Learning Gene Regulatory Networks via Globally Regularized Risk Minimization

Yuhong Guo and Dale Schuurmans

Department of Computing Science,
University of Alberta, Edmonton T6G 2E8, Canada
{yuhong,dale}@cs.ualberta.ca

Abstract. Learning the structure of a gene regulatory network from time-series gene expression data is a significant challenge. Most approaches proposed in the literature to date attempt to predict the regulators of each target gene individually, but fail to share regulatory information between related genes. In this paper, we propose a new globally regularized risk minimization approach to address this problem. Our approach first clusters genes according to their time-series expression profiles—identifying related groups of genes. Given a clustering, we then develop a simple technique that exploits the assumption that genes with similar expression patterns are likely to be co-regulated by encouraging the genes in the same group to share common regulators. Our experiments on both synthetic and real gene expression data suggest that our new approach is more effective at identifying important transcription factor based regulatory mechanisms than the standard independent approach and a prototype based approach.

1 Introduction

Genes and their products do not work independently in the cell. Rather, they are jointly regulated in a coordinated fashion, both internally and externally, to achieve proper cell function. One of the key mechanisms of gene regulation takes place at the mRNA transcription level. With the emergence of high-throughput microarray techniques, the mRNA expression levels of thousands of genes can be measured simultaneously. Using computational techniques to learn gene regulatory networks from high-throughput time-series gene expression data has been an active area of research in recent years. The goal of such research is to discover the causal control relationships between genes, which would offer a fundamental understanding of how biological processes are coordinated in the cell.

A variety of computational approaches have been proposed in the literature to model gene regulatory networks from expression data. Many approaches have been based on the use of linear models to express dependence between time series profiles. For example, D'Haeseleer et al. [1] studied a straightforward linear model for this purpose; Chen et al. [2] and De Jong et al. [3] investigated linear differential equations for gene regulatory network modeling. All of these approaches suffer from risks of over-fitting, however, since they fit a number of

parameters that is proportional to the size of the data itself. To counter the risk of over-fitting, other linear approaches have taken advantage of sparseness of the regulatory relationship between genes; that is, that any one gene is regulated by a small subset of the other genes. De Hoon et al. [4] have proposed to use "Akaike's Information Criterion" (AIC) to determine the nonzero coefficients in the linear system. Similarly, Li & Yang [5] used "L1 regularization" to conduct feature selection on the linear parent set.

Another popular approach to learning gene regulatory network structure is to exploit various forms of (dynamic) Bayesian network structure learning methods. A Bayesian network is a graphical representation of the causal relationships underlying a set of variables that provides a sound probabilistic framework for representing and inferring probabilistic relationships. *Dynamic* Bayesian networks are a natural extension of Bayesian networks to modeling time-series data. Learning the structure of a Bayesian network from data generally requires one of two approaches to be followed: a score-based approach—where a heuristic search is performed through the space of causal network structures to identify the most likely structure explaining the data—and a constraint-based approach—where conditional independence tests are used to determine whether a direct causal relationship should be postulated between two variables. Many variants of these techniques have been applied to gene regulatory network learning, including search-based approaches [6–8], information-theoretic approaches [9], parameter-tying based approaches [10], and conventional dynamic Bayesian network learning approaches [11, 12].

Although these previous techniques have achieved some promising results, the fundamental limitation of the amount of data available relative to the large number of parameters estimated (e.g. distinct parameters used to predict the expression level of each gene given other genes) severely constrains their effectiveness. This difficulty is inherent to the task: orders of magnitude more expression data would be required for naive estimation approaches devoid of background knowledge and biologically relevant assumptions to succeed on this problem.

One common shortcoming in the current literature, whether using linear modeling or using Bayesian network structure learning, is that nearly all proposed approaches attempt to determine the regulation structure for each target gene independently. Yet it is well known that genes that share the same expression pattern are likely to be involved in the same regulatory process, and therefore share the same (or at least a similar) set of regulators [13]. The main question we investigate is how to exploit biologically significant knowledge about co-regulation to improve the inference of the underlying gene regulatory network from expression data. Although a few previous investigators, such as van Someren et al. (2000), have proposed to group genes with similar expression profiles in a single prototypical "gene", and then model the relations between prototypical genes instead of modeling the genes individually, this is a somewhat oversimplified approach that ultimately ignores the individual differences between genes in the same group, and puts a particular high requirement on the clustering step.

