Published online October 24, 2005

Nucleic Acids Research, 2005, Vol. 33, No. 18 **5781–5798**
doi:10.1093/nar/gki875

Protein–DNA binding specificity predictions with structural models

Alexandre V. Morozov*, James J. Havranek[1], David Baker[1] and Eric D. Siggia

Center for Studies in Physics and Biology, The Rockefeller University, 1230 York Avenue, New York, NY 10021, USA and [1]Department of Biochemistry, University of Washington, Box 357350, Seattle, WA 98195-7350, USA

Received July 13, 2005; Revised and Accepted September 13, 2005

ABSTRACT

Protein–DNA interactions play a central role in transcriptional regulation and other biological processes. Investigating the mechanism of binding affinity and specificity in protein–DNA complexes is thus an important goal. Here we develop a simple physical energy function, which uses electrostatics, solvation, hydrogen bonds and atom-packing terms to model direct readout and sequence-specific DNA conformational energy to model indirect readout of DNA sequence by the bound protein. The predictive capability of the model is tested against another model based only on the knowledge of the consensus sequence and the number of contacts between amino acids and DNA bases. Both models are used to carry out predictions of protein–DNA binding affinities which are then compared with experimental measurements. The nearly additive nature of protein–DNA interaction energies in our model allows us to construct position-specific weight matrices by computing base pair probabilities independently for each position in the binding site. Our approach is less data intensive than knowledge-based models of protein–DNA interactions, and is not limited to any specific family of transcription factors. However, native structures of protein–DNA complexes or their close homologs are required as input to the model. Use of homology modeling can significantly increase the extent of our approach, making it a useful tool for studying regulatory pathways in many organisms and cell types.

INTRODUCTION

Gene regulation is mediated in part by protein transcription factors (TFs) binding to *cis*-regulatory regions of the genome. Accurate genomewide characterization of TF binding sites is thus a necessary prerequisite to deciphering complex gene expression patterns. Probabilistic models of TF binding profiles, often called position-specific weight matrices (PWMs), are typically used as input to such predictions (1–3). With the weight matrix representation of TF binding sites, the probability $P(S\,|\,p)$ that sequence S is a binding site for the TF represented by p is given by

$$P(S\,|\,p) = \prod_{i=1}^{L} p^{i}_{s_i}, \qquad \mathbf{1}$$

where L is the length of the binding site in base pairs, s_i is the base at position i and p^i_α is the probability of base α (A, C, G or T) at position i in the weight matrix, subject to normalization: $\sum_{\alpha=1}^{4} p^i_\alpha = 1$ (any i). PWMs can be constructed from an alignment of binding sites obtained by footprinting methods, gel-shift analysis and reporter constructs. In addition, TF binding sites can be characterized using SELEX, DNA microarray, genomewide location (ChIP-chip) and other *in vitro* techniques (1). However, TF binding specificity data of this type cannot be used to rationalize the mechanism by which proteins interact with DNA. Moreover, for many TFs there is little or no experimental binding site information.

In principle, one should be able to predict the TF binding site profile from a structure of the protein–DNA complex or its close homolog. At first it was anticipated that structural studies would reveal a universal protein–DNA recognition code, which could be used for predicting TF binding sites based on amino acid identities at the protein–DNA interface (4,5). It became apparent, however, when more protein–DNA structures were solved and classified that despite some predominantly occurring interactions, such as Lys-G, the energetics of amino acid–base contacts depends on their structural context and, in particular, on the structural family of the DNA-binding protein (6–10). Many amino acids were observed to form favorable contacts with different bases, making it necessary to generalize a deterministic recognition code to a probabilistic binding profile based, for example, on maximizing the likelihood of observed protein–DNA contacts (11–13).

*To whom correspondence should be addressed. Tel: +1 212 327 8139; Fax: +1 212 327 8544; Email: morozov@edsb.rockefeller.edu

Probabilistic recognition codes are more accurate when developed for a specific structural family, thereby implicitly taking protein–DNA structural context into account. Indeed, binding site profiles based on the classification of TFs into families were found to be useful in bioinformatics pattern detection algorithms (14). However, data availability has so far limited knowledge-based PWM predictions to the C_2H_2 zinc finger family (15,16).

An alternative approach to specificity and binding affinity predictions is based on all-atom modeling of protein–DNA complexes (17–19). Starting from a known structure of the protein bound to its consensus DNA sequence, an ensemble of models is created by threading novel DNA sequences onto the binding site. Protein–DNA binding energies, ΔG, are then evaluated for each member of the structural ensemble. ΔG predictions can be either used to directly infer high-affinity binding sites in genomic sequence or converted into PWM probabilities using the Boltzmann formula. In the latter case, it is only necessary to compute ΔG values for all one-point mutations from the consensus-binding site. The main limiting factor of the structural approach to TF binding site specificity predictions is the availability of experimentally determined structures of protein–DNA complexes. The range of applicability of structural methods will significantly increase if the DNA-binding proteins can be modeled by homology. Homology modeling involves threading a protein amino acid sequence onto a suitable structural template chosen on the basis of its sequence similarity to the protein of interest. The threading procedure creates a new protein–DNA binding interface, for which ΔG and PWM probability calculations are then carried out as in native structures.

Here we present a computational model for predicting protein–DNA binding affinities and specificities. The model can be applied to a wide variety of DNA-binding proteins for which there is either a native protein–DNA structure or a sufficiently close homolog. The model is based on a simple free energy function, which consists of the protein–DNA interaction energy and the DNA conformational energy. The protein–DNA interaction energy is used to describe direct readout of the DNA sequence by the protein, whereas the DNA conformational energy takes into account distortion of B-DNA shape caused by protein binding. We carried out a series of tests of our PWM and binding energy predictions. First, we checked the ability of the model to reproduce experimental binding free energy measurements. We also assessed the accuracy of the pairwise additivity approximation in our analysis. Second, we checked the ability of our algorithm to discriminate experimentally known TF binding sites from random ensembles of sequences. Third, we carried out PWM predictions for a number of TFs and compared them with experimental PWMs. For all these predictions native protein–DNA complexes were used as structural templates. Finally, the extent of applicability of homology modeling to protein–DNA binding affinity predictions was explored with several representative PWM calculations. The relative accuracy and computational efficiency of our approach allowed us to carry out numerous predictions of TF binding affinities and specificities, facilitating future experimental and computational studies of transcriptional regulation in different organisms and biological systems.

METHODS

Binding stability and weight matrix predictions

Free energy model of protein–DNA interactions. We extended Rosetta protein–nucleic acid model developed in Ref. (20) by adding sequence-specific DNA conformational energy. The free energy function employed in protein–DNA binding affinity calculations consists of the protein–DNA interaction component describing intermolecular readout of the DNA sequence by the protein, and of the DNA deformation component describing intramolecular readout of the binding site sequence:

$$G_{\text{prot-dna}} = G_{\text{pd}} + G_{\text{d}}. \qquad \mathbf{2}$$

Protein–DNA interactions are modeled as a linear combination of the Lennard–Jones potential (which switches to a linear repulsive potential at short distances) (20), the orientation-dependent hydrogen bonding potential describing amino acid side chain–amino acid side chain and amino acid side chain–DNA base hydrogen bonds (21), the Generalized Born electrostatics and solvation model (22), and the implicit solvation model developed by Lazaridis and Karplus (23):

$$\begin{aligned} G_{\text{pd}} = {} & w_{\text{LJ}_{\text{rep}}} E_{\text{LJ}_{\text{rep}}} + w_{\text{LJ}_{\text{attr}}} E_{\text{LJ}_{\text{attr}}} + w_{\text{hb}} E_{\text{hb}} \\ & + w_{\text{el}} G_{\text{el}} + w_{\text{LK}_{\text{sol}}} G_{\text{LK}_{\text{sol}}}, \end{aligned} \qquad \mathbf{3}$$

where each energy is a sum over all protein–DNA and protein–protein atomic pairs, and $\{w\}$ is a set of fitting weights. Parameterization of the Lennard–Jones potential, the hydrogen bonding potential and the Lazaridis and Karplus solvation model is carried out as described in Ref. (20), whereas the Generalized Born model with AMBER *parm99* atomic partial charges (24) is adopted from Ref. (22). Other free energy components, including the knowledge-based potential derived from residue pair statistics and the solvent accessible surface area term, were tested but omitted from the final model since they did not significantly improve the predictions of protein–DNA binding energies. The model requires an atomic representation of protein–DNA complexes including all hydrogen and heavy atoms. Hydrogen atoms and heavy atoms missing from experimental structures were built using standard CHARMM27 bond lengths and bond angles, with all rotatable bonds connecting hydrogens to heavy atoms optimized using Monte Carlo search with the free energy function described above.

The DNA sequence-dependent conformational energy model is based on the effective harmonic potential developed in Ref. (25). The DNA conformational energy is computed as a sum over all base pair and base step energies:

$$G_{\text{d}} = w_{\text{dna-bp}} \sum_{\text{bp}} E^{\alpha\beta}_{\text{dna-bp}} + w_{\text{dna-bs}} \sum_{\text{bs}} E^{\alpha\beta}_{\text{dna-bs}}, \qquad \mathbf{4}$$

where the first sum is over all base pairs in the double helical DNA (α,β denote bases in a base pair), the second sum is over all consecutively stacked base pair steps (α,β denote base pairs in a base step), and $w_{\text{dna-bp}}$, $w_{\text{dna-bs}}$ are fitting weights. Base pairs and base steps are counted once in the $5'$–$3'$ direction. The first term in Equation 4 is necessary to enforce reasonable base pairing in the process of DNA minimization, which occurs in the space of DNA torsional angles and not in the space of effective degrees of freedom defined below. Both $E_{\text{dna–bs}}$ and $E_{\text{dna–bp}}$ are approximated with harmonic functions (25):

$$E^{\alpha\beta}_{\text{dna–bs/dna–bp}} = \frac{1}{2}\sum_{i=1}^{6}\sum_{j=1}^{6} f^{\alpha\beta}_{ij}\delta\theta^{\alpha}_{i}\delta\theta^{\beta}_{j}, \tag{5}$$

where the sum runs over effective degrees of freedom $\{\theta\}$ (Twist, Tilt, Roll, Shift, Slide and Rise for base steps; Opening, Buckle, Propeller, Shear, Stretch and Stagger for base pairs) (26). All effective geometric parameters are calculated as described in Ref. (27). Every set of six geometric parameters completely describes the mutual orientation of two bases or base pairs represented as rigid bodies. $\delta\theta^{\alpha}_{i}$ is a deviation of θ^{α}_{i} from its value averaged over a set of experimental observations (DNA structures from a non-homologous set of protein–DNA complexes): $\delta\theta^{\alpha}_{i} = \theta^{\alpha}_{i} - \langle\theta^{\alpha}_{i}\rangle$. The force constants $f^{\alpha\beta}_{ij}$ are evaluated by inverting the covariance matrix of $\delta\theta_{i}{}^{\alpha}$ averaged over the same set of DNA structures: $f^{\alpha\beta}_{ij} = \langle\delta\theta^{\alpha}_{i}\delta\theta^{\beta}_{j}\rangle^{-1}$.

All θ distributions are self-consistently trimmed by removing all data points for which at least one of the geometric parameters is more than three standard deviations away from the average, followed by updating all averages and standard deviations. This procedure is repeated until convergence, which usually requires 2–3 iterations and removes just a few percent of the data points (28). A non-homologous, manually curated set of 101 structures used in Ref. (20) was employed to derive the force constants for the DNA conformational potential.

Given free energies of the protein–DNA complex and its unbound partners, the binding free energy is computed as follows:

$$\Delta G = G_{\text{prot–dna}} - G_{\text{prot}} - G_{\text{dna}}. \tag{6}$$

Mutations of one or several base pairs in the binding site result in changes of protein–DNA binding free energies:

$$\Delta\Delta G = \Delta G^{\text{mut}} - \Delta G^{\text{wt}}, \tag{7}$$

where $\Delta G^{\text{mut}}(\Delta G^{\text{wt}})$ refers to the mutated and wild-type DNA sequence, respectively.

Weight matrix predictions based on binding free energies

Computation of weight matrix probabilities $p^{i}_{s_i}$ requires an assumption of the pairwise additivity of base pair energies:

$$\Delta G = \sum_{i=1}^{L}\Delta G^{i}_{s_i}, \tag{8}$$

where L is the length of the binding site in base pairs, s_i is the base at position i and ΔG^{i}_{α} is the binding energy of base α (A, C, G or T) at position i, which is assumed to be independent of all other bases. In our model, all energy terms except for the base step energies are explicitly pairwise independent. However, even the energies that are pairwise independent by construction (i.e. computed as sums over individual atom–atom interactions) can become non-additive if conformational rearrangement is allowed at the protein–DNA binding interface. Thus, it is important to check explicitly if the energies in our model are approximately pairwise additive.

With the pairwise additivity assumption, a set of $4L$ predicted energies ΔG^{i}_{α} can be converted into weight matrix probabilities using the Boltzmann formula (29):

$$p^{i}_{\alpha} = \frac{\exp\left(-\beta\Delta G^{i}_{\alpha}\right)}{\sum_{\gamma=1}^{4}\exp\left(-\beta\Delta G^{i}_{\gamma}\right)}, \tag{9}$$

where $\beta = 1/RT$ is the inverse temperature used as a fitting parameter [β was changed in steps of 0.25 (kcal/mol)$^{-1}$ in all fits].

