BIOINFORMATICS

Vol. 22 no. 14 2006, pages e141–e149
doi:10.1093/bioinformatics/btl223

Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE

Barrett C. Foat[1], Alexandre V. Morozov[2] and Harmen J. Bussemaker[1,3,*]

[1]Department of Biological Sciences, Columbia University, New York, NY 10027, USA, [2]Center for Studies in Physics and Biology, The Rockefeller University, New York, NY 10021, USA and [3]Center for Computational Biology and Bioinformatics, Columbia University, New York, NY 10032, USA

ABSTRACT

Motivation: Regulation of gene expression by a transcription factor requires physical interaction between the factor and the DNA, which can be described by a statistical mechanical model. Based on this model, we developed the MatrixREDUCE algorithm, which uses genome-wide occupancy data for a transcription factor (*e.g.* ChIP-chip) and associated nucleotide sequences to discover the sequence-specific binding affinity of the transcription factor. Advantages of our approach are that the information for all probes on the microarray is efficiently utilized because there is no need to delineate “bound” and “unbound” sequences, and that, unlike information content-based methods, it does not require a background sequence model.

Results: We validated the performance of MatrixREDUCE by inferring the sequence-specific binding affinities for several transcription factors in *S. cerevisiae* and comparing the results with three other independent sources of transcription factor sequence-specific affinity information: (i) experimental measurement of transcription factor binding affinities for specific oligonucleotides, (ii) reporter gene assays for promoters with systematically mutated binding sites, and (iii) relative binding affinities obtained by modeling transcription factor-DNA interactions based on co-crystal structures of transcription factors bound to DNA substrates. We show that transcription factor binding affinities inferred by MatrixREDUCE are in good agreement with all three validating methods.

Availability: MatrixREDUCE source code is freely available for non-commercial use at http://www.bussemakerlab.org/. The software runs on Linux, Unix, and Mac OS X.

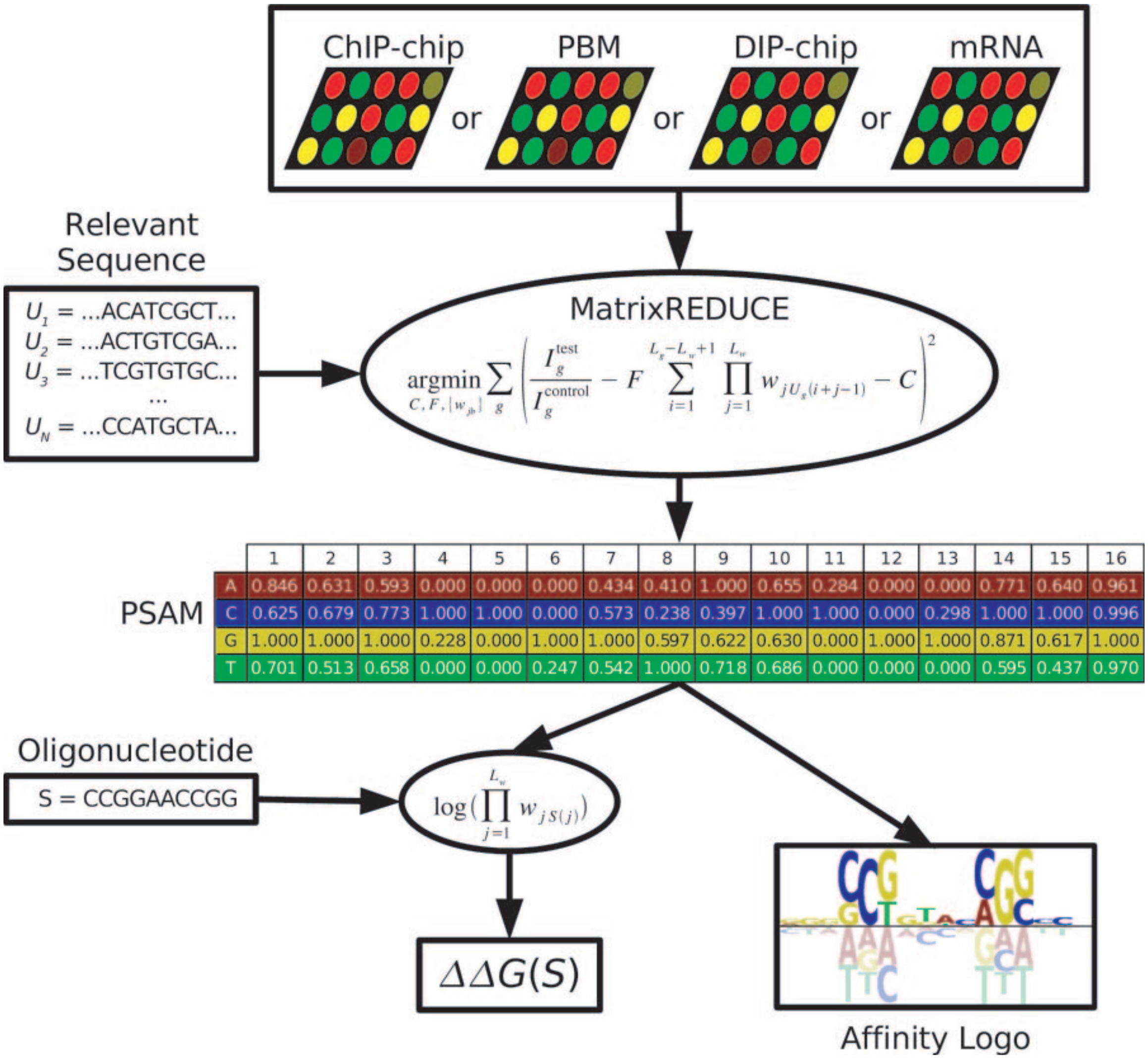
Contact: Harmen.Bussemaker@columbia.edu

1 INTRODUCTION

The sequence-specific regulatory activity of a transcription factor (TF) is the result of energetically favorable interactions between the amino acids exposed in the DNA binding domain and portions of nucleic acid bases exposed in the grooves of the DNA. A computational method for discovering the binding specificity of a TF cannot provide a quantitative description of TF binding unless it considers the physical underpinnings of the TF-DNA interaction. Most physically motivated computational methods discover over-represented patterns in a set of nucleotide sequences that are considered to be bound by the TF (for review see Stormo, 2000). These methods use the information content of nucleotide patterns as a proxy for the free energy contributions of the bases found in the TF binding site (Berg and von Hippel, 1987; Stormo and Fields, 1998). Other computational methods infer physically-based TF binding specificities from measured TF binding affinities for a small set of oligonucleotides (Liu and Clarke, 2002) or from structural modeling of protein-DNA interaction (Paillard and Lavery, 2004; Endres *et al*., 2004; Morozov *et al*., 2005). However, genome-scale, quantitative measurements of TF occupancies of intergenic regions are now available due to the advent of *in vivo* chromatin immunoprecipitation microarrays (Ren *et al*., 2000; Iyer *et al*., 2001; Lieb *et al*., 2001; Simon *et al*., 2001; Lee *et al*., 2002; Harbison *et al*., 2004), *in vitro* protein binding microarrays (PBM; Mukherjee *et al*., 2004), and DNA immunoprecipitation microarrays (DIP-chip; Liu *et al*., 2005). Thus, it is no longer necessary to rely on small data sets, availability of protein-DNA structures, or the analogy between information content and statistical mechanics to infer free energy representations of transcription factor binding sites.

We have developed a method, implemented as the program MatrixREDUCE (Foat *et al*., 2005), that infers the sequence specificity of a TF directly and accurately from genome-wide TF occupancy data by fitting a statistical mechanical model for TF-DNA interaction (Figure 1). The sequence specificity of the TF’s DNA-binding domain is modeled using a position-specific affinity matrix (PSAM), representing the change in the binding affinity (K_d) whenever a specific position within a reference binding sequence is mutated. To validate the physical model of MatrixREDUCE, we discovered the PSAMs for several TFs in *S. cerevisiae* and compared the results with three other independent sources of TF sequence-specific affinity information: (i) experimentally measured K_d’s as determined by *in vitro* methods (Gailus-Durner *et al*., 1996; Liu and Clarke, 2002; Pierce *et al*., 2003), (ii) *lacZ* reporter assays for promoters with systematically mutated binding sites (Gailus-Durner *et al*., 1996; Pierce *et al*., 2003), and (iii) relative K_d’s obtained by using a physical model of protein-DNA interaction that makes binding affinity predictions starting from a co-crystal structure of the protein-DNA complex (Morozov *et al*., 2005). We find a surprising level of agreement between MatrixREDUCE-predicted TF binding affinities, experimental measurements, and structural predictions, suggesting that MatrixREDUCE is a

[*]To whom correspondence should be addressed.


