BIOINFORMATICS

Vol. 22 no. 14 2006, pages e368–e374
doi:10.1093/bioinformatics/btl253

An equilibrium partitioning model connecting gene expression and *cis*-motif content

Joe Mellor[1,*] and Charles DeLisi[1]

[1]Program in Bioinformatics, Boston University, Boston, Massachusetts, USA 02215

ABSTRACT

Thermodynamic favorability of transcription factor (TF) binding to DNA is a significant factor in the control of gene expression. Theoretical and *in vitro* measures link the relative equilibrium energy of a particular DNA binding protein to the sequence variation among binding sites in a genome. Extending this principle, we investigate whether biological variation in expression levels of active proteins leads to regulation of different sets of genes, based on inferred affinities of sites upstream of those genes. The TF-concentration-dependent variation in the repertoire of genes regulated by a particular TF is expected to follow patterns of chemical partitioning over DNA sites having differing affinity, and we develop a new modeling approach to test this hypothesis. Based on computational TF binding site discovery and genome-wide expression data available in *Saccharomyces cerevisiae,* we explore motif content for sets of genes and conditions having varying concentrations of different transcription factors which turn those genes on or off. We find cases of significant correlation between the level of intragenomic motif sequence variation and modeled TF protein levels that actuates regulation of corresponding sets of genes, and discuss the observed TF motif variants for several yeast transcription factors, as well as the potential biological functions of genes that are regulated by differential response to these high and low concentrations of particular TFs. These findings suggest that motif sequences of transcription factor binding sites may often be linked with the expression state of corresponding DNA-binding proteins.

Contact: mellor@bu.edu

1 INTRODUCTION

Response to the internal and external cellular millieu is often facilitated by energetically favored binding between at least one regulatory protein and several specific, high affinity consensus motifs in the intergenic DNA. In addition to protein-DNA contacts, transcription factor (TF) binding affinity can be affected by cooperativity, chromatin structure and changes in binding site accessibility. The magnitude of binding affinity at protein-DNA interfaces has been shown to correlate with features including level of sequence variation (1), the presence of multiple motif copies in the intergenic neighborhood (2), and the level of binding found by large-scale chromatin immuno-precipitation (ChIP) experiments (3,4). Transcription factor binding often initiates mechanisms of regulation in RNA production, and while some genes having more primitive regulation mechanisms might escape this paradigm, an understanding of large systems will come in an unraveling of various parts of this regulatory machine.

*To whom correspondence should be addressed.

The biological question we address here is summarized as follows: To what extent are differences in the regulation of various genes mediated by functional differences in a TF's affinity to different upstream intergenic sequences, or the level of the TF needed to bind sets of these sequences? Various forms of evidence suggest that the affinity between proteins and DNA often governs the specificity of regulation (3-11), and we hypothesize that affinity for specific sequences ought to be related to the TF concentration which is needed to turn genes next to those sequences on or off.

Without practical means to address the mechanistic questions directly—that is, to measure *in vivo* affinities for such proteins to all possible motifs in the genome—we instead query for site affinity by proxy. Using a historically observed relationship between motif sequence variation and expected binding affinity, we seek to show how atomistic properties of simple regulatory schemes (*e.g.* TF binding) can be effectively estimated from aggregate measurements of gene expression among system components. We then examine whether the content of *cis*-regulatory elements explains significant differences in responsiveness of downstream genes to various levels of TF. To address these questions we introduce a modeling approach and its application to the *cis*-regulatory sites of the yeast *Saccharomyces cerevisiae*, based on gene expression and upstream sequence information.

Following previous convention (12), a transcription factor's affinity for sequences on DNA can be represented with a position weight matrix (PWM), which indicates preferences for protein binding to any of $b = 4$ possible nucleotides at k independent positions in a set of DNA k-mers. The information content I of the PWM (Eq. 1.1) is a measure of the overall degeneracy (or entropy) of the sequences to which a protein binds.

$$\sum_k \sum_b f(k,b) \frac{\log\left(f(k,b)\right)}{\log\left(f(b)\right)} = I(k,b)_{PWM} \qquad (1.1)$$

Berg and von Hippel showed that, given assumptions of independent contributions of each base at each position in a motif, the PWM equates via statistical thermodynamics to the expected relative free energy ($\Delta\Delta G$) of the binding event at the motif (13), and therefore also to the relative equlibrium binding affinity compared to all possible binding sites in the genome, as shown in Eq. 1.2.

$$I(k,b)_{PWM} \propto \ln(K_{bind}^{eff}) = -\frac{\Delta\Delta G}{RT} \qquad (1.2)$$

The correlation between motif degeneracy and protein-DNA binding free energy leads us to consider whether all motif sites that seem 'allowable' for binding by a transcription factor (that is, with a known PWM) are in fact actually bound by it under

in vivo equilibrium conditions in the cell. Cellular conditions can potentially differ, for example, by having active transcription factor present at different equilibrium concentrations, by having sites that are more or less accessible on the chromosome, or perhaps by post-translational changes to transcription factor activity. The mechanism of regulation is important, as well; for example, some instances of regulation by a protein may require cooperativity with another protein, but other instances may not. To a certain approximation the effective binding constant, K_{bind}^{eff}, of a set of regulatory sites that are *cis* to some collection of genes will, for some fixed concentration of transcription factor protein, determine the fraction of those sites which are bound and likely to involve gene regulation. Similarly, given a constant value of K_{bind}^{eff} across many genes, the ratio of bound to unbound genes might determine the concentration of transcription factor needed to activate those genes. Thus the *in vivo* regulatory program could by this representation be a fairly complex function of the affinity and availability of different *cis* sites, and fluctuations of concentration and interactions between regulating TFs.

