A fuzzy entropy based approach for development of Gene Prediction Networks (GPNs): Detecting altered dependency in carcinogenic state

Anupam Ghosh
Department of Computer Science and Engineering,
Netaji Subhash Engineering College,
Kolkata, India
anupam.ghosh@rediffmail.com

Rajat K. De *
Machine Intelligence Unit,
Indian Statistical Institute,
Kolkata, India
rajat@isical.ac.in

ABSTRACT

In this article, the dependencies among the genes have been identified from microarray gene expression data. Here we propose a methodology for identifying the dependencies among the genes that have deviated quite significantly from normal stage to diseased stage with respect to their expression patterns. This idea leads to predict the disease mediating genes along with their deviated dependencies. The proposed methodology involves measuring information content of individual genes using fuzzy entropy, conditional fuzzy entropy of a gene on another, dependencies of a pair of genes in both normal and diseased states, and finally identifying the dependencies being deviated from normal to carcinogenic state. The deviated dependencies among the genes have been represented using a network, called gene prediction network (GPN), in which each node represents a gene and a directed edge signifies deviated dependency between a pair of nodes (genes).

The methodology has been demonstrated on two gene expression data sets dealing with human lung cancer and breast cancer. The results are appropriately validated by earlier investigations in terms of gene regulation. We have also used some statistical techniques like t-test, accuracy in terms of sensitivity and specificity to validate the results.

Keywords

Fuzzy set, Entropy, GPN, True Positive, t-test

1. INTRODUCTION

Transcriptional regulatory networks are crucial in the understanding of fundamental cellular processes and functions. The determination of factors that control expression level can offer further insight into the miss regulated expression that is common in many human diseases [19, 9]. There exist many investigations on identifying transcription factors including those through sequence similarity [13, 20], motif binding [14, 8, 4, 5] and through the dynamics of gene expression patterns [16, 18].

Using microarray data, reverse engineering of gene regulatory networks is one of the essential roles in elucidating transcriptional systems. Various statistical approaches have been developed to capture gene regulations using dynamic Bayesian networks [7, 11], vector autoregressive models [3], and state space models [6, 21] based on statistical causality, among others. Traditional approaches include formulation of a set of coupled differential equations and their solutions, with the objective to obtain a deeper understanding of the exact nature of the regulatory circuits and their regulation mechanisms [17]. Various alternative methods, like relevance networks [2] and graphical Gaussian models [15] have been proposed and applied to the inference of gene regulatory networks from gene expression data. Thus many investigations are there for predicting gene regulatory networks. But we could not find any investigations on the dependencies among the genes, in terms of regulation, which have changed from normal to diseased state.

In the present article, we concentrate on dependencies among the genes obtained from microarray gene expression patterns. Here we develop a methodology for identifying the dependencies among the genes, which have changed from normal to diseased state. In this way, we can predict disease mediating genes along with their altered dependencies. The methodology involves measuring information content of individual genes using fuzzy entropy, conditional fuzzy entropy of a gene on another, dependencies of a pair of genes in both normal and diseased states, and finally identifying the dependencies that have altered from normal to diseased state. These deviated dependencies have been shown using a network in which nodes representing genes and directed edges representing deviated dependencies between pairs of nodes. We call this network as Gene Prediction Network (GPN). The effectiveness of the methodology has been demonstrated on two gene expression data dealing with human lung cancer, breast cancer. The results have been validated with some existing investigations as well as some statistical parameters.

*corresponding author

ACM-BCB '11, August 1-3, Chicago, IL, USA

2. METHODS

Here, we formulate the methodology for developing a gene prediction network based on the gene expression patterns of normal and diseased samples. First of all, two gene dependency matrices (GDMs) are formed for gene expression profiles of normal and diseased samples respectively. Based on these two matrices, a prediction network is created involving the genes, where dependency has been changed from normal to diseased state. This prediction network provides an idea on the genes along with their dependencies for mediating carcinogenic development. In other words, the network may predict the responsible genes and their altered behavior. Then we find out some possible genes mediating the development of cancer. We have used fuzzy set theoretic entropy to determine the gene dependency matrices. Using these matrices, we build the gene prediction network.

