Nucleic Acids Research

Accuracy and application of the motif expression decomposition method in dissecting transcriptional regulation

Zhihua Zhang and Jianzhi Zhang

Nucleic Acids Res. 36:3185-3193, 2008. First published 14 Apr 2008;
doi:10.1093/nar/gkn127

Published online 14 April 2008

Nucleic Acids Research, 2008, Vol. 36, No. 10 ***3185–3193***
doi:10.1093/nar/gkn127

Accuracy and application of the motif expression decomposition method in dissecting transcriptional regulation

Zhihua Zhang and Jianzhi Zhang*

Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor MI 48109, USA

Received January 9, 2008; Revised March 5, 2008; Accepted March 6, 2008

ABSTRACT

Understanding transcriptional regulation is a major goal of molecular biology. Motif expression decomposition (MED) was recently introduced to describe the expression level of a gene as the sum of the products of the binding strengths of its *cis*-regulatory motifs and the activities of the corresponding *trans*-acting transcription factors (TFs). Here, we use computer simulation to examine the accuracy of MED. We found that although MED accurately rebuilds gene expression levels from decomposed motif binding strengths and TF activities, estimates of motif binding strengths and TF activities are unreliable. Nonetheless, MED provides accurate estimates of relative binding strengths of the same motif in different genes and relative activities of the same TF under different conditions. We found that reasonably accurate results are achievable with genome-wide expression data from only 30 conditions and that MED results are robust to the existence of unknown occurrences of known motifs, although they are less robust to the presence of unknown motifs. With these understandings, judicious use of MED will likely provide useful information about eukaryotic transcriptional regulation. As an example, MED results are used to demonstrate that motifs generally have higher binding strengths when appearing in multiple copies than appearing in one copy per promoter.

INTRODUCTION

Understanding how gene expression is regulated is a major task of molecular biology. Jacob and Monod (1) pioneered the study of transcriptional regulation at the level of interaction between *cis*-regulatory motifs (or elements) in a gene's promoter region and *trans*-acting transcription factors (TFs) in the cell. Based on their idea, one may describe the log-transformed expression level of a gene at a given cellular condition by a function of the motifs present in the gene's promoter region and the TF activities present in the condition, as given in Equation (1) in Methods section [see also (2–4)]. The availability of several high-throughput technologies such as gene-expression microarrays and chromatin immunoprecipitation on microarray chips (ChIP-chip), and rapid progress in genomics and computational biology make it possible to study patterns of transcriptional regulation at the genomic scale (5–8). For example, large architectural differences in the yeast regulatory network among different cellular conditions have been identified (7,9). Recently, Nguyen and D'Haeseleer used Jacob and Monod's model to analyze microarray gene expression data obtained from multiple conditions in order to decipher principles of transcriptional regulation (10). Their method, called motif expression decomposition (MED), decomposes a matrix ($\boldsymbol{E}$) of gene expression levels at multiple conditions into the product of two matrices: the first ($\boldsymbol{M}$) contains the condition-independent binding strength of each motif (in each promoter) with its corresponding TF, while the second ($\boldsymbol{A}$) contains the activity of each TF at each condition studied. Some interesting patterns were observed from the analysis of the $\boldsymbol{M}$ matrix. For instance, the same motif with different orientations relative to the transcriptional direction may have different binding strengths, and the same motif with different physical distances from the transcriptional starting site may also have different strengths. Such findings, if correct, are invaluable for understanding the structure, function and evolution of promoters as well as those of transcriptional regulatory networks (11). Nguyen and D'Haeseleer examined the performance of MED by a cross-validation procedure, showing that the product of the decomposed $\boldsymbol{M}$ and $\boldsymbol{A}$ matrices is reasonably well correlated with the microarray gene expression levels. Although this result suggests that the method can be used to predict the expressions of some genes at a given condition when the expressions of many other genes are known at the same condition, it does

*To whom correspondence should be addressed. Tel: 734-763-0527; Fax: 734-763-0544; Email: jianzhi@umich.edu

not necessarily mean that the decomposed $\boldsymbol{M}$ and $\boldsymbol{A}$ matrices are accurate, as the same $\boldsymbol{E}$ may be decomposed into many different combinations of $\boldsymbol{M}$ and $\boldsymbol{A}$ (see subsequently). Because it is the $\boldsymbol{M}$ and $\boldsymbol{A}$ matrices that are of interest to most biologists, we decide to examine whether these matrices decomposed by the MED method are reliable. Because the true values of $\boldsymbol{M}$ and $\boldsymbol{A}$ matrices are unknown for any organism, here we employ a computer simulation approach. Our simulation results show that MED-derived $\boldsymbol{M}$ and $\boldsymbol{A}$ matrices are unreliable. Although this limitation of MED prohibits the direct use of $\boldsymbol{M}$ and $\boldsymbol{A}$ matrices, we find that MED accurately predicts the relative binding strengths of the same motif in different genes and relative activities of the same TF under different conditions. The performance of MED was also examined under limited expression data or partial knowledge of motifs. With improved understanding of MED, we applied MED in yeast to demonstrate at the genomic scale that motifs with >1 copy per motif have significantly higher binding strengths than the same motifs with 1 copy per motif.

METHODS

Generation of gene expression data

Based on Jacob and Monod's model of transcriptional regulation (1), the log-transformed expression level (E_{gc}) of gene g under condition c equals the sum of the products of the binding strength of each motif and the activity of its corresponding TF. That is,

$$E_{gc} \approx \sum_{j \in \Omega_g} M_{gj} A_{jc} \quad \mathbf{1}$$

Here, Ω_g is the set of motifs occurring in gene g's promoter region, M_{gj} is the binding strength of motif j in the promoter of gene g, A_{jc} is the activity of TF j, which binds to motif j, under condition c. A positive M indicates an enhancer motif, whereas a negative M indicates a repressor motif. Similarly, a positive A means activation, whereas a negative A means suppression. Following Nguyen and D'Haeseleer, we write Equation (1) in a matrix format for all genes, all motifs and all conditions, as

$$\boldsymbol{E} = \boldsymbol{M} \cdot \boldsymbol{A}, \quad \mathbf{2}$$

where $\boldsymbol{E}$ is a $m \times n$ matrix that gives m genes' expression levels at n conditions, $\boldsymbol{M}$ is a $m \times k$ matrix that gives the condition-independent binding strengths of k motifs in m genes' promoter regions and $\boldsymbol{A}$ is a $k \times n$ matrix that gives the activities of k TFs under n conditions.

