doi:10.1016/j.jmb.2004.09.029 J. Mol. Biol. (2004) **344**, 59–70

JMB

Available online at www.sciencedirect.com
SCIENCE @ DIRECT®

ELSEVIER

A Simple Physical Model for the Prediction and Design of Protein–DNA Interactions

James J. Havranek†, Carlos M. Duarte† and David Baker*

Howard Hughes Medical Institute and Department of Biochemistry, University of Washington, Seattle, WA 98195 USA

**Corresponding author*

Protein–DNA interactions are crucial for many biological processes. Attempts to model these interactions have generally taken the form of amino acid–base recognition codes or purely sequence-based profile methods, which depend on the availability of extensive sequence and structural information for specific structural families, neglect side-chain conformational variability, and lack generality beyond the structural family used to train the model. Here, we take advantage of recent advances in rotamer-based protein design and the large number of structurally characterized protein–DNA complexes to develop and parameterize a simple physical model for protein–DNA interactions. The model shows considerable promise for redesigning amino acids at protein–DNA interfaces, as design calculations recover the amino acid residue identities and conformations at these interfaces with accuracies comparable to sequence recovery in globular proteins. The model shows promise also for predicting DNA-binding specificity for fixed protein sequences: native DNA sequences are selected correctly from pools of competing DNA substrates; however, incorporation of backbone movement will likely be required to improve performance in homology modeling applications. Interestingly, optimization of zinc finger protein amino acid sequences for high-affinity binding to specific DNA sequences results in proteins with little or no predicted specificity, suggesting that naturally occurring DNA-binding proteins are optimized for specificity rather than affinity. When combined with algorithms that optimize specificity directly, the simple computational model developed here should be useful for the engineering of proteins with novel DNA-binding specificities.

Keywords: protein–DNA interactions; binding specificity; computational modeling; protein design; binding site prediction

Introduction

Sequence-specific interactions between proteins and DNA are critical for the maintenance and expression of genomic information. The ability to modulate the function of existing interfaces and to engineer novel interfaces would be enormously useful for a number of biological and medical applications. Likewise, the ability to predict transcription factor binding sites accurately would have great utility for understanding transcriptional regulatory networks.

An accurate and computationally efficient model of protein–DNA interactions would greatly expand our capability to engineer interfaces. At present, the most dramatic successes in protein–DNA interface engineering have been achieved with phage display of zinc finger proteins.[1–4] Zinc finger proteins are ideal for phage display: the domains are modular and amenable to catenation and rearrangement, and DNA recognition is mediated by a handful of residues, well within the library size limits of phage display.[5] However, many systems of interest involve large, elaborately interconnected protein–DNA interfaces whose complexity cannot be thoroughly explored by phage display.[6] This high complexity would not be a problem for a computational protein design approach that could search sequence spaces much larger than those accessible to genetically encoded libraries.

† J.J.H. & C.M.D. contributed equally to this work.

E-mail address of the corresponding author: dabaker@u.washington.edu

An accurate model of protein–DNA interfaces would also benefit transcription factor binding site prediction. Current methods for predicting binding sites in non-coding regions of genomes use a position-specific scoring matrix representation of the binding preferences of a protein.[7] The elements of this matrix are often determined by time-consuming *in vitro* binding measurements; a rapid method for generating this information computationally would greatly accelerate the prediction process.

A common approach to modeling protein–DNA interfaces for both prediction and design has been to develop a recognition code between amino acids and bases.[8–10] This attractive simplification possesses several drawbacks.[11,12] First, the most successful of such models are specific for a single structural family. Development of such models is not possible for all families due to insufficient data. Second, the models assume a single binding mode for each family, but in reality backbone and side-chain conformational rearrangements undermine this assumption. Finally, although mononucleotide recognition codes can be extended to include internucleotide dependencies that are not captured in these models,[13,14] there is often insufficient data to specify the increased number of parameters required.[15]

