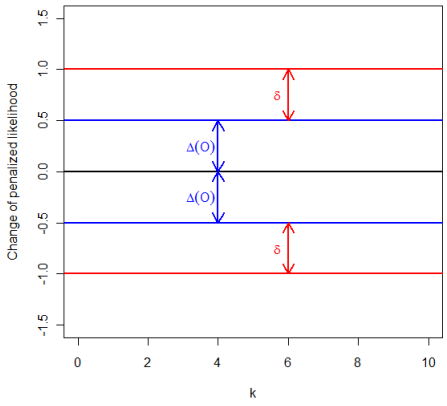


Figure 2.3: Convergence of unperturbed penalized likelihood function. Assuming $\tau = 1$. $\Delta(\mathcal{O})$ is the part regarding the change of unperturbed penalized likelihood. δ is the part measuring the goodness of approximation between the derivative of perturbed penalty and original penalty.

we find the estimates obtained by setting $\hat{\beta}_j^{(k+1)} = 0$ if $\left|\frac{\partial \mathcal{O}(\hat{\beta})}{\partial \beta_j}\right|_{\hat{\beta}_j^{(k+1)}} > \tau$ have quite bad performance under the NJ, NG and NIG priors in terms of correctly selected predictors (results not shown in this thesis). Finally, under this scheme, when the coefficient is estimated to 0, it could re-enter the model after tuning ϵ smaller, because we allow zero estimations to compete with others as long as convergence has not been arrived. The tolerance is decomposed as shown in Figure 2.3.

2.3.3 Connections between algorithms

The connection between the EM, MM2 and MM3 algorithms are summarized below:

- The EM algorithm requires to compute the expected complete-data log-likelihood for the E-step, while the MM algorithm needs a majorize function. The subsequent step of these algorithms are essentially the same: optimizing the transferred surrogate function. Hence, it is easy to see that the E-step is a special case of constructing an arbitrary sur-

rogate function. In fact, the solution to the EM algorithm is identical to that of LQA/MM1 algorithms

- If the marginal density of β is infinitely spiked around 0, then the EM and MM2 algorithms have the problem as forward/backward selection procedures, i.e. variables discarded cannot be re-introduced into models in later stage. The proposed MM3 algorithm can avoid this local optimality by adding a small perturbation to the derivatives while retaining the monotonicity of the EM and MM2 algorithms.

- The choice of ϵ, and thus the majorize function, is adapted to the estimated parameters. Hence, the choice of ϵ can be controlled by the information from data and is robust to its initial choice. As ϵ is dynamically tuning smaller, the MM3 usually takes a longer time to converge than the EM and the MM2 based on empirical experience.

- It is not difficult to see that $\beta = 0$ is a local optimum for the penalized linear regression with the EM, MM2, and MM3 algorithms. The EM with all priors and the MM2 with NJ, NG, and NIG priors have to drop the variables once they are estimated to be zero, and thus will be stuck at the origin. However, a simulation study will demonstrate that the MM3 with all priors has the ability to jump from the zero nodal points and correctly identify the true variables.

- It is well known that an L1 penalty can be viewed as the following constrained optimization problem:

$$\text{minimize } \frac{1}{2}||y - X\beta||_2^2$$

subjected to

$$\sum_{j=1}^{p} |\beta_j| \leq t$$

The penalized regression is equivalent to the constrained regression problem, in that, given a $\lambda \geq 0$, there exists a $t \geq 0$ such that these two problems share the same solution. Generally speaking, however,

the penalized formulation is not the Lagrangian function for the constrained formulation. However, for NG and NIG model, we can interpret the MM solutions to penalized LS as log-barrier functions (see Table A.1). The idea of the MM algorithms here is the same as that of the log-barrier function solution to the L1 penalized LS: they all prevent variables from becoming exactly 0 while remaining a good approximation to the original penalized functions. Specifically for L1, the log-barrier function can defined as

$$g(x) = \frac{1}{2}||y - X\beta||_2^2 + \lambda \sum_{j=1}^{p} |\beta_j| - \epsilon \log ||x||_2^2$$

When ϵ is sufficiently small, minimizers of $g(x)$ corresponds to minimizers of $\mathcal{O}(\beta, \lambda|X, y)$. For MM2 and MM3, the surrogate functions have the same role as $g(x)$, and we can prove that as ϵ goes down sufficiently close to 0, the original objective functions can be arbitrarily approximated by surrogate functions.

Chapter 3

Results on comparing across models

3.1 Simulation studies

To compare the performances of models with different penalizations and algorithmic solutions, we simulate 100 data sets, each with $n = 100$ points and $\sigma = 1$. Following the setup of [32], the true coefficients are set to be $\beta = (3, 0, 0, 0, 1.5, 0, 0, 0, 2, 0, 0, 0)^T$, with $p = 12$. This is the scenario to simulate 3 true TFs out of 12 candidate TFs for one gene. Because genes are usually dependent, we introduce three correlation levels between covariates: $\rho = 0.1, 0.5, 0.9$. Data X is scaled to mean 0 and unit variance and y is centered. For both MM2 and MM3 algorithms, the initial ϵ is set to 10^{-8} and *effective-zero* $\tau = 10^{-8}$. The hyper-parameters of all models are set to be $\alpha = 0.01$ and $c = 2^{1:10}$ (see Appendix A Methods for the notations). All models and algorithms are initialized with all 0's and the Ordinary Least Square (OLS) or the Maximum Likelihood Estimation (MLE). Optimal parameters are determined by 5-fold CV as used in [56]. The performances of various models are measured by the mean square error (MSE) (can be calculated in closed form by $(\hat{\beta}-\beta)^T \text{cov}(X)(\hat{\beta}-\beta)$), the number of correctly estimated zeros (i.e. equals 9), the number of correctly estimated non-zeros (i.e. equals 3) and the number of incorrectly estimated zeros. Note that we do not implement the SCAD prior with the EM algorithms.

First, we look at the results with all models initialized at zeros at different correlation levels, see Figure 3.1. The true number of non-zero coefficients is three, and it is easy to observe that the MM3 uniformly dominates the

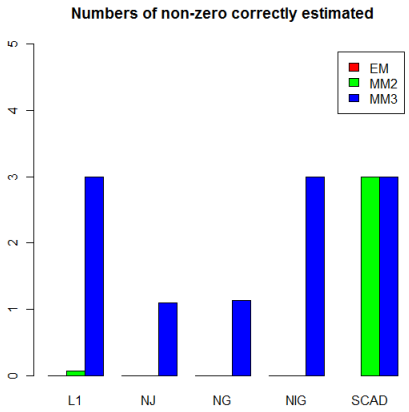
Table 3.1: Mean errors for linear models, averaged over 100 simulations. MSE is the mean square error. C is the number of correctly estimated zeros and I is the number of incorrectly estimated zeros. Boldfaced methods are best results.

	MSE	**Zeros**		MSE	**Zeros**		MSE	**Zeros**	
	Median	C	I	Median	C	I	Median	C	I
Model	$\rho = 0.1$			$\rho = 0.5$			$\rho = 0.9$		
NJ MM3	0.1104	8.54	0	0.1099	8.52	0	0.1254	8.51	0
NJ MM2	0.1104	8.54	0	0.1099	8.52	0	0.1255	8.51	0
NJ EM	0.1105	8.54	0	0.1099	8.52	0	0.1254	8.51	0
NG MM3	0.0992	8.75	0	0.0945	8.79	0	0.1280	8.77	0
NG MM2	0.1067	8.56	0	0.1055	8.54	0	0.1278	8.57	0
NG EM	0.0971	8.77	0	0.0975	8.75	0	0.1337	8.77	0
NIG MM3	0.0926	0	0	0.0872	0	0	0.1185	0	0
NIG MM2	0.0937	0.78	0	0.0846	0.9	0	0.1230	2	0
NIG EM	0.0905	0	0	0.0872	0	0	0.1256	0	0
L1 MM3	0.1364	4.18	0	0.1457	4.60	0	0.1692	4.86	0
L1 MM2	0.1533	5.36	0.32	0.1589	5.34	0.30	0.2275	6.06	0.46
L1 EM	0.1462	4.36	0	0.1481	4.63	0	0.1751	4.97	0
SCAD MM3	0.1454	4.31	0	0.1387	4.84	0	0.1568	5.21	0
SCAD MM2	0.1204	6.35	0	0.1414	6.68	0	0.1758	7.32	0

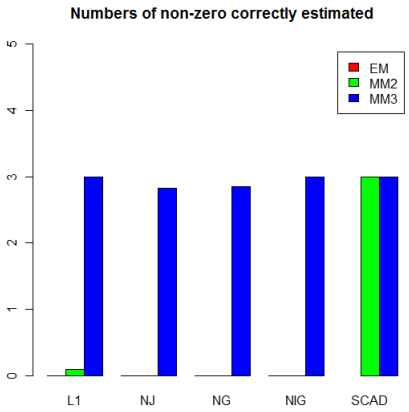
EM and the MM2 in terms of the number of correctly identified non-zeros. Comparing across models, the EM is always stuck at 0's for all models and the MM2 only selects variables in case of L1 and SCAD. For MM3, starting from zeros has no significant effect on the selected variables. Hence, the local optimum issue can be well handled by the MM3 in this extreme situation.

Secondly, we compare the results of algorithms starting from OLS estimations. From Table 3.1, we can see that the NJ and NG priors are performing among the best with all three algorithms, in terms of both correctly and incorrectly estimated zeros. The NIG has the lowest MSE but its estimates are not a thresholding rule, i.e. the NIG prior is not a variable selection prior and can produce many coefficients with smallest absolute values (c.f. Appendix A Methods). However, one can read small coefficients as zeros and thus corresponding variables are not selected. Further, the NJ models for the EM, MM2, and MM3 algorithms are essentially the same and

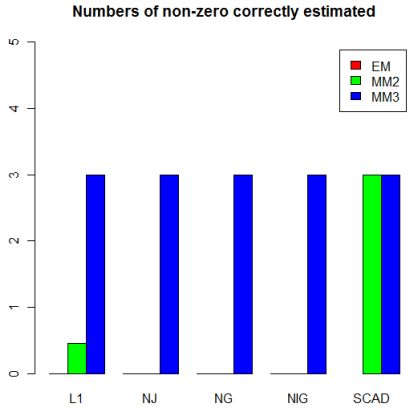
Figure 3.1: Simulation comparisons, initialized with all zeros start. Simulation averaged over 100 times



(a) $\rho = 0.1$



(b) $\rho = 0.5$



(c) $\rho = 0.9$

the L1 and SCAD priors have larger MSEs. Figure 3.2 shows that the number of zeros included in the NG and the NJ is higher than L1 and SCAD. Therefore, the NG, NJ, and NIG models tend to produce sparser models than L1 and SCAD. Finally, in general these variables selection models with the MM3 algorithm tend to have lower MSEs than those with the MM2 and the EM algorithms. In conclusion, local optimality is unlikely to occur if the starting points are not carefully chosen and the MM3 is better but the improvement is marginal in this case.

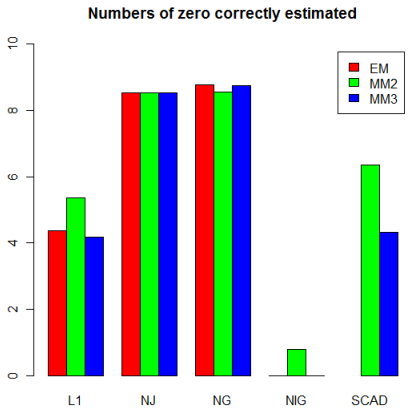
3.2 *E.coli* data set

In [18], Faith *et.al.* showed how RegulonDB database [50] can serve as a *ground truth* of regulatory network for *E. coli*. RegulonDB contains 3216 experimentally confirmed regulatory interactions among 1058 genes and 153 TFs. [18] assembled a compendium of 445 new and previously published *E. coli* K12 Affymetrix Antisense2 microarray expression profiles collected under various conditions. Compared with the ground truth (see Figure 3.3), this compendium is an ideal real dataset that we can evaluate the performance of various models. This dataset can be downloaded at the Many Microbe Microarrays database (M^{3D}) Web site (http://m3d.bu.edu/).

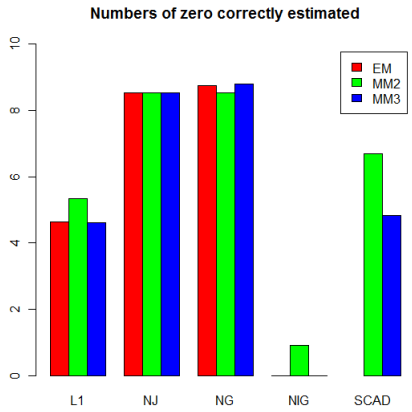
We also compare our linear models with the state-of-the-art gene regulatory network construction algorithms, i.e. *Context Likelihood Relatedness* (CLR) algorithm [18]. CLR is a mutual information (MI)-based method. The CLR algorithm estimates the likelihood of the MI score for a particular pair of genes, i and j, by comparing the MI value for that pair of genes to a background distribution of MI values (the null model). There are three major steps in CLR algorithms:

1. Computing raw MI values for every pair of genes
2. For each gene, normalizing the raw MI values
3. Estimating the joint likelihood measure (i.e. a significance measure of MI values), Z-score, of MI between every pair of gene

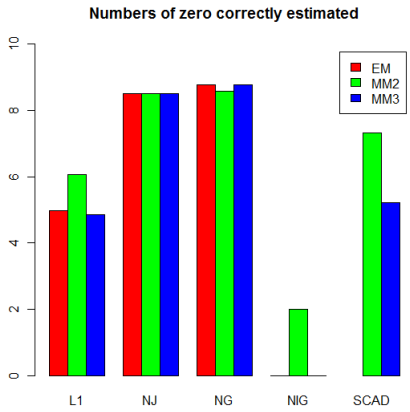
Figure 3.2: Simulation comparisons, initialized with MLE. Simulation averaged over 100 times



(a) $\rho = 0.1$



(b) $\rho = 0.5$



(c) $\rho = 0.9$

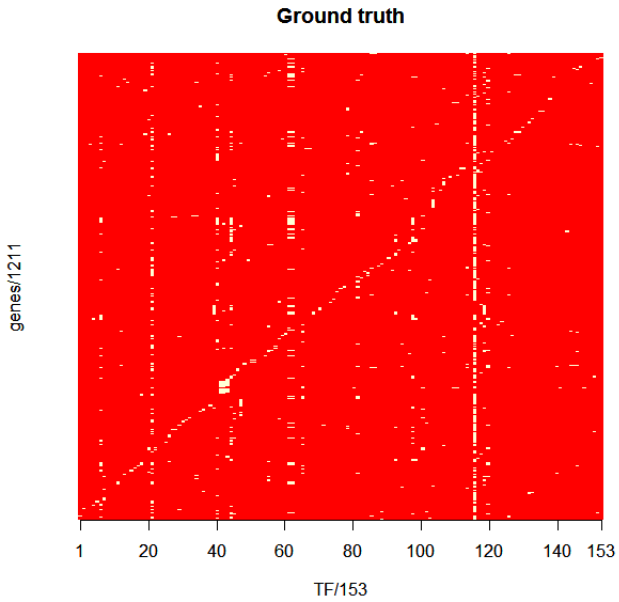


Figure 3.3: The real regulatory network of *E.coli* presented in a binary matrix M. A bright spot at M_{ij} is a regulatory relationship from TF j to gene i.

3.2.1 Precision-recall curves

The performances of all methods are measured by the area under the precision-recall curve (PR curve). *Precision* is defined as the fraction of true positives out of the total predicted positives, and *recall* the fraction of true positives out of the actual total positives. By formulation, they can be expressed as following:

$$precision = \frac{TP}{TP + FP} \tag{3.1}$$

$$recall = \frac{TP}{TP + FN} \tag{3.2}$$

Hence, the larger area under the PR curve, the better performance an algorithm has.

Faith [18] have applied various algorithms to genes on the whole *E.coli*

Table 3.2: Characteristics of 60% and 80% precise networks inferred from models with top perfermance in *E.coli*.

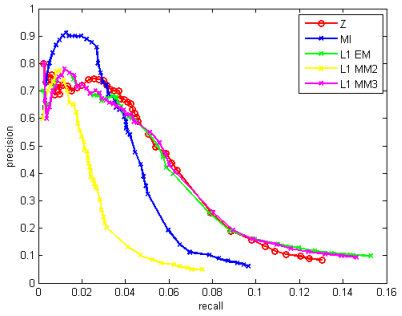
	60% precise network			80% precise network		
Model	TP	FP	Threshold	TP	FP	Threshold
L1 MM3	132	88	0.4530(Coeff value)	39	11	0.7531(Coeff value)
SCAD MM3	137	93	0.4435(Coeff value)	39	11	0.7610(Coeff value)
L1 EM	132	88	0.4396(Coeff value)	39	11	0.7585(Coeff value)
NIG EM	94	66	0.6446(Coeff value)	31	9	0.8954(Coeff value)
CLR	147	103	5.7905(Z-score)	8	2	10.658(Z-score)

Antisense2 microarray data, but we only apply linear models and MI-based methods on the set of genes which have representation in RegulonDB database, i.e. 1058 targets + 153 TFs = 1211 genes. This is because [18] can biologically validate the predicted and presumably novel interactions which are not curated in RegulonDB, while we have no such advantages and our main focus in this thesis is on the statistical modeling side. Thus, our results from CLR are different from those in the original paper applied on *E.coli* [18].

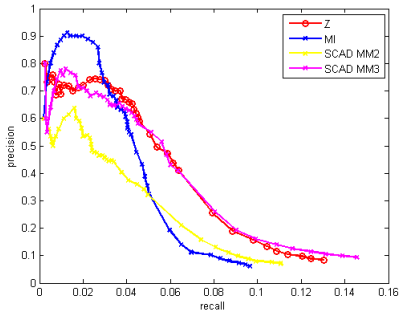
From the PR curve in Figure 3.4, we can see that the L1 penalty fitted with the EM and the MM3, the SCAD penalty with the MM3, and the NIG penalty with the EM, perform best in most algorithms and are comparable with the CLR algorithm. The NJ prior for all three algorithms are essentially overlapping each other. Performances of the NG prior for three algorithms are very close, although they are all inferior to the CLR algorithm. The network characteristics corresponding to models with top performance are summarized in Table 3.2.

To access the goodness-of-fit of various models and algorithms, scatter plots of the fitted expression values of target genes against their residuals are shown in Figure 3.5. It displays that the correlation between the actual expression values and fitted values is quite high. Moreover, the correlations from the L1 and SCAD models are better than the NIG model fitted using the EM algorithm.

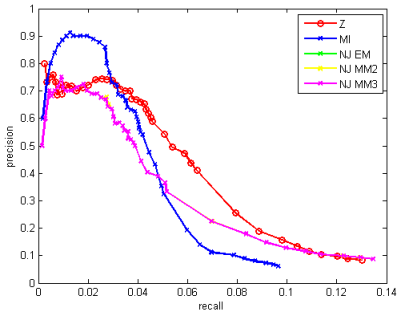
Figure 3.4: Precision-recall curves for various models and algorithms on *E.coli* data set. MI is the raw MI method. Z is CLR algorithm.