We begin with the simplest situation, that of a single protein binding alone to DNA, where the binding site affinity has an inverse relationship with transcription factor concentration at across a range of conditions (TF concentrations) in which regulated genes are bound. This is shown in Eq. 1.3 (assuming uniform concentration of DNA).

$$K_{bind}^{eff} = \frac{[TF]_{bound}}{[TF]_{free}} \tag{1.3}$$

If the amount of bound TF is assumed to be constant across some set of sites near regulated genes, the TF concentration and affinity for those sites should be inversely related. That is, sites with high affinity will be selected on the basis of their thermodynamic favorability at low levels of active TF, but this partitioning between sites will relax at higher concentrations of TF, where sites of lower affinity will bind as well. With a fixed level of bound TF ($[TF]_{bind}$) needed to turn genes on or off, it is possible to consider subpopulations of regulated genes (and their *cis*-sites) which could vary by relative affinity, and therefore potentially respond at relatively higher or lower concentrations of active TF.

Binding (attaining a level of $[TF]_{bound}$) in the simple cases is mediated by the relationship between $[TF]_{free}$ and K_{bind}^{eff} seen in Eq. 1.3. In this study we develop a statistical model to infer whether simple cases such as this exist in the data. We devise a model which addresses whether genes divided into categories modulated by high or low concentrations of a given TF have *cis* sites that can be similarly classified as having 'strong' (high K_{bind}^{eff}) or 'weak' (low K_{bind}^{eff}) affinity. For any fixed set of genes, we assume that the information entropy $I(K, b)$ of *cis* sites near those genes can be used as a proxy for the measurement of K_{bind}^{eff} across those sites, and by Eq. 1.1, we then ask whether sites near those genes turned on or off by high levels of a TF contain different amounts of information than sites near the genes turned on by low levels.

The simple model we propose notwithstanding, it is useful to consider possible alternatives that would deviate from the relation predicted by Eq. 1.3. One such case might be when TF binding to an otherwise low-affinity site is preferred because of an added affinity cause by cooperative binding to another, different regulator. Combinatorial and cooperative mechanisms of transcriptional regulation are abundant in eukaryotes, and in many of these situations, a transcription factor's affinity for a site is potentially mediated by separate, often cooperative, binding events between two or more regulating proteins. Little is known, generally, about the interplay between the affinity of proteins to other proteins or DNA in such cases. Two potentially opposite modes of affinity, one cooperative and the other not, are summarized in Figure 1.

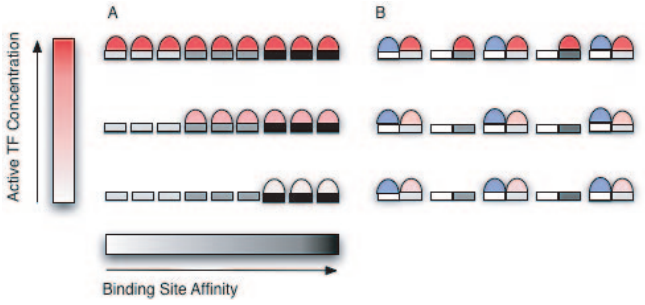

Fig. 1. Binding Site Occupancy Models as a Function of Site Affinity and TF Concentration. (A) In the simplest case, concentration of the active TF (red) controls various gene sets depending on an average affinity of the sites (shades of gray) near of those genes. Under equilibrium conditions, and with many sites, the information content of sites should correspond to their average affinity. (B) Cases of cooperativity between factors possibly deviate from this behavior, when binding is mediated by an additional protein (blue) which binds nearby, changing the effective affinity of the sites. The relationship between motif content, TF concentration and site occupancy could be altered to favor low affinity sites at low TF concentration.

DNA sites with affinity for cooperative proteins are not necessarily bound solely based on their intrinsic affinity, or the availability of the active TF, but also due to favorable TF binding to other proteins, and of these proteins to other, neighboring sites. In cases where the thermodynamic selection for a ''weak'' site by a protein is preferred because that site neighbors a site of another cooperating protein to which the first protein binds, ''Strong'' sites, on the other hand, which have higher affinity, can bind TFs in the absence of the cooperating protein. This type of 'neighboring site effect' has been recently shown to play a role in governing the content of *cis* elements for several TFs in yeast(5).

