Learning Gene Regulatory Networks via Globally Regularized Risk Minimization

Yuhong Guo and Dale Schuurmans

Department of Computing Science
University of Alberta
{yuhong,dale}@cs.ualberta.ca

Abstract

Learning the structure of a gene regulatory network from time-series gene expression data is a significant challenge, but an important problem that has recently gained attention. Most approaches proposed in the literature to date—from linear modeling to Bayesian network structure learning—attempt to predict the regulators of each target gene individually, but fail to effectively share regulatory information between related genes. Modeling each gene independently does not make appropriate use of a limited data supply, since there are typically many more genes than time points sampled in any expression data set. In this paper, we propose a new globally regularized risk minimization approach to learn gene regulatory networks from gene expression data. Our approach first clusters genes according to their time-series expression profiles—identifying related groups of genes—and then exploits the assumption that genes with similar expression patterns are likely to be co-regulated. Given a clustering, we then develop a simple technique that exploits similarity information by encouraging similar genes to share common regulators. Our method is based on a new *global* feature selection approach, wherein potential regulators compete to explain entire groups of related genes. We formulate a new risk minimization objective based on this intuition and show how it can be minimized. 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.

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. (?) studied a straightforward linear model for this purpose; Chen et al. (?) and De Jong et al. (?) 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. (?) have proposed to use "Akaike's Information Criterion" (AIC) to determine the nonzero coefficients in the linear system. Similarly, Li & Yang (?) 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 (?; ?), information-theoretic approaches (?), parameter-tying based approaches (?), and conventional dynamic Bayesian network learning approaches (?; ?).

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 (?). 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. (?), 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 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 (?). 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. (?) 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* (?). 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 (?). 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.

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

[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.

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

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

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 (**??**).

2.2 Coping with Time Lags via Time Shifting

Unfortunately, the naive linear modeling approach (**??**) 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 (**??**) 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 $\mathbf{y}_j$

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

(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 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

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

2.3 Feature Selection via L1 Regularized Risk Minimization

Although the modified linear approach (**??**) 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 (?) 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 $\mathbf{w}_j$:

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

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 (?).

2.4 Regulation Sharing via Globally Regularized Risk Minimization

Simply resolving the minimization problem in (??) brings no advantages over the approaches proposed in the literature, since it suffers from the same problem of short of data points versus large number of genes. Fortunately, (?) recently proposed a novel convex Bayesian network structure learning approach, which introduced a set of global feature selection indication variables to control the global feature selection. Borrowing this global feature selection idea, we define a set of variables $\boldsymbol{\eta} = \{\eta_1, ..., \eta_n\}^\top$ corresponding with the gene set $\mathbf{z} = \{z_1, ..., z_n\}$, where each $\eta_i \in \{0, 1\}$ is an indicator variable. Thus if we have gene j as the target gene, then $y_j = z_j$, $\mathbf{x}_j = \{z_1, ..., z_{j-1}, z_{j+1}, ..., z_n\}$ and $\boldsymbol{\eta}_j = \{\eta_1, ..., \eta_{j-1}, \eta_{j+1}, ..., \eta_n\}$. Mapping the features back to the minimization problem (??), the new features we should use are $\mathbf{x}'_j = \mathbf{x}_j N_j$, where $N_j = \text{diag}(\boldsymbol{\eta}_j)$. Adding a new global regularization term on the feature selection variables, we obtain:

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

where $\mathbf{u}$ is a positive weight vector that provides a way to incorporate prior knowledges 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 knowledges are available. Note that the global regularization term $\lambda \mathbf{u}^\top \boldsymbol{\eta}$ is in fact a L0 norm regularizer, and it will automatically force a sparse solution that chooses only a small set of global features for all the target genes, while local feature selection based on the L1 norm can still make individual local choice for each specific target gene from the global chosen feature set. Therefore, if all or parts of the target genes share some common regulators, the global feature selection function will be very helpful to pick them out. As it has been observed that genes with similar expression pattern are usually co-regulated and involved in the same functional process, we choose to first cluster the target genes into groups of similar expression patterns, and then perform optimization (??) on each cluster separately (but attempting to encourage parent sharing within a cluster). We used a straightforward K-means technique to cluster gene expression profiles in our implementation.

