DNA-Protein Binding and Gene Expression Patterns

Hongyu Zhao, Baolin Wu and Ning Sun

Abstract

Although many clustering methods have been applied to analyze gene expression data, genes in the same cluster may have neither common functions nor common regulation. As a result, computational approaches have been developed to identify motifs in the regulatory regions of a cluster of genes or of genes with similar gene expression levels that are responsible for DNA-protein binding and similar gene expression levels. However, these motifs are neither sufficient nor necessary for a transcription factor to bind to the promoter region of a gene with these motif patterns. More recently, molecular methods have been developed to directly measure DNA-protein binding at the genomic level. In this article, we first evaluate the predictive power of computational approaches to predict DNA-protein binding from a study involving nine transcription factors in the cell cycle. We then compare how much variation in gene expression levels can be explained either by the observed DNA-protein binding or by the binding predicted through computational approaches. We find that current computational approaches may be limited both in predicting DNA-protein binding as well as in predicting gene expression levels. We also observe indirectly that the correspondence between gene expression levels and protein levels may be rather poor, which suggests that there may be difficulty in modeling genetic networks purely through gene expression data. To better understand gene expression patterns, an integrated approach to incorporating different kinds of information should be developed.

Keywords: gene expression; DNA-protein binding; motif; microarray

1 Introduction

With the completion of the Human Genome Project, large-scale gene expression experiments have become common practice in the scientific community. Such experiments normally have different objectives: (1) to identify differentially expressed genes, (2) to identify genes expressed in a coordinated manner across a set of conditions, (3) to identify gene expression patterns that distinguish different samples (*e.g.* normal versus tumor tissues), and (4) to define global biological pathways. Genomics research is different from traditional molecular biology in that traditional approaches focus on the study of individual genes considered in isolation, whereas functional genomics allows researchers to determine the principles underlying complex biological processes (*e.g.*

development) by examining the expression patterns of large numbers of genes in parallel, taking into consideration temporal, as well as anatomical, patterns. Identification and characterization of regulatory *cis*-elements and *trans*-factors of a gene is essential for understanding the mechanisms of the control of gene expression, which can further shed light on gene function.

Currently, three types of statistical methods are under active development for gene expression data, including methods to identify differentially expressed genes (*e.g.* [8, 20] for cDNA arrays and [9, 22] for Affymetrix arrays), methods to identify clusters of genes with correlated expression patterns, *e.g.* [3, 10, 16, 21, 31], and methods to use gene expression patterns to distinguish phenotypes and predict clinical outcomes, *e.g.* [7, 13, 15, 32]. Although clustering methods have given some insight into gene function, similar gene expression patterns imply neither similar functions nor similar regulation for a group of genes. In addition, clustering results strongly depend on the set of experiments used to define similarities among genes, and results from different clustering algorithms may disagree with each other [12].

In contrast to standard statistical treatments of microarray data where data are mostly treated as a two-dimensional matrix, bioinformatics tools have been developed to use other information, mostly sequence information, to assist in the interpretation of gene expression patterns. For example, motif searches have been integrated in gene expression analysis in yeast studies, *e.g.* [4, 5, 24, 30]. The rationale is that genes having similar expression patterns are more likely to share common regulatory motifs in their promoter regions. These methods represent integration of expression data with sequence information. A more ambitious goal has been taken by some researchers to develop computational methods to reconstruct genetic networks, *e.g.* correlation metric construction [2], Boolean networks [1, 23, 28], and Bayesian networks [11, 14]. Unfortunately, most of these computational methods were not developed specifically for the analysis of gene expression data; therefore, it is difficult to incorporate biological information in these methods. They may generate results that are both hard to interpret and to verify, and they impose assumptions that are likely to be violated in real biological systems. This computational approach is in contrast to biologically driven approaches to dissecting pathways [18]. It has become clear that "the combination of predictive modeling with systematic experimental verification will be required to gain a deeper insight into living organisms, therapeutic targeting and bioengineering" [6].