Experimental datasets

Binding free energies. Table 1 collects structural data for protein–DNA complexes with binding free energy measurements available from the ProNIT database (http://dna01.bse.kyutech.ac.jp/jouhou/pronit/pronit.html) and from the literature. Each dataset in Table 1 consists of a structure of the protein–DNA complex and a series of binding free energy measurements ΔG for wild-type and mutant DNA sequences. In several cases, association or dissociation constants reported by the authors were converted into binding free energies using

$$\Delta\Delta G = \mp RT\ln\left(\frac{K^{\text{mut}}_{a,d}}{K^{\text{wt}}_{a,d}}\right), \tag{10}$$

with $RT = 0.59$ kcal/mol. In the case of the MAT **a**1/α2 TF, *in vivo* repression levels of the heterologous reporter promoter construct were used as a measure of binding affinity:

$$\Delta\Delta G = -RT\ln\left(\frac{R^{\text{mut}}}{R^{\text{wt}}}\right), \tag{11}$$

where R is the repression ratio in the presence and absence of the wild-type or mutant **a**1/α2 binding site (30). Eleven

Table 1. Experimental binding affinity dataset

Name	PDB code	Method (Res., Å)	ΔΔG data points	Organism	Reference
Zif268	1aay	X-ray (1.6)	15 (8)	*Mus musculus*	(42)
Zif268	1aay	X-ray (1.6)	6 (6)	*M.musculus*	(67)
Zif268 D20A	1jk1	X-ray (1.9)	6 (6)	*M.musculus*	(67)
Tus	1ecr	X-ray (2.7)	20 (20)	*Escherichia coli*	(68)
LacR	1efa	X-ray (2.6)	5 (5)	*E.coli*	(69)
λR	1lmb	X-ray (1.8)	51 (51)	λ-Phage	(51)
TrpR	1tro	X-ray (1.9)	9 (9)	*E.coli*	(70)
ER	1hcq	X-ray (2.4)	7 (7)	*Homo sapiens*	(71)
CroR	6cro	X-ray (3.0)	56 (56)	λ-Phage	(52)
EcoRI	1ckq	X-ray (1.85)	13 (13)	*E.coli*	(41)
Crp	1run	X-ray (2.7)	15 (15)	*E.coli*	(72)
BamHI	1bhm	X-ray (2.2)	23 (0)	*Bacillus amyloliquefaciens*	(49)
PU.1 ETS	1pue	X-ray (2.1)	25 (0)	*M.musculus*	(50)
Ndt80	1mnn	X-ray (1.4)	26 (0)	*Saccharomyces cerevisiae*	(34)
MAT **a**1/α2	1yrn	X-ray (2.5)	54 (0)	*S.cerevisiae*	(30)
c-Myb	1mse	NMR (NA)	27 (0)	*M.musculus*	(47)
AtERF1	1gcc	NMR (NA)	21 (0)	*Arabidopsis thaliana*	(45)

For protein–DNA structures solved by X-ray crystallography, resolution (Å) is shown in parentheses. The total number of ΔΔG measurements are shown for each dataset, with the number of data points used in free energy function parameterization listed in parentheses.

Table 2. Experimental binding site and weight matrix dataset

Name	PDB code	Method (Res., Å)	N_{seq}	Organism	Reference
λR	1lmb	X-ray (1.8)	–[a]	λ-Phage	(51)
CroR	6cro	X-ray (3.0)	–[a]	λ-Phage	(52)
AtERF1	1gcc	NMR (NA)	–[a]	*A.thaliana*	(45)
c-Myb	1mse	NMR (NA)	–[a]	*M.musculus*	(47)
Zif268	1aay	X-ray (1.6)	6[b]	*M.musculus*	(31)
Ndt80	1mnn	X-ray (1.4)	8[b]	*S.cerevisiae*	(34)
Gcn4p	1ysa	X-ray (2.9)	9[c]	*S.cerevisiae*	(35,36)
MAT **a**1/α2	1yrn	X-ray (2.5)	19[c]	*S.cerevisiae*	(30)
EcR/Usp	1r0o	X-ray (2.24)	33[c]	*Drosophila melanogaster*	(57)
Ttk	2drp	X-ray (2.8)	16[c]	*D.melanogaster*	–
Prd(homeo)	1fjl	X-ray (2.0)	15[c]	*D.melanogaster*	(37)
Ubx/Exd	1b8i	X-ray (2.4)	4[b]	*D.melanogaster*	–
Trl	1yui	NMR (NA)	5[c]	*D.melanogaster*	–
MetJ	1mj2	X-ray (2.4)	16[c]	*E.coli*	(32)
TrpR	1tro	X-ray (1.9)	15[c]	*E.coli*	(32,33)
PhoB	1gxp	X-ray (2.5)	16[c]	*E.coli*	(32)
Ihf	1ihf	X-ray (2.5)	27[c]	*E.coli*	(32)
DnaA	1j1v	X-ray (2.1)	9[c]	*E.coli*	(32)
PurR	2puc	X-ray (2.7)	23[c]	*E.coli*	(32)
Crp	1run	X-ray (2.7)	50[c]	*E.coli*	(32)

For protein–DNA structures solved by X-ray crystallography, resolution (Å) is shown in parentheses. N_{seq} is the total number of aligned binding site sequences (including the DNA sequence from the protein–DNA complex).

[a] PWM is obtained from $\Delta\Delta G$ data for all one-point mutations of the binding site.

[b] PWM is obtained separately from the binding sites listed in the table, by SELEX experiments or independently published surveys of genomic sites.

[c] Alignment of binding sites listed in the N_{seq} column is used to create an experimental PWM (binding site from the protein–DNA structure is not included into PWM; pseudocounts are set to 0.0).

datasets in the top part of Table 1 (total of 196 data points) are used to train the weights in the free energy function.

Binding sites and weight matrices. Table 2 collects protein–DNA structures for which binding site data and experimental weight matrix data are available. The Zif268 weight matrix was taken from a selection experiment (31). The experiment generated the G-C-G-T/g-G/A-G-G-C/a/t-G-G/T consensus sequence, which we used to create five binding site variants (in addition to the consensus sequence from the Zif268 structure). Binding sites and weight matrices for seven *Escherichia coli* TFs were obtained from the DPInteract database (32). For TrpR, 4 sites from DPInteract were augmented with 10 additional sites from RegulonDB (33). For Ndt80 in *Saccharomyces cerevisiae* we used the weight matrix from Ref. (34); Ndt80 binding sites were collected by searching promoters of the genes known to be regulated by Ndt80 with the YGNCACAAAA consensus sequence. A collection of 17 naturally occurring MAT **a**1/α2 binding sites plus the synthetic consensus sequence were obtained from Ref. (30). Eight Gcn4p sites were assembled from Ref. (35) and the TRANSFAC database (36). Finally, binding sites and weight matrices for the Prd homeodomain homodimer were taken from Ref. (37), whereas binding data for the rest of *Drosophila melanogaster* TFs were collected by E. D. Siggia and E. Emberly (unpublished data).

Prediction testing

Protein–DNA interaction model based on the number of interface atomic contacts. In order to test our $\Delta\Delta G$ and PWM predictions we developed a simple null model that exploits the structure of the protein–DNA complex but does not require any detailed predictions of protein–DNA energetics. This so-called 'contact' model constructs a weight matrix from the consensus DNA sequence and the number of atomic contacts N between protein side chains and DNA base pairs. In particular, we assume that the three non-consensus bases occur with equal probabilities, whereas the consensus base is favored over any non-consensus base by $(N/N_{\max})$ if $N \leqslant N_{\max}$. If $N > N_{\max}$, the consensus base becomes absolutely conserved:

$$p^i_{s_i}(N) = \begin{cases} \frac{1}{4}(1 - N/N_{\max}) \text{ if } N \leqslant N_{\max},\ 0 \text{ if } N > N_{\max}(i \neq wt); \\ \frac{1}{4}(1 + 3N/N_{\max}) \text{ if} N \leqslant N_{\max},\ 1 \text{ if } N > N_{\max}(i = wt) \end{cases} \qquad \mathbf{12}$$

Here $p^i_{s_i}(N)$ is the probability of the base pair type $s_i = \{$A, C, G, T$\}$ in the PWM column i, wt denotes a consensus base pair found in the protein–DNA complex at position i, N is the number of protein–DNA base atomic contacts summed over the base pair i (protein and DNA atoms are defined to be in contact if they are separated by ⩽4.5 Å; hydrogen atoms are excluded from the counts), and $N_{\max}$ is the number of contacts above which the native base pair is always conserved. $N_{\max}$ is treated as a free parameter in the probability model.

Probabilities defined by Equation 12 are converted into energies using:

$$\Delta G^i_{s_i}(N) = \begin{cases} E_{\max}\left[\log p^i_{s_i}(N) - \log p^i_{s_i}(0)\right] \Big/ \\ \left[\log p^i_{s_i}(N_{\max} - 1) - \log p^i_{s_i}(0)\right] \text{ if} N < N_{\max}; \\ E_{\max} \quad \text{if} N \geqslant N_{\max} \end{cases} \qquad \mathbf{13}$$

The $\log p^i_{s_i}(0) = \log(0.25)$ offset ensures that mutations from the consensus sequence are not penalized in the absence of protein–DNA contacts. The maximum energy penalty is capped at $E_{\max}$, which together with $N_{\max}$ constitute the free parameters of the model. In all contact model fits $N_{\max}$ was changed in steps of 5 kcal/mol and $E_{\max}$ was changed in steps of 1.0 kcal/mol.

Significance test of PWM predictions. Statistical significance of PWM predictions is estimated using the ψ-test, which is a generalization of the well-known χ^2-test (38):

$$\psi(p, q) = \frac{1}{L}\left[\sum_{j=1}^{L} \sum_{i=\{\text{A, C, G, T}\}} q^j_i \ln \frac{q^j_i}{p^j_i}\right], \qquad \mathbf{14}$$

where $\{p^j\}$ are predicted probabilities, $\{q^j\}$ are experimental frequencies and L is the length of the binding site in base pairs. Both p and q distributions are smoothed by adding 0.05 to all PWM entries and re-normalizing. The quality of PWM predictions is further estimated by comparing $\psi(p, q)$ with the average value of ψ computed for an ensemble of 10 000 alignments of random weight matrices p^j_{random} to the experimental weight matrix frequencies $\{q^j\}$ $[\langle\psi(p_{\text{random}}, q)\rangle]$. Each column in the random weight matrix is obtained by uniformly sampling four numbers in the (0,1) interval and enforcing normalization afterwards. The difference between $\langle\psi\ (p_{\mathrm{random}},q)\rangle$ and $\psi\ (p,q)$ can be viewed as a measure of success of our predictions.

RESULTS AND DISCUSSION

Binding free energy predictions

All-atom free energy models. DNA-binding proteins employ two complementary mechanisms of binding site recognition. The intermolecular readout mechanism is based on direct interactions of protein side chains with DNA bases, whereas the intramolecular readout mechanism involves sequence-specific deformation of the DNA site by the bound protein. We developed a free energy function that takes both these mechanisms into account. Interactions of protein side chains with DNA are modeled using an all-atom representation of both protein and DNA, including all the hydrogen atoms. The protein–DNA interaction energy is a weighted sum of terms describing shape complementarity and packing at the interface, polar interactions (electrostatics and hydrogen bonds), van der Waals forces and solvation energies (Equation 3). The DNA conformational energy is calculated using a reduced geometric representation in which DNA bases and base pairs are represented by rigid bodies and their mutual orientation serves as a measure of deviation from the B-form DNA. The DNA conformational energy is a weighted sum of two terms describing base pairing and stacking of consecutive base pairs. Using this free energy function, we developed the following approach to predicting protein–DNA binding affinities. First, a suitable protein–DNA complex is identified as a structural template for computational modeling. Second, each novel DNA sequence (i.e. from a set for which experimental-binding affinity measurements are available) is threaded onto the DNA phosphate backbone with fixed DNA torsional angles. The result of this procedure is a set of initial structural models with novel DNA sequences. Third, binding free energies, ΔG, for each member of the set are computed in either of two different ways.