2.5 Optimization Procedure

Equation (??) encodes an integer *min-min* optimization problem. To attempt to solve it efficiently, we first relax it to an optimization on continuous variables, by relaxing $\eta_j \in \{0, 1\}$

to continuous $\eta_j \in [0,1]$. This leads to solve the following *min-min* optimization:

$$\begin{aligned} \min_{\boldsymbol{\eta}} \quad & \min_{\mathbf{w}} \sum_j \left[\mathbf{w}_j^\top N_j \mathbf{x}_j^\top \mathbf{x}_j N_j \mathbf{w}_j - 2\mathbf{y}_j^\top \mathbf{x}_j N_j \mathbf{w}_j + \alpha \|\mathbf{w}_j\|_1 \right] + \lambda \mathbf{u}^\top \boldsymbol{\eta} && (6) \\ s.t. \quad & 0 \le \boldsymbol{\eta} \le 1. && (7) \end{aligned}$$

In fact, this step relaxes the original L0 norm regularizer into a L1 norm regularizer. Therefore we kept the feature selection capacity, while gaining computational efficiency.

We conduct the optimization in two alternative steps: $\min_{\mathbf{w}}$ and $\min_{\boldsymbol{\eta}}$. Each $\min_{\mathbf{w}}$ step is simply a minimization of least square regression error with L1 norm regularization. This can be implemented as a quadratic program (**?**), or one can use a fast grafting algorithm (**?**). For the $\min_{\boldsymbol{\eta}}$ step, we use a quasi-Newton BFGS method to perform the optimization (**?**).

3 Experiments and Results

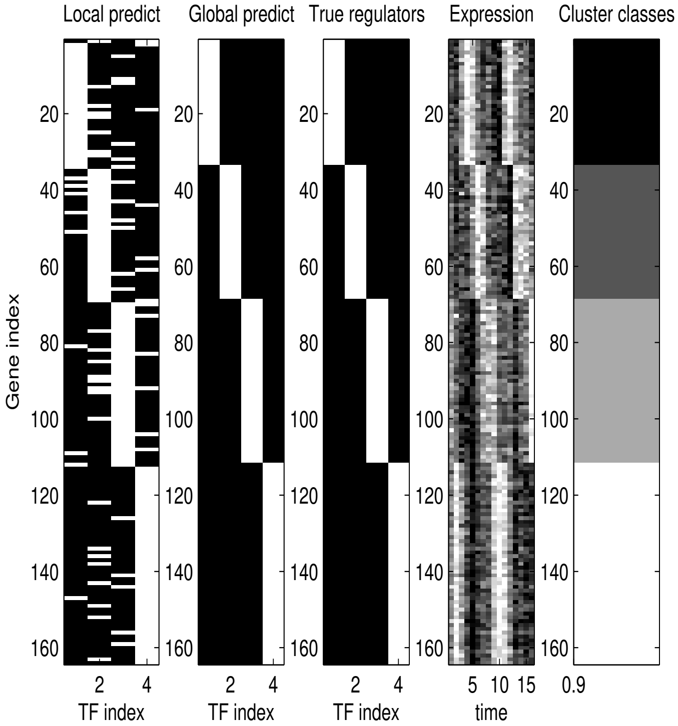
We conducted experiments on both synthetic and real cell cycle data to evaluate our approach. Synthetic experiments are useful to gauge the potential effectiveness of the approach under controlled conditions where the ground truth is available. Once the intuitive behavior of the technique is understood, we then apply the method to inferring the structure of the regulatory network of the yeast cell cycle.

