

Towards Understanding the Relations Among Co-expressions, Annotations and Co-regulations

Raja Loganantharaj* and Jun Chung†

*Department of Computer Science, University of Louisiana at Lafayette, LA, USA

†LSU Health Sciences Center at Shreveport, LA, USA

*Contact author: logan@cacs.louisiana.edu

Abstract—Microarray technology provides an opportunity to view transcriptions at genomic level under different conditions controlled by an experiment. From an array experiment, few hundreds of differentially expressed genes are selected and are clustered using one of several standard algorithms. Generally, co-expressed genes, which are members of the same cluster, are expected to have similar function, but unfortunately it is not often true. A functional annotation often refers to molecular function and biological processes and there is no specific way to combine the functional annotation when it comes to associating them with clusters. We will show that the proposed method based on singular value decomposition performs better than a typical smoothing technique. Co-regulated genes are often expressed simultaneously and are expected to have similar function. In this paper, we will propose a systematic method to study the relations among co-expression, co-regulation and functional annotations. We will illustrate the method with a recent microarray experimental conducted for identifying transcriptional targets of integrin $\alpha\beta4$ for breast cancer progression.

Index Terms—Information content, gene ontology, annotation.

I. INTRODUCTION

A micro array experiment is conducted to study expression profiles of genes in a specimen under different experimental conditions, or over several different time periods. Statistical tests are conducted to filter valid signals first and then a subset of genes called differentially expressed genes is selected based on their expression levels. The differentially expressed probes, which roughly correspond to genes, are reduced to few hundreds while the total number of probes of an experiment is in the order of 40 to 50 thousands.

The differentially expressed genes are grouped together based on the expression similarity using one of standard clustering algorithms such as hierarchical clustering or k-mean clustering. The genes of a cluster have similar expression patterns and hence are called co-expressed genes. When a set of transcription factors regulates a pair of genes, then they are said to be co-regulated and the pair of genes are often co-expressed. Also, co-regulated genes are likely to have similar function. On the other hand, co-expressed genes are not necessarily co-regulated and hence do not have to have similar function.

In gene ontology, which is explained in section II, a gene is associated with cellular component, biological process and molecular function. A functional annotation of a gene often denotes its molecular function or biological processes. We will explore the relationship between the expression patterns and

its functional annotations in this paper.

Co-regulated genes share transcriptions factors; hence these genes share binding sites. Since it is hard to obtain or find the experimentally determined binding sites for many genes, we will use putative binding sites for this experiment.

We provide background materials on gene ontology in Section II, and on data fusion in Section III. It is followed by a section on binding sites and data. It is followed by a section on our approach, which discusses the proposed methods in details. We summarize and discuss the work in section VII followed by brief conclusions.

II. GENE ONTOLOGY

The gene ontology (GO) [1] project provides structured controlled vocabularies to address gene products consistently over several databases including FlyBase (*Drosophila*), the *Saccharomyces* Genome Database (SGD) and the Mouse Genome Database (MGD). The ontology describes gene products in terms of their associated biological processes, cellular components and molecular functions for each annotated gene. Each description of a gene product is arranged in a hierarchy from more general to very specific and the corresponding graph forms a directed acyclic graph (DAG) in which each node corresponds to a GO term and the label on the arc corresponds to the relationship between the terms. The relationship between a pair of GO terms includes *part_of* and *is_a*. The terms that are of one hop distance from the node corresponds to molecular function, say n_{mf0} , are said to be terms at level 1. The terms that are at level k are at least k hop distance from n_{mf0} . Similarly, the terms that are of one hop distance from the node corresponds to biological process, say n_{bp0} , are said to be terms at level 1. The terms that are at level k are at least k hop distance from n_{bp0} .

If a gene is annotated, the information is given by the corresponding GO accession numbers. Since a gene may be associated with many functions, it may have many GO accessions or GO terms.

III. DATA FUSION

Data fusion is a technique that combines different outcomes of classifiers and produces an outcome that is better than any of the individual outcome of a classifier. The classifiers are trained with different sample of the training set or using different features of training set. The problem of functional

prediction may be viewed as classification problem of assigning functional class to a given gene. Based on the underlying approaches for fusing data, it can be broadly categorized into (1) ensemble based fusion, (2) kernel based fusion and (3) singular value decomposition (SVD) based fusion. We plan to investigate and experiment with these data fusion technology for their suitability for improving function prediction accuracy. We briefly describe each method.