While combinatorial effects may be common among eukaryotic mechanisms of regulation, TFs and *cis*-sites near genes obeying the simpler relationship of Eq 2 (*i.e.*, Figure 1a) are still quite interesting. First, they represent cases where the dynamics of regulatory behavior are approximately reducible to TF concentrations and estimated binding strengths alone. Second, these are cases where regulation mechanisms are possibly more accessible to efforts in engineering of synthetic systems; their relatively simple behavior makes them ideal for manipulation in novel systems.

2 METHODS

A yeast TF is designated as either an activator or repressor based on primary evidence collected in the *Saccharomyces* Genome Database (14). As is the case in many previously described expression-based models (15-17), we'll assume transcription of genes by this TF can be used as a predictor of the TFs protein-level activity. Though the strength of this predictor may vary by TF, we'll assume it is a uniform predictor regardless of what genes a particular TF regulates. We'll also posit that the protein-level activity of a TF can be ascertained in many cases from its own gene expression data. Simplifications

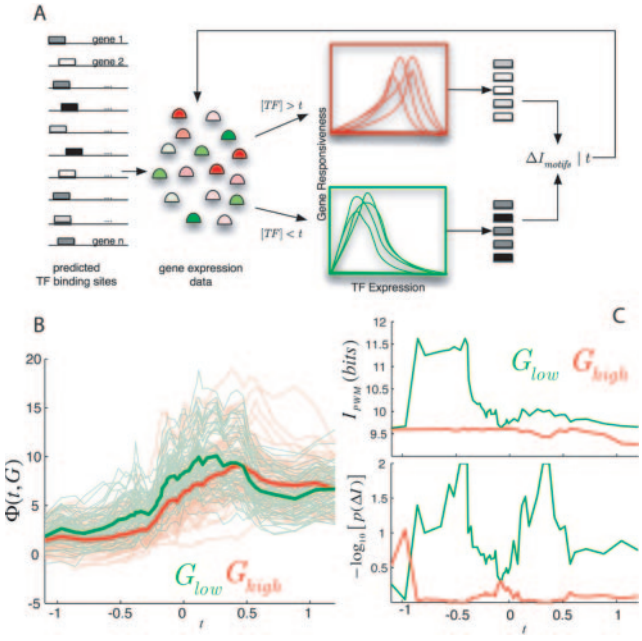

Fig. 2. Regulation Modeling Procedure and Analysis of Expression-Based Motif Sets. **(A)** Sets of genes with motifs for a given transcription factor are analyzed to produce an estimate of the difference in motif information content as a function of the TF expression threshold parameter *t*. The method identifies cases where the divergence in motif content corresponds to a divergence in regulatory response to the TF. **(B)** An example of the regulation scoring function profile for genes with at least one motif for the amino acid biosynthesis regulator *GCN4*, for *t=0.25*. Genes in ''high'' group (red) are maximally responsive to levels of *GCN4* greater than *t*; genes in the ''low'' group (green) to levels less than *t*. **(C)** The upper plot shows information content (in bits) of motifs in high and low gene groups as a function of *t*. Lower half shows the probability of observing a difference in information as great as that between low and high groups at values of *t*.

such as this will fail to account for some quite relevant biological situations, for example where a TF is constitutively expressed or active, where the level of TF or target activity is controlled exclusively by post-translational modifications, or where regulation in general is not mediated by mRNA production. Using mRNA measurements alone potentially underestimates, but not likely contradicts, the importance of these more complex mechanisms in our model. And while the model is much simpler than a fuller physical model of the processes involved, it usefully isolates variables which can be easily measured by experiments such as gene expression arrays and protein-DNA ChIP assays.

2.1 Computation of gene regulatory models

The outline of our procedure is summarized in Figure 2. We begin with a collection of hypotheses, or putatively independent regulatory models, that follow from any pair of genes *A* and *G*, where a high-scoring sequence element for a motif representing transcription factor *a* (encoded by *A*) is found by motif scanning in the 1000 bp region upstream of the start codon of *G*. The set $\mathbf{G}_A = \{G_1, G_2, \ldots, G_n\}$ then represents the collection of such genes for a TF *A*; we then ask whether expression exhibited by the transcription factor expression is significantly associated with a net activation or repression of the target gene. Transcription factor binding sites (TFBS) were predicted from PWMs derived for a number of yeast transcription factors based on intergenic binding (ChIP) in the recent paper by Harbison, *et al* (3). We used MotifScanner (18) with these matrices to scan 1000 bp regions in the 5′ upstream area of each open reading frame in the *Saccharomyces cerevisiae* genome. Each gene hit, and the corresponding motif sequences in the input to the expression analysis.