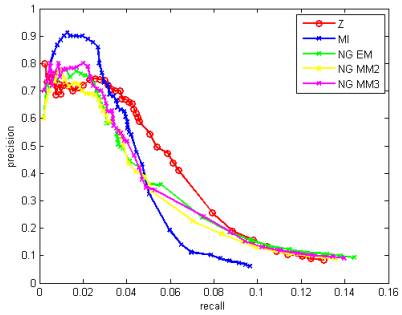
(a) L1



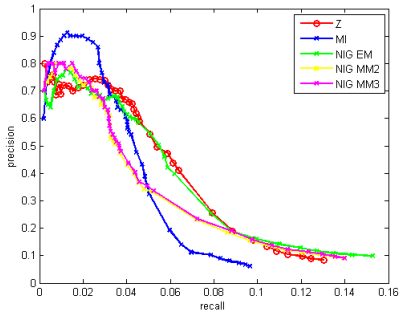
(b) SCAD



(c) NJ

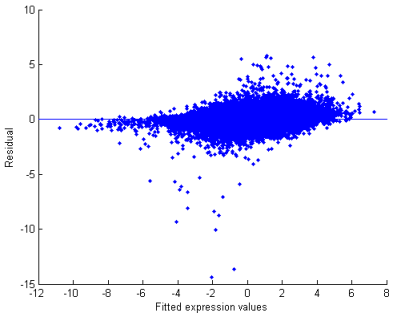


(d) NG

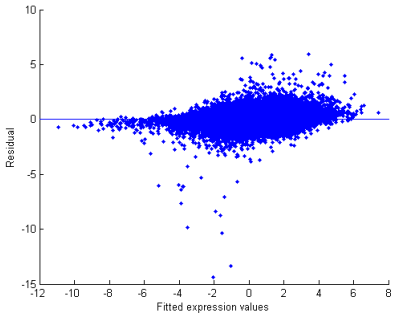


(e) NIG

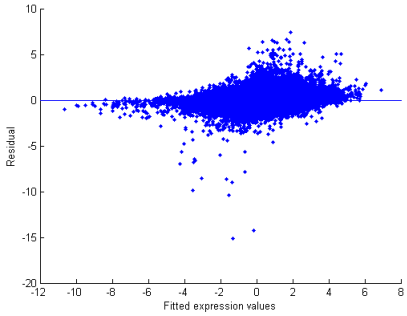
Figure 3.5: Scatter plots of the fitted expression values of target genes against their residuals.



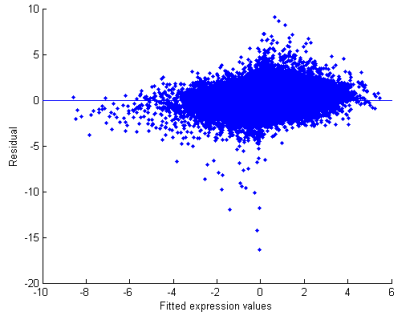
(a) L1 MM3



(b) SCAD MM3



(c) L1 EM



(d) NIG EM

Table 3.3: Number of targets regulated by transcription factors in the 60% precise network with $p \geq 5$ predicted targets with top performance algorithms, in *E.coli* network.

TFs	Targets #	Targets # inferred				
	in RegulonDB	CLR	L1 MM3	L1 EM	SCAD MM3	L1 NIG
fliA	42	40	43	44	44	44
lexA	16	6	7	7	7	7
hycA	7	10	9	14	9	14
gatR	6	6	7	7	7	7
yhiE	5	11	10	10	11	10

3.2.2 Visualizing and analyzing inferred networks

Networks are visualized through the `Graphviz` software, and the networks corresponding to Table 3.2 are shown in Figure 3.6, 3.7, 3.8, 3.9, 3.10, 3.11, 3.12, 3.13, 3.14, and 3.15. Specifically, to visualize the actual learned networks, we first extract the 80% precise networks as [18] did from the CLR algorithm, see Figure 3.6, 3.9, and 3.10 (i.e. the network is extracted by thresholding which gives 80% precision.). At this precision level, the learned networks from the LASSO and the NIG penalties are much better than the CLR algorithm. This agrees with the result in the PR curve, since at 80% precision, CLR almost has a recall of 0, which means there are almost no TP edges inferred. In fact, by trading off between the true positive and false positive rate in selecting a threshold for identifying significant regulatory interactions, a higher precision level leads to a small network with fewer false positives, while a lower threshold will include more false positive features. Hence, next we lower down the precision level to visualize the 60% precise networks, the resulting networks are shown in Figure 3.11, 3.12, 3.13, 3.14, and 3.15. The numbers in the nodes are the gene indexes in the expression matrix and each number can be uniquely mapped to one gene.

We summarized the TFs in the 60% precise network with $p \geq 5$ predicted operon targets with these top performance algorithms, and the results are presented in Appendix C Supplementary Materials C.1, C.2, C.3, and C.4. Here, we report 5 TFs with $p \geq 5$ connectivities to their tar-

get genes supported by all 5 algorithms with high performance, measured by PR curve (see Table 3.3). Their gene names are: fliA_b1922_at (fliA, gidx: 345), lexA_b4043_at (lexA, gidx: 615), hycA_b2725_at (hycA, gidx: 532), gatR_2_b2090_f_at (gatR, gidx: 413), and yhiE_b3512_at (yhiE, gidx: 1170). These genes are well documented in the literature. We suumarize their functions from the literature.

- **fliA:** we observed that it is a hub gene in both 80% and 60% precise networks and contols the transcriptional activities of many downstream genes. Very recently, fliA is reported to be one of the two global regulators (rssB and fliA) in *E.coli* through genetic screen experiments and the products of these two genes are involved in the regulation of major genetic networks [21]. It is known that even in the absence of identifiable exogenous stress, there remains a measurable, basal death frequency in *E.coli* populations. But the underlying mechanisms still remains unclear compared with those under stress conditions. [21] showed that mutant of the fliA gene affects the levels of different sigma factors within the cell and results reduced death frequencies in *E.coli* populations. Specifically, the inactivation of the fliA gene encodes the flagellar sigma factor. This results in the lack of expression of a number of genes involved in motility and chemotaxis, and consequently, non-motile cells. This loss of motility results in greater absolute availability of both RNA polymerase and energy for other processes within the cell, and consequently the cell may gain viability through a number of potential mechanisms by losing the motility pathway.

- **lexA:** a major regulator of DNA repair and DNA-binding transcriptional repressor of SOS regulon, is known to have a single well-conserved DNA-binding motif. It is one of the best-perturbed regulators in the microarray compendium due to the compendium's emphasis on DNA-damaging conditions. 16 regulatory interactions were collected in the RegulonDB database. The target genes that are supported by these 5 algorithms are: *dinF, recA, recN, sulA*. Interestingly, all of these four

genes are involved in the DNA repair and SOS response, which are important functions of *lexA* in *E.coli*.

- **hycA:** a formate hydrogenlyase regulatory protein. The target genes that are supported by these 5 algorithms are: *hycB, hycC, hycD, hycE, hycF, hycG, hycH, hydN*. The products of these target genes (*hycB-H*) are the subunits of hydrogenase, and *hydN* is involved in electron transport from formate to hydrogen. Thus, all genes have a common biological function: form and maturate the formate hydrogenlyase protein complex, which is an important molecule for energy production and conversion in *E.coli*.

- **gatR:** annotated to be a pseudogene, repressor for *gat* operon in NCBI. The target genes that are predicted by these 5 algorithms are: *gatA, gatB, gatC, gatD, gatY, and gatZ*. It is known that these genes form a *gat* operon with 7 ORFs in *E.coli* EC3132 and involve in galactitol metabolism. A mutation in *gatR* in the *E.coli* K12 strain implies constitutive expression of *gatABCDYZ* [8]. This is in fact a negative regulation effect of *gatR* on these genes.

- **yhiE:** *yhiE* is a hypothetical protein, and currently there is no literature describing the functions of this gene with experimental support. The target genes that are commonly predicted by these 5 algorithms are: *gadA, gadB, hdeA, hdeB, hdeD, slp, and yhiD*. However, by comparing the functions of the genes targeted by *yhiE* annotated in NCBI, it might be that *yhiE* involves functions of regulating acid-resistance (*hdeA-D*) and/or glutamate decarboxylase (*gadA-B*).

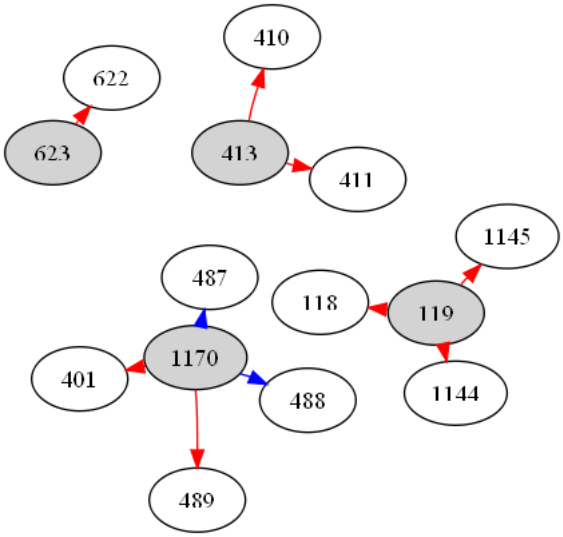


Figure 3.6: 80% precise network for CLR algorithm. Red lines are correctly inferred edges and blue lines are false positives. Grey nodes are TFs.

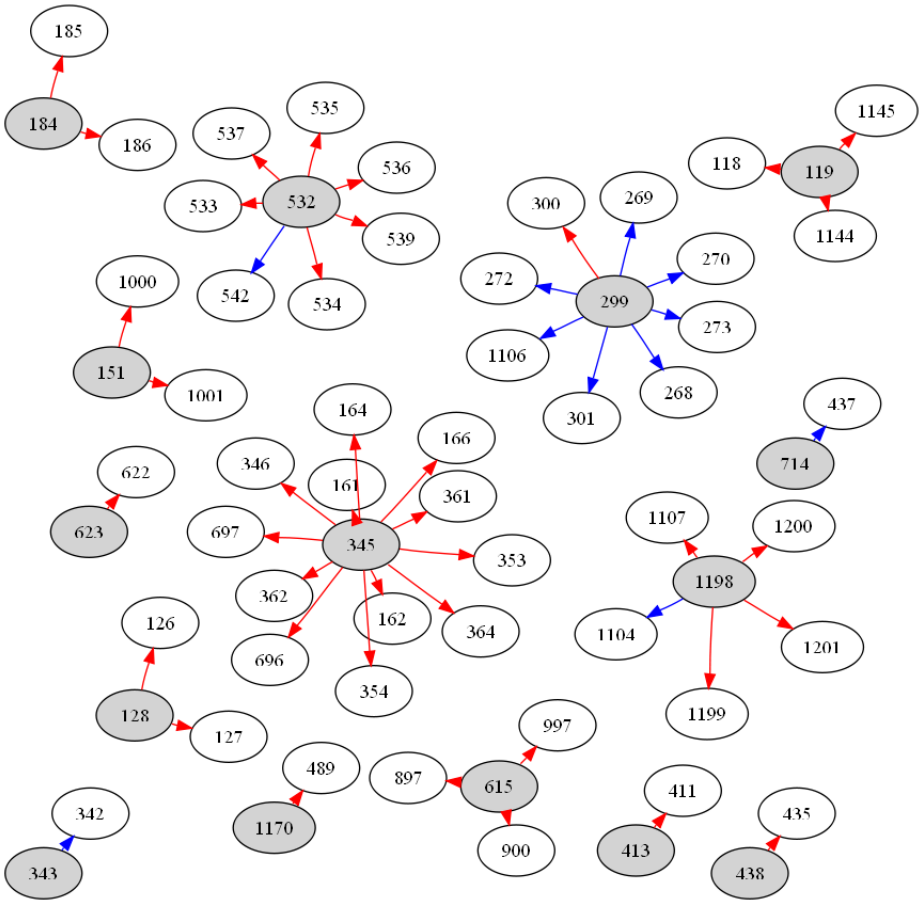


Figure 3.7: 80% precise network for L1 penalty from MM3 algorithm. Red lines are correctly inferred edges and blue lines are false positives. Grey nodes are TFs.

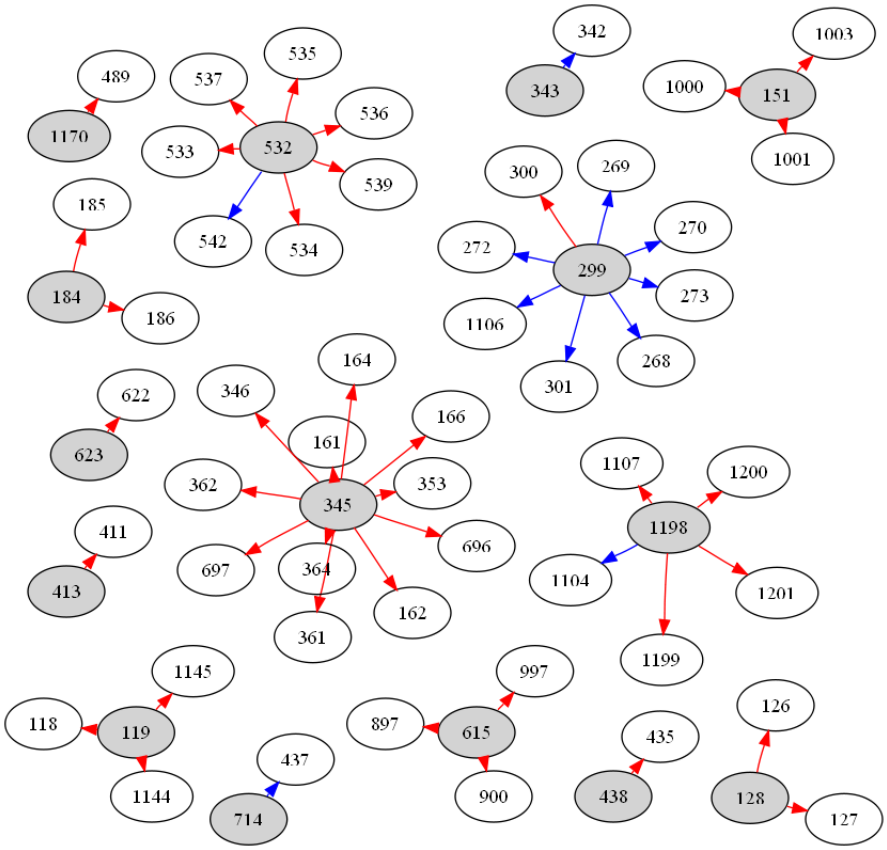


Figure 3.8: 80% precise network for SCAD penalty from MM3 algorithm. Red lines are correctly inferred edges and blue lines are false positives. Grey nodes are TFs.

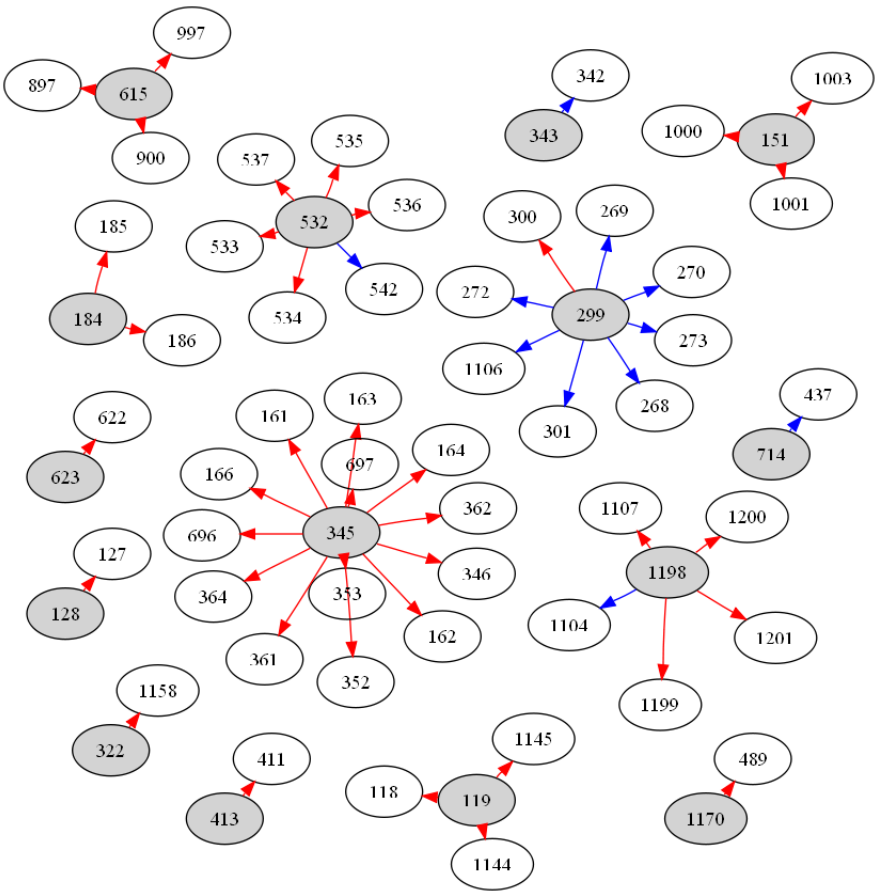


Figure 3.9: 80% precise network for LASSO penalty from EM algorithm. Red lines are correctly inferred edges and blue lines are false positives. Grey nodes are TFs.

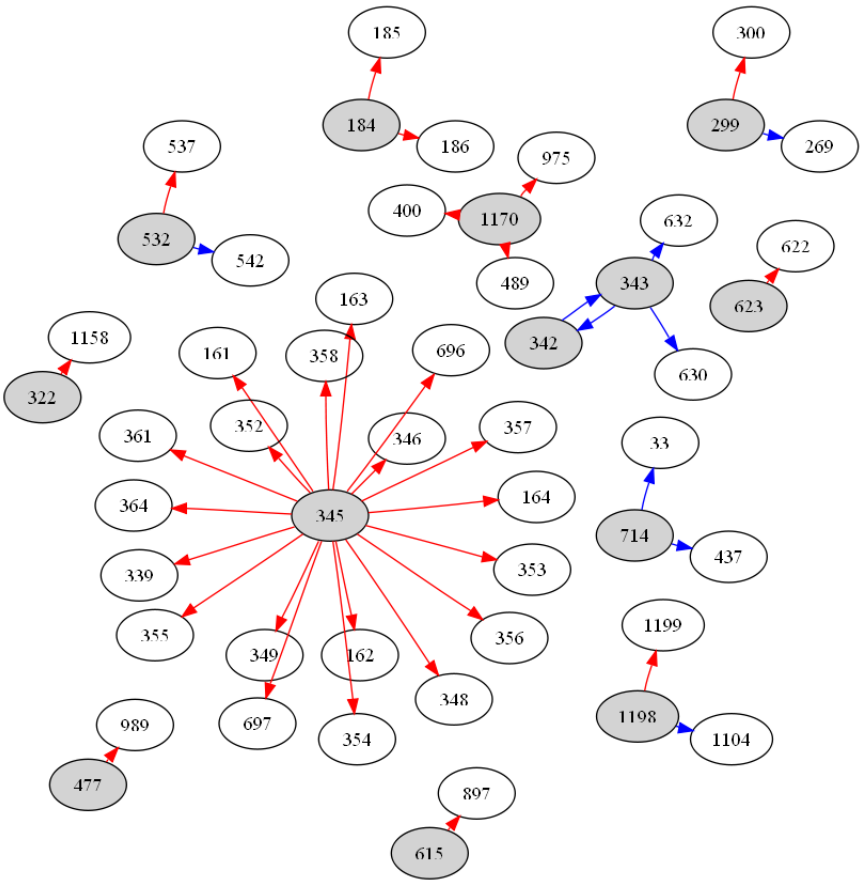


Figure 3.10: 80% precise network for NIG penalty from EM algorithm. Red lines are correctly inferred edges and blue lines are false positives. Grey nodes are TFs.