A. Ensemble Based Fusion

Suppose there are N different prediction systems or classifiers and each of them is trained individually with different portions of the training set or different aspects of the training set. Each classifier has created its decision boundaries and has its own error in prediction. The function prediction may be viewed as a classifier that assigns labels to an unlabeled instance. There are several variations among the algorithms on how the data is being distributed among different classifiers, and how the outputs of the classifiers are combined. Each classifier may be assigned some weight reflecting the input choice and its prediction accuracy. The easiest way to combine the output would be a linear combination of the weighted averages of the outcome, or enforcing voting mechanism. Having diverse classifiers in the pool will improve the overall prediction accuracy. Tumer et al. [2] has shown that the overall accuracy will improve with diverse or uncorrelated classifiers.

The many ways of combing the results of individual classifier in the ensemble include ranking, voting, sum, product, combination of posterior probability and decision templates. Theoretical analysis of combining results are discussed in [3, 4].

B. Kernel Based Fusion

Kernel based methods [5, 6] embed an object or a data item, say vector x , in high dimensional feature space Γ denoted by the mapping $\Phi(x)$ and the embedding in the feature space is implicitly specified by inner product for the feature space. A kernel is a symmetric positive semidefinite matrix with elements, say a kernel function $K(x_i, x_j)$, represents a vector inner product of the embedded data $\langle \Phi(x_i), \Phi(x_j) \rangle$. Kernels also have an interesting property that addition, multiplication and exponentiation of kernels will also result into another kernel by preserving symmetry and positive semidefiniteness. This property makes it very attractive for data fusion.

Lanckriet et al. [7] have demonstrated the utility of kernel based data fusion for classification of ribosomal and membrane proteins of yeast.

C. Singular Value Decomposition based Fusion

Singular Value Decomposition (SVD) based methods has been used very successfully in information retrieval [8-11] especially in the context of latent semantic analysis (LSA) and the technique found to be very useful for filtering and linguistic ambiguity in information retrieval. SVD handles

synonym (several words referring to the same concept) and polysemy (one word refer to several concepts) by mapping key words into a lower dimensional space of singular vectors called eigen-keywords and eigen-document.

Let X denote a $m \times n$ of real valued matrix. The equation for the singular value decomposition of X is given as the following:

$X = USV^t$ where U is a $m \times n$ matrix, S is $n \times n$ diagonal singular matrix, and V is $n \times n$ matrix. Let $S = \text{diag}(s_1, s_2, \dots, s_n)$. Also we have $s_i \geq s_{i+1}$ for $i=1$ to $(n-1)$. All the variance in the data set in X is captured by all the non zero members, say s_1 through s_r of the singular values of S . To capture 90% of the variance of X , the first r components of the singular values are selected such that $\sum_{k=1}^{k=r} s_k^2 > 0.9 * \sum_{k=1}^{k=m} s_k^2$.

Suppose we want to fuse different relevant features, say f_1, f_2, \dots, f_r pertaining to the gene function prediction of m genes and these are represented by $m_{f_1}, m_{f_2}, \dots, m_{f_r}$ matrices respectively. Each feature matrix is normalized and the composite matrix X is created by column concatenation of m_{f_1}, m_{f_2}, \dots , and m_{f_r} . Clustering or appropriate classification is done on the projected reduced dimension and it seems to have a better results in a limited study done by us [12] for yeast genome.

IV. BINDING SITES

The JASPAR CORE database [13, 14] contains a curated, non-redundant set of 138 profiles of binding sites of 18 eukaryotes and provides open data access. All these profiles are derived from published collections of experimentally defined transcription factor binding sites for multi-cellular eukaryotes. We preferred to use Jaspas for its open access policy even though TRANSFAC [15] provides larger collection of binding sites.

V. DATA

We are focusing on genes of Homo sapiens and their expressions for this experiment. From Affymetrix site at <http://www.affymetrix.com>, we have downloaded the annotations (HG-U133A_2.na22.annot) for the genes that are tested in a microarray experiment. To predict the binding sites, we have downloaded the known or experimentally determined promoter sequences (-1000 to +300 bp) from computational biology and bioinformatics lab in Cold Spring Harbor Laboratory at <http://rulai.cshl.edu/>.

Jun Chung and his associates at the LSU health Science Center of Shreveport had conducted a microarray experiment using the affymetrix HG-U133A_2 to identify transcriptional targets of integrin $\alpha 6 \beta 4$. The goal of their study is to identify $\alpha 6 \beta 4$ transcriptional targets important for breast cancer progression. The $\alpha 6 \beta 4$ integrin, an epithelial-specific integrin, functions as a receptor for the members of the laminin family of extracellular matrix proteins. While the primary known function of $\alpha 6 \beta 4$ is to contribute to tissue integrity through its ability to mediate the formation of hemidesmosomes (HD) there is growing evidence suggesting that this integrin also

plays a pivotal role in functions associated with cancer progression. For example, high expression of this integrin in women with breast cancer has been shown to correlate significantly with mortality and disease states. However, therapeutic targets of breast cancer that over-express $\alpha6\beta4$ are not yet well characterized. For this reason, it is essential to elucidate the mechanism by which $\alpha6\beta4$ contributes to breast cancer progression.

