KERNEL-BASED GENE REGULATORY NETWORK INFERENCE

Yi Shi *
Department of Computing Science, University of Alberta,
Edmonton, Alberta, T6G2E8, Canada
Email: ys3@cs.ualberta.ca

Yuhong Guo
Department of Computer and Information Sciences, Temple University,
Philadelphia, Pennsylvania 19122, United States
Email: yuhong@temple.edu

Guohui Lin, Dale Schuurmans
Department of Computing Science, University of Alberta,
Edmonton, Alberta, T6G2E8, Canada
Email: guohui, dale@cs.ualberta.ca

We propose a kernel-based method for inferring regulatory networks from gene expression data that exploits several important factors previously neglected in the literature, including expression clustering, nonlinear regulator-gene relationships, variable time lags and gene competition. In particular, our approach infers regulatory relationships by encouraging genes with similar expression patterns to share common regulators. Furthermore, since regulator-gene expression relationships are not typically linear but instead obey a wider class of canalyzing relationships, we map the relations between transcription factors to an implicit higher dimensional space that is able to model more complex interactions. Such a kernel-based approach avoids explicit enumeration of canalyzing regulations while allowing nonlinear relations between transcription factors to be discovered. Third, to address the problem of varying regulation time lags, we exploit cubic spline interpolation to identify more precise lag times from discretely sampled expression levels, leading to more robust inference of regulatory delays. Finally, we model competition between genes; an effect that has not been explicitly modeled by previous methods. The combination of these extensions leads to more accurate inference of regulator-gene causal relations from gene expression data.

Keywords: Gene regulatory networks, cubic spline interpolation, sparse regression, feature selection, kernels, gene competition

1. INTRODUCTION

Genes play a central role in governing biological system function. Rather than acting individually, genes and their products cooperate to form dynamic regulatory networks. Revealing and understanding these complex networks of interaction is critical to understanding life at a fundamental systems level, enabling advances in systems biology, drug design, health care, and so on. With the emergence and development of high-throughput gene profiling technology and ChIP-on-chip technology, gene regulatory network research has greatly expanded, but reliable gene regulatory networks are still being sought in the research community. The goal of such research is to discover the causal control relationships between genes, leading to a fundamental understanding of how biological processes are coordinated in the cells.

Various computational approaches have been proposed in the literature for inferring gene regulatory networks from expression data. Most approaches have used linear models to express dependence between time series profiles. For example, D'Haeseleer et al. [5] proposed a simple linear model, while De Jong et al. [10] and Chen et al. [2] studied linear differential equations for modeling gene regulatory networks. Unfortunately, these methods suffer from over-fitting, since the number of parameters fit is proportional to the size of the data itself. Other linear approaches attempt to exploit sparse-

*Corresponding author.

ness of the regulatory relationships between genes to counteract the risk of over-fitting. These models employ the idea that any one gene is regulated by a small subset of the other genes. De Hoon et al. [8] proposed to use Akaike's Information Criterion (AIC) to determine the nonzero coefficients. Similarly, Li and Yang [16] used L_1 regularization to select features of the linear parent set. Recently, Zhang et al. [28] proposed a new multiple linear regression model that makes use of the scale-free property of a real biological network. After using this scale-free property and some appropriate statistical tests, redundant interactions among the genes were removed, then a model constructed by minimizing the gap between the observed and predicted data. Kabir et al. [12] proposed a new linear time-variant model where self-adaptive differential evolution, a versatile and robust evolutionary algorithm, is used as the learning paradigm.

Another popular approach in gene regulatory network induction is to infer an underlying (dynamic) Bayesian network structure that accounts for the causal relationships between gene expression levels. Bayesian networks provide a graphical representation of the causal relationships between a set of variables, yielding a well formed probabilistic framework for representing and inferring probabilistic relationships. Generally, there are two approaches in learning the structure of a Bayesian network from data. The first approach is 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. The second approach is a constraint-based approach where conditional independence tests are used to determine whether a direct causal relationship should be postulated between two variables. In the literature, many variants of these techniques have been applied to gene regulatory network inference, including search-based approaches [7; 26; 25], information-theoretic approaches [3], parameterizing based approaches [21], and conventional dynamic Bayesian network learning approaches [1; 29]. Recently, Li et al. [15] proposed a network-based empirical Bayes method for analyzing genomic data in the framework of linear models, where the dependency of genes is modeled by a discrete Markov random field defined on a predefined biological network.

Although the above approaches have achieved some promising results, their effectiveness has been severely constrained by the limited amount of data available relative to the large number of parameters estimated (i.e. the distinct parameters used to predict the expression level of each gene given other genes). This difficulty appears inherent to the task, since only a limited amount of background knowledge and biologically relevant assumptions are normally applied. Unfortunately, orders of magnitude more data would be required for such "knowledge free" approaches to yield accurate models.