In this paper, we propose a novel approach for predicting the regulators for a given group of genes with similar mRNA expression patterns, by minimizing a globally shared regularized prediction risk that encourages similar genes to share regulators. The models we learn, however, are otherwise standard linear models. The novelty of the approach is to first cluster the genes based on their time series expression profiles, and then minimize a loss determined on a set of global indicator variables associated with the common set of possible regulatory variables. We evaluate the performance of our approach on both synthetic data and the cell cycle time-series gene expression data of [14]. Our synthetic results show that our approach is able to learn the correct structure far more effectively than the typical approach that does not take into account co-regulation knowledge. Our results on the Cho et al. (1998) cell cycle data suggests our approach can identify the important transcription factors in the cell cycle genes more accurately by exploiting the co-regulation knowledge.

2 Method

The core of our method is based on using linear regression models to infer the expression level of each target gene from the expression levels of a set of potential regulator genes. However, even though linear prediction provides a simple and elegant foundation for modeling time series expression data, it cannot be applied naively. At least three significant issues need to be addressed before reasonable results can be achieved. First, time lags exist in the regulatory pathways controlling gene expression. These time lags vary between pathways and remain generally unknown *a priori* [12]. Second, the number of parameters required by a simple linear model (one parameter for each target-regulator combination) is far too many to be estimated reliably from available time series gene expression data. Some sort of effective feature selection mechanism must be employed [5]. Third, genes that serve related or synchronized functions tend to share common regulatory mechanisms. That is, related genes tend to share common regulators, and this knowledge must be exploited somehow to improve the quality of the regulation networks that are inferred. Failure to take into account any of these issues causes the linear prediction (or any other) approach to perform poorly.

We take into account all three of the above issues and modify the linear prediction approach to infer gene regulatory networks from time series expression data. The first two issues have been handled in varying ways in existing research—although we propose particularly simple and elegant ways to handle them in this paper. The third issue comprises the main observation we make, and motivates our use of a novel form of global risk minimization that is able to share regulatory information between similar genes while simultaneously allowing individual differences.

2.1 Linear Modeling

First, to establish the basic linear prediction approach consider an $n \times t$ matrix Y of time series gene expression data, where each column corresponds to the

expression levels of a single gene measured over a series of n time points; hence, Y stores the expression profiles for t genes. For each gene, we would like to identify which other genes measured in Y are likely to be regulators. The fundamental hypothesis is that the expression levels of a regulator gene should be predictive of the expression levels for a regulated target gene, possibly subject to time lag and the presence of co-regulators or absence of inhibitors.[1]

A straightforward linear prediction approach proceeds as follows. Assume for a target expression profile $\mathbf{y}_j$ given by an $n \times 1$ column vector from Y, we have a set of candidate regulator profiles stored in an $n \times k$ matrix X_j consisting of k distinct columns selected from Y. (We will discuss below how such a set of candidate profiles might be inferred for a given target $\mathbf{y}_j$.) The quality of this set of candidate regulators can be assessed by how well their expression levels predict the expression levels of the target, which can be determined by solving for the combination weights of the regulator profiles that best reconstruct the target profile

$$\min_{\mathbf{w}_j} \ \|X_j \mathbf{w}_j - \mathbf{y}_j\|_2^2. \tag{1}$$

Here the $k \times 1$ vector of combination weights $\mathbf{w}_j$ describes how the expression levels of the regulator genes in X_j interact to best explain the target expression levels $\mathbf{y}_j$, and the quality of the fit can be assessed by the residual error in (1).

2.2 Coping with Time Lags via Time Shifting

Unfortunately, the naive linear modeling approach (1) suffers from the three major drawbacks mentioned above. The first problem is that it does not account for any time lag between the expression of a regulating gene and the expression of its downstream target. In fact, the naive approach (1) implicitly assumes that regulation occurs instantaneously, and therefore performs quite poorly at identifying any regulatory relationship that exhibits delayed effects. To cope with this shortcoming, we modify the approach to first take into account any potential time lag between the expression of a regulator and its downstream target. In particular, for each candidate regulator measured in X_j, given by an $n \times 1$ vector $\mathbf{x}_{ij}$, we first compute an optimal shift back in time that best aligns $\mathbf{x}_{ij}$ individually with the target $\mathbf{y}_j$

$$s_{ij}^* = \arg \min_{s \in \{0,1,2,3\}} \|\mathbf{x}_{ij}(1, ..., n-s) - \mathbf{y}_j(s+1, ..., n)\|_2^2. \tag{2}$$