The term *gene dependency* may be defined as follows. Let us consider two genes g_i and g_j. If the change in expression level of gene g_j causes the change in expression level of gene g_i, then we say that gene g_i depends on gene g_j. In other words, gene g_j regulates the expression level of gene g_i. Thus we have an $n \times n$ matrix GDM corresponding to the gene dependencies involving n genes. The (i, j)th entry of GDM represents the degree of dependency of g_i on g_j. In this way, GDMs are formed for normal samples as well as for diseased samples. In diseased state, this dependency may change from normal ones. From these two gene dependency matrices, we may have an idea of a gene regulatory network depicting transcriptional regulation for normal and diseased states.

The methodology involves five steps. In Step 1, the information content of a gene is computed using the notion of fuzzy entropy from gene expression data set. The conditional entropy of a pair of genes is computed in Step 2. In Step 3, the information gain is computed pairwise and also generates the gene dependency matrix using the concept of symmetrical uncertainty. The values of the gene dependency matrix are quantized, in Step 4, using a threshold. Finally, in Step 5, the gene prediction matrix and the corresponding network are generated.

2.1 Step 1 – Measuring information content (entropy) of a gene

Let us consider an expression data for a set of these n genes $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, for each of which m expression values are given. Let G be the set of n genes $\{g_1, g_2, \ldots, g_n\}$. For each gene g_i, there is an m-dimensional vector $\mathbf{x}_i$, where x_{il} is the l-th expression value of g_i. Let us also consider a set of m microarray experiments (measurements) $\mathbf{Y} = \{e_1, e_2, \ldots, e_m\}$. For each experiment, we have n expression values corresponding to n genes in G.

Here we consider a fuzzy set around a gene g_i and the membership function $U_{g_i}(l)$ signifies the degree of belongingness of the lth expression value of gene g_i to this fuzzy set. In other words, $U_{g_i}(l)$ represents the extent by which g_i is expressed in lth sample. Then we compute the entropy (uncertainty) associated with this fuzzy set, i.e., associated with gene g_i, as

$$H(g_i) = \sum_{l=1}^{m} U_{g_i}(l)(1 - U_{g_i}(l)) \quad (1)$$

Here the membership function $U_{g_i}(l) \in [0, 1]$ is defined as

$$U_{g_i}(l) = \exp(-|x_{il} - \bar{x_i}|) \quad (2)$$

where $\bar{x_i}$ is given by

$$\bar{x_i} = 1/m \sum_{l=1}^{m} x_{il} \quad (3)$$

2.2 Step 2 – Measuring conditional entropy of a gene on another

In the previous step, we have calculated the uncertainty associated with each gene, i.e., information content of an individual gene. Now the entropy associated with gene g_i given that gene g_j has attained some expression value, is defined as

$$H(g_i|g_j) = \sum_{l=1}^{m} U_{g_j}(l) \sum_{l'=1}^{m} U_{g_i|g_j}(l')(1 - U_{g_i|g_j}(l')) \quad (4)$$

where $U_{g_i|g_j}(l) \in [0, 1]$ is defined as

$$U_{g_i|g_j}(l) = \exp(-||\mathbf{s}_{ij}(l) - \bar{\mathbf{s}}_{ij}||)/\exp(-|x_{jl} - \bar{x}_j|) \quad (5)$$

Here $\mathbf{s}_{ij}(l) = [x_{il},\ x_{jl}]^T$ and $\bar{\mathbf{s}}_{ij} = [\bar{x}_i,\ \bar{x}_j]^T$. That is, we are writing $U_{g_i|g_j}(l)$ as $U_{g_i g_j}(l)/U_{g_j}(l)$, where $U_{g_i g_j}(l)$ is a two dimensional membership function of the fuzzy set formed by genes g_i and g_j together.

2.3 Step 3 – Measuring information gain and building gene dependency matrix

The entropy $H(g_i)$ of the gene g_i has two parts. The part $H(g_i|g_j)$ represents the entropy of gene g_i given that gene g_j has attained some expression value. The remaining part represents the information gain of gene g_i provided by gene g_j, and is defined as [12],

$$IG(g_i|g_j) = H(g_i) - H(g_i|g_j) \quad (6)$$

The amount by which the entropy of g_i decreases reflects additional information about g_i provided by g_j and is called information gain. According to this measure, a gene g_j is regarded as more correlated to gene g_i than gene g_k, if $IG(g_i|g_j) > IG(g_k|g_j)$. Information gain may be biased towards the genes with more expression values. We normalize IG-values to get Normalized IG-values or NIG-values as [12],

$$NIG(g_i, g_j) = 2 \times |IG(g_i|g_j)|/(H(g_i) + H(g_j)) \quad (7)$$

$NIG(g_i, g_j)$ signifies the strength of dependency, i.e., to what extent the expression of gene g_i depends on the expression value of g_j. $NIG(g_i, g_j) = 1$ indicates that knowledge of the expression level of either one completely predicts that of the other, and $NIG(g_i, g_j) = 0$ indicating that g_i and g_j are independent. With these NIG-values, we get the gene dependency matrix GDM of order $n \times n$ and with (i, j)th element as $NIG(g_i, g_j)$.

