Multi-class Biclustering and Classification Based on Modeling of Gene Regulatory Networks

Ilias Tagkopoulos
Department of Electrical Engineering
Princeton University
iliast@princeton.edu

Nikolai Slavov
Department of Molecular Biology
Princeton University
nslavov@princeton.edu

S.Y. Kung
Department of Electrical Engineering
Princeton University
kung@princeton.edu

Abstract— **The attempt to elucidate biological pathways and classify genes has led to the development of numerous clustering approaches to gene expression. All these approaches use a single metric to identify genes with similar expression levels. Until now, the correlation between the expression levels of such genes has been based on phenomenological and heuristic correlation functions, rather than on biological models. In this paper, we derive six distinct correlation functions based on explicit thermodynamic modeling of gene regulatory networks. We then combine these correlation functions with novel biclustering algorithms to identify functionally enriched groups. The statistical significance of the identified groups is demonstrated by precision-recall curves and calculated p-values. Furthermore, comparison with chromatin immunoprecipitation data indicates that the performance of the derived correlation functions depends on the specific regulatory mechanisms. Finally, we introduce the idea of multi-class biclustering and with the help of support vector machines we demonstrate its improved classification performance in a microarray dataset.**

I. Introduction

One of the main challenges in computational biology is the classification of genes in functional categories and biological pathways based on their expression profiles. The expression data are often massively produced by microarray experiments and formulated as a matrix where each row is the logarithmic representation of a gene's relative expression with respect to the expression of a control, throughout different conditions. The control may be either the same gene in a predetermined condition or the average mRNA concentration of all genes. The columns of the microarray data matrix represent the different experiments (environmental conditions, individual cases, tissues or diseases).

A number of traditional clustering techniques such as K-means [25], self organizing maps [6], and hierarchical clustering [24] have been used over the years in order to cluster together genes or conditions. One promising approach, called biclustering, aims at discovering clusters of genes that are correlated in only a subset of conditions. The term biclustering (also known as co-clustering [4], direct clustering [7], box clustering [16]) was used by Mirkin (1996) to describe "simultaneous clustering of both row and column sets in a data matrix". The first time biclustering was used to cluster microarray data was in 2000, where a soft clustering algorithm on both genes/conditions was proposed [3] [12]. Since then, numerous microarray biclustering algorithms have been developed that either used a different metric to search for a similarity class [5] [11] [19] [23] [22], or proposed an alternative search method [21] [26]. A recent review and categorization of most biclustering methods so far can be found in [20].

Despite the plethora of metrics used in various clustering algorithms, the number of correlation functions that they aim to find is rather limited and unsubstantiated by any biological model. Most approaches try to cluster genes that either (a) have the same expression level through different conditions, (b) their expression levels differ by a constant or (c) their expression levels exhibit a linear relation. However, there is no biological reason to believe that the expression profile of coregulated genes should be strictly following one of the correlation functions mentioned above. Furthermore, due to the existence of various coregulation mechanisms, it is very likely that from gene group to gene group the observed correlation will be different.

We addressed the above concerns by associating correlation functions to transcription factor (TF) regulatory models. It is well established that TF regulatory networks have a profound, dominant effect in the transcription regulation. Therefore it is expected that correlation functions derived directly from them would better capture the common features among coregulated genes.

In the following sections, we derive directly from thermodynamic principles six distinct correlation classes that correspond to the most frequent regulation scenarios. We then introduce several metrics and two biclustering algorithms that are used to search for different gene groups and we propose a method of combining multiple classes together. Finally, we evaluate the performance of various correlation classes and biclustering methods by analyzing the functional enrichment of the resulting biclusters. Throughout the paper, when referring to expression data the words "gene" and "condition" are synonyms to "row" and "column" of the data matrix. Supplementary figures and proofs are available online [27].

II. Methods

A. Correlation Classes

The transcription of gene$_i$ under condition$_j$ will result in a specific mRNA concentration [mRNA]$_{ij}$. The simplest regulatory case occurs when a group of genes has exactly the same cis and trans regulatory elements, the corresponding mRNAs

Proceedings of the 5th IEEE Symposium on Bioinformatics and Bioengineering (BIBE'05)
0-7695-2476-1/05 $20.00 © 2005 **IEEE**

have the same lifetimes and the binding of trans regulatory elements to the promoter sites induces the same rate of transcription. In this scenario, the steady-state concentrations of the mRNAs for all members in the group will be equal for any condition. This corresponds to class 1 similarity with correlation function given by

$$\log{[mRNA]_{ij}} = \log{[mRNA]_{kj}} \tag{1}$$

However, any deviation from the above regulatory mechanism will introduce a displacement in the expression level between two genes. This leads to class 2 with correlation function

$$\log{[mRNA]_{ij}} = w_0 + \log{[mRNA]_{kj}} \tag{2}$$

The constant displacement in the transcription rates may arise from (a) genes having different epigenetic modifications that result in being transcribed at different rates [13] [14], (b) different topology of the promoter region that causes the genes to have different efficiency of transcription initiation or (c) different half lifetimes of the mRNAs [15]. More specifically, if several genes are transcribed at the same rate, but their mRNAs have different stabilities, their steady-state concentrations will be scaled and the corresponding logarithms will be displaced by a constant.