Although many computational approaches have been proposed to identify DNA-protein binding motifs from gene expression patterns, such analyses may only provide indirect inference on binding. In addition, binding motifs are neither necessary nor sufficient for a given transcription factor to bind to the regulatory region of a gene [29]. Regulatory networks cannot be accurately deduced from expression profiles, partly because it is difficult to distinguish direct and indirect effects. Recently, experimental procedures have been developed to directly identify the *in vivo* genome binding sites for known transcription factors [19, 26]. Using this method, Simon *et al.* [29] studied genomic targets of nine known cell cycle transcription activators: Swi4, Swi6, Mbp1,

Fkh1, Fkh2, Mcm1, Ndd1, Ace2, and Swi5. MBF (Swi4 and Swi6) and SBF (Mbp1 and Swi6) control late G1 (cell cycle gap 1 phase) genes. Mcm1, together with Fkh1 or Fkh2, recruits the Ndd1 protein in late G2 (cell cycle gap 2) and controls G2/M (cell cycle gap 2 and mitosis phases) genes. Mcm1 is involved in M/G1 genes, whereas Swi5 and Ace2 control late M and early G1 genes [29]. Although Simon *et al.* [29] were able to infer binding motifs for each factor based on their data, they noted that the putative binding motifs are neither sufficient nor necessary to identify binding sites for a transcription factor.

In this article, using both gene expression data and binding data, we study how much DNA binding information explains gene expression levels through two approaches. In the first approach, we directly model expression levels as a function of the empirically measured binding of known transcription factors. In the second approach, we first infer putative motifs for each transcription factor based on the binding data, then predict binding based on these putative motifs, and finally model expression levels as a function of the predicted binding. Therefore, the second approach is an "indirect" computational method. We found that although the existing computational approaches yield significant associations between gene expression levels and predicted binding, the proportion of variation explained by these computational methods are much lower than those explained by empirically measured binding data. Our results suggest that better computational models and methods are needed to identify binding motifs and then to predict DNA-protein binding in the analysis and interpretation of gene expression data.

2 Methods

2.1 Gene expression data

We analyze cell cycle gene expression data reported in Spellman *et al.* [30], where yeast cell cultures were synchronized by three independent methods: α factor arrest, elutriation, and arrest of a cdc15 temperature-sensitive mutant. Approximately 800 genes, $>10\%$ of all yeast protein-coding genes, were identified as cell cycle regulated. In this article, we analyze the time course data from the α factor based synchronization experiment and gene expression levels of cell cycle regulated genes. The expression patterns of these genes were studied in detail by Spellman *et al.* [30] and a number of clusters of genes based on expression levels were investigated; this investigation included the identification of motifs for each gene cluster.

2.2 DNA binding data

The DNA binding data used in this article are those collected by Simon *et al.* [29]; the details of their experiments and statistical analysis of binding data can be found in Ren *et al.* [26]. Each experiment was done in triplicate. An estimate of the ratio of binding intensities of two fluorescents was calculated for each promoter region for a given transcription factor. This ratio, called the *binding ratio* here, is a measure of

the binding intensity of the given transcription factor. A statistical procedure was used by Simon *et al.* [29] to evaluate the statistical significance of the binding. In this article, we use their estimated p-values to assess statistical evidence for binding. These data revealed that genes encoding several of the cell cycle transcriptional regulators are themselves bound by other cell cycle regulators. Their data also suggested partial functional redundancy between homologous activators.

2.3 Motif searches

We use AlignACE [17, 27] to identify motifs that are over-represented in the upstream regulatory regions of a set of genes. In this article, we apply AlignACE to nine sets of genes, each of which were bound by the nine transcription factors, respectively. We then use CompareACE to identify those putative motifs that are similar to known motifs in yeast. Finally, ScanACE, a program that searches a genome for close matches to a motif found by AlignACE [17], is used to scan the upstream regions of the cell cycle regulated genes to identify those containing putative motifs. For each putative motif, each gene is defined as either having (coded "1") or not having (coded "0") this motif.

3 Results

3.1 Gene clusters based on binding data and gene clusters based on gene expression data

Transcription factors induce expression levels of cell cycle genes at different stages of the cell cycle. Simon *et al.* [29] observed consistency between DNA binding and gene expression levels. For example, SBF (Swi4 and Swi6) and MBF (Mbp1 and Swi6) are important activators of late G1 genes, and the expression levels for most of the genes bound by Swi4, Swi6, or Mbp1 are highest at the late G1 stage in the cell cycle [30]. When we cluster the nine transcription factors according to their binding ratios across the genome, transcription factors active at the same stage of the cell cycle are clustered together (Figure 1). However, when these factors are clustered according to their expression levels reported in Spellman *et al.* [30], there is no such ordering among them (Figure 2). This indicates that gene expression levels of the nine transcription levels are rather uninformative for correlating their functions in the cell cycle.