2.4 Step 4 – Quantizing the elements of gene dependency matrix

All entries of GDM are in $[0, 1]$ and provide the degrees of dependency. Here we introduce three types of dependencies (i.e., low, medium, high) between a pair of genes. To represent each type, we have to fix a threshold value to generate the quantized gene dependency matrix $D = [d_{ij}]_{n \times n}$.

The domain of the matrix elements is divided into three partitions: [0,α] as the first partition, (α,β] as the second one, and (β,1] as the third partition. In this case, α and β are treated as user defined thresholds. Now d_{ij} values, where $i \neq j$, are defined as

$$\begin{aligned} d_{ij} &= 0.0 \quad \text{if } 0 \leq NIG(g_i, g_j) \leq \alpha \\ &= 0.5 \quad \text{if } \alpha < NIG(g_i, g_j) \leq \beta \\ &= 1.0, \quad \text{if } \beta < NIG(g_i, g_j) \leq 1, \end{aligned} \tag{8}$$

and $d_{ii} = 1$, for all i. Thus the quantized gene dependency matrix is formed with the values 0.0, 0.5 and 1.0. Note that, $d_{ij} = 1.0$ means gene g_i is highly dependent on gene g_j. $d_{ij} = 0$ signifies that the dependency of g_i on g_j is low, and $d_{ij} = 0.5$ implies that this dependency is medium.

2.5 Step 5 – Building gene prediction network

In the previous step, we have defined gene dependency matrix D. This matrix is computed for control (normal) samples as well as for test (diseased) samples. We denote the gene dependency matrix by $D^{(N)}$ for normal samples, and by $D^{(D)}$ for diseased samples.

Using these two matrices, we generate a matrix $P = D^{(N)} - D^{(D)}$ of order $n \times n$. The matrix $P = [p_{ij}]_{n \times n}$ consists of the values of 0.0, -1.0, 1.0, -0.5 and 0.5. The significance of these values are discussed below.

$p_{ij} = 0.0$: In this case, $d_{ij}^{(N)} = d_{ij}^{(D)}$. That is, the diseased state does not affect the dependency of gene g_i on gene g_j.

$p_{ij} = -1.0$: In this case, $d_{ij}^{(N)} = 0.0$ and $d_{ij}^{(D)} = 1.0$. It indicates that two genes g_i and g_j of low dependency in normal state become highly dependent in the diseased state.

$p_{ij} = 1.0$: That is, $d_{ij}^{(N)} = 1.0$ and $d_{ij}^{(D)} = 0.0$ indicates that two highly dependent genes g_i and g_j in normal state become independent in diseased state.

$p_{ij} = 0.5$ (-0.5): In this case, dependency of g_i on g_j in diseased condition is partially lost (gained).

From these p_{ij} values, we have created a gene prediction network (GPN) containing n nodes (for n genes) where p_{ij} represents the weight of the link connecting between two nodes depicting ith and jth genes respectively. In case of $p_{ij} = 0.0$, there is no edge between ith and jth gene in the gene prediction network. When dependency between ith and jth gene highly lost (i.e., $p_{ij} = -1.0$) in diseased state then there is a dotted edge directed from ith gene to jth gene in GPN. Similarly, a continuous edge directed from ith gene to jth gene will be placed in GPN, when $p_{ij} = 1.0$ (i.e., dependency between ith and jth gene in diseased state is highly gained). We didn't consider any representation in GPN for $p_{ij} = 0.5$ (-0.5). In this way, we have built the gene prediction network (GPN) from gene dependency matrix.

3. RESULTS AND DISCUSSION

In this section, the effectiveness of the methodology is demonstrated on two cancer gene expression data sets. Here we have considered $\alpha = 0.3$ and $\beta = 0.6$ for all the data sets for both normal and cancer samples. This is followed by validation of the results.