ChIP-chip
PBM
DIP-chip
mRNA
or
or
or
Relevant Sequence
U_1 = ...ACATCGCT...
U_2 = ...ACTGTCGA...
U_3 = ...TCGTGTGC...
...
U_N = ...CCATGCTA...
MatrixREDUCE
$\underset{C,F,\{w_{jb}\}}{\operatorname{argmin}} \sum_g \left(\frac{I_g^{\text{test}}}{I_g^{\text{control}}} - F \sum_{i=1}^{L_g - L_w + 1} \prod_{j=1}^{L_w} w_{jU_g(i+j-1)} - C \right)^2$
PSAM
Oligonucleotide
S = CCGGAACCGG
$\log\left(\prod_{j=1}^{L_w} w_{jS(j)}\right)$
$\Delta\Delta G(S)$
Affinity Logo

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
A	0.846	0.631	0.593	0.000	0.000	0.000	0.434	0.410	1.000	0.655	0.284	0.000	0.000	0.771	0.640	0.961
C	0.625	0.679	0.773	1.000	1.000	0.000	0.573	0.238	0.397	1.000	1.000	0.000	0.298	1.000	1.000	0.996
G	1.000	1.000	1.000	0.228	0.000	1.000	1.000	0.597	0.622	0.630	0.000	1.000	1.000	0.871	0.617	1.000
T	0.701	0.513	0.658	0.000	0.000	0.247	0.542	1.000	0.718	0.686	0.000	0.000	0.000	0.595	0.437	0.970

Fig. 1. The flow of data. A microarray measurement of TF occupancies (ChIP-chip, PBM, DIP-chip, or differential mRNA expression data) and relevant nucleotide sequences for each microarray feature are used as input to MatrixREDUCE. MatrixREDUCE performs a least-squares fit to a statistical-mechanical model of TF-DNA interaction to discover the relative contributions to the free energy of binding for each nucleotide at each position in the generalized TF binding site. These contributions are represented as a position specific affinity matrix (PSAM) containing the relative equilibrium constants of the TF-DNA interaction, with the highest affinity nucleotide at each position scaled to a value of one ($\Delta\Delta G = 0$). The PSAM can be converted into an affinity logo that graphically represents the $\Delta\Delta G$'s for each nucleotide at each position relative to the average $\Delta\Delta G$ at the respective positions. The PSAM can also be used to predict the relative TF occupancy of any nucleotide sequence, allowing the PSAMs inferred by MatrixREDUCE to be compared with experimental measurements of TF binding affinities for particular oligonucleotides.

powerful and accurate tool for the elucidation of physically accurate TF sequence-specific binding affinities.

2 RELATED WORK

In contrast to information theory-based methods of defining nucleotide-binding protein specificities, MatrixREDUCE belongs to a small but growing class of methods that infer binding affinities by directly fitting a physical model to experimental data. The first such method was introduced by Stormo *et al*. (1986) who noted, ''When quantitative data are known for many sequences one can solve for the matrix elements that give the best fit between the sequences and those data.'' For Stormo *et al*. (1986) the quantitative data were β-galactosidase activities for genes containing mutated binding sites of the *E. coli su*2 amber stop codon suppressor. A similar type of analysis was performed by Liu and Clarke (2002) who fit a physical model for transcription factor binding to electrophoretic mobility shift assay (EMSA) data measuring affinity of the *S. cerevisiae* Leu3 TF for several oligonucleotides. The physical model behind MatrixREDUCE is the same as that employed by Stormo *et al*. (1986) and Liu and Clarke (2002). However, our ''quantitative data'' are microarray probe intensities, which measure TF occupancy over long chromosomal regions with unknown binding site locations. Thus, the MatrixREDUCE model integrates the binding signal over the entire length of the sequence. The GOMER method of Granek and Clarke (2005) performs a similar integration of signal over long regulatory sequences relevant to measured microarray intensities. However, GOMER was only used to test hypotheses about the regulatory mechanisms of TFs for which a binding site weight matrix had already been defined by other methods. Granek and Clarke (2005) did not attempt to fit the GOMER model directly to experimental data to infer the binding affinities of TFs. Finally, also of note are the QPMEME algorithm of Djordjevic *et al.* (2003) and the work of Djordjevic and Sengupta (2006) which use maximum likelihood procedures to infer PSAMs by fitting physical models to known TF binding sites and SELEX data, respectively, but rely on the prior delineation of ‘‘bound’’ sequences.

3 METHODS

3.1 Modeling TF-DNA interaction

We will develop the statistical-mechanical model used by MatrixREDUCE starting with a transcription factor P that binds to a DNA sequence S to form the TF-DNA complex PS:

$$P + S \underset{k_{\text{off}}}{\overset{k_{\text{on}}}{\rightleftarrows}} PS \tag{1}$$

The affinity of the TF for the sequence can be expressed in terms of its equilibrium dissociation constant $K_d(S)$:

$$K_d(S) = \frac{[P][S]}{[PS]} = \frac{k_{\text{off}}}{k_{\text{on}}} = e^{\Delta G/RT}, \tag{2}$$

which is directly related to ΔG, the Gibbs free energy of binding per mole (R is the gas constant and T is temperature). The occupancy $N(S)$ of sequence S by transcription factor P can be expressed as the concentration of TF-DNA complex divided by the total concentration of DNA (bound or unbound):

$$N(S) = \frac{[PS]}{[PS] + [S]} = \frac{[P]}{[P] + K_d(S)}. \tag{3}$$

For simplicity, we will assume that the TF concentration $[P]$ is much smaller than $K_d(S)$. This assumption seems physiologically plausible because in this regime, the highest affinity binding sites in the genome will be the most responsive to a change in the nuclear concentration of active TF. Thus, the occupancy becomes:

$$N(S) \approx \frac{[P]}{K_d(S)} = [P]K_a(S), \tag{4}$$

where

$$K_a(S) \equiv K_d^{-1}(S). \tag{5}$$

Consider a single point mutation from the original reference sequence S_{ref} to base b at position j resulting in the mutated sequence S_{mut}. Such a mutation will give rise to an additive change $\Delta\Delta G$ in the free energy of binding or, equivalently, a multiplicative change w_{jb} in $K_a(S_{\text{ref}})$:

$$w_{jb} = \frac{K_a(S_{\text{mut}})}{K_a(S_{\text{ref}})} = e^{\Delta\Delta G/RT}, \tag{6}$$

where

$$\Delta\Delta G = \Delta G(S_{\text{ref}}) - \Delta G(S_{\text{mut}}). \tag{7}$$

To be able to generalize the binding of transcription factor P to a sequence S_{mut} with more than one point mutation, we assume that the free energy contributions for each position in the binding site are independent (Benos *et al.*, 2002) and therefore additive. Equivalently, we can multiply the w_{jb}'s for any nucleotide sequence to obtain the overall $K_a(S_{\text{mut}})/K_a(S_{\text{ref}})$ ratio. Thus, the occupancy of a particular binding site S_{mut} of length L_w with nucleotide sequence $S_{\text{mut}}(1, 2, \ldots, L_w) = (b_1, b_2, \ldots, b_{L_w})$ is:

$$N(S_{\text{mut}}) = [P]K_a(S_{\text{ref}}) \prod_{j=1}^{L_w} w_{jS_{\text{mut}}(j)}. \tag{8}$$

The occupancy $N(U_g)$ for the entire promoter region U_g of gene g equals the sum of occupancies for each binding site window of length L_w at each position i over the length L_g of the sequence U_g:

$$N(U_g) = [P]K_a(S_{\text{ref}}) \sum_{i=1}^{L_g - L_w + 1} \prod_{j=1}^{L_w} w_{jU_g(i+j-1)}, \tag{9}$$

where $U_g(i)$ is the base at position i in sequence U_g.