We randomly generate a $m \times k$ matrix designated as $\boldsymbol{M}_{\mathrm{O}}$; each element in column i of $\boldsymbol{M}_{\mathrm{O}}$ is a random variable drawn from the normal distribution $N(b_i, \sigma)$, where $i = 1, 2, 3, \ldots, k$, and b_i and σ are the mean and standard deviation of the normal distribution, respectively. Each b_i is a random variable drawn from the normal distribution $N(B, \sigma)$. We set H_{g}, the number of motifs in gene g, by drawing a Poisson random variable with mean equal to 3. We then randomly pick H_{g} of the k motifs in gene g and leave their corresponding entries in row g of $\boldsymbol{M}_{\mathrm{O}}$ unchanged but set zero to all other entries in row g of $\boldsymbol{M}_{\mathrm{O}}$. We further make sure that each row and each column has at least one non-zero entry. If there is a row or column that contains all zeros, we randomly choose an entry and reverse the value to that in the original $\boldsymbol{M}_{\mathrm{O}}$. The matrix generated after these steps is referred to as $\boldsymbol{M}$. We randomly generate a $k \times n$ matrix designated as $\boldsymbol{A}$. The elements in the ith row of $\boldsymbol{A}$ are random variables drawn from the normal distribution $N(C_i, \phi)$, where $i = 1, 2, 3, \ldots, k$, and C_i is a random variable drawn from the normal distribution $N(C, \phi)$. We then generate gene expression data $\boldsymbol{E}$ using Equation (2). Because gene expression has stochastic variations (12) and because measurement of gene expression has errors, the observed gene expression level will differ from the above computed $\boldsymbol{E}$. Hence, we add an error term to each expression value. For entry E_{ij}, the error is a random variable drawn from $N(0, \varepsilon E_{ij})$, where ε is the noise level fixed in each simulation. We have used $\varepsilon = 0$, 5, 10, 20, 30, 40, 50, and 100% in different simulations. After this step, the $\boldsymbol{E}$ matrix is referred to as the observed or true expressions. MED requires an initial $\boldsymbol{M}$ matrix designated as $\boldsymbol{M}_{\mathrm{I}}$ to start the decomposition. We generate $\boldsymbol{M}_{\mathrm{I}}$ by replacing all non-zero entries in $\boldsymbol{M}$ to 1. Unless otherwise stated, this $\boldsymbol{M}_{\mathrm{I}}$ is used in our simulations. As will be described later, in some occasions, we also used an $\boldsymbol{M}_{\mathrm{I}}$ where each non-zero entry is -1 and an $\boldsymbol{M}_{\mathrm{I}}$ where each non-zero entry is either 1 or -1, with equal probabilities.

Simulation

Because Nguyen and D'Haeseleer's study focused on the yeast *Saccharomyces cerevisiae*, we use parameters appropriate for yeast in our simulation. Using the approach outlined in the above section, we randomly generate expression data for 4500 genes under 300 conditions. The total number of TFs in the organism is set to be 100. In the dataset analyzed by Nguyen and D'Haeseleer, there were expression data from 5719 genes under 255 conditions and the total number of TFs was 62. Using the MED method (10), we decompose the expression data (matrix $\boldsymbol{E}$) into $\boldsymbol{M}'$ and $\boldsymbol{A}'$ matrices and then compute $\boldsymbol{E}'$ using $\boldsymbol{E}' = \boldsymbol{M}' \cdot \boldsymbol{A}'$. We then compare $\boldsymbol{E}'$ with $\boldsymbol{E}$, $\boldsymbol{M}'$ with $\boldsymbol{M}$ and $\boldsymbol{A}'$ with $\boldsymbol{A}$, as they represent the MED-derived matrices and the true matrices, respectively. At each noise level, we repeat the simulation 10 times. This number of replications is sufficient because our results are highly reproducible.

RESULTS

Performance in predicting expression levels

Using computer simulation as described in Methods section, we generated motif binding strength ($\boldsymbol{M}$) and TF activity ($\boldsymbol{A}$) matrices for 4500 genes under 300 conditions, including information for 100 different TFs and their corresponding motifs. We first used $B = 2.5$ and $\sigma = 10$ in generating the $\boldsymbol{M}$ matrix and used $C = 0$ and $\phi = 10$ in generating $\boldsymbol{A}$. Our B and σ values are similar to the $\boldsymbol{M}$ matrix decomposed from the yeast expression data (10). Our C and ϕ are different from the decomposed values

Table 1. Pearson's correlation coefficients (± standard deviation) between the true values and MED-predicted values of expression levels (***E***), motif binding strengths (***M***) and TF activities (***A***)

Noise level (%)	***E***	***M***	***M*** ratio (within-column)[a]	***M*** ratio (between-column)[b]	***A***	***A*** ratio (within-row)[c]	***A*** ratio (between-row)[d]
0	1.000 ± 0.000	0.120 ± 0.997	0.998	0.289	0.120 ± 0.997	0.996	−0.044
5	0.997 ± 0.001	0.179 ± 0.988	0.986	−0.028	0.179 ± 0.988	0.992	0.200
10	0.991 ± 0.005	0.119 ± 0.997	0.988	0.101	0.119 ± 0.997	0.964	0.020
20	0.962 ± 0.026	0.119 ± 0.996	0.942	0.004	0.119 ± 0.995	0.930	0.048
30	0.929 ± 0.036	0.199 ± 0.981	0.904	0.081	0.199 ± 0.981	0.862	−0.045
40	0.872 ± 0.063	0.059 ± 0.997	0.848	−0.028	0.059 ± 0.995	0.771	0.103
50	0.834 ± 0.067	0.178 ± 0.979	0.812	−0.031	0.179 ± 0.977	0.714	0.110
100	0.606 ± 0.099	0.300 ± 0.890	0.587	0.064	0.303 ± 0.893	0.435	0.170