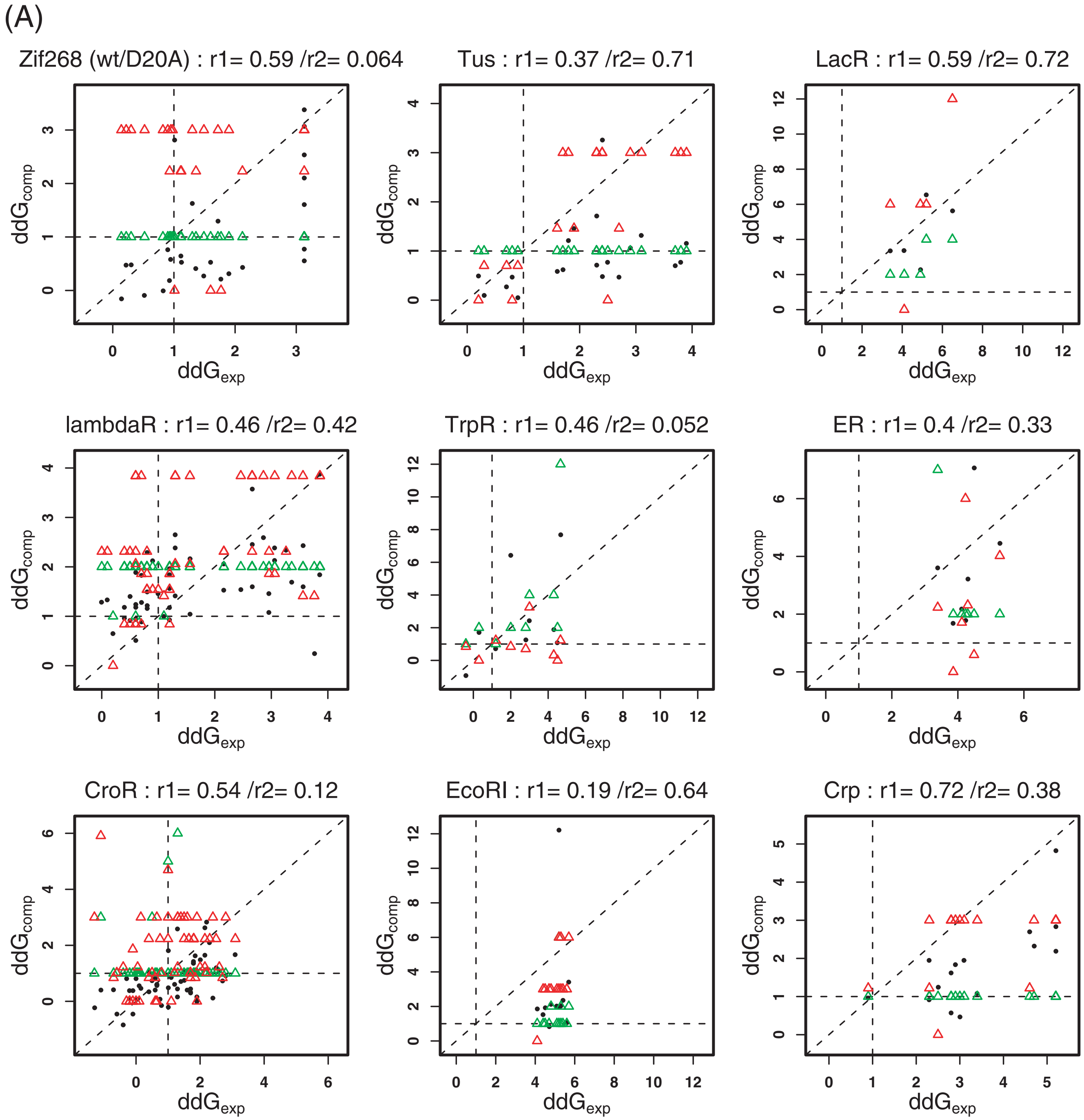
One approach, which we shall call the static model, does not allow any side chain or DNA conformational rearrangements in the protein–DNA complex. The free energy is computed once for each initial model, and the difference in binding affinity between mutant and wild-type DNA sequences is calculated as $\Delta\Delta G = G^{\mathrm{mut}}_{\mathrm{prot-dna}} - G^{\mathrm{wt}}_{\mathrm{prot-dna}}$, where $G_{\mathrm{prot\text{–}dna}}$ is the free energy of the protein–DNA complex. The relative weights of the free energy terms in Equation 2 are found by the least-squares fit to a set of experimentally observed protein–DNA binding affinities. The experimental dataset used in the fitting consists of 11 series of ΔG measurements with a total of 196 data points (Table 1). For each series of measurements there is a crystal structure of the protein–DNA complex with $\leqslant 3.0$ Å resolution used as a template for base pair threading and binding affinity predictions. The set of weights obtained in this way is cross validated by removing parts of the dataset and refitting the weights. Very similar weights are obtained in each case (data not shown). The ratio of protein–DNA to DNA conformational energies obtained through the least squares fit to experimental $\Delta\Delta G$ data is necessarily averaged over protein families included in the fit. Although there are not enough data to carry out separate fits for each protein family, restricting the fit to proteins known to significantly bend and twist DNA results in a larger contribution of the DNA conformational energies.

In the other approach, called the dynamic model, we minimize the total free energy of the protein–DNA complex starting from the initial model. The protein backbone stays fixed during minimization, whereas the torsional angles of DNA and interface side chains are allowed to relax (the protein–DNA interface is defined based on amino acid-dependent distance cutoffs). The conformational search used in $G_{\mathrm{prot\text{–}dna}}$ minimization consists of 10 two-step iterations. (i) Simulated annealing of amino acid side chains at the protein–DNA interface with side chains represented as discrete backbone-dependent rotamers (39) on a fixed protein backbone, and frozen DNA conformation. (ii) Continuous minimization of amino acid side chains at the protein–DNA interface together with simultaneous conformational relaxation of DNA. Amino acid side chains are no longer represented by rotamers at this step.

Experimental binding affinity data available to us are insufficient to reliably fit the weights by iterations to self-consistency when conformational rearrangement is allowed. Instead, we obtain the weights for components of the protein–DNA free energy function by maximizing the recovery of native amino acid side chains at all interface positions in a non-homologous set of protein–DNA complexes. In other words, we adopt a strategy used in protein sequence design in which rotamer conformations for all amino acids are substituted at all interface positions, and the probability of native amino acids is maximized by varying the weights (20,40). Similar to the static model, the ratio of the protein–DNA energy to the DNA conformational energy is expected to be protein family dependent and was estimated on average by requiring that the typical fluctuations from the equilibrium shape observed in the database of protein–DNA complexes be on the order of RT: $3\bar{f}_{ii}\langle\bar{\theta}_i^{\,2}\rangle \sim RT$ (overbar denotes an average over all base step/base pair types; we neglect off-diagonal coupling of the effective degrees of freedom). The estimated ratio of the protein–DNA to intra-DNA energies is consistent with the non-iterative least squares fit to experimental $\Delta\Delta G$ data from Table 1. In the dynamic model, $\Delta\Delta G$ is calculated as $\Delta\Delta G = (G^{\mathrm{mut}}_{\mathrm{prot-dna}} - G^{\mathrm{mut}}_{\mathrm{dna}}) - (G^{\mathrm{wt}}_{\mathrm{prot-dna}} - G^{\mathrm{wt}}_{\mathrm{dna}})$, where $G^{\mathrm{mut/wt}}_{\mathrm{prot-dna}}$ is the minimized free energy of the protein–DNA complex, and $G^{\mathrm{mut/wt}}_{\mathrm{dna}}$ is the reference free energy of the unbound DNA ($G^{\mathrm{mut}}_{\mathrm{prot}} = G^{\mathrm{wt}}_{\mathrm{prot}}$ since the protein sequence is fixed). $G^{\mathrm{mut/wt}}_{\mathrm{dna}}$ is computed by continuous minimization of the DNA conformation in the absence of the protein. Including separately minimized unbound DNA rather than ideal B-DNA as a reference state was found to be beneficial in most cases where DNA was not significantly distorted from its equilibrium shape. However, in several more extreme cases, such as BamHI endonuclease and the PU.1 ETS domain, local DNA minimization in the absence of bound protein was found to be insufficient for relaxing DNA conformation and was omitted from the model. Better conformational sampling might be provided by simulated annealing of DNA degrees of freedom.

Measures of prediction success and the contact model. We use three alternative measures to assess the quality of ΔG predictions: a linear correlation coefficient r, an average unsigned error ϵ between predicted and experimental binding free energies, and a fraction of correct predictions F. Although the first two measures are computed using standard formulae, the fraction of correct predictions is based on a binary function: a prediction is considered to be correct if both $\Delta\Delta G_{\mathrm{comp}}$ and $\Delta\Delta G_{\mathrm{exp}}$ are <1.0 kcal/mol, or >1.0 kcal/mol, or else separated by <0.3 kcal/mol. The threshold value of 1.0 kcal/mol corresponds to a ~5-fold reduction in binding affinity at room temperature. ΔG predictions are labeled as correct if they are successfully classified to be favorable or destabilizing, even if the absolute magnitudes of binding energies are not perfectly reproduced.

We compare calculations of binding energies described above with predictions based on a simple contact model. Instead of modeling detailed energetics of protein–DNA complexes, the contact model uses only the consensus sequence and the number of protein–DNA base atomic contacts at the binding interface (see Methods). The energy penalty for each mutation from the consensus sequence is given by Equation 13; it is a function of N_{max} (the minimum number of contacts at which a given consensus base is assumed to be absolutely conserved, cf. Equation 12) and of E_{max} (the maximum energy penalty for a mutation of the consensus base). N_{max} and E_{max} are adjusted to simultaneously maximize the fraction of correct predictions F and minimize the average error ϵ over the $\Delta\Delta G$ dataset. The latter requirement is necessary since many $N_{\mathrm{max}}/E_{\mathrm{max}}$ pairs result in similar



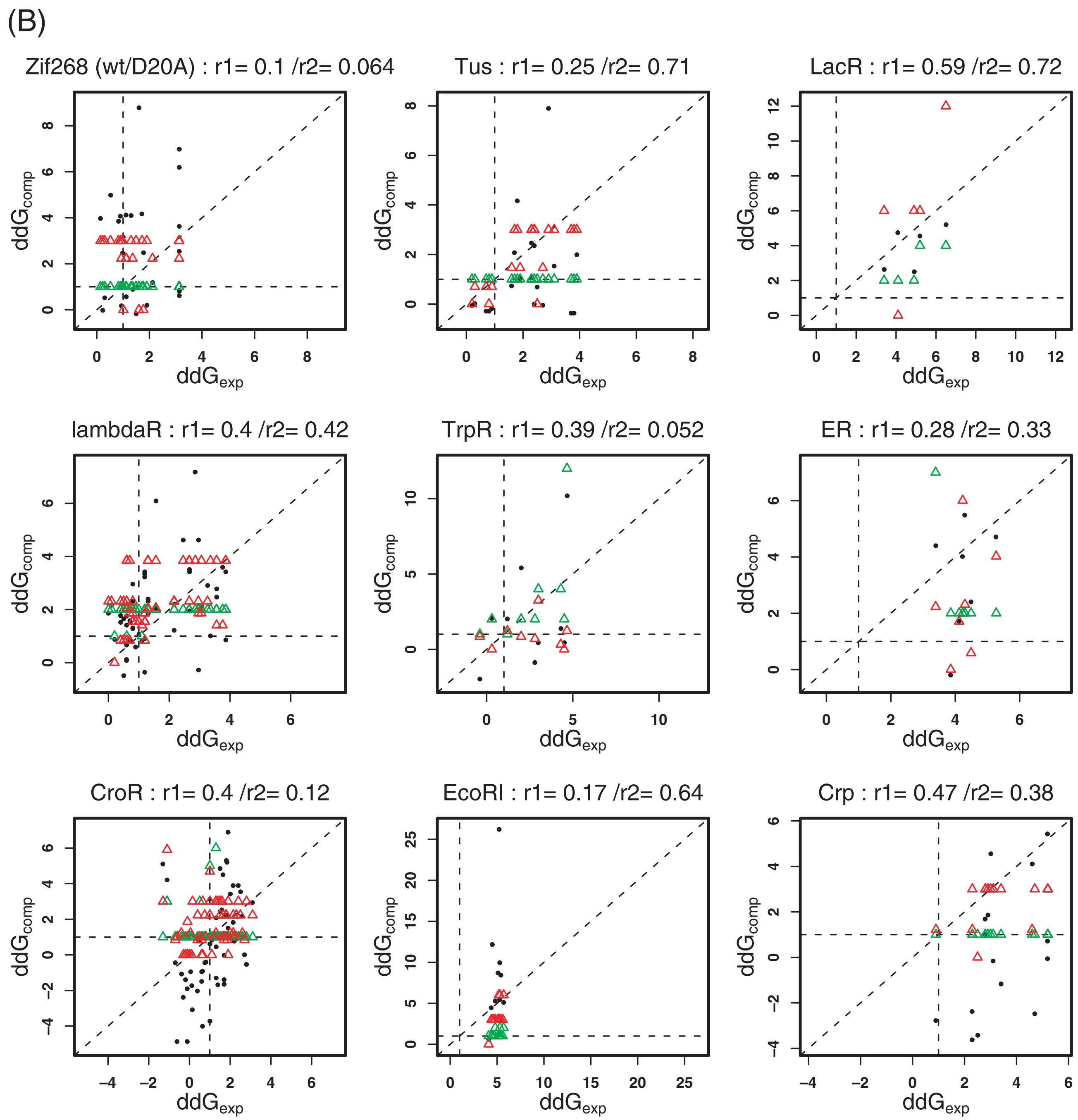


Figure 1. $\Delta\Delta G$ predictions (ddG_{comp}) versus experimental measurements (ddG_{exp}). (**A**) Static model binding energy predictions for the set of experimental measurements used in fitting static model weights. (**B**) Dynamic model binding energy predictions for the same set of experimental measurements. Closed circles, static/dynamic model; red triangles, contact model; green triangles, number of mutations from the consensus sequence. r1-Linear correlation coefficient for the static/dynamic model and r2-linear correlation coefficient for the contact model. Three Zif268 datasets from Table 1 [two for Zif268 wild-type (42,67) and one for Zif268 D20A mutant (67)] are combined into one panel.

values of F. The minimum value of $\epsilon = 1.73$ kcal/mol is obtained for $N_{max} = 15$ and $E_{max} = 3.0$ kcal/mol. We find that the contact model provides a stringent test of more complicated models because, as demonstrated below, it is fairly successful in binding affinity and weight matrix predictions.