3.2 Modeling genome-wide TF occupancy data

Recent technologies such as ChIP-chip (Ren *et al.*, 2000; Iyer *et al.*, 2001; Lieb *et al.*, 2001; Simon *et al.*, 2001; Lee *et al.*, 2002; Harbison *et al.*, 2004), PBM (Mukherjee *et al.*, 2004), and DIP-chip (Liu *et al.*, 2005) provide indirect but quantitative information about the TF occupancy of large genomic regions. For each segment of DNA there are two microarray intensities. The test intensity I_g^{test} is equal to a background intensity α^{test} plus a term that, to first approximation, is proportional (γ) to the occupancy $N(U_g)$ by the TF, either because the amount of TF bound to the probe contributes directly to the signal intensity (PBM) or because it determines the proportion at which an immunoprecipitated TF-DNA fragment is present in the sample (ChIP-chip or DIP-chip). The control intensity I_g^{control} is only the result of background signal α^{control}. Allowing for experimental noise ϵ_g, we obtain:

$$\frac{I_g^{\text{test}}}{I_g^{\text{control}}} = \frac{\gamma N(U_g) + \alpha^{\text{test}}}{\alpha^{\text{control}}} + \epsilon_g \equiv \beta N(U_g) + C + \epsilon_g \tag{10}$$

Using Equation 9 for the occupancy $N(U_g)$, we obtain:

$$\frac{I_g^{\text{test}}}{I_g^{\text{control}}} = F \sum_{i=1}^{L_g - L_w + 1} \prod_{j=1}^{L_w} w_{jU_g(i+j-1)} + C + \epsilon_g, \tag{11}$$

where

$$F = \beta[P]K_a(S_{\text{ref}}). \tag{12}$$

Note that β, $[P]$, and $K_a(S_{\text{ref}})$ cannot be determined separately without additional information such as the real protein concentration or $K_a(S_{\text{ref}})$.

MatrixREDUCE discovers the set of w_{jb} elements as well as F and C by performing a least squares fit to the measured intensity ratios:

$$(C, F, \{w_{jb}\}) = \underset{C, F, \{w_{jb}\}}{\text{argmin}}$$

$$\sum_g \left(\frac{I_g^{\text{test}}}{I_g^{\text{control}}} - F \sum_{i=1}^{L_g - L_w + 1} \prod_{j=1}^{L_w} w_{jU_g(i+j-1)} - C \right)^2. \tag{13}$$

The $4 \times L_w$ matrix of K_a ratios w_{jb} ($3L_w$ parameters plus L_w reference nucleotide values) for all nucleotides at all positions in the binding site is referred to as the position specific affinity matrix (PSAM). Each position j in the PSAM is rescaled such that the largest w_{jb} is equal to unity, without loss of generality.

Differential mRNA expression microarray data, which measures the change in mRNA concentrations in cells from two different experimental conditions, can be used in place of genome-wide TF occupancy data. This substitution is reasonable since, to first approximation, the transcription rate of genes is proportional to the total TF occupancy along the associated promoter regions. Genome-wide occupancy data is preferable, however, since it is a more direct measure of TF-DNA interaction and since the design of the experiments provides the TF identities for the discovered PSAMs.

3.3 MatrixREDUCE implementation and parameters

MatrixREDUCE was implemented in Perl and C as outlined above and as previously described (Foat *et al.*, 2005) with some modifications. Briefly, MatrixREDUCE takes microarray intensities and corresponding nucleotide sequence data as input. It first finds a gapped dyad motif (*e.g.* Leu3: CCG-4nt-CGG), out of all possible dyad motifs of a fixed number of nucleotides and a range of gap sizes, whose occurrences best correlate with the measured intensities for the same sequences. The best dyad motif is then converted into a seed matrix by filling in the gap with N's and extending out a user defined number of flanking N's on either side of the best-scoring dyad. In the $4 \times L_w$ seed matrix, acceptable nucleotides (all nucleotides for N's, a single specific nucleotide at positions within the top scoring motif) are given K_a ratios of one and unacceptable nucleotides are given a very small K_a ratio $w_{\min}$. This seed matrix serves as the starting point for a quasi-Newton numerical minimization of Equation 13 to find the optimal PSAM. The new version of MatrixREDUCE uses a k-fold cross-validation to determine the significance of each discovered PSAM. After converging on a PSAM, the input data is split into k random subsets of array features with associated sequences. The optimal PSAM is then used to seed each of k re--optimizations of the PSAM. A t-value (Pearson correlation) for the goodness of fit is calculated for the optimal PSAM of each subset. Finally, the P-value corresponding to the average t-value for the k re-optimizations is used to test whether the originally optimized PSAM should be kept. This procedure does not test the significance of the optimal PSAM itself, but rather it tests whether the data contains widely distributed, explainable variance. Thus, false PSAMs due to a few outliers are prevented. While not relevant to the current study, MatrixREDUCE can iteratively build a linear model of multiple PSAMs that best explain a particular data set (see Foat *et al.*, 2005).

The parameters for the runs of MatrixREDUCE were as follows: For all runs, the length of each of the two dyads of the seed motifs was three, the length of the added flanks on each side of the dyad was three, the minimum gap was zero, the k cross-validations were two, and $w_{\min}$ was 10^{-5}. For all runs on ChIP-chip and PBM data, the maximum acceptable P-value was 10^{-3} and the maximum dyad gap was twenty. For all runs on DIP-chip data, the maximum acceptable P-value was 10^{-6} and the maximum dyad gap was ten. For all runs on differential mRNA expression data, the maximum acceptable P-value was 10^{-3} and the maximum dyad gap was eleven.

3.4 Microarray and sequence data

All microarray data was gathered from publication supplements. We chose specific TFs to analyze based on the availability of experimental K_d data or crystal structure data. PSAMs were inferred by MatrixREDUCE for chromatin immunoprecipitation microarrays (ChIP-chip) using the microarray data and microarray feature sequences from Harbison *et al.* (2004). These ChIP-chip experiments were performed under a variety of culture conditions, including rich media (YPD); sulfometuron methyl (SM), an inhibitor of amino acid biosynthesis; and treatment with rapamycin (RAPA). PSAMs were inferred for PBM experiments using the microarray data from Mukherjee *et al.* (2004) and the feature sequence data from Harbison *et al.* (2004) as the two studies used the same array features. PSAMs were inferred for Leu3 using the DIP-chip microarray data and feature sequences from Liu *et al.* (2005). Liu *et al.* (2005) performed DIP-chip experiments using two different concentrations of Leu3, 4nM and 40nM, and PSAMs were inferred for each concentration. The PSAM for Ndt80 was inferred from differential mRNA expression microarray data measuring the sporulation response in a *ndt*80 deletion strain versus a wild-type strain (Chu *et al.*, 1998). The sequence data for the Ndt80 PSAM inference was the 800 bp upstream of every yeast gene, retrieved from the *Saccharomyces* Genome Database (Issel-Tarver *et al.*, 2002) and purged of redundant sequences as previously described (Foat *et al.*, 2005). All microarray intensities were analyzed as the ratio of the experimental sample intensity to the control sample intensity with the exception of the *ndt*80 deletion data, which was analyzed as the $\log_2$-ratio. All microarray data was purged of extreme outliers before being analyzed by MatrixREDUCE (Grubbs' test, P-value $= 10^{-10}$; Grubbs, 1969).