Recent advances in protein design methodology suggest an alternative method for analyzing protein–DNA interfaces. The combination of simple physical models of macromolecular energetics and rapid algorithms for sampling side-chain conformations could provide a powerful, quantitative description of protein–DNA interfaces in their entirety. Similar models are already successful in describing the determinants of stability in globular proteins and affinity in protein–protein interactions.[16,17] Such simple physical models are capable of making predictions that extend beyond their input data, yielding novel intermolecular interactions.[18,19] Here, we develop such a model for protein–DNA interactions using a simple physically-based energy function, fixed DNA and protein backbone conformations, and a rotamer-based description of protein side-chain conformation. The success of the model in recovering native amino acid sequences at interfaces suggests that it may be applied to the design of novel protein–DNA interactions. The model is successful in recapitulating protein–DNA binding preferences in known complexes. Prediction of binding preferences in structurally homologous complexes is limited by sensitivity to protein and DNA backbone orientation; further work incorporating backbone sampling or docking techniques will be required before the model is suitable for the prediction of binding preferences on a genomic scale.

Results

We begin this section with a brief overview of our computational model of protein–nucleic acid interactions, and then describe the performance of the model in two tests based on crystal structures of protein–DNA complexes. In the first test, the DNA is held fixed in structure and composition while the identity and/or the conformation of the amino acids at the interface are optimized in searches over either all rotamers of all amino acids or over all rotamers of the naturally occurring amino acid. We investigate the extent to which both the experimentally observed side-chain conformations and the native amino acid identities are recovered (i.e. have lower energy than the alternative conformations or identities). In the second test, the identity of the amino acids at the interface is held fixed and the sequence of the DNA is varied. We investigate the extent to which the DNA sequences predicted to bind with highest affinity correspond to known DNA binding specificities. We then illustrate the complications that can arise when homology models rather than native crystal structures are used for specificity calculations, which indicate clear avenues for future work to improve the model. Finally, we investigate in the context of the model the long-standing issue of whether transcription factors are optimized for specificity or affinity.

Overview of computational model

The model is described in detail in Materials and Methods; here we give only a brief overview. An all-atom description of both the DNA and protein is used. The backbone of the protein and DNA are held fixed, the amino acid side-chains are allowed to sample all conformations in the Dunbrack backbone-dependent rotamer library,[20] and the bases are allowed to vary in identity but not in conformation. The energy is computed for all possible rotamers for fixed DNA sequence (test I) or for all possible DNA sequences for fixed amino acid composition (test II). The rapidly computable energy function is similar to that used in previous prediction and design work on proteins and protein–protein interactions. The dominant terms are an orientation-dependent hydrogen bonding term, an implicit solvation model, and a Lennard–Jones potential; hence the lowest energy complexes identified tend to be rich in side-chain–base hydrogen bonds, have few buried polar atoms that do not make hydrogen bonds, and to be relatively well packed. The overall amino acid composition is controlled through reference energies for each amino acid, which incorporate effects not included in the model such as long-range electrostatics (which favors positively charged residues at protein–DNA interfaces).

Test I: recovery of native protein sequences and conformations for the test set

For each amino acid at the protein–DNA interface in a large test set of DNA-binding proteins, the energy of each rotamer for all 20 amino acids was determined and the amino acid with the

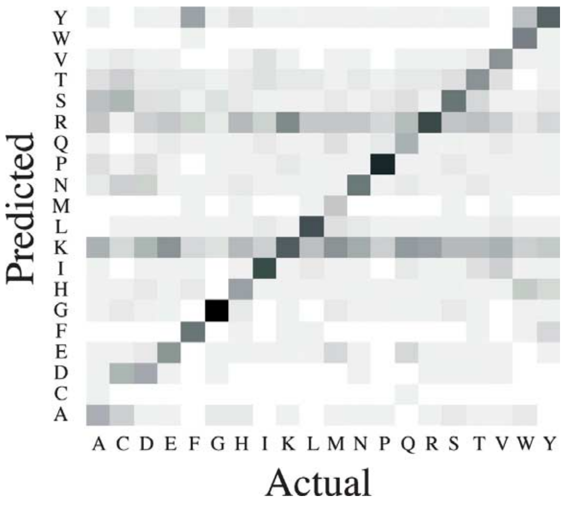

Figure 1. Amino acid recovery at protein–DNA interfaces. The columns represent the native amino acids from the test set and the rows represent the amino acids selected by our model. Single letter amino acid codes identify each row and column. Boxes are shaded by grayscale ranging from 0% recovery (white) to the maximum observed recovery of 76% for glycine (black).