3.1 Experiments on Synthetic Data

For the synthetic experiments, we set up a small system to simulate the cell cycle process. Figure 1 shows the results of an experiment on a synthetic system consisting of 164 genes, 4 of which are critical transcription factors, TF1-4. The remaining genes are divided into 4 groups, where each group are regulated by one of the TFs respectively. The 4 TFs are also be regulated by each other in a cyclic network. We generate the expression data by first simulating ideal expression levels for the 4 TFs for two complete cell cycles, totaling 16 time steps. Then for each remaining gene we generate its expression profile by a delayed response from a single parent TF, plus Gaussian noise. Repeating this procedure, we can generate synthetic time-series data.

Figure 1 shows the expression profiles for the 164 genes after their profiles have been clustered into 4 groups. We then learn the regulators for genes in each group, using the formulation above to encourage genes in the same group to share parents. This was compared to the standard approach of learning the parent regulators for each gene separately. Here, we used the same L1 regularized approach to learn parents for each gene separately. This serves as a control, since any performance differences are due solely to the novel parent sharing mechanism introduced in Section 2. For a fair comparison, the regularization parameters were chose to yield the highest F-measure values in each case.

The table in Figure 1 shows that without the global feature selection, many genes picked the wrong parent TF from the four potential parents because of the noise in the data. However, by using the global feature selection strategy, the accuracy is greatly improved through sharing information between similar genes. Clearly, the global regularization approach is capable of using shared information to improve the result of recognizing parents under co-regulation. The main question is whether this leads to improved results on real gene expression data, which we consider next.



Accuracy comparison	Local regularization	Global regularization
precision (%)	70.8	100
recall (%)	97.6	100
F-measure	82.1	100

Figure 1: Results on synthetic data. Column 1 in the figure indicates parents predicted by local regularization (predicting parents for each gene independently). Column 2 indicates parents predicted by the global regularization approach, which encourages parent sharing within a cluster of genes. Column 3 shows the true baseline parents for each gene. Column 4 shows the expression data. Column 5 shows the gene clusters discovered based on the expression data. In the table, *precision* measures true positive predictions (tp) divided by true positives plus false positive predictions (fp). That is, $precision = tp/(tp + fp)$. Similarly, *recall* can be measured in terms of the number of false negative predictions (fn), and is given by $recall = tp/(tp + fn)$. Finally, *F-measure* is a standard combination of both precision (p) and recall (r), given by $F\text{-}measure = 2\,p\,r/(p + r)$.

3.2 Experiments on Real Data

Yeast cell cycle gene expression microarray data typically contains more than 6000 genes, while only a subset of them are cell cycle regulated. It has been known there are 9 important TFs (transcription factors) that regulate the cell cycle process (**?**). The 9 TFs are: SWI4, SWI6, MPB1, FKH1, FKH2, NDD1, MCM1, ACE2 and SWI5. The yeast data contains a lot of previous identified gene regulatory relationships, therefore it is usually used to evaluate gene regulatory network learning approaches. Here we choose to use Cho et al.'s data, and our task is to identifying the regulators from the 9 TFs for each cell cycle gene (**?**). In order to better evaluate our approach, we choose a subset of 267 cell cycle regulated genes from Cho's data, while we can obtain the confirmed regulation relationships from previous literature (**?**; **?**) or the potential regulation relationships from the binding data (**?**) for 150 genes among them. We rescale the expression data and then cluster them into 9 clusters. Finally we can run our algorithm on each cluster. Figure 2 shows the prediction results on all 267 genes, showing both the locally regularized with the globally regularized predictions and how they compare to known TF-based regulatory relationships. Since the regulatory mechanisms are still not known for most genes, the third column in Figure 2 is quite sparse—making it difficult to assess the quality of the methods. Therefore, Figure 3 shows the same data, but restricted to the specific 150 genes (rows) that have known TF-based regulators.

Figure 3, including the table at the bottom, show that, again, the globally regularized approach can improve the quality of the standard locally regularized approach. The parent information shared from within the clustering structure has some capacity to correct errors, due to noise, made by the local regularizer. For example, in Figure 3, the group of genes indexed between 6-35, one can see that a large set of the errors produced by the standard approach (Column 1) have been corrected by sharing parent information throughout the cluster. 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 errors remain elsewhere, clearly the overall quality of the parent prediction has been improved substantially. The prediction quality achieved by these methods on this data is still somewhat limited, but has improved over the past few years, and in some sense is remarkable given the noise exhibited in the expression profiles (Column 4).