From a set of expression measurements (two-color microarrays), we denote expression values for a TF *A* and gene *G* as the joint distribution $\theta_{A,G}$. We used the Rosetta deletion compendium of ~300 expression conditions (19), normalizing the expression of each gene in this set to have zero mean and unit variance. We use a suitable function $\Phi(t_A, G)$ (Eq 2.1) to represent the magnitude of activation or repression of *G* with respect to different expression values of *A*. We denote this function as the log-likelihood ratio that some expression change in *G* occurs when the expression of *A* exceeds a threshold value $t_A(G)$.

$$\Phi(t_A, G) = -\delta \log_{10}\left(\frac{p[(\theta_G \mid \theta_{A>t_A}) < (\theta_G \mid \theta_{A<t_A})]}{p[(\theta_G \mid \theta_{A>t_A}) > (\theta_G \mid \theta_{A<t_A})]}\right) \quad (2.1)$$

Values obtained by Eq 2.1 are calculated using bootstrap samplings from the distribution of gene *G*, from conditions where the normalized expression level of TF *A* is either above or below a threshold value t_A. Probability terms representing the differences in the conditional distributions of *G* are calculated using the Kolmogorov-Smirnov test. A sign parameter $\delta = (-1, +1)$ indicates whether the TF is a known to be an activator (+1) or repressor (−1), based on literature annotation—high values of $\Phi(t_A, G)$ indicate significant shifts in *G* expression consistent with the TF's known function.

We estimate the optimal threshold $t_A(G)$ that maximizes the value of the statistic Φ for gene *G* by calculating $\Phi(t_A, G)$ for many values of $t_A(G)$ based on 100 bootstrapped samplings from the original expression data. Results of this sampling procedure are then aggregated to produce a mean value of the scoring statistic $\bar{\Phi}(t_A, G)$, corresponding to an average value of the threshold parameter $\bar{t}_A(G)$. The score $\bar{\Phi}(t_A, G)$ then represents an optimum in the average 'responsiveness' of gene *G* to the expression level of gene *A* at expression level $\bar{t}_A$, but doesn't imply that *G* is only regulated at this level. Because we average over many random selected models created from the data, the estimate of $\bar{\Phi}(t_A, G)$ and $\bar{t}_A$ summarizes the most persistent effects over many perturbations and conditions in the original data.

2.2 Computation of TFBS partitioning

Each TF and gene combination produces a model fit by the above procedure, with varying values of $\bar{\Phi}(t_A, G)$ and $\bar{t}_A$. Because the motif scanning procedure exhibits known low specificity, and because expression values between unrelated gene can be weakly correlated at random, it is appropriate to filter the results of the expression-based modeling to select combinations for which both motif and regulation data are both present. Numerous iterations of the procedure found $\bar{\Phi}(t_A, G) = 3.0$ to be an effective value for removing spurious associations at minimal cost to further analysis. This filtered set of modeled genes is $\mathbf{G}_A^{filtered} = \{G_1, G_2, \ldots, G_n\}$. In general, any over-fitting which is likely encountered by having many multiple independently parameterized models is avoided by using models averaged over many sets of conditions, and as we describe next, many sets of genes.

The next step of our procedure seeks to partition or permute the filtered set of modeled genes, $\mathbf{G}_A^{filtered}$, into sets having significant differences in *cis* motif content. These sets will be explicitly dependent on modeled values of $\bar{t}_A$, which are obtained in the previous stage. By selecting genes having average modeled values of $\bar{t}_A$ greater than (*high*) or less than (*low*) some threshold value t_A^{crit}, we produce two new sets $\mathbf{G}_A^{high}$ and $\mathbf{G}_A^{low}$, whose aggregate maximum $\bar{\Phi}(t_A, G)$ occurs at relatively high and low levels of TF relative to t_A^{crit}. Each value of t_A^{crit} yields a new pair of sets of aligned motifs *cis* to the genes in $\mathbf{G}_A^{high}$ and $\mathbf{G}_A^{low}$, from which the information weight matrix (*I*) can be calculated by using the PWM of each set of motifs:

$$I_{PWM} = \sum_l \sum_b f(b,l) \log_2 (f,l) \quad (2.2)$$

This process is then repeated for successive values of t_A^{crit}. In addition, providing a more robust estimate the mean and variance of the information content from substituent motifs in the gene set, we additionally bootstrap the

selection of the substituent genes (x100) from both high and low sets at each chosen value of $\bar{t}$. The value of t_A^{crit} producing the maximum bootstrap significance (bootstrap $p \leq 0.01$) in PWM information difference between set $\mathbf{G}_A^{high}$ and $\mathbf{G}_A^{low}$ defines two optimal average PWMs (PWM_{high} and PWM_{low}) for the two sets. Finally, a more precise estimate of the average information difference $\Delta I_A = I(\mathbf{G}_A^{low}) - (\mathbf{G}_A^{high})$between motifs in the sets, and the significance of this final difference is estimated from a last much larger bootstrapped sampling (x1000) of gene sets having equal size as final sets $\mathbf{G}_A^{high}$ and $\mathbf{G}_A^{low}$ made from the original set $\mathbf{G}_A^{filtered}$. This probability estimate final gives the level of surprisal that the information content of PWM_{high} and PWM_{low} would diverge as much as is observed over many random samplings from our original large set of motifs.