Table 1. Side-chain conformation recovery for test set

Amino acid	χ1 (%)	χ2 (%)	χ3 (%)	χ4 (%)
ALA				
CYS	90.0			
ASP	85.7	50.0		
GLU	70.2	55.3	31.9	
PHE	100.0	89.7		
GLY				
HIS	80.8	61.5		
ILE	88.9	79.6		
LYS	74.6	53.2	41.3	28.6
LEU	87.5	69.6		
MET	70.6	70.6	47.1	
ASN	82.7	75.0		
PRO	66.7			
GLN	76.2	59.5	35.7	
ARG	75.6	59.5	33.6	20.6
SER	63.6			
THR	87.9			
VAL	78.3			
TRP	100.0	88.2		
TYR	84.6	76.9		
Total	79.2	64.8	36.9	24.5
Protein total	83.4	69.7	40.5	26.7

The definition of a correct χ angle is cumulative. Thus, for χ3 to be assessed as correct, both χ1 and χ2 must also be correct.

lowest-energy rotamer was selected and compared to the naturally occurring amino acid. As indicated in Figure 1 (and see Table 5 in the Supplementary Data), the lowest-energy amino acid (Figure 1, vertical axis) was very often the naturally occurring amino acid (Figure 1, horizontal axis). For example, the vertical stripe in Figure 1 above the R indicates the distribution of lowest-energy amino acids for sites that have arginine in the native complexes; it is evident that the most commonly predicted amino acids at such sites are arginine and the chemically similar lysine. The overall native sequence recovery rate of 42.3% is comparable to rates seen for similar experiments conducted on single-domain proteins (52% for buried positions, 26% for all positions).[16]

The native amino acid was the most frequently identified substitution for all amino acids except cysteine, methionine and glutamine. Cysteine was recovered most frequently as aspartate or serine. This likely reflects the frequent role of cysteine in coordinating zinc in protein-DNA complexes; as metals are not included in our calculations, these cysteine residues were replaced with amino acids that can hydrogen-bond to other liganding residues. Methionine and glutamine were recovered as lysine, probably because of the favorable reference energies for lysine, which stems from the high abundance of this amino acid at protein-DNA interfaces due to the overall negative charge of DNA. It is notable that besides the frequent choices of lysine or arginine, a large portion of the "incorrect" predictions consists of conservative substitutions (e.g. tyrosine for phenylalanine).

Although the energy function parameterization procedure was optimized for the recovery of native amino acid sequences, the resultant force-field also reproduces native side-chain conformations with great fidelity. The extent of recovery of the native side-chain χ angles for each amino acid in the test set is shown in Table 1. The recovery percentages are comparable with those obtained from repacking a set of over 300 high-resolution, non-redundant monomeric protein structures (a subset of the data set from Lovell *et al.*[21]) with a similar energy function optimized to recover the native sequences of proteins.

Water-mediated hydrogen bonds are a significant contributor to sequence-specific recognition at many interfaces. Water molecules were not, however, included in the calculations on the test set or in the generation of parameterization weights from the training set. The end goal of these experiments is the design of novel protein–DNA interfaces and for these purposes results in the absence of water are more relevant, since it is unlikely that water molecules will occupy the same positions in different interfaces.

Sequence recovery for E-DreI endonuclease

An important application for protein–DNA interface design will be to design novel highly specific endonucleases. To illustrate the recovery of native sequences in more detail, we describe the results with the E-DreI endonuclease–DNA interface, which is the largest unique (i.e. non-palindromic) and specific DNA recognition site in our dataset. E-DreI is a computationally designed chimera of the homing endonucleases I-DmoI and I-CreI.[18] During the original E-DreI design procedure, amino acids at the protein–protein interface between the I-DmoI and I-CreI subunits were altered, but the protein–DNA interfaces were unaltered and not included in the computational design process. As a result, E-DreI specifically recognizes a 22 base-pair DNA sequence that is a combination of the native I-DmoI and I-CreI recognition sites.