Our study here describes the gene expression profile obtained from MDA-MB-435 mock transfectants ($\alpha6\beta4$ negative breast carcinoma cell line) and MDA-MB-435 b4 integrin transfectants ($\alpha6\beta4$ positive breast carcinoma cell lines). Out of oligonucleotide probe sets representing approximately 22,277 genes, expression of b4 integrin in MDA-MB-435 cells upregulates 149 genes by two fold or higher. 193 genes are down regulated by over two fold change. We anticipate that microarray data will lead to not only the identification of $\alpha6\beta4$ target genes that are important for breast cancer cell growth, survival and invasion, but also discover signaling pathways lead to the expression of these genes.

A. Data preprocessing

The experiment is repeated three times and in each repetition the expressions of genes under the following two conditions are measured: (1) integrin negative cell line (control), and (2) integrin positive cell line. Out of the 22,277 genes we have selected only 8,512 genes that have valid signal in all measurements. The average of the log ratio between the integrin positive and the control expression in all the repetitions is taken as the expression of a gene. From the expressions, we could create different expression patterns based on the values such as up regulated fold changes over 2 to 3, 3 to 4, and over 4. Among the down regulated genes, we may have the similar groups. For simplicity, we have taken only two patterns namely up regulated, and down regulated genes. The up regulated genes are those that have fold changes over 2 and the down regulated are those that have less than 0.5 fold changes.

VI. OUR APPROACH

A. Relationship between expression and functional annotation

Let us start with focusing on the ability of functional annotations in discriminating expression patterns; the up regulated and the down regulated groups. The functional annotations at level 2 of the up and down regulated genes are obtained from the GO ontology. There are 164 molecular functional terms at level 2 of molecular function, and 278 biological process terms at level 2 of biological processes. Not all the selected genes are associated with many of the selected terms. The functional terms that are associated with over 5% of the selected genes are retained, and the genes that do not have any annotations in the retained terms are also eliminated in this analysis.

After applying the filtering, we are left with the following 41 biological process terms for analysis: regulation of molecular function, establishment of localization, positive

regulation of biological process, macromolecule metabolic process, regulation of metabolic process, system process, establishment of localization in cell, primary metabolic process, response to chemical stimulus, anatomical structure development, negative regulation of cellular process, cell motility, cell cycle, cellular localization, cell communication, cellular metabolic process, regulation of cellular process, cell cycle process, anatomical structure formation, biosynthetic process, response to stress, positive regulation of cellular process, cell proliferation, catabolic process, anatomical structure morphogenesis, cellular developmental process, negative regulation of biological process, gene expression, macromolecule localization, regulation of biological process, cell adhesion, transport, cellular component organization and biogenesis, response to external stimulus, regulation of developmental process, cell development, establishment of protein localization, multicellular organismal development, localization of cell, death, and regulation of biological quality.

Applying the similar filtering leaves the following 10 functional terms for further analysis: transferase activity, nucleic acid binding, signal transducer activity, protein binding, oxidoreductase activity, nucleotide binding, substrate-specific transporter activity, ion binding, transmembrane transporter activity, and hydrolase activity.

To determine the extent to which either one of these functional groups can identify or discriminate expression patterns, we have applied the following set of well established machine learning algorithms[16]: naive Bayes, random forest and J48 (variation of C4.5 decision tree). Each gene is represented by a feature vector of either the retained molecular function terms, or biological process terms. Suppose a gene, say g_k , has functional annotations, say f_1 and f_3 , out of 5 retained functional term. The corresponding feature vector of gene g_k is represented as

$g_k^T = (1,0,1,0,0)$. We use 10 fold cross validation which involves randomly generating 10 different partition of the labeled feature vectors and training the model with 9 partition and test with the other one partition. This experiment is continued 10 times so that every instance will be participating as a training instance and as well as a testing instance in different cycles of the process. The confusion matrix of each machine learning algorithm with 10 fold cross validation is given in tables 1 through 3.

TABLE I
CONFUSION MATRIX OF NAÏVE BAYES CLASSIFICATION WITH 10 FOLD CROSS VALIDATION WHEN MOLECULAR FUNCTIONAL FEATURES ARE USED FOR PREDICTING EXPRESSION PATTERNS.

	Pred. up reg.	Pred. dn reg.
True up reg. genes	42	77
True dn reg. genes	35	114

The prediction accuracy of Naïve Bayes with functional features is 58.21%.

TABLE 2
CONFUSION MATRIX OF J48 CLASSIFICATION WITH 10 FOLD
CROSS VALIDATION WHEN MOLECULAR FUNCTIONAL
FEATURES ARE USED FOR PREDICTING EXPRESSION
PATTERNS.