Nearly all proposed approaches using either linear modeling or Bayesian network structure learning have a common shortcoming in that they attempt to determine the regulation structure for each target gene independently, while 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 [5]. A few investigators, such as [24], have previously proposed to group genes with similar expression profiles into a single prototypical "gene", then model the relations between prototypical genes instead of modeling the genes individually. However, this is a somewhat oversimplified approach that ultimately ignores the individual differences between genes in the same group, and establishes a particularly high requirement on the clustering step. In previous work [6], Guo and Schuurmans attempted to employ biologically significant knowledge about co-regulation to improve the inference of the underlying gene regulatory network from expression data. The novelty of this approach was to first cluster the genes based on their time series expression profiles, then minimize a loss determined on a set of global indicator variables associated with the common set of possible regulatory variables. The performance of this approach on both synthetic data and the cell cycle time-series gene expression data of [4] was quite promising, showing that important transcription factors (TFs) in the cell cycle genes could be identified more accurately. Unfortunately, this previous work was based on using standard linear models and a very crude form of discrete time lag inference.

In this paper, we extend the previous work in

four key respects. First, we replace linear model inference with a more flexible kernel based approach. Second, we use spline interpolation to more accurately infer regulator-gene time lags. Third, we explicitly take into account gene competition for regulator mRNA within a gene cluster. Finally, we also investigate the use of the cross entropy loss as an alternative to the standard least squares in the minimization objective. The remainder of the paper is organized as follows. Section 2 discusses the motivations of the above four approaches and presents the methods in detail. Section 3 then presents the experimental results on a Yeast dataset, which is followed by a discussion and conclusions in Section 4.

2. METHOD

The core of our work is to pursue a kernel-based approach rather than a simple linear regression model for regulatory network inference, so that non-linear relations among transcription factors can be considered. Our goal is to allow the model to consider non-linear interactions between regulators to allow more accurate inference of TF-gene causal relationships. Note that a linear model implicitly assumes that multiple transcription factors independently regulate target genes, but this is not an accurate model of the regulatory relationship when multiple transcription factors interact through complexes. We therefore propose to map the linear relations between transcription factors to more complex relations in high dimensional spaces via kernelization. Furthermore, to enhance the accuracy of the approach we adopt three further extensions over previous work: first, we introduce spline interpolation to more accurately infer TF-gene time lags; second, we model gene competition within gene clusters; and finally we briefly consider using alternative loss models to the standard least squares objective used in previous research.

2.1. Clustered Linear Model [6]

We first re-introduce the clustered linear model described in previous work [6]. 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, and each row corresponds to a single time point measured over a set of t. For each gene, one wants to identify which other genes measured in Y are likely to be it regulators. The fundamental hypothesis we would like to follow in the present work 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.

In [6], a simple linear prediction approach was developed. Assume that for a target expression profile 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. The potential regulators could be determined by solving for the combination weights of the regulator profiles that best reconstruct the target profile

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

where the $k \times 1$ vector of combination weights w_j describes how much each of the k regulator genes in X_j contribute to best fit the target expression levels y_j, and the quality of the fit is assessed by the residual error in (1).

However, since the set of candidate regulators for a given gene is usually much larger than the number of time points, a large set of combination weights $\mathbf{w}_j$ need to be inferred from a limited amount of data with such an approach. Moreover, only a tiny fraction of the candidate regulators are expected to be true regulators for any given gene, which means most of the weights should be set to 0 to indicate non-regulation. Therefore, [6] proposed to use an L_1 norm regularizer (rather than the traditional L_2 regularizer) to perform feature selection while inferring the linear model; a well known and effective method from the machine learning literature [18; 23; 6]. In particular, in this approach, one adds a penalty to the risk (the reconstruction objective) which encourages small values for $\mathbf{w}_j$:

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

where α is a parameter that trades off the influence of the risk with the regularizer. The regularizer encourages many of the weights to become exactly zero in the solution.

Considering that genes with similar expression patterns are likely co-regulated and involved in the same functional process, [6] proposed to first cluster the target genes based on their expression profiles. (A straightforward K-means method was used.) Then, for each cluster, one wants 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. For this, [6] introduced a set of auxiliary indicator variables to control global feature selection, and used 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, the idea is that the common regulators can be more accurately identified. Specifically, for a set of target genes $Y = \mathbf{y}_1, ..., \mathbf{y}_m$, we want 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 $\boldsymbol{\eta} = \eta_1, ..., \eta_l{}^T$, 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 $N = \Delta(\boldsymbol{\eta})$, where $\Delta(\boldsymbol{\eta})$ denotes putting the column vector $\boldsymbol{\eta}$ on the main diagonal of a square matrix. Then, one can form a globally regularized version of the minimization problem by introducing the selection variables $\boldsymbol{\eta}$ and adding a new global regularization term on these variables:

$$\min_{\boldsymbol{\eta}\in\{0,1\}^n} \min_{\mathbf{w}_j} \sum_j \left(\|XN\mathbf{w}_j - \tilde{\mathbf{y}}_j\|_2^2 + \alpha\|\mathbf{w}_j\|_1\right) + \lambda\mathbf{u}^T\boldsymbol{\eta}, \tag{3}$$

where $\mathbf{u}$ is a positive weight vector that allows one to incorporate prior knowledge about the importance of each global feature and is simply set to 1's. Since $\boldsymbol{\eta}$ is boolean, the global regularization term $\lambda\mathbf{u}^T\boldsymbol{\eta}$ acts as an L_0 norm regularizer which automatically force a sparse solution that selects only a small set of global features for the set of target genes in a cluster. The local L_1 norm regularizer $\alpha\|\mathbf{w}_j\|$ however, will still make individual choices of regulators for each specific target gene; choosing these regulators from the globally selected features identified by $\boldsymbol{\eta}$.

2.2. Kernelized Extension

The linear models used in (1)–(3) assume that, if a target gene (or a set of target genes) has multiple regulators, then each regulator can independently regulate the target genes, and their overall regulation effect is a weighted addition of their individual effects. Obviously, such an assumption would be false in cases where different transcription factors need to form a single protein complex and possibly cooperate with other such protein complexes to regulate the target genes. That is, linear models are quite limited in that they only consider *OR*-gate style regulation, while in nature regulation rules are more generally thought to obey the so-called canalyzing rules [13; 19; 20; 14]: a combination of *OR* and *AND* gate regulation. It follows that we need to consider more complex relations between transcription factors than the existing linear models considered by [6] and earlier work. However, enumerating all possible canalyzing rules over *OR* and *AND* gates is computationally impractical, therefore, so we propose a kernelized extension to the basic linear model. The goal is to map the complex relations between transcription factors to simple linear relations in high dimensional feature space. The use of a kernel-based model can side-step the problem of enumerating all canalyzing regulations, while discovering non-linear relations between transcription factors that better construct the target gene profile and hence more accurately infer the TF-gene causal relations.

First, to kernelize the L_2 norm term in (3) it suffices to consider the linear substitution $\mathbf{w}_j = NX^T\mathbf{b}_j$, where we re-express the weight optimization over $\mathbf{w}_j$ in terms of an expanded dual variable $\mathbf{b}_j$.

Second, to kernelized the L_1 norm term we consider the trick proposed in [17], and reformulate $\|\mathbf{w}_j\|_1$ using the relationship:

$$\begin{aligned}\|\mathbf{w}_j\|_1 &= \sum_{p=1}^{k} |w_{jp}| \\ &= \min_{\boldsymbol{\gamma}_j \geq 0} \frac{1}{2}\sum_{p=1}^{k}\left(\frac{w_{jp}^2}{\gamma_{jp}} + \gamma_{jp}\right) \\ &= \min_{\boldsymbol{\gamma}_j \geq 0} \frac{1}{2}\mathbf{w}_j^T G_j^{-1}\mathbf{w}_j + \boldsymbol{\gamma}_j^T\mathbf{1}, \end{aligned} \tag{4}$$

where we use the definition $G_j = \Delta(\boldsymbol{\gamma}_j)$. Now by

substituting $\mathbf{w}_j$ and (4) into (3), we obtain:

$$\min_{\boldsymbol{\eta}\in\{0,1\}^n} \min_{\mathbf{b}_j} \min_{\boldsymbol{\gamma}_j\geq 0} \sum_j \|XN^2X^T\mathbf{b}_j - \tilde{\mathbf{y}}_j\|_2^2 + \alpha\sum_j \left(\frac{1}{2}\mathbf{b}_j^T XNG_j^{-1}NX^T\mathbf{b}_j + \boldsymbol{\gamma}_j^T\mathbf{1}\right) + \mathbf{u}^T\boldsymbol{\eta}. \quad (5)$$

If we let $\Psi = XN$ and $\tilde{\Psi}^j = XNG_j^{-1/2}$ then (5) can be written as:

$$\min_{\boldsymbol{\eta}\in\{0,1\}^n} \min_{\mathbf{b}_j} \min_{\boldsymbol{\gamma}_j\geq 0} \sum_j \|\Psi\Psi^T\mathbf{b}_j - \tilde{\mathbf{y}}_j\|_2^2 + \alpha\sum_j \left(\frac{1}{2}\mathbf{b}_j^T\tilde{\Psi}^j\tilde{\Psi}^{jT}\mathbf{b}_j + \boldsymbol{\gamma}_j^T\mathbf{1}\right) + \mathbf{u}^T\boldsymbol{\eta}. \quad (6)$$