The output for each TF, therefore, the optimal difference in PWM information entropy, ΔI_A as a function of the threshold of expression for the TF,$\bar{t}_A$. The process implicitly yield two sets of genes, $\mathbf{G}_A^{high}$ and $\mathbf{G}_A^{low}$, whose regulation appears to respond in aggregate to correspondingly high or low expression levels of TF *A*. As noted earlier, an increase in this measure of information among sets of protein-bound sequences corresponds to an increase in the statistical thermodynamic estimate of a relative binding affinity between the protein and DNA, K_{bind}^{eff}. The procedure therefore gives us way to examine the relationship between the modeled concentration of TF needed to regulate its genes, and the affinity of sites upstream of those genes.

3 RESULTS

We applied the outlined procedure to several yeast transcription factors for which PWM data was sufficiently available. We excluded transcription factors for which we could find no sites, or TFs for which predicted sites were identical across all genes; obtaining informative subsets is impossible in these cases. The results for seven TFs are shown in Table 1.

3.1 Site content and transcription factor level at transcriptionally active genes

We were able to recover a number of instances where the information content of TFBS motifs could be partitioned into significant subsets depending on the modeled level of transcription factor which corresponded to activation of different genes in these sets. A surprising result of the study was that most transcription factors (10 out of 12 cases) showed some type of significant change in information content ($p \leq 0.05$) as a function of the modeled value of t_A. The probabilities reported in Table 1 are not Bonferroni correct, however, and therefore possibly of marginal significance when judge in total, but the individual probabilities of a several cases are at 0.01 or better, and have supporting evidence as we discuss further. Based on this evidence it is likely that TF expression corresponds in some cases to changes in the information of *cis* sites on genes, but it apparent that this change can be positive or negative.

Half of the significant cases we found (5 of 10), the change in information was positive – Gcn4, Leu3, Hap1, Msn4 and Skn7 made up this set. The remaining five (Rpn4, Mcm1, Swi4, Abf1 and Ume6) had significant negative change in PWM information content between 'high' and 'low' TF-regulated gene sets. Based on the predictions of our basic model, we expect cases where the change in information is positive between high versus low TF levels to reflect those situations where concentration and binding site affinity are dominant in governing regulation. The other cases show an opposite tendency, suggesting certain TFs regulate genes with lower site affinity even at lower TF levels, perhaps due to cooperativity.

The two transcription factors shown in Table 1 (Mcm1 and Rpn4) which exhibit strong negative change in gene TFBS information content (ΔI) at low TF levels are both known participants with other factors during DNA binding and regulation. Rpn4p participates in a proteosomal auto-regulation pathway (20), while Mcm1p alternately binds to alpha1, alpha2 and Ste12p during different stages of mating and mate-type gene expression (21). The preference of Rpn4 for high-affinity sites at higher concentrations could be linked to the observed negative feedback mechanism whereby the targets of Rpn4 encode proteins that ultimately degrade the TF itself. The transcription factors which exhibit positive change in motif information content, Gcn4, Leu3, Hap1, Msn4 and Skn7 are generally non-complexed protein regulators of gene expression, although some (*e.g.* Gcn4p and Hap1p) are known to dimerize before binding DNA.

The results of the modeling procedure are suggest that potential binding mechanisms can be seen in the increased preference for certain nucleotides in averaged PWMs corresponding to low and high information sets. For example, in Skn7 sites, the additional information gained between site partitions is also entirely due to the conservation of C at position 2. The Rpn4 motif shows an increased preference for a triplet G in the beginning of the motif, and Hap1 sites show increased conservation in the middle positions between two highly conserved ends.

3.2 Biological significance of motif partitioning

Our approach assumes that the affinity of transcription factor sites can be functionally segregated based on regulatory pattens among their downstream *cis* genes. If this is true, then we should expect cases where a biological or functional interpretation can be associated with this type of partitioning. As shown in Table 1 we investigated high and low sets of modeled genes for each TF to see if they were particularly enriched for binding in large-scale ChIP or for functional information in the Gene Ontology. These tests potentially provide indirect evidence that the genes sets of different sets exhibit different specificities in binding assays, or are involved in different types of cellular processes. We report for each high and low set the most-enriched TF in ChIP binding assays and the most common functional category in the Gene Ontology. In some cases (*Gcn*4, *Mcm*1, *Msn*4) ChIP results returned the identical TF for least one of the sets. For many other cases, however, the results returned other high-scoring TFs, suggesting that the binding of these different TFs may have cross-specificity or target gene overlap. Whether this represents binding of related TFs to the same or similar upstream sequences in ChIP experiments, or the binding of TFs to each other, remains to be explored.