(Note that the shifts only allow time lags forward in time from the expression of the regulator to the expression of the target.) Repeating this for each candidate regulator profile in X_j, yields a series of optimal time lags. We can then reformulate the expression matrix X_j for the candidate regulators by applying the

[1] To mitigate the effect of measurement errors and outliers in the expression data, we generally assume the columns of Y have been rescaled to values between 0 and 1, and thus we are only searching for explanations of *relative* increases or decreases in expression level.

optimal shift to each column, and truncating the columns to a common length based on the maximum shift, obtaining an $(n-s_{max})\times k$ time-lag aligned matrix Φ_j. The target expression profile $\mathbf{y}_j$ is then also truncated to a corresponding $(n-s_{max})\times 1$ vector $\tilde{\mathbf{y}}_j$, where $\tilde{\mathbf{y}}_j = \mathbf{y}_j(s_{max}, ..., n)$. The quality of the candidate regulators can then be assessed by the more appropriate aligned reconstruction

$$\min_{\mathbf{w}_j} \ \|\Phi_j \mathbf{w}_j - \tilde{\mathbf{y}}_j\|_2^2. \tag{3}$$

2.3 Feature Selection via L1 Regularized Risk Minimization

Although the modified linear approach (3) appropriately handles time lags between regulator and target expression patterns, it still suffers from a major drawback: the set of candidate regulators for a given gene is usually very large (e.g. the complete set of remaining genes), while the number of time points sampled in a time series experiment is usually quite small (on the order of 20 to 30). Therefore a large set of combination weights $\mathbf{w}_j$ need to be inferred from a limited amount of data. Moreover, only a tiny fraction of the candidate regulators are expected to be true regulators for any given gene, meaning that, ideally, most of the weights should be set to 0 to indicate non-regulation. The bottom line is that some sort of effective form of *feature selection* is required for this problem. From a large set of candidate regulator expression profiles, most need to be discarded, and a small number retained to provide a good explanation of the target expression profile.

It is well known in the machine learning literature [15] that using the L1 norm (rather than the more conventional L2 norm) for regularization is very effective for feature selection. In this approach, one adds a penalty to the risk (the reconstruction objective) which encourages small values for $\mathbf{w}_j$:

$$\min_{\mathbf{w}_j} \ \|\Phi_j \mathbf{w}_j - \tilde{\mathbf{y}}_j\|_2^2 + \alpha\|\mathbf{w}_j\|_1, \tag{4}$$

where α is a parameter that trades off the influence of the risk with the regularizer. Crucially, this regularizer encourages many of the weights to become exactly zero in the solution. To see why, note that the regularization term is non-differentiable at zero, but any movement of a weight from zero immediately creates a derivative of magnitude α encouraging movement back to zero. Thus, if the magnitude of the derivative of the risk is not greater than α, then the weight will remain at zero. These intuitions lead to an efficient optimization procedure known as grafting [16].

2.4 Regulation Sharing via Globally Regularized Risk Minimization

Simply solving the minimization problem in (4) provides no advantage over the approaches proposed in the literature however, since it does not address the problem of facing a shortage of data while trying to make inferences about a large number of genes. To mitigate this problem we propose to share regulatory

information across sets of target genes. Given the hypothesis that genes with similar expression patterns are usually co-regulated and involved in the same functional process, we propose to first cluster the target genes based on their expression patterns. (This clustering can be performed in many different ways. In our implementation below we simply used a straightforward K-means method.) Then, for each cluster, our goal is to identify a set of regulators that is shared among the entire set of genes in the cluster, while still allowing for differences among the regulation of individual genes. Achieving this type of information sharing in the context of regularized linear modeling (4) however, requires some novel technical developments.

In [17] we recently developed a novel convex Bayesian network structure learning approach based on introducing a set of auxiliary indicator variables to control global feature selection. Adapting this idea to the current context, we propose to use a global regularization scheme on auxiliary selection variables to help identify the common candidate regulators among a group of target genes with similar expression profiles. Given that there is much more data available for sets of similar genes, as opposed to individual genes, we hope that the common regulators can be more accurately identified.