3.5 Gel shift and *lacZ* expression data

While prone to their own inaccuracies, experimentally measured *in vitro* binding affinities and changes in *lacZ* expression served as our ''gold standards'' to assess the validity of our MatrixREDUCE model. The electrophoretic mobility shift assay (EMSA) is able to provide direct estimates of K_d's for a TF binding to particular oligonucleotides (Fried and Crothers, 1981). The ratio of the EMSA-measured K_d of a reference oligonucleotide S_{ref} to the K_d of one of the other tested oligonucleotides S_{mut} provides the same information as the product across the MatrixREDUCE PSAM over the same sequence for the same TF. In the simplifying scenario where the length of the oligonucleotides is the same as the length L_w of the PSAM, we have

$$\frac{K_d(S_{\text{ref}})}{K_d(S_{\text{mut}})} = \prod_{j=1}^{L_w} w_{jS_{\text{mut}}(j)}. \tag{14}$$

While the biological processes involved are considerably more complex, *lacZ* expression data can be employed to the same end. If we assume that β-galactosidase activity, concentration of β-galactosidase, the amount of mRNA expressed, the specific recruitment of RNA polymerase to the promoter, and the promoter occupancy by the TF are all proportional to each other, then relative K_d's are reflected in the ratio of β-galactosidase activities between the assay using the reference binding site and another assay using a different binding site. Thus, we used *lacZ* reporter expression assay data in a similar manner to EMSA-derived K_d data to confirm the results of MatrixREDUCE.

Experimentally determined *in vitro* binding affinities and *lacZ* reporter expression activity data were gathered from publications. The K_d data and *lacZ* expression data for Abf1 are from Gailus-Durner *et al.* (1996); K_d data for Leu3 are from Liu and Clarke (2002); and K_d data and *lacZ* expression data for Ndt80 and Sum1 are from Pierce *et al.* (2003).

To compare the experimental K_d measurements with MatrixREDUCE PSAMs, all experimental K_d and *lacZ* expression data was first converted to K_a ratios by normalizing with respect to the value of the highest affinity oligonucleotide. The K_a ratios were then log-transformed to obtain the $\Delta\Delta G$ values. MatrixREDUCE PSAMs for each TF were converted to $\Delta\Delta G$'s relative to the highest affinity oligonucleotide from the respective experiment. The sum of the $\Delta\Delta G$ values was calculated for the best PSAM-matching window in each of the experimentally tested sequences. If a sequence was shorter than the PSAM, the sum was taken over only the best matching positions within the PSAM. All experimental $\Delta\Delta G$'s were then compared to the PSAM $\Delta\Delta G$'s by plotting and by calculating Pearson correlations.

BioProspector (Liu *et al.*, 2001) and MDscan (Liu *et al.*, 2002) are popular information theory-based methods for determination of TF binding specificities. To compare the quality of the results from these methods with MatrixREDUCE results, position-specific scoring matrices (PSSMs) were derived from BioProspector and MDscan outputs by calculating the frequencies of each base at each position in the putative binding sites and then dividing by a background frequency for each respective base. Two different background frequencies were tested: equal nucleotide probabilities and nucleotide probabilities for intergenic sequences in *S. cerevisiae*. Once the PSSMs had been created, they were tested against experimental EMSA and *lacZ* data in the same manner as the MatrixREDUCE PSAMs above.

3.6 Structural modeling

DNA binding affinities and specificities of TFs are determined by the forces of electrostatics, solvation, the hydrogen bonding patterns, and shape complementarity at the binding interface. The magnitude of these contributions to the binding free energy can in principle be calculated given a structure of the protein bound to its cognate DNA site. Therefore, it should be possible to predict PSAMs starting from the experimentally available structure of the protein-DNA complex (solved by either X-ray diffraction or NMR), or, in the absence of the exact structure, from a suitably constructed homology model. Under the assumption that the base pair energies contribute approximately independently to the total binding affinity (Benos *et al.*, 2002), all one-point base pair mutations are introduced into the DNA binding site. Protein-DNA binding energies $\Delta G =$

$G_{\mathrm{prot-dna}} - G_{\mathrm{prot}} - G_{\mathrm{dna}}$ are then evaluated for each mutation. Mutations in the reference binding site result in changes of protein-DNA binding energies ($\Delta\Delta G$; Equation 7). A table of $\Delta\Delta G$ values can be used to construct a PSAM that is directly comparable with MatrixREDUCE predictions.

We have previously developed two alternative approaches for predicting TF binding affinities and specificities starting from the protein-DNA structure (Morozov *et al*., 2005). In one approach, the ''all atom model'' (which builds on the ROSETTA protein-nucleic acid interaction model of Havranek *et al*., 2004), both direct and indirect readout mechanisms contribute to the recognition of the DNA binding site: $\Delta G = \Delta G_{\mathrm{direct}} + \Delta G_{\mathrm{indirect}}$. Direct readout is mediated by protein amino acid-DNA base interactions, while indirect readout is encoded in the shape of the DNA site imparted by the bound protein, primarily through non-specific amino acid-DNA phosphate backbone contacts. Direct protein-DNA interactions are modeled as a linear combination of the repulsive and attractive parts of the Lennard-Jones potential, the orientation-dependent hydrogen bonding potential (Kortemme *et al*., 2003), and the Generalized Born electrostatics and solvation model (Onufriev *et al*., 2004):

$$\begin{aligned}\Delta G_{\mathrm{direct}} = {} & w_{\mathrm{LJrep}} E_{\mathrm{LJrep}} + w_{\mathrm{LJattr}} E_{\mathrm{LJattr}} \\ & + w_{\mathrm{hb}} E_{\mathrm{hb}} + w_{\mathrm{el}} G_{\mathrm{el}},\end{aligned} \tag{15}$$

where each term is a sum over all protein-DNA and protein-protein atomic pairs, and $\{w\}$ is a set of fitting weights. Indirect readout is modeled using an effective harmonic representation of the DNA conformational energy (Olson *et al*., 1998):

$$\Delta G_{\mathrm{indirect}} = w_{\mathrm{dna-bp}} \sum_{\mathrm{bp}} E^{\alpha\beta}_{\mathrm{dna-bp}} + w_{\mathrm{dna-bs}} \sum_{\mathrm{bs}} E^{\alpha\beta}_{\mathrm{dna-bs}}, \tag{16}$$

where the first sum is over all base pairs in the DNA site (α, β denote bases in a base pair), and the second sum is over all consecutively stacked base pair steps (α, β denote base pairs in a base step). Base pairs and base steps are counted once in the $5'$ to $3'$ direction. The first term penalizes deviations from canonical base pairing, while the second term captures base stacking energies. The quadratic energy terms are given by:

$$E^{\alpha\beta}_{\mathrm{dna-bs/dna-bp}} = \frac{1}{2} \sum_{i=1}^{6} \sum_{j=1}^{6} f^{\alpha\beta}_{ij} \delta\theta^{\alpha}_{i} \delta\theta^{\beta}_{j}, \tag{17}$$

where the sums run over six geometric degrees of freedom θ_i (Twist, Tilt, Roll, Shift, Slide and Rise for base pair steps; Opening, Buckle, Propeller, Shear, Stretch and Stagger for base pairs; Lu and Olson, 2003). The DNA potential is a quadratic expansion in $\delta\theta_i$ (deviations of the degrees of freedom θ_i from their average values computed using a set of non-homologous protein-DNA complexes). The force constants f_{ij} are evaluated by inverting the covariance matrix of $\delta\theta_i$ obtained with the same protein-DNA dataset: $f^{-1}_{ij} = \langle \delta\theta_i \delta\theta_j \rangle$. All six weights are simultaneously fit to experimental $\Delta\Delta G$ data using a generalized linear model (implemented in the statistical software package R): $(w_{\mathrm{LJrep}}, w_{\mathrm{LJattr}}, w_{\mathrm{hb}}, w_{\mathrm{el}}, w_{\mathrm{dna-bp}}, w_{\mathrm{dna-bs}}) = (0.00, 0.46, 0.77, 0.27,\ 0.03, 0.03)$. No conformational flexibility is allowed at the protein-DNA interface. Further details on the fitting procedure and comprehensive tests of the all-atom free energy function can be found in Morozov *et al*. (2005).