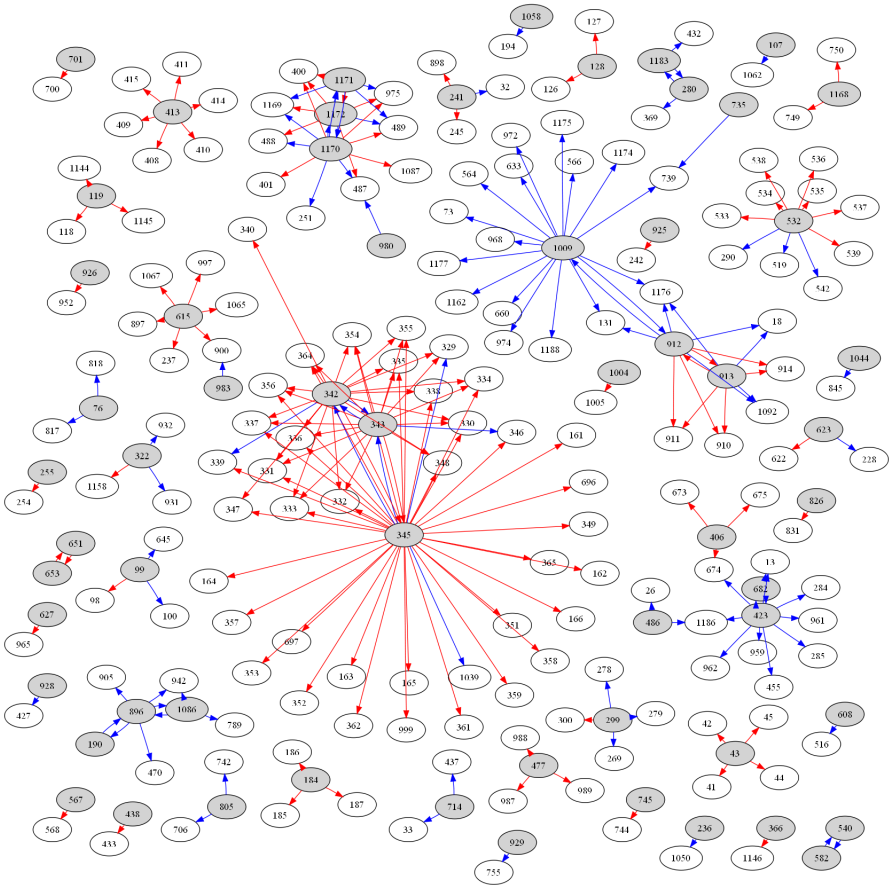


Figure 3.11: 60% precise network for CLR algorithm. Red lines are correctly inferred edges and blue lines are false positives. Grey nodes are TFs.

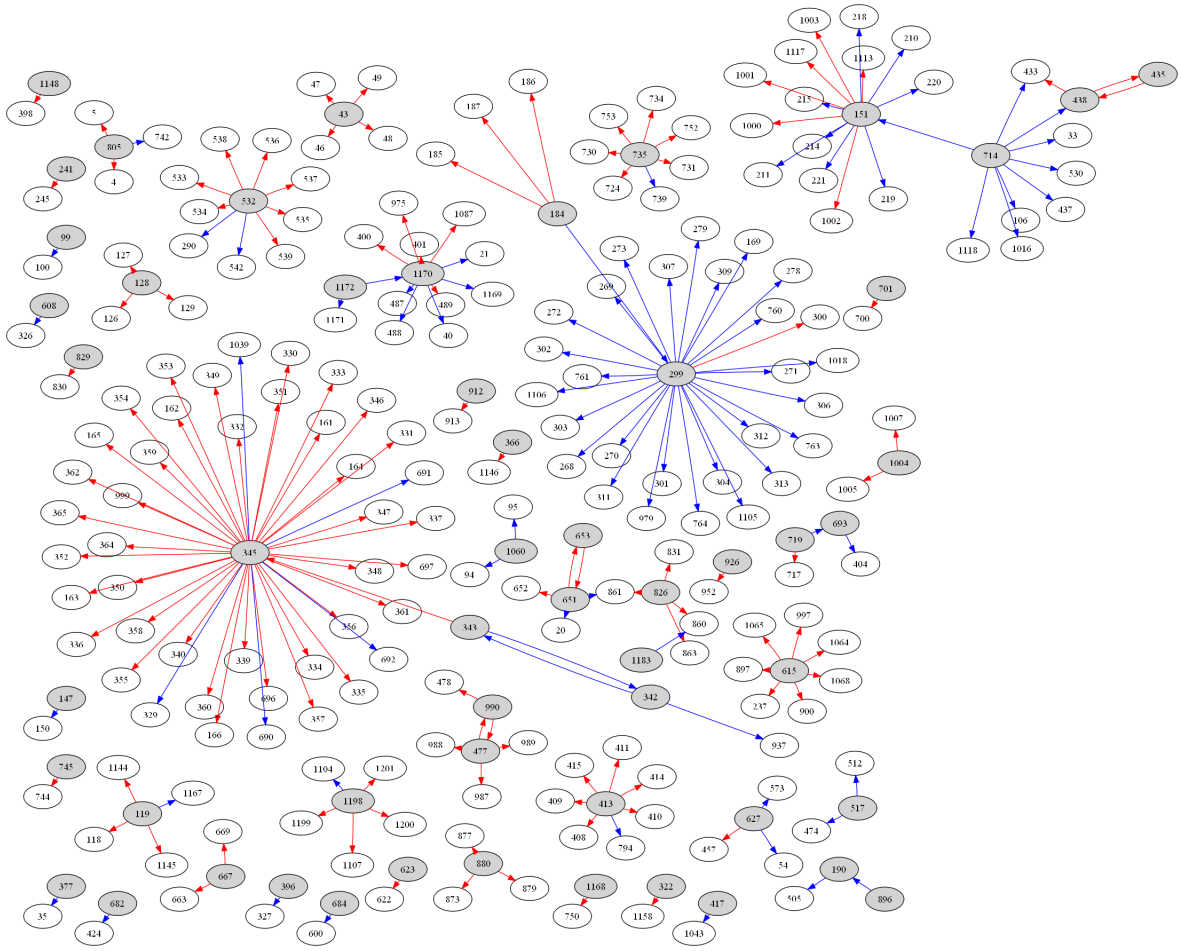


Figure 3.12: 60% precise network for L1 penalty from MM3 algorithm. Red lines are correctly inferred edges and blue lines are false positives. Grey nodes are TFs.

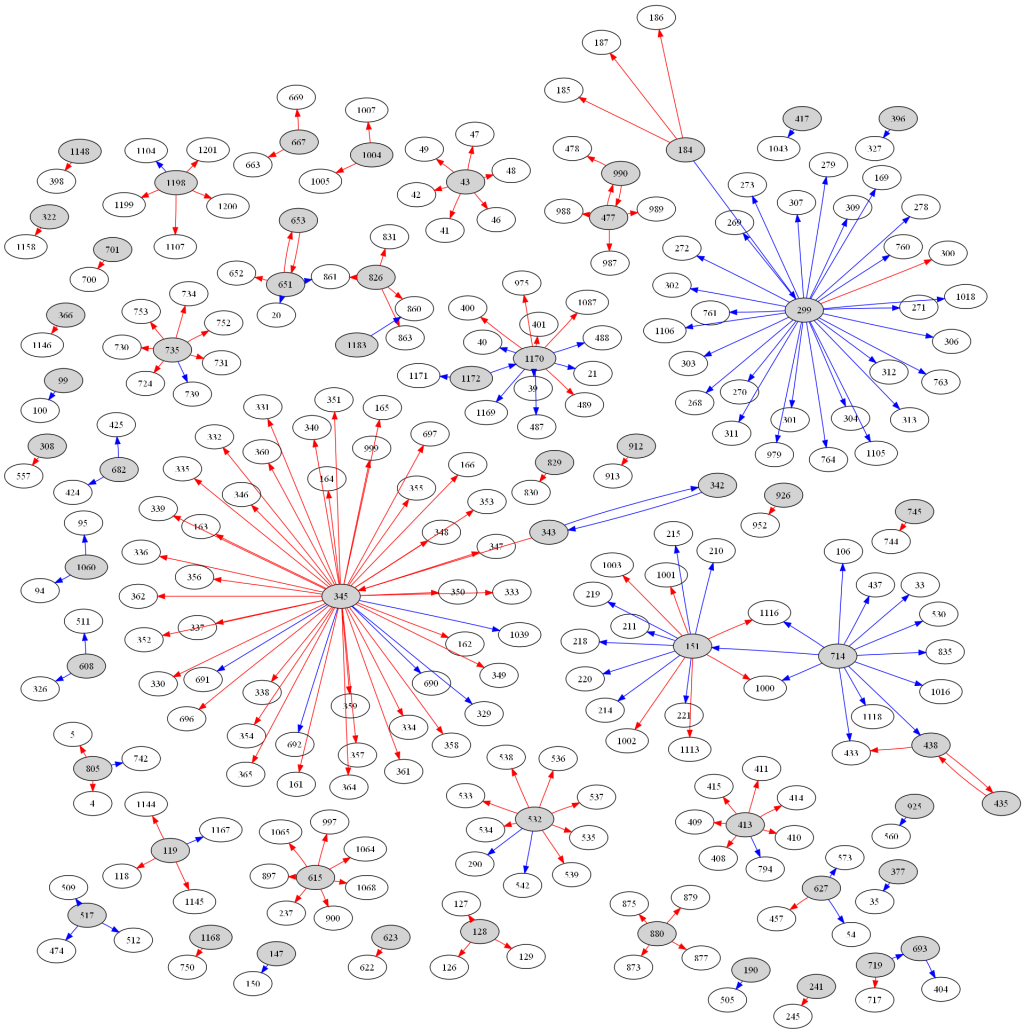


Figure 3.13: 60% precise network for SCAD penalty from MM3 algorithm. Red lines are correctly inferred edges and blue lines are false positives. Grey nodes are TFs.

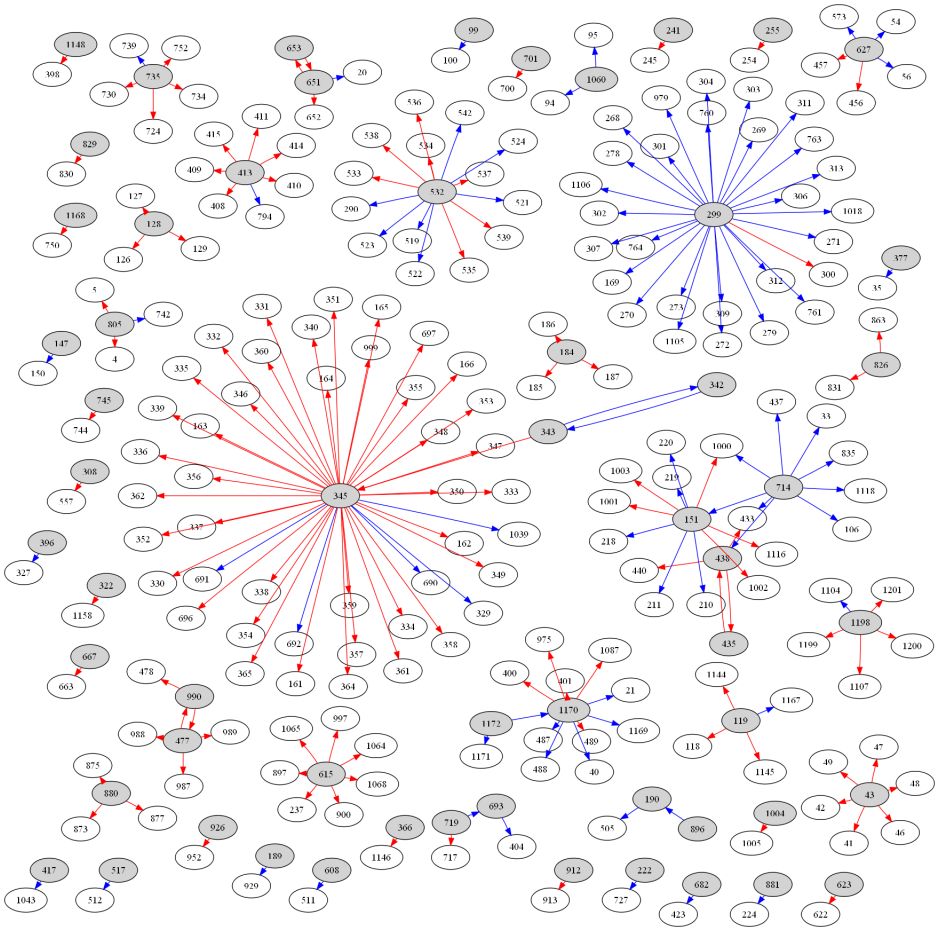


Figure 3.14: 60% precise network for LASSO penalty from EM algorithm. Red lines are correctly inferred edges and blue lines are false positives. Grey nodes are TFs.

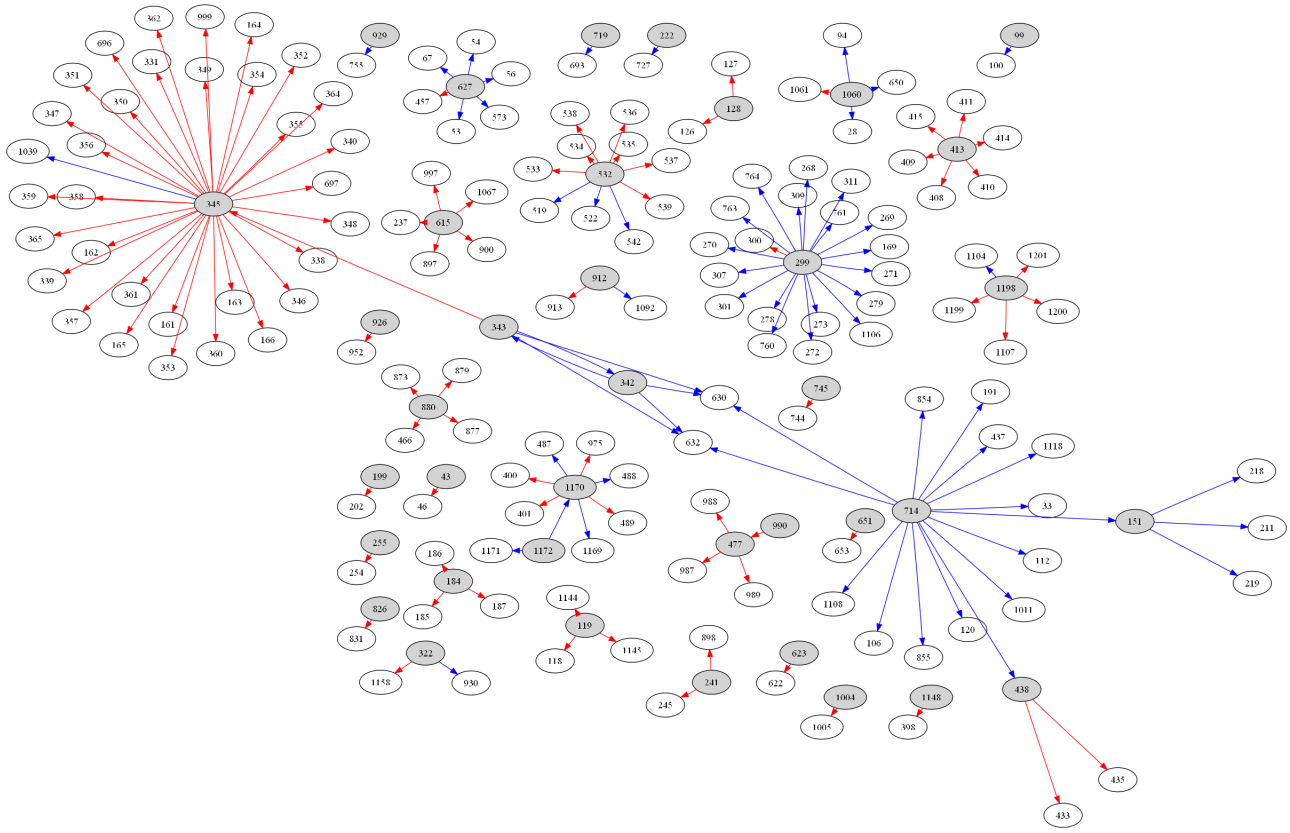


Figure 3.15: 60% precise network for NIG penalty from EM algorithm. Red lines are correctly inferred edges and blue lines are false positives. Grey nodes are TFs.

Chapter 4

Comparing models coupled with prior structural information

4.1 NCI-60 Data preparation

NCI-60 is a 60 human cancer cell line data set by Developmental Therapeutics Program of the National Cancer Institute to screen more than 100,000 chemical compounds since 1990, includeing leukemias, melanomas and cancers of ovarian etc [10, 53]. Until now, the NCI-60 cell lines have been characterized more extensively than any other set of cells in existence [53]. The purpose of our study for NCI-60 data set will be focused on the p53 hub, which is a central TF controlling the development of a variety of cancers in mammals. Hence, the ways we collected the data are including the set of genes which are targeted by p53 and potentially higher-level TFs regulating p53. Only transcript consistent re-mapped probe sets were used, i.e. probe sets that consistently match the same sets of transcripts. Further, as mentioned in the introduction to span the network across different molecular levels, we also collect a set of expression data of miRNAs in NCI-60. And we only consider the regulatory direction of microRNA→gene (including TF). Finally, to incorporate the information from other sources (e.g. from sequence comparative analysis), binary structural data, i.e. TF→gene and miRNA→gene, are also collected, respectively.

On one hand for TF→gene, the selection was made initially of targets from a ChIP-PET (Chromatin ImmunoPrecipitation and sequencing

of paired-end di-tags) experiment in colon cancer cells (treated 6hrs with 5-fluorouracil that elicits a p53 response), the most likely targets were selected from Table S4 in [59] and also from data deposited at UCSC genome browser table *GIS ChIP-PET - Genome Institute of Singapore ChIP-PET*. Potential targets were limited to genes downstream of ChIP binding sites. These tables select the most probable mappings and also looks for the presence of p53-TFBS. More targets are added from Ingenuity knowledge base [2] (both transcription activation and repression) and from known targets in the literature referred to [59]. Hence, the TFs chosen as potential targets of TP53 or regulators of TP53 are from three sources: collected from ChIP data, ingenuity and p53 knowledge base. On the other hand for miRNA→gene, the potential targets selection was made from PITA version 5 database [1]. The prediction of PITA based on two levels. The stringent one is to choose top selections based on high *conservation score* across species (giving 2.24% positive interactions), which takes into account the structure around the *seed* sites. The other selection is much more noisy, which includes all predictions without filtering (giving 23.45% positive interactions). For our analysis, only the *top* predictions were used where conservation is taken into account.

In summary, we have four input data sets:

- X: a 16143×59 matrix containing mRNA expression levels for 16143 genes across 59 experiments
- Z: a 278×59 matrix containing miRNA expression levels for 278 miRNAs across 59 experiments
- D: a 16143×183 binary matrix containing TF→gene prior structural information based sequence motif analysis
- C: a 16143×278 binary matrix containing miRNA→gene prior structural information based sequence motif analysis

Note that there is one patient removed from the study because the measured expression values are constant over all treatments.

To the same end to reduce computation as in [31], we fit various models within the positive interactions in prior binary data and select interactions

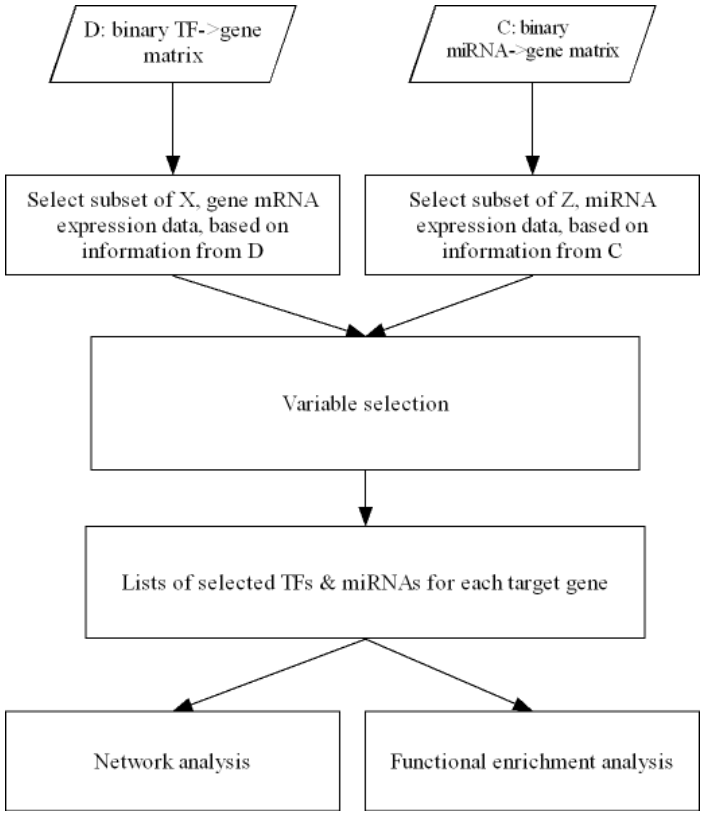


Figure 4.1: Flowchart of learning regulatory network on NCI-60 data set.

as predicted positives with supports from both prior structural information and the resulting sparse network.

$$\{(i,j) : LD_{ij} = 1 \cap D_{ij} = 1\}$$
$$\{(k,l) : LC_{kl} = 1 \cap C_{kl} = 1\}$$

where LD and LC is the learned structures for TF→gene and miRNA→gene by variable selection procedures, respectively.