Specifically, given a set of target genes $Y = \{\mathbf{y}_1, ..., \mathbf{y}_m\}$, we would like to identify a common set of regulators from the set of candidates $X = \{\mathbf{x}_1, ..., \mathbf{x}_l\}$. Define a set of indicator variables $\boldsymbol{\eta} = \{\eta_1, ..., \eta_l\}^\top$, corresponding to the candidate set $X = \{\mathbf{x}_1, ..., \mathbf{x}_l\}$, such that each $\eta_i \in \{0, 1\}$ indicates whether a regulator X_i is selected as an active regulator. Let $N = \text{diag}(\boldsymbol{\eta})$. Then, we can form a globally regularized version of the minimization problem (4) by introducing the selection variables $\boldsymbol{\eta}$ and adding a new global regularization term on these variables:

$$\min_{\boldsymbol{\eta} \in \{0,1\}^n} \min_{\mathbf{w}} \sum_j \left(\|\Phi N \mathbf{w}_j - \tilde{\mathbf{y}}_j\|_2^2 + \alpha \|\mathbf{w}_j\|_1 \right) + \lambda \mathbf{u}^\top \boldsymbol{\eta}, \tag{5}$$

where $\mathbf{u}$ is a positive weight vector that allows one to incorporate prior knowledge about the importance of each global feature. Although we simply set this vector to 1 in our later experiments, it will be very useful whenever prior knowledge is available. Note that the global regularization term $\lambda \mathbf{u}^\top \boldsymbol{\eta}$ is in fact an L0 norm regularizer, which will automatically force a sparse solution that selects only a small set of global features for the set of target genes in a cluster. Nevertheless, the local L1 norm regularizer, $\alpha \|\mathbf{w}_j\|_1$, will still make individual choices of regulators for each specific target gene; choosing these regulators from the globally selected features identified by $\boldsymbol{\eta}$. Therefore, if the target genes in a cluster share some common regulators, the global feature selection process will be very helpful to pick them out, while the ability to individually model the regulation of each gene has not been diminished.

2.5 Optimization Procedure

Equation (5) encodes a *min-min* integer optimization problem. Unfortunately, integer optimization problems of this form are generally NP-hard. To attempt to

solve the problem efficiently, we first relax it into an optimization over continuous variables, by relaxing each $\eta_i \in \{0,1\}$ to be continuous $\eta_i \in [0,1]$. This leads to solve the following relaxed *min-min* optimization:

$$\min_{\boldsymbol{\eta}} \min_{\mathbf{w}} \sum_j \left(\|\Phi N \mathbf{w}_j - \tilde{\mathbf{y}}_j\|_2^2 + \alpha \|\mathbf{w}_j\|_1 \right) + \lambda \mathbf{u}^\top \boldsymbol{\eta}$$
$$s.t. \quad 0 \leq \boldsymbol{\eta} \leq 1. \tag{6}$$

In fact, this formulation has relaxed the original L0 norm regularizer over $\boldsymbol{\eta}$ into a L1 norm regularizer. In this way we maintain feature selection ability, while gaining computational efficiency.

In our implementation below, we conduct the optimization in two alternating steps: $\min_{\mathbf{w}}$ and $\min_{\boldsymbol{\eta}}$. Each $\min_{\mathbf{w}}$ step is simply a minimization of least squares regression error with L1 norm regularization, which can be implemented as a quadratic program [18], or by using a fast grafting algorithm [16]. For the $\min_{\boldsymbol{\eta}}$ step, we use a quasi-Newton BFGS method to perform the optimization [19].

3 Experiments and Results

We conducted experiments on both synthetic and real cell cycle data to evaluate our approach. In particular, we compared our global regularization approach to the standard independent local predication approach, and a prototype based linear regression method adapted from [20]. Synthetic experiments are useful to gauge the potential effectiveness of the approach under controlled conditions where the ground truth is available. Once the intuitive behavior of the technique is understood, we then apply the method to inferring the structure of the regulatory network of the yeast cell cycle.

In our experiments, we assume all transcription regulations work through activators, instead of inhibitors; that is, we assume the $\mathbf{w}$ parameters are non-negative in the linear regressions. Also, to keep the $\mathbf{w}$ parameters from becoming too small and causing a threshold selection problem, we included the additional constraint $\|\mathbf{w}_j\|_1 \geq 1$ in the three linear regression algorithms.