$\Delta\Delta G$ predictions and weight fitting. The extent to which the static model reproduces experimental $\Delta\Delta G$ measurements from the 196 point dataset used for static model weight fitting is shown in Figure 1A (closed circles). The overall correlation between experimental and predicted binding free energies is 0.57 (Table 3). This is better than the correlation of 0.42 predicted with the contact model (Figure 1A, red triangles). The number of mutations from the consensus sequence is also shown (Figure 1A, green triangles). With the static model, the fraction of correct predictions is 73%, somewhat higher than

Table 3. $\Delta\Delta G$ predictions summary

N	r	F	ϵ	Description
1	0.57	143/196	1.58	static model: full
2	0.57	125/196	1.80	static model: protein–DNA only
3	0.43	133/196	2.90	dynamic model: full
4	0.41	128/196	2.93	dynamic model: protein–DNA only
5	0.39	117/196	1.86	dynamic model: full, fitted weights
6	0.42	137/196	1.73	contact model

r, Linear correlation coefficient; F, fraction of successful predictions (see text); and ϵ (kcal/mol), average unsigned error between predicted and experimental-binding affinities.

the 70% fraction for the contact model. The average unsigned error between experiment and prediction (ϵ) is 1.58 kcal/mol for the static model. Most of the errors (with an exception of the λ repressor) occur when the experimentally measured reduction in binding affinity is underpredicted by our method. This will lead to false positive hits in genomic sequence scans, but will not miss true binding sites. Finally, we observe that including DNA conformational energy into the model is beneficial on average even though the relative degree of its importance is probably protein family specific: prediction quality decreases when it is excluded from the fit (Table 3).

The EcoRI endonuclease example clearly demonstrates why the fraction of correct prediction is a relevant measure of success. The linear correlation between prediction and experiment in a series of destabilizing mutations of the EcoRI endonuclease (41) is only 0.19, but all of the mutations are correctly predicted to be destabilizing (Figure 1A, EcoRI panel). The experiment shows binding free energy reduction in the 4.1–5.7 kcal/mol range, whereas static model predictions range from 0.83 to 12.21 kcal/mol (the latter prediction is for the inverted 6 bp consensus sequence: GAATTC → CTTAAG; experiment shows 5.2 kcal/mol reduction in binding energy in this case). Two binding free energy changes are underpredicted at 0.83 and 1.06 kcal/mol, but the rest penalize unfavorable mutations with 1.5 kcal/mol or higher. Therefore, high-affinity sequences will be correctly identified by our method even though the degree of suppression of unfavorable mutations is not perfectly reproduced. In accordance with this observation, only 1 out of 13 predictions is labeled as incorrect in the binary measure. Another example is a series of seven destabilizing mutations of the zinc finger consensus sequence (Figure 1A, Zif268 panel) (42). Apparent binding dissociation constants (determined by a nitrocellulose filter binding assay) were too low to be measured accurately; the experiment only provides the lower bound of 3.13 kcal/mol on the binding energy reduction. Our calculations underpredict the magnitude of destabilization in two out of seven cases (giving 0.55 and 0.77 kcal/mol rather than $\geqslant$3.13 kcal/mol), but the other five mutations are correctly identified as being very unfavorable. The linear correlation coefficient is less relevant in this case since experimental binding affinities are not exactly known.

Figure 1B shows the dynamic model predictions for the same $\Delta\Delta G$ dataset. Unlike the static model, relative weights of energy terms are not adjusted using these data. Overall, the prediction quality is somewhat decreased compared with the static model, both in terms of the correlation coefficient and the fraction of correct predictions (Table 3). There is a striking increase in the average error (2.90 kcal/mol versus 1.58 kcal/mol), caused by the absence of the least squares fit to the data. Indeed, the error decreases considerably when such a fit is carried out (Table 3). The decision to 'turn on' conformational rearrangements and energy minimization depends on several crucial factors, such as the quality of the structural template (e.g. its X-ray resolution), the degree of DNA bending (in some extreme cases such as IHF in *E.coli* DNA is distorted so much that our effective quadratic description of DNA conformational energies may be no longer accurate), and the necessity to model changes in DNA shape caused by amino acid substitutions at the protein–DNA binding interface. It is remarkable that in most cases where the protein–DNA crystal structure is available, reasonable quality $\Delta\Delta G$ predictions can be made simply by computing the free energy of the initial, non-minimized complex. The static model is very efficient computationally and is nearly pairwise additive by construction, making possible rapid scans of longer genomic sequences for high-affinity binding sites. However, it is more likely to fail when DNA conformational change is required to predict novel binding sites by homology, because the DNA shape from the structure will always favor the original binding site sequence unless it is allowed to relax.

In Figure 2 we show four binding free energy predictions with the static model. Static model weights are not fitted on this dataset. The first prediction is for Ndt80, the primary transcriptional activator of the middle sporulation genes in budding yeast (43). Ndt80 binds to the middle sporulation element (MSE) with the gNCRCAAAW consensus sequence found in promoters of middle sporulation genes. Its *in vitro* binding affinity was studied by Pierce *et al.* (34) for a number of one-point mutations from the GTCACAAAT MSE variant and its flanking sequence. The measurements were carried out using the electrophoretic mobility shift assay (EMSA) and reported as ratios of protein concentrations bound to the mutant sequence and the wild-type sequence. We converted these ratios into changes in free energy using the Boltzmann formula, and carried out binding affinity predictions starting from the Ndt80 crystal structure with 1.4 Å resolution and the GACACAAAA site. The correlation between theoretical predictions and experimental measurements is 0.67, the average error is 0.58 kcal/mol, and the fraction of correct predictions is 19/26. When the dynamic model is used for predictions the correlation improves to 0.74, but the fraction of correct predictions drops to 15/26 (data not shown).

The second prediction is for the MAT **a**1/α2 homeodomain heterodimer. **a**1 and α2 TFs bind cooperatively to repress transcriptional activity of haploid-specific genes in diploid **a**/α cells (44). Jin *et al.* (30) investigated effects of mutations in the **a**1–α2 binding site on *in vivo* repression of a heterologous promoter assayed for β-galactosidase activity in wild-type diploid **a**/α cells. The presence of the **a**1–α2 binding site in the promoter causes MAT **a**1/α2 dependent repression of *lacZ* expression. Repression ratios relative to wild type are converted into energies using the Boltzmann formula (see Methods) and compared to the experimental predictions. For the static model, the correlation coefficient is 0.45, the average error is 0.62 kcal/mol, and 44 out of 54 measurements are predicted correctly according to our definition. All but one of the incorrect predictions result in energies that are lower than corresponding experimental energies

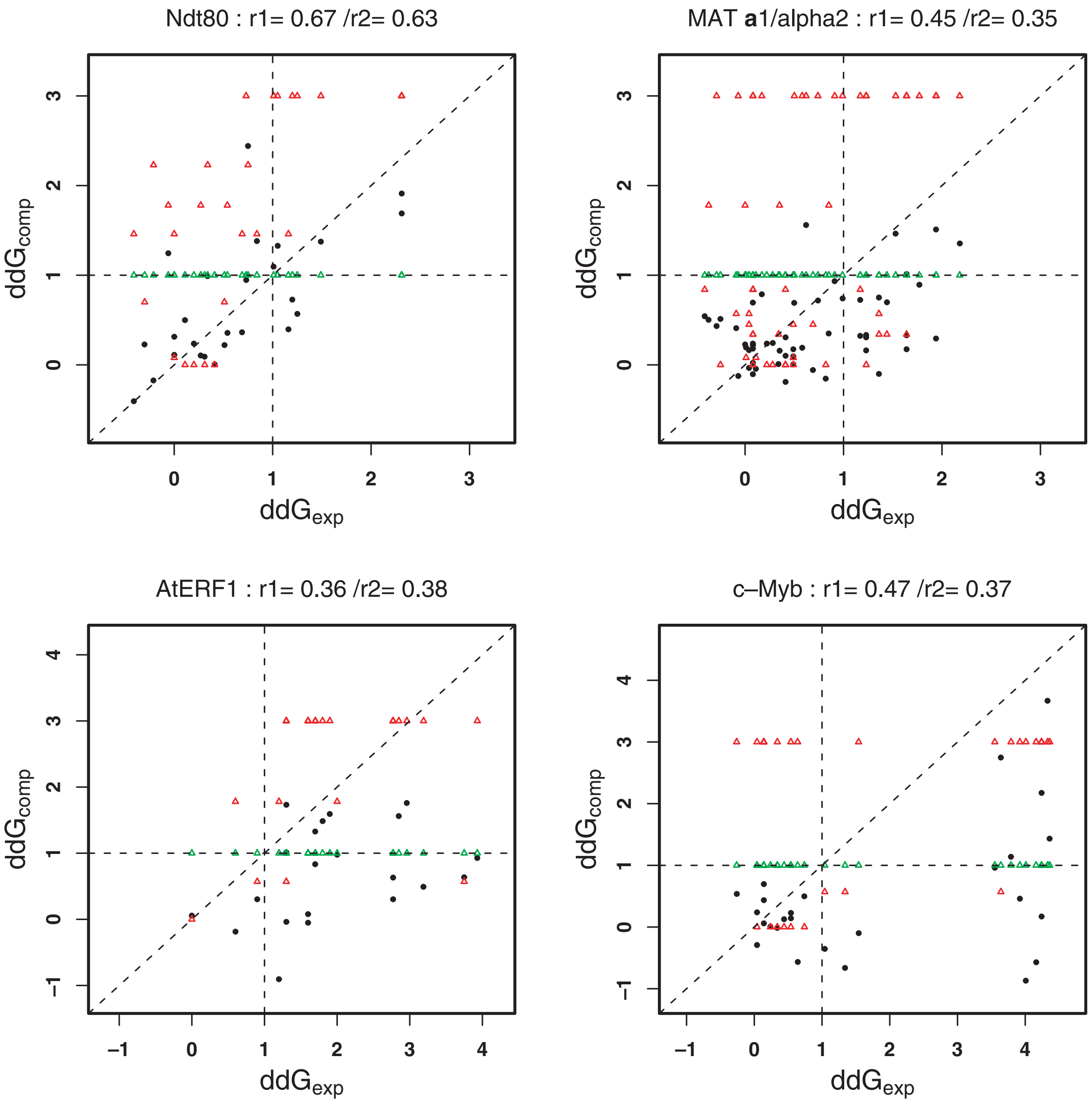


Figure 2. $\Delta\Delta G$ predictions (ddG$_{comp}$) versus experimental measurements (ddG$_{exp}$). Static model binding energy predictions for Ndt80, (34) MAT **a**1/α2, (30) AtERF1 (45) and c-Myb. (47) Closed circles, static model; red triangles, contact model; green triangles, number of mutations from the consensus sequence. r1-Linear correlation coefficient for the static model and r2-linear correlation coefficient for the contact model.

(Figure 2). The dynamic model is less successful in this case ($r = 0.37$; data not shown), probably because the X-ray structural template is only resolved to 2.5 Å and the 19 bp long binding site presents a challenge for DNA conformational modeling.

For our third and fourth predictions the only available structural templates were solved by NMR rather than X-ray crystallography. AtERF1 is the ERF DNA-binding domain from *Arabidopsis*, which mediates gene regulation by the plant hormone ethylene (45). c-Myb is a product of the mouse protooncogene *c-myb* essential in proliferation and differentiation of hematopoietic cells (46). In both cases the binding affinity data were obtained by using EMSA (45,47). Surprisingly, the static model is capable of reasonable quality predictions (Figure 2). The fraction of correct predictions is 12/21 for AtERF1 and 20/27 for c-Myb. All incorrect predictions classify corresponding mutations as too favorable, probably as a result of experimental uncertainty in atomic coordinates of NMR structures. This observation is confirmed by a significant drop in prediction quality when the DNA conformation is allowed to relax (data not shown). We generally find that attempts at refinement of side chain and DNA atomic positions starting from NMR structures do not lead to better $\Delta\Delta G$ and weight matrix predictions, in agreement with a previous observation that X-ray structures are consistently more accurate than NMR-derived structures (48).