Amino acids at DNA interfaces can interact with DNA directly through hydrogen bonds and van der Waals contacts, indirectly through water-mediated hydrogen bonds, or through a combination of both.[22] In sequence recovery experiments with E-DreI, all of the amino acid residues (6/6; Table 2) with only direct native contacts to the DNA were recovered (Figure 2(A)). There is a drop in fidelity for amino acids that have a combination of direct and water-mediated contacts (8/11) and poorer recovery for amino acids with only water-mediated contacts (3/7). Several of the amino acids that interact with the DNA through water-mediated contacts are replaced by larger charged or polar amino acids. For example, the aspartic acid residue at position 172 is replaced by a glutamic acid residue. The conformation of the glutamic acid side-chain positions it such that the charged moiety resides in the position occupied by a water molecule in the native conformation (Figure 2(B)). This allows the side-chain to replace the water-mediated contact with a direct hydrogen-bonding contact. The substitution of direct amino acid-base contacts for water-mediated contacts in redesign calculations will likely be a recurrent feature for our physical model. This has the potential to yield proteins with higher affinities or greater specificities for their cognate DNA sequences.

Table 2. Sequence recovery for E-DreI endonuclease

Residue	Contact type (Dir/H_2O/Comb)	Substitution
Y25	Dir	NAT
Y29	Dir	NAT
R37	Dir	NAT
T76	Dir	NAT
Y130	Dir	NAT
R167	Dir	NAT
G31	H_2O	NAT
Q70	H_2O	L
E79	H_2O	NAT
S137	H_2O	Q
T139	H_2O	R
Y163	H_2O	NAT
D172	H_2O	E
R33	Comb	NAT
E35	Comb	NAT
D75	Comb	R
R77	Comb	NAT
R81	Comb	NAT
Q123	Comb	NAT
K125	Comb	NAT
N127	Comb	NAT
Q135	Comb	NAT
Q141	Comb	K
R165	Comb	Y

Dir—direct contact, H_2O—water-mediated contact, Comb—combination of direct and water-mediated contacts, NAT-native amino acid.

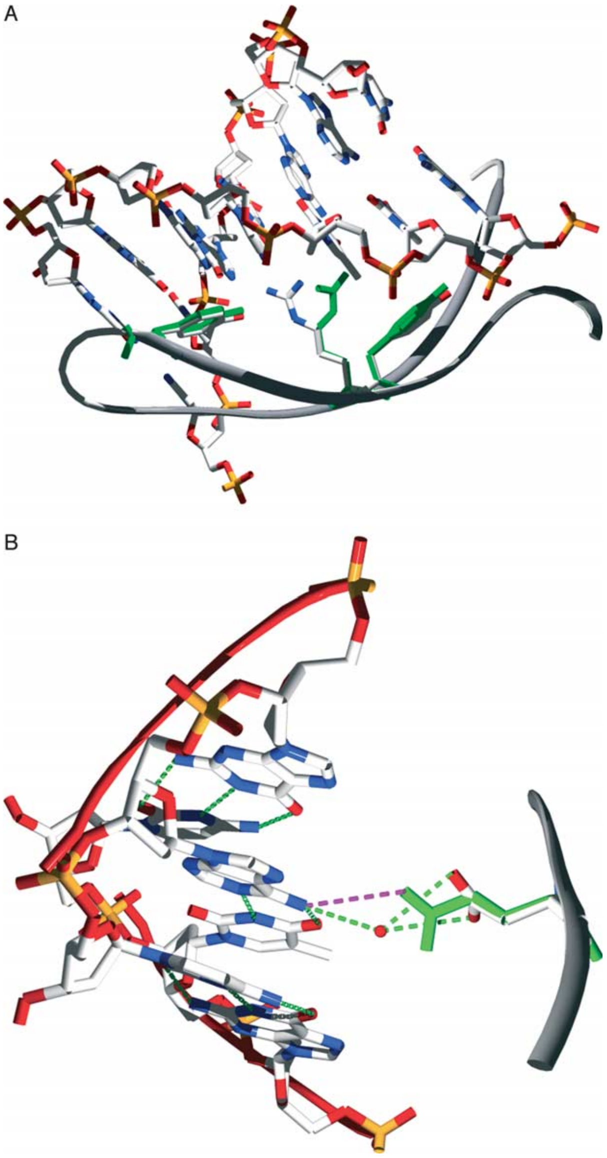

Figure 2. Examples of sequence redesign for E-DreI endonuclease. (A) Recovery of amino acids participating in direct contacts. Three representative amino acids from E-DreI that make direct contacts to DNA and were correctly recovered by the model are shown: (from left to right) Tyr29, Arg37, and Tyr25. For the DNA all atoms are shown. Predicted conformations for the side-chains are shown in green, superimposed on the side-chain conformations of the crystal structure 1MOW.[18] (B) Redesign of a water-mediated contact. Asp172 forms a water-mediated contact with DNA in the crystal structure. Green dashes indicate hydrogen bonds from the native structure; the red sphere represents the native water. The designed replacement is shown in green; red dashes indicate the designed hydrogen bond that replaces the original DNA–water interaction.