A more general class, class 3, incorporates the event of dissimilar TF binding stoichiometry in a group. A simple example is when only a single TF molecule is needed to regulate a certain gene, but two or more TF molecules are needed for the regulation of another. This class also incorporates the case where a modulator molecule(s) binds to multiple locations on the TF protein. The number of modulator molecules and the position they are bound affect regulation of target genes. In both cases, the mRNA concentrations of two coregulated genes will have a power law dependency, which gives rise to a class 3 correlation function

$$\log{[mRNA]_{ij}} = w_0 + w_1 \log{[mRNA]_{kj}} \tag{3}$$

A TF regulates multiple genes by identifying and binding to short conserved sequences that usually exist in the upstream regions of the regulated genes. Nevertheless, although these sequences are similar, they are almost never the same. In addition, the orientation of the consensus sequences and their distance from the start of translation may also differ from gene to gene. This binding site diversity results in different binding affinities of the TF-DNA interaction and, thus, alters the regulation dynamics of the system. The above regulatory mechanism gives rise to class 4 with correlation function

$$\log{[mRNA]_{ij}} = log(w_0[mRNA]_{kj}) + \log{(w_1[mRNA]_{kj} + w_2)} \tag{4}$$

The analytical expression of class 4 encompasses many other biological scenarios. For example, assume that TF$_m$ and TF$_n$ regulate genes *i* and *k* respectively. If TF$_m$ and TF$_n$ are also correlated by class 1 or class 2 correlation functions in a regulation cascade, then the expression profile of genes *i* and *k* will be described by correlation function 4. Hence, class 4 can detect correlations in the expression levels of mRNAs regulated by different TFs.

Transcription factors can be enzymatically modified [10] [8] [9]. Some modifications include phosphorylation [2] [10], acetylation [9], and methylation [18]. A TF modification usually changes the conformation and binding affinity to the promoter site, which may result in activation or inactivation of TF's regulatory role for a specific gene. It is also possible that the different conformational states of the TF regulate different gene groups. To encompass the effects of enzymatic modification we introduce class 5, whose correlation function is given by

$$\log{[mRNA]_{ij}} = \log{\frac{w_0 + w_1[mRNA]_{kj}}{w_2 + w_3[mRNA]_{kj}}} \tag{5}$$

It is interesting to note that we can also derive class 5 if the TF-induced transcription rate is comparable to the basal transcription rate, i.e., the transcription rate in the absence of any bound TF.

A sixth class emerges when a TF is a repressor of one gene and activator of another, as in the bacteriophage lambda system. The same class is derived when a TF has the same binding affinity for both genes. Its correlation function is linear in the mRNA domain and is given by

$$\log{[mRNA]_{ij}} = \log{(w_0 + w_1[mRNA]_{kj})} \tag{6}$$

In describing the six classes, we considered only one TF. However, all derivations can be generalized to cases where the gene expression is regulated by multiple TFs with invariant expression, i.e unaffected by the cluster conditions. For these general cases, the unchanging TFs will appear as a constant that is identical for all genes in a group. However, the expression of a gene group regulated by multiple non-shared TFs that change across the cluster conditions will not conform to the any of the six classes. Such cases with more than one dominant TF have to be modeled explicitly with multivariable correlation functions.

By assuming a real-valued expression matrix $X = \{x_{ij} | 1 \leq i \leq M, 1 \leq j \leq N\}$, with $x_{ij} = \log{[mRNA]_{ij}}$, the derived classes are summarized in the following expressions:

$$\begin{cases} x_{ij} = x_{kj} & 1^{st}Class \\ x_{ij} = w_0 + x_{kj} & 2^{nd}Class \\ x_{ij} = w_0 + w_1 x_{kj} & 3^{rd}Class \\ x_{ij} = w_0 + x_{kj} + \log{(w_1 e^{x_{kj}} + w_2)} & 4^{th}Class \\ x_{ij} = \log{\frac{w_0 + w_1 e^{x_{kj}}}{w_2 + w_3 e^{x_{kj}}}} & 5^{th}Class \\ x_{ij} = \log{(w_1 e^{x_{kj}} + w_2)} & 6^{th}Class \end{cases} \tag{7}$$

B. Metrics

In order to select genes that belong to one of the correlation classes mentioned above, different proximity metrics have been considered. We combined metrics with the appropriate data manipulation/preprocessing step for each correlation class. Table I depicts three proximity metrics and preprocessing techniques, along with the similarity class that they aim for. We use two data manipulation techniques as preprocessing steps: (a) row normalization, the substraction of the row mean from each element and (b) row standardization, the substraction of row mean from each element and its subsequent division by the row standard deviation:

$$\begin{array}{ll} \hat{x}_{ij} = x_{ij} - \bar{x}_i & RowNormalization \\ \tilde{x}_{ij} = \frac{\hat{x}_{ij}}{\sigma_i} = \frac{x_{ij}-\bar{x}_i}{\sigma_i} & RowStandardization \end{array} \qquad (8)$$

where $\bar{x}_i = \frac{1}{N}\sum_{j=1}^{N} x_{ij}$ and $\sigma_i = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(x_{ij} - \bar{x}_i)^2}$.

Euclidian distance is less complex and more flexible since it can search for first, second, and third class similarities. Using euclidian distance without any preprocessing will result in a class 1 bicluster. On the other hand, if row normalization is used, the resulting cluster will exhibit class 2 correlation. Finally, a standardization preprocessing step and subsequent use of the euclidian distance finds class 3 biclusters. Figure 1 depicts how a class 1,2, and 3 bicluster looks like, both in raw expression levels and after one of the two preprocessing steps.

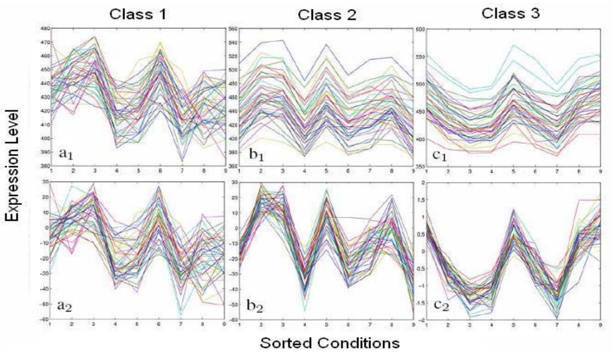

Fig. 1. Class 1,2 and 3 biclusters formed using euclidian distance and different preprocessing steps. Experimental conditions(x-axis) have been sorted in error ascending order in each subplot. The gene expression level is mapped in the y-axis. Each bicluster has 41 members and every line corresponds to the expression level of a gene in nine conditions. A class 1 bicluster is depicted in subplots (a_1-a_2). The gene expression level of a class 2 bicluster is shown in (b_1), whereas the expression level of the same bicluster after row normalization is shown in (b_2). Finally (c_1) and (c_2) depict the expression level of a class 3 bicluster before and after standardization accordingly.