3.1 Analysis of the results

Lung expression data [1] contains 10 normal samples and 86 tumor samples for expression values of 7129 genes. We have applied the methodology to this data set. It has been found that 187 (162) genes have lost (gained) their dependencies from normal samples to tumor samples; 275 (303) genes have lost (gained) their dependencies partially. Similarly for human breast expression data set (expression levels of 22645 genes for 2 normal breast epithelial cells and 4 samples for breast cancer cells) [10], 356 (371) genes have been found to loose (gain) their dependencies from normal to cancer cells; 510 (419) genes have lost (gained) their dependencies partially.

In order to restrict the size of the article, we have included only one figure (Figure 1) for the lung expression data set. Now we consider gene expression profiles of cancer samples in different stages. For lung carcinoma data, we have applied the methodology on various diseased states of human lung adenocarcinoma (86 tumor samples including 67 stage I tumor samples and 19 stage III tumor samples). In this case, 32 (41) genes have lost (gained) their dependencies from stage I tumor samples to stage III tumor samples. It has also been noted that 55 (48) genes have lost (gained) their dependencies from stage I to stage III partially.

3.2 Statistical validation

As already mentioned earlier, we have developed a methodology to establish the concept of gene dependency from gene expression data sets. This concept leads to develop a gene prediction network (GPN) that shows the altered dependencies among a set of genes from normal to carcinogenic samples. Unfortunately, we didn't get any information in literature to validate our results obtained from GPN. In this section, we try to validate the results in a different way. From GPN, we find out the set of *influential genes*. Here the term *influential genes* is defined as the genes that are involved in at least one altered dependency in GPN.

We apply t-test on the set of influential genes. Here we report three sets of genes with the levels of significance as 99.9%, 99% and 95%. For lung expression data, we have identified 263 influential genes (Table 1). Out of them 63% (167 out of 263) influential genes result in 99.9% level of significance, and 82% (218 out of 263) and 95% (251 out of 263) correspond to 99% and 95% levels of significance. Similarly, these figures for human breast expression data are 62% (99.9% level of significance), 70% (99% level of significance) and 80% (95% level of significance). All these results indicate that these influential genes have changed their expression level from normal state to diseased state quite significantly. Thus we can say that these genes have a significant role in mediating the disease.

Based on the aforesaid influential genes, we now try to validate the results in another way. For lung expression data (7129 genes), we have found that 241 (out of 263) *influential genes* are supported by some earlier investigations (http://www.ncbi.nlm.nih.gov/Database) in terms of disease mediation. Here we denote these 241 *influential genes* as a *true positive* (TP) results. However, we didn't get any information in literature for remaining 22 *influential genes*. We denote these genes as a *false positive* (FP). Similarly, we have got 51 other genes in literature supported by some earlier investigations, which are totally absent in our result. We identify these set of genes as *false negative* (FN) genes. Lastly, we have found 6815 ($= 7129 - (241 + 22 + 51)$) genes as *true negative* (TN), which are neither present in our result nor supported by literature in web. Based on these parameters, we compute the *specificity* ($= \frac{TN}{TN+FP}$),

Table 1: Statistical validation using t-test

Datasets (total number of influential genes)	Number of genes with the level of significance		
	99.9%	**99%**	**95%**
lung (263)	167	218	251
breast (403)	252	286	326

sensitivity ($= \frac{TP}{TP+FN}$), *precision* ($= \frac{TP}{TP+FP}$) and *accuracy* ($= \frac{TP+TN}{TP+FP+TN+FN}$).

From these data, we have got high *sensitivity* for all the data sets (Table 2) that measures the ability of the methodology to correctly identify the presence of the disease mediating genes. *Specificity* of 99% for all the data sets signifies the fact that the method correctly identifies the set of genes which are not responsible for the disease. A test with high *specificity* refers to a few *false positives* (FP). *Precision* measures the ability of the method to correctly identify the set of *true positive* (TP) genes from the *influential gene* set. High *precision* indicates that the method results is a very few *false positive* (FP) genes. Lastly, about 99% *accuracy* for all the data sets signifies that the method is capable of finding out the less number of *false positive* (FP) and *false negative* (FN) genes.

3.3 Validation of the results in terms of gene regulation

The present method finds some genes along with the dependencies among them, which have changed from normal to carcinogenic samples. We could not find any article in literature, which deals with similar investigation. That is why, we have tried to validate our results based on gene regulation.