	Pred. up reg.	Pred. dn reg.
True up reg. genes	23	96
True dn reg. genes	29	120

The prediction accuracy of J48 with functional features is 53.36%.

TABLE 3
CONFUSION MATRIX OF RANDOM FOREST WITH 10 FOLD
CROSS VALIDATION WHEN MOLECULAR FUNCTIONAL
FEATURES ARE USED FOR PREDICTING EXPRESSION PATTERNS.

	Pred. up reg.	Pred. dn reg.
True up reg. genes	51	68
True dn reg. genes	51	98

The prediction accuracy of random forest with functional features is 55.6%.

Similarly, we have created feature vectors of genes using biological process terms and have used the same classifiers with 10 fold cross validation. The confusion matrix for the classifiers are given in Tables 4 through 6.

TABLE 4
CONFUSION MATRIX OF NAÏVE BAYES CLASSIFICATION WITH
10 FOLD CROSS VALIDATION WHEN BIOLOGICAL PROCESSES
FEATURES ARE USED FOR PREDICTING EXPRESSION PATTERNS.

	Pred. up reg.	Pred. dn reg.
True up reg. genes	67	47
True dn reg. genes	66	88

The prediction accuracy of Naïve Bayes with functional features is 57.8%.

TABLE 5
CONFUSION MATRIX OF J48 CLASSIFICATION WITH 10 FOLD
CROSS VALIDATION WHEN BIOLOGICAL PROCESSES FEATURES
ARE USED FOR PREDICTING EXPRESSION PATTERNS.

	Pred. up reg.	Pred. dn reg.
True up reg. genes	64	50
True dn reg. genes	56	98

The prediction accuracy of J48 with functional features is 60.45%.

TABLE 6
CONFUSION MATRIX OF RANDOM FOREST WITH 10 FOLD
CROSS VALIDATION WHEN MOLECULAR FUNCTIONAL
FEATURES ARE USED FOR PREDICTING EXPRESSION PATTERNS.

	Pred. up reg.	Pred. dn reg.
True up reg. genes	69	45
True dn reg. genes	55	99

The prediction accuracy of random forest with functional features is 62.69%.

We will summarize the prediction accuracy of the three classifiers when a gene is represented with either the biological process, or molecular function feature vector.

TABLE 7
THE PREDICTION ACCURACY OF THE CLASSIFIERS FOR EACH
VECTOR REPRESENTATION WHEN APPLIED WITH 10 FOLD CROSS
VALIDATION.

Classifiers	Prediction accuracy	
	Molecular function feature vector	Biological process feature vector
Naïve Bayes	58.12%	57.8%
J48	53.36%	60.45%
Random forest	55.6%	62.69%

The results clearly indicate that neither the biological process nor the molecular function alone has good predictive power or to discriminate expression patterns. Let us normalize the feature vector and combine them and test the predictive power of the expression pattern. The results of the normalized vector and corresponding confusion matrices are shown in Tables 8 and 9. We have not used naïve Bayes classifier for feature vectors with floating point value.

TABLE 8
CONFUSION MATRIX OF J48 CLASSIFICATION WITH 10 FOLD
CROSS VALIDATION WHEN BIOLOGICAL PROCESSES FEATURES
ARE USED FOR PREDICTING EXPRESSION PATTERNS.

	Pred. up reg.	Pred. dn reg.
True up reg. genes	67	36
True dn reg. genes	52	76

The prediction accuracy of J48 with functional features is 61.9%.

TABLE 9
CONFUSION MATRIX OF RANDOM FOREST WITH 10 FOLD
CROSS VALIDATION WHEN MOLECULAR FUNCTIONAL
FEATURES ARE USED FOR PREDICTING EXPRESSION PATTERNS.

	Pred. up reg.	Pred. dn reg.
True up reg. genes	67	36
True dn reg. genes	44	84

The prediction accuracy of random forest with functional features is 65.37%.

When combining the normalized feature vectors of biological process with molecular function, the overall prediction has improved while the size of feature vector has increases to 51. To reduce the dimensionality and to reduce redundancy in the data set, we use SVD to the gene expression matrix M as USV^t . The variance of M is reflected in the singular values of S . Throughout the experiment, we select the singular values so as to cover 90% of the variance of the data set. For the data set, the first seven singular values cover more than 90% of the variance. The projected feature vector has size of 23. Using the reduced feature vector, we ran the classifiers with 10 fold cross validation and the corresponding the confusion matrices are given in tables 10 and 11.

TABLE 10

CONFUSION MATRIX OF J48 CLASSIFICATION WITH 10 FOLD CROSS VALIDATION WHEN BIOLOGICAL PROCESSES FEATURES ARE USED FOR PREDICTING EXPRESSION PATTERNS.