An alternative metric for third class similarities, as well as the basic metric for the fourth and fifth class, is the mean squared deviation from a Least Squares regression. Assume vectors

$$\vec{x}_k = \begin{pmatrix} x_{k1} \\ x_{k2} \\ \dots \\ x_{km} \end{pmatrix}, \vec{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \dots \\ x_{im} \end{pmatrix}, \vec{w} = \begin{pmatrix} w_0 \\ w_1 \\ \dots \\ w_d \end{pmatrix}$$

and matrix $X_k = [1_m|\vec{x^1}_k|\dots|\vec{x^d}_k]$, where 1_m is a m-dimensional vector of ones, $k\epsilon[1, 2, \dots, L]$ and $\vec{x^d}_k$ is the vector created when each element of $\vec{x}_k$ is raised in the d^{th} power. We want to determine $\vec{w}$ so that the sum of squared errors is minimum, that is

$$\Theta = \underset{\vec{w}}{\operatorname{argmin}} \sum_{k=1}^{L} \|\vec{x}_i - X_k\vec{w}\|^2$$

The solution is given by:

$$\vec{w} = \left(\sum_{k=1}^{L} X_k^T X_k\right)^{-1} \sum_{k=1}^{L} X_k^T \vec{x}_i \qquad (9)$$

Therefore, we can easily evaluate the correlation of any new (gene) vector x_i to a group of L vectors.

C. Algorithms

Most biclustering algorithms start from a random matrix row or column and try to end up with a bicluster whose members have a certain type of correlation. Moreover, the number of conditions and members in the resulting bicluster is usually picked by an arbitrary user-defined threshold. Although this approach may work adequately for gene groups that exhibit strong correlation, it severely compromises the performance in the general case. Our approach makes use of the already known biological information to find functionally enriched biclusters and estimates the optimum initial parameters for our biclustering algorithm. The first is implemented by starting from a subset of genes that are biologically proven to be related, whereas the latter is achieved by introducing a scoring function.

Furthermore, combination of different metrics in 2D space and use of a non-linear kernel classifier such as support vectors machines, yields higher performance in all cases.

1) Single-metric classification: The basic idea in both algorithms presented here is identical: Given a subset of genes that are known to be correlated, cluster together genes whose expression can be described using the correlation functions (1-5). Algorithm 1 clusters genes with correlation of type 1,2, and 3 , whereas algorithm 2 is used for searching class 3,4, and 5 similarities. In algorithm 1 all metrics of table I can be used, by adjusting the preprocessing step accordingly.

Algorithm 2 uses solution (9) to calculate $\vec{w}$. Then, it finds genes and conditions that approximate better the correlation function of a certain class. The correlation functions of classes 3 and 6 are linear functions in the log(mRNA) and mRNA domain respectively, hence the application of a least squares method is obvious. In addition, both class 4 and 5 correlation functions can also be approximated by polynomials. More specifically, exponentiation of 4 results to

$$\begin{array}{rcl} [mRNA]_{ij} & = & w_0 w_1 [mRNA]_{kj}^2 + w_1 w_2 [mRNA]_{kj} \\ & = & w_2' [mRNA]_{kj}^2 + w_1' [mRNA]_{kj} \end{array}$$

Thus, in the [mRNA] domain the correlation function of class 4 is a second order polynomial with the zero order term absent.