The whole process is illustrated in Figure 4.1.

Remark: In current pipeline, we do not compare with the variable selection procedure without prior for every gene. We think of the prior structure information have taken account into most biological relevant facts and be-

sides this information, there are also noises in the prior information, i.e. we assume that the probability of analogous type I error is small and under the positives in prior, use variable selection to denoise the irrelevances that are not supported by our expression data (remember we subset the learned network from prior network). Thus, the methods with prior information should be better than those without prior, i.e. if the model can detect the signals contained in the prior, then the model should be also able to and easier to detect the same interaction in smaller range. Three specific examples are taken to compare between with and without prior information in the next section to justify this, see Table 4.2.

4.2 Results

4.2.1 Prediction of p53 related microRNAs and genes

We took several specific examples that of particular interests and related to p53 (TP53) regulation [55]. Some of them have biological support and the others are predicted from bioinformatics algorithms with high conservation scores (e.g. PITA) and suspected to have some biological functions that are interesting to us.

1. **TP53→GAS1.** Growth arrest-specific 1 (GAS1) plays a role in growth suppression. GAS1 blocks entry to S phase and prevents cycling of normal and transformed cells. GAS1 is a putative tumor suppressor gene and TP53 as a protein has a domain important for the activity of GAS1 as a suppressor of the cell cycle, i.e. an anti-proliferative function [49].

2. **hsa-miR-34→GAS1.** hsa-miR-34, conjectured to combine with p53, coregulates the transcription activities of GAS1 gene. hsa-miR-34→GAS1 is predicted by PITA taking into consideration the structure around the *seed* sites, and one of the top predictions where conservation is taken into account. In our analysis, we took two members of its subfamily, i.e. hsa-miR-34a and hsa-miR-34c.

3. **ETS1→TP53 and TP53→ETS1.** There are some evidences that TP53 and ETS1 interact each other at protein level, regulating downstream target genes [38]. Two members of the ETS family of transcription factors, ETS1 and ETS2, have been shown to bind to a palindromic ETS binding site and were able to trans-activate a heterologous promoter containing the binding site [13]. However, we are also interested in the direction TP53→ETS1 because nearly all of our models claimed this interaction (see below).

4. **TP53→RB1.** p53 is known to suppress RB1 transcription through inhibition of the basal promoter activity [63]. Additionally, post-transcriptionally, p53 might suppress RB1 further by inducing mir-106a, a known suppressor of RB1 [55].

These predicted interactions are summarized in Table 4.1. Here, we do not include the CLR because the thresholding is not easy to be determined and justified without ground truth interactions known. By rows, we can see that L1 and SCAD models can detect most of the interactions. NG and NJ models tend to produce sparser regulations. Although NIG can identify most non-zero elements, but the effects of these coefficients are very small and negligible. In fact, we can threshold out these small coefficients and read them as *effective zeros*. Hence, NIG is similar to the NG and NJ models. Moreover, the fitted models tend to have MSEs for L1 and SCAD which are uniformly smaller over NG, NJ, and NIG models. This is not surprising as the L1 and SCAD are less sparser than the others. By columns, we view the results for each specific interaction. For TP53→GAS1 and TP53→ETS1, it seems most of the models can correctly identify the interaction (there are some exceptions such as NG MM3, NG EM, and L1 MM2 etc). For hsa-miR-34→GAS1 (including two family members), L1 MM3, L1 EM and SCAD MM3 select it as positive interaction, and the regulatory strength is negative due to the degradation nature of miRNAs. For TP53→RB1 is an interesting example, which can be correctly identified in some models with any solution (L1 and SCAD), but in other models none of solutions do the right job (NIG, NG and NJ). As we mentioned above, p53 is known to have

Table 4.1: MSE and estimated coefficients for the specific interactions predicted from NCI-60 data set. Bold numbers represent the estimated interactions agreeing with the literatures.

Interaction	TP53→GAS1		hsa-miR-34a→GAS1		hsa-miR-34c→GAS1	
	MSE	Coefficient	MSE	Coefficient	MSE	Coefficient
L1 MM3	.16	**.23**	.16	**-.18**	.16	**-.49**
L1 MM2	.18	0	.18	0	.18	0
L1 EM	.16	**.23**	.16	**-.18**	.16	**-.49**
NIG MM3	.18	0	.18	0	.18	0
NIG MM2	.18	1E-6	.18	-2E-6	.18	-1E-6
NIG EM	.18	1E-6	.18	-2E-6	.18	-2E-6
NG MM3	.18	0	.18	0	.18	0
NG MM2	.17	**.17**	.17	0	.17	0
NG EM	.18	0	.18	0	.18	0
NJ MM3	.17	**.21**	.17	0	.17	0
NJ MM2	.17	**.17**	.17	0	.17	0
NJ EM	.17	**.21**	.17	0	.17	0
SCAD MM3	.16	**.23**	.16	**-.18**	.16	**-.49**
SCAD MM2	.17	**.36**	.17	0	.17	0

Interaction	TP53→RB1		ETS1→TP53		TP53→ETS1	
	MSE	Coefficient	MSE	Coefficient	MSE	Coefficient
L1 MM3	.059	**.13**	.085	0	.16	**.37**
L1 MM2	.049	**.17**	.20	0	.28	0
L1 EM	.056	**.14**	.059	<1E-8	.16	**.37**
NIG MM3	.086	0	.18	0	.20	**.33**
NIG MM2	.086	3E-7	.15	0	.20	**.37**
NIG EM	.086	3E-7	.18	1E-8	.21	**.28**
NG MM3	.080	0	.13	0	.18	**.39**
NG MM2	.080	0	.11	0	.17	**.47**
NG EM	.086	0	.16	0	.19	**.38**
NJ MM3	.079	0	.11	0	.17	**.51**
NJ MM2	.079	0	.11	0	.17	**.47**
NJ EM	.079	0	.11	0	.17	**.51**
SCAD MM3	.056	**.14**	.058	0	.16	**.37**
SCAD MM2	.038	**.22**	1E-5	0	.097	**.90**

a negative regulatory effect on RB1 in inhibiting the basal promoter activity, but for cancer cell lines, there might be a positive effect as well, although currently there is no literature to support this claim. Finally, in the p53 knowledge base, ETS1 is documented to regulate TP53, but here none of our models detect this interaction. Surprisingly, nearly all of our models (except L1 MM2) claim there is an opposite regulatory strength TP53→ETS1. After we searched through the literature, there indeed are experimental evidence in terms of apoptosis function to support this claim. For example, in embryonic stem (ES) cells, mouse ES cells lacking ETS1 are deficient in their ability to undergo UV-induced apoptosis, similar to p53 null ES cells. Chromatin immunoprecipitations demonstrated that ETS1 was required for the formation of a stable p53-DNA complex under physiological conditions and activation of histone acetyltransferase activity. These demonstrate that ETS1 is an essential component of a UV-responsive p53 transcriptional activation complex in ES cells and suggests that ETS1 may contribute to the specificity of p53-dependent gene transactivation in distinct cellular compartments [62].

Furthermore, to see how the prior structural information can improve the prediction ability, we select GAS1 gene as the target gene and there are 463 possible regulators in our data set, including all TFs and miRNAs. The reason of not doing variable selection for each gene without prior information is mainly a computational concern as stated before. Therefore, we just selected a few examples to demonstrate the prediction improvement. The comparison result is summarized in Table 4.2. Essentially without prior information, all the three interested interactions are not selected among the 463 candidate regulators. We conjecture that there is so many variables that by including them all, it is like finding a needle in a haystack.

In the work presented above, we started from a conjectured *bona fide* interaction and compare the performance of various models. However, in most realities, we think of the problem differently: given a set of genes and possible regulatory networks, how can we infer the biological functions of those genes and networks? In the next section, we apply functional enrichment analysis to check whether or not our predictions are functionally significant.

Table 4.2: Estimated coefficients for the specific interactions predicted from NCI-60 data set, compared between with and without prior information. Coefficients with absolute values less than 10^{-8} are thresheld to 0. Bold numbers represent the estimated interactions agreeing with the literatures.

Interaction	TP53→GAS1		hsa-miR-34a→GAS1		hsa-miR-34c→GAS1	
	No prior	Prior	No prior	Prior	No prior	Prior
L1 MM3	1E-07	**.23**	0	**-.18**	-1E-08	**-.49**
L1 MM2	0	0	0	0	0	0
L1 EM	0	**.23**	0	**-.18**	0	**-.49**
NIG MM3	0	0	0	0	0	0
NIG MM2	0	1E-6	0	-2E-6	0	-1E-6
NIG EM	0	1E-6	0	-2E-6	0	-2E-6
NG MM3	0	0	0	0	0	0
NG MM2	0	**.17**	0	0	0	0
NG EM	0	0	0	0	0	0
NJ MM3	0	**.21**	0	0	0	0
NJ MM2	0	**.17**	0	0	0	0
NJ EM	0	**.21**	0	0	0	0
SCAD MM3	3E-08	**.23**	0	**-.18**	0	**-.49**
SCAD MM2	0	**.36**	0	0	0	0

4.2.2 Functional enrichment analysis

Functional over-representation analysis was performed to objectively identify biological processes potentially affected by p53 and miRNA target genes. Specifically, the learned network in previous section is cut into small subgroups. For each group, we test the significance of every possible function assumed in these sub-networks by counting the occurrence number of a function with a given gene ontology (GO) annotation. The background is set to be the domain which contains all the genes touched by the prior structure data. A significant p value ($p < 0.05$) associated with Fisher's exact test indicates that the observed percentage of the target genes with a given annotation could not likely occur by chance given the frequency of prior network with the same annotation. Note that we do not use genes covering the whole genome as the background to avoid potential underestimate, because genes filtered by prior network never have a chance to enter the model and thus will have no influence on the learned sub-network.

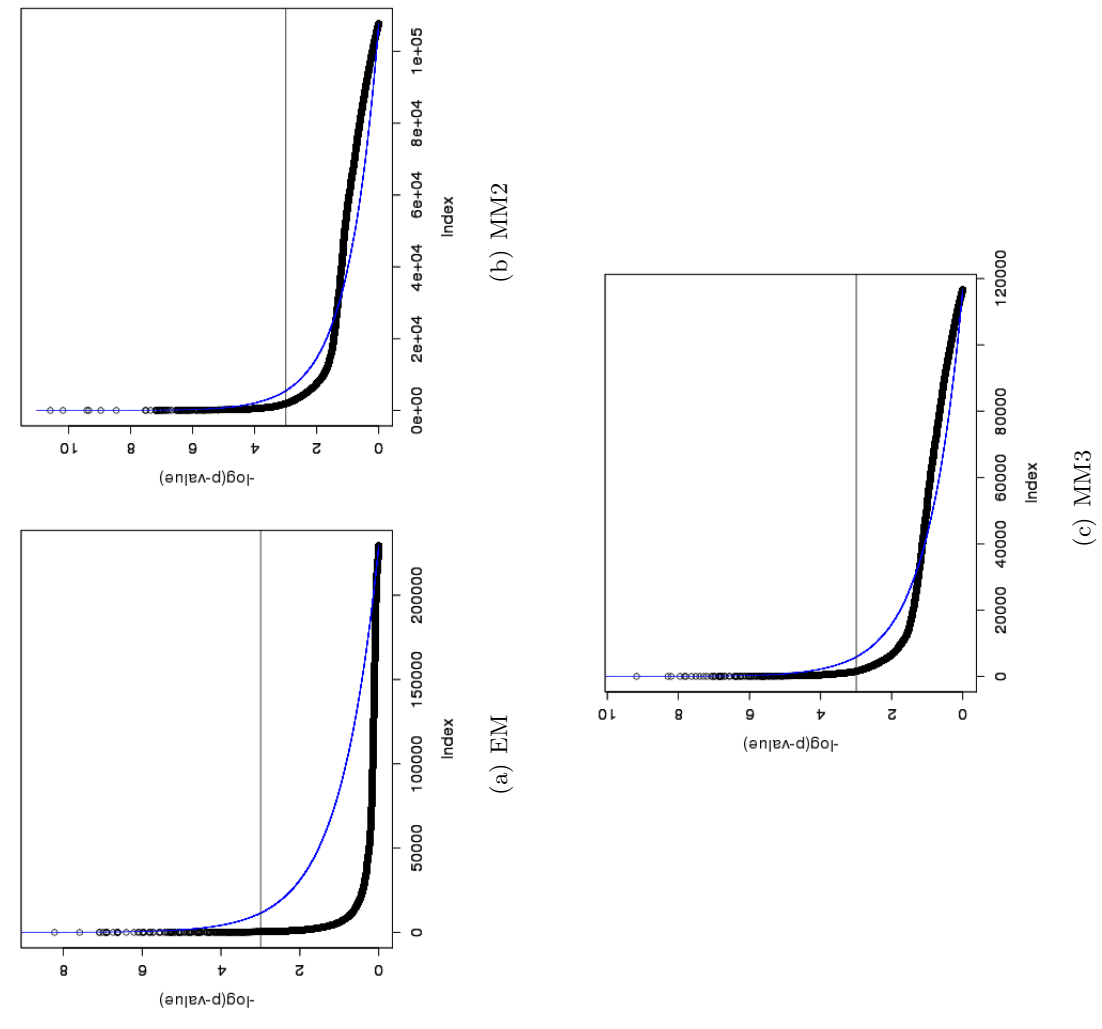
Since apoptosis and growth arrest are common known consequences of p53 activation, we tested whether p53/miRNA tend to target apoptotic (a form of programmed cell death in multicellular organisms) and cell proliferation related genes. [55] has shown that while cell cycle regulation was among the top in the p53/miRNA targets functional enrichment, apoptosis was not significant. Interestingly, this agrees with our results, i.e. in all scenarios, we found the apoptotic function annotation is not enriched by setting p-value cutoff at 0.05. Hence, we use p53/miRNA target gene functional enrichment analysis to identify specific p53/miRNA that target significant cell proliferation genes. The results for L1 model are shown in Figure 4.2 and Table 4.3. The enrichment analysis results for other models are presented in Appendix C Supplementary Materials section: Figure C.1, C.2, C.3, C.4, Table C.6, C.7, C.8 and C.9.

Obviously, genes with this enriched function that the EM algorithm has identified concentrate only on p53 regulator, and consequently EM can detect many genes that have cell proliferation annotation. On the contrary, MM2 and MM3 have a different landscape: they identified many regulators

Table 4.3: Identified significant p53/miRNAs that target known cell proliferation genes from L1 prior.

Method	Regulator	Total # of predicted targets	Cell proliferation genes	*p*-value	Gene names
EM	TP53	1167	43	5.152E-3	CDK4,PTCH1,PTEN,TGFBI,PAX3,IL1A, IL1B,ADRA1D,GAS6,PRL,BUB1B,CDK6, CTF1,IGFBP4,CYR61,AREG,EGF,FGF2, FGF7,FTH1,MDM4,MKI67,PIM1,PRKD1,MAP2K1, RAF1,STIL,TGFA,TXN,CUL1,BUB1,EPS8,PLK1, E2F1,IFI16,REST,TPX2,DLG7,PDGFC,OSM,E2F8, ZEB1,TNFSF13B
MM2	ACADVL	11	2	3.571E-2	CDK4,TYR
	hsa-miR-125a	68	3	4.898E-2	MAP3K11,MAPRE2,PES1
	hsa-miR-155	108	4	4.350E-2	PTEN,FGF2,SPOCK1,BCAT1
	PAX5	30	6	3.910E-2	IL7,FLT3LG,FGF17,CDKN1B,FLT3,HES1
	T	9	2	3.061E-2	LIF,SUZ12
	hsa-miR-27b	124	3	4.437E-2	DAZAP2,SUZ12,PLEKHK1
	hsa-miR-296	18	2	4.496E-2	NR2E1,IGF2BP1
	hsa-miR-302b	161	6	4.960E-2	IGF1,CCND2,LIF,DAZAP2,SUZ12,IGF2BP1
	hsa-miR-372	125	4	2.159E-2	NR2E1,CUL3,DAZAP2,SUZ12
	ACADVL	5	2	4.247E-2	FTH1,HMOX1
	PSMC3	6	2	4.455E-2	RB1,MDM4
	hsa-miR-202	111	4	3.706E-2	PTEN,NDN,ST13,GPC3
	hsa-miR-429	198	5	1.686E-2	PTEN,NDN,TIMP2,TOB1,HDAC4
	hsa-miR-507	60	3	3.703E-2	BTG1,HDAC4,MNT
	hsa-miR-519b	88	5	1.280E-2	PTEN,HDAC4,TSG101,ADAMTS1,STK38
MM3	ACADVL	15	3	2.806E-2	CDK4,CDK6,E2F1
	hsa-miR-125a	88	4	1.030E-2	RPS27,MAP3K11,MAPRE2,PES1
	hsa-miR-135a	180	6	4.151E-2	DAB2,MNAT1,CDK5R1,EVI5,PIM2,MAPRE2
	hsa-miR-181a	111	7	2.427E-2	TGFBI,CKS1B,GNAI2,IRS2,UCHL1,BHLHB3,ZAK
	hsa-miR-202*	66	4	1.229E-2	PTEN,EPS15,FGF2,CUL5
	hsa-miR-221	145	6	3.813E-3	RPS4X,RPS27,CYR61,GNAI2,FRAT2,PDGFC
	hsa-miR-27b	159	6	4.075E-3	RPS4X,EPS15,PRKD1,STIL,TRIB1,BHLHB3
	hsa-miR-410	128	8	1.257E-3	COL4A3,RPS4X,RPS27,FGF2,PRKD1,IRS2,EPS8,HDGFRP3
	ACADVL	26	5	3.149E-2	BCL2,TGFB1,CXCL10,CCND2,CDKN1B
	PAX5	34	7	9.723E-3	IL7,FLT3LG,FGF17,CDKN1B,FLT3,HES1,STAT5B
	hsa-miR-191	98	3	4.904E-2	IGF1R,HES1,PBEF1
	hsa-miR-206	128	3	4.772E-2	IRS2,FOXA1,PBEF1
	hsa-miR-296	15	2	3.086E-2	NR2E1,IGF2BP1
	hsa-miR-378	36	3	1.244E-2	KLF5,CCND2,NTN1
	hsa-miR-524*	251	6	3.570E-2	FGA,IRS2,HES1,SUZ12,HOXC10,TNS3
	hsa-miR-96	133	3	4.808E-2	PBEF1,MAB21L2,STAT5B
	SP1	41	5	1.151E-2	TGFB1,PTGS2,EREG,HDAC4,ETS1
	HDAC2	20	3	3.331E-2	IL1B,TNF,TGFB1
	IRF1	52	8	2.657E-2	IL1B,TNF,TGFB1,PTGS2,EIF2AK2,JAK2,OSM,IL1RN
	MYCN	90	5	1.023E-2	GLI3,RB1,TGFB1,TIMP2,LEPRE1
	hsa-miR-186	383	8	2.285E-2	PTEN,HOXB2,SKAP2,SMAD4,TOB1,HDAC4,CHERP,TOB2
	hsa-miR-195	60	4	3.754E-2	PPM1D,CHERP,SESN1,PDS5B
	hsa-miR-202	99	4	2.415E-2	PTEN,NDN,ST13,DLGAP2
	hsa-miR-202*	66	4	3.020E-2	PTEN,FGF2,CUL5,SMAD4
	hsa-miR-31	135	7	9.240E-3	TIMP2,SKAP2,JAK2,HDAC4,STK38,PDS5B,ETS1
	hsa-miR-34c	116	3	3.342E-2	BTG1,NOTCH2,ETS1
	hsa-miR-429	211	5	2.239E-2	NDN,TIMP2,TOB1,HDAC4,SESN1
	hsa-miR-495	166	5	3.464E-2	PTEN,SSTR1,NDN,PAWR,ETS1
	hsa-miR-499	153	4	3.169E-2	CUL5,CUL1,PPM1D,FOXO4
	hsa-miR-518e	53	3	4.911E-2	BMP2,MED17,PDS5B
	hsa-miR-7	115	5	1.376E-2	PPM1D,CNOT8,SETDB1,ETS1,GJB6

including TFs and miRNAs targeting some genes which have cell proliferation annotation, but for each regulator the total number of genes which was known to be cell proliferation is much smaller than that of EM. However, this case only happens with the L1 prior model and with the NIG model, only MM3 gets much more enriched regulators, see Table C.7. Again, comparing among priors, we found that results of NJ model for the three algorithms are essentially the same, which agrees with the simulation studies. Most of the models and algorithms have p53 regulator enriched, which makes sense since the cell proliferation is an important function of p53 for suppressing tumor genesis.