3.1 Experiments on Synthetic Data

For the synthetic experiments, we set up a small system to simulate a cell cycle process controlled by a small number of critical transcription factors (TFs). We defined 10 TFs that regulated the expression levels of 212 genes in 4 phases of a synthetic cell cycle. These 10 TFs were divided into 4 regulatory groups, with 3, 2, 3, and 2 TFs in each group respectively. Each group of TFs was associated with a specific phase of the cell cycle, and regulated the expression of 53 genes, as well as the TFs in the next phase of the the cycle. In our setting, we assumed that one gene (including the TFs themselves) can be regulated by either one TF or a combination of two TFs. We generated the expression data by first simulating ideal expression levels for the TFs in a selected phase for two complete cell

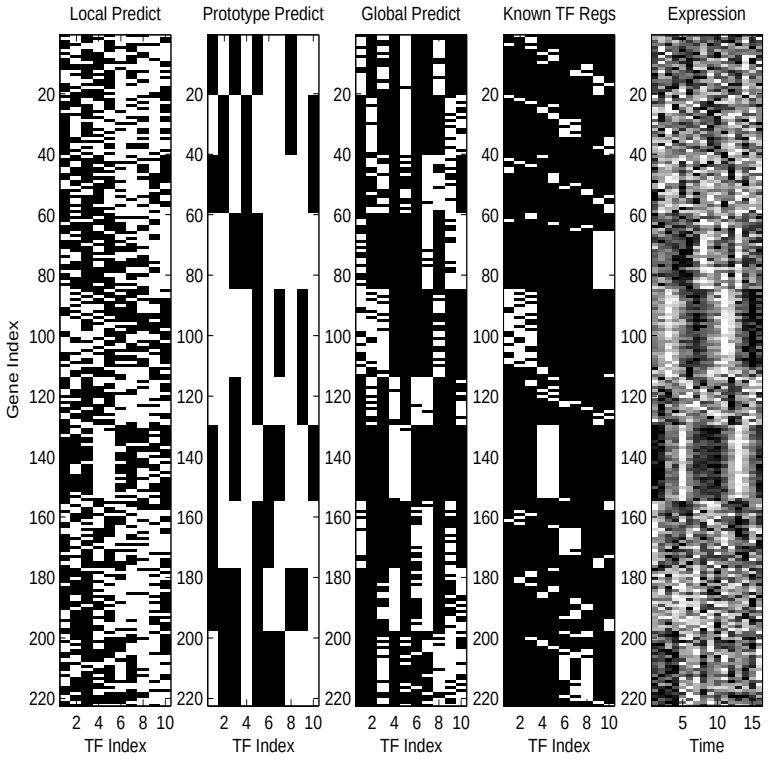
cycles, totaling 16 time steps. Then we generated the expression profiles of the genes (or TFs) in the next phase by a 2 time step delayed response from the combination ("and") of m ($m \leq 2$) randomly selected TFs in their previous phase, plus Gaussian noise. Repeating this procedure for all the phases in the cycle in turn, we generated synthetic time-series profiles for the entire set of TFs and genes.

Both our global regularization approach and the prototype based method require the genes to first be clustered based on their expression profiles. Although the number of clusters used has a minor effect on the performance of both algorithms, the impact is not significant provided that the cluster number is not extreme (neither extremely big nor extremely small). For our synthetic experiments, we simply choose to use 10 as the number of clusters.

Column 5 in Figure 1 shows the expression profiles for the genes and TFs after their profiles have been clustered into 10 groups. We then learn the regulators for the genes in each group, using our globally regularized linear regression to encourage genes in the same group to share parents. We compared the results of the global approach to both the standard "local" approach of learning the parent regulators for each gene separately, and the prototype based approach of forcing all the genes in one group to have the exactly same set of parents. The comparison algorithms serve as controls at the two opposite extremes. We used the same L1 regularized method for parent selection in all of the algorithms. After obtaining the $\mathbf{w}$ parameters from each algorithm, all the parents indicated by $\mathbf{w} > 10^{-5}$ are determined as predicted regulators for the corresponding genes. For a fair comparison, the regularization parameters (α and λ) were chosen to yield the highest F-measure values in each case.

Columns 1–3 in Figure 1 show the regulator prediction results for the three algorithms respectively; comparing them with the true regulation information in Column 4. The x-axis for each column indicates the candidate TFs from which a subset is selected as the set of regulators for each gene. The y-axis for each column indexes the individual target genes. Each row plots the predicted regulators for each gene based on the corresponding $\mathbf{w}$ parameters for that gene, where white indicates a large value (indicating a regulator), while dark indicates a value close to 0 (indicating no regulation).