For lung expression dataset, gene like HBB have shown stronger dependency on HBA1 in carcinogenic state. It is also noticed that genes like TP53, IGF1, IGFBP1 have strong influence in regulating gene IGFBP3 in cancer state whereas there is no dependency among them in normal state. In other words, TP53, IGF1, IGFBP1, IGF1R may regulate the expression value of IGFBP3 in tumor samples. It has been found that gene TP53 also regulates TNF, and gene TNF regulates genes TP53, HLA-B and PTEN in lung adenocarcinoma samples. In this way, we have identified a set of genes that have shown their strong dependencies in cancer state of lung expression data. Likewise, genes KRAS, EGFR, VEGFA regulate the gene TNF in normal state, whereas in carcinogenic state the expression levels of these genes may be almost independent of the expression level of TNF.

Regarding human breast expression data, we have identified the gene BRCA1 that has shown strong association in terms of dependency on genes like PTEN, TP53 in malignant tumor state. But in normal state, there is no dependency among them. Similarly, genes like STAT3 and CDKN2A have strong influence in regulating the gene BRCA1 in normal state. In other words, STAT3 and CDKN2A may regulate the expression of BRCA1 in normal state. On the other hand, NPM1 has shown strong dependency on gene BRCA1 in normal state. Gene KRAS has shown strong dependency on genes like TP53, PTEN in cancer state, whereas in normal state CDKN2A, BCL2, ERBB2, BRAF have strong influence on gene KRAS. It has been found that genes like

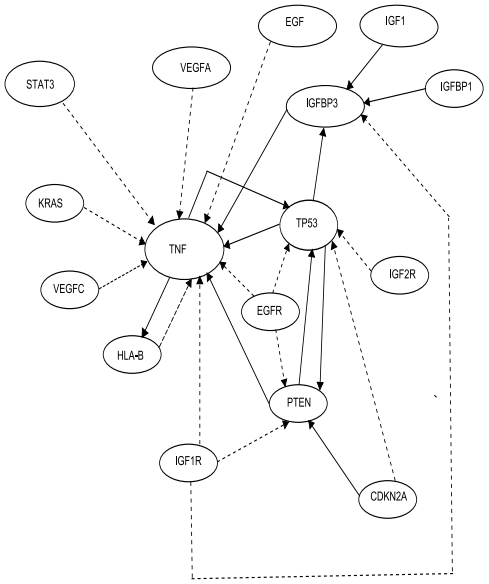

Figure 1: Gene Prediction Network (GPN) for lung expression data. Continuous (dashed) arrow indicates that dependencies between genes in carcinogenic samples have gained (lost). Here we consider $p_{ij} = 1$ or -1 only

HNF1A, HNF4A, CREBBP have strong influence in regulating the gene H3F3A in normal state, whereas in diseased state there is no such dependency. But there is no information in literature to our knowledge about these genes. This result suggests that the aforesaid genes may have impact on human breast cancer.

4. CONCLUSION

In this article, we have developed a methodology that shows how the dependencies among the genes have changed from normal state to diseased state. The algorithm has identified the dependencies among the genes, which have altered quite significantly from normal state to diseased state. The methodology involves measuring information content of individual genes using fuzzy entropy, dependencies of a pair of genes in both normal and diseased states using conditional fuzzy entropy. Finally, the dependencies that have altered from normal to carcinogenic state have been identified. The altered dependencies among the genes have been represented using network, called gene prediction network (GPN), in which each node represents a gene and a directed edge signifies altered dependency between a pair of nodes (genes).

In this way, we have identified the responsible genes as well as the altered dependencies among them. We have applied the algorithm on two cancer data sets (lung and breast).

Table 2: Statistical validation using some other parameters

Datasets (no of influential genes)	TP	FP	FN	TN	Specificity	Sensitivity	Precision	Accuracy
lung (263)	241	22	51	6815	0.99	0.82	0.91	0.98
breast (403)	352	51	59	22183	0.99	0.85	0.87	0.99

As a result, we have identified the gene prediction network for each of the data sets. We could not find any article in literature, which deals with similar investigation. So we have tried to validate our results based on gene regulation. In this context, we have used two statistical techniques to validate the results. From all the results, we have found that the method has been able to correctly identify many true positive and true negative genes. As a consequence, we can say that these set of identified genes along their deviated dependencies, have a significant role of mediating the disease. Hence, these results may facilitate the biologists and researchers carrying out the biochemical analysis to do further study on gene regulatory networks and how the entire network structure changes from normal state to diseased state.