(a) EM (b) MM2 (c) MM3

Figure 4.2: Fisher's exact test p-values with the L1 prior. The blue line is the simulated uniform background p-values under null hypothesis and horizontal line is the threshold at p-value at 0.05.

Figure 4.2 shows the negative logarithm of p-values in a descent order. The larger the log(p-value), the more significant of corresponding enriched function in the subnetwork. Clearly in L1 case, the p-values are skewed and not uniformly distributed, which means the learned associations are unlikely to happen by random chance. EM for L1 has many p-values shifted to the non-significance level, and this again agrees with Table 4.3, which reflects that L1 with EM is concentrated on identifying a few large sub-networks. For MM2 and MM3, the differences in the tails are not significant as EM. For other priors in general, the different shapes of log(p-value) the plots confirm the different landscapes listed in tables. For example, EM and MM2 are similar to each other with the NIG prior in terms of both the plots and enriched genes (see Figure C.2 and Table C.7), while MM3 significantly differs from those two.

Note that we do not perform a multiple comparison correction procedure here. The reason is that we test the significance of each possible function by conditioning on its local sub-network and its annotation, this will introduce very complicated dependency structures among different testings. Thus, the independence assumption of multiple comparison fails here and it is very difficult to control the false positive error rate. For this reason no multiple testing is taken into account, the significance of the p-values should be taken with a grain of salt.

Although we correctly inferred the TP53→GAS1 regulatory relationship in most models, this interaction is not enriched in over-representation analysis here.

On the other hand, by comparing to miRNAs which target cell proliferation genes and were captured by [55], we predicted several regulatory interactions that have been reported in the literature as differentially regulated in various cancerous tissues or cancer cell lines. For example, CKS1B, IRS2, and TGFBI, target genes of miR-181a, overlap the results found in [55]. We did a literature search and there are indeed very recent evidences to support the predicted cell proliferation function [45, 61, 66]. Speficially, overexpression of CKS1B is linked to a poor prognosis in multiple myeloma and contributes to increased p27Kip1 turnover, cell proliferation, and a poor

prognosis in many tumor types [66]. [66] showed that CKS1B influences myeloma cell growth and survival through SKP2- and p27Kip1-dependent and -independent mechanisms and that therapeutic strategies aimed at abolishing CKS1B function, which in our case is the possible degradation through miR-181a regulation, may hold promise for the treatment of high-risk disease for which effective therapies are currently lacking. Very recently, The protein levels of IRS2, insulin receptor substrate-2, has been showed that reduced IRS2 levels in the islets by high dosage chlorpromazine contributes to apoptosis of pancreatic b-cells and decreased proliferation [45]. TGFBI, transforming growth factor β-induced, can reduce cell proliferation or be down-regulated in certain type of cancers, including bladder cancer [61].

Another example addressed in [55] is miR-195, which was also reported by our model. This miR-195 is important in that it targets genes that may function as both apoptosis and cell proliferation and these two functions are ranked very significantly from their model. Our models, for example L1, have also predicted this miR-195 and part of its target genes are overlapped with [55], i.e. PPMID and SESN1 genes. A recent study also implicated miR-195, along with miR-23a and miR-24, in cardiac hypertrophy and reported that these miRNAs were regulated in response to stress signaling in the heart [57].

GAS6, a gene targeted by p53 and enriched from L1 prior with EM algorithm in Table 4.3, like GAS1 is a member of the growth arrest sequences (GAS) family genes (known modulators of cell cycle progression and survival). Serial analysis of gene expression from aggressive mammary tumors derived from transplantable p53 null mouse mammary outgrowth lines revealed significant up-regulation of GAS6 transcripts. The minimal region of amplification contained genes CUL4a, LAMP1, TFDP1, and GAS6, highly overexpressed in the p53 null mammary outgrowth lines at preneoplastic stages, and in all its derived tumors [3].

As we have shown many examples above, combining the variable selection and over-representation analysis can improve the quality of predicted functional interactions, and thus reduce the number of interactions without support from biological interpretations. We presented examples with pre-

dicted interactions without enriched functions, although these predictions have some direct/indirect biological supports. On the other hand, there are also many predicted interactions with enriched annotations, while those interactions are not found in the literature.

Besides, comparing across various models can further enhance and provide clues for biological analysis. For example, using EM and sometimes MM2 (e.g. NIG prior) tend to focus on fewer regulators with a large population of target genes, while using MM2 and MM3 can detect more enriched regulators (notably most of them are miRNAs).

Chapter 5

Conclusions and discussions

In this thesis, our ultimate goal is integrating multiple data sources and using statistical methods to construct regulatory networks spanning different molecular levels. The regulatory interactions are modeled as multivariate linear regression models. However, linear model applied to high-dimensional data are usually over-parametrized, unclear for interpretation, and have high variance for estimated parameters. Motivated from this fact, we formulate the problem as fitting a sparse model to reduce its dimensionality. To achieve this, we used penalized regression models and look for the corresponding penalized MLEs (pMLEs). Because the penalization can be alternatively viewed as prior regularization, different functional forms of penalties can be mapped to priors on parameters and the pMLE is also the corresponding MAP estimator. Based on this interpretation, various scaled mixture Gaussian prior distributions are proposed such that the marginal densities of coefficients/regulatory strengths can be given in closed form. The practical question is to determine the MAP estimation. Currently, there exists several optimization approaches. For example, EM and MM algorithm. Applying these two methods to a variable selection prior which is unbounded at origin could possibly entail the local optimum problem (although this case rarely happens). To address this issue, we propose a variant MM algorithm to avoid this. The following simulation studies were performed to compare the performance of various models with the three algorithms. Then, these methods were applied to an *E.coli* data set with ground truth interactions believed to be known. The performances are measured in terms of precision-recall curve. The results showed that L1 MM3, SCAD MM3, L1 EM, and NIG EM methods are performing best among all models, and their performances are comparable to the state-of-the-art regulatory net-

work construction algorithm CLR. This shows that a simple linear regression network with a properly chosen prior can compete with the current best method. Further, five *hub* genes were selected with common predictions from high performance algorithms and we showed that at functional level, the learned networks agree with current literature. Finally, to arrive at our ultimate goal, we coupled our models with collected prior structural information from various databases. A sub-network of the prior network was learned for each model, and we took several important interactions (possibly conjectured) to demonstrate the detecting ability of our models. To identify predicted interactions which are significantly enriched with cell proliferation function annotation, we subsequently did an over-representation analysis of our prediction results. Several interactions were taken as examples to show that some of predictions indeed exert cell proliferation function in abnormal cell lines, as confirmed by literature and agreeing with the other studies.

On the variable selection side, beyond the models and methods used in this project there are many other ways to estimate both a compact submodel and parameters. For parameter estimation post-model-selection, see [47]. Recently, there are also heavy literature volumes concerning simultaneously do model selection and parameter estimation [11, 26, 42, 56, 60, 68], both in frequencist and Bayesian contexts.

The combinatorial modeling was discussed at the beginning of the thesis, and we did not implement this combined effects in this project. This is a computational concern. Grouped variable selection may be applied to reduce the optimization burden, but we have no extra time to extend on this topic. It is certainly a future topic worthy to study the combined effects of regulators.

To conclude, by comparing different statistical models and integrating multiple data information can improve the reliability and facilitate interpretation of constructed regulatory networks, in terms of prediction variability and biological functions.

Bibliography

[1] `http://genie.weizmann.ac.il/pubs/mir07/mir07_data.html`.

[2] Ingenuity systems. 2007.

[3] Martin C. Abba, Victoria T. Fabris1, Yuhui Hu, Frances S. Kittrell, Wei-Wen Cai, Lawrence A. Donehower, Aysegul Sahin, Daniel Medina, and C. Marcelo Aldaz. Identification of novel amplification gene targets in mouse and human breast cancer at a syntenic cluster mapping to mouse ch8a1 and human ch13q34. *Cancer Res.*, 67(9):4104–12, 2007.

[4] M. Abramowitz and I. Stegun. *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*. Dover Publications, tenth edition, 1972.

[5] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–23, 1974.

[6] A Antoniadis. Wavelets in statistics: a reivew (with discussion). *J. Italian Statistical Society*, 6:97–144, 1997.

[7] K. Azra and et al. Combinatorial microrna target predictions. *Nat. Genetics*, 37:495–500, 2005.

[8] Nobelmann B and Lengeler JW. Molecular analysis of the gat genes from *Escherichia coli* and of their roles in galactitol transport and metabolism. *J Bacteriol.*, 178(23):6790–5, 1996.

[9] P. Billingsley. *Probability and Measure*. Jone Wiley & Sons, Inc., second edition, 1986.

[10] Paul E. Blower and et al. Microrna expression profiles for the nci-60 cancer cell panel. *Mol Cancer Ther*, 6(5), 2007.

[11] P. Brown, M. Vannucci, and T. Fearn. Multivariate bayesian variable selection and prediction. *J. Royal. Statist. Soc B.*, 60:627–641, 1998.

[12] F. Caron and A. Doucet. Sparse bayesian nonparametric regression. *International Conference on Machine Learning(ICML), Accepted*, 2008.

[13] Reisman D and Loging WT. Transcriptional regulation of the p53 tumor suppressor gene. *Semin Cancer Biol.*, 8:317–324, 1998.

[14] A. Dempster. Covariance selection. *Biometrics*, 28(1):157–175, 1972.

[15] A. Dempster, N. Laird, and D. Robin. Maximum likelihood from incomplete data via the em algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B*, 39:1–38, 1977.

[16] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.

[17] Gao. F, Foat. B, and Bussemaker. H. Defining transcriptional networks through integrative modeling of mrna expression and transcription factor binding data. *BMC Bioinformaticss*, 5(31), 2004.

[18] J. Faith and et al. Large-scale mapping and validation of *escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biol.*, 5(1):e8, 2007.

[19] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96(456):1348–1360, 2001.

[20] M. Figueiredo. Adaptive sparseness for supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25:1150–1159, 2003.

[21] Fanette Fontaine, Eric J.Stewart, Ariel B.Lindner, and François Taddei. Mutations in two global regulators lower individual mortality in *Escherichia coli*. *Mol Microbiol.*, 67(1):2–14, 2008.

[22] I. Frank and J. Friedman. A statistical view of some chemometrics regression tools (with discussion). *Technometrics*, 35:109–148, 1993.

[23] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics, to be appeared*, 2007.

[24] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using bayesian network to analyze expression data. *J. Computational Biology*, 7:601–620, 2000.

[25] W. Fu. Penalized regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7(3):397–416, 1998.

[26] E. George and R. McCulloch. Variable selection via gibbs sampling. *J. Amer. Statist. Assoc.*, 88(423):881–9, 1993.

[27] J. Griffin and P. Brown. Alternative prior distributions for variable selection with very many more variables than observations. *Technical Report, University of Kent*, 2005.

[28] J. Griffin and P. Brown. Bayesian adaptive lassos with non-convex penalization. *Technical Report, University of Kent*, 2007.

[29] G. Grimmett and D. Stirzaker. *Probability and Random Processes*. Oxford Unviersity Press, third edition, 2001.

[30] A. Hoerl and R. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(3):55–67, 1970.

[31] J. Huang, B. Morris, and J. Frey. Bayesian inference of microrna targets from sequence and expression data. *J. Comp. Biol.*, 14(14):550–563, 2007.

[32] D. Hunter and R. Li. Variable selection using mm algorithms. *Annals of Statistics*, 33:1617–1642, 2005.

[33] Hartemink. J, Gifford D, Jaakkola. T, and Young. R. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Nature Reviews Genetics*, 6:422–33, 2001.

[34] Keene. J. Rna regulons: coordination of post-transcriptional events. *Nature Reviews Genetics*, 8:533–543, 2007.

[35] B. John, A. Enright, A. Aravin, T. Tuschl, C. Sander, and et al. Human microrna targets. *PLoS Biol.*, 2(11):e363, 2004.

[36] K. Knight and W Fu. Asymptotics for lasso-type estimators. *Annals of Statistics*, 28(5):1356–78, 2000.

[37] S. Konishi and G Kitagawa. *Information Criteria and Statistical Modeling*. Springer, first edition, 2008.

[38] Sebum Lee and Hriday K. Das. Inhibition of basal activity of c-junnh2-terminal kinase (jnk) represses the expression of presenilin-1 by a p53-dependent mechanism. *Brain Res.*, 1207:19–31, 2008.

[39] B. Lewis, C. Burge, and D. Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microrna targets. *Cell*, 120:15–20, 2005.

[40] Buck. M and Lieb. J. Chip-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, 83:349C360, 2004.

[41] Eisen. M, Spellman. P, Browndagger. P, and Botstein. D. Cluster analysis and display of genome-wide expression patterns. *Nature Reviews Genetics*, 95:14863–8, 1998.

[42] J. Marin and C. Robert. *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*. Springer-Verlag, first edition, 2007.

[43] N. Meinshausen and P. Buhlmann. High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3):1436–62, 2006.

[44] M. Osborne, B. Presnell, and B. Turlach. On the lasso and its dual. *Journal of Computational and Graphical Statistics*, 9(2):319–337, 2000.

[45] Sunmin Park, Sang Mee Hong, Ji Eun Lee, So Ra Sung, and Sung Hoon Kim. Abstract chlorpromazine attenuates pancreatic β-cell function and mass through irs2 degradation, while exercise partially reverses the attenuation. *J Psychopharmacol*, 2008.

[46] D. Pe'er, A. Regev, G. Elidan, , and Friedman N. Inferring subnetworks from preturbed expression profiles. *Bioinformatics*, 17:S215–S224, 2001.

[47] B. Potscher. Effects of model selection on inference. *Econometric Theory*, 7(2):163–85, 1991.

[48] Jensen. S, Chen. G, and Stoeckert. C. Bayesian variable selection and data integration for biological regulatory networks. *Ann. Appl. Stat.*, 1(2):612–633, 2007.

[49] G Del Sal, E M Ruaro, R Utrera, C N Cole, A J Levine, and C Schneider. Gas1-induced growth suppression requires a transactivation-independent p53 function. *Mol Cell Biol*, 15(12):7152–60, 1995.

[50] H. Salgado and et al. Regulondb (version 5.0): *Escherichia coli* k-12 transcriptional regulatory network, operon organization, and growth conditions. *Nuclear Acid Res.*, 34:D394–397, 2006.

[51] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–4, 1978.

[52] R. Shalgi, D. Lieber, M. Oren, and Y. Pilpel. Global and local architecture of the mammalian microrna-transcription factor regulatory network. *PLoS Comput. Biol.*, 3(7):e131, 2007.

[53] U. Shankavaram and et.al. Transcript and protein expression profiles of the nci-60 cancer cell panel: an integromic microarray study. *Molecular Cancer Therapeutics*, 6(3):820–32, 2007.

[54] J. Shao. Linear model selection by cross-validation. *J. Amer. Statist. Assoc.*, 88:486–94, 1983.

[55] Amit U Sinha, Vivek Kaimal, Jing Chen, and Anil G Jegga. Dissecting microregulation of a master regulatory network. *BMC Genomics*, 9(88), 2008.

[56] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.*, 58:267–288, 1996.

[57] Eva van Rooij, Lillian B. Sutherland, Ning Liu, Andrew H. Williams, John McAnally, Robert D. Gerard, James A. Richardson, and Eric N. Olson. A signature pattern of stress-responsive micrornas that can evoke cardiac hypertrophy and heart failure. *Proc Natl Acad Sci U S A*, 103(48):18255–60, 2006.

[58] W. Wasserman and A. Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nat. Genetics*, 5:276–287, 2004.

[59] C. Wei, Q. Wu, and et al. A global map of p53 transcription-factor binding sites in the human genome. *Cell*, 124(1):207–19, 2006.

[60] W. Wilks, S. Richardson, and D. Spiegelhalter. *Markov Chain Monte Carlo In Practice*. London: Chapman & Hall, first edition, 1995.

[61] Wang X, Colby JK, Rengel RC, Fischer SM, Clinton SK, and Klein RD. Overexpression of cyclooxygenase-2 (cox-2) in the mouse urinary bladder induces the expression of immune- and cell proliferation-related genes. *Mol Carcinog*, 2008.

[62] Dakang Xu, Trevor J. Wilson, David Chan, Elisabetta De Luca1, Jiong Zhou1, Paul J. Hertzog1, and Ismail Kola. Ets1 is required for p53 transcriptional activity in uv-induced apoptosis in embryonic stem cells. *The EMBO Journal*, 21:4081C409, 2002.

[63] Shiio Y, Yamamoto T, and Yamaguchi N. Negative regulation of rb expression by the p53 gene product. *Proc Natl Acad Sci*, 89(12):5206–5210, 1992.

[64] Stephen Yeung, Jesper Tegnęr, and James Collins. Reverse engineering gene networks using singular value decomposition and robust regression. *PNAS*, 99(9):6163–8, 2002.

[65] A. Zellner and A. Siow. Posterior odds ratios for selected regression hypotheses. *In Bayesian Statistics: Proceedings of the First International Meeting held in Valencia (Spain), Valencia: University Press*, pages 585–603, 1980.

[66] Fenghuang Zhan, Simona Colla, Xiaosong Wu, Bangzheng Chen, James P. Stewart, W. Michael Kuehl, Bart Barlogie, and Jr John D. Shaughnessy. Cks1b, overexpressed in aggressive disease, regulates multiple myeloma growth and survival through skp2- and p27kip1-dependent and -independent mechanisms. *Blood*, 109(11):4995–5001, 2007.

[67] P. Zhao and B. Yu. On model selection consistency of lasso. *J. of Machine Learning Res.*, 7:2541–2563, 2006.

[68] H. Zhou and T. Hastie. Regularization and variable selection via the elastic net. *J. Royal. Statist. Soc B.*, 67(2):301–20, 2005.

Appendix A

Methods

A.1 Penalties

A.1.1 Information criteria

AIC and BIC are motivated from model misspecification and model selection area [37]. AIC [5] and BIC [51] corresponds to the penalty function with form $p_\lambda = 0.5\lambda^2\mathbb{I}(\beta \neq 0)$, where $\lambda = \sqrt{2}$ and $\lambda = \sqrt{\log n}$, respectively. Thus, the penalization is based on the complexity of assumed model and discontinuous w.r.t. parameter β. Moreover, it is well known that model selection procedures based on AIC are inconsistent in terms of the probability of selecting an over-parameterized model is asymptotically positive [47].

A.1.2 Hard thresholding

Let $p_\lambda = \lambda^2 - (|\beta| - \lambda)^2\mathbb{I}(|\beta < \lambda|)$. Then this is the hard thresholding penalty [6], which is a smoothed version of entropy penalization.