In another approach, we developed a ''contact model'' that exploits the structure of the protein-DNA complex bound to a high affinity reference DNA sequence but does not require detailed predictions of protein-DNA interaction energies. In the contact model each mutated base in the PSAM column i incurs equal energy cost relative to the consensus base from the reference sequence:

$$\Delta\Delta G^{i}(N) = \begin{cases} E_{\max}\ [f_1(N_{\max}) \log(1 - N/N_{\max}) \\ \quad - f_2(N_{\max}) \log(1 + 3N/N_{\max})]\ (N < N_{\max}) \\ E_{\max}\ (N \geq N_{\max}) \end{cases} \tag{18}$$

Here, N is the number of protein amino acid-DNA base atomic contacts summed over the base pair i (atomic contact is defined by a distance of less than 4.5 Å; hydrogen atoms are excluded from the counts), and $N_{\max}$ is the number of contacts above which the maximum energy penalty $E_{\max}$ is imposed. $f_1(N_{\max})$ and $f_2(N_{\max})$ are fixed prefactors defined in Morozov *et al*. (2005). $E_{\max}$ together with $N_{\max}$ constitute the free parameters of the contact model and are adjusted simultaneously to maximize the fraction of correct predictions and minimize the average error over the $\Delta\Delta G$ data set identical to that used in fitting the all-atom model. The fraction of correct predictions is based on a binary function: a prediction is considered to be correct if both computational and experimental $\Delta\Delta G$'s are less than 1.0 kcal/mol, or greater than 1.0 kcal/mol, or else separated by less than 0.3 kcal/mol. The global minimum for the fit is found by exhaustive search; the best fit is obtained with $N_{\max} = 15$, $E_{\max} = 3.0$ kcal/mol.

3.7 Affinity logos

Information content-based weight matrices are usually displayed as sequence logos (Schneider and Stephens, 1990) However, MatrixREDUCE weight matrices are discovered without a background sequence model. Thus, an appropriate logo should display the actual relative free energies of binding for each nucleotide at each position rather than information content. Therefore, we created affinity logos, which are constructed as follows: For each position in the PSAM, the average $\Delta\Delta G$ is calculated. Then, the difference between each individual $\Delta\Delta G$ and the average $\Delta\Delta G$ at that position is computed; the absolute value of this difference is the height of the character representing that nucleotide. If the difference is positive (more favorable than average), the letter is placed above a horizontal black line through the center of the logo. If the difference is negative (less favorable than average) the letter is placed below the black line. Larger letters are stacked on smaller letters moving outward from the black line. The height of the letter can be interpreted as free energy difference from the average in units of *RT*. Thus, an intuitive high amplitude is given to the nucleotide positions that most contribute to the sequence specificity of the TF. To highlight that the characters representing the high affinity nucleotides are above the black line, the characters representing the low affinity nucleotides are made partially transparent. However, maintaining the representation of the poor affinity nucleotides below the center line allows the viewer to immediately see which nucleotide substitutions are most unfavorable to binding.

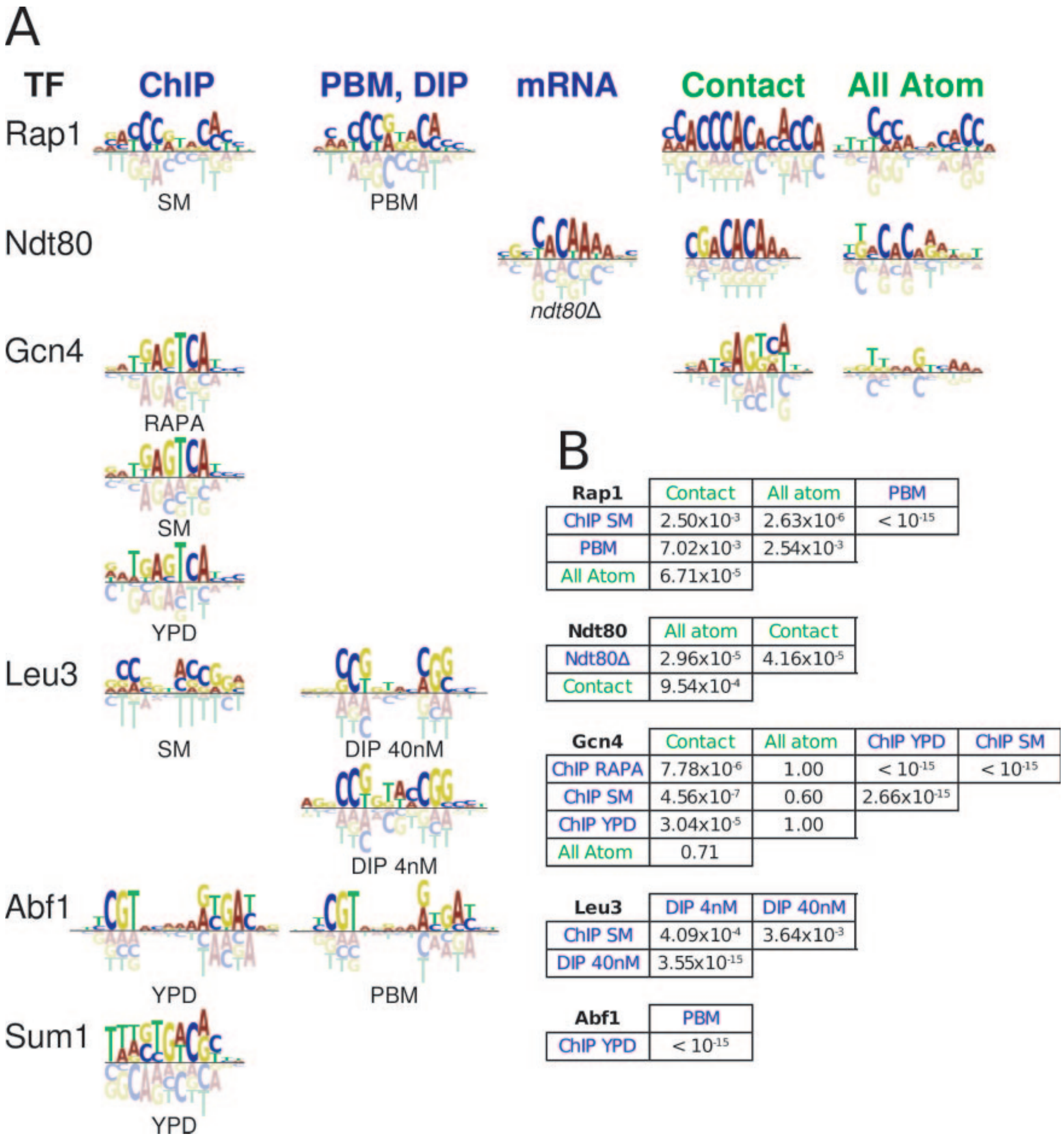
3.8 PSAM to PSAM alignments and correlations

By inspection of affinity logos, one can make qualitative observations about the similarity between any two PSAMs. However, a quantitative measure of similarity allows for more objective comparisons. Before two PSAMs can be compared, they must first be aligned. Pearson correlations were calculated between the $\Delta\Delta G$ values for each nucleotide at each position for every possible overlap of the two PSAMs for both the forward and the reverse complement alignments. After the best overlap position and strand was determined from the best correlation *P*-value, the $\Delta\Delta G$'s of the two PSAMs were recentered relative to a common reference consensus sequence. Finally, the *P*-value for the Pearson correlation between the two optimally aligned and transformed PSAMs was calculated and subjected to a Bonferroni correction for the number of alignments that were tested.