Now let $K = \Psi\Psi^T$ and $\tilde{K}^j = \tilde{\Psi}_j\tilde{\Psi}_j^T$, so that each element $K_{i\ell}$ in matrix K is the inner product of the i'th and ℓ'th row of Ψ, namely $K_{i\ell} = \Psi_{i:}\Psi_{\ell:}^T$. Similarly, $\tilde{K}_{i\ell}^j = \tilde{\Psi}_{i:}^j\tilde{\Psi}_{\ell:}^{j\,T}$. Because K and $\tilde{K}_j$ are symmetric and positive semi-definite matrices, according to the *Mercer*'s theorem, they correspond to dot products in a high dimensional implicit feature space. In our case, we use the widely adopted *RBF* kernel, whereby

$$K_{i\ell} = e^{\frac{-\|\Psi_{i:}-\Psi_{\ell:}\|^2}{2\sigma^2}} \quad (7)$$

$$\tilde{K}_{i\ell}^j = e^{\frac{-\|\tilde{\Psi}_{i:}^j-\tilde{\Psi}_{\ell:}^j\|^2}{2\sigma^2}}. \quad (8)$$

Substituting the two *RBF* kernel matrices into model (6), we obtain the kernelized version:

$$\min_{\boldsymbol{\eta}\in\{0,1\}^n} \min_{\mathbf{b}_j} \min_{\boldsymbol{\gamma}_j\geq 0} \sum_j \|K\mathbf{b}_j - \tilde{\mathbf{y}}_j\|_2^2 + \alpha\sum_j \left(\frac{1}{2}\mathbf{b}_j^T\tilde{K}^j\mathbf{b}_j + \boldsymbol{\gamma}_j^T\mathbf{1}\right) + \mathbf{u}^T\boldsymbol{\eta}. \quad (9)$$

In this formulation, notice that the local (within-cluster) feature selection is governed by $\boldsymbol{\gamma}_j$, whereas the global (between-cluster) feature selection is governed by $\boldsymbol{\eta}$. In the final solution, for target gene j and candidate regulator ℓ, a value of $\gamma_{j\ell} = 0$ or $\eta_\ell = 0$ indicates that regulator ℓ is not selected by target gene j.

Note that Equation (9) encodes a mixed-integer minimization 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 minimization:

$$\min_{\boldsymbol{\eta}} \min_{\mathbf{b}_j} \min_{\boldsymbol{\gamma}_j\geq 0} \sum_j \|K\mathbf{b}_j - \tilde{\mathbf{y}}_j\|_2^2 + \alpha\sum_j \left(\frac{1}{2}\mathbf{b}_j^T\tilde{K}^j\mathbf{b}_j + \boldsymbol{\gamma}_j^T\mathbf{1}\right) + \mathbf{u}^T\boldsymbol{\eta}$$

$$s.t. \quad 0 \leq \boldsymbol{\eta} \leq 1. \quad (10)$$

This formulation has actually relaxed the original L_0 norm regularizer over boolean $\boldsymbol{\eta}$ to an L_1 norm regularizer over continuous $\boldsymbol{\eta}$. In this way we maintain feature selection ability, while gaining computational efficiency.

In our implementation, we jointly minimize over all the $\boldsymbol{\eta}$, $\mathbf{b}_j$ and $\boldsymbol{\gamma}_j \geq 0$ in (10) using the Matlab optimization toolbox *fmincon*, where $\boldsymbol{\eta}$ appears implicitly in K through the appearance of Ψ in the formula (7), and both $\boldsymbol{\eta}$ and $\boldsymbol{\gamma}_j$ appear implicitly in $\tilde{K}^j$ through the appearance of Ψ^j in formula (8).

2.3. Continuous Time Lags

Notice that neither of the above models account for any time lag between the expression of a regulating gene and the expression of its downstream target. In fact, they implicitly assume regulation occurs instantaneously, which performs poorly at identifying regulatory relationships that exhibit delayed effects. To cope with this shortcoming, we previously proposed a simple time-shifting method [6], in which, for each candidate regulator measured in X_j, given by an $n \times 1$ vector x_{ij}, one first computes an optimal backward shift in time that best aligns x_{ij} individually with the target y_j:

$$s_{ij}^* = \arg\min_{s\in\{1,2,3\}} \|\mathbf{x}_{ij}(1,...,n-s) - \mathbf{y}_j(s+1,...,n)\|_2^2. \quad (11)$$

Repeating this for each candidate regulator profile in X_j yields a series of optimal time lags. We then reformulated the expression matrix X_j for the candidate regulators by applying the 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 $\mathbf{y}_j$ was 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 above approach has two major problems. First, it is unlikely that real time lags are aligned with the sampling period, and more typically occur within a time period. Previous studies [27; 11] have shown that continuous representation of discrete time series data has positive effects. Second, because each transcription factor may regulate multiple target genes, it is more reliable to consider the time lag between a TF and its entire set target genes rather than a single gene. That is, a common time lag can then be used between a TF and its entire target set. Considering that there is always noise in microarray measurements, the TF/gene-set time lag should be more robust than a single TF-gene time lag estimate. In practice, we first cluster the genes, then use cubic spline interpolation to represent all the gene profiles in continuous cubic polynomial functions. The optimal time lag between a TF $\mathbf{x}_{ij}$ and a target gene $\mathbf{y}_j$ is then given by:

$$s^*_{ij} = \underset{s}{\operatorname{argmin}} \int_{s+1}^{n-s} \sum_{p \in cluster(j)} |f(\mathbf{x}_{ij}) - f(\mathbf{y}_{jp})| \tag{12}$$

where s can be arbitrarily small time step (we use 1/10 of the original time period). We then use the same procedure to construct Φ_j for each target gene $\mathbf{y}_j$ with time point dense n inverse proportional to s.