3.2 Binding motifs and binding ratios

To investigate how much computational methods can offer in predicting binding ratios, we apply AlignACE to genes bound by the same transcription factor to identify common motifs in the upstream promoter regions of these genes. We then select those putative motifs that are similar to known motifs, and run ScanACE on all cell cycle regulated genes to determine whether these motifs occur in the promoter regions of these genes. After this step, for each transcription factor, we fit a linear regression model

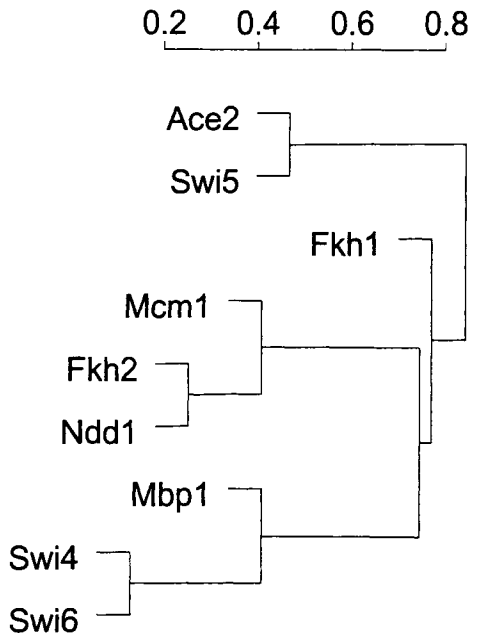

Figure 1: Clustering of nine transcription factors based on DNA binding data

with the observed binding ratios for cell cycle regulated genes as the response variable and the presence or absence of each putative motif in each gene as predictors, *i.e.*

$$y_i = \beta_1 M_{i1} + ... + \beta_k M_{ik} + e_i,$$

where y_i is the observed binding ratio for the ith gene, M_{ij} is a binary variable representing the presence ($M_{ij} = 1$) or absence ($M_{ij} = 0$) of the jth putative motif for this transcription factor in the i^{th} gene, and k is the number of putative motifs for this factor. In addition to this additive model, we also consider interactions among the M_{ij}, *i.e.*

$$y_i = \sum_{j=1}^{k} \beta_j M_{ij} + \sum_{j=1}^{k-1} \sum_{l=j+1}^{k} \gamma_{jl} M_{ij} M_{il} + e_i.$$

The results are summarized in Table 1, where all significant predictors for each transcription factor are listed, together with the proportion of variation in binding ratios explained by these predictors (R^2).

There are a few common features across all factors. First, SCB is the most commonly shared motif in these factors. Second, there are significant interaction terms for

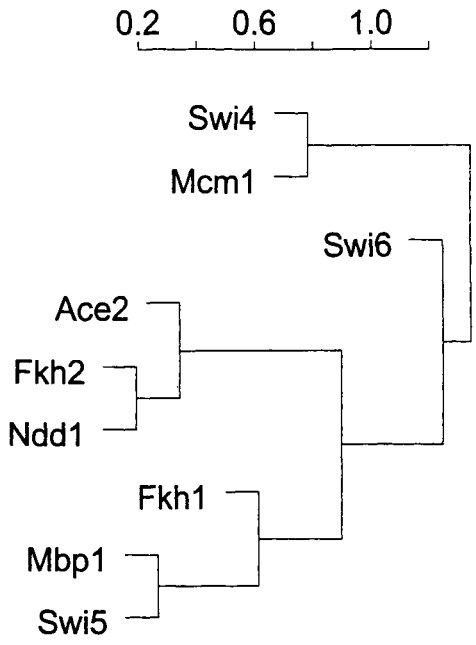

Figure 2: Clustering of nine transcription factors based on gene expression data

all the factors, which suggests that these putative motifs may interact with each other to recruit factors to the promoter regions. It is also clear from this table that the proportion of variation explained by the putative binding motifs varies across different transcription factors, with the variation in the binding ratios for Swi4, Swi6, and Ace explained most by the putative motifs. But overall, the R^2 is rather modest, which suggests that either there is substantial amount of measurement variation in binding ratios, or the motif search and binding prediction methods are far from satisfactory, or both.