A.1.3 $\mathcal{L}^p$ penalty

Choosing the $\mathcal{L}^p$ penalty, i.e. $p_\lambda = \lambda|\beta|^p$, is the solution of the bridge regression [22, 25], in linear regression context. Typically, $p \in [0, 2]$. Moreover the penalized likelihood has variable selection property when $p \leq 1$ and the optimization problem of $\mathcal{O}(\cdot)$ function in Eq.(2.2) is convex only if $p \geq 1$. Particularly, $p = 0$ corresponds to traditional model selection, $p = 1$ yields LASSO [56] and $p = 2$ ridge regression [30]. Under $\mathcal{L}^0$ penalty, Shao [54] has shown that the cross-validated choice of the penalty parameter λ is consistent for model selection under certain conditions on the size of the testing data set. But this oracle property does not hold for $\mathcal{L}^1$ penalty. Under a

set of assumptions (c.f. Assumption 1-6 in [43]), however, Meinshausen and Buhlmann [43] showed that $\mathcal{L}^1$ penalty is a consistent (in probability) approximation of jointly modeling the dependency in the Gaussian graphical model context [14, 23]. Zhao and Yu [67] also showed the probability of LASSO choosing the true model goes to 1 at the exponential rate, provided the regularity conditions and *Strong Irrepresentable Condition* holds. Furthermore, Knight and Fu [36] showed that LASSO even retains the model estimation consistency and asympototic normality when $\lambda_n = o(n)$, where $f(n) = o(n)$ means $\lim_{n\to\infty} \frac{f(n)}{n} = 0$.

Based on these work, we can further extend the consistency of LASSO in a stronger sense, i.e. *almost sure (a.s.)* convergence. Before establishing the results, the definitions are presented below:

Definition A.1.1. (Model Selection Consistency)

1. In probability: $P(sgn(\hat{\beta}(n)) = sgn(\beta)) \to 1$, as $n \to \infty$, notated as $sgn(\hat{\beta}(n)) \xrightarrow{P} sgn(\beta)$.
2. Almost sure: $P(\lim_{n\to\infty} sgn(\hat{\beta}(n)) = sgn(\beta))) = 1$, notated as $sgn(\hat{\beta}(n)) \xrightarrow{a.s.} sgn(\beta)$.

Proposition A.1.1. *Fix p, under the regularity conditions as in [36]: as* $n \to \infty$

$$C_n = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T \quad \to \quad C \succ 0 \tag{A.1}$$

$$\frac{1}{n}\max_{i=1:n} x_i^T x_i \quad \to \quad 0 \tag{A.2}$$

where $C \succ 0$*, meaning* C *is a positive definite matrix. and if Strong Irrepresentable Condition holds [67], then* $sgn(\hat{\beta}_{L1}(n)) \xrightarrow{a.s.} sgn(\beta)$.

The proof is given in Appendix B Proofs, and essentially an application of Borel-Cantelli lemma ([9] p53-57). It reveals the fact that if the inconsistency decreases at a rapid rate when the number of data points tends to infinity, then there could be only a finite number of incorrected estimated models, which implies the *a.s.* convergence.

A.1.4 SCAD penalty

Let the continous differentiable penalty function has the form

$$p_\lambda^{'}(\theta) = \lambda\{\mathbb{I}(\theta \leq \lambda) + \frac{(a\lambda - \theta)_+}{\lambda(a-1)}\mathbb{I}(\theta > \lambda)\} \tag{A.3}$$

with some $a > 2$ and $\theta > 0$, then p_λ is called the Smoothly Clipped Absolute Deviation (SCAD) penalty [19]. x_+ is the positive part of x, i.e. $x = x$ if $x \geq 0$; $x = 0$ if $x < 0$. SCAD penalty is very similar to $\mathcal{L}^1$ penalty in the region where the absolute values of estimated parameters are small. But SCAD penalty is constant for large absolute values of estimated parameters to avoid over-penalization or biasness. See Figure 2.2(b).

A.1.5 Bayesian linear regression

The penalization for likelihood function has Bayes interpretation. The penalty term can be alternatively viewed as the prior regularization of the sampling model. Therefore, to maximize the penalized likelihood function is equivalent to find the mode of corresponding posterior distribution. For example, LASSO estimator is the posterior mode resulted from Gaussian likelihood coupled with double-exponential prior. Griffin and Brown [27, 28] proposed a family of Bayesian linear regression models using a scale mixture of Gaussian prior on regression coefficients, which includes Normal-Gamma (NG) and Normal-Inverse Gaussian (NIG) model. This hierarchical model family can be extended to include Zellner's g-prior [65], the parameter-free Normal-Jeffreys (NJ) prior [20], and LASSO, which is a double-exponential mixture of Gaussian as a special case of NG model. The MAP estimator can be found using EM algorithm [12, 28]. Next we briefly review this hierarchical Bayesian regression model family. In general, the marginal prior distribution of regression coefficients in this family can be factored as a scale mixture of Gaussian density:

$$\pi(\beta_j) = \int \mathcal{N}(\beta_j;\, 0,\, \tau_j^2)\, p(\tau_j^2)\, d\tau_j^2 \tag{A.4}$$

where $\mathcal{N}(x;\mu,\sigma^2)$ denotes the normal distribution of x, with mean μ and variance σ^2. Depending on different prior distribution specified for τ_j^2, we can introduce several different models. It is easy to check that LASSO, NG, and NJ priors satisfy the conditions of Theorem 2.3.1 and they are also singular at the origin [19], i.e. $|\beta_j| + p_\lambda^{'}(|\beta_j|)$ is strictly positive and attains minimum at $\beta_j = 0$.

Normal-Jeffreys prior

If choosing the Jeffreys' non-informative prior for τ_j^2, i.e. $p(\tau_j^2) \propto \frac{1}{\tau_j^2}$, then we get the NJ prior. The resulting marginal p.d.f. for β_j is

$$p(\beta_j) \propto \frac{1}{|\beta_j|} \tag{A.5}$$

Normal-Gamma prior

If choosing $\tau_j^2 \sim \mathcal{G}(\frac{\alpha}{K}, c)$, where $\mathcal{G}(x;\, a,\, b)$ is the Gamma p.d.f. with mean $\frac{a}{b}$ and variance $\frac{a}{b^2}$. Then the marginal p.d.f. for β_j is given by

$$\pi(\beta_j) = \sqrt{\frac{2}{\pi}} \frac{c^{\frac{\alpha}{K}}}{\Gamma(\frac{\alpha}{K})} \left(\frac{\beta_j^2}{2c}\right)^{\frac{\frac{\alpha}{K}-\frac{1}{2}}{2}} \mathcal{K}_{\frac{\alpha}{K}-\frac{1}{2}}(\sqrt{2c\beta_j^2}) \tag{A.6}$$

where $\mathcal{K}_v(\beta)$ is the modified Bessel function of the second kind [4]. Note that if $\frac{\alpha}{K} = 1$, the NG prior reduces to double-exponential prior, and thus yeilds LASSO model with $\lambda = \sqrt{2c}$. And NJ prior is the limiting case when $\frac{\alpha}{K} \to 0$ and $c \to 0$.

Normal-Inverse Gaussian prior

If choosing $\tau_j^2 \sim \mathcal{IG}(\frac{\alpha}{K}, c)$, where $IG(\frac{\alpha}{K}, c)$ is the Inverse-Gaussian p.d.f. Then the marginal p.d.f. for β_j is given by

$$\pi(\beta_j) = \frac{c\alpha}{\pi K} \exp\left(\frac{c\alpha}{K}\right) \left(\frac{\alpha^2}{K^2} + |\beta_j|^2\right)^{-\frac{1}{2}} \mathcal{K}_1\left(c\sqrt{\frac{\alpha^2}{K^2} + |\beta_j|^2}\right) \tag{A.7}$$

Table A.1: Penalizations/log(prior) and their first order derivatives evaluated at $|\beta_j|$. Notation: $q_j = \sqrt{(\frac{\alpha}{K})^2 + |\beta_j|^2}$

	$p_\lambda(\lvert\beta_j\rvert)$	$p_\lambda^{'}(\lvert\beta_j\rvert)$
LASSO	$\lambda\lvert\beta_j\rvert$	λ
NJ	$\log\lvert\beta_j\rvert$	$\frac{1}{\lvert\beta_j\rvert}$
NG	$(\frac{1}{2} - \frac{\alpha}{K})\log\lvert\beta_j\rvert - \log\mathcal{K}_{\frac{\alpha}{K}-\frac{1}{2}}(\sqrt{2c}\lvert\beta_j\rvert)$	$\frac{\sqrt{2c}\mathcal{K}_{\frac{\alpha}{K}-\frac{3}{2}}(\sqrt{2c}\lvert\beta_j\rvert)}{\mathcal{K}_{\frac{\alpha}{K}-\frac{1}{2}}(\sqrt{2c}\lvert\beta_j\rvert)}$
NIG	$\log q_j - \log\mathcal{K}_1(cq_j)$	$\frac{2\lvert\beta_j\rvert}{q_j^2} + \frac{c\lvert\beta_j\rvert\mathcal{K}_0(cq_j)}{q_j\mathcal{K}_1(cq_j)}$
SCAD	$\lambda\lvert\beta_j\rvert\mathbb{I}(\lvert\beta_j\rvert \le \lambda) + \frac{1+\alpha}{2}\lambda^2\mathbb{I}(\alpha\lambda < \lvert\beta_j\rvert)$ $+\left[\lambda^2 + \frac{\alpha\lambda(\lvert\beta_j\rvert-\lambda)-\frac{1}{2}(\lvert\beta_j\rvert^2-\lambda^2)}{\alpha-1}\right]\mathbb{I}(\lambda < \lvert\beta_j\rvert \le \alpha\lambda)$	$\lambda\{\mathbb{I}(\lvert\beta_j\rvert \le \lambda)$ $+\frac{(\alpha\lambda-\lvert\beta_j\rvert)_+}{(\alpha-1)\lambda}\mathbb{I}(\lvert\beta_j\rvert > \lambda)\}$

Table A.2: Properties of sparsity promoting priors. Source: François Caron.

Name	Range	Finite value at 0	Sparsity	Convexity of $p_\lambda(\beta)$
Laplace	$c > 0$	yes	yes	Weakly convex for $\gamma > 0$
NJ	None	no	yes	Strictly concave
NG	$\frac{\alpha}{K} > 0, c > 0$	$\frac{\alpha}{K} \ge \frac{1}{2}, c > 0$	$\frac{\alpha}{K} \le 1$	Strictly convex for $\frac{\alpha}{K} > 1$ Weakly convex for $\frac{\alpha}{K} = 1$ Strictly concave for $\frac{\alpha}{K} < 1$
NIGauss	$\frac{\alpha}{K} > 0, c > 0$	yes	no	

Because of $\lim_{|\beta_j|\downarrow 0}(|\beta_j| + p_\lambda^{'}(|\beta_j|)) = 0$, the NIG prior does not satisfy the variable selection criterion.

Finally, Table A.1 summarizes these penalizations/log(prior) and their corresponding first order derivatives and table A.2 summarizes the properties of sparsity promoting priors (also including NIG prior).

Appendix B

Proofs

Before proving the theorems, we quote and present some useful lemmas.

B.1 Lemmas

Lemma B.1.1. *(c.f. [29] p310) Let X_n, X be r.v.'s. Fix $\epsilon > 0$, let*

$$A_n(\epsilon) = \{|X_n - X| > \epsilon\}$$

If $\sum_{n=1}^{\infty} A_n(\epsilon) < \infty$, $\forall \epsilon$, then $X_n \overset{a.s.}{\to} X$.

B.2 Proof of Proposition A.1.1

Proof. Define $\forall \epsilon > 0$

$$A_n(\epsilon) = \{|sgn(\hat{\beta}_{L1}(n)) - sgn(\beta)| > \epsilon\} \tag{B.1}$$

By Lemma B.1.1, hence suffices to show $\sum_n A_n(\epsilon) < \infty$. Under the hypothesis, Zhao and Yu [67] showed for some $0 \leq c < 1$

$$P(sgn(\hat{\beta}_{\lambda_n}(n)) = sgn(\beta)) = 1 - o(e^{-n^c}) \tag{B.2}$$

But it is straightforward to check that

$$\sum_n A_n(\epsilon) = \sum_n (1 - P(sgn(\hat{\beta}_{\lambda_n}(n)) = sgn(\beta))) = \sum_n o(e^{-n^c}) < \infty$$

□

B.3 Proof of Theorem 2.3.1

Proof. For part (1), clearly $\Phi_3(\theta_0, \theta_0) = p_{\lambda 3}(|\theta_0|)$. $p_{\lambda 3}(|\theta|)$ and $\Phi_3(\theta, \theta_0)$ are even functions on $\mathbb{R}$.

Let $x > 0$. Consider

$$\begin{aligned}
\Delta(x) &= \frac{d\left[\Phi_3(x, \theta_0) - p_{\lambda 3}(|x|)\right]}{dx} \\
&= \frac{x p_\lambda^{'}((|\theta_0| + \epsilon)+)}{|\theta_0| + \epsilon} - p_\lambda^{'}(x + \epsilon) + \epsilon \frac{p_\lambda^{'}(x + \epsilon)}{x + \epsilon} \\
&= x \left[\frac{p_\lambda^{'}((|\theta_0| + \epsilon)+)}{|\theta_0| + \epsilon} - \frac{p_\lambda^{'}(x + \epsilon)}{x + \epsilon} \right]
\end{aligned}$$

Now for $\theta \in (0, \infty)$, letting $x \downarrow \theta$, we have

$$\begin{aligned}
\Delta(\theta) &= \lim_{x \downarrow \theta} \Delta(x) \\
&= \theta \left[\frac{p_\lambda^{'}((|\theta_0| + \epsilon)+)}{|\theta_0| + \epsilon} - \frac{p_\lambda^{'}((\theta + \epsilon)+)}{\theta + \epsilon} \right]
\end{aligned}$$

Note that under hypothesis, it is straightforward to conclude that $\forall \epsilon > 0$, $\frac{p_\lambda^{'}(\theta+)}{\epsilon+\theta}$ is non-decreasing, positive of $\theta > 0$. So, $\Delta(\theta) \leq 0$ for $0 < \theta < |\theta_0|$; and $\Delta(\theta) \geq 0$ for $\theta > |\theta_0|$. Hence, $\Phi_3(x, \theta_0) - p_{\lambda 3}(|x|)$ attains its minimum at $|\theta_0|$, and therefore at $\pm|\theta_0|$.

For part (2), following definition

$$\begin{aligned}
|p_{\lambda 3}(\theta) - p_\lambda(\theta)| &= \left| p_\lambda(|\theta| + \epsilon) - p_\lambda(|\theta|) - \epsilon \int_0^{|\theta|} \frac{p_\lambda^{'}(\epsilon + t)}{\epsilon + t} dt \right| \\
&\leq |p_\lambda(|\theta| + \epsilon) - p_\lambda(|\theta|)| + \epsilon \left| \int_0^{|\theta|} \frac{p_\lambda^{'}(\epsilon + t)}{\epsilon + t} dt \right| \\
&\leq |p_\lambda(|\theta| + \epsilon) - p_\lambda(|\theta|)| + \epsilon p_\lambda^{'}(\epsilon+) \left| \int_0^{|\theta|} \frac{1}{\epsilon + t} dt \right| \\
&= |p_\lambda(|\theta| + \epsilon) - p_\lambda(|\theta|)| + \epsilon p_\lambda^{'}(\epsilon+) \log\left(1 + \frac{|\theta|}{\epsilon}\right)
\end{aligned}$$

By the piecewise differentiability assumption of $p_\lambda(\cdot)$ on $(0, \infty)$, we have $p'_\lambda(\epsilon+) < \infty$, $\forall \epsilon > 0$. Further, by compactness assumption, $p_\lambda(\cdot)$ is uniformly continuous on $\mathcal{C}$. Hence, sending $\epsilon \downarrow 0$ yields claimed result. $\square$

Appendix C

Supplementary Materials

Table C.1: *E.coli* Transcription factors in the 60% precise network with $p \geq 5$ predicted operon targets by CLR algorithm.

Regulator	Targets in RegulonDB	Targets # inferred by CLR
flhC_b1891_at	30	20
flhD_b1892_at	46	17
fliA_b1922_at	42	40
gatR_2_b2090_f_at	6	6
glcC_b2980_at	5	10
hycA_b2725_at	7	10
lexA_b4043_at	16	6
rcsB_b2217_at	11	5
rhaR_b3906_at	5	9
rhaS_b3905_at	5	7
tdcR_b3119_at	7	17
yhiE_b3512_at	5	11
yhiW_b3515_at	4	6
yhiX_b3516_at	13	8

Table C.2: *E.coli* Transcription factors in the 60% precise network with $p \geq 5$ predicted operon targets by L1 prior with MM3 algorithm.

Regulator	Targets in RegulonDB	Targets # inferred by L1 MM3
cbl_b1987_at	9	14
fecI_b4293_at	6	28
fliA_b1922_at	42	43
gatR_2_b2090_f_at	6	7
hycA_b2725_at	7	9
lexA_b4043_at	16	7
nac_b1988_at	12	9
narL_b1221_at	84	7
yhiE_b3512_at	5	10
ylcA_b0571_at	4	5

Table C.3: *E.coli* Transcription factors in the 60% precise network with $p \geq 5$ predicted operon targets by SCAD prior with MM3 algorithm.

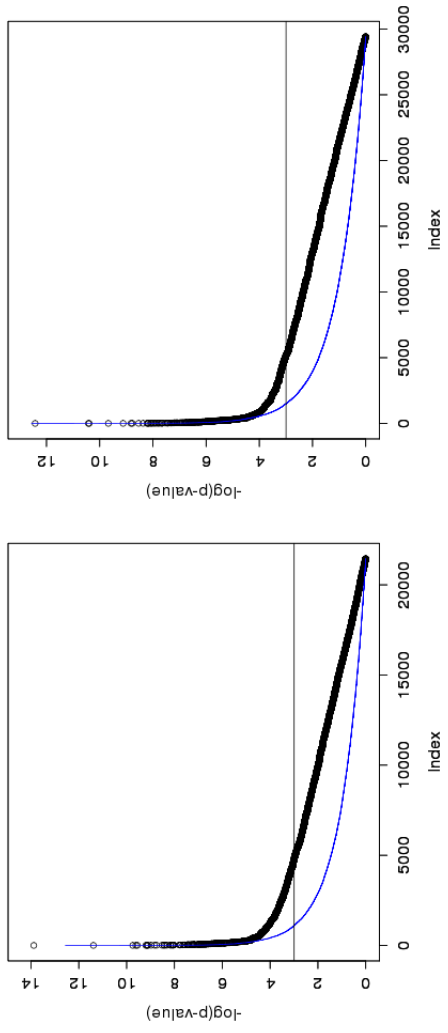
Regulator	Targets in RegulonDB	Targets # inferred by SCAD MM3
araC_b0064_at	8	6
cbl_b1987_at	9	14
fecI_b4293_at	6	28
fliA_b1922_at	42	44
gatR_2_b2090_f_at	6	7
hycA_b2725_at	7	9
lexA_b4043_at	16	7
nac_b1988_at	12	12
narL_b1221_at	84	7
yhiE_b3512_at	5	11
ylcA_b0571_at	4	5

Table C.4: *E.coli* Transcription factors in the 60% precise network with $p \geq 5$ predicted operon targets by L1 prior with EM algorithm.