Different gene sets obtained by our partitioning method are often enriched for separate GO categories despite sharing the same TF, suggesting some degree of functional heterogeneity exists among genes responsive to different TF levels. For example Skn7 targets were most represented in metabolism (high *SKN*7 expression) and biogenesis (low *SKN*7 expression). Often, at least one of the gene sets for each TF represented functional categories that correspond to processes regulated by the TF. For example Leu3 target genes are involved in amino acid biosynthesis (22), Rpn4 genes in proteolysis (20), Gcn4 genes in amino acid biosynthesis (23), Hap1 in catalytic activity (24), *etc*., are consistent with literature evidence of function for these regulators.

3.3 Biological validation of motif variants

We further analyzed examples found as part of this study to see if supporting evidence exists for the binding and function of motif

Table 1. Summary of modeled PWM content relating to transcriptional activities

TF	$\langle PWM_{high} \rangle$	$\langle I_{high} \rangle$ (bits)	top ChIP^2 % (prob) top GO (prob)	$\langle PWM_{low} \rangle$	$\langle I_{low} \rangle$ (bits)	top ChIP^2 % (prob) top GO (prob)	$\Delta\langle I_{eff} \rangle$ (bits)	$\Delta\langle t \rangle$	prob (ΔI_{eff}) (random)
Gcn4		8.29 ± 0.18	***GCN4* 52.8% (0.008)** amino acid biosynthesis (5.3e-05)		9.27 ± 0.21	***GCN4* 19.5% (6.5e-12)** amino acid biosynthesis (1.6e-21)	**0.94 ± 0.28**	0.77 ± 0.45	0.005
Rpn4		12.88 ± 0.34	***ARG81* 11.1% (0.08)** proteolysis (0.02)		11.75 ± 0.19	***NRG1* 10.9% (0.02)** transcription factor complex (0.0008)	**-1.13 ± 0.39**	1.51 ± 0.49	0.007
Leu3		8.43 ± 0.14	***RIM101* 2% (0.005)** amino acid metabolism (0.01)		9.31 ± 0.34	***RPN4* 8.8% (0.009)** transport (0.0007)	**0.96 ± 0.36**	0.69 ± 0.26	0.015
Hap1		13.50 ± 0.17	***GCN4* 9.5% (0.002)** catalytic activity (7.0e-05)		14.71 ± 0.51	***TEC1* 5.0% (0.02)** catalytic activity (0.05)	**1.15 ± 0.55**	1.38 ± 0.75	0.011
Mcm1		11.66 ± 1.65	***NDD1* 27.3% (0.004)** site of polarized growth (0.02)		9.08 ± 0.20	***MCM1* 10.9% (5.0e-5)** helicase activity (5e-06)	**-2.76 ± 1.66**	2.05 ± 0.83	0.025
Msn4		10.27 ± 0.47	***CUP9* 5.0% (0.05)** carbohydrate metabolism (0.002)		11.11 ± 0.20	***MSN4* 14.1% (0.01)** generation of precursor metabolites and energy (5e-10)	**0.77 ± 0.50**	0.93 ± 0.38	0.041
Skn7		12.42 ± 0.41	***XBP1* 5.8% (0.004)** metabolism (0.0001)		13.46 ± 0.46	***IFH1* 6.7% (0.06)** cytoplasm organization and biogenesis (6e-06)	**1.08 ± 0.48**	1.51 ± 0.77	0.011

Sequence logos depict the model-averaged position-weight matrices (PWMs) for sets of genes whose regulation occurs at corresponding high and low expression of the transcription factor (TF). Mean information content of motifs in the ''high'' and ''low'' sets are given, along with the probability of measuring the observed information difference in bootstrap randomized subsets of gene sets of equal size. Measured enrichment corresponding to binding in large-scale ChIP assays (3,4), as well as the most enriched category (30) in the Gene Ontology (31), is reported for both gene sets. ΔI and Δt are the mean PWM information content and average TF expression level among modeled genes in the low minus the high sets.

variants in experimental literature. We were able to find evidence to support three of the observed motifs for TFs partitioned on the basis of regulatory activity.

Leu3
An experimental *in vitro* selection assay of dissociation constants (K_D) of 43 variants of the binding site of Leu3 was recently performed by Liu and Clarke (25), who found that several variants had higher affinity than the accepted consensus motif for this TF ('CCGGTACCGG'). The classical description of the Leu3p binding site is of two everted 'CCG' repeats separated by four nucleotides. The PWM for motifs found by scanning the yeast genome indicate that the majority of positional information is indeed found in the repeats, which are bound by two domains of the leucine zipper Leu3, regardless of the gene group targets. Our analysis predicted that a higher affinity version of this motif (Table 1) has increased positional information at the two bases at positions four and five immediately after the first repeat. The binding experiments by Liu and Clarke confirms the affinity conferred by a T at position four has over three-fold higher affinity than the consensus motif. The other T, at position five, has marginally stronger affinity than the consensus G reported for that position in the Leu3 site.