2.4. Gene Competition

To the best of our knowledge, almost all previous research on inferring TF-gene regulations has focused on the TF-gene relations only, while the potential competition effects between genes that share common TFs are neglected. For example, if a gene y_1 and gene y_2 are co-activated by a common set of transcription factors X_1, then the expression profile of gene y_1 may not solely depend on the expression profile of X_1, but rather the expression profile of y_2 may dampen the expression profile of y_1. We therefore take gene competition effects into account in our linear model (3) since it is not obvious how to adopt this idea in the kernelized model.

Specifically, consider the formulation (3). For a cluster of k genes, the corresponding set of weights W consists of an $l \times k$ matrix, where each column indicates how much one of the k genes responds to the l candidate regulators, and each row indicates how much a regulator is selected by the genes. To model the gene competition effect, one can force the sum of each row of W to be smaller than a given threshold. To properly determine the values of the these thresholds requires prior knowledge, but we simply set them to be all 1's in our initial investigation.

2.5. Alternative Loss Minimization

To this point we have focused on training under an L_2 norm reconstruction loss. In our implementation, we also investigated the cross entropy loss, as it is commonly used as an alternative to the L_2 norm when output targets are bounded. For an observed vector $\mathbf{y}$ and an estimated vector $\hat{\mathbf{y}}$, the cross entropy loss is defined as:

$$loss_{CrsEtp}(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_i y_i \ln \hat{y}'_i + (1 - y_i) \ln (1 - \hat{y}'_i)$$

where $\hat{y}'_i = \frac{1}{1+e^{-\hat{y}_i}}$.

3. Experimental Results

We conducted experiments on real cell cycle data to evaluate our approach. In particular, we compared our kernel method (both L_2 loss and cross entropy loss) to the global regularization approach (with and without gene competition effect), the standard independent local predication approach, and a prototype based linear regression method adapted from [24]. We applied the methods to inferring the structure of the regulatory network of the yeast cell cycle. In our experiments, we assumed all transcription regulations work through activators rather than inhibitors; that is, we assumed the w parameters are nonnegative in the linear regressions.

Yeast 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 [22], namely: SWI4, SWI6, MPB1, FKH1, FKH2, NDD1, MCM1, ACE2 and SWI5. Because many gene regulatory relationships in Yeast have already been identified, it is a commonly used test-bed for evaluating regulatory

network inference methods. Here we use Cho et al.'s data [4], 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. We choose a subset of 127 cell cycle regulated genes from the Cho et al. data [4] to clearly evaluate our approach, where we could obtain confirmed regulatory relationships from the previous literature [22; 9], or could obtain potential regulation relationships from existing binding data [22]. We re-scaled the expression data to values between 0 and 1, then clustered the genes into 15 clusters using K-means. (In the images shown in Figure 1 and 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. After obtaining the w parameters from each algorithm, all the parent TFs indicated by values of $w > 10^{-5}$ were 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 1 and Figure 2 show the prediction results on 127 genes for all the six algorithms with and without applying the cubic spline interpolation for the time lag problem respectively: locally regularized prediction, prototype based prediction, globally regularized prediction without considering the within cluster gene competition effects, globally regularized prediction considering the within cluster gene competition effects, the kernel model prediction, and the kernel model prediction with cross entropy loss. The images compare the performance of the six methods on inferring regulators from among the 9 candidate TFs, and shows how they related to the known TF-based regulatory relationships.

Table 1 and Table 2 compare the performance of the six algorithms with and without applying the cubic spline interpolation for the time lag problem. 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 $recall = tp/(tp + fn)$. The F-measure is a standard combination of both precision (p) and recall (r), given by $F - measure = 2pr/(p + r)$. The accuracy score measures the proportion of the correct predictions. That is, $accuracy = (tp + tn)/(tp + tn + fp + fn)$.

These results show that the kernel approaches improve the quality of the regulation relation inference in general. The globally regularized approaches have the ability to share regulatory information between genes within a cluster, leading to better noise robustness than the local approach. The kernel approaches outperform the globally regularized approaches in terms of precision, recall and F-measure in table 1. For example, in Figure 1, in the group of genes indexed between 20-50, one can see that a large set of TFs that were not picked out by any other approaches are picked out by the kernel approach (Column 5). Considering the effect of within cluster gene competition seems not lead to significant improvement in our case. However, with appropriate upper bound threshold setting and applied in the kernel methods, we believe it would contribute to better prediction performance. The idea of using cubic spline interpolation to better address the time lag problem is effective, as the overall performance of all methods is improved.