3.3 Gene expression levels and empirically measured binding ratios

We consider how useful the binding ratios are to predict gene expression levels for cell cycle regulated genes. We analyze two sets of genes separately. The first set of genes includes all cell cycle regulated genes defined by Spellman *et al.* [30], whereas the second set of genes includes only those 298 genes that were found to be significantly bound (p-value < 0.001) by at least one of the nine transcription factors [29]. At each time point, for each set of genes, we first fit regression models with gene expression levels as the response variable and the observed binding ratios as predictors, *i.e.*

Table 1: Significant binding motifs as well as significant interactions among these motifs in the prediction of the binding ratios for each of the nine transcription factors studied. The last column is the proportion of variation explained by the joint effects of the binding motifs on the observed binding ratios

TF	Significant motifs and interactions	R^2
Swi4	SCB LEU MCB PDR SCB:MCB PDR:MCB	19%
Swi6	MCB STRE SCB:MCB MCB:STRE	14%
Mbp1	MCB RRPE MCB:RRPE	4%
Fkh1	RRPE STRE RRPE:STRE	1%
Fkh2	SCB RPN SCB:RPN	2%
Mcm1	SCB LYS SCB:LYS	4%
Ndd1	SCB REB SCB:REB	6%
Ace2	SCB LEU RAP STRE SCB:RAP SCB:STRE LEU:RAP	21%
Swi5	SCB STRE SCB:STRE	6%

$$y_i = \beta_1 R_{i1} + ... + \beta_9 R_{i9} + e_i,$$

where R_{ij} is the binding ratio between the ith gene and the jth factor, β_j is the regression parameter for the jth factor, and y_i is the observed expression level of the ith gene. The R^2 of the model for each of the 18 time points in the cell cycle are plotted in Figures 3 and 4 (solid lines).

It can be seen from these figures that the proportion of variation explained by the binding ratios is a function of time in the cell cycle, with the most variation explained at the S/G2 phase. The R^2 is increased if we focus on the subset of genes with each gene bound by at least one of the nine transcription factors. In the above analyses, we only consider the additive effects of different transcription factors. When interactions among factors are included in the model, we observe a significant increase in the R^2 for all time points. The comparisons between the additive models and the models with two-way interactions for the second set of genes are summarized in Figure 5, and the significant individual factors as well as significant interacting factors at each time point are summarized in Table 2. It can be seen that some interaction terms are significant, and that including interaction terms does improve the overall proportion of variation explained by the binding of these nine factors.

3.4 Gene expression levels and computationally predicted binding ratios

To evaluate the power of the predicted binding ratios in explaining gene expression levels, we fit regression models with the same response variable, *i.e.* gene expression

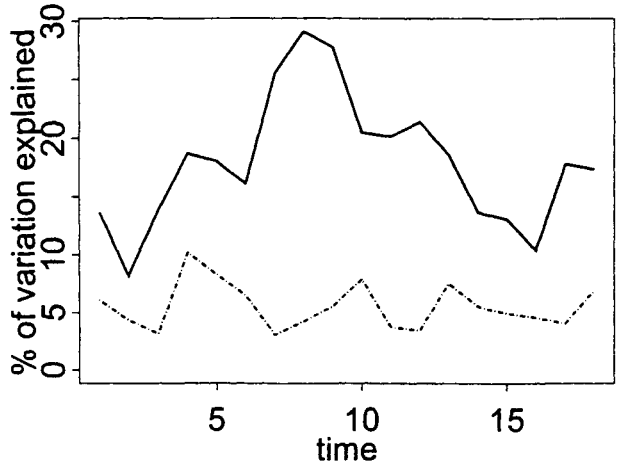

Figure 3: The proportions of variation explained by the observed binding data (solid line) and by the predicted binding (dotted line) for all cell cycle genes

levels, as above, but we use the predicted binding ratios through putative motifs as predictors this time. The R^2 of the model is plotted in Figures 3 and 4 (dashed line). It is clear from this figure that the observed binding data provides better information to explain expression levels. When interactions are included in the models, the overall R^2 is improved, but is still lower than that based on the empirically measured binding ratios. Therefore, although computational approaches are able to identify binding motifs that explain a statistically significant proportion of the variation in gene expression levels, their utility is limited compared to the directly observed binding data. Because we only consider nine transcription factors here, the unexplained proportion of the variation may be due to the effects from those transcription factors not included in the analysis, measurement errors in binding ratios, and sample variation in gene expression levels. Despite these other uncertainties, it is remarkable that these nine factors could explain up to 56% of the total variation at certain time points.