Test II: recovery of DNA binding specificity

EcoRI endonuclease

We evaluated the ability of our model to recover native base sequences in protein–DNA interfaces.

This test is complementary to the previous example, in which recovery of native amino acid sequences was the objective. We tested the ability of our model to identify sequence preferences for the EcoRI restriction endonuclease system. By necessity, restriction enzymes display a high level of specificity for their recognition sequences under physiological conditions.[23] Models were generated for variants of the EcoRI–DNA complex in which the wild-type 6 bp recognition sequence was replaced with all possible 6 bp palindromes (Figure 3(A)) (M. Horvath *et al.*, unpublished results). The amino acid side-chains in the interface were built with ideal bond and angle geometries, with side-chain conformations taken from a rotamer library. Thus, the models did not include coordinates from the crystal structure for amino acid side-chains, eliminating a possible source of bias. The side-chain conformations of amino acids in the protein–DNA interface were optimized using a Monte Carlo rotamer search, and the free energy of the resulting complexes was evaluated. Each calculation was repeated ten times because of the stochastic nature of the rotamer search.

The complex with the native recognition sequence was identified correctly as the lowest in energy (Figure 3(B)). There is an energy gap of 2.6 kcal/mol between the lowest energy for the complexes containing the native recognition sequence and the lowest energy for any other complex (1 cal=4.184 J). In no case did any of the 63 non-native complexes have a lower energy than that of the native complex. The DNA sequences of the lowest energy non-native complexes are similar to the native sequence (GAATTC), with the three lowest differing from the native at only one of the three independent positions each (GAGCTC, TAATTA, CAATTG). No sequence differing from the native at all three positions is found within the ten lowest energy non-native complexes.

Prediction of zinc finger DNA-binding specificity

We selected the zif268 zinc finger transcription factor as a system for evaluating the model's ability to predict binding specificity for complexes whose structures may only be inferred from homologues. The zif268 transcription factor consists of three zinc finger domains, each of which primarily recognizes a 3 bp binding site (Figure 4(A)). This protein has been subjected to combinatorial mutagenesis, and the resultant libraries selected for altered binding specificity by phage display.[24–26] The interacting partners in each case differ from those in the structurally characterized native complex in both their protein and DNA sequences.[27] The ability