DNA conformational energy and sequence specificity. Finally, we studied how well the DNA base step energy captures sequence specificity owing to indirect readout in BamHI endonuclease (49) and the PU.1 ETS domain (50). Engler *et al*. (49) mutated 3 bp sequences flanking the GGATCC core BamHI

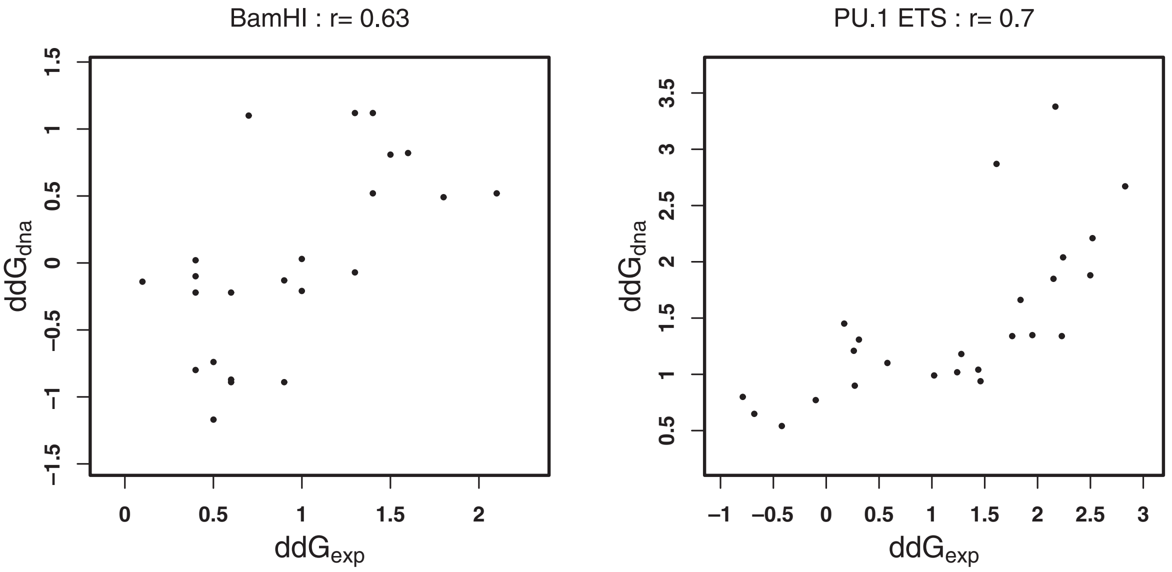


Figure 3. Experimental binding affinities conferred by indirect readout can be explained with DNA conformational energies alone. Dynamic model predictions of DNA base step energies (ddG$_{dna}$) versus experimental binding free energies (ddG$_{exp}$) for BamHI endonuclease (49) and PU.1 ETS domain (50).

recognition site and measured resulting binding free energy changes. Inspection of the protein–DNA structure reveals that outside of the core binding site sequence specificity is mostly imparted by protein-phosphate backbone contacts and thus should be ‘recorded’ in the DNA shape. Using the 2.2 Å crystal structure of the BamHI/DNA complex as a modeling template we were able to reproduce experimental binding affinities measured for flanking sequence mutations with a correlation coefficient of 0.63 (Figure 3, left panel). Interestingly, this prediction requires DNA minimization in the context of the protein–DNA complex, probably owing to DNA geometry artifacts in the initial structure (the correlation drops to 0.24 if DNA is not allowed to relax). For the PU.1 ETS domain, up to 3 bp upstream and/or 2 bp downstream of the GGAA core recognition site were mutated and corresponding free energy changes measured using EMSA (50). We again observe that most of the flanking sequence specificity is due to the changes in DNA shape conferred by protein-phosphate backbone contacts and captured by the DNA base step energy (Figure 3, right panel). DNA minimization is not as critical in this case (the correlation is 0.69 for the static model and 0.70 for the dynamic model).

Since most of the binding specificity for base pairs dominated by indirect readout can be captured by the DNA conformational energy, it is interesting to know what fraction of the base pairs in protein–DNA complexes has mostly phosphate backbone contacts. For these base pairs, the DNA conformational energy could be more important than the energies of direct interactions between protein side chains and DNA bases. A survey of protein–DNA contacts in structures from Figure 1 shows that out of 143 bp which have at least one contact with the protein, 54 bp have a ratio of phosphate backbone contacts to DNA base contacts of 3 or higher. When the static model $\Delta\Delta G$ fit is restricted to mutations at positions with high backbone/base contact ratio, the fitted DNA weights are indeed several times greater than in the fit with all data points, indicating the increased importance of DNA conformational energies in this limit (data not shown).

Additivity in protein–DNA interactions

The assumption of independence of DNA base pair probabilities at each position in the binding site forms a basis of the weight matrix description of binding specificity (1,2). Independence of DNA base pair probabilities implies that the binding energy of a given DNA sequence is a sum of energies associated with each base pair (Equation 8). Although there is direct experimental evidence indicating that pairwise additivity is a reasonable approximation for λ, Cro and Mnt repressors (51–53), there are also reports in the literature emphasizing the importance of dinucleotide and higher order correlations (54,55). Nonetheless, it is generally believed that pairwise additive energies do provide a reasonable approximation to true protein–DNA interaction energies (1,56).

In protein–DNA energy, non-additivity can only arise in the dynamic model because all atomic potentials are pairwise. When conformational rearrangement is allowed, the degree of additivity will depend on the range of protein–DNA interactions. For example, long-range electrostatic interactions may cause more deviations from pairwise additivity than relatively short-range van der Waals and hydrogen bonding interactions. The degree of conformational change at the protein–DNA interface in turn depends on the quality of experimental structures and the number of water molecules at the interface (as they are not explicitly modeled in our approach). In DNA conformational energy the base stacking term is non-additive by construction. If, however, the total binding energy predicted by our model turns out to be approximately pairwise additive, the search for binding sites in genomic sequence can be considerably simplified by constructing a weight matrix or a table of energies for all one-point mutations from the consensus sequence, instead of independently computing ΔG values for each putative binding site.

In Figure 4, we compare protein–DNA binding energies computed directly and by assuming pairwise additivity. In the latter approach, the protein–DNA binding energy change

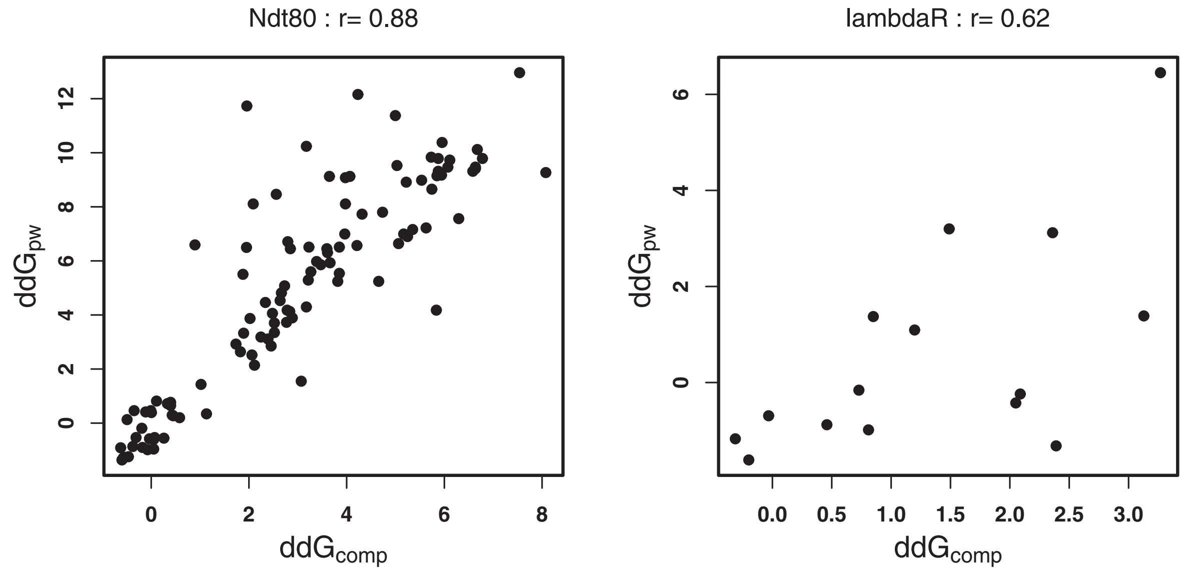


Figure 4. Degree of pairwise additivity in binding energies predicted with the dynamic model. Comparison of binding energies computed after making multiple base pairs substitutions (ddG$_{comp}$) with the sum of binding energies computed for corresponding one-point mutations of the DNA site from the protein–DNA structure (ddG$_{pw}$).

caused by an arbitrary number of binding site mutations is computed as a sum over energy changes $\Delta\Delta G^i_\alpha$ associated with one-point mutations from the consensus sequence:

$$\Delta\Delta G(S) = \sum_{i=1}^{L}\sum_{\alpha=1}^{4}\Delta\Delta G^i_\alpha \delta_{\alpha,s_i}, \tag{15}$$

where L is the length of the binding site in base pairs, s_i is the base at position i in sequence S, $\alpha = \{\mathrm{A, C, G, T}\}$, and $\delta_{\alpha,s_i} = 1$ if $\alpha = s_i$ and 0 otherwise. In the left panel of Figure 4 we plot the energies of all adjacent dinucleotide mutations of the Ndt80 binding site using the dynamic model (all terms in the static model except the base stacking energy are exactly pairwise additive). The protein–DNA interaction energies of the mutant sequences are quite well reproduced by the additive model. One discrepancy is in the absolute magnitude of binding energies which is predicted to be too large in the additive model. In the right panel of Figure 4, a similar comparison is carried out for a set of sequences including naturally occurring λ repressor operators and multiple mutations of the O_R1 site. The agreement is reasonable but somewhat less pronounced, perhaps reflecting the increased number of mutations from the O_R1 consensus sequence. For each sequence in this set, $\Delta\Delta G$ and $\Delta\Delta G^i_\alpha$ measurements were also made (51). A similar plot using experimental $\Delta\Delta G$ measurements shows better correlation (51), demonstrating limits of applicability of computational models to predicting the experimental degree of pairwise additivity.

Binding site discrimination

In order to carry out successful predictions of genomewide transcriptional regulation, computational models of protein–DNA binding free energy should be able to discriminate TF binding sites from random ensembles of sequences. Since we demonstrated that the static model works best for $\Delta\Delta G$ predictions, we assess its discriminatory power here by computing binding free energies of 16 TFs for which multiple DNA-binding sites are available (Table 2). Because pairwise additivity is nearly exact for the static model, we can compute binding energies of sites with arbitrary sequences using only binding energies for one-point mutations from the consensus sequence as input (Equation 15). The degree of discrimination of TF binding sites from random sites is given by the Z-score:

$$Z(S) = \frac{\Delta\Delta G(S) - \langle\Delta\Delta G\rangle}{\sigma}, \tag{16}$$

where $\Delta\Delta G(S)$ is the binding energy of sequence S measured relative to the wild-type binding energy, $\langle\Delta\Delta G\rangle$ is the average binding energy for the ensemble of all possible 4^L sequences (L is the length of the binding site) and σ is the standard deviation for the same ensemble.

In Table 4 we show Z-scores for the binding site from the protein–DNA complex (Z_{PDB}) and Z-scores averaged over all sites from Table 2 ($\langle Z_{\mathrm{site}}\rangle$). In addition, binding energies with sites from protein–DNA complexes are ranked for all TFs with $L < 15$. The overall quality of predictions is quite good, consistent with our previous ΔG predictions. Interestingly, the energy of the binding site from the protein–DNA structure is more favorable than the average energy of all binding sites in all cases except Trl (1yui; Table 4), showing that most experimentally characterized binding sites have lower affinities than the consensus site. Nonetheless, almost all of the inspected sites have highly favorable binding energies and thus low Z-scores. One notable exception is the integration host factor protein (Ihf) for which the average Z-score for 27 sites from Table 2 is only −1.12. The large discrepancy between the average Z-score and the PDB Z-score might be explained in this case by the major role of indirect readout in Ihf binding: the crystal structure of the Ihf–DNA complex shows that DNA is bent almost 180° and has relatively few direct contacts between side chains and DNA bases, especially in the 5′ region of the binding site. The experimental PWM is relatively

Table 4. Summary of binding site energy predictions

PDB	Z_{PDB}	$\langle Z_{\mathrm{site}} \rangle$	Rank	L
1mnn	−3.69	−2.54	14	12
1ysa	−3.38	−2.53	1	7
1yrn	−4.30	−3.34	–	19
1aay	−3.70	−3.60	13	10
1b8i	−3.22	−2.74	8	10
1r0o	−3.80	−2.88	–	15
2drp	−3.77	−2.76	6	11
1yui	−2.38	−2.48	91	7
1fjl	−3.22	−2.93	4507	13
1gxp	−4.17	−3.06	–	20
1ihf	−3.41	−1.12	–	34
1j1v	−4.20	−3.36	13	13
1mj2	−3.09	−2.13	–	16
1run	−3.50	−2.32	–	22
1tro	−3.73	−2.57	–	18
2puc	−3.88	−3.68	–	16

Z_{PDB} is the Z-score (Equation 16) for the protein–DNA binding energy with the binding site found in the protein–DNA structure; $\langle Z_{\mathrm{site}} \rangle$ is the average Z-score for protein–DNA binding energies with binding sites listed in Table 2; Rank is the rank of the binding energy for the structural site in the ensemble of 4^L sequences (L is the binding site length). Rank was computed for all binding sites with $L < 15$. Static model was used in all predictions. TF binding sites (including sequences from protein–DNA complexes) are as listed in Table 2 (see Methods for details).

non-specific, and different sites or groups of sites might utilize significantly different binding modes that are not captured well by our approach.

Ranking binding sites from the protein–DNA structure versus ensembles of random sites provides further illustration of the accuracy of our predictions: for example, the native binding site of the Ndt80 (1mnn) is 14th out of 16 777 216 sequences, whereas in Gcn4p (1ysa) the native binding site is the lowest in energy among 16 384 sequences (Table 4). Similar to $\Delta\Delta G$ predictions, binding site discrimination from random sequences strongly depends on the quality of the structural template: Trl (1yui) native site is ranked only 91st out of 16 384, most probably because the structure of the protein–DNA complex was determined by NMR.