Hap1
Hap1 is a member of the Zn_2Cys_6 family of binuclear cluster TFs which bind as homodimers, typically to CGG repeats. Our analysis of Hap1 binding sites predicted that sites bound by higher levels of expression of Hap1 have enrichment for information in the spacer region at positions 3 to 8. This observation is corroborated by a recent experimental analysis of the composition of the Hap1p binding site performed by Wang, *et al* (26), who found that deletion of the spacer region between repeats lowered the effective affinity of the TF for sequences *in vitro*. They also report a binding preference for a dinucleotide TA in this spacer region.

Gcn4
The sequence specificity of the bZip family member Gcn4 has been previously reported to bind to a half site in DNA (27) with a consensus seven base-pair sequence 'TGA(C/G)TCA'(28). The preference for T and G at the first and second positions of this core, and for A at the last position, are more variable among sites in the yeast genome than the four base pairs at positions three through six. This observation agrees with the prediction made by our analysis among gene sets regulated by high and low levels of Gcn4 expression; specifically, we find there is a preference for the canonical binding motif among gene regulated at lower levels of the transcription factor, and this is weakened among gene regulated at higher levels of Gcn4 expression. An overall difference in PWM information content of 0.94 ± 0.28 (bits), as seen in Table 1, is due largely to a preference for TG at the first two positions in the core motif, as well as A at the last position.

4 DISCUSSION

The equilibrium partitioning of transcription factors among sites of different binding affinity across the genomes is a simple but potentially important mechanism that plausibly controls whole sets of genes across different conditions. Using a novel, thermodynamically motivated approach, we've presented preliminary evidence based on predicted TF binding sites and expression data in *Saccharomyces cerevisiae* that this general effect might play some role between certain transcription factors and their targeted genes. In these cases, a significant inverse relationship was noted between the aggregate information content of a set of motifs, and the expression level of the TF that putatively binds these motifs. Gene regulation in these cases is plausibly tuned to respond to various conditions by the interplay between available transcription factors and the affinity of sites to which they bind.

Our results also suggest that site affinity plays a more complicated role in the specificity of TFs acting in tandem, however, where the affinity of sites is possibly dependent on the activity of other proteins that can bind at or near the same region of DNA. The cases where we find strong evidence that regulation follows a simpler model are potentially more attractive targets for forward engineering in synthetic systems.

The analysis we've shown doesn't prove the actual physical affinities of partitioned sites, and for this further experiments will clearly be necessary. A variety of assumptions must be made in linking the information in the sites with the biophysical interpretation of binding preferences, but the observed correlations between modeled target gene activity and motif content lend support to the model we've used, and suggest that functional knowledge of biological systems can be gained by this simplification. There are also clear limitations where binding site partitioning, whether statistical or thermodynamic, is effectively impossible, for example if a particular site is completely invariant across all genes or has uniform affinity.

We note that, in general, more complicated cooperative effects aren't incompatible with the model of binding and regulation we describe here, and in fact these cases might adequately be described as modulations to the single protein equilibrium. These modulations can be obtained by changing the effective affinity of proteins for sites secondary binding events, having protein co-localization, co-orientation, *etc*. Signatures of cooperativity might therefore be detectable as in the content of combined *cis* regulatory signals, as well, or the expression and activity of combinations multiple TFs, leading to an enriched understanding of regulatory logic. The basic model we've used here can be extended to consider combinations and sets of transcription factors, where the affinity of combinations of TFs is influenced by motif content, as well as relative orientation. In such cases, factors other than the information content of motifs may play a much more important role. In work along such lines, Beer and Tavazoie (29) showed that spatial patterns in motif organization are sufficient to predict the regulatory response of many genes. The extraction of mechanistic rules in systems of combinatorial regulation is a remaining challenge for this and many other modeling approaches.

To summarize, despite a of lack direct methods to (a) verify the *in vivo* affinity of any particular *cis* sites in the genome, (b) to understand mechanisms of affinity at arbitrary protein-DNA interfaces, (c) know the effective protein concentrations of different active TFs, or (d) measure the availability of TF sites on DNA, we can still approach some of these questions with modeling methods based on available data. In this study we examined the relationships between transcription factor affinity and regulatory efficacy by using model based on an assumed physio-chemical partitioning that occurs during binding and regulation. In testing this model, and whether genes are aggregately activated or repressed in response to

high and low levels of a transcription factor's expression, we found interesting signatures of the dynamical processes involved in gene regulation. These patterns are in several cases sufficient to identify significant and functional differences between *cis*-elements to which a transcription factor binds.

ACKNOWLEDGEMENTS

The authors thank Melissa Landon, Evan Snitkin and Yaoyu Wang for helpful discussions.

REFERENCES

Moses,A.M., Chiang,D.Y., Kellis,M., Lander,E.S. and Eisen,M.B. (2003) Position specific variation in the rate of evolution in transcription factor binding sites. *BMC Evol Biol*, **3**, 19.

Frith,M.C., Li,M.C. and Weng,Z. (2003) Cluster-Buster: Finding dense clusters of motifs in DNA sequences. *Nucleic Acids Res*, **31**, 3666–3668.

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.

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.