Although some local errors remain in this region (and elsewhere), clearly the overall quality of the parent prediction has been improved substantially by the kernel method. 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.

4. Conclusion

In this paper, we have proposed a new kernelized version of 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, then learns the regulatory relationships by encouraging genes with similar expression patterns to share regulators. Considering that natural TF-gene

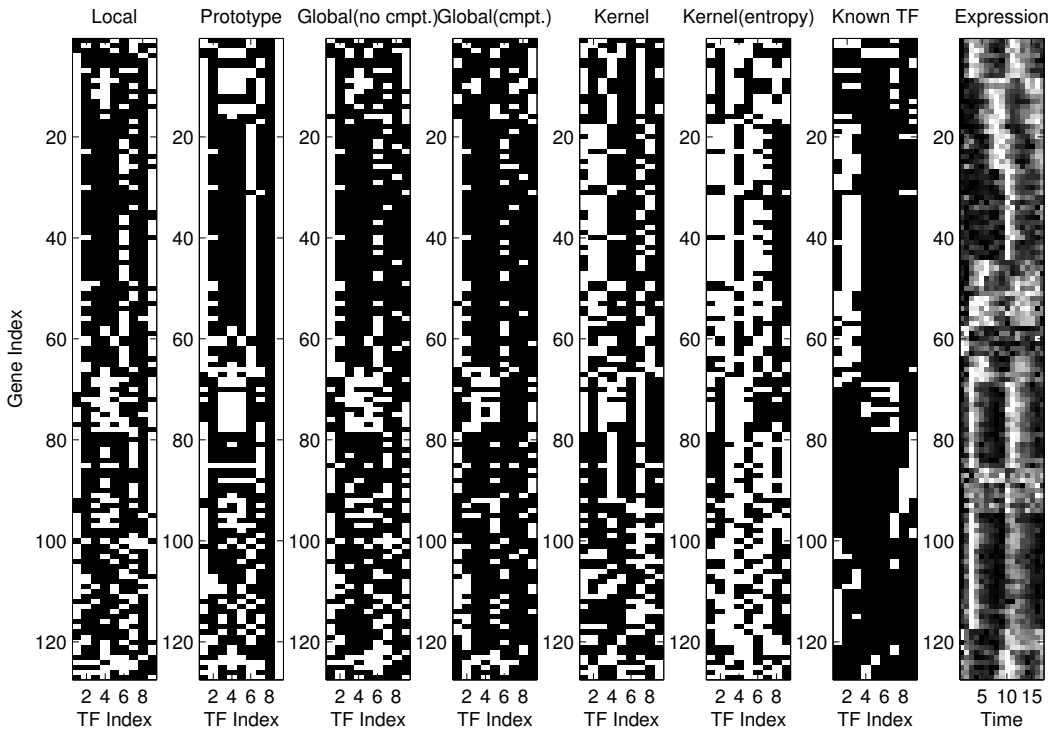

Fig. 1. Results after applying the cubic spline interpolation for regulation time lag problem on the subset of the real gene expression data from [4], restricted to genes where TF-based regulation information is known or can be inferred from other sources [22; 9]. Rows denote target genes in the synthetic 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 without gene competition output. Column 4: global prediction with gene competition output. Column 5: Kernel method prediction. Column 6: Kernel method prediction with cross entropy loss. Column 7: ground truth regulatory relationships. Column 8: expression level data used as input.

Performance comparison	Local	Prototype	Global (no compete)	Global (compete)	Kernel	Kernel (CE loss)
Accuracy(%)	59.9	55.7	68.3	**70.6**	66.9	54.1
Precision(%)	22.1	20.6	26.9	28.3	**31.4**	25.4
Recall(%)	42.5	45.2	37.1	33.9	60.2	**71.0**
F-measure	29.1	28.3	31.2	30.9	**41.3**	37.4

Performance comparison	Local	Prototype	Global (no compete)	Global (compete)	Kernel	Kernel (CE loss)
Accuracy(%)	57.5	55.0	**67.7**	67.6	49.7	54.6
Precision(%)	22.1	20.9	**29.7**	29.2	25.1	22.2
Recall(%)	47.5	47.5	48.9	47.5	**81.0**	53.8
F-measure	30.2	29.0	36.9	36.2	**38.4**	31.4

regulation rules are likely to obey canalyzing rules [13; 19; 20; 14], we proposed to kernelize the linear model to map the independent linear relations between transcription factors to more complex relations