PWM predictions

Results from the previous section show a reasonable degree of additivity in protein–DNA binding free energy predictions, most probably because of the limited role of long-range interactions in our model. Therefore, we can convert binding energies into weight matrices without significant loss of information and test our PWM predictions against experimental data. Similar to $\Delta\Delta G$ predictions, we compare all-atom static and dynamic models with the simpler contact model, which uses the number of atomic contacts between protein side chains and DNA base pairs as a measure of binding specificity (Equation 12).

Experimental PWMs and prediction testing. We constructed experimental PWMs for 20 TFs using several alternative approaches. PWMs for λR (1lmb), CroR (6cro), c-Myb (1mse) and AtERF1 (1gcc) were created by converting experimental binding free energy measurements for all one-point mutations of the binding site (Table 1) into probabilities using the Boltzmann formula at room temperature. The resulting PWMs are then constructed in a way directly comparable to computational predictions. Another and more commonly used method for constructing PWMs is based on the alignment of binding sites obtained from either a SELEX experiment or a set of promoter sequences of genes regulated by the TF (2). The quality of weight matrices obtained from such alignments depends on the number of sequences used in the alignment. In our 20 PWM dataset shown in Table 2, 4 PWMs are created from $\Delta\Delta G$ measurements, 14 PWMs are based on genomic sites available from the literature, the Zif268 PWM is from a SELEX experiment (31), and the EcR/Usp PWM comes from a combination of SELEX (57) and genomic sites. PWM predictions are analyzed using the ψ-test (Equation 14). $\psi(p, q)$ is a non-negative measure of the ‘goodness of fit’ between computational probabilities and experimental frequencies (38). It is a monotonic function of prediction quality.

Free parameters of PWM models. The contact model replaces detailed calculations of protein–DNA energetics with an assumption that the probability of each base in the consensus sequence is directly proportional to the number of contacts made between the base pair and all protein side chains (Equation 12). As in the binding free energy contact model, protein–DNA contacts are defined as protein–DNA base atomic pairs within $R_{\max} = 4.5$ Å; contacts to the phosphate backbone are ignored. As the number of contacts N increases, the consensus base becomes more and more specific, with the rest of the probability evenly divided between the other three bases. The probability of the consensus base becomes 1 for $N = N_{\max}$, and all other probabilities become 0. Given $R_{\max}$, $N_{\max}$ is a free parameter to be adjusted by minimizing $\psi(p, q)$ averaged over a subset of TFs from Table 2. PWMs constructed from fewer than 10 binding site sequences [including Gcn4p (1ysa), Trl (1yui) and DnaA (1j1v)] are removed from the $N_{\max}$ fit. The average value of ψ is at a minimum for $N_{\max} = 20$.

Using the Boltzmann formula to convert energies into probabilities involves an inverse temperature factor $\beta = 1/RT$, which can also be viewed as an adjustable scaling factor. The specificity of PWM predictions depends on β: at low temperatures only a few lowest energy binding sites contribute to weight matrix probabilities, whereas at high temperatures a broader spectrum of sites is included. Therefore, incorrect predictions of low-energy sites will result in higher fitted temperatures. The static model fit with the 17 TF dataset used for adjusting $N_{\max}$ in the contact model gives $\beta = 2.25$ (kcal/mol)$^{-1}$. For the dynamic model fit, we additionally exclude all NMR structures (1gcc and 1mse), crystal structures with >2.5 Å resolution (6cro, 2drp, 1run and 2puc) and the Ihf–DNA complex (1ihf) because its DNA conformation cannot be reasonably expected to be modeled by relaxation with the quadratic DNA potential. The dynamic model fit over 10 remaining TFs results in $\beta = 0.75$ (kcal/mol)$^{-1}$.

PWM predictions and comparison with experiment. The fitted values of $N_{\max}$ and β are used to make PWM predictions for all 20 TFs (16 for the dynamic model as 3 NMR structures and Ihf are excluded). In Table 5 we show values of ψ computed using the contact, static and dynamic models. We compare model predictions with the average value of ψ for an ensemble of randomly generated weight matrices ($\langle \psi_{\mathrm{random}} \rangle$ column in

Table 5. The ψ-test of PWM predictions for the contact, static, dynamic and random models

PDB	ψ_{contact}	ψ_{static}	ψ_{dynamic}	$\langle\psi_{\text{random}}\rangle$
1mnn	0.14	0.12	0.20	0.68
1ysa	0.16	0.31	0.39	0.91
1yrn	0.20	0.26	0.36	0.73
1lmb	0.10	0.09	0.14	0.47
6cro	0.07	0.10	0.21	0.47
1mse	0.26	0.24	–	0.55
1gcc	0.15	0.12	–	0.57
1aay	0.13	0.19	0.35	0.95
1b8i	0.35	0.34	0.36	0.87
1r0o	0.22	0.25	0.38	0.72
2drp	0.19	0.24	0.23	0.69
1yui	0.07	0.26	–	0.95
1fjl	0.30	0.32	0.51	0.83
1gxp	0.24	0.28	0.41	0.68
1ihf	0.21	0.19	–	0.43
1j1v	0.21	0.22	0.36	0.70
1mj2	0.21	0.38	0.33	0.69
1run	0.10	0.17	0.38	0.51
1tro	0.30	0.31	0.39	0.71
2puc	0.18	0.26	0.63	0.81

Random model uses randomly sampled normalized PWM entries (see Methods). Three NMR structures and Ihf (1ihf) are excluded from the dynamic model predictions.

Table 5). ψ is lower than the corresponding random value if a prediction is successful. For the contact model the average value of ψ over all TFs is 0.19, significantly better than the random value of 0.70. For the static and dynamic models the average over all TFs increases to 0.23 and 0.35, respectively. Note, however, that even for the least successful predictions (1b8i for the contact model, 1mj2 for the static model and 2puc for the dynamic model) ψ is much smaller than the corresponding random ψ (Table 5). Surprisingly, it is the contact model that is the most successful on average: the static model has a lower ψ in only 6 cases out of 20. This finding demonstrates that PWM predictions may not require detailed models of protein–DNA energetics if native protein–DNA complexes are available. Furthermore, allowing conformational change generally makes predictions worse, consistent with our earlier observations regarding $\Delta\Delta G$ predictions.

The ψ-test provides only an average measure of success and does not necessarily reflect all relevant details of probability distributions in specific columns. Hence it is useful to analyze several PWM predictions in more detail. In Figure 5, PWM WebLogo representations (58) are shown for 2 TFs: Ndt80 (A) and Zif268 (B). The total height of each weight matrix column is constant and the relative height of each letter is proportional to its probability in a given PWM column. All bases are sorted by probability. For each panel in Figure 5 the top logo is the experimental PWM and the other three logos are predictions with the contact model, the static model and the dynamic model, respectively.

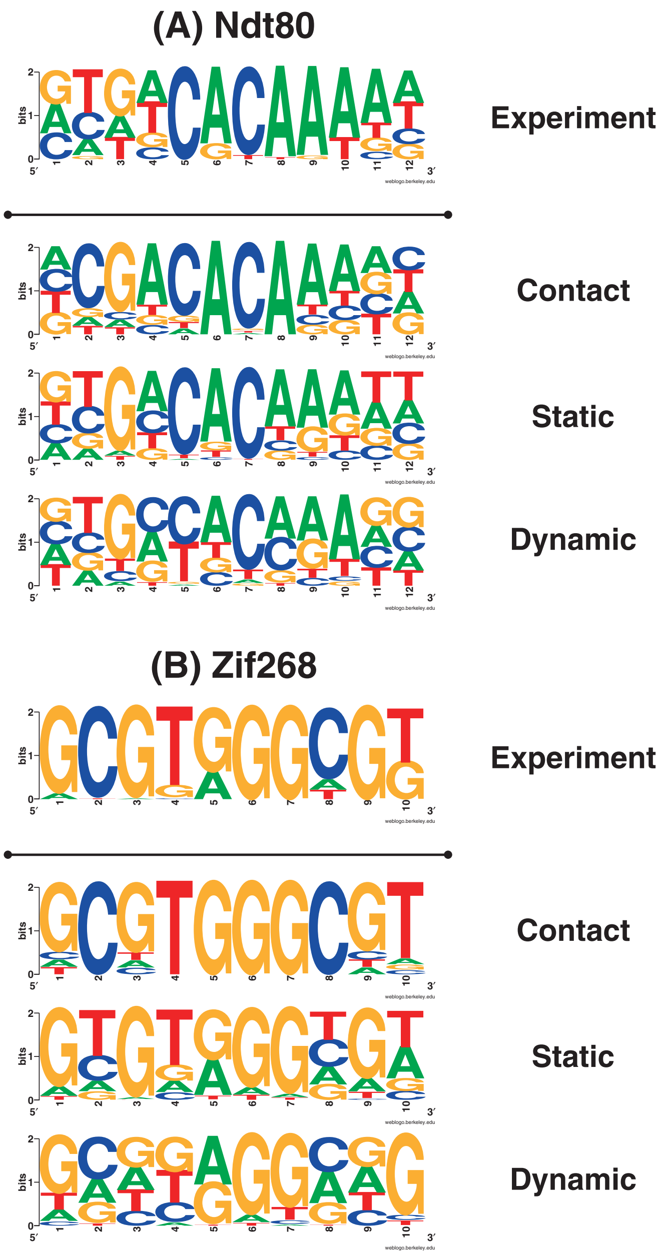


Figure 5. PWM predictions for Ndt80 (**A**) and Zif268 (**B**). From top to bottom: experiment, contact model based on the consensus sequence and the number of protein–DNA contacts, static model and dynamic model (see text for details). PWMs are displayed using the uniform height WebLogo representation(58): the height of each letter in the column is proportional to its probability in the PWM.

For Ndt80, ψ is 0.14 for the contact model, 0.12 for the static model and 0.20 for the dynamic model (Table 5). Figure 5A shows that relative specificities of C at position 5 and A/g at position 6 are best reproduced with the static model (the contact model underpredicts specificity of C5 and overpredicts specificity of A6, whereas in the dynamic model both C5 and A6 are insufficiently conserved). All three computational models overpredict specificity of guanine at position 3 and underpredict specificity of adenine at position 9. Specificity of adenine at position 8 is predicted best with the contact model. Our second prediction is for the Zif268 zinc finger with the experimental PWM from a SELEX experiment (Figure 5B) (31). The value of ψ is 0.13, 0.19 and 0.35 for the contact, static and dynamic models, respectively. Because almost all positions in the experimental PWM are strongly conserved, the contact model with the appropriate $N_{\max}$ cutoff is capable of a near perfect prediction of experimental results. The high degree of specificity in the SELEX experiment could be an artifact of the stopping criteria in the selection rounds. The main advantage of more complicated models is in making some columns less specific in better agreement with the experimental data, notably G/a at position 5. Interestingly, the ψ-test yields 0.13 for the weight matrix predicted using zinc finger energy matrices from Benos *et al.* (15), comparable to our predictions with contact and static models.

In summary, reliable PWM predictions can be carried out if the native structure of the protein–DNA complex is used as a starting point for computational modeling. The contact model which assigns base pair specificity based on the number of atomic contacts between protein amino acids and DNA bases is surprisingly successful. In some cases, the static model provides the best results: although the agreement with the experimental data is somewhat worse on average compared with the contact model, core motif probabilities are often better reproduced (Figure 5). Finally, the dynamic model is typically worse than the others, especially in cases where a high-resolution crystal structure of the protein–DNA complex is not available, or where ordered water molecules mediating protein–DNA interactions play a significant role. Water molecules are not explicitly modeled in our approach and, thus, their removal from the interface may result in artificially increased conformational freedom of neighboring bases and side chains. Limited utility of side chain conformational rearrangement in computational modeling of intermolecular binding interfaces was previously noted in the context of protein–protein complexes (59).

Homology modeling

Binding affinity and specificity predictions described in the previous sections require native structures of protein–DNA complexes. However, even in well-studied organisms such as *S.cerevisiae* and *D.melanogaster* there are only 10–15 suitable structures in the database. Furthermore, in most cases experimentally available protein–DNA complexes are not focused on any specific biological pathway (such as regulation of the *Drosophila* segmentation gene network), being instead distributed across a range of regulatory pathways and cell types. Hence the ability to model protein–DNA interactions by homology is crucial to future practical applications of our approach.

A suitable modeling template is initially identified by sequence similarity to protein–DNA complexes in the structural database using Ginzu (60). Besides sequence similarity, the quality of experimental structures (such as X-ray resolution or missing atoms) is taken into account. Amino acid substitutions at the DNA-binding interface are identified by a sequence–structure alignment with the template using K*Sync (60) or ClustalW (61). The approximate pairwise additivity of base pair energies and the relatively short range nature of the free energy function make it possible to ignore amino acid substitutions elsewhere because they are less likely to mediate protein–DNA interactions. An obvious consequence of this assumption is that binding specificity does not change if all amino acids at the protein–DNA interface are conserved. To give a specific example, essentially all DNA contacting amino acids are conserved in the Rel homology region family (8) and, thus, binding sites for the embryonic polarity protein dorsal in fly bear a strong resemblance to nuclear factor κB sites from mouse and human [e.g. a dorsal weight matrix from Ref. (62) gives the T-G/T-G-G-A/T-T-T-T-T/C-C-C consensus sequence, very close to the T-G-G-G-A-A-T-T-C-C-C binding site from the structure of the mouse nuclear factor κB p50 homodimer bound to DNA]. A classification of protein–DNA complexes into families and subfamilies has been carried out on the basis of identical DNA contacting amino acids (8). The definition of interface amino acids is somewhat arbitrary: typically, amino acids are considered to be at the interface based on distance cutoffs and/or visual identification of hydrogen bonds, van der Waals contacts and favorable electrostatic interactions with DNA atoms. We adopt a simple interface definition based on the 4.5 Å cutoff between protein side chain atoms and DNA base or phosphate backbone atoms.