Bilu,Y. and Barkai,N. (2005) The design of transcription-factor binding sites is affected by combinatorial regulation. *Genome Biol*, **6**, R103.

Nachman,I., Regev,A. and Friedman,N. (2004) Inferring quantitative models of regulatory networks from expression data. *Bioinformatics*, **20 Suppl 1**, I248–I256.

Granek,J.A. and Clarke,N.D. (2005) Explicit equilibrium modeling of transcription-factor binding and gene regulation. *Genome Biol*, **6**, R87.

Djordjevic,M., Sengupta,A.M. and Shraiman,B.I. (2003) A biophysical approach to transcription factor binding site discovery. *Genome Res*, **13**, 2381–2390.

Benos,P.V., Bulyk,M.L. and Stormo,G.D. (2002) Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res*, **30**, 4442–4451.

Bulyk,M.L., Johnson,P.L. 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.

Mukherjee,S., Berger,M.F., Jona,G., Wang,X.S., Muzzey,D., Snyder,M., Young,R.A. and Bulyk,M.L. (2004) Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat Genet*, **36**, 1331–1339.

Schneider,T.D., Stormo,G.D., Gold,L. and Ehrenfeucht,A. (1986) Information content of binding sites on nucleotide sequences. *J Mol Biol*, **188**, 415–431.

Berg,O.G. and von Hippel,P.H. (1987) Selection of DNA binding sites by regulatory proteins Statistical-mechanical theory and application to operators and promoters. *J Mol Biol*, **193**, 723–750.

Cherry,J.M., Adler,C., Ball,C., Chervitz,S.A., Dwight,S.S., Hester,E.T., Jia,Y., Juvik,G., Roe,T., Schroeder,M. *et al.* (1998) SGD: Saccharomyces Genome Database. *Nucleic Acids Res*, **26**, 73–79.

Friedman,N., Linial,M., Nachman,I. and Pe'er,D. (2000) Using Bayesian networks to analyze expression data. *J Comput Biol*, **7**, 601–620.

Pe'er,D., Regev,A., Elidan,G. and Friedman,N. (2001) Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, **17 Suppl 1**, S215–224.

Pilpel,Y., Sudarsanam,P. and Church,G.M. (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat Genet*, **29**, 153–159.

Aerts,S., Thijs,G., Coessens,B., Staes,M., Moreau,Y. and De Moor,B. (2003) Toucan: deciphering the cis-regulatory logic of coregulated genes. *Nucleic Acids Res*, **31**, 1753–1764.

Hughes,T.R., Marton,M.J., Jones,A.R., Roberts,C.J., Stoughton,R., Armour,C.D., Bennett,H.A., Coffey,E., Dai,H., He,Y.D. *et al.* (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.

Xie,Y. and Varshavsky,A. (2001) RPN4 is a ligand, substrate, and transcriptional regulator of the 26S proteasome: a negative feedback circuit. *Proc Natl Acad Sci U S A*, **98**, 3056–3061.

Zhong,H., McCord,R. and Vershon,A.K. (1999) Identification of target sites of the alpha2-Mcm1 repressor complex in the yeast genome. *Genome Res*, **9**, 1040–1047.

Friden,P. and Schimmel,P. (1988) LEU3 of Saccharomyces cerevisiae activates multiple genes for branched-chain amino acid biosynthesis by binding to a common decanucleotide core sequence. *Mol Cell Biol*, **8**, 2690–2697.

Hinnebusch,A.G. (2005) Translational regulation of GCN4 and the general amino acid control of yeast. *Annu Rev Microbiol*, **59**, 407–450.

Becerra,M., Lombardia-Ferreira,L.J., Hauser,N.C., Hoheisel,J.D., Tizon,B. and Cerdan,M.E. (2002) The yeast transcriptome in aerobic and hypoxic conditions: effects of hap1, rox1, rox3 and srb10 deletions. *Mol Microbiol*, **43**, 545–555.

Liu,X. and Clarke,N.D. (2002) Rationalization of gene regulation by a eukaryotic transcription factor: calculation of regulatory region occupancy from predicted binding affinities. *J Mol Biol*, **323**, 1–8.

Wang,L.L., Denman,I. and Junker,M. (2004) Control of Hap1-DNA site recognition through the interplay of multiple distinct intermolecular interactions. *Biochemistry*, **43**, 13816–13826.

Hollenbeck,J.J. and Oakley,M.G. (2000) GCN4 binds with high affinity to DNA sequences containing a single consensus half-site. *Biochemistry*, **39**, 6380–6389.

Oliphant,A.R., Brandl,C.J. and Struhl,K. (1989) Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast GCN4 protein. *Mol Cell Biol*, **9**, 2944–2949.

Beer,M.A. and Tavazoie,S. (2004) Predicting gene expression from sequence. *Cell*, **117**, 185–198.

Boyle,E.I., Weng,S., Gollub,J., Jin,H., Botstein,D., Cherry,J.M. and Sherlock,G. (2004) GO::TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, **20**, 3710–3715.

Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*, **32**, D258–261.