4 RESULTS

4.1 PSAMs inferred by MatrixREDUCE agree well with experimental measurements of TF binding affinity

We discovered the position specific affinity matrices (PSAMs) for the *Saccharomyces cerevisiae* TFs Rap1, Ndt80, Gcn4, Leu3, Abf1, and Sum1 by applying MatrixREDUCE to genome-wide TF occupancy data and, in the case of Ndt80, differential mRNA expression microarray data (Figure 2A). Experimental measurements of relative K_d's (EMSA or *lacZ* expression) for specific oligonucleotides were available for Abf1, Leu3, Ndt80, and Sum1. EMSA has long been employed to determine the



Rap1	Contact	All atom	PBM
ChIP SM	2.50×10^{-3}	2.63×10^{-6}	$< 10^{-15}$
PBM	7.02×10^{-3}	2.54×10^{-3}	
All Atom	6.71×10^{-5}		

Ndt80	All atom	Contact
Ndt80Δ	2.96×10^{-5}	4.16×10^{-5}
Contact	9.54×10^{-4}	

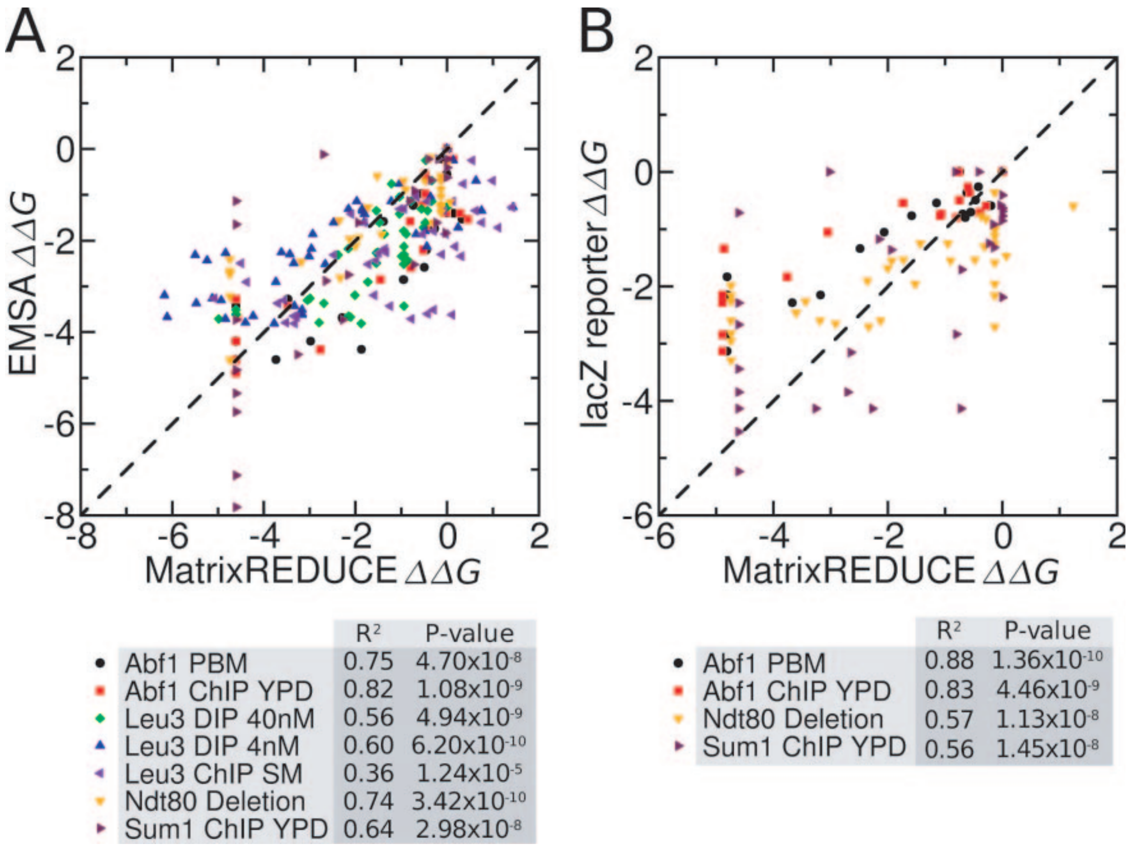
Gcn4	Contact	All atom	ChIP YPD	ChIP SM
ChIP RAPA	7.78×10^{-6}	1.00	$< 10^{-15}$	$< 10^{-15}$
ChIP SM	4.56×10^{-7}	0.60	2.66×10^{-15}	
ChIP YPD	3.04×10^{-5}	1.00		
All Atom	0.71			

Leu3	DIP 4nM	DIP 40nM
ChIP SM	4.09×10^{-4}	3.64×10^{-3}
DIP 40nM	3.55×10^{-15}	

Abf1	PBM
ChIP YPD	$< 10^{-15}$

Fig. 2. Comparison of PSAMs—affinity logos and correlations. (**A**) The PSAMs represented in the columns with blue headers were inferred by Matrix-REDUCE from ChIP-chip, PBM, DIP-chip, or mRNA differential expression microarray data. YPD (rich media), SM (sulfometuron methyl), and RAPA (rapamycin) refer to the environmental conditions to which the test sample was exposed before the ChIP-chip experiment. The DIP-chip experiments were performed with two different concentrations of Leu3, 4nM and 40nM. An *ndt80* deletion (*ndt80*Δ) versus wild-type mRNA expression experiment (mRNA) was used to obtain the Ndt80 PSAM. The PSAMs represented in the columns green headers were inferred by modeling TF-DNA interactions based on crystal structures of the TFs using two different methods, a contact-only model and an all atom model. (**B**) All PSAMs for each TF were aligned pairwise and the Pearson correlation between the ΔΔ*G* values of both PSAMs for the best alignment was calculated. The *P*-value for this correlation is a measure of similarity between the PSAMs. Again, blue labels indicate PSAMs inferred by MatrixREDUCE PSAMs and green labels indicate structurally inferred PSAMs.

DNA-binding affinities of TFs in vitro. Likewise, the *lacZ* reporter assay has long been used to measure the difference in activities of TF binding sites. We claim that a PSAM inferred by MatrixREDUCE from genome-wide TF occupancy data can be used to predict the relative binding affinities of the measured TF to any sequence. Therefore, EMSA and *lacZ* expression data provide nearly ideal data sets for validation of the MatrixREDUCE approach. For each combination of experimentally tested sequence, experimental method (EMSA or *lacZ*), and TF, we compared the experimental ΔΔ*G* with the ΔΔ*G* predicted from a PSAM for the same TF (Figure 3). In every case, the experimental ΔΔ*G* values strongly correlated with the PSAM-predicted ΔΔ*G* values, with R^2's ranging from 0.36 to 0.88. Thus, PSAMs inferred by MatrixREDUCE seem to be good models of the true relative DNA binding affinities of the corresponding TFs. Unexpectedly, all of the regressions of experimental ΔΔ*G*'s on MatrixREDUCE ΔΔ*G*'s have



	R^2	P-value
● Abf1 PBM	0.75	$4.70\text{x}10^{-8}$
■ Abf1 ChIP YPD	0.82	$1.08\text{x}10^{-9}$
◆ Leu3 DIP 40nM	0.56	$4.94\text{x}10^{-9}$
▲ Leu3 DIP 4nM	0.60	$6.20\text{x}10^{-10}$
◀ Leu3 ChIP SM	0.36	$1.24\text{x}10^{-5}$
▼ Ndt80 Deletion	0.74	$3.42\text{x}10^{-10}$
▶ Sum1 ChIP YPD	0.64	$2.98\text{x}10^{-8}$

	R^2	P-value
● Abf1 PBM	0.88	$1.36\text{x}10^{-10}$
■ Abf1 ChIP YPD	0.83	$4.46\text{x}10^{-9}$
▼ Ndt80 Deletion	0.57	$1.13\text{x}10^{-8}$
▶ Sum1 ChIP YPD	0.56	$1.45\text{x}10^{-8}$

Fig. 3. Comparison of experimentally measured ΔΔG's with MatrixREDUCE PSAM-predicted Δ*ΔG*'s. Experimental measurements of Δ*ΔG*'s were derived from EMSA (**A**) and *lacZ* reporter assays (**B**). The experimental Δ*ΔG* values are plotted along the vertical axes. Predicted Δ*ΔG*'s were calculated from the PSAM for each tested TF for the same oligonucleotide sequences that were measured in each experiment. The MatrixREDUCE-predicted Δ*ΔG* values are plotted along the horizontal axes. In this representation, the higher affinity oligonucleotides have more positive Δ*ΔG*'s. The diagonal dashed line represents experimental Δ*ΔG* equal to MatrixREDUCE Δ*ΔG*. Δ*ΔG*'s are in units of *RT*, where *R* is the gas constant and *T* is the temperature. The R^2 and *P*-values for the Pearson correlations between the experimental and predicted Δ*ΔG*'s are presented for each PSAM-experimental data pair.