TABLE I
SIMILARITY - PROXIMITY METRICS

Name	Metric	PREPROCESSING		
		without Normalization	with row Normalization	with row Standardization
Euclidian$(\vec{x}_i, \vec{x}_k)$	$\sqrt{\sum_{j=1}^{J}(x_{ij} - x_{k,j})^2}$	Class 1	Class 2	Class 3
Mean squared residue(X)	$\frac{1}{MN}\sum_{i\epsilon M, j\epsilon N}(x_{ij} - \bar{x}_i - \bar{x}_j - \bar{x})^2$	Class 2	Class 2	Class 3
Pearson Correlation$(\vec{x}_i, \vec{x}_k)$	$\frac{\sum_{j\epsilon J}(x_{ij} - \bar{x}_i)(x_{kj} - \bar{x}_k)}{\sqrt{\sum_{j\epsilon J}(x_{ij} - \bar{x}_i)^2 \sum_{j\epsilon J}(x_{kj} - \bar{x}_k)^2}}$	Class 3	Class 3	Class 3

```
Algorithm 1 - Clusters of Correlation Class 1,2,3
Input(X,C,L,K,I):
    A data matrix $X$ of size MxN, the correlation class $C$, an initial expression submatrix $L$ of correlated genes in all N conditions, the number $K$ of initial columns and $I$, the number of members of the final bicluster.
Output(B):
    An IxJ bicluster $B$, whose members exhibit class $C$ correlation.
Preprocessing:
    Depending on $C$ and the metric used, perform normalization or standardization on X, as depicted on Table I.
Initialization:
    • Calculate the metric distance for submatrix $L$ in all N columns.
    • Sort the columns in ascending distance order (smallest distance first).
    • Form initial $I'xJ$ bicluster B, where I'=rows(L) and J=K.
Row addition:
    • $\forall i\epsilon M$ and $\forall j\epsilon J$, compute the mean row metric distance $E_{row_i}$ to the center of B. Example: If metric=Euclidian, $B_{ij}$ and $X_{ij}$ elements of i$^{th}$ row and j$^{th}$ column of matrices B and X, then compute
          $E_{row_i} = \frac{1}{J}\sum_{j=1}^{J}(X_{ij} - \frac{1}{I'}\sum_{o=1}^{I'}B_{oj})^2$
    • Add the row with the smallest distance $E_i$ to B and remove it from initial matrix.
    • Set $M = M - 1, I' = I + 1$. Continue until $I' = I$.
Column addition
    Compute the column variance $Var_{col_j}$ for the rest $N - K$ columns in X. Add any column j with a column variance less than the mean column variance of the bicluster, i.e. satisfy: $Var_{col_j} \leq \frac{1}{K}\sum_{o=1}^{K}Var_{col_o}$
```

```
Algorithm 2 - Search for Correlation Class 3,4,5,6
Input(X,C,L,K,I):
    A data matrix $X$ of size MxN, the correlation class $C$, an initial expression submatrix $L$ of correlated genes in all N conditions, the number $K$ of initial columns and $I$, the number of members of the final bicluster.
Output(B):
    An IxJ bicluster $B$, whose members exhibit class $C$ correlation.
Preprocessing:
    For class 4,5 or 6, exponentiate each element in matrix X.
Initialization:
    • Calculate for submatrix $L$ the mean deviation from the correlation function of class C.
    • Sort the columns in ascending distance order (smallest distance first).
    • Form initial $I'xJ$ bicluster B, where I'=rows(L) and J=K.
    • Set $x_1 = [B_{11}B_{12}...B_{1K}]$.
    • Set
        – $X_k = [1_K|\vec{x}_i]$ for class 3 or 6
        – $X_k = [\vec{x}_i|\vec{x}_i^2]$ for class 4
        – $X_k = [1_K|\vec{x}_i|\vec{x}_i^2]$ for class 5
    • compute $\vec{w}$ from equation (9), i.e.
          $\vec{w} = \left(\sum_{k=2}^{L}X_k^TX_k\right)^{-1}\sum_{k=2}^{L}X_k^T\vec{x}_1$
Iteration:
    • calculate $E_i = \sum_{k=1}^{I'}\|\vec{x}_i - X_k\vec{w}\|^2\forall i\epsilon M.$
    • Add the row with the smallest distance $E_i$ to B and remove it from initial matrix.
    • Set $M = M - 1, I' = I + 1$. Continue until $I' = I$.
    • Calculate mean correlation function deviation for the rest N-K columns and add any column with less than the mean deviation of the K columns.
```

Similarly, after a first order Taylor expansion in the mRNA domain, class 5 gives rise to a second order polynomial:

$$\begin{aligned}[mRNA]_{ij} &= \tfrac{w_0+w_1[mRNA]_{kj}}{w_2+w_3[mRNA]_{kj}} \\ &\cong \tfrac{w_0+w_1[mRNA]_{kj}}{w_2}(1 - \tfrac{w_3}{w_2}[mRNA]_{kj} + \ldots) \\ &= w_2''[mRNA]_{kj}^2 + w_1''[mRNA]_{kj} + w_0''\end{aligned}$$

where $w_2'', w_2', w_1'', w_1', w_0'', w_0'$ are constants.

An interesting question is how to find the optimum initial parameters, namely the number of initial columns K and final row size I of the bicluster. To address this question, we introduce the following scoring function:

$$Score = w_1Sensitivity + w_2Specificity + w_3Precision \quad (10)$$

where w_1, w_2, w_3 are weights equal to $\frac{1}{3}$ for the unbiased case. The initial parameters(*K,I*) that maximize expression (10) are the ones that yield the most enriched bicluster. Figure 2 demonstrates how the score calculated by (10) identifies the initial parameters for correlation class 1,2, and 3, starting from 2 initial genes(L=2). For example, for class 1 it turns out that the resulting bicluster is highly enriched when (K,I)=(10,115).

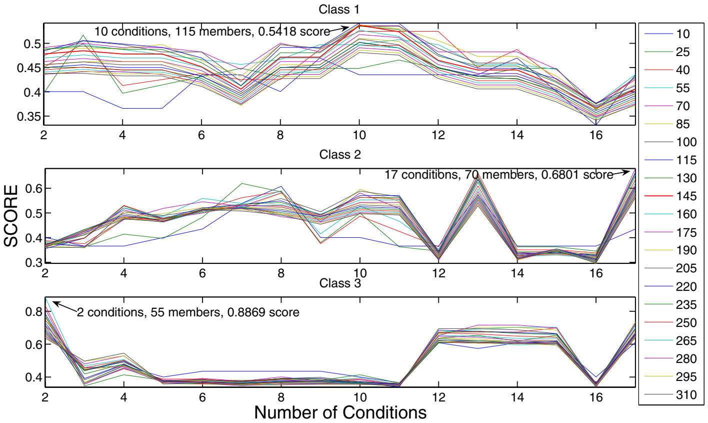

Fig. 2. Scoring function as selection method for the size of final bicluster and number of initial columns. Each line represents a bicluster of a certain size. The number of conditions are sorted in ascending scoring order and are mapped on x-axis. Each subplot corresponds to a class, namely class 1,2 and 3. The legend shows the number of members (I) of the final bicluster for each case.

2) Multi-metric classification: Although correlation functions based on TF regulatory networks help identify groups with members that share the same TF mechanism, there is no guaranty that they will perform equally well in groups with higher regulatory complexity. Since many functional category groups have multiple substructures, one may expect that any single correlation function algorithm will be able to recover a specific subgroup at most.

For this reason we developed a multi-class/multi-metric methodology methodology similar to [17]:

1) Split the microarray data matrix in a training and a test set.
2) For certain initial parameters and for each correlation class, create a bicluster by using Algorithms 1 and 2 on data from the training set.
3) Compare the number of true positive genes in all biclusters and the amount of overlap. Select the two with the minimum amount of overlap.
4) For all genes in the data matrix, compute the mean deviation from the center of the selected biclusters (as in algorithms 1 and 2). This corresponds to two representative numbers (correlation errors) E_{class_i}, E_{class_j} for each gene.
5) Train a classifier (SVM or any other nonlinear classifier) with E_{class_i}, E_{class_j} as features. Finally, evaluate its performance on the test set.