Note: The simulated expression data are from 300 conditions.
[a]Relative binding strengths of the same motif in two genes.
[b]Relative binding strengths of two different motifs.
[c]Relative activities of the same TF under two different conditions.
[d]Relative activities of two different TFs.

in (10), because MED has a normalization step that artificially equalizes the average activity of each TF such that the actual TF activities cannot be seen from the decomposed ***A*** in (10). Nonetheless, even when we use $C = 0$ and $\phi = 0.1$, similar to those observed from the decomposed ***A*** in (10), our results remain unchanged.

We then generated the gene expression matrix ***E*** by multiplying ***M*** and ***A*** matrices followed by addition of different levels of expression noise. The ***E*** matrix was decomposed into ***M′*** and ***A′*** matrices using the MED method. We conducted a total of 10 simulation replications. Because the results are essentially identical among the replicates, subsequently we describe our findings from the first replication.

There are three expectations if the MED method performs well. First, predicted gene expressions (***E′***, or the product of ***M′*** and ***A′***) should be close to the observed expressions (***E***). Second, predicted motif binding strengths (***M′***) should be close to their true values (***M***). Third, predicted TF activities (***A′***) should be close to their true values (***A***). To measure the agreement between predicted and true values of expression levels, we computed Pearson's correlation coefficient (r) between ***E*** and ***E′*** for each gene (row), and then computed the average r value across the 4500 genes and the standard deviation of r. Similarly, to measure the agreement between predicted and true values of motif binding strengths and TF activities, we computed r between ***M*** and ***M′*** for each motif (column) and r between ***A*** and ***A′*** for each TF (row), and then take averages across all motifs and all TFs, respectively.

As shown in Table 1, r between ***E*** and ***E′*** gradually declines as the noise level rises. Nonetheless, $r > 0.80$ even when the noise is as high as 50% of the true value and is greater than 0.90 when the noise level is <30%. These results suggest that expression levels predicted by MED are reliable. Indeed, for individual genes under individual conditions, Figure 1 shows that the predicted expression levels match the true values for the majority of genes under the majority of conditions. Figure 1 is based on the simulation results with a noise level of 30%. Qualitatively similar patterns were obtained when different levels of noise (5–100%) were introduced.

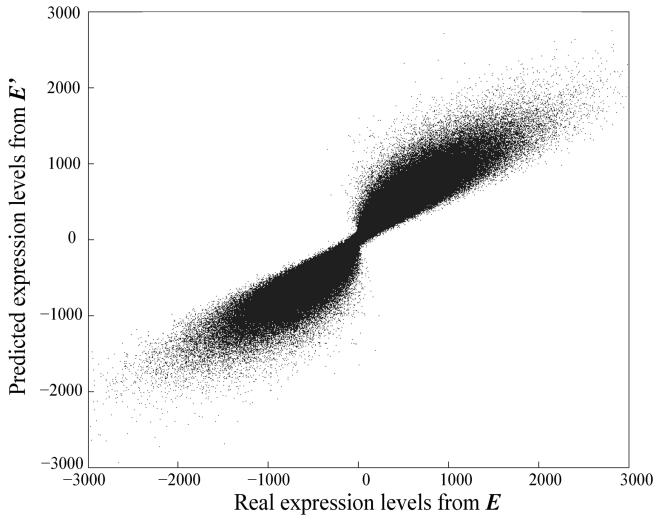

Figure 1. Comparison between the true (***E***) and MED-predicted (***E′***) gene expression levels. The noise level is 30%. Note that the expression levels are log-transformed and thus can be negative.

Performance in predicting motif binding strengths and TF activities

To our disappointment, however, the r values between ***M*** and ***M′*** matrices are low (<0.3) regardless of the level of noise (Table 1). Figure 2A shows that the motif binding strength values in ***M*** and ***M′*** are dramatically different. Similarly, the r values between ***A*** and ***A′*** matrices are low (Table 1) and the TF activity values in ***A*** and ***A′*** are quite different (Supplementary Figure S1A). These observations suggest that although ***E′*** is close to ***E***, ***M′*** is not close to ***M*** and ***A′*** is not close to ***A***. It is easy to show that if ***M*** and ***A*** form one solution, multiplying column i of ***M′*** by a and row i of ***A′*** by $1/a$ ($a \neq 0$) generates another solution. Because a can be 1, −1 or any non-zero number, there are infinite numbers of decomposition solutions. The original proof of the uniqueness of the MED decomposition solution was based on the arbitrary assumption that each TF has a mean activity of 1 across all conditions (i.e., the mean of each row in the ***A′*** matrix is fixed at 1) (10).