3.5 Estimation of transcription factor levels

These binding data also allow us to estimate relative protein expression levels for the transcription factors if we make the simple assumption that the effects of each transcription factor on inducing other genes' expression levels are proportional to the protein levels of the transcription factors in the cell. To estimate the protein levels of the nine transcription factors, we find the regression coefficients in the following regression model for each time point:

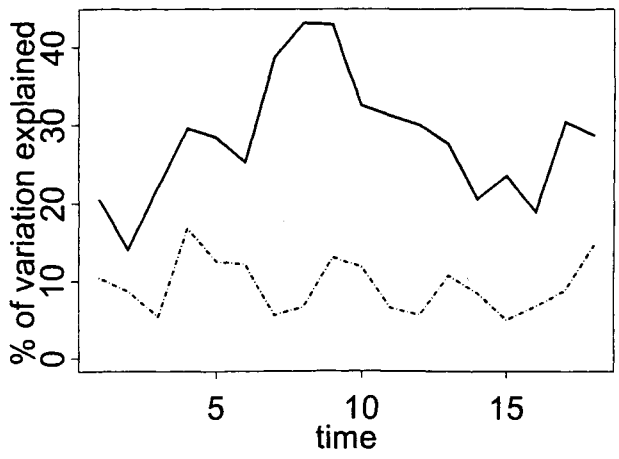

Figure 4: The proportions of variation explained by the observed binding data (solid line) and by the predicted binding (dotted line) for the 298 genes significantly bound by at least one of the nine transcription factors

$$y_i = B_{i1}L_1 + ... + B_{i9}L_9 + e_i,$$

where B_{ij} is the binding ratio between the ith gene and the jth transcription factor for the ith gene, and y_i is the gene expression level for the ith gene. Then the estimated L_j is the estimated protein expression level. Note that because the binding data only measure the relative levels, we should interpret the estimated L_j as equal to some constant times the protein level. Because we normalize the levels for the same protein across different time points in our summary (Figure 6), this is a reasonable approach to examine how the protein levels change across time. In Figure 6, we plot the observed gene expression levels and estimated protein expression levels at all 18 time points for each of the nine transcription factors. It can be seen from this figure that the correspondence between gene expression levels and protein levels is rather poor for some genes (*e.g.* Ace2), strong for some genes (*e.g.* Fkh1, Fkh2, and Ndd1), and a phase delay for other genes (*e.g.* Swi4 and Swi5).

4 Conclusions

We have first studied how well computational approaches can predict empirically observed DNA-protein interactions. Although we found that the computational approaches can yield results that are statistically significantly associated with the observed data, the

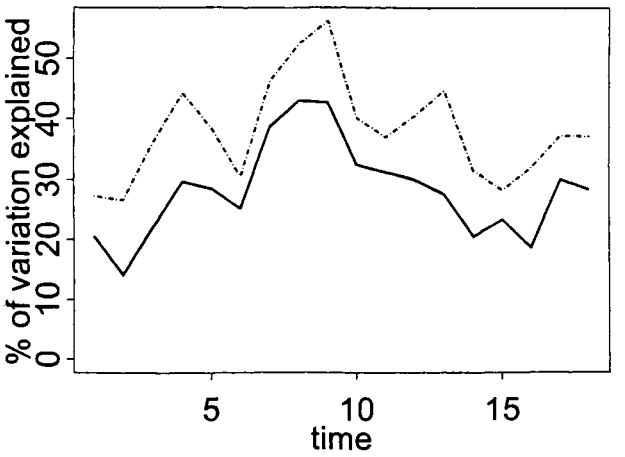

Figure 5: The proportions of variation explained by the observed binding data when only additive models are considered (solid line) and when interactions among factors are also considered (dotted line) for the 298 genes significantly bound by at least one of the nine transcription factors

correlation is rather modest. Current computational methods search for binding motifs separately; however, our results suggest the presence of interactions among putative binding motifs to jointly determine binding ratios. Similar observations were made by Pilpel *et al.* [25]. This suggests that interaction effects need to be taken into account in the search for binding motifs. Overall, even after interactions are taken into account, the proportion of variation in binding ratios explained by binding motifs through linear models is low. Therefore, there is ample room for methodology developments to predict DNA-protein binding.

We studied how well gene expression levels can be explained by DNA binding through two approaches. We found that a significant proportion of expression level variation across genes can be explained by the empirically measured DNA binding data. Similarly, computationally predicted binding also explain a significant proportion of the observed expression variation, but at much lower levels. We also investigated whether the predicted binding provide extra information to explain gene expression levels in addition to the observed binding by including both the observed binding and the predicted binding in the model. We found that the improvement of the model by the inclusion of the predicted binding was not significant (data not shown). Because it is well known that other transcription factors are involved in the cell cycle, we expect that the availability of binding data from other factors will further improve the prediction of the model. We also found that there is statistically significant evidence that

Table 2: Significant transcription factors and interacting terms in the prediction of gene expression levels at different time points