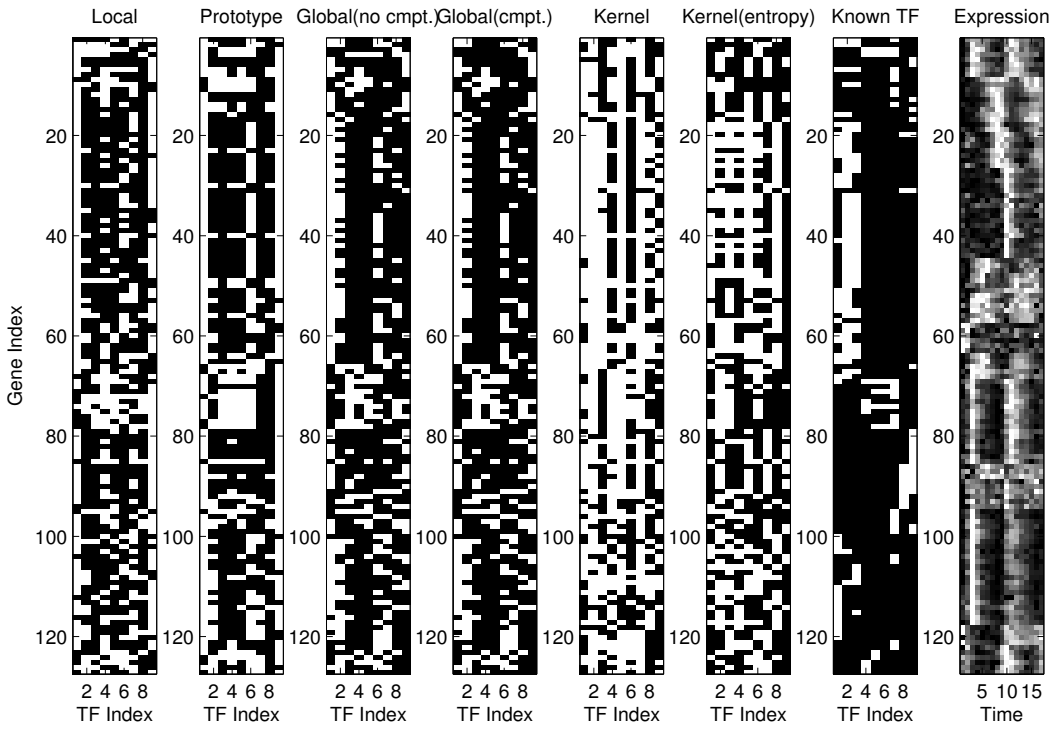

Fig. 2. Results without applying the cubic spline interpolation for regulation time lag problem on the subset of the real gene expression data from [4], restricted to genes where TF-based regulation information is known or can be inferred from other sources [22; 9]. Rows denote target genes in the synthetic 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 without gene competition output. Column 4: global prediction with gene competition output. Column 5: Kernel method prediction. Column 6: Kernel method prediction with cross entropy loss. Column 7: ground truth regulatory relationships. Column 8: expression level data used as input.

in high dimensional space. We conjecture that the kernelized model can avoid the problem of enumerating all the canalyzing regulations and can discover the non-linear relations between transcription factors to better construct the target gene profile and more accurately infer the TF-gene causal relations. To address the regulation time lag problem, we proposed to use cubic spline interpolation to extend discrete-time gene profiles to continuous-time profiles, allowing a higher resolution search for alignments between transcription factors and gene clusters. This makes the time lag searching more robust to noise. We also considered the within-cluster gene competition effect that is neglected by most gene regulatory network inference methods in the literature.

Our experimental results yeast cell cycle data show that the kernel approach is more effective at identifying important (transcription factor based) regulatory mechanisms than the standard independent approach, the prototype based approach, and the globally regularized 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 the within-cluster gene competition mRNA upper bound, binding information [22], or other forms of prior knowledge beyond the co-regulation assumption made here. Moreover, as an effective strategy, the kernel method combined with the L_1 norm feature selection might be extended to resolve other similar problems in bioinformatics area and other re-

search areas.