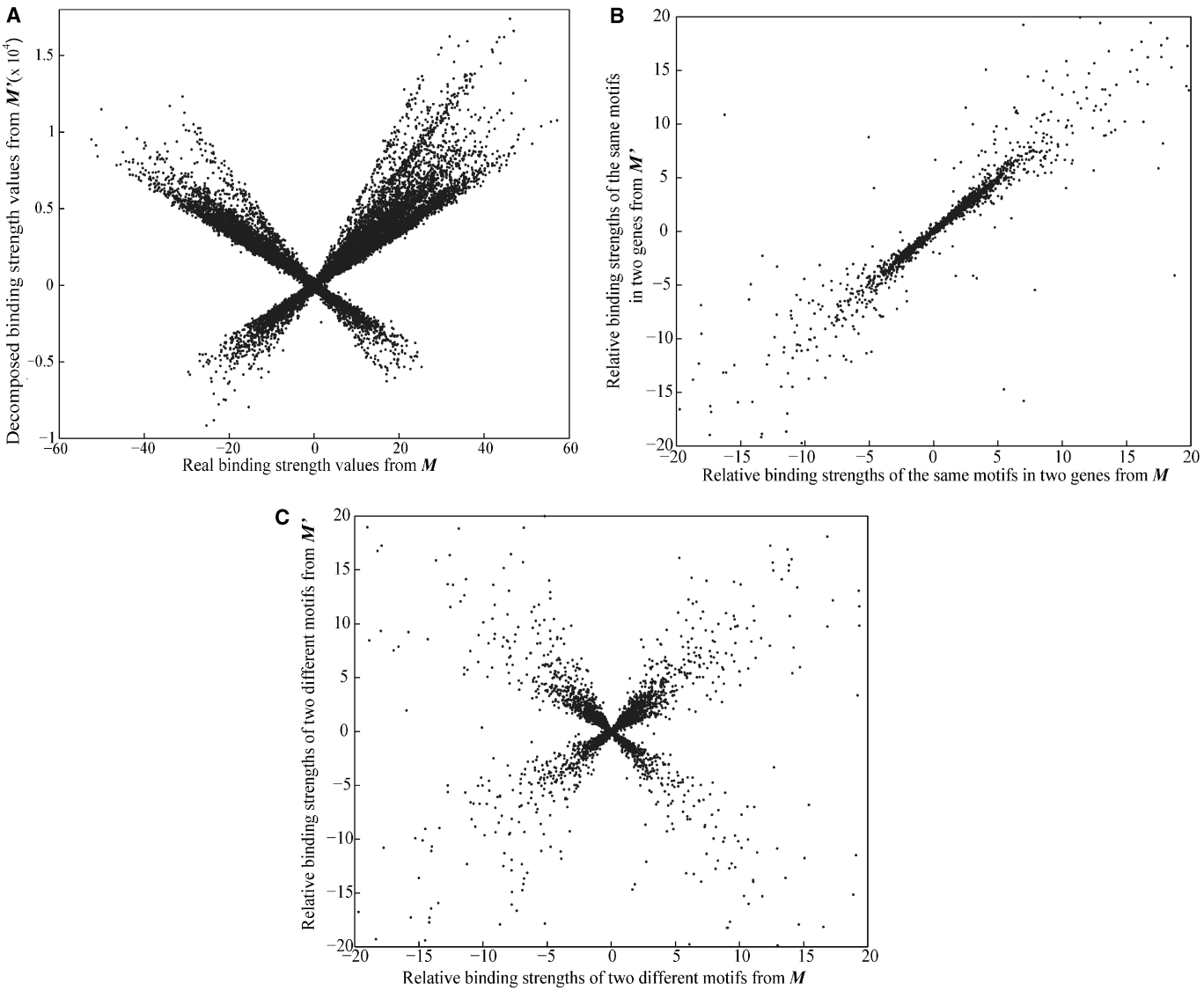

Figure 2. Comparison between true (***M***) and MED-predicted (***M′***) motif binding strengths. The noise level is 30%. (**A**) The scatter plot for true and predicted motif binding strengths. Note the difference in scale between *X*-axis and *Y*-axis. (**B**) True and predicted relative binding strengths of the same motifs in different genes. (**C**) True and predicted relative binding strengths of pairs of different motifs.

Although there is only one decomposition solution under this arbitrary assumption, the solution is not guaranteed to be the right one. In fact, our simulations showed that it is generally not the right solution. Nonetheless, our above consideration predicts that the ratio of any two entries within the same column (motif) of ***M′*** can still be close to the corresponding ratio in ***M***, while the ratio of any two entries from different columns of ***M′*** should not correlate with the corresponding ratio in ***M***. Similar predictions can be made for rows (TFs) of ***A*** and ***A′***. These predictions were indeed confirmed in our simulations. That is, between ***M*** and ***M′***, within-column ratios are highly correlated, whereas between-column ratios are not (Table 1; Figure 2B and C). In parallel, between ***A*** and ***A′***, within-row ratios are highly correlated, whereas between-row ratios are not (Table 1; Supplementary Figure S1B and C). Note that in this article, we measured Pearson's correlation between true and predicted ratios by using only ratios falling in the range of [−20, 20], which account for >95% of all ratios. This treatment is preferred over the use of all ratios because of the existence of a small number of ratios with extreme values, which affects the measure of Pearson's correlation coefficient. Similar results were obtained when all ratios were considered in Spearman's rank correlation.

As stated earlier, if ***M′*** and ***A′*** form one solution, multiplying column i of ***M′*** by a and row i of ***A′*** by $1/a$ ($a \neq 0$) generates another solution. Because a can be either positive or negative, it is expected that the r between a column in ***M*** and its corresponding column in ***M′*** should be close to 1 or −1 when the noise level is low. This is indeed the case. For example, in the simulation with 30% noise, between ***M*** and ***M′***, 60% of columns have $r > 0.98$, while 40% of columns have r lower than −0.98 (same for