	Pred. up reg.	Pred. dn reg.
True up reg. genes	76	27
True dn reg. genes	60	68

The prediction accuracy of J48 with functional features is 62.33%.

TABLE 11

CONFUSION MATRIX OF RANDOM FOREST WITH 10 FOLD CROSS VALIDATION WHEN MOLECULAR FUNCTIONAL FEATURES ARE USED FOR PREDICTING EXPRESSION PATTERNS.

	Pred. up reg.	Pred. dn reg.
True up reg. genes	68	35
True dn reg. genes	44	84

The prediction accuracy of random forest is 65.8%

In table 12, we summarize the prediction accuracy of the three classifiers when a gene is represented with either the biological process, or and molecular function feature vector.

TABLE 12

THE SUMMARY OF PREDICTION ACCURACY OF THE 3 CLASSIFIERS WITH DIFFERENT FEATURE VECTORS. ALL THE RESULTS ARE OBTAINED FOR 10 FOLD CROSS VALIDATION. MF AND BP RESPECTIVELY STANDS FOR MOLECULAR FUNCTION AND BIOLOGICAL PROCESS.

Classifiers	Prediction Accuracy in percentage			
	mf. alone	Bp alone	Normalized mf and bp	SVD proj 90% mf. and bp
Naïve bayes	58.12	57.8	-	-
J48	53.36	60.45	61.9	62.34
Random forest	55.6	62.69	65.37	65.8

We cluster the genes using the reduced feature vectors using the following clustering algorithms (1) hierarchical clustering, (2) k-nearest neighbor and (3) self organization map. We

prefer hierarchical clustering for its visual appealing and interactive user control.

The result of hierarchical clustering of projected genes using Pearson correlation coefficient, and average cluster distance with similarity index over 0.788 is shown in Figure 1. This resulted in 28 clusters and out of which 17 clusters have 5 or more genes.

B. Relationship between expression and co-regulation

The genes that are regulated together is said to be co-regulated and therefore have a common subset of transcription factors. Therefore, the co-regulated genes must share binding sites. In the absence of experimentally determined binding sites for the differentially expressed genes, we determine the common binding sites computationally. We have downloaded the known promoter sequences of Homo sapiens from computational biology and bioinformatics lab in Cold Spring Harbor Laboratory. We have also downloaded profiles of common binding sites of Homo sapiens from Jaspar, which provides open access to all the data sets. The profile of a binding site is a matrix, say M , in which each row corresponds to one of the four nucleotides $a, c, g, \text{ or } t$ while each column represents the position of the binding site and an entry, say $M[a, j]$, denotes the counts of nucleotide a in position j . The background probability is obtained by scanning the promoter regions of the downloaded sequences. Using the background probability and the profile, we have computed positional specific weighted matrix (PSWM) of each common core binding sites.

A putative binding sites, say BS_k of length m , is determined or located in a promoter sequence when the total weights of a subsequence of length m exceeded the threshold that was set for the binding site. Setting a proper threshold for determining putative binding site is very important since lower threshold will lead to false positive while higher value may lead to excluding potential binding sites. Through simulations, we have set appropriate threshold so as to achieve putative binding sites with respectable lower p vale. The selected 49 common binding sites and the threshold are given in Table 13.

TABLE 13

THE DETAILS OF BINDING SITES BEING USED FOR THE STUDY.

Binding Sites	Threshold	Length
GATA2	7.30	5
GATA3	9.16	6
MZF1_1-4	10.28	6
SPI1	9.32	6
YY1	9.38	6
ETS1	8.48	6
ZNF354C	10.05	6
SPIB	11.43	7
USF1	12.55	7
NKX3-1	13.23	7
BRCA1	9.01	7
E2F1	14.80	8

FOXC1	8.82	8
FOXD1	14.23	8
FOXL1	10.11	8
ELK4	16.28	9
Lhx3	15.70	9
SOX9	13.95	9
SRY	13.18	9
TFAP2A	10.52	9
ELK1	12.88	10
GABPA	16.34	10
MAX	14.59	10
MEF2A	18.67	10
MIZF	16.15	10
MZF1_5-13	13.92	10
REL	14.29	10
RELA	17.02	10
RORA_1	15.95	10
SP1	13.56	10
MYC-MAX	16.15	11
NFIL3	16.53	11
NFKB1	17.04	11
CREB1	14.77	12
FOXI1	15.48	12
HLF	15.30	12
NHLH1	16.51	12
IRF1	17.64	12
Myf	17.12	12
PBX1	16.76	12
SRF	18.33	12
TEAD1	17.40	12
TAL1-TCF3	16.51	12
NR2F1	16.66	14
FOXF2	16.34	14
Pax6	16.08	14
RORA_2	18.91	14
TLX1-NFIC	19.67	14
STAT1	18.55	14