D. Dataset and Group Selection

All simulations were performed on a Saccharomyces cerevisiae cell cycle expression data, initially produced by Cho et al. and normalized by Church [3]. The dataset contains 2884 genes in 17 conditions. Missing values were replaced by random numbers for simplicity.

The functional category evaluation presented here was focused on four distinct gene groups downloaded from Stanford's Saccharomyces cerevisiae database [28], namely the structural constituent of ribosome, structural molecular activity, dna binding, rna binding. The group selection was based on the number of members, diversity and expected correlation. Moreover, we tested our method in groups of genes that bind to common TFs as identified by ChIP experiments [18].

For comparison purposes, the metric used with Algorithm 1 for the creation of all graphs was euclidian distance, and the initial "seed" subset had only two members(genes), randomly selected from all positive samples. Classes one to five were evaluated in this study. In multi-class evaluation, the training to testing ratio was 2:1 for both positives and negative samples.

III. Results

A. Single Class

Figure 3 illustrates the performance of Algorithm 1 and 2 when searching for the first five classes. For class 3, both algorithms where used (Algorithm 1 and 2 are mapped to "class3" and "class3/LS" accordingly). Table II illustrates the number of true positive (TP) and false positive (FP) members, as well as the p-values (hypergeometric distribution) for the resulting biclusters corresponding to all five correlation classes, with initial values picked by the presented scoring function.

A general observation that arises from figure 3 is the absence of a universal class, a class that outperforms all others in classification. For example, in the ribosomal group the dominant class is 3, whereas in the RNA binding it is class 5. This is somewhat expected: The regulatory mechanism of ribosomal genes is much less complex than its RNA binding counterpart. Thus, a more general class, such as class 5 is required in the latter case in order to capture the features of the RNA binding group. It has also been reported that a TF for ribosomal genes is Fhl1. We verified this observation by analyzing the ChIP data available in [18] and extended the ribosomal specificity to the fusion of other two TFs, ADR1 and ARG80 (out of 52 common targets 46 are ribosomal genes). On the other hand, RNA and DNA binding genes do not have a universal dominant TF. Hence, although ribosomal expression is highly correlated, an observation that explains the good overall performance of all classes in figure 3, we cannot expect this to be true for the DNA/RNA binding group.

The classification performance of the Algorithm 2 - class 3 combination for the DNA binding group performs better than most metric-class combinations, including Algorithm 1 - class 3 combination. Algorithm 2 is sensitive to anti-correlated genes, i.e. genes with odd symmetry in their expression levels. Anti-correlated expression is more likely to happen in complex regulatory mechanisms, where TF cascades can be formed. This is what is also happening in the DNA binding case. Moreover, by comparing the plots of less complex groups (ribosomal and structural activity groups, plots (a) and (b)) to these with higher regulatory complexity (DNA and RNA