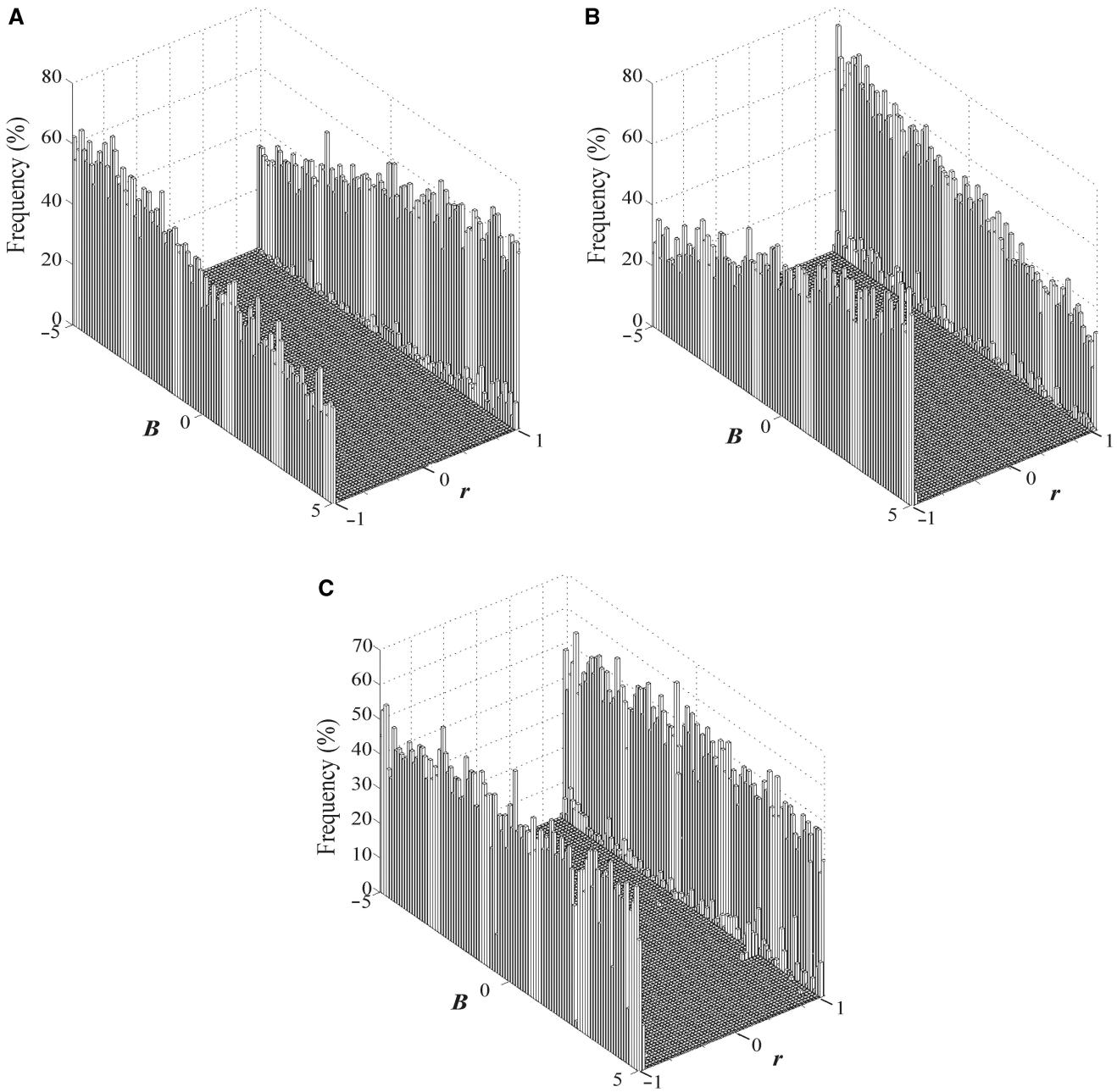

Figure 3. The distribution of Pearson's correlation coefficient between columns (motifs) of $\boldsymbol{M}$ and $\boldsymbol{M}'$, when all non-zero entries in $\boldsymbol{M}_{\mathrm{I}}$ are (**A**) 1, (**B**) −1, and (**C**) randomly assigned to be either 1 or −1, with equal probabilities. B is the mean motif binding strength in $\boldsymbol{M}$.

rows between $\boldsymbol{A}$ and $\boldsymbol{A}'$). This is why we observed low average r values and high standard deviations for both motif biding strengths and TF activities (Table 1).

Because MED only supplies one of infinite numbers of solutions of $\boldsymbol{M}'$ and $\boldsymbol{A}'$, which particular solution does it provide? This question is equivalent to asking what a values MED uses. We found that the initial matrix ($\boldsymbol{M}_{\mathrm{I}}$) used to start the decomposition process affects a. We conducted three sets of simulations, each containing 50 individual simulations. In the first set of 50 simulations, we started with an $\boldsymbol{M}_{\mathrm{I}}$ where every non-zero entry was set to be 1, as used by the original authors of MED (10). The $\boldsymbol{M}$ matrix was generated with parameter B changing from −5 to 5 in a step size of 0.2 in the 50 simulations. The $\boldsymbol{A}$ matrix was generated as usual. In the second set of 50 simulations, we started with an $\boldsymbol{M}_{\mathrm{I}}$ where every non-zero entry was set to be −1. In the third set of 50 simulations, we started with a $\boldsymbol{M}_{\mathrm{I}}$ where every non-zero entry was randomly set to be either 1 or −1, with equal probabilities. Figure 3A–C shows the distributions of Pearson's correlation coefficients between columns of $\boldsymbol{M}$ and $\boldsymbol{M}'$ for all the simulations in the three sets, respectively. They clearly show that the entries in $\boldsymbol{M}'$ tend to have the same sign as in $\boldsymbol{M}_{\mathrm{I}}$. For example, when B is

Table 2. Pearson's correlation coefficients between true values and MED-predicted values of expression levels (***E***), relative motif binding strengths (***M***) and relative TF activities (***A***), when the expression data are obtained from 300, 100 and 30 conditions, respectively

Noise level (%)	***E***			***M*** ratio (within-column)[a]			***A*** ratio (within-row)[b]		
	300 conditions	100 conditions	30 conditions	300 conditions	100 conditions	30 conditions	300 conditions	100 conditions	30 conditions
0	1.000 ± 0.000	0.997 ± 0.004	1.000 ± 0.000	0.998	0.993	0.976	0.996	0.998	0.996
5	0.997 ± 0.001	0.997 ± 0.001	0.998 ± 0.001	0.986	0.989	0.946	0.992	0.987	0.993
10	0.991 ± 0.005	0.990 ± 0.006	0.989 ± 0.009	0.988	0.933	0.867	0.964	0.956	0.976
20	0.962 ± 0.026	0.964 ± 0.025	0.967 ± 0.027	0.942	0.845	0.699	0.930	0.906	0.916
30	0.929 ± 0.036	0.930 ± 0.037	0.934 ± 0.049	0.904	0.840	0.586	0.862	0.873	0.818
40	0.872 ± 0.063	0.880 ± 0.061	0.887 ± 0.076	0.848	0.760	0.579	0.771	0.798	0.744
50	0.834 ± 0.067	0.833 ± 0.078	0.841 ± 0.098	0.812	0.611	0.404	0.714	0.680	0.623
100	0.606 ± 0.099	0.595 ± 0.125	0.652 ± 0.164	0.587	0.361	0.224	0.435	0.359	0.314