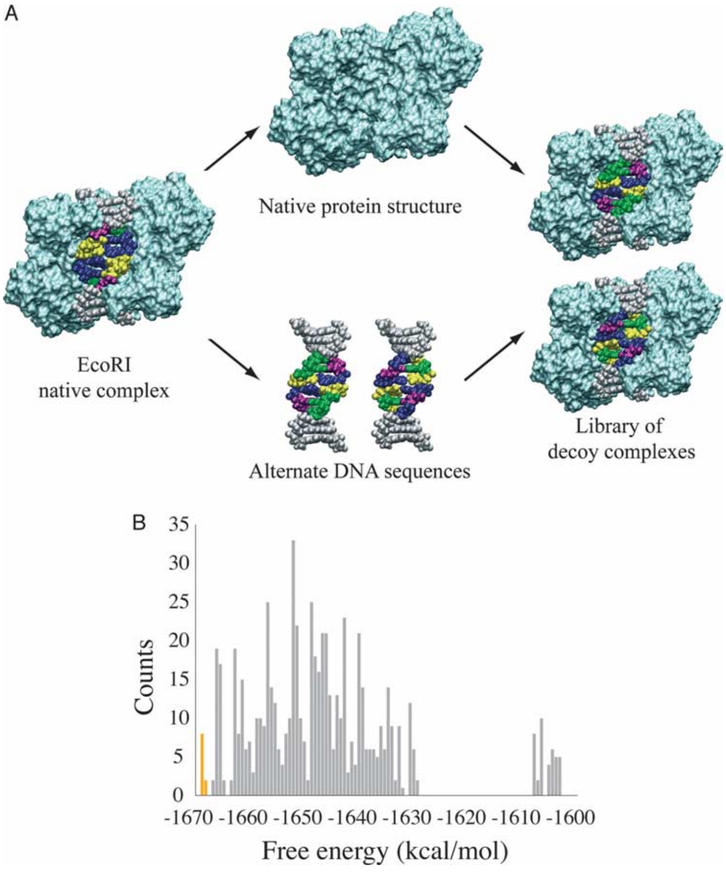

Figure 3. Computational assessment of DNA-binding preferences of EcoRI endonuclease. (A) A library of EcoRI–DNA complexes. Atomic coordinates for the cognate DNA duplex were taken from the crystal structure 1CKQ (M. Horvath *et al.*, unpublished results). Models for all possible 6 bp palindromic variants of the recognition sequence were built using standard bond and angle geometries from CHARMM27[39] and torsion angles retained from the crystal structure (a total of 64 duplexes). Replacement of the DNA in the crystal with each of these alternate DNA binding sites gives rise to a library of protein–DNA complexes. The free energy of each complex in the library described above was evaluated after the conformations of the amino acid side-chains in the interface were determined using a Monte Carlo-based rotamer search. Each calculation was repeated ten times to account for the stochastic nature of the search. (B) Distribution of minimized energies. Each of the 640 energies resulting from the Monte Carlo conformational searches is shown. Energies corresponding to those obtained from the ten complexes with the native recognition sequence are shown in yellow.

to identify binding sites for proteins using structural information from homologues is important if models of protein–DNA interactions are to contribute new information, rather than rationalize what is already known.

We generated models for all possible 3 bp binding site sequences for the first and second zinc finger domains of zif268. The sequences for a series of zinc finger mutants selected to bind to all possible 5′-GNN-3′ binding sites[24,25] were threaded onto the protein backbones of both the first and second zinc finger domains of zif268 (Table 3). The side-chain conformations for these amino acids were predicted using a Monte Carlo rotamer search in complex with each of the recognition sequences, and the energies evaluated.

As shown in Table 4, the experimental results were reproduced reasonably well when finger I was used as a template. The model performs well for mutants that bind their recognition sequences with high affinity and specificity (no experimentally observed cross-reactivity). The model selects the correct sequence among the top four (of 64) for each mutant that exhibits no cross-reactivity and a K_D below 10 nM. The success of the model for each mutant is likely to depend on how well the combination of separate interactions in the interface are modeled. For instance, the interaction between the 5′ G of the recognition site and the Arg at the +6 helical position is modeled well. Likewise, residues at position +3 of the recognition helix are known to interact with the middle base of the recognition sequence. The four mutants whose targeted DNA sequences are ranked lowest, and only these mutants, have acidic residues at the +3 helical position. It is possible that the interactions that