References

1. A. Bernard and A. Hartemink. Informative structure priors: joint learning of dynamic regulatory netwokrs from multiple types of data. *Pac. Symp. Biocomput.*, pages 459–470, 2005.
2. K.C. Chen, T.Y. Wang, H.H. Tseng, C.Y.F. Huang, and C.Y. Kao. A stochastic differential equation model for quantifying transcriptional regulatory network in saccharomyces cerevisiae. *Bioinformatics*, 21:2883–2890, 2005.
3. X. Chen, G. Anantha, and X. Wang. An effective structure learning method for constructing gene networks. *Bioinformatics*, 22:1367–1374, 2006.
4. R.J. Cho, M.J. Campbell, E.A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T.G. Wolfsberg, A.E. Gabrielian, D. Landsman, D.J. Lockhart, and R.W. Davis. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell*, 2:65–73, 1998.
5. P. D'Haeseleer, X. Wen, S. Fuhrman, and R. Somogyi. Linear modeling of mrna expression levels during cns development and injury. *Pac. Symp. Biocomput.*, pages 41–52, 1999.
6. Y. Guo and D. Schuurmans. Learning gene regulatory networks via globally regularized risk minimization. *RECOMB Satellite Workshop on Comparative Genomics (LNCS)*, 4751:83–95, 2007.
7. A. Hartemink, D. Gifford, T. Jaakkola, and R. Young. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Pac. Symp. Biocomput*, pages 422–444, 2001.
8. M.J.L. De Hoon, S. Imoto, K. Kobayashi, N. Ogasawara, and S. Miyano. Inferring gene regulatory networks from time-ordered gene expression data of bacillussubtilis using differential equations. *Pac. Symp. Biocomput.*, pages 17–28, 2003.
9. V.R. Iyer, C.E. Horak, C.S. Scafe, D. Botstein, M. Snyder, and P.O. Brown. Genomic binding sites of the yeast cell-cycle transcriptiona factors sbf and mbf. *Nature*, 409:533–538, 2001.
10. H. De Jong, J.L. Gouzé, C. Hernandez, M. Page, T. Sari, and J. Geiselmann. Qualitative simulation of genetic regulatory networks using piecewise-linear models. *Bull. Math. Biol.*, 66:301–340, 2004.
11. Z.B. Joseph. Analyzing time series gene expression data. *Bioinformatics*, 20(16):2493–2503, 2004.
12. M. Kabir, N. Noman, and H. Iba. Reverse engineering gene regulatory network from microarray data using linear time-variant model. *BMC Bioinformatics*, 11:(APBC) supp 1., 2010.
13. S. Kauffman. *The Origins of Order: Self-organization and Selection in Evolution*. Oxford University Press, 1993.
14. S. Kauffman, C. Peterson, B. Samuelsson, and C. Troein. Genetic networks with canalyzing boolean rules are always stable. *Procl Natl. Acad. Sci. U.S.A*, 101:17102–17107, 2004.
15. C. Li, Z. Wei, and H. Li. Network-based empirical bayes methods for linear models with applications to genomic data. *J. Biopharm. Stat.*, 20:209–222, 2010.
16. F. Li and Y. Yang. Recovering genetic regulatory networks from micro-array data and location analysis data. *Genome Informatics*, 15:131–140, 2004.
17. C.A. Micchelli and M. Pontil. Learning the kernel function via regularization. *J. Mach. Learn. Res.*, 6:1099–1125, 2006.
18. A. Ng. Feature selection, l1 vs l2 regularization, and rotational invariance. *International Conf. on Mach. Learn. (ICML)*, 69:78, 2004.
19. S. Nikolajewa, M. Friedel, and T. Wilhelm. Boolean networks with biologically relevant rules show ordered behavior. *Biosystems.*, 90:40–47, 2007.
20. U. Paul, V. Kaufman, and B. Drossel. Properties of attractors of canalyzing random boolean networks. *Phys. Rev. E. Stat. Nonlin. Soft. Matter. Phys.*, 73:026118, 2006.
21. E. Segal, D. Pe'er, A. Regev, D. Koller, and N. Friedman. Learning module netwokrs. *J. Mach. Learn. Res.*, 6:557–588, 2005.
22. I. Simon, J. Barnett, N. Hannett, C.T. Harbison, N.J. Rinaldi, T.L. Volkert, J.J. Wyrick, J. Zeitlinger, D.K. Gifford, T.S. Jaakkola, and R.A. Young. Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell*, 106:697–708, 2001.
23. P. Simon, L. Kevin, and T. James. Grafting: Fast, incremental feature selection by gradient descent in function space. *J. Mach. Learn. Res.*, 3:1333–1356, 2003.
24. E. Van Someren, L. Wessels, and M. Reinders. Learning modeling of genetic networks from experimental data. *Intelligient Systems for Molecular Biology (ISMB)*, pages 355–366, 2000.
25. S. Wang. Reconstructing genetic networks from time ordered gene expression data using Bayesian method with global search algorithm. *J. Bioinform. Comput. Biol.*, 2:441–458, 2004.
26. J. Yu, V. Smith, P. Wang, A. Hartemink, and E. Jarvis. Advances to Bayesian network inference for generating causual networks from observational biological data. *Bioinformatics*, 20:3594–3603, 2004.
27. Z.B.Joseph, G.K. Gerber, D.K. Gifford, T.S. Jaakkola, and I. Simon. Continuous representations of time-series gene expression data. *J. Comput. Biol.*, 10:341–356, 2003.
28. S.Q. Zhang, W.K. Ching, N.K. Tsing, H.Y. Leung, and D. Guo. A new multiple regression approach for the construction of genetic regulatory networks. *Artif Intell Med.*, 48:153–160, 2010.
29. M. Zou and S. Conzen. A new dynamic Bayesian network (DBN) approach for identifying gene regulatory netwokrs from time course microarray data. *Bioinformatics*, 21:71–79, 2005.