[a]Relative binding strengths of the same motif in two genes.
[b]Relative activities of the same TF under two different conditions.

positive and most entries in ***M*** are positive, use of the $\boldsymbol{M}_{\mathrm{I}}$ with positive entries tends to give more positive r values (Figure 3A) than use of the $\boldsymbol{M}_{\mathrm{I}}$ with negative entries (Figure 3B). Similar patterns are observed in ***A*** (Supplementary Figure S2).

Combining all the simulation results, we now have a better understanding of MED. The MED algorithm is designed in such a way that only one of infinite numbers of solutions is provided and this solution depends on the initial values used in decomposition. Knowing this property, it becomes clear that the MED-decomposed binding strengths for a given motif (across genes) are not true strengths, but are expected to be true strengths multiplied by an unknown number. Furthermore, this unknown number can be different for different motifs. The relative binding strengths of the same motif in different genes can be reliably estimated by MED. However, MED cannot distinguish between enhancers and repressors, neither can it distinguish between activation and suppression TF activities. Moreover, MED-predicted binding strengths cannot be compared among different motifs, and MED-predicted TF activities cannot be compared among different TFs.

Robustness of MED

MED relies on the input of gene expression data and *cis*-motif information. It is important to examine the influences of these factors on the performance of MED. In the above simulations, we simulated expression data from 4500 genes at 300 conditions. A practical question is how large the expression data have to be for MED to produce reliable values of ***E′***, ***M′*** and ***A′***. We do not reduce the gene number because most eukaryotes have >4500 genes. Rather, we reduce the number of conditions from 300 to 100 and 30, respectively, with the rationale that the cost for generating expression data can be significantly reduced if 100 or even 30 conditions are sufficient for predicting motif bind strengths and TF activities. Table 2 gives the results for 30 and 100 conditions, in comparison with 300 conditions. One can see that the reliability of the MED method in rebuilding ***E′*** is not reduced when fewer conditions are used. But, for predicting relative binding strengths and TF activities, use of fewer conditions worsens the MED performance. However, if the noise level is <10%, use of 30 conditions can still provide reasonably good predictions (Table 2).

Detection of TF-binding sites is a much studied topic in the past decade (13–16). However, not all *cis*-regulatory motifs can be detected by current methods (13). We examined the accuracy of MED in two situations when some motifs in the genome are undetected. In the first situation, for a given TF, a fraction of its corresponding *cis*-motifs in the genome are assumed to be undetected. In the simulation, we fixed a random set of non-zero entries in $\boldsymbol{M}_{\mathrm{I}}$ at 0. We repeated the simulation 10 times, as in each replication a different set of non-zero entries from the same $\boldsymbol{M}_{\mathrm{I}}$ were fixed at 0. We examined r between ***M*** and ***M′*** for relative binding strengths of the same motif in two genes. Note that presumably undetected motifs were not considered in computing r. We assumed that 0, 5, 10, 20, 30, 40 and 50% of motifs are undetected in seven sets of simulations, respectively. The results show that undetected motifs slightly worsen the performance of MED in predicting relative motif binding strengths (Figure 4A). The same is true for the relative TF activities (Supplementary Figure S3A).

In the second situation, we assumed that for most TFs, all of their corresponding motifs are known, while for the rest of the TFs, none of their motifs are known. In the simulation, we fixed all the entries of a random set of columns in $\boldsymbol{M}_{\mathrm{I}}$ at 0. We repeated the simulation 10 times, as in each replication a different set of columns from the same $\boldsymbol{M}_{\mathrm{I}}$ were fixed at 0. We examined r between ***M*** and ***M′*** for relative binding strengths of the same motif in two genes. Again, presumably undetected motifs were not considered in computing r. We also assumed that 0, 5, 10, 20, 30, 40 and 50% of motifs are undetected in seven sets of simulations, respectively. The results show that this type of ignorance of motifs has a great impact on the prediction of relative motif binding strengths (Figure 4B). The same is true for the relative TF activities (Supplementary Figure S3B). Nonetheless, the predictions are not too bad (mean $r > 0.65$) when motifs corresponding to up to 10% of TFs are completely unknown and the noise level is not >30%.