We were able to get the promoter sequences only for a subset of differentially expressed genes. We search for putative binding sites in the upstream (-500bp) of the transcription start sites and constructed binding site matrix. From the matrix we have filtered out the binding sites that were found less than 1% of the promoter sequences. The binding sites reduced to the following 13: GATA2, GATA3, ETS1, ZNF354C, SPIB, USF1, E2F1, RORA_1, SP1, CREB1, FOXI1, Myf, and TAL1-TCF3. Also we have eliminated the genes that do not have any binding sites. With the filtering, the set of genes that associated with some binding sites reduced to 83 out off which 38 are up regulated and the remaining 45 are down regulated. The p-values and the percentage of sequences having these putative binding sites are given in table 14.

TABLE 14
BINDING SITE INFORMATION WITH P-VALUE.

Binding site	Width	Sequence %	P-value
GATA2	5	42.17%	3.74E-01
GATA3	6	4.82%	9.74E-02
ETS1	6	21.69%	2.85E-01
ZNF354C	6	10.84%	1.26E-01
SPIB	7	15.66%	2.49E-02
USF1	7	15.66%	3.26E-02
E2F1	8	2.41%	9.32E-03
RORA_1	10	1.20%	9.82E-04
SP1	10	6.02%	2.06E-03
CREB1	12	1.20%	3.95E-03
FOXI1	12	1.20%	3.86E-03
Myf	12	1.20%	1.86E-03
TAL1-TCF3	12	1.20%	2.30E-03

To test the relationship between the expression patterns and the co-expression, we represent each gene with binding sites as the feature vector along with its label: up or down regulated. Similar to the previous experiment, we have applied the three classifiers namely the Naïve Bayes, J48, and random forest with 10 fold cross validation. The confusion matrices are given in Tables 14 through 17.

TABLE 15
CONFUSION MATRIX OF NAÏVE BAYES CLASSIFICATION WITH 10 FOLD CROSS VALIDATION WHEN BINDING SITES ARE USED FOR PREDICTING EXPRESSION PATTERNS.

	Pred. up reg.	Pred. dn reg.
True up reg. genes	12	26
True dn reg. genes	22	23

The prediction accuracy of Naïve Bayes with binding site features is 42.16 %.

TABLE 16
CONFUSION MATRIX OF J48 CLASSIFICATION WITH 10 FOLD CROSS VALIDATION WHEN BINDING SITES ARE USED FOR PREDICTING EXPRESSION PATTERNS.

	Pred. up reg.	Pred. dn reg.
True up reg. genes	8	30
True dn reg. genes	12	33

The prediction accuracy of J48 with binding site features is 49.4%.

TABLE 17:
CONFUSION MATRIX OF RANDOM FOREST WITH 10 FOLD CROSS VALIDATION WHEN BINDING SITES ARE USED FOR PREDICTING EXPRESSION PATTERNS.

	Pred. up reg.	Pred. dn reg.
True up reg. genes	14	24
True dn reg. genes	15	30

The prediction accuracy of random forest with binding site features is 53%.

VII. SUMMARY AND DISCUSSION

As the cost on conducting microarray experiment has come down, the numbers of array experiments have increased so as to obtain biomarkers, to study the underlying mechanism of regulation, or to obtain annotation of a gene from another set of annotated genes. It is important to understand the relationships among co-regulation, functional annotation and co-regulation.

We have outlined a general approach to study the relations among gene expression patterns, annotations and co-regulation. With the lack of experimentally determined co-regulated information, we assume that genes sharing the binding sites are the candidates of being co-regulated. We have been using molecular function and biological processes terms from gene ontology as the annotation of genes.

To illustrate the approach, we have been using the micro array data set of Jun Chung and his associates for identifying transcriptional targets of integrin $\alpha\beta4$. The goal of their study is to identify $\alpha\beta4$ transcriptional targets important for breast cancer progression. The expression of genes with $\alpha\beta4$