The table in Figure 1 compares the performance of the three algorithms. The *precision* score measures true positive predictions (tp) divided by true positives plus false positive predictions (fp). That is, $precision = tp/(tp + fp)$. Similarly, *recall* score is measured in terms of the number of false negative predictions (fn), and is given by $recall = tp/(tp + fn)$. *F-measure* is a standard combination of both precision (p) and recall (r), given by $F\text{-}measure = 2\,p\;r/(p + r)$. The *accuracy* score measures the proportion of the correct predictions. That is, $accuracy = (tp + tn)/(tp + tn + fp + fn)$. Here we can see that the global regularization approach greatly outperforms both the local regularization and prototype based methods with respect to both accuracy and F-measure. The local predication method is not able to effectively identify the true regulators due to the noise in the data and the limited number of time points. The proto-



Performance comparison	Local regularization	Prototype method	Global regularization
accuracy (%)	57.6	47.2	73.0
precision (%)	21.4	18.1	30.0
recall (%)	71.5	75.0	63.8
F-measure	33.0	29.2	40.8

Fig. 1. Results on synthetic data. Rows denote target genes in the synthetic experiment. Columns denote candidate regulators (transcription factors). A white cell denotes a large weight ($w_{ij} > 10^{-5}$) connecting a TF j to a target gene i in the estimated linear model, indicating that j is inferred to regulate i. A black cell denotes a small weight ($w_{ij} \leq 10^{-5}$), indicating that j is not inferred to regulate i. Column 1: local prediction output. Column 2: prototype prediction output. Column 3: global prediction output. Column 4: ground truth regulatory relationships. Column 5: expression level data used as input.

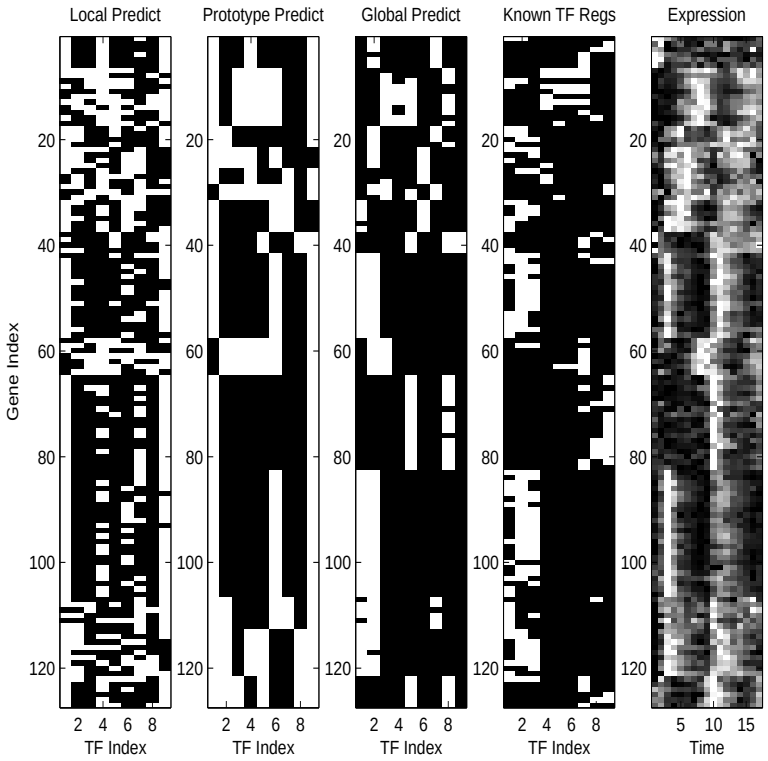
type base method also has difficulty identifying correct regulatory relationships, and tends to choose too many parents for each gene. The reason for this is clear however. Since the prototype method is forced to choose a single set of regulators for controlling a large set of genes, it naturally chooses the union of the prospective regulators for each gene, leading to subsequently low precision and accuracy. Thus, the prototype approach depends heavily on having a more refined and accurate set of clusters from which it can make accurate regulatory inferences, but an accurate clustering is very hard to achieve in practice. Figure 1 shows, on the other hand, that the global regularization approach can effectively remove irrelevant candidate TFs by sharing co-regulation information within a group, while simultaneously reducing the number of spurious regulators being inferred by allowing individual differences between genes in a given cluster. The overall result is a much more accurate (albeit far from perfect) recovery of the underlying regulatory structure.

The main question that remains is whether the higher quality inference on this synthetic model leads to improved results on real gene expression data, which we consider next.