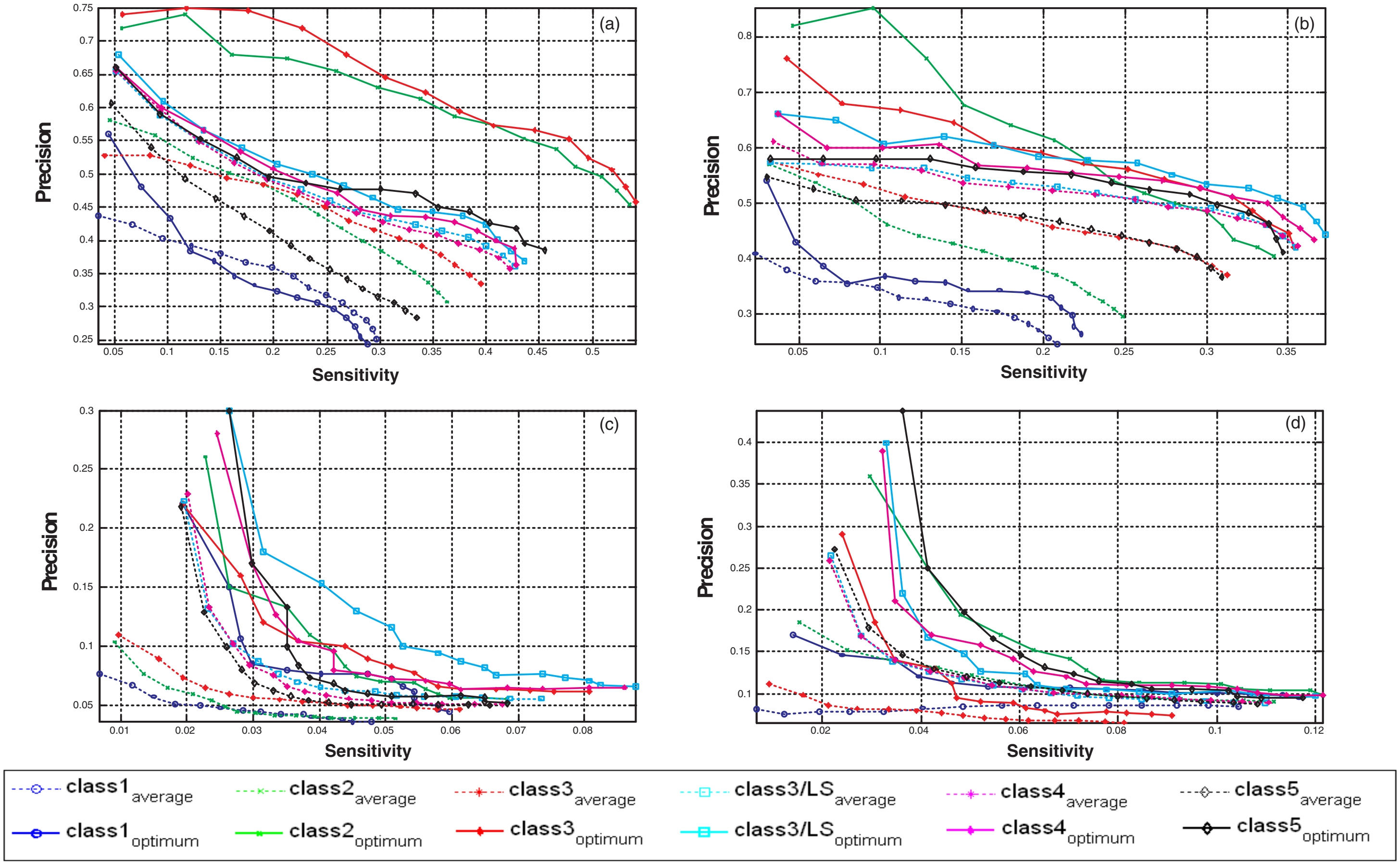

Fig. 3. Precision - Recall Curves for all five correlation classes and for four distinct functional groups. Each graph is created by running algorithms 1 and 2 with 20 random initial 'seed' subsets L, that containing only two genes each. Bicluster size *I* ranges from 10 to 150 genes and initial condition number *K* ranges from 2 to 17. The initial two genes (the "seed" subset) are selected from the positive sample pool. The positive sample pool is also the reference in respect to which the true/false positive/negative numbers are calculated in each subplot and it consists of (a) ribosomal genes (126 members) (b) structural activity genes (178 members) (c) DNA binding activity genes (114 members) (d) RNA binding activity genes (121 members). Solid lines represent the average precision-sensitivity over the 20 random seeds when the scoring function was utilized in order to determine the optimum (*I* and *K*) parameters as described in the methods section. Dotted lines depict the average precision-sensitivity over 20 random seeds L, all bicluster sizes *I* and initial condition numbers *K*.

binding) it becomes clear that Algorithm 1 - class 3 works better in the first whereas Algorithm 2 - class 3 in the latter.

From the precision - recall curves, the importance of the parameter selection step that was presented in algorithmic section becomes clear. The biclusters that are formed with initial parameter selection are significantly more enriched than the average case. This observation is also consistent with our ChIP data findings (not shown here).

B. Multi-class

Figure 4 illustrates the classification of ribosomal genes, using the error score of two correlation classes. By combining two different classes, we were able to improve the clustering performance in the ribosomal group by an average of 12.6% when compared to the best single Algorithm - class combination (Algorithm 1 - Class 3). Note that the various correlation error pairs have non-similar spatial topology. Thus, the pair we select to serve as a feature can significantly change the performance of the method.

The potential of the multi-class fusion methodology stems not only from the fact that it takes into account different regulatory mechanisms, but also from its ability to exploit any special relationship our data may have (noise correlation and canceling, linear dependency, etc). For example, in figure 4(d), the true positive samples exhibit an almost linear relationship to class 2 and 3.

IV. Conclusions

By explicit modeling of regulatory networks that control gene expression, we derived six classes that capture distinct correlation functions between the expression levels of genes regulated by common TFs. Each class predicts the expression of genes that are regulated by the specific mechanisms outlined in the methodology section. Therefore, each class is mechanism-specific and preferentially identifies genes whose expression is regulated by the mechanisms used in its derivation. The expression levels of genes regulated by common TFs are predicted with highest sensitivity and precision by a

TABLE II

FUNCTIONAL ENRICHMENT - RIBOSOMAL BICLUSTER

		Class 1		Class 2		Class 3 - Alg. 1		Class 3 - Alg. 2		Class 4		Class 5	
Funct. Group	Members	TP/FP	P	TP/FP	P	TP/FP	P	TP/FP	P	TP/FP	P	TP/FP	P
Ribosomal	121	66/32	-74.68	68/27	-80.29	81/32	-98.80	71/26	-85.61	70/28	-82.92	73/22	-91.21
Structural	178	13/5	-12.10	19/1	-22.11	55/8	-60.93	78/27	-78.08	17/2	-18.70	76/37	-70.37
DNA Binding	114	7/7	-6.47	5/15	-5.22	6/14	-4.11	6/8	-5.12	4/12	-2.57	4/10	-2.80
RNA Binding	121	19/14	-18.09	10/4	-10.99	6/9	-4.77	8/8	-7.14	8/8	-7.14	8/8	-7.14

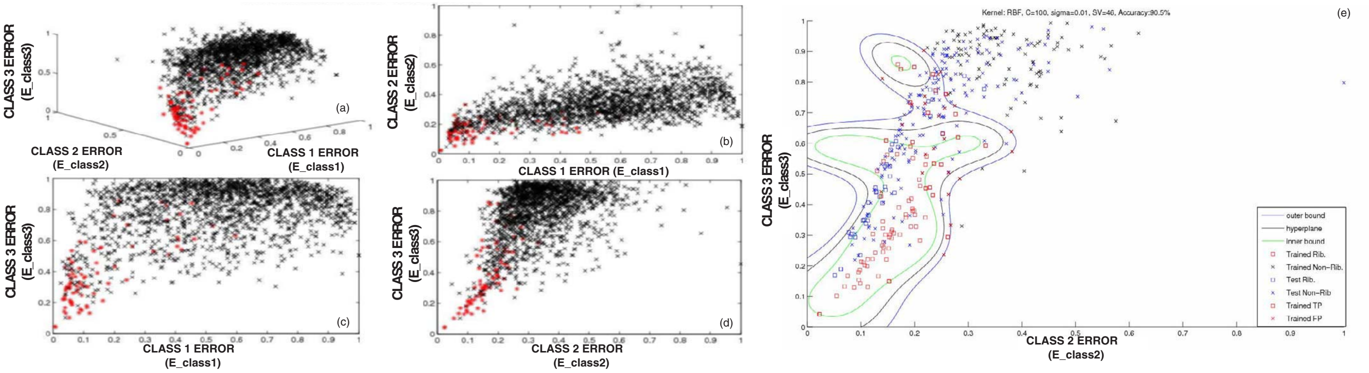

Fig. 4. Representation of correlation error (E_{class}) for class 1,2, and 3. Red stars are the ribosomal genes (positives samples), whereas black crosses represent the non-ribosomal genes in the data set. In (e), the ribosomal/non-ribosomal genes are depicted with squares/crosses, the color red is used for the training and the color blue for the test samples. A RBF kernel was used with N-fold cross validation for the selection of kernel parameters.

single class. In addition, different coregulated gene groups are predicted best by different classes.