5. REFERENCES

[1] G. D. Beer et. al. Gene-expression profilespredict survival of patients with lung adenocarcinoma. *Nature Medicine.*, 8:816–823, 2002.

[2] A. S. Butte and I. S. Kohane. Relevance networks: a first step toward finding genetic regulatory networks within microarray data. *The Analysis of Gene Expression Data, Springer.*, pages 428–446, 2003.

[3] A. Fujita, J. R. Sato, H. M. Garay-Malpartida, P. A. Morettin, M. C. Sogayar, and C. E. Ferreira. Time-varying modeling of gene expression regulatory networks using the wavelet dynamic vector autoregressive method. *Bioinformatics.*, 23:1623–1630, 2007.

[4] N. Grabe. Alibaba2: context specific identification of transcription factor binding sites. *In Silico Biol.*, 2:S1–S15, 2002.

[5] M. S. Halfon, Y. Grad, G. M. Church, and A. M. Michelson. Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model. *Genome Res.*, 12:1019–1028, 2002.

[6] O. Hirose, R. Yoshida, S. Imoto, R. Yamaguchi, T. Higuchi, S. D. Charnock-Jones, C. Print, and S. Miyano. Statistical inference of transcriptional module-based gene networks from time course gene expression profiles by using state space models, module finder on gene expression profiles. *Bioinformatics.*, 24:932–942, 2008.

[7] S. Kim, S. Imoto, and S. Miyano. Dynamic bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data. *Biosystems.*, 75:57–65, 2004.

[8] W. Krivan and W. W. Wasserman. A predictive model for regulatory sequences directing liver-specific transcription. *Genome Res.*, 11:1559–1566, 2001.

[9] D. H. Ly, D. J. Lockhart, R. A. Lerner, and P. G. Schultz. Mitotic misregulation and human aging. *Science*, 287:2486–2492, 2000.

[10] B. H. Mecham, G. T. Klus, J. Strovel, M. Augustus, D. Byrne, P. Bozso, D. Z. Wetmore, T. J. Mariani, I. S. Kohane, and Z. Szallasil. Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements. *Nucleic Acids Res.*, 32:e74, 2004.

[11] I. Nachman, A. Regev, and N. Friedman. Inferring quantitative models of regulatory networks from expression data. *Bioinformatics.*, 4:i248–i256, 2004.

[12] J. Quinlan. Programs for machine learning. *Morgan Kaufmann.*, 1993.

[13] J. L. Riechmann, J. Heard, G. Martin, L. Reuber, C. Jiang, J. Keddie, L. Adam, O. Pineda, O. J. Ratcliffe, and R. R. Samaha. Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. *Science.*, 290:2105–2110, 2000.

[14] E. Roulet, I. Fisch, T. Junier, P. Bucher, and N. Mermod. Evaluation of computer tools for the prediction of transcription factor binding sites on genomic dna. *In Silico Biol.*, 1:21–28, 1998.

[15] J. Schafer and K. Strimmer. An empirical bayes approach to inferring largescale gene association networks. *Bioinformatics.*, 21:754–764, 2005.

[16] O. Schuldiner, C. Yanover, and N. Benvenisty. Computer analysis of the entire budding yeast genome for putative targets of the GCN4 transcription factor. *Curr. Genet.*, 33:16–20, 1998.

[17] J. D. Storey and R. Tibshirani. Statistical significance for genomwide studies. *Proc. Natl. Acad. Sci.*, 100:9440–9445, 2003.

[18] K. Tan, G. Moreno-Hagelsieb, J. Collado-Vides, and G. D. Stormo. A comparative genomics approach to prediction of new members of regulons. *Genome Res.*, 11:566–584, 2001.

[19] R. Tupler, G. Perini, M. A. Pellegrino, and M. R. Green. Profound misregulation of muscle-specific gene expression in facioscapulohumeral muscular dystrophy. *Proc. Natl. Acad. Sci.*, 96:12650–12654, 1999.

[20] E. Wingender, X. Chen, E. Fricke, R. Geffers, R. Hehl, I. Liebich, M. Krull, V. Matys, H. Michael, and R. Ohnhauser. The transfac system on gene expression regulation. *Nucleic Acids Res.*, 29:281–283, 2001.

[21] R. Yamaguchi, R. Yoshida, S. Imoto, T. Higuchi, and S. Miyano. Finding modulebased gene networks in time-course gene expression data with state space models. *IEEE Signal Processing Magazine.*, 24:37–46, 2007.