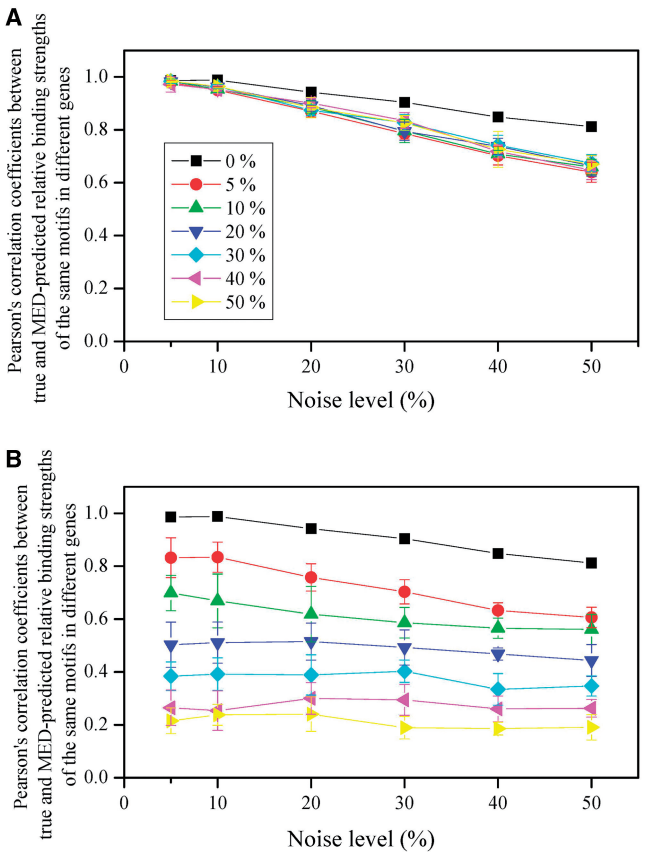

Figure 4. Performance of the MED method in predicting relative motif binding strength when some motifs in the genome are undetected. The mean correlation coefficient from 10 simulations and the associated standard deviation are presented for each condition examined. In (**A**), a fraction of motifs (from 0% to 50%) for each TF are undetected in the genome. In (**B**), all motifs of a fraction of TFs (from 0% to 50%) are undetected in the genome. Different colors show different fractions.

An application of MED

After knowing what MED can do and cannot do, we decided to use MED to address an important question in gene regulation. It is frequently observed in eukaryotic promoters that a motif appears with multiple tandem copies (6). Although it has been frequently assumed that a motif with multiple copies in a promoter has stronger binding strength than the same motif with only one copy (2,17), whether this assumption is valid at the genomic scale has not been empirically tested. This question is ideal for MED to tackle, because it only requires the mean binding strength of a given motif in one set of genes, relative to that in another set of genes. Using the same yeast dataset used by Nguyen and D'Haeseleer (10), we separated the genes into two groups for each motif. The first group includes genes that each has only one copy of this motif, whereas the second group includes genes that each has multiple copies of the motif. Of the 62 motifs that can be separated into two groups, we found 18 motifs for which the average binding strengths for the two groups have opposite signs (i.e. one is positive and other is negative). These inconsistent results are likely due to MED errors and thus are removed. For each of the remaining

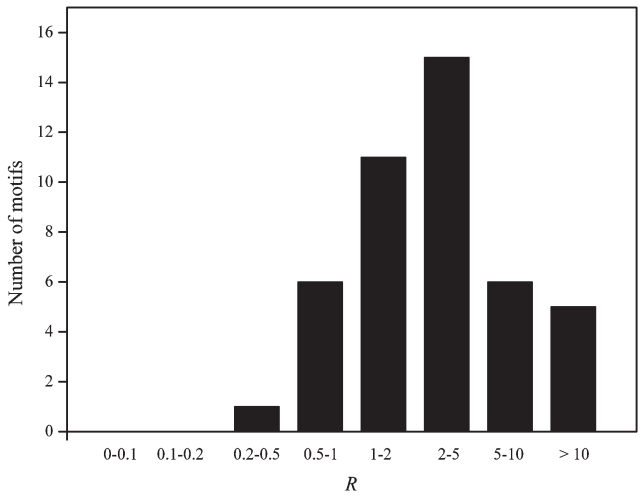

Figure 5. Frequency distribution of the ratio (***R***) between the mean binding strength of a motif in promoters where it has multiple copies to the mean binding strength of the same motif in promoters where it has one copy. The distribution is from 44 different motifs in yeast.

44 motifs, we calculated the ratio (R) between the average binding strength of the second group and that of the first group. We then tested the null hypothesis that $R = 1$, against the alternative hypothesis that $R > 1$. We found that the average R of the 44 motifs is 4.517 ± 0.897, significantly greater than 0 ($P < 10^{-5}$; t-test; Figure 5). Furthermore, 37 motifs, significantly more than half of the 44 motifs, have $R > 1$ ($P = 3 \times 10^{-6}$; binomial test; Figure 5). These results indicate that motifs with multiple copies in promoters generally have greater binding strengths than the same motifs with single copies (Figure 5).

DISCUSSION

The exponential growth of available functional genomic data opens the possibility to understand biological processes at the genomic and systems levels (6,18,19). One major advance in this endeavor is the development of methods for identifying *cis*-regulatory motifs in promoters of all genes in a genome. Using genome-wide microarray gene expression data and motif information, Nguyen and D'Haeseleer invented the MED method, which decomposes the gene expression data into motif binding strength data and TF activity data (10). The knowledge of binding strengths and TF activities can be used to decipher principles of transcriptional regulation. Thus, it is important to know how well MED performs. In this work, we conducted computer simulations to evaluate the MED method. Our results showed that at realistic levels of noise, which includes both expression stochasticity and microarray errors, MED-predicted gene expression levels are highly reliable. This result is not unexpected, as MED decomposes $\boldsymbol{E}$ into $\boldsymbol{M'}$ and $\boldsymbol{A'}$, which are then used to rebuild $\boldsymbol{E'}$.

For both binding strengths and TF activities, however, MED cannot provide accurate predictions. Furthermore, MED cannot differentiate between enhancer and

repressor motifs and cannot differentiate between activation and suppression TF activities. MED results cannot be used to compare binding strengths among different motifs and compare activities among different TFs. Nevertheless, the relative binding strengths of the same motif in different genes and the relative activities of the same TF under different conditions can be estimated with fairly high accuracy. If we have external information that a motif is an enhancer or repressor or that a TF activity under a given condition is activation or suppression (relative to the control condition), such information can be combined with MED results to provide better predictions. We note that relative binding strengths of the same motif in different genes and relative activities of the same TF under different conditions can provide much information that is valuable to our understanding of principles of transcriptional regulation. One such example is the comparison between binding strengths of the same motif when it has one copy per promoter versus multiple copies per promoter. Using MED results, we demonstrated that for the majority of motifs (84%), the binding strength is greater when a motif appears in multiple copies than when it appears in one copy. This may explain why many motifs have multiple copies in a promoter. However, we caution that this result was based on an analysis of motifs corresponding to only 62 TFs, about a third of all TFs in yeast. Because our simulation showed that MED is not robust to the ignorance of all motifs of even 10% of TFs in the genome, the validity of our result should be further examined when larger data become available.