Time	Significant Terms
1	Ndd1, Ace2, Mbp1, Swi4, Mcm1:Swi4, Mcm1:Swi6, Mbp1:Swi6, Ndd1:Mcm1
2	Fkh1, Fkh2, Mbp1, Swi4, Mcm1:Swi4, Mcm1:Swi6, Mbp1:Swi6, Fkh1:Ndd1
3	Fkh2, Ndd1, Mbp1, Swi6, Ndd1:Swi5, Ace2:Swi5, Ndd1:Mbp1, Fkh1:Ndd1, Mcm1:Swi6
4	Fkh2, Ace2, Mbp1, Swi6, Fkh2:Ndd2, Ndd1:Swi6, Mcm1:Swi4
5	Ndd1, Mcm1, Ace2, Swi5, Mbp1, Swi6, Mbp1:Swi6, Ndd1:Mcm1
6	Fkh2, Mcm1, Ace2, Swi5, Swi6, Mbp1:Swi6
7	Fkh2, Ndd1, Mcm1, Ace2, Swi5, Swi4, Swi6, Fkh1:Fkh2, Mcm1:Swi6, Mbp1:Swi4, Swi5:Swi6
8	Fkh1, Fkh2, Ndd1, Mcm1, Ace2, Swi6, Fkh1:Ndd1, Mcm1:Swi6, Ace2:Swi5
9	Ndd1, Ace2, Swi5, Swi4, Swi6, Mcm1:Swi6, Fkh1:Ndd1, Ndd1:Mcm1
10	Ndd1, Mcm1, Ace2, Swi5, Swi4, Swi6, Ace2:Swi4
11	Fkh1, Mcm1, Swi5, Swi4, Fkh1:Swi4
12	Fkh2, Mcm1, Ace2, Swi5, Swi6, Mcm1:Swi6, Fk2:Ndd1, Fkh2:Ace2, Ndd1:Ace2
13	Ace2, Swi4, Swi6, Ace2:Swi4, Mcm1:Swi4, Ndd1:Swi5
14	Mbp1, Ace2, Swi4, Ace2:Swi4, Ndd1:Mcm1
15	Fkh2, Mcm1, Ace2, Swi5, Swi4, Ace2:Swi4
16	Fkh1, Fkh2, Ndd1, Swi6, Mcm1:Ace2, Fkh2:Ndd1, Ace2:Swi5, Ndd1:Swi5, Ndd1:Swi6
17	Fkh2, Ndd1, Ace2, Swi5, Mbp1, Swi6, Fkh2:Ndd1, Mcm1:Swi6, Mcm1:Ace2
18	Ndd1, Swi5, Swi6, Fkh2:Mcm1, Mcm1:Swi6, Fkh2:Swi6

different transcription factors interact with each other to contribute to the levels of gene expression. The interacting pairs not only include those known to work as a complex or present at the same stage of the cell cycle, they also include other pairs, suggesting that the interactions among these factors may be far more complex than currently thought.

In our analysis, we observed that the variation explained by the nine transcription factors is a function of time in the cell cycle. This indicates the importance of these nine transcription factors, as a group, varies at different stages of the cell cycle.

From the observed gene expression levels for different genes and the binding ratios between each gene and each factor, under a simple assumption, we were able to estimate the relative protein levels of the nine transcription factors studied. We found that although there is good correspondence between expression levels and "protein" levels for some factors, the correspondence is rather weak for others. There is no general relationship, and it appears that the relationship is both gene specific and time specific. The lack of consistency between gene expression data and protein expression data was noted by Ideker *et al.* [18]. However, factors with similar functions, *e.g.* Fkh1 and Fkh2, seem to have similar patterns between the observed gene expression data and the estimated protein expression levels. From the generally weak correlations between gene expression data and the estimated protein levels, we expect that computational models that only use gene expression data to reconstruct biological pathways may have limited power to make precise quantitative predictions. On the other hand, other types of data, such as the binding information, will be very useful in such efforts.

Another question that is of biological interest but has not been addressed in this paper is to examine how much of the gene expression similarities among a group of