3.2 Experiments on Real Data

Gene expression microarray data for the yeast cell cycle typically contains more than 6000 genes, while only a subset of these genes are cell cycle regulated. It is known there are 9 important transcription factors (TFs) that regulate the cell cycle process [21], namely: SWI4, SWI6, MPB1, FKH1, FKH2, NDD1, MCM1, ACE2 and SWI5. Since a lot of gene regulatory relationships have already been identified for yeast, this model is commonly used to evaluate learning approaches that attempt to infer gene regulatory networks from data. Here we use Cho et al.'s data [14], and focus on the task of identifying the subset of regulators from the 9 candidate TFs, for each yeast gene that is cell cycle regulated. To clearly evaluate our approach, we chose a subset of 267 cell cycle regulated genes from the Cho et al. data [14], while we could obtain confirmed regulatory relationships from the previous literature [21, 22], or could obtain potential regulation relationships from existing binding data [21] for 127 genes among them. We rescaled the expression data to values between 0 and 1, and then clustered the genes into 15 clusters using K-means. (In the images shown in Figure 2, the genes are grouped vertically into the clusters. The number of clusters is chosen by using visual judgment to achieve a smooth clustering effect.) Finally, we tested our algorithms on each cluster. As in the synthetic experiments, after obtaining the $\mathbf{w}$ parameters from each algorithm, all the parents indicated by $\mathbf{w} > 10^{-5}$ are determined as predicted regulators for the corresponding genes. For a fair comparison, the regularization parameters (α and λ) were chosen to yield the highest F-measure values in each case.

Since the regulatory mechanisms are still not known for a portion of the 267 genes, we therefore can only evaluate the results over the 127 genes for which regulatory relationships are presumed known. Figure 2 shows the prediction results



Performance comparison	Local regularization	Prototype method	Global regularization
accuracy (%)	57.8	55.4	73.9
precision (%)	22.3	21.2	35.7
recall (%)	47.5	48.0	43.4
F-measure	30.4	29.4	39.2

Fig. 2. Results on the subset of the real gene expression data from [14], restricted to genes where TF-based regulation information is known or can be inferred from other sources [21, 22]. Rows denote target genes in the synthetic experiment. Columns denote candidate regulators (transcription factors). A white cell denotes a large weight ($w_{ij} > 10^{-5}$) connecting a TF j to a target gene i in the estimated linear model, indicating that j is inferred to regulate i. A black cell denotes a small weight ($w_{ij} \leq 10^{-5}$), indicating that j is not inferred to regulate i. Column 1: local prediction output. Column 2: prototype prediction output. Column 3: global prediction output. Column 4: ground truth regulatory relationships. Column 5: expression level data used as input.

on 127 genes for all the three algorithms: locally regularized prediction, prototype based prediction, and globally regularized prediction. The images compare the performance of the three methods on inferring regulators from among the 9 candidate TFs, and shows how they related to the known TF-based regulatory relationships. These results show that the globally regularized approach can significantly improve the quality of both the standard locally regularized approach and the prototype based approach adapted from [20]. As in the synthetic case, the globally regularized approach has the ability to share regulatory information between genes within a cluster, leading to better noise robustness than the local approach. Here too, the global technique also overcomes the problem of being overly dependent on clustering quality, like the prototype approach, by allowing regulation differences with a cluster. For example, in Figure 2, in the group of genes indexed between 42-58, one can see that a large set of the errors produced by the standard independent approach (Column 1) have been corrected by sharing parent information throughout the cluster (Column 3). The global regularizer correctly recognizes that this set of late-G1 genes is regulated by a subset of SWI4/SWI6 and MBP1/SWI6. Although some local errors remain in this region (and elsewhere), clearly the overall quality of the parent prediction has been improved substantially in the global method. For these genes, the prototype based method (Column 2) recognizes two additional parents, perhaps due to noise.

Overall, the prediction quality achieved by these methods on this data is still somewhat limited, but has improved significantly over the past few years, and in some sense is remarkable given the noise exhibited in the expression profiles (Column 5).

4 Conclusions

In this paper, we have proposed a new globally regularized risk minimization objective for learning regulatory networks from gene expression data. Exploiting the assumption that genes with similar expression patterns are likely to be co-regulated, our approach first clusters the genes, and then learns the regulatory relationships by encouraging genes with similar expression patterns to share regulators. Our experimental results on both synthetic data and real cell cycle data show that this new approach is more effective at identifying important (transcription factor based) regulatory mechanisms than the standard independent approach, and a prototype based approach.

Thus far, we have only considered using gene expression data in the learning process. Further prediction improvements are likely to come from incorporating further sources of biologically relevant data, such as binding information [21], or other forms of prior knowledge beyond the co-regulation assumption made here. These informations can be nicely incorporated into our global risk minimization approach by using the $\mathbf{u}$ parameter vector. Moreover, as an effective feature selection strategy, it might be useful to extend this approach resolving other feature selection bioinformatics problems.