Regulator	Targets in RegulonDB	Targets # inferred by L1 EM
araC_b0064_at	8	6
cbl_b1987_at	9	10
fecI_b4293_at	6	28
fliA_b1922_at	42	44
gatR_2_b2090_f_at	6	7
hycA_b2725_at	7	14
lexA_b4043_at	16	7
lrp_b0889_at	61	5
nac_b1988_at	12	9
narL_b1221_at	84	5
yhiE_b3512_at	5	10
ylcA_b0571_at	4	5

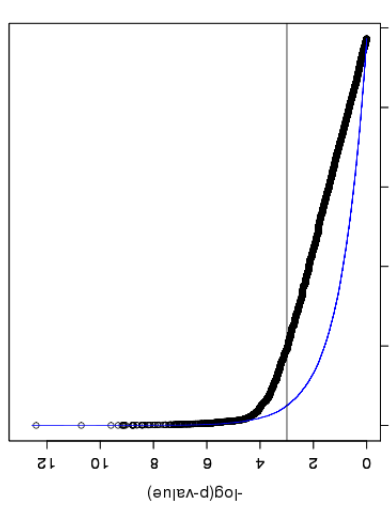
Table C.5: *E.coli* Transcription factors in the 60% precise network with $p \geq 5$ predicted operon targets by NIG prior with EM algorithm.

Regulator	Targets in RegulonDB	Targets # inferred by NIG EM
fecI_b4293_at	6	19
fliA_b1922_at	42	33
gatR_2_b2090_f_at	6	6
hycA_b2725_at	7	10
lexA_b4043_at	16	5
lrp_b0889_at	61	6
nac_b1988_at	12	15
yhiE_b3512_at	5	7
ylcA_b0571_at	4	5



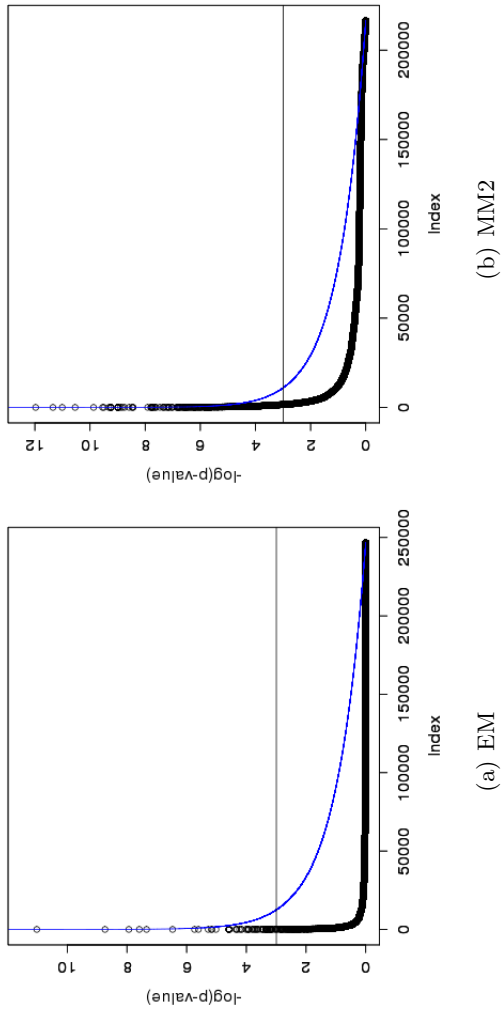
(a) EM

(b) MM2



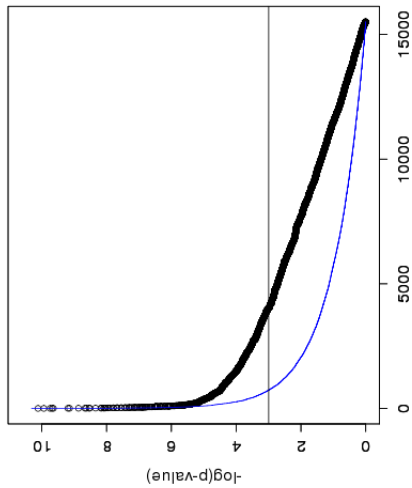
(c) MM3

Figure C.1: Fisher's exact test p-values with the NG prior. The blue line is the simulated uniform background p-values under null hypothesis and horizontal line is the threshold at p-value at 0.05.



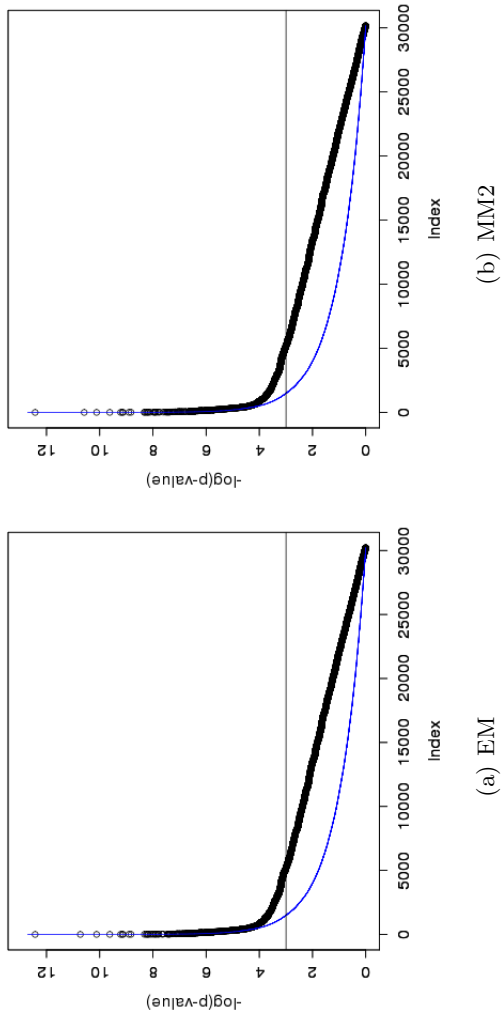
(a) EM

(b) MM2



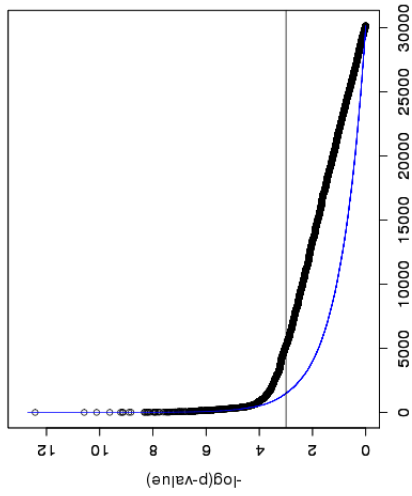
(c) MM3

Figure C.2: Fisher's exact test p-values with the NIG prior. The blue line is the simulated uniform background p-values under null hypothesis and horizontal line is the threshold at p-value at 0.05.



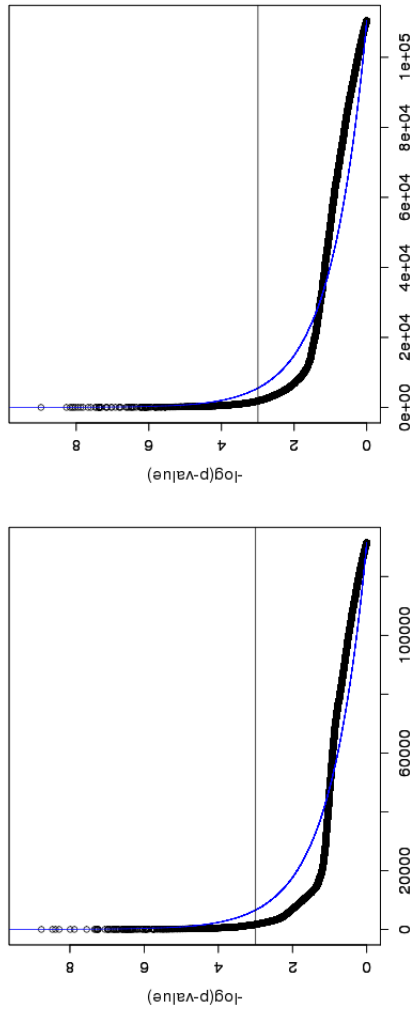
(a) EM

(b) MM2



(c) MM3

Figure C.3: Fisher's exact test p-values with the NJ prior. The blue line is the simulated uniform background p-values under null hypothesis and horizontal line is the threshold at p-value at 0.05.



(a) MM2 (b) MM3

Figure C.4: Fisher's exact test p-values with the SCAD prior. The blue line is the simulated uniform background p-values under null hypothesis and horizontal line is the threshold at p-value at 0.05.

Table C.6: Identified significant p53/miRNAs that target known cell proliferation genes from NG prior.

Method	Regulator	Total # of predicted targets	Cell proliferation genes	*p*-value	Gene names
EM	TP53	70	9	1.586E-04	TGFBI,IGFBP4,CYR61,AREG,EPS8, IFI16,TPX2,DLG7,PDGFC
	ACADVL	1	1	3.571E-02	TYR
	hsa-miR-130b	16	3	7.496E-04	CUL5,PDGFC,ChGn
	hsa-miR-145	15	2	2.803E-02	FSCN1,PDGFC
	hsa-miR-155	12	2	1.178E-02	FGF2,SPOCK1
	hsa-miR-15a	4	1	4.726E-02	FGF2
	hsa-miR-181a	7	2	1.572E-02	GNAI2,UCHL1
	hsa-miR-221	24	3	4.478E-03	CYR61,FRAT2,PDGFC
	hsa-miR-339	3	1	2.222E-02	ELF4
	hsa-miR-367	16	2	1.348E-02	HDGFRP3,BHLHB3
	hsa-miR-429	9	2	8.134E-03	CYR61,IRS2
	hsa-miR-485-5p	3	1	4.950E-02	SPOCK1
	hsa-miR-513	18	2	3.284E-02	EPS8,BHLHB3
	hsa-miR-522	18	2	4.320E-02	CYR61,EPS8
	ACADVL	1	1	3.846E-02	SMARCA4
	hsa-miR-1	13	2	3.358E-03	IRS2,FOXA1
	hsa-miR-30c	7	2	2.724E-03	GNAI2,PBEF1
	hsa-miR-375	8	1	5.000E-02	LRP5
	hsa-miR-410	17	2	1.030E-02	IRS2,HOXC10
	E2F1	6	2	2.293E-02	IL1B,TGFB1
	MYCN	10	2	1.704E-02	TIMP2,LEPRE1
	hsa-let-7i	3	1	8.772E-03	ETS1
	hsa-miR-202*	2	1	4.293E-02	SMAD4
	hsa-miR-33	8	1	3.309E-02	HDAC4
	hsa-miR-506	18	2	2.643E-02	PCAF,ETS1
	hsa-miR-518c*	6	1	3.877E-02	ETS1
	hsa-miR-7	14	2	1.920E-02	CNOT8,ETS1
MM2	TP53	105	10	8.167E-04	TGFBI,IGFBP4,CYR61,AREG,MKI67, EPS8,IFI16,TPX2,DLG7,PDGFC
	ACADVL	1	1	3.571E-02	TYR
	hsa-miR-130b	24	2	3.632E-02	CUL5,ChGn
	hsa-miR-145	19	2	4.388E-02	FSCN1,IRS2
	hsa-miR-155	13	2	1.383E-02	FGF2,SPOCK1
	hsa-miR-181a	13	3	4.940E-03	GNAI2,IRS2,UCHL1
	hsa-miR-221	31	4	5.859E-04	CYR61,GNAI2,FRAT2,PDGFC
	hsa-miR-27b	27	2	4.676E-02	TRIB1,BHLHB3
	hsa-miR-320	16	2	1.532E-02	PTEN,BHLHB3
	hsa-miR-339	4	1	2.963E-02	ELF4
	hsa-miR-367	26	2	3.471E-02	HDGFRP3,BHLHB3
	hsa-miR-429	22	2	4.640E-02	CYR61,IRS2
	hsa-miR-488	10	2	4.237E-02	SPOCK1,HDGFRP3
	hsa-miR-517*	4	1	3.695E-02	GPC4
	hsa-miR-518f	6	2	9.557E-03	PTEN,SPOCK1
	ACADVL	1	1	3.846E-02	SMARCA4
	hsa-miR-1	22	2	9.713E-03	IRS2,FOXA1
	hsa-miR-145	19	2	2.181E-02	IRS2,PBEF1
	hsa-miR-181a	13	2	4.727E-03	GNAI2,IRS2
	hsa-miR-206	18	2	6.501E-03	IRS2,PBEF1
	hsa-miR-221	31	2	3.805E-02	GNAI2,PDGFC
	hsa-miR-30c	20	2	2.263E-02	GNAI2,PBEF1
	hsa-miR-410	21	2	1.564E-02	IRS2,HOXC10
	MYCN	13	2	2.872E-02	TIMP2,LEPRE1
	hsa-let-7i	12	1	3.509E-02	ETS1
	hsa-miR-144	23	2	4.463E-02	TOB1,BTG3
	hsa-miR-186	31	3	3.672E-03	PTEN,HOXB2,SKAP2
	hsa-miR-219	5	1	3.546E-02	BMP2
	hsa-miR-26a	23	3	1.682E-03	TIMP2,SKAP2,HDAC4
	hsa-miR-26b	24	2	2.989E-02	PTEN,TIMP2

Table C.6: (continued)

Method	Regulator	Total # of predicted targets	Cell proliferation genes	*p*-value	Gene names
	hsa-miR-33	19	2	1.487E-03	NDN,HDAC4
	hsa-miR-34c	20	2	8.619E-03	NOTCH2,ETS1
	hsa-miR-429	22	3	7.382E-04	NDN,TIMP2,HDAC4
	hsa-miR-492	2	1	1.942E-02	ADAMTS1
	hsa-miR-498	9	2	1.332E-03	PTEN,SESN1
	hsa-miR-506	19	2	2.936E-02	PCAF,ETS1
	hsa-miR-518c*	7	1	4.516E-02	ETS1
	hsa-miR-7	17	2	2.805E-02	CNOT8,ETS1
MM3	TP53	83	10	1.138E-04	TGFBI,IGFBP4,CYR61,AREG,MKI67, EPS8,IFI16,TPX2,DLG7,PDGFC
	ACADVL	1	1	3.571E-02	TYR
	hsa-miR-130b	19	3	1.275E-03	CUL5,PDGFC,ChGn
	hsa-miR-145	16	2	3.172E-02	FSCN1,PDGFC
	hsa-miR-155	11	2	9.875E-03	FGF2,SPOCK1
	hsa-miR-181a	6	2	1.142E-02	GNAI2,UCHL1
	hsa-miR-221	28	4	3.867E-04	CYR61,GNAI2,FRAT2,PDGFC
	hsa-miR-339	3	1	2.222E-02	ELF4
	hsa-miR-367	21	2	2.300E-02	HDGFRP3,BHLHB3
	hsa-miR-429	14	2	1.965E-02	CYR61,IRS2
	hsa-miR-485-5p	3	1	4.950E-02	SPOCK1
	hsa-miR-488	7	2	2.095E-02	SPOCK1,HDGFRP3
	hsa-miR-513	22	2	4.805E-02	EPS8,BHLHB3
	ACADVL	1	1	3.846E-02	SMARCA4
	hsa-miR-145	16	2	1.559E-02	PBEF1,PDGFC
	hsa-miR-221	28	2	3.129E-02	GNAI2,PDGFC
	hsa-miR-23b	21	2	3.560E-02	IRS2,STAT5B
	hsa-miR-30c	11	2	6.949E-03	GNAI2,PBEF1
	hsa-miR-410	18	2	1.154E-02	IRS2,HOXC10
	ACADVL	5	2	3.271E-02	MDM4,TSG101
	MYCN	13	2	2.872E-02	TIMP2,LEPRE1
	hsa-let-7i	8	1	2.339E-02	ETS1
	hsa-miR-186	21	2	2.012E-02	PTEN,HOXB2
	hsa-miR-26a	16	2	1.361E-02	TIMP2,HDAC4
	hsa-miR-33	15	2	9.134E-04	NDN,HDAC4
	hsa-miR-429	14	3	1.803E-04	NDN,TIMP2,HDAC4
	hsa-miR-492	1	1	9.709E-03	ADAMTS1
	hsa-miR-506	17	2	2.365E-02	PCAF,ETS1
	hsa-miR-7	14	2	1.920E-02	CNOT8,ETS1

Table C.7: Identified significant p53/miRNAs that target known cell proliferation genes from NIG prior.

Method	Regulator	Total # of predicted targets	Cell proliferation genes	*p*-value	Gene names
EM	TP53	1261	45	5.680E-03	CDK4,PTCH1,PTEN,TGFBI,PAX3,IL1A, IL1B,ADRA1D,GAS6,PRL,BUB1B,CDK6, CTF1,IGFBP4,CYR61,AREG,CKS1B,EGF, FGF2,FGF7,FTH1,MDM4,MKI67,PIM1, PRKD1,MAP2K1,RAF1,STIL,TGFA,TXN, CUL1,BUB1,EPS8,PLK1,CSF1R,E2F1, IFI16,REST,TPX2,DLG7,PDGFC,OSM, E2F8,ZEB1,TNFSF13B
	TP53	1261	25	1.886E-02	KIT,PAX3,IGF1,BCL2,TGFB1,ADRA1D, IGF1R,CDC6,CTF1,EGF,F2R,FGF7, LIF,ODC1,ST8SIA1,TTK,CDKN1B,MYCN,

Table C.7: (continued)

Method	Regulator	Total # of predicted targets	Cell proliferation genes	*p*-value	Gene names
					HES1,PURA,PDGFC,ITGB1,TNFSF13B, SERTAD1,BIRC6
MM2	TP53	1181	41	4.859E-02	CDK4,PTCH1,PTEN,TGFBI,PAX3,IL1B, ADRA1D,GAS6,PRL,BUB1B,CTF1,IGFBP4, CYR61,AREG,CKS1B,EGF,FGF7,FTH1, MDM4,MKI67,PRKD1,MAP2K1,RAF1,STIL, TGFA,TXN,CUL1,BUB1,EPS8,PLK1,CSF1R, E2F1,IFI16,REST,TPX2,DLG7,PDGFC, OSM,E2F8,ZEB1,TNFSF13B
	FOS	421	29	4.859E-02	TGFBI,IL2RA,IL1A,IL1B,ADRA1D,ADRA1B, PRL,CTF1,DAB2,AREG,EGF,FGF2, GCG,GNB1,MAP2K1,RAF1,TGFA,TGFB3, TXN,IRS2,CDK5R1,E2F1,MPL,NAB2, S100B,CD160,FZD3,OSM,SFRP2
	TP53	1181	24	2.149E-02	KIT,PAX3,IGF1,BCL2,TGFB1,ADRA1D, CDC6,CTF1,EGF,F2R,FGF7,LIF, ODC1,ST8SIA1,TTK,CDKN1B,MYCN, HES1,PURA,PDGFC,ITGB1,TNFSF13B, SERTAD1,BIRC6
MM3	TP53	53	6	4.340E-03	TGFBI,CYR61,AREG,TPX2,DLG7,PDGFC
	hsa-miR-135b	3	1	4.822E-02	DAB2,DAB2
	hsa-miR-155	8	2	5.123E-03	FGF2,SPOCK1,FGF2,SPOCK1
	hsa-miR-15a	3	1	3.561E-02	FGF2,FGF2
	hsa-miR-181a	3	2	2.404E-03	GNAI2,UCHL1,GNAI2,UCHL1
	hsa-miR-221	19	3	2.214E-03	CYR61,FRAT2,PDGFC,CYR61,FRAT2,PDGFC
	hsa-miR-302c*	9	1	3.688E-02	GPC4,GPC4
	hsa-miR-367	9	2	4.188E-03	HDGFRP3,BHLHB3,HDGFRP3,BHLHB3
	hsa-miR-485-5p	3	1	4.950E-02	SPOCK1,SPOCK1
	hsa-miR-513	12	2	1.484E-02	EPS8,BHLHB3,EPS8,BHLHB3
	hsa-miR-522	16	3	2.730E-03	DAB2,CYR61,EPS8,DAB2,CYR61,EPS8
	ACADVL	1	1	3.846E-02	SMARCA4
	hsa-miR-128a	8	1	3.437E-02	PDGFC
	hsa-miR-128b	9	1	3.862E-02	PDGFC
	hsa-miR-181a	3	1	2.721E-02	GNAI2
	hsa-miR-191	4	1	4.460E-02	PBEF1
	hsa-miR-375	5	1	3.125E-02	LRP5
	hsa-miR-384	3	1	3.007E-02	IGF2BP1
	hsa-miR-410	10	2	3.506E-03	IRS2,HOXC10
	hsa-miR-429	6	1	3.994E-02	IRS2
	hsa-let-7i	1	1	2.924E-03	ETS1
	hsa-miR-106b	3	1	3.690E-02	TXNIP
	hsa-miR-181a	3	1	3.393E-02	NOTCH2
	hsa-miR-181b	4	1	4.506E-02	ETS1
	hsa-miR-222	4	1	2.878E-02	ETS1
	hsa-miR-33	7	1	2.898E-02	HDAC4
	hsa-miR-506	9	2	6.585E-03	PCAF,ETS1

Table C.8: Identified significant p53/miRNAs that target known cell proliferation genes from NJ prior.