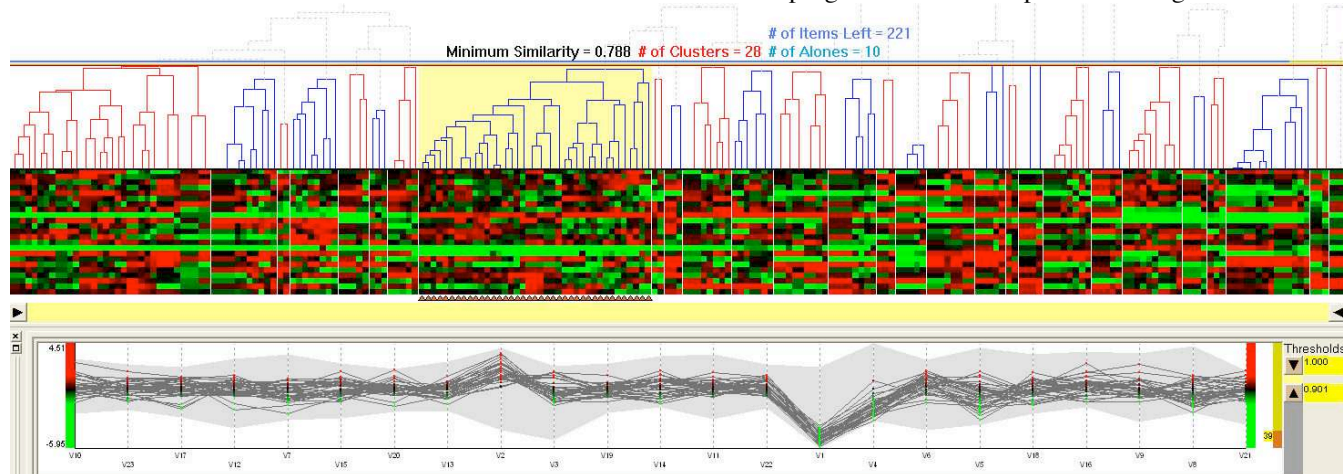


Figure 1: Hierarchical cluster of projected feature vectors of biological processes and molecular function

positive breast carcinoma cell lines and $\alpha\beta4$ negative breast carcinoma cell line (control) were obtained. The log of the expression ratio ranges from -3.55 to 2.88 (fold changes from 0.085 through 7.36). We were focused on the differentially expressed genes whose fold changes over 2 (up regulated) and as well as those having fold changes less than 0.5 (log ratio less than -1). For this experiment, the expression pattern could have based on the ranges of fold changes, such as, 2 to 3, 3 to 4, and over 4. Or simply up regulated and down regulated. We have taken the simple approach of having two patterns namely up regulated and down regulated.

Each gene is associated with many functional terms of varying degrees of specificity. In the GO terms hierarchy the terms that is furthest away from the root (either the biological process term, or molecular function term) are the most specific compared to the ones closer to the root. A term, of course, inherits all the annotations of its predecessors. For our analysis, we have considered the annotation of terms at the 2nd level of the hierarchy, which has 164 molecular functional terms, and 278 biological processes terms). The differentially expressed genes have 41 biological terms and 10 molecular functional terms.

To study the relationship between the expression patterns and the functional annotations, we modeled a gene as a feature vector of either the molecular function or biological processes. If there is a close relationship between the expression patterns and the associated function, then one expects the feature vectors will provide a clear separation between the expression patterns, in other word, prediction accuracy of expression

pattern will be higher when using classifiers with the feature vector. The random forest yield the best prediction accuracy of 55.6% with molecular function feature vector, and 62.69 with biological processes as feature vector. The details are described in Section VI A. When we have normalized and combine these feature vectors, the prediction accuracy increased to 65.37 along with an increase of feature vector dimension, which is 51. We have applied data fusion based on SVD[12] to capture the variation up to 90% that reduced the dimension to 23. The prediction accuracy with the projected feature vector is 65.8 which is slightly better than the previous results of concatenation of feature vectors of molecular functions and biological processes.

We take the projected feature vectors and cluster the genes using hierarchical clustering algorithm [17] using Pearson correlation coefficient, and average cluster distance with similarity index over 0.788. The illustrative figure of the cluster is shown in Figure 1. The algorithm produced 28 clusters out of which 17 have more than 5 genes and these 17 clusters are taken for further study.

In the absence of experimentally determined binding sites, we have computed putative binding sites of those genes that have known promoter sequences. The binding sites that have over 1% sequence coverage were retained for the study. The small motifs of length 5 or 6 have very high p value when scanning 500bp of promoter sequences, which is often the problem with the short motifs. We provide the details of the binding site coverage and the p-values in table 14. The genes that share binding sites indirectly related to co-regulated genes.

To study the relationship between the expression patterns and the co-regulation, we modeled a gene with binding site as a feature vector along with expression pattern label as either up regulated or down regulated, and use the set of classifiers; Naïve bayes, J48 and random forest. The confusion matrices are given in Tables 14 through 17. The best prediction accuracy with binding site feature vector is 53%, which is quite low.

VIII.CONCLUSIONS

It is undisputed that microarray technology has been successfully used for determining biomarkers; highly up regulated and as well as highly down regulated genes may be used as possible bio-markers. On the other hand, inferring either functional annotations or co-regulation with high confidence from similar expression patterns is questionable. However, combining or fusing different modalities of data and cluster the genes seem to have a better cohesive cluster so as to create new hypothesis for further study. We plan to combine the binding site information along with molecular function and biological process annotation and cluster the genes. The binding sites that we have considered in this study may be limited to what was available at the Jaspar site. We plan to extend these binding sites from other resources.