Acknowledgments

Research supported by NSERC, MITACS, CFI, the Alberta Ingenuity Centre for Machine Learning, and the Canada Research Chair program. We thank the anonymous referees for their helpful comments.

References

1. D'Haeseleer, P., Wen, X., Fuhrman, S., Somogyi, R.: Linear modeling of mrna expression levels during cns development and injury. Pac. Symp. Biocomput. (1999) 41–52
2. Chen, K., Wang, T., Tseng, H., Huang, C., Kao, C.: A stochastic differential equation model for quantifying transcriptional regulatory network in saccharomyces cerevisiae. Bioinformatics **21** (2005) 2883–2890
3. De Jong, H., Gouze, J., Hernandez, C., Page, M., Sari, T., Geiselmann, J.: Qualitative simulation of genetic regulatory networks using piecewise-linear models. Bull. Math. Biol. **66** (2004) 301–340
4. De Hoon, M., Imoto, S., Kobayashi, K., Ogasawara, N., Miyano, S.: Inferring gene regulatory networks from time-ordered gene expression data of bacillussubtilis using differential equations. Pac. Symp. Biocomput. (2003) 17–28
5. Li, F., Yang, Y.: Recovering genetic regulatory networks from micro-array data and location analysis data. Genome Informatics **15** (2004) 131–140
6. Hartemink, A., Gifford, D., Jaakkola, T., Young, R.: Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. Pac. Symp. Biocomput. (2001) 422–433
7. Yu, J., Smith, V., Wang, P., Hartemink, A., Jarvis, E.: Advances to Bayesian network inference for generating causal networks from observational biological data. Bioinformatics **20** (2004) 3594–3603
8. Wang, S.: Reconstructing genetic networks from time ordered gene expression data using Bayesian method with global search algorithm. J. Bioinform. Comput. Biol. **2** (2004) 441–458
9. Chen, X., Anantha, G., Wang, X.: An effective structure learning method for constructing gene networks. Bioinformatics **22** (2006) 1367–1374
10. Segal, E., Pe'er, D., Regev, A., Koller, D., Friedman, N.: Learning module networks. J. Mach. Learn. Res. **6** (2005) 557–588
11. Bernard, A., Hartemink, A.: Informative structure priors: joint learning of dynamic regulatory networks from multiple types of data. Pac. Symp. Biocomput. (2005) 459–470
12. Zou, M., Conzen, S.: A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. Bioinformatics **21** (2005) 71–79
13. D'Haeseleer, P., Wen, X., Fuhrman, S., Somogyi, R.: Genetic network inference: from co-expression clustering to reverse engineering. Bioinformatics **16** (2000) 707–726
14. Cho, R.J., Campbell, M.J., Winzeler, E.A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Landsman, D., Lockhart, D.J., Davis, R.W.: A genome-wide transcriptional analysis of the mitotic cell cycle. Mol. Cell. **2** (1998) 65–73
15. Ng, A.: Feature selection, L1 vs L2 regularization, and rotational invariance. In: International Conf. on Mach. Learn. (ICML). (2004)

16. Simon, P., Kevin, L., James, T.: Grafting: Fast, incremental feature selection by gradient descent in function space. J. Mach. Learn. Res. **3** (2003) 1333–1356
17. Guo, Y., Schuurmans, D.: Convex structure learning for Bayesian networks: Polynomial feature selection and approximate ordering. In: Conf. on Uncertainty in Artif. Intell. (UAI). (2006)
18. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge Univ. Press (2004)
19. Bertsekas, D.: Nonlinear Optimization. Athena Scientific (1995)
20. van Someren, E., Wessels, L., Reinders, M.: Linear modeling of genetic networks from experimental data. Intelligent Systems for Molecular Biology (ISMB 2000) (2000) 355–366
21. Simon, I., Barnett, J., Hannett, N., Harbison, C., Rinaldi, N., Volkert, T., Wyrick, J.J. Zeitlinger, J., Gifford, D., Jaakkola, T., Young, R.: Serial regulation of transcriptional regulators in the yeast cell cycle. Cell **106** (2001) 697–708
22. Iyer, V.R., Horak, C.E., Scafe, C.S., Botstein, D., Snyder, M., Brown, P.O.: Genomic binding sites of the yeast cell-cycle transcription factors sbf and mbf. Nature **409** (2001) 533–8