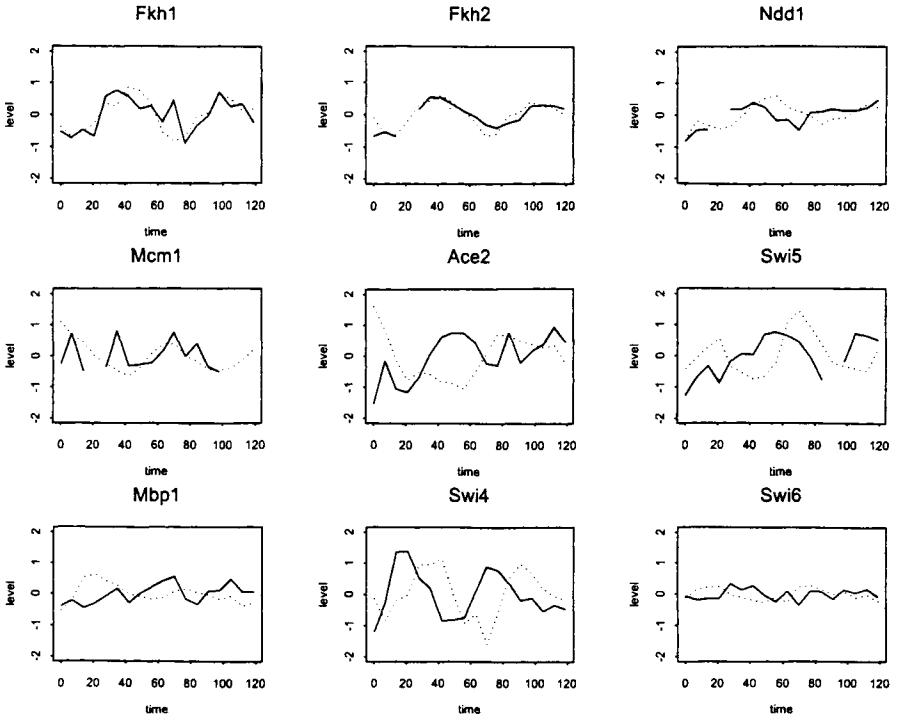

Figure 6: The observed expression levels of the nine transcription factors and their estimated protein expression levels. The data are normalized so that gene expression levels and protein expression levels have the same variance for a given transcription factor

genes can be explained by their regulation through a set of transcription factors. We can study this issue by comparing clusters derived purely from gene expression data and clusters derived purely from DNA binding data. Consistency between the two types of clusters would imply that the studied transcription factors may explain the regulation of these genes well, whereas a poor correlation implies that there are major mechanisms that drive the gene expression patterns but have not been uncovered or included in the study.

We have mainly used AlignACE and ScanACE to identify binding motifs for a group of genes. There are other computer programs available for motif findings and they may offer results better than we have found here. In addition, we have only considered those putative motifs that are similar to known motifs for the nine transcription factors. Although this procedure may exclude some unknown motifs that could play some role in determining DNA-protein binding, the likelihood of missing motifs with strong effects is small: these factors have been under intensive study by yeast geneticists, thus we expect that motifs with strong effects would have been identified. We are currently conducting a more thorough analysis to assess the importance of these

unmatched putative motifs.

Here we have considered the binding as a continuous measurement using the estimated binding ratios from replicate experiments. When we tried to dichotomize the binding data through the p-values reported by Simon *et al.* [29] (0 for the absence of binding and 1 for the presence of binding), the overall fit of the models is not as good as those we reported above (data not shown). This suggests that the continuous measurements do have more information on the regulation and interactions between genes and the transcription factors.

The ultimate goal of genomics studies is to understand biological pathways. In this article, we have shown the limitation of one existing computational method for studying gene regulation and the need to integrate gene expression data with other types data to dissect biological pathways. Incorporating DNA binding data is only the first step to move beyond purely statistical approaches for gene expression analysis.

Acknowledgements

We thank the reviewer for careful reading of this manuscript. Research supported in part by NIH grants GM59507, DK58776, and Grant IRG–58–012–45 from the American Cancer Society.

Hongyu Zhao, Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, hongyu.zhao@yale.edu

Baolin Wu, Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, baolin.wu@yale.edu

Ning Sun, Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, ning.sun@yale.edu

References

[1] T. Akutsu, S. Miyano, and S. Kuhara. Inferring qualitative relations in genetic networks and metabolic pathways. *Bioinformatics*, 16:727–734, 2000.

[2] A. Arkin, P. Shen, and J. Rose. A test case of correlation metric construction of a reaction pathway from measurements. *Science*, 277:1275–1279, 1997.

[3] A. Ben-Dor and Z. Yakhini. Clustering gene expression patterns. *S. Istrail, P. Pevzner, and M. S. Waterman (eds) Recomb 99, ACM Press, Washington, DC*, pages 188–197, 1999.

[4] H. J. Bussemaker, H. Li, and E. D. Siggia. Regulatory element detection using correlation with expression. *Nature Genetics*, 27:167–171, 2001.

[5] R. J. Cho, M. J. Campbell, L. Steinmetz E. A. Winzeler, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lockhart, and R. W. Davis. A genome wide transcriptional analysis of mitotic cell cycle. *Molecular Cell*, 2:65–73, 1998.