The approach we have proposed is general and can be applicable to any other array experiment with rich expression patterns.

REFERENCES

- [1] M. A. Harris, J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall, J. Richter, G. M. Rubin, J. A. Blake, C. Bult, M. Dolan, H. Drabkin, J. T. Eppig, D. P. Hill, L. Ni, M. Ringwald, R. Balakrishnan, J. M. Cherry, K. R. Christie, M. C. Costanzo, S. S. Dwight, S. Engel, D. G. Fisk, J. E. Hirschman, E. L. Hong, R. S. Nash, A. Sethuraman, C. L. Theesfeld, D. Botstein, K. Dolinski, B. Feierbach, T. Berardini, S. Mundodi, S. Y. Rhee, R. Apweiler, D. Barrell, E. Camon, E. Dimmer, V. Lee, R. Chisholm, P. Gaudet, W. Kibbe, R. Kishore, E. M. Schwarz, P. Sternberg, M. Gwinn, L. Hannick, J. Wortman, M. Berriman, V. Wood, N. de la Cruz, P. Tonellato, P. Jaiswal, T. Seigfried, and R. White, "The Gene Ontology (GO) database and informatics resource," *Nucleic Acids Res*, vol. 32, pp. D258-61, 2004.
- [2] K. Tumer and J. Ghosh, "Estimating the Bayes Error through classifier combining," presented at International Conference on Pattern Recognition, Vienna, Austria, 1996.
- [3] G. Fumera and F. Roli, "A theoretical and experimental analysis of linear combiners for multiple classifier systems," *IEEE Trans Pattern Anal Mach Intell*, vol. 27, pp. 942-56, 2005.
- [4] L. I. Kuncheva, "A Theoretical Study on Six Classifier Fusion Strategies," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, pp. 281-286, 2002.
- [5] N. Cristianini and J. Shawe-Taylor, *An introduction to support vector machines: and other kernel-based learning methods*. Cambridge, U.K.; New York: Cambridge University Press, 2000.
- [6] B. Schölkopf and A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. Cambridge, Mass.: MIT Press, 2002.
- [7] G. R. Lanckriet, T. De Bie, N. Cristianini, M. I. Jordan, and W. S. Noble, "A statistical framework for genomic data fusion," *Bioinformatics*, vol. 20, pp. 2626-35, 2004.
- [8] C. G. Chute and Y. Yang, "An evaluation of concept based latent semantic indexing for clinical information retrieval," *Proc Annu Symp Comput Appl Med Care*, pp. 639-43, 1992.
- [9] S. Deerwater, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman., "Indexing by latent semantic analysis," *Journal of the Society for Information Science*, vol. 41, pp. 391-407, 1990.
- [10] R. Homayouni, K. Heinrich, L. Wei, and M. W. Berry, "Gene clustering by latent semantic indexing of MEDLINE abstracts," *Bioinformatics*, vol. 21, pp. 104-15, 2005.
- [11] Y. Yuan, L. Lin, Q. Dong, X. Wang, and M. Li, "A Protein Classification Method Based on Latent Semantic Analysis^{*}," *Conf Proc IEEE Eng Med Biol Soc*, vol. 7, pp. 7738-41, 2005.
- [12] R. Loganathanaraj, "Beyond Clustering of Array Expressions," presented at Biotechnology and Bioinformatics Symposium, Colorado Spring, Colorado, 2007.
- [13] D. Vlieghe, A. Sandelin, P. J. De Bleser, K. Vleminckx, W. W. Wasserman, F. van Roy, and B. Lenhard, "A new generation of JASPAR, the open-access repository for transcription factor binding site profiles," *Nucleic Acids Res*, vol. 34, pp. D95-7, 2006.
- [14] J. C. Byrne, E. Valen, M. H. Tang, T. Marstrand, O. Winther, I. da Piedade, A. Krogh, B. Lenhard, and A. Sandelin, "JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update," *Nucleic Acids Res*, vol. 36, pp. D102-6, 2008.
- [15] TRANSFAC, "The Transcription Factor Database <http://www.gene-regulation.com/pub/databases.html>."
- [16] I. H. Witten and E. Frank, *Data mining: practical machine learning tools and techniques*, second ed. San Francisco, Calif.: Morgan Kaufmann, 2005.
- [17] J. Seo, H. Gordish-Dressman, and E. P. Hoffman, "An interactive power analysis tool for microarray hypothesis testing and generation," *Bioinformatics*, 2006.