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.

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 (**?**), 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 vec-

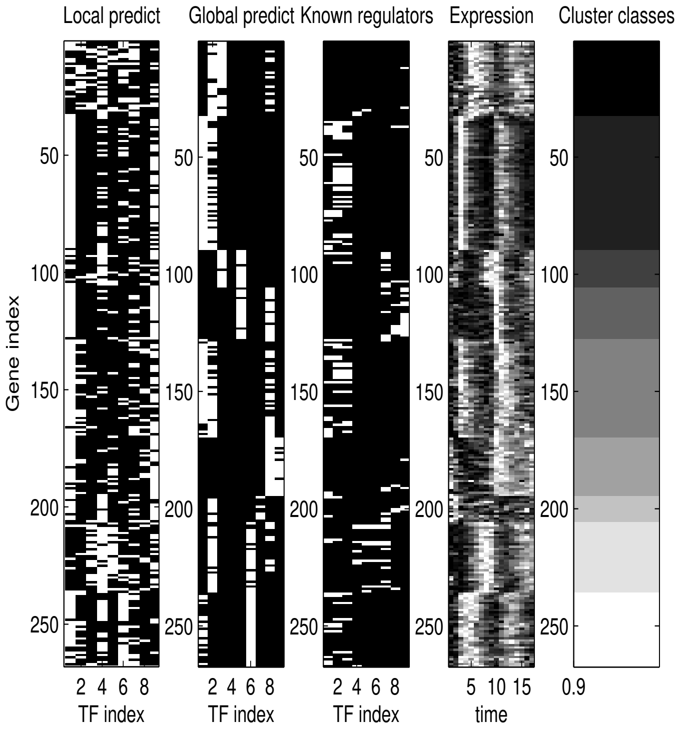
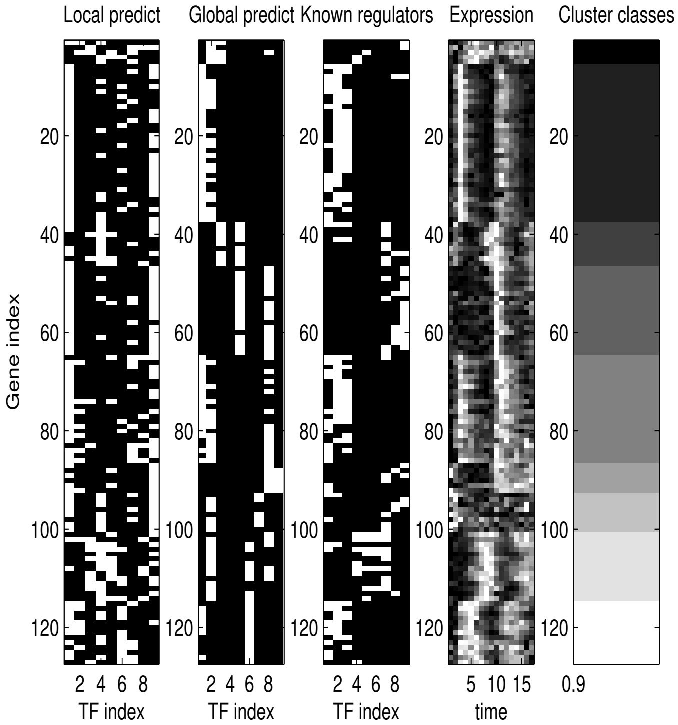


Figure 2: Results on the (?) data.

tor. Moreover, as an effective feature selection strategy, it might be useful to extend this approach resolving other feature selection bioinformatics problems.



Accuracy comparison	Local regularization	Global regularization
precision (%)	23.7	32.3
recall (%)	36.2	38.9
F-measure	28.6	35.3

Figure 3: Results on the subset of the (?) data where TF-based regulation information is known.