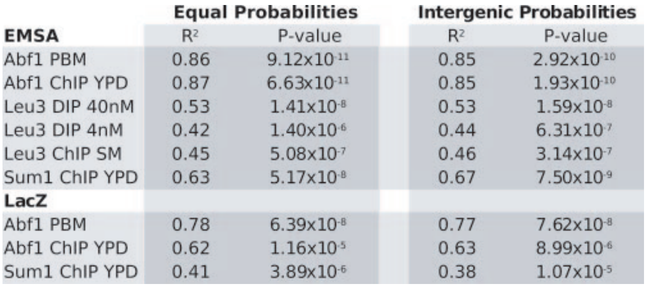
slopes less than one (range: 0.31 to 0.94). It seems that MatrixREDUCE produces a slightly larger range of predicted Δ*ΔG*'s than is realized in experiments. Nonetheless, the MatrixREDUCE PSAM-predicted Δ*ΔG*'s are close to the experimentally inferred Δ*ΔG*'s in most cases, especially among the highest affinity sequences.

4.2 PSAMs inferred by MatrixREDUCE agree well with PSAMs inferred by structural models

Both genome-wide TF occupancy data and crystal structures of protein-DNA complexes are available for Ndt80, Gcn4, and Rap1. Thus, we were able to compare MatrixREDUCE PSAMs with those based on *ab initio* structural models (see Methods; Figure 2A). The structurally inferred PSAMs for Ndt80 were obtained from its co-crystal structure bound to a high affinity GACACAAAA site, solved at 1.4 Å resolution (Lamoureux *et al*., 2002). Figure 2A shows a reasonable agreement between Δ*ΔG* predictions carried out with MatrixREDUCE and structural models. The close correspondence with the contact model, which is a function of the number of protein side chains in contact with DNA base pairs, is especially remarkable, showing that the MatrixREDUCE approach is capable of reproducing structural details of the binding interface based only on the genomic sequence and genome-wide TF occupancy data.

Gcn4 is a TF of the bZIP class. It is a homodimer with the basic region mediating sequence specific DNA binding and the leucine zipper region required for dimerization (O'Shea *et al*., 1991). For deriving the Gcn4 structural PSAMs, we used a 2.9 Å crystal structure of the TF bound to the ATGAGTCAT site (Ellenberger *et al*., 1992). The symmetry of the binding site (two reverse complement 4 bp half-sites separated by G in the middle) is a reflection of the homodimeric binding and is captured well in MatrixREDUCE predictions. While contact model and MatrixREDUCE predictions are similar, the all-atom model is less successful, probably due to the low resolution of the crystal structure, which leads to considerable uncertainty in side chain positions with respect to the neighboring DNA bases.

Finally, Rap1 binds DNA as a homodimer in a way that makes its DNA site a tandem repeat. The crystal structure of the Rap1 homodimer in complex with a telomeric DNA site has been solved to 2.25 Å resolution (Konig *et al*., 1996). Comparison of MatrixREDUCE PSAMs and structural PSAMs reveals good agreement with the all atom model. The contact model overpredicts binding specificity at the intermediate positions in the binding site (located



	Equal Probabilities		Intergenic Probabilities	
EMSA	R^2	P-value	R^2	P-value
Abf1 PBM	0.86	$9.12x10^{-11}$	0.85	$2.92x10^{-10}$
Abf1 ChIP YPD	0.87	$6.63x10^{-11}$	0.85	$1.93x10^{-10}$
Leu3 DIP 40nM	0.53	$1.41x10^{-8}$	0.53	$1.59x10^{-8}$
Leu3 DIP 4nM	0.42	$1.40x10^{-6}$	0.44	$6.31x10^{-7}$
Leu3 ChIP SM	0.45	$5.08x10^{-7}$	0.46	$3.14x10^{-7}$
Sum1 ChIP YPD	0.63	$5.17x10^{-8}$	0.67	$7.50x10^{-9}$
LacZ				
Abf1 PBM	0.78	$6.39x10^{-8}$	0.77	$7.62x10^{-8}$
Abf1 ChIP YPD	0.62	$1.16x10^{-5}$	0.63	$8.99x10^{-6}$
Sum1 ChIP YPD	0.41	$3.89x10^{-6}$	0.38	$1.07x10^{-5}$

Fig. 4. Correlations of experimentally measured $\Delta\Delta G$'s with information theory-predicted $\Delta\Delta G$'s. Experimental measurements of $\Delta\Delta G$'s were derived from EMSA and *lacZ* reporter assays. The R^2 and P-values for the Pearson correlations between the experimental and predicted $\Delta\Delta G$'s are presented for each PSSM-experimental data pair. PSSMs were derived and tested using two different background nucleotide frequencies: equal probabilities and intergenic probabilities.

between tandem repeats), likely because it assigns similar specificities to protein-DNA contacts in the loop region and in the DNA binding domains.

4.3 PSAM to PSAM correlations

Upon visual inspection of Figure 2A, the similarities are immediately apparent between affinity logos for the same factor inferred using different experimental and computational methods. However, a quantitative measure of these similarities can be obtained by aligning the PSAMs (see Methods) and calculating the correlation of their $\Delta\Delta G$ values. The P-value for this correlation between two PSAMs serves as our similarity metric (Figure 2B). Overall, the similarity between the PSAMs from MatrixREDUCE are the most significant. There is extreme similarity between the Rap1 PSAMs inferred from ChIP-chip and PBM data. The PSAMs inferred for Gcn4 for three different ChIP-chip conditions are all very similar as well. The Leu3 PSAMs inferred from the DIP-chip data are much more similar to each other than they are to the Leu3 PSAM inferred from the ChIP-chip data, but they are still both significantly similar (P-value < 0.01) to the ChIP-chip Leu3 PSAM. The significance of the correlations between MatrixREDUCE PSAMs and structurally inferred PSAMs is more variable. Both the all atom model PSAM and the contact model PSAM for Rap1 and Ndt80 have significant similarities with the respective MatrixREDUCE PSAMs (P-value < 0.01). However for Gcn4, while the contact model PSAM has strong similarities with all of the other PSAMs, the all atom PSAM has insignificant similarities with all other PSAMs.

4.4 How good is the information theory approximation?

In the original papers describing the ChIP-chip (Harbison *et al*., 2004), PBM (Mukherjee *et al*., 2004), and DIP-chip (Liu *et al*., 2005) data, the authors used BioProspector (Liu *et al*., 2001) or MDscan (Liu *et al*., 2002) to define weight matrix representations of TF binding sites. These two methods use the set of sequences that the experimenters label as ''bound'' to produce a list of potential binding sites. When interpreted through information theory, the nucleotide frequencies at each individual position in the binding sites divided by the ''background'' frequencies for the respective bases provide an estimate of a PSAM in the form of a position-specific scoring matrix (PSSM). Since we had already compiled the EMSA and *lacZ* expression data, we had the opportunity to experimentally verify the results of these PSSMs.

We gathered the BioProspector and MDscan results from the original, published analyses, transformed them into PSSMs, and used them to predict $\Delta\Delta G$'s for the EMSA and *lacZ* experimentally tested sequences. We performed this comparison using two different ''background'' nucleotide frequency models: one using equal nucleotide probabilities and one using nucleotide probabilities derived from *S. cerevisiae* intergenic sequences. The R^2 and P-values for the correlations between these predicted $\Delta\Delta G$'s and the experimental results are displayed in Figure 4. Overall, the quality of the results from the information theory PSSMs and the MatrixREDUCE PSAMs were similar. However, the results for the PSSMs are *different* depending on the choice of equal or intergenic nucleotide frequencies. While we did not test this scenario, the information theory results would also change depending on the probe intensity threshold chosen to label genes as ''bound.'' Thus, while MatrixREDUCE performs comparably with existing information theory methods, it conveniently avoids having to choose several *ad hoc* parameters required by the other methods.