In Figure 6 we show PWM predictions by homology for two TFs from the *Drosophila* segmentation gene network: the maternal factor bicoid (Bcd) and the gap factor giant (Gt) (63). Bcd is a homeodomain TF; its experimentally characterized binding sites suggest monomeric binding. The experimental PWM for Bcd obtained from the alignment of 30 binding sites shows a strongly conserved C/T-T-A-A-T-C-C/T-C/G consensus sequence. We use structures of the Engrailed homeodomain–DNA wild-type complex (3hdd) and its Q50K mutant (2hdd) as modeling templates (64,65). The amino acid at position 50 is an important specificity determinant for the 2 bp 3′ of the core TAAT homeodomain motif: Lys-50 makes favorable contacts with two guanines complementary to the CC dinucleotide immediately 3′ of TAAT, contributing to the difference between Q50 wild-type Engrailed (GTAATTAC) and Engrailed Q50K (GTAATCCC) binding sites from the structures.

Using the Q50K Engrailed mutant as a structural template for Bcd makes homology modeling relatively easy: all amino acids are conserved at the DNA-binding interface, even though there are 28/55 amino acid substitutions and a 2 residue gap in the alignment. Because Q50K Engrailed and Bcd DNA-binding interfaces are virtually identical, the experimental PWM for Bcd is reasonably well reproduced by the contact model. Prediction of the motif 3′ of the TAAT core is further improved with the static model, but the TAAT motif becomes less specific (Figure 6A). Structural analysis of homeodomain–DNA complexes reveals that homeodomain binding causes distortion in the DNA conformation (37,66). The conformational change serves to enclose the recognition helix within the major groove, increasing the surface area of the protein–DNA interface and resulting in the B_{eg}–DNA (enlarged groove) DNA form (66). The distortion of the DNA site is captured by the DNA conformational component of the free energy function (Figure 6A; note that DNA conformational energy weights from the static model are multiplied by 5). Surprisingly, DNA shape confers binding specificity not only to the TAAT core homeodomain motif, but also to the 3′ CCC motif.

Owing to the differences between Bcd- and En-binding specificities, the contact model is not very successful in predicting the bicoid PWM starting from the wild-type En–DNA complex. The dynamic model reproduces PWM columns 8 and 9 significantly better, but in column 7 adenine is favored over cytosine, and in column 4 adenine is mixed with guanine

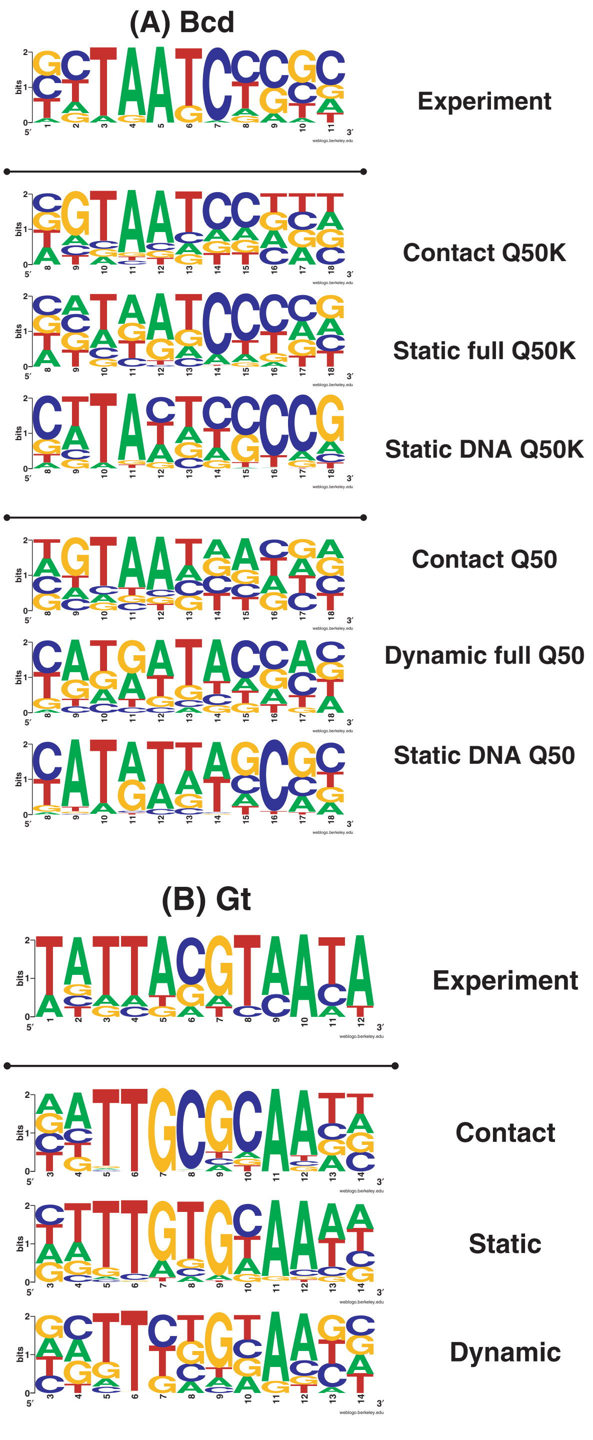


Figure 6. PWM predictions by homology for *D.melanogaster* TF bicoid (Bcd) (A) and giant (Gt) (B). From top to bottom: (**A**) panel 1: experiment; panels 2–4: contact model, static model with full energy function and static model with DNA conformational energy only (with dna–bp and dna–bs weights multiplied by 5) using *D.melanogaster* Engrailed homeodomain Q50K (2hdd) as a template; panels 5–7: contact model, dynamic model with full energy function (reference DNA energies are not subtracted since DNA is bent in homeodomains) and static model with DNA conformational energy only (with dna–bp and dna–bs weights multiplied by 5) using *D.melanogaster* Engrailed wild-type homeodomain (3hdd) as a template. (**B**) Experiment, contact model, static model and dynamic model using *Homo sapiens* nuclear factor NF-IL6 (C/EBP-β;1gu4) as a template. All amino acids substituted at the protein–DNA interface are repacked in the static model. PWMs are displayed using the uniform height WebLogo representation (58).

to the extent not corroborated by the experiment (Figure 6A). One possible reason for this discrepancy is that the model does not fully reproduce experimentally observed conformational differences between wild-type and Q50K DNA sites (64). This difference is reflected in wild-type Engrailed DNA conformational energies which favor T/A at position 7 and G at position 8 (Figure 6A; the static model DNA weights are multiplied by 5 as before). Better DNA conformational sampling should improve the accuracy of PWM predictions with homology models.

Giant is a TF from the leucine zipper family. It binds DNA as a homodimer, with the TTAC consensus motif at positions 3–6 and its inverted complement GTAA at positions 7–10 (Figure 6B). The experimental PWM for Gt is constructed using just 7 binding sites and thus might be too specific at positions 1 and 2 and 11 and 12, where there are few contacts with protein side chains (as is evident from the contact model prediction). We used the 1.8 Å resolution crystal structure of the human nuclear factor NF-IL6 bound to the GATTGCGCAATA site (1gu4) as the homology modeling template. For each monomer in the homodimer there are 5 amino acid mutations at the protein–DNA binding interface: 1 DNA base contact and 4 phosphate backbone contacts. However, only one of these contacts is non-homologous (K contacting phosphate backbone in NF-IL6 is mutated to V in Gt), suggesting that NF-IL6- and Gt-binding specificities should not be very different. The mutated LYS makes contacts with the phosphate group joining a dinucleotide complementary to TG at positions 4 and 5 in one chain, and to CA at positions 8 and 9 in the other chain (Figure 6B, panel 2 from top; column numbers are from the experimental PWM). In accordance with this observation, comparison of the experimental PWM for Gt with the contact model shows that the specificity change is largely restricted to positions 5 (G→A) and 8 (C→T). By construction, the contact model is incapable of accommodating this change, whereas static and dynamic models are partially successful, but at the price of making the prediction at position 6 in the experimental PWM less specific (Figure 6B).

The Bcd and Gt examples described above are representative of the future applications of our approach. Binding site specificity predictions based on structural modeling can be used in conjunction with existing bioinformatic algorithms to study regulatory gene networks in many species. Even though the requirement of having homologous structures of protein–DNA complexes with a limited number of mutations at the protein–DNA binding interface is much less restrictive, it is still likely to be the main limitation of our approach. PWM prediction by homology becomes less tractable if the number of mutations at the interface is >2 or 3, probably because the current implementation of our model does not allow protein backbone degrees of freedom to relax. Modeling TF binding specificities using distant homologs may require including rigid body motion of the TF and sampling over multiple docking conformations.

CONCLUSIONS

We developed a computational all-atom approach for predicting protein–DNA binding affinities and TF weight matrices. Protein–DNA energetics is described with the empirical free energy function that accounts for protein–DNA interactions (including electrostatics, solvation, hydrogen bonding, van der Waals interactions and packing) and distortion of the DNA shape caused by protein binding. Each term in the free energy function is multiplied by a weight which is adjusted to optimize the performance of the model on an experimental dataset. Free energy minimization and conformational rearrangement at the protein–DNA binding interface are either not employed at all (static model), or limited to repacking interface side chains and DNA minimization (dynamic model). Protein–DNA docking orientation and protein backbone conformation are kept fixed during energy minimization. Our approach is computationally efficient and can be applied on the genomewide scale. We demonstrated its utility by carrying out a number of $\Delta\Delta G$ and PWM predictions using native protein–DNA complexes as structural templates.

Proteins bind DNA in a sequence-specific manner by utilizing two distinct interaction mechanisms. The mechanism of direct readout is mediated by protein side chains directly contacting DNA base atoms. Favorable protein–DNA base contacts result in base pair preferences at corresponding positions in the binding site. The mechanism of indirect readout is mediated by side chain contacts with the DNA phosphate backbone. These contacts are typically as numerous as direct protein–DNA base contacts and can exploit DNA flexibility by twisting and bending it into the shape that fits best with the binding interface presented by the protein. Since some DNA sequences are more flexible than others, DNA conformational change confers additional sequence specificity to the binding site. In cases where indirect readout predominates, our model predicts a major contribution of the DNA conformational energy to the overall binding specificity (Figure 3).

None of the terms in the DNA base pair energy depend on neighboring base pairs in the absence of conformational rearrangement (except the base stacking energy) and, thus, DNA base pair energies are nearly independent in the static model and only weakly coupled in the dynamic model (Figure 4). Thus, we can convert binding affinity predictions for one-point mutations into weight matrix probabilities without significant information loss. Results in Table 5 demonstrate that we are reasonably successful in predicting experimental PWMs for a variety of TFs starting from the native protein–DNA complex. Surprisingly, a simple contact model based on the consensus sequence from the protein–DNA complex works quite well, even though specific examples in Figure 5 make evident some of its limitations compared with all-atom models of protein–DNA interactions.

The number of protein–DNA complexes currently available in the structural database is insufficient for modeling transcriptional regulation on a large scale. Therefore, the range of applicability of our approach depends on its accuracy in modeling TF binding specificities starting from homologous structures. Owing to the relatively short-range nature of our free energy function, it is sufficient to substitute amino acids only at the DNA-binding interface when creating protein–DNA homology models. Homology modeling should be easiest when there are no dissimilar amino acid substitutions at the interface, because in many instances TFs with conserved interfaces have identical binding specificities. Our model makes accurate predictions in such cases, but changes in binding specificity resulting from amino acid mutations are often predicted less accurately (Figure 6). Refinement of protein–DNA interfaces is a challenging problem which is strongly affected by the quality of experimental structural data and the presence of ordered water molecules mediating interactions across the protein–DNA binding interface. Homology-based predictions should be improved if interface waters are explicitly modeled and multiple protein docking conformations are allowed. Furthermore, homology modeling should be aided by better DNA conformational sampling (i.e. using simulated annealing techniques rather than minimization towards the nearest local minimum).

In summary, the computational algorithm developed here is useful for binding affinity and weight matrix predictions if either a native structure of the protein–DNA complex or its sufficiently close homolog is available. Unlike previously reported knowledge-based approaches (15,16), our algorithm is not limited to any specific TF family and is not as data intensive. However, its accuracy strongly depends on the quality of the experimental structure used as the modeling template, and the number of amino acid substitutions at the DNA-binding interface. In future, we intend to combine structurally predicted PWMs with motif detection algorithms in order to identify TF binding sites on the genomic scale.

SOFTWARE AND DATA AVAILABILITY

The protein–nucleic acid interaction module is implemented in C++ in the ROSETTA software package (http://www.bakerlab.org). ROSETTA software package is freely available to academic users. We hope to foster further development of computational algorithms for protein–DNA binding specificity predictions by providing all experimental datasets used in this study (including ΔG measurements, PWMs and TF binding sites) as Supplementary Data.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank Jonathan Widom for careful reading of the manuscript and many useful suggestions, and Eldon Emberly for providing experimental binding data for several TFs. A.V.M. would like to acknowledge a fellowship from the Leukemia and Lymphoma Society. J.J.H. is a fellow of the Jane Coffin Childs Memorial Fund for Medical Research. Funding to pay the Open Access publication charges for this article was provided by National Institutes of Health (NIH).