An encouraging finding from our simulations is that at realistic levels of noise, MED requires expression data from as few as 30 conditions to provide reasonably accurate predictions of relative motif binding strengths and relative TF activities. Thus, even a small lab may be able to generate sufficient data for a genome-wide estimation of motif binding strengths in a non-model organism. Another encouraging finding is that even when some motifs (e.g. 20%) in the genome are undetected, MED can still make reasonable good predictions, as long as the majority of motifs are detected for each TF. When all motifs of some TFs are unknown, MED will have much reduced accuracy. Thus, from the perspective of MED performance, it is more important to identify most motifs for each TF than to identify all motifs for some TFs.

It should be noted, however, that the simulation results presented here were based on a number of simplified assumptions that warrant discussion. First, we assumed a simple logic of transcriptional regulation as described by Equation (1) in Methods section. If this assumption is violated, MED predictions will be less accurate. One potentially important violation is interaction between motifs or interaction between TFs, which have been observed (20,21). Second, epigenetic factors are known to affect gene expression differently for different genes under different conditions (22). Third, we assumed a relatively simple form of expression stochasticity and microarray noise. If expression errors are much larger and/or more complex, MED predictions may be less accurate. We believe that a better understanding of the molecular mechanisms of gene expression regulation will assist the development of more powerful computational tools, which in turn help further understand gene expression regulation.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Meg Bakewell and Ben-Yang Liao for valuable comments. This work was supported by research grants from National Institutes of Health and University of Michigan Center for Computational Medicine and Biology to J.Z. Funding to pay the Open Access publication charges for this article was provided by National Institutes of Health.

Conflict of interest statement. None declared.

REFERENCES

1. Jacob,F. and Monod,J. (1961) Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.*, **3**, 318–356.
2. Bussemaker,H.J., Li,H. and Siggia,E.D. (2001) Regulatory element detection using correlation with expression. *Nat. Genet.*, **27**, 167–171.
3. Liao,J.C., Boscolo,R., Yang,Y.L., Tran,L.M., Sabatti,C. and Roychowdhury,V.P. (2003) Network component analysis: reconstruction of regulatory signals in biological systems. *Proc. Natl Acad. Sci. USA*, **100**, 15522–15527.
4. Tran,L.M., Brynildsen,M.P., Kao,K.C., Suen,J.K. and Liao,J.C. (2005) gNCA: a framework for determining transcription factor activity based on transcriptome: identifiability and numerical implementation. *Metab. Eng.*, **7**, 128–141.
5. Lee,T.I., Rinaldi,N.J., Robert,F., Odom,D.T., Bar-Joseph,Z., Gerber,G.K., Hannett,N.M., Harbison,C.T., Thompson,C.M., Simon,I. *et al.* (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, **298**, 799–804.
6. Harbison,C.T., Gordon,D.B., Lee,T.I., Rinaldi,N.J., Macisaac,K.D., Danford,T.W., Hannett,N.M., Tagne,J.B., Reynolds,D.B., Yoo,J. *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.
7. Luscombe,N.M., Babu,M.M., Yu,H., Snyder,M., Teichmann,S.A. and Gerstein,M. (2004) Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, **431**, 308–312.
8. MacIsaac,K., Wang,T., Gordon,D.B., Gifford,D., Stormo,G. and Fraenkel,E. (2006) An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics*, **7**, 113.
9. Zhang,Z., Liu,C., Skogerbø,G., Zhu,X., Lu,H., Chen,L., Shi,B., Zhang,Y., Wang,J., Wu,T. *et al.* (2006) Dynamic changes in subgraph preference profiles of crucial transcription factors. *PLoS Comput. Biol.*, **2**, e47.
10. Nguyen,D.H. and D'Haeseleer,P. (2006) Deciphering principles of transcription regulation in eukaryotic genomes. *Mol. Syst. Biol.*, **2**, 2006.0012.
11. Bussemaker,H.J. (2006) Modeling gene expression control using Omes Law. *Mol. Syst. Biol.*, **2**, 2006.0013.
12. Elowitz,M.B., Levine,A.J., Siggia,E.D. and Swain,P.S. (2002) Stochastic gene expression in a single cell. *Science*, **297**, 1183–1186.
13. Tompa,M., Li,N., Bailey,T.L., Church,G.M., De Moor,B., Eskin,E., Favorov,A.V., Frith,M.C., Fu,Y., Kent,W.J. *et al.* (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotech.*, **23**, 137–144.
14. Elnitski,L., Jin,V.X., Farnham,P.J. and Jones,S.J. (2006) Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques. *Genome Res.*, **16**, 1455–1464.

15. Kim,S.Y. and Kim,Y. (2006) Genome-wide prediction of transcriptional regulatory elements of human promoters using gene expression and promoter analysis data. *BMC Bioinformatics*, **7**, 330.
16. Maston,G.A., Evans,S.K. and Green,M.R. (2006) Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.*, **7**, 29–59.
17. van Helden,J., Andre,B. and Collado-Vides,J. (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.*, **281**, 827–842.
18. Ideker,T., Galitski,T. and Hood,L. (2001) A new approach to decoding life: systems biology. *Annu. Rev. Genomics Hum. Genet.*, **2**, 343–372.
19. Brazma,A., Krestyaninova,M. and Sarkans,U. (2006) Standards for systems biology. *Nat. Rev. Genet.*, **7**, 593–605.
20. Bulyk,M.L., Johnson,P.L.F. and Church,G.M. (2002) Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res.*, **30**, 1255–1261.
21. Bulyk,M.L., McGuire,A.M., Masuda,N. and Church,G.M. (2004) A motif co-occurrence approach for genome-wide prediction of transcription-factor-binding sites in Escherichia coli. *Genome Res.*, **14**, 201–208.
22. Allis,C.D., Jenuwein,T. and Reinberg,D. (2007) *Epigenetics*, *1st edn.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.