5 DISCUSSION

Overall, position specific affinity matrices (PSAMs) as inferred by MatrixREDUCE from genome-wide TF occupancy data are good approximations of the real sequence-specific DNA binding affinities. Discrepancies between the computationally predicted and the experimentally inferred binding affinities may be due to either the computational or the experimental methods. EMSA has known problems with ''caging'' of the TFs by the gel while electrophoresis is proceeding (Fried and Crothers, 1981). This could lead to inferred $\Delta\Delta G$'s of smaller magnitude. Likewise, *lacZ* reporter assays are a very indirect way of measuring relative binding affinities as they require transcription, translation, and β-galactosidase reactions in order to make measurements, and noise could be introduced at each step. Structural model predictions are strongly dependent on the quality of input structures and are affected by errors in the energy function. The current MatrixREDUCE model may also give rise to systematic biases. First, it makes the approximation that nucleotides contribute independently to the free energy of TF binding (Benos *et al*., 2002). Second, it makes the assumption that the concentration of TF is much smaller than the K_d, which may not be correct for some TFs. Finally, all consecutive positions in the PSAM are currently treated as parameters to be estimated, which may lead to overfitting. We plan to address these issues in a future version of the algorithm.

Despite these current limitations, PSAMs discovered using the current implementation of MatrixREDUCE are good approximations of the relative nucleotide binding affinities of assayed TFs. Especially for microarray methods like PBM and DIP-chip, where the objective is to define nucleotide-binding specificities, MatrixREDUCE may be the most physically accurate method available to analyze the data. Even for less direct reflections of TF binding affinities like ChIP-chip or differential mRNA expression data, it will still provide good approximations of the sequence-specific binding affinities of TFs relevant to those data. Preliminary results also suggest that MatrixREDUCE performs well on data from higher eukaryotes including *D. melanogaster* and mammals. Finally, MatrixREDUCE has two key advantages over most other computational methods for defining nucleotide binding specificities: (i) it uses the information for all probes in genome-wide TF occupancy data, and (ii) it does not require a background sequence model.

ACKNOWLEDGMENTS

This work was supported by National Institutes of Health grants HG003008 (to H.J.B) and CA121852. A.V.M. is a fellow of the Leukemia and Lymphoma Society. The authors thank Jason Lieb, Neil Clarke, Martha Bulyk, Ernest Fraenkel, Xiao Liu, and Michael Berger for graciously answering questions and supplying additional data from their publications. The authors also thank Gabor Halasz, Ronald Tepper, Junbai Wang, Xiang-Jun Lu, and Lucas Ward for valuable conversations, and Xiang-Jun Lu for reimplementing the MatrixREDUCE software for distribution.

REFERENCES

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.

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.

Chu,S., DeRisi,J., Eisen,M., Mulholland,J., Botstein,D., Brown,P.O. and Herskowitz,I. (1998) The transcriptional program of sporulation in budding yeast. *Science*, **282**, 699–705.

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

Djordjevic,M. and Sengupta,A.M. (2006) Quantitative modeling and data analysis of SELEX experiments. *Phys. Biol.*, **3**, 13–28.

Ellenberger,T.E., Brandl,C.J., Struhl,K. and Harrison,S.C. (1992) The GCN4 basic region leucine zipper binds DNA as a dimer of uninterrupted alpha helices: crystal structure of the protein-DNA complex. *Cell*, **71**, 1223–1237.

Endres,R.G., Schulthess,T.C. and Wingreen,N.S. (2004) Toward an atomistic model for predicting transcription-factor binding sites. *Proteins*, **57**, 262–268.

Foat,B.C., Houshmandi,S.S., Olivas,W.M. and Bussemaker,H.J. (2005) Profiling condition-specific, genome-wide regulation of mRNA stability in yeast. *Proc. Natl. Acad. Sci. USA*, **102**, 17675–17680.

Fried,M. and Crothers,D.M. (1981) Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis. *Nucleic Acids Res.*, **9**, 6505–6525.

Gailus-Durner,V., Xie,J., Chintamaneni,C. and Vershon,A.K. (1996) Participation of the yeast activator Abf1 in meiosis-specific expression of the HOP1 gene. *Mol. Cell. Biol.*, **16**, 2777–2786.

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

Grubbs,F. (1969) Procedures for detecting outlying observations in samples. *Technometrics,* **11**, 1–21.

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., Jennings,E.G., Zeitlinger,J., Pokholok,D.K., Kellis,M., Rolfe,P.A., Takusagawa,K.T., Lander,E.S., Gifford,D.K., Fraenkel,E. and Young,R.A. (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.

Havranek,J.J., Duarte,C.M. and Baker,D. (2004) A simple physical model for the prediction and design of protein-DNA interactions. *J. Mol. Biol.,* **344**,59–70.

Issel-Tarver,L., Christie,K.R., Dolinski,K., Andrada,R., Balakrishnan,R., Ball,C.A., Binkley,G., Dong,S., Dwight,S.S., Fisk,D.G., Harris,M., Schroeder,M., Sethuraman,A., Tse,K., Weng,S., Botstein,D. and Cherry,J.M. (2002) Saccharomyces Genome Database. *Methods Enzymol.*, **350**, 329–346.

Iyer,V.R., Horak,C.E., Scafe,C.S., Botstein,D., Snyder,M. and Brown,P.O. (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature*, **409**, 533–538.

Konig,P., Giraldo,R., Chapman,L. and Rhodes,D. (1996) The crystal structure of the DNA-binding domain of yeast RAP1 in complex with telomeric DNA. *Cell,* **85**, 125–136.

Kortemme,T., Morozov,A.V. and Baker,D. (2003) An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *J. Mol. Biol.*, **326**, 1239–1259.

Lamoureux,J.S., Stuart,D., Tsang,R., Wu,C. and Glover,J.N.M. (2002) Structure of the sporulation-specific transcription factor Ndt80 bound to DNA. *EMBO J.*, **21**, 5721–5732.

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., Zeitlinger,J., Jennings,E.G., Murray,H.L., Gordon,D.B., Ren,B., Wyrick,J.J., Tagne,J.B., Volkert,T.L., Fraenkel,E., Gifford,D.K. and Young,R.A. (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, **298**, 799–804.

Lieb,J.D., Liu,X., Botstein,D. and Brown,P.O. (2001) Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nat. Genet.*, **28**, 327–334.

Liu,X.S., Brutlag,D.L. and Liu,J.S. (2002) An algorithm for finding protein-dna binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat. Biotechnol.*, **20**, 835–839.

Liu,X., Brutlag,D.L. and Liu,J.S. (2001) Bioprospector: discovering conserved dna motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput.* 127–138.

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.

Liu,X., Noll,D.M., Lieb,J.D. and Clarke,N.D. (2005) DIP-chip: rapid and accurate determination of DNA-binding specificity. *Genome Res.*, **15**, 421–427.

Lu,X.J. and Olson,W.K. (2003) 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.*, **31**, 5108–5121.

Morozov,A.V., Havranek,J.J., Baker,D. and Siggia,E.D. (2005) Protein-DNA binding specificity predictions with structural models. *Nucleic Acids Res.,* **33**, 5781–5798.

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.

Olson,W.K., Gorin,A.A., Lu,X.J., Hock,L.M. and Zhurkin,V.B. (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl. Acad. Sci. USA*, **95**, 11163–11168.

Onufriev,A., Bashford,D. and Case,D.A. (2004) Exploring protein native states and large-scale conformational changes with a modified Generalized Born model. *Proteins*, **55**, 383–394.

O'Shea,E.K., Klemm,J.D., Kim,P.S. and Alber,T. (1991) X-ray structure of the GCN4 leucine zipper, a two-stranded, parallel coiled coil. *Science*, **254**, 539–544.

Paillard,G. and Lavery,R. (2004) Analyzing protein-DNA recognition mechanisms. *Structure*, **12**, 113– 122.

Pierce,M., Benjamin,K.R., Montano,S.P., Georgiadis,M.M., Winter,E. and Vershon,A.K. (2003) Sum1 and Ndt80 proteins compete for binding to middle sporulation element sequences that control meiotic gene expression. *Mol. Cell. Biol.*, **23**, 4814–4825.

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

Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.

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

Stormo,G.D. and Fields,D.S. (1998) Specificity, free energy and information content in protein-DNA interactions. *Trends Biochem. Sci.*, **23**, 109–113.

Stormo,G.D., Schneider,T.D. and Gold,L. (1986) Quantitative analysis of the relationship between nucleotide sequence and functional activity. *Nucleic Acids Res.*, **14**, 6661–6679.

Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.