[6] P. D'haeseleer, S. Liang, and R. Somogyi. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, 16:707–726, 2000.

[7] S. Dudoit, J. Fridlyand, and T. P. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97:77–87, 2002a.

[8] S. Dudoit, Y. H. Yang, T. P. Speed, and M. J. Callow. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, 12:111–139, 2002b.

[9] B. Efron, R. Tibshirani, J. D. Storey, and V. Tusher. Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96:1151–1160, 2001.

[10] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of National Academy of Sciences USA*, 95:14863–14868, 1998.

[11] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using bayesian networks to analyze expression data. *Journal of Computational Biology*, 7:601–620, 2000.

[12] D. R. Goldstein, D. Ghosh, and E. M. Conlon. Statistical issues in the clustering of gene expression data. *Statistica Sinica*, 12:219–240, 2002.

[13] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

[14] A. J. Hartemink, D. K. Gifford, T. S. Jaakkola, and R. A. Young. Bayesian methods for elucidating genetic regulatory networks. *IEEE Intelligent Systems*, March/April:37–43, 2002.

[15] T. Hastie, R. Tibshirani, D. Botstein, and P. Brown. Supervised harvesting of expression trees. *Genome Biology*, 2:research0003.1–0003.12, 2001.

[16] T. Hastie, R. Tibshirani, M. B. Eisen, A. Alizadeh, R. Levy, L. Staudt, W. C. Chan, D. Botstein, and P. Brown. 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biology*, 1:research0003.1–0003.21, 2000.

[17] J. D. Hughes, P. W. Estep, S. Tavazoie, and G. M. Church. Computational identification of *cis*-regulatory elements associated with functionally coherent groups of genes in *Saccharomyces cerevisiae*. *Journal of Molecular Biology*, 296:1205–1214, 2000.

[18] T. Ideker, V. Thorsson, J. A. Ranish, R. Christmas, J. Buhler, J. K. Eng, R. Bumgarner, D. R. Goodlett, R. Aebersold, and L. Hood. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, 292(5518):929–934, 2001.

[19] V. R. Iyer, C. E. Horak, C. S. Scafe, D. Botstein, M. Snyder, and P. O. Brown. Genomic binding sites of the yeast cell-cycle transcription factors sbf and mbf. *Nature*, 409:533–538, 2001.

[20] M. K. Kerr, M. Martin, and G. A. Churchill. Analysis of variance for gene expression microarray data. *Journal of Computational Biology*, 7:819–837, 2000.

[21] L. Lazzeroni and A. Owen. Plaid models for gene expression data. *Statistica Sinica*, 12:61–86, 2002.

[22] C. Li and W. H. Wong. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of National Academy of Sciences USA*, 98:31–36, 2001.

[23] S. Liang, S. Fuhrman, and R. Somogyi. Reveal, a general reverse engineering algorithm of genetic network architectures. *Pacific Symposium on Biocomputing*, 3:18–29, 1998.

[24] X. Liu, D. L. Brutlag, and J. S. Liu. Bioprospector: Discovering conserved DNA motifs in upstream regulatory regions. *Pacific Symposium on Biocomputing*, 6:127–138, 2001.

[25] Y. Pilpel, P. Sudarsanam, and G. M. Church. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genetics*, 29:153–159, 2001.

[26] B. Ren, F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T. L. Volkert, C. J. Wilson, S. P. Bell, and R. A. Young. Genome-wide location and function of DNA binding proteins. *Science*, 290:2306–2309, 2000.

[27] F. P. Roth, J. D. Hughes, P. W. Estep, and G. M. Church. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotechnology*, 16:939–945, 1998.

[28] I. Schmulevich, E. R. Dougherty, S. Kim, and W. Zhang. Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, 18:261–274, 2002.

[29] I. Simon, J. Barnett, N. Hannett, C. T. Harbison, N. J. Rinaldi, T. L. Volkert, J. J. Wyrick, J. Zeitlinger, D. K. Gifford, T. S. Jaakkola, and R. A. Young. Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell*, 106:697–708, 2001.

[30] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycles regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell*, 9:3273–3297, 1998.

[31] P. Tamayo, P. D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. Golub. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proceedings of National Academy of Sciences USA*, 96:2907–2912, 1999.

[32] M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J. R. Marks, and J. R. Nevins. Predicting the clinical status of human breast cancer using gene expression profiles. *Proceedings of National Academy of Sciences USA*, 98:11462–11467, 2001.