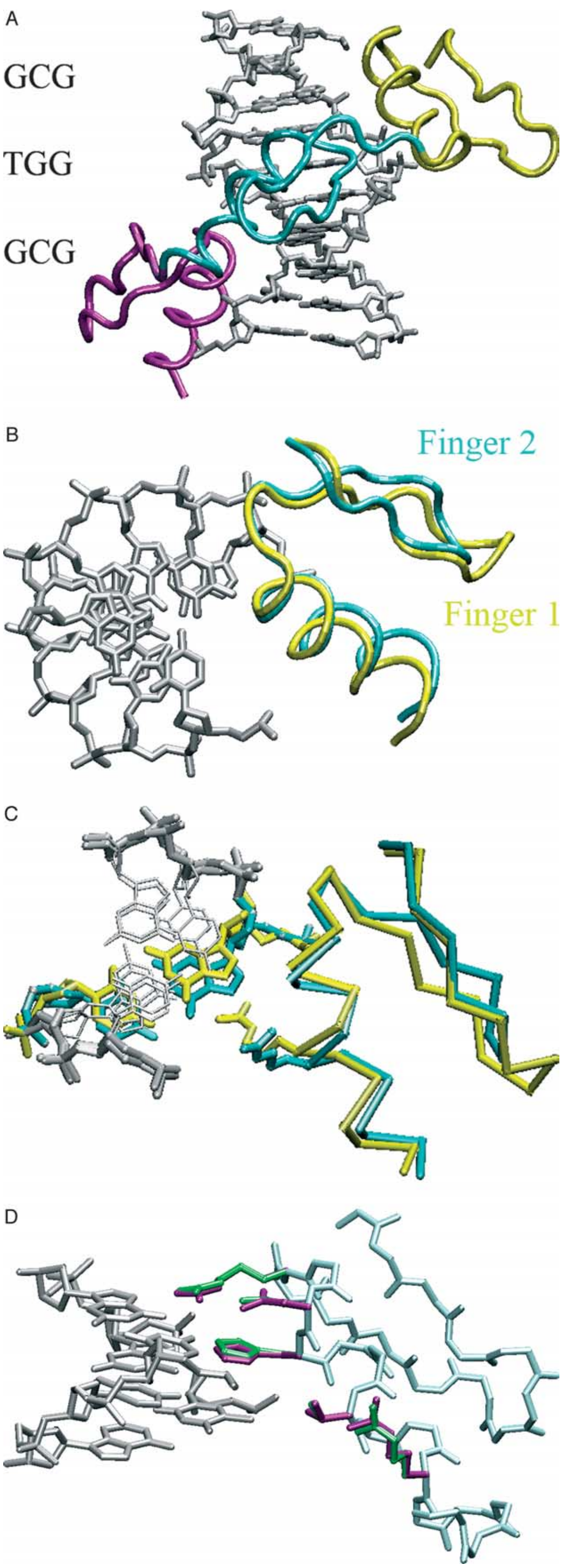

Figure 4. Conformation, specificity and sequence redesign calculations for the zif268 transcription factor. (A) Zif268 transcription factor. The three zinc finger domains are shown in yellow, cyan, and magenta. The recognition sequences for each are shown at left. (B) Alternate zinc finger binding modes. The recognition sequences for zinc fingers 1 and 2 from zif268 were structurally aligned, and the resulting superposition of protein domains is shown. (C) Backbone dependence of arginine-guanine interaction. The first and second zinc finger domains of zif268 and their three base-pair recognition sequences are shown aligned by their DNA backbone atoms. Finger one is shown in yellow and finger two in cyan. The backbone orientation of the 5′ base-pair of the recognition sequence of finger two is distorted relative to that of the binding site for finger one. When a guanine base is modeled in this position of the finger two binding site (shown in cyan; the analogous interaction in finger is shown in yellow) it moves too close to the protein backbone to form favorable interactions with an arginine residue at position +6 of the recognition helix without clashing. To remove this unfavorable interaction, our model removes the arginine side-chain from the interface, allowing it to interact with the phosphate backbone instead. (D) Side-chain conformation recovery for wild-type zif268 with cognate DNA. Crystallographically determined side-chain conformations for amino acids involved in specific base-pair recognition are rendered in magenta. The side-chain conformations selected by a Monte Carlo rotamer-based search using our physical model are rendered in green.

Table 3. Calculated binding preferences for zinc finger mutants from Segal *et al.*[24]

Target sequence	Amino acid sequence[a]	Experimental cross-reactivity	Experimental K_D (nM)	Calculated rank/(energy gap)[b]	
GGA	SQRAHLER	No	3	1	1.16
GGG	SRSDKLVR[c]	No	6	1	0.8
GGT	STSGHLVR[c]	Some	15	1	0.7
GAA	SQSSNLVR[c]	No	0.5	3	(0.79)
GAG	SRSDNLVR[c]	Some	1	3	(1.37)
GAT	STSGNLVR	No	3	3	(1.62)
GGC	SDPGHLVR[c]	No	40	4	(1.23)
GAC	SDPGNLVR[c]	No	3	4	(2.85)
GTT	STSGSLVR[c]	Yes	5	7	(1.97)
GTC	SDPGALVR	No	40	9	(1.97)
GCG	SRSDDLVR[c]	Yes	9	9	(2.59)
GTA	SQSSSLVR	No	25	11	(2.01)
GCA	SQSGDLRR	Yes	2	12	(1.51)
GCC	SDCRDLAR	No	80	15	(2.65)
GTG	SRSDELVR[c]	Yes	15	18	(3.65)
GCT	STSGELVR[c]	Some	65	26	(3.77)

[a] Amino acid sequence beginning with position −2 in the recognition helix.