Conflict of interest statement. None declared.

REFERENCES

1. Bulyk,M. (2003) Computational prediction of transcription-factor binding site locations. *Genome Biol.*, **5**, 201.
2. Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.
3. Siggia,E.D. (2005) Computational methods for transcriptional regulation. *Curr. Opin. Genet. Dev.*, **15**, 214–221.

4. Seeman,N.C., Rosenberg,J.M. and Rich,A. (1976) Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl Acad. Sci. USA*, **73**, 804–808.
5. Suzuki,M. and Yagi,N. (1994) DNA recognition code of transcription factors in the helix–turn–helix, probe helix, hormone receptor, and zinc finger families. *Proc. Natl Acad. Sci. USA*, **91**, 12357–12361.
6. Matthews,B.W. (1988) Protein–DNA interaction. No code for recognition. *Nature*, **335**, 294–295.
7. Luscombe,N.M., Laskowski,R.A. and Thornton,J.M. (2001) Amino acid–base interactions: a three-dimensional analysis of protein–DNA interactions at an atomic level. *Nucleic Acids Res.*, **29**, 2860–2874.
8. Luscombe,N.M. and Thornton,J.M. (2002) Protein–DNA interactions: amino acid conservation and the effects of mutations on binding specificity. *J. Mol. Biol.*, **320**, 991–1009.
9. Pabo,C.O. and Sauer,R.T. (1992) Transcription factors: structural families and principles of DNA recognition. *Annu. Rev. Biochem.*, **61**, 1053–1095.
10. Pabo,C.O. and Nekludova,L. (2000) Geometric analysis and comparison of protein–DNA interfaces: Why is there no simple code for recognition? *J. Mol. Biol.*, **301**, 597–624.
11. Benos,P.V., Lapedes,A.S. and Stormo,G.D. (2002) Is there a code for protein–DNA recognition? Probab(ilistical)ly *Bioessays*, **24**, 466–475.
12. Mandel-Gutfreund,Y. and Margalit,H. (1998) Quantitative parameters for amino acid–base interaction: implications for prediction of protein–DNA binding sites. *Nucleic Acids Res.*, **26**, 2306–2312.
13. Kono,H. and Sarai,A. (1999) Structure-based prediction of DNA target sites by regulatory proteins. *Proteins*, **35**, 114–131.
14. Sandelin,A. and Wasserman,W.W. (2004) Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *J. Mol. Biol.*, **338**, 207–215.
15. Benos,P.V., Lapedes,A.S. and Stormo,G.D. (2002) Probabilistic code for DNA recognition by proteins of the EGR family. *J. Mol. Biol.*, **323**, 701–727.
16. Kaplan,T., Friedman,N. and Margalit,H. (2005) *ab initio* prediction of transcription factor targets using structural knowledge. *PLoS Comput. Biol.*, **1**, e1.
17. 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.
18. Paillard,G. and Lavery,R. (2004) Analyzing protein–DNA recognition mechanisms. *Structure*, **12**, 113–122.
19. Paillard,G., Deremble,C. and Lavery,R. (2004) Looking into DNA recognition: zinc finger binding specificity. *Nucleic Acids Res.*, **32**, 6673–6682.
20. 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.
21. 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.
22. Onufriev,A., Bashford,S.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.
23. Lazaridis,T. and Karplus,M. (1999) Effective energy function for proteins in solution. *Proteins*, **35**, 133–152.
24. Wang,J., Cieplak,P. and Kollman,P.A. (2000) How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem.*, **21**, 1049–1074.
25. Olson,W.K., Gorin,A.A., Lu,X., 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.
26. Lu,X. and Olson,W.K. (2003) 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structure. *Nucleic Acids Res.*, **31**, 5108–5121.
27. Lu,X., El Hassan,M.A. and Hunter,C.A. (1997) Structure and conformation of helical nucleic acids: analysis program (SCHNAaP). *J. Mol. Biol.*, **273**, 668–680.
28. Gromiha,M.M., Siebers,J.G., Selvaraj,S., Kono,H. and Sarai,A. (2004) Intermolecular and intramolecular readout mechanisms in protein–DNA recognition. *J. Mol. Biol.*, **337**, 285–294.
29. 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.
30. Jin,Y., Zhong,H. and Vershon,A.K. (1999) The yeast a1 and α2 homeodomain proteins do not contribute equally to heterodimeric DNA binding. *Mol. Cell. Biol.*, **19**, 585–593.
31. Swirnoff,A.H. and Milbrandt,J. (1995) DNA-binding specificity of NGFI-A and related zinc finger transcription factors. *Mol. Cell. Biol.*, **15**, 2275–2287.
32. Robison,K., McGuire,A. and Church,G. (1998) A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete *Escherichia coli* K-12 genome. *J. Mol. Biol.*, **284**, 241–254.
33. Salgado,H., Gama-Castro,S., Martínez-Antonio,A., Díaz-Peredo,E., Sánchez-Solano,F., Peralta-Gil,M., Garcia-Alonso,D., Jimenez-Jacinto,V., Santos-Zavaleta,A., Bonavides-Martinez,C. *et al.* (2004) RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12. *Nucleic Acids Res.*, **32**, D303–D306.
34. 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.
35. Natarajan,K., Meyer,M.R., Jackson,B.M., Slade,D., Roberts,C., Hinnebusch,A.G. and Marton,M.J. (2001) Transcriptional profiling shows that Gcn4p is a master regulator of gene expression during amino acid starvation in yeast. *Mol. Cell. Biol.*, **21**, 4347–4368.
36. Wingender,E., Chen,X., Fricke,E., Geffers,R., Hehl,R., Liebich,I., Krull,M., Matys,V., Michael,H., Ohnhauser,R. *et al.* (2001) The TRANSFAC system on gene expression regulation. *Nucleic Acids Res.*, **29**, 281–283.
37. Wilson,D.S., Guenther,B., Desplan,C. and Kuriyan,J. (1995) High resolution crystal structure of a paired (Pax) class cooperative homeodomain dimer on DNA. *Cell*, **82**, 709–719.
38. Jaynes,E.T. (2003) *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge, UK.
39. Dunbrack,R.L.,Jr and Cohen,F.E. (1997) Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci.*, **6**, 1661–1681.
40. Kuhlman,B. and Baker,D. (2000) Native protein sequences are close to optimal for their structures. *Proc. Natl Acad. Sci. USA*, **97**, 10383–10388.
41. Lesser,D.R., Kurpiewski,M.R. and Jen-Jacobson,L. (1990) The energetic basis of specificity in the EcoRI endonuclease–DNA interaction. *Science*, **250**, 776–786.
42. Hamilton,T.B., Borel,F. and Romaniuk,P.J. (1998) Comparison of the DNA binding characteristics of the related zinc finger proteins WT1 and EGR1. *Biochemistry*, **37**, 2051–2058.
43. 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.
44. Dranginis,A.M. (1990) Binding of yeast a1 and α2 as a heterodimer to the operator DNA of a haploid-specific gene. *Nature*, **347**, 682–685.
45. Hao,D., Ohme-Takagi,M. and Sarai,A. (1998) Unique mode of GCC box recognition by the DNA-binding domain of ethylene-responsive element-binding factor (ERF domain) in plant. *J. Biol. Chem.*, **273**, 26857–26861.
46. Lipsick,J.S. and Wang,D.-M. (1999) Transformation by v-Myb. *Oncogene*, **18**, 3047–3055.
47. Tanikawa,J., Yasukawa,T., Enari,M., Ogata,K., Nishimura,Y., Ishii,S. and Sarai,A. (1993) Recognition of specific DNA sequences by the c-*myb* protooncogene product: role of three repeat units in the DNA-binding domain. *Proc. Natl Acad. Sci. USA*, **90**, 9320–9324.
48. Lee,M.R. and Kollman,P.A. (2001) Free-energy calculations highlight differences in accuracy between X-ray and NMR structures and add value to protein structure prediction. *Structure*, **9**, 905–916.
49. Engler,L.E., Sapienza,P., Dorner,L.F., Kucera,R., Schildkraut,I. and Jen-Jacobson,L. (2001) The energetics of the interaction of BamHI endonuclease with its recognition site GGATCC. *J. Mol. Biol.*, **307**, 619–636.
50. Poon,G.M.K. and Macgregor,R.B.,Jr (2003) Base coupling in sequence-specific site recognition by the ETS domain of murine PU.1. *J. Mol. Biol.*, **328**, 805–819.
51. Sarai,A. and Takeda,Y. (1989) λ Repressor recognizes the ∼2-fold symmetric half-operator sequences asymmetrically. *Proc. Natl Acad. Sci. USA*, **86**, 6513–6517.

52. Takeda,Y., Sarai,A. and Rivera,V.M. (1989) Analysis of the sequence-specific interactions between Cro repressor and operator DNA by systematic base substitution experiments. *Proc. Natl Acad. Sci USA*, **86**, 439–443.
53. Fields,D.N., He,Y., Al-Uzri,A.Y. and Stormo,G.D. (1997) Quantitative specificity of the Mnt repressor. *J. Mol. Biol.*, **271**, 178–194.
54. Man,T.-K. and Stormo,G.D. (2001) Non-independence of *Mnt* repressor–operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. *Nucleic Acids Res.*, **15**, 2471–2478.
55. Bulyk,M.L., Johnson,P. and Church,G. (2002) Nucleotides of transcription factor binding sites exert inter-dependent effects on the binding affinities of transcription factors. *Nucleic Acids Res.*, **30**, 1255–1261.
56. 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.
57. Vogtli,M., Elke,C., Imhof,M.O. and Lezzi,M. (1998) High level transactivation by the ecdysone receptor complex at the core recognition motif. *Nucleic Acids Res.*, **26**, 2407–2414.
58. Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.
59. Kortemme,T. and Baker,D. (2002) A simple physical model for binding energy hot spots in protein–protein complexes. *Proc. Natl Acad. Sci. USA*, **99**, 14116–14121.
60. Kim,D.E., Chivian,D. and Baker,D. (2004) Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res.*, **32**, W526–W531.
61. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
62. Rajewsky,N., Vergassola,M., Gaul,U. and Siggia,E.D. (2002) Computational detection of genomic *cis*-regulatory modules applied to body patterning in the early *Drosophila* embryo. *BMC Bioinformatics*, **3**, 30.
63. Schroeder,D., Pearce,M., Fak,J., Fan,H., Unnerstall,U., Emberly,E., Rajewsky,N., Siggia,E.D. and Gaul,U. (2004) Transcriptional control in the segmentation gene network of *Drosophila*. *PLoS Biol.*, **2**, 1396–1410.
64. Fraenkel,E., Rould,M.A., Chambers,K.A. and Pabo,C.O. (1998) Engrailed homeodomain–DNA complex at 2.2 Å resolution: a detailed view of the interface and comparison with other Engrailed structures. *J. Mol. Biol.*, **284**, 351–361.
65. Tucker-Kellogg,L., Rould,M.A., Chambers,K.A., Ades,S.E., Sauer,R.T. and Pabo,C.O. (1997) Engrailed (Gln50->Lys) homeodomain–DNA complex at 1.9 Å resolution: structural basis for enhanced affinity and altered specificity. *Structure*, **5**, 1047–1054.
66. Nekludova,L. and Pabo,C.O. (1994) Distinctive DNA conformation with enlarged major groove is found in Zn-finger-DNA and other protein–DNA complexes. *Proc. Natl Acad. Sci. USA*, **91**, 6948–6952.
67. Miller,J.C. and Pabo,C.O. (2001) Rearrangement of side-chains in a Zif268 mutant highlights the complexities of zinc finger-DNA recognition. *J. Mol. Biol.*, **313**, 309–315.
68. Coskun-Ari,F.F. and Hill,T.M. (1997) Sequence-specific interactions in the Tus–*Ter* complex and the effect of base pair substitutions on arrest of DNA replication in *Escherichia. coli*. *J. Biol. Chem.*, **272**, 26448–26456.
69. Frank,D.E., Saecker,R.M., Bond,J.P., Capp,M.W., Tsodikov,O.V., Melcher,S.E., Levandoski,M.M. and Record,M.T.,Jr (1997) Thermodynamics of the interactions of *Lac* repressor with variants of the symmetric *Lac* operator: effects of converting a consensus site to a non-specific site. *J. Mol. Biol.*, **267**, 1186–1206.
70. Grillo,A.O., Brown,M.P. and Royer,C.A. (1999) Probing the physical basis for *trp* repressor-operator recognition. *J. Mol. Biol.*, **287**, 539–554.
71. Boyer,M., Poujol,N., Margeat,E. and Royer,C.A. (2000) Quantitative characterization of the interaction between purified human estrogen receptor α and DNA using fluorescence anisotropy. *Nucleic Acids Res.*, **28**, 2494–2502.
72. Gunasekera,A., Ebright,Y.W. and Ebright,R.H. (1992) DNA sequence determinants for binding of the *Escherichia coli* catabolite gene activator protein. *J. Biol. Chem.*, **267**, 14713–14720.