Method	Regulator	Total # of predicted targets	Cell proliferation genes	*p*-value	Gene names
	TP53	109	10	1.101E-03	TGFBI,IGFBP4,CYR61,AREG,MKI67, EPS8,IFI16,TPX2,DLG7,PDGFC

Table C.8: (continued)

Method	Regulator	Total # of predicted targets	Cell proliferation genes	*p*-value	Gene names
	EGR1	22	4	4.628E-02	IL1B,IRS2,NAB2,PDGFC
	ACADVL	1	1	3.571E-02	TYR
	hsa-miR-130b	26	2	4.222E-02	CUL5,ChGn
	hsa-miR-155	13	2	1.383E-02	FGF2,SPOCK1
	hsa-miR-181a	14	4	3.999E-04	GNAI2,IRS2,UCHL1,BHLHB3
	hsa-miR-181c	7	2	1.572E-02	UCHL1,BHLHB3
	hsa-miR-221	31	4	5.859E-04	CYR61,GNAI2,FRAT2,PDGFC
	hsa-miR-339	5	1	3.704E-02	ELF4
	hsa-miR-367	27	2	3.730E-02	HDGFRP3,BHLHB3
	hsa-miR-488	10	2	4.237E-02	SPOCK1,HDGFRP3
	hsa-miR-517*	4	1	3.695E-02	GPC4
	hsa-miR-517c	4	2	6.162E-03	AREG,ChGn
	hsa-miR-518f	6	2	9.557E-03	PTEN,SPOCK1
	hsa-miR-522	38	4	4.423E-03	DAB2,CYR61,EPS8,PDGFC
	hsa-miR-526b	14	2	2.832E-02	ChGn,TXNDC1
	ACADVL	1	1	3.846E-02	SMARCA4
	hsa-miR-1	22	2	9.713E-03	IRS2,FOXA1
	hsa-miR-145	21	2	2.645E-02	IRS2,PBEF1
	hsa-miR-181a	14	2	5.498E-03	GNAI2,IRS2
	hsa-miR-206	19	2	7.247E-03	IRS2,PBEF1
	hsa-miR-221	31	2	3.805E-02	GNAI2,PDGFC
	hsa-miR-30c	20	2	2.263E-02	GNAI2,PBEF1
	hsa-miR-410	20	2	1.421E-02	IRS2,HOXC10
	MYCN	13	2	2.872E-02	TIMP2,LEPRE1
	hsa-let-7i	14	1	4.094E-02	ETS1
	hsa-miR-144	24	2	4.828E-02	TOB1,BTG3
	hsa-miR-186	33	3	4.403E-03	PTEN,HOXB2,SKAP2
	hsa-miR-219	5	1	3.546E-02	BMP2
	hsa-miR-26a	24	3	1.912E-03	TIMP2,SKAP2,HDAC4
	hsa-miR-26b	24	2	2.989E-02	PTEN,TIMP2
	hsa-miR-33	19	2	1.487E-03	NDN,HDAC4
	hsa-miR-34c	19	2	7.772E-03	NOTCH2,ETS1
	hsa-miR-429	23	3	8.455E-04	NDN,TIMP2,HDAC4
	hsa-miR-492	2	1	1.942E-02	ADAMTS1
	hsa-miR-498	8	2	1.036E-03	PTEN,SESN1
	hsa-miR-506	19	2	2.936E-02	PCAF,ETS1
	hsa-miR-518c*	7	1	4.516E-02	ETS1
	hsa-miR-7	17	2	2.805E-02	CNOT8,ETS1
	TP53	109	10	1.101E-03	TGFBI,IGFBP4,CYR61,AREG,MKI67, EPS8,IFI16,TPX2,DLG7,PDGFC
	EGR1	22	4	4.628E-02	IL1B,IRS2,NAB2,PDGFC
	ACADVL	1	1	3.571E-02	TYR
	hsa-miR-130b	27	2	4.530E-02	CUL5,ChGn
	hsa-miR-155	13	2	1.383E-02	FGF2,SPOCK1
	hsa-miR-181a	14	4	3.999E-04	GNAI2,IRS2,UCHL1,BHLHB3
	hsa-miR-181c	7	2	1.572E-02	UCHL1,BHLHB3
	hsa-miR-221	31	4	5.859E-04	CYR61,GNAI2,FRAT2,PDGFC
	hsa-miR-339	5	1	3.704E-02	ELF4
	hsa-miR-367	27	2	3.730E-02	HDGFRP3,BHLHB3
	hsa-miR-429	22	2	4.640E-02	CYR61,IRS2
	hsa-miR-488	10	2	4.237E-02	SPOCK1,HDGFRP3
	hsa-miR-517*	4	1	3.695E-02	GPC4
	hsa-miR-517c	4	2	6.162E-03	AREG,ChGn
	hsa-miR-518f	6	2	9.557E-03	PTEN,SPOCK1
	hsa-miR-522	38	4	4.423E-03	DAB2,CYR61,EPS8,PDGFC
	hsa-miR-526b	14	2	2.832E-02	ChGn,TXNDC1
	ACADVL	1	1	3.846E-02	SMARCA4
MM2	hsa-miR-1	22	2	9.713E-03	IRS2,FOXA1
	hsa-miR-145	21	2	2.645E-02	IRS2,PBEF1
	hsa-miR-181a	14	2	5.498E-03	GNAI2,IRS2

Table C.8: (continued)

Method	Regulator	Total # of predicted targets	Cell proliferation genes	*p*-value	Gene names
	hsa-miR-206	19	2	7.247E-03	IRS2,PBEF1
	hsa-miR-221	31	2	3.805E-02	GNAI2,PDGFC
	hsa-miR-30c	20	2	2.263E-02	GNAI2,PBEF1
	hsa-miR-410	20	2	1.421E-02	IRS2,HOXC10
	MYCN	13	2	2.872E-02	TIMP2,LEPRE1
	hsa-let-7i	12	1	3.509E-02	ETS1
	hsa-miR-144	23	2	4.463E-02	TOB1,BTG3
	hsa-miR-186	35	3	5.217E-03	PTEN,HOXB2,SKAP2
	hsa-miR-219	5	1	3.546E-02	BMP2
	hsa-miR-26a	24	3	1.912E-03	TIMP2,SKAP2,HDAC4
	hsa-miR-26b	25	2	3.230E-02	PTEN,TIMP2
	hsa-miR-33	20	2	1.653E-03	NDN,HDAC4
	hsa-miR-34c	19	2	7.772E-03	NOTCH2,ETS1
	hsa-miR-429	22	3	7.382E-04	NDN,TIMP2,HDAC4
	hsa-miR-492	2	1	1.942E-02	ADAMTS1
	hsa-miR-498	9	2	1.332E-03	PTEN,SESN1
	hsa-miR-506	20	2	3.241E-02	PCAF,ETS1
	hsa-miR-518c*	7	1	4.516E-02	ETS1
	hsa-miR-7	17	2	2.805E-02	CNOT8,ETS1
MM3	TP53	109	10	1.101E-03	TGFBI,IGFBP4,CYR61,AREG,MKI67, EPS8,IFI16,TPX2,DLG7,PDGFC
	EGR1	22	4	4.628E-02	IL1B,IRS2,NAB2,PDGFC
	ACADVL	1	1	3.571E-02	TYR
	hsa-miR-130b	27	2	4.530E-02	CUL5,ChGn
	hsa-miR-155	13	2	1.383E-02	FGF2,SPOCK1
	hsa-miR-181a	14	4	3.999E-04	GNAI2,IRS2,UCHL1,BHLHB3
	hsa-miR-181c	7	2	1.572E-02	UCHL1,BHLHB3
	hsa-miR-221	31	4	5.859E-04	CYR61,GNAI2,FRAT2,PDGFC
	hsa-miR-339	5	1	3.704E-02	ELF4
	hsa-miR-367	27	2	3.730E-02	HDGFRP3,BHLHB3
	hsa-miR-429	22	2	4.640E-02	CYR61,IRS2
	hsa-miR-488	10	2	4.237E-02	SPOCK1,HDGFRP3
	hsa-miR-517*	4	1	3.695E-02	GPC4
	hsa-miR-517c	4	2	6.162E-03	AREG,ChGn
	hsa-miR-518f	6	2	9.557E-03	PTEN,SPOCK1
	hsa-miR-522	38	4	4.423E-03	DAB2,CYR61,EPS8,PDGFC
	hsa-miR-526b	14	2	2.832E-02	ChGn,TXNDC1
	ACADVL	1	1	3.846E-02	SMARCA4
	hsa-miR-1	22	2	9.713E-03	IRS2,FOXA1
	hsa-miR-145	21	2	2.645E-02	IRS2,PBEF1
	hsa-miR-181a	14	2	5.498E-03	GNAI2,IRS2
	hsa-miR-206	19	2	7.247E-03	IRS2,PBEF1
	hsa-miR-221	31	2	3.805E-02	GNAI2,PDGFC
	hsa-miR-30c	20	2	2.263E-02	GNAI2,PBEF1
	hsa-miR-410	20	2	1.421E-02	IRS2,HOXC10
	MYCN	13	2	2.872E-02	TIMP2,LEPRE1
	hsa-let-7i	12	1	3.509E-02	ETS1
	hsa-miR-144	23	2	4.463E-02	TOB1,BTG3
	hsa-miR-186	35	3	5.217E-03	PTEN,HOXB2,SKAP2
	hsa-miR-219	5	1	3.546E-02	BMP2
	hsa-miR-26a	24	3	1.912E-03	TIMP2,SKAP2,HDAC4
	hsa-miR-26b	25	2	3.230E-02	PTEN,TIMP2
	hsa-miR-33	20	2	1.653E-03	NDN,HDAC4
	hsa-miR-34c	19	2	7.772E-03	NOTCH2,ETS1
	hsa-miR-429	22	3	7.382E-04	NDN,TIMP2,HDAC4
	hsa-miR-492	2	1	1.942E-02	ADAMTS1
	hsa-miR-498	9	2	1.332E-03	PTEN,SESN1
	hsa-miR-506	20	2	3.241E-02	PCAF,ETS1
	hsa-miR-518c*	7	1	4.516E-02	ETS1

Table C.8: (continued)

Method	Regulator	Total # of predicted targets	Cell proliferation genes	*p*-value	Gene names
	hsa-miR-7	17	2	2.805E-02	CNOT8,ETS1

Table C.9: Identified significant p53/miRNAs that target known cell proliferation genes from SCAD prior.

Method	Regulator	Total # of predicted targets	Cell proliferation genes	*p*-value	Gene names
MM2	TP53	471	21	2.503E-02	PTCH1,TGFBI,PAX3,IL1B,CTF1,IGFBP4, CYR61,AREG,MKI67,PIM1,PRKD1,STIL, TXN,BUB1,EPS8,IFI16,REST,TPX2, DLG7,PDGFC,ZEB1
	FOXO1	67	4	2.638E-02	PTEN,PRL,IRS2,IFI16
	ACADVL	44	4	4.859E-02	PAX3,DAB2,ELF4,PIM1
	ACADVL	133	9	2.085E-02	PTEN,PAX3,ADRA1D,ADRA1B,CTF1, EGF,ISG20,PIM1,CSF1R
	hsa-miR-135b	178	6	3.935E-02	RPS4X,RPS27,DAB2,MNAT1,CDK5R1,MAPRE2
	hsa-miR-15a	128	4	2.095E-02	RPS4X,RPS27,FGF2,FZD3
	hsa-miR-200c	163	5	1.716E-02	PTEN,RPS4X,RPS27,EPS15,MAPRE1
	RB1	104	9	1.741E-02	KIT,IGF1,TGFB1,IGF1R,CDC6, CCND2,CDKN1B,MYCN,SUZ12
	EGR1	83	9	1.926E-02	TGFB1,EGF,F2R,FLT1,LYN,TIMP1,IRS2,PDGFC,CHRM1
	PAX5	36	7	1.509E-02	IL7,FLT3LG,FGF17,CDKN1B,FLT3,HES1,STAT5B
	hsa-miR-106b	283	7	2.611E-02	BCL2,CCND2,LIF,AKAP5,POU3F2,DERL2,IGF2BP1
	hsa-miR-130b	205	6	4.212E-02	IGF1,BCL2,CDKN1B,POU3F2,SUZ12,IGF2BP1
	hsa-miR-181b	174	6	6.323E-03	BCL2,LIF,CUL3,AKAP5,POU3F2,DAZAP2
	hsa-miR-186	406	7	3.294E-02	IGF1R,ITPR1,CDKN1B,AKAP5,POU3F2,SUZ12,IGF2BP1
	hsa-miR-296	16	2	3.526E-02	NR2E1,IGF2BP1
	hsa-miR-301	178	7	2.639E-03	IGF1,BCL2,IGF1R,CDKN1B,SUZ12,PDGFC,IGF2BP1
	hsa-miR-30b	259	7	1.385E-02	GLDC,KLF5,IRS2,SUZ12,GLI2,BIRC6,PLEKHK1
	hsa-miR-330	107	4	4.969E-02	MAB21L1,POU3F2,STAT5B,IGF2BP1
	hsa-miR-372	162	5	5.496E-03	IGF1R,NR2E1,CUL3,DAZAP2,SUZ12
	hsa-miR-452	146	3	3.781E-02	CCND2,GNAI2,IRS2
	hsa-miR-520a*	126	4	4.548E-02	IGF1R,CCND2,SUZ12,IGF2BP1
	hsa-miR-522	312	9	1.486E-02	GLDC,IGF1,CCND2,GNAI2,CDKN1B, SUZ12,PDGFC,IGF2BP1,PLEKHK1
	hsa-miR-9	296	6	2.436E-02	CCND2,NR2E1,FGF18,CDKN1B,PURA,ANGEL1
	HES1	27	3	2.329E-02	AGT,IL1B,TGFB1
	LITAF	6	2	3.448E-02	IL1A,TNF
	ACADVL	29	4	3.295E-02	TNF,TGFB1,FGF2,FOXO4
	ACADVL	14	3	2.748E-02	POU1F1,TNF,TGFB1
	FOSB	19	3	1.770E-02	IL1B,TNF,BMP2
	HDAC3	20	3	4.596E-02	PTEN,TNF,GDF11
	ACADVL	40	4	2.638E-02	APC,IL3,TGFB1,BMP2
	IRF1	47	8	1.302E-02	IL1B,TNF,TGFB1,PTGS2,EIF2AK2,JAK2,OSM,IL1RN
	MYCN	79	5	5.248E-03	GLI3,RB1,TGFB1,TIMP2,LEPRE1
	ACADVL	74	5	3.488E-02	PTEN,IL1A,IL1B,TNF,TGFB1
	hsa-miR-31	149	7	1.709E-02	FRK,TIMP2,JAK2,HDAC4,STK38,PDS5B,ETS1
	hsa-miR-34c	127	3	4.396E-02	BTG1,NOTCH2,ETS1
	hsa-miR-507	61	3	3.894E-02	BTG1,HDAC4,MNT
	hsa-miR-519b	245	8	3.132E-02	PTEN,RB1,CUL5,HDAC4,TSG101,ADAMTS1,STK38,RFX3
MM3	ACADVL	12	2	4.286E-02	CDK4,TYR
	ACADVL	12	3	1.357E-02	CDK4,CDK6,E2F1
	hsa-miR-125a	87	4	9.832E-03	RPS27,MAP3K11,MAPRE2,PES1
	hsa-miR-135b	135	5	4.667E-02	RPS4X,DAB2,MNAT1,PIM2,MAPRE2
	hsa-miR-155	110	4	4.655E-02	PTEN,FGF2,SPOCK1,BCAT1
	hsa-miR-181a	110	7	2.308E-02	TGFBI,CKS1B,GNAI2,IRS2,UCHL1,BHLHB3,ZAK

Table C.9: (continued)

Method	Regulator	Total # of predicted targets	Cell proliferation genes	*p*-value	Gene names
	hsa-miR-202*	61	4	9.055E-03	PTEN,EPS15,FGF2,CUL5
	hsa-miR-221	147	6	4.146E-03	RPS4X,RPS27,CYR61,GNAI2,FRAT2,PDGFC
	hsa-miR-27b	152	6	3.146E-03	RPS4X,EPS15,PRKD1,STIL,TRIB1,BHLHB3
	hsa-miR-320	127	4	4.963E-02	PTEN,DAB2,STAT4,BHLHB3
	hsa-miR-410	122	8	8.823E-04	COL4A3,RPS4X,RPS27,FGF2,PRKD1,IRS2,EPS8,HDGFRP3
	hsa-miR-504	27	3	3.525E-02	RPS27,GPC4,NAB2
	hsa-miR-518f	46	4	2.105E-02	PTEN,RPS4X,RPS27,SPOCK1
	hsa-miR-522	223	8	3.862E-02	RPS4X,DAB2,CYR61,GNAI2,MAP2K1,EPS8,PDGFC,BCAR1
	ACADVL	24	5	2.113E-02	BCL2,TGFB1,CXCL10,CCND2,CDKN1B
	PAX5	33	7	7.721E-03	IL7,FLT3LG,FGF17,CDKN1B,FLT3,HES1,STAT5B
	hsa-miR-206	115	3	3.531E-02	IRS2,FOXA1,PBEF1
	hsa-miR-296	18	2	4.496E-02	NR2E1,IGF2BP1
	ACADVL	14	3	2.748E-02	POU1F1,TNF,TGFB1
	SP1	42	5	1.295E-02	TGFB1,PTGS2,EREG,HDAC4,ETS1
	HDAC2	19	3	2.832E-02	IL1B,TNF,TGFB1
	IRF1	49	8	1.754E-02	IL1B,TNF,TGFB1,PTGS2,EIF2AK2,JAK2,OSM,IL1RN
	MYCN	85	5	7.638E-03	GLI3,RB1,TGFB1,TIMP2,LEPRE1
	PSMC3	6	2	4.455E-02	TGFB1,MDM4
	SMAD3	57	6	4.923E-02	IL1B,TGFB1,EREG,IFNB1,TGFB3,SMAD4
	hsa-miR-128b	140	5	2.906E-02	PHB,FOXO4,MYO16,PDS5B,ING5
	hsa-miR-144	239	7	3.864E-02	PRKRA,PCAF,JAK2,TOB1,HDAC4,BTG3,ETS1
	hsa-miR-186	375	8	1.997E-02	PTEN,HOXB2,SKAP2,SMAD4,TOB1,HDAC4,CHERP,TOB2
	hsa-miR-195	54	4	2.600E-02	PPM1D,CHERP,SESN1,PDS5B
	hsa-miR-202*	61	4	2.265E-02	PTEN,FGF2,CUL5,SMAD4
	hsa-miR-31	132	6	4.215E-02	FRK,TIMP2,SKAP2,JAK2,HDAC4,STK38
	hsa-miR-34c	109	3	2.768E-02	BTG1,NOTCH2,ETS1
	hsa-miR-429	187	5	1.303E-02	NDN,TIMP2,TOB1,HDAC4,SESN1
	hsa-miR-499	150	4	2.944E-02	CUL5,CUL1,PPM1D,FOXO4
	hsa-miR-518e	44	3	2.886E-02	BMP2,MED17,PDS5B
	hsa-miR-7	117	5	1.493E-02	PPM1D,CNOT8,SETDB1,ETS1,GJB6