[b] For correctly identified sequences (ranked first), this is the energy gap between the correct sequence and the closest non-specific sequence. When the correct sequence is not identified as the lowest energy, the energy is shown in parentheses and denotes the difference in energy between the correct sequence and the sequence incorrectly identified as lowest in energy.

[c] Amino acid sequences were modified from the phage-selected sequence (see Segal *et al.*[24]).

these residues participate in are mediated by water or are otherwise treated poorly by our model.

Challenges facing specificity prediction from homology models

The experimental results were less well recapitulated when finger 2 was used as a template. Examination of the predicted structures revealed that a common interaction motif between arginine at position +6 of the recognition helix and guanine at the 5′ position of the 3 bp recognition sequence could not be made in the context of the second zinc finger. We investigated the structural differences between the two zinc finger domains that determined whether the Arg–G interaction could be formed. Differences in protein backbone conformation were not the cause; the C^{α} atoms of the DNA-contacting residues (helical positions −2 to +6) of the two fingers can be superimposed with an rmsd of 0.3 Å. We tested whether a difference in the relative orientation between the protein and DNA backbones could account for the discrepancy (see Figure 4(B)).[5] However, even when finger 2 is forced into the same orientation relative to the DNA as finger 1, rotameric side-chain conformations of Arg at position +6 still fail to make the expected interactions with G at the 5′ position of the recognition sequence. Instead, the inability to form the G–Arg interaction is due to a distortion in the DNA backbone at this position. The 5′ position of the recognition sequence is occupied by a T in the wild-type sequence. This base is rotated outward from the major groove towards the protein, and stacks with a His at the helical position 3. Our DNA models retain this outward rotation and, as a result, guanine bases modeled at this site cannot be contacted by the Arg side-chain at position +6 (Figure 4(C)).

Comparison with zinc finger binding specificity data from Bulyk et al.*[26]*

In another experimental study, the second zinc finger of zif268 was subjected to combinatorial mutagenesis and selected for altered binding specificity. The specificities of the resulting mutants

Table 4. Calculated binding preferences for zif268 mutants

			Ordinal rank[a] (out of 64)	
Zif268 variant[b]	Highest-affinity sequence	K_d^{app} (nM)[c]	Finger 2	Finger 1
Wild-type	TGG	3.0±5.7	1	4
RGPD	GCG	17±4.0	28	3
REDV	GCG	11±4.3	9	3
LRHN	TAT	6.3±1.6	2	13
KASN	AAT	250±28	61	49

[a] Since the library of structures has 64 members, a rank order less than 32 (a rank of one denoting the most favorable interaction) for a given protein sequence represents a preference for the DNA sequence over the library as a whole.

[b] The sequences of these proteins from positions −1 to +9 of the recognition helix are: wild-type, (RSDHLTTHIR); RGPD, (RGPDLARHGR); REDV, (REDVLIRHGK); LRHN, (LRHNLETHMR); KASN, (KASNLVSHIR).

[c] From Supplemental Material for Bulyk *et al.*[26]

were measured *in vitro* by double-stranded DNA microarray binding assays.[26] We calculated the specificities for the wild-type zif268 sequence and for each of the mutants in the contexts of both the first and second finger domains (Table 4). Not surprisingly, the two mutants with a +6 helical position residue of arginine and guanine at the 5′ position of their recognition sequences (mutants RGPD and REDV) were modeled well in the context of the first, but not the second, zinc finger domain. Similarly, the wild-type protein and one of the mutants (LHRN) were modeled well in the context of the second finger domain, and were modeled less well in the first finger context. The preferred sequence for the LHRN mutant (TAT) is similar to the wild-type sequence (TGG). We note that the side-chain conformations for the wild-type sequence are in good agreement with the crystal structure (Figure 4(D)). The fourth mutant (KASN) was modeled poorly in the context of either backbone orientation, perhaps because this domain was experimentally found to be both the least specific and weakest binding mutant in the set, with an affinity two orders of magnitude weaker than wild-type (Table 4).[26] Alternatively, the complex between this mutant and its preferred sequence (AAT) may be unlike those adopted by