These findings are consistent with the experimental evidence suggesting that gene regulatory networks can be diverse and gene specific. Since the performance of a certain correlation class depends on the underlying regulatory mechanism, our approach can also be applied in order to determine possible regulatory mechanisms within groups and further evaluate their statistical significance by making use of microarray expression data.

Moreover, we presented two effective biclustering algorithms and a multi-class classification technique that utilizes the correlation functions mentioned above. The proposed framework and algorithms can easily be expanded to include gene groups regulated by multiple TFs and more complex regulatory mechanisms. Such future extensions are likely to have improved performance in classification of more diverse gene groups and identification of their regulatory elements. Additionally, a class integration criterion, possibly estimating interclass mutual information, would boost the performance multi-class classification technique. Finally, it would be interesting to combine the newly derived gene classes with more appropriate condition/experiment correlation functions to further improve the classification performance.

V. APPENDIX

A. Thermodynamics model

For a gene transcription system with M repressors and N activators, we define:

- **s** as the state number.
- **ΔG** as the difference of free energy in respect to the ground state.
- k_s the transcription initiation rate for a specific state s.
- a_i^s The multiplicity (number of molecules) of activator i in state s.
- r_j^s The multiplicity (number of molecules) of repressor j in state s.

In equilibrium and allowing the multiplicity being different than one, the probability that the system is in each of the possible $(a_1+1)(a_2+1)\dots(a_N+1)(r_1+1)(r_2+1)\dots(r_M+1)2^{N+M}$ states can be described by partition functions. The probability of the system being in any distinct state s_0 is given by:

$$P(s_0) = \frac{e^{\frac{-\Delta G}{RT}}[Repr_1]_{s_0}^{r_1^{s_0}} \dots [Repr_M]_{s_0}^{r_M^{s_0}} [Act_1]_{s_0}^{a_1^{s_0}} \dots [Act_N]_{s_0}^{a_N^{s_0}}}{\sum_s e^{\frac{-\Delta G}{RT}}[Repr_1]_s^{r_1^s} \dots [Repr_M]_s^{r_M^s} [Act_1]_s^{a_1^s} \dots [Act_N]_s^{a_N^s}} \tag{11}$$

If k_s is the transcription initiation rate for the state s and RNA polymerase is abundant the total transcription rate will be

$$F(Act_1, Repr_1, \dots) = \sum_s k_s P(s) \tag{12}$$

The chemical equation for $[mRNA]_i$ can be modeled with a differential equation [1]:

$$\frac{d([mRNA]_i)}{dt} = F(Act_1, Repr_1, \dots) - k_{deg}[mRNA]_i \tag{13}$$

where k_{deg} is the degradation rate of $[mRNA]_i$ and F is given by (12). In equilibrium, equation (13) becomes

$$[mRNA]_i = \frac{F(Act_1, Repr_1, \ldots)}{k_{deg}} \tag{14}$$

By using equations (11-14), any possible combination of transcription factors, binding affinity, binding site multiplicity, transcription initiation rate can be solved and their corresponding classes studied.

B. Class Derivation

Without loss of generality, we assume one common TF and multiplicity equal to one. State '1' is the unbound state and k_1 is the basal transcription initiation rate. From (11 - 14) we have:

$$[mRNA]_i = \frac{k_1 + k_2 e^{\frac{-\Delta G_2}{RT}}[TF]}{k_{deg}(1 + e^{\frac{-\Delta G_2}{RT}}[TF])} \tag{15}$$

and

$$[mRNA]_k = \frac{k_1' + k_2' e^{\frac{-\Delta G_2'}{RT}}[TF]}{k_{deg}'(1 + e^{\frac{-\Delta G_2'}{RT}}[TF])} \tag{16}$$

Solving (15) for [TF] and substituting it to (16) yields

$$[mRNA]_k = \frac{A[mRNA]_i + B}{C[mRNA]_i + D} \tag{17}$$

where
- $A = (k_1' k_{deg} e^{\frac{-\Delta G_2}{RT}} - k_2' k_{deg}' e^{\frac{-\Delta G_2'}{RT}})$
- $B = (k_1 k_2' e^{\frac{-\Delta G_2'}{RT}} - k_2 k_1' e^{\frac{-\Delta G_2}{RT}})$
- $C = k_{deg} k_{deg}' (e^{\frac{-\Delta G_2}{RT}} - e^{\frac{-\Delta G_2'}{RT}})$
- $D = k_{deg}'(k_1 e^{\frac{-\Delta G_2'}{RT}} - k_2 e^{\frac{-\Delta G_2'}{RT}})$

Equation (17) represents the general case and provides the relationship between $[mRNA]_k$ and $[mRNA]_i$. Since the terms in parenthesis are constants, the above result to correlation function 5. Now we consider the following special cases :

- TF is an activator for both genes and basal level of expression is minimal when compared to the expression level of the state where TF is bound. In other words $k_1 = k_1' = 0$, which transforms (17) into a class 4 correlation function.
- TF has the same binding affinity for both genes, i.e. $\Delta G' = \Delta G$, or TF is activator for the one gene and repressor for the other. This yields a class 6 correlation function.
- TF has the same binding affinity for both genes and the basal level of expression is minimal compared to the TF bound expression level. Thus, with $\Delta G' = \Delta G$ and $k_1 = k_1' = 0$, we get from (17) a class 2 correlation function.
- When binding affinity, basal level of expression and degradation rate is the same, we get class 1 similarity (all previous class 2 constrains and $k_{deg} = k_{deg}'$).

VI. Acknowledgements

We would like to thank Eli Berici for the data preparation, Ioannis Avramopoulos and Alexander Solovyov for the helpful discussions. This work was supported by the Burroughs-Wellcome Graduate Training Program in Biological Dynamics.

References

[1] Bower J., Bolouri H., ”Computational Modeling of Genetic and Biochemical Networks”, MIT Press, 2001.

[2] Bhoumik A, Takahashi S, Breitweiser W, Shiloh Y, Jones N and Ronai Z. ATM-Dependent Phosphorylation of ATF2 Is Required for the DNA Damage Response. Mol.Cell 18(5), pages 577-587, 2005.

[3] Yizong Cheng and George M. Church. Biclustering of expression data. Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB00), pages 93-103, 2000.

[4] Dhillon S. Co-clustering documents and words using bipartite spectral graph partitioning. Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD01), pages 269-274, 2001.

[5] Getz G.,Levine E., and Domany E. Coupled two-way clustering analysis of gene microarray data. PNAS, pages 12079-12084, 2000.

[6] Eisen M. B., Spellman P. T., Brown P. O., Botstein D.: Cluster analysis and display of genome-wide expression patterns. PNAS USA, 95, pages 14863-14868, 1998.

[7] Hartigan J. A. Direct clustering of a data matrix. Journal of the American Statistical Association (JASA), 67:(337) pages 123-129, 1972.

[8] Huo X. and Zhang J. Important roles of reversible acetylation in the function of hematopoietic transcription factors. J.Cell.Mol.Med. 9:(1) 103-112, 2005.

[9] Huq MM. and Wei LN. Post-translational modification of nuclear co-repressor receptor interacting protein 140 by acetylation. Mol.Cell.Proteomics 2005.

[10] El-Kady A. and Klenova E. Regulation of the transcription factor, CTCF, by phosphorylation with protein kinase CK2. FEBS Lett. 579:(6) 1424-1434, 2005.

[11] Yuval Klugar, Ronen Basri, Joseph T. Chang, and Mark Gerstein. Spectral biclustering of microarray data: coclustering genes and conditions. Genome Research, volume 13, pages 703-716, 2003.

[12] Lazzeroni L. and Owen A.B. Plaid models for gene expression data. Technical report, Stanford, 2000.

[13] Ma K, Chan JK, Zhu G and Wu Z. Myocyte enhancer factor 2 acetylation by p300 enhances its DNA binding activity, transcriptional activity, and myogenic differentiation. Mol.Cell.Biol. 25:(9) 3575-3582, 2005.

[14] Magis W, Fiering S, Groudine M and Martin DI. An upstream activator of transcription coordinately increases the level and epigenetic stability of gene expression. PNAS 93:(24) 13914-13918, 1996.

[15] Meins F,Jr. RNA degradation and models for post-transcriptional gene-silencing. Plant Mol.Biol. 43:(2-3) 261-273, 2000.

[16] Mirkin, B. Mathematical Classification and Clustering. Dordrecht: Kluwer, 1996.

[17] S.Y.Kung, M.W. Mak, and I. Tagkopoulos, ”Multi-Metric and Multi-Substructure Biclustering Analysis for Gene Expression Data,” IEEE Computational Systems Bioinformatics Conference (CSB05), Stanford University, California, 2005

[18] Lee, T.I., Rinaldi, N., Roberts, F., Odom, D., Bar-Joseph, Z., Gerber, G., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., et al. Transcriptional regulatory networks in Saccharomyces cerevisiae. Science 298, pages 799-804, 2002.

[19] Eran Segal, Ben Taskar, Audrey Gasch, Nir Friedman, and Daphne Koller. Rich probabilistic models for gene expression. Bioinformatics, volume 17 (Suppl. 1), pages 243-252, 2001.

[20] Madeira S., Oliveira A. Biclustering Algorithms for Biological Data Analysis: A Survey. IEEE/ACM Trans. Comput. Biol. Bioinformatics, volume 1, pages 24-45, 2004.

[21] Qizheng Sheng, Yves Moreau, and Bart De Moor. Biclustering micrarray data by gibbs sampling. Bioinformatics, volume 19 (Suppl. 2), pages 196-205, 2003.

[22] Amos Tanay, Roded Sharan, and Ron Shamir. Discovering statistically significant biclusters in gene expression data. Bioinformatics, volume 18 (Suppl. 1), pages 136-144, 2002.

[23] Chun Tang, Li Zhang, Idon Zhang, and Murali Ramanathan. Interrelated two-way clustering: an unsupervised approach for gene expression data analysis. Proceedings of the 2nd IEEE International Symposium on Bioinformatics and Bioengineering, pages 41-48, 2001.

[24] Tamayo P. et al.: Interpreting patterns of gene expression with self-organizing maps, method and application to hematopoietic differentiation. PNAS, 96, pages 2907-2912, 1999.

[25] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, G. M. Church. Systematic Determination of Genetic Network Architecture. Nature Genetics, 22, pages 281-285, 1999.

[26] Jiong Yang, Wei Wang, Haixun Wang, and Philip Yu. Enhanced biclustering on expression data. Proceedings of the 3rd IEEE Conference on Bioinformatics and Bioengineering, pages 321-327, 2003.

[27] Supplemental Information is available at:
http://www.princeton.edu/∼iliast/bibe2005

[28] Saccharomyces Genome Database.
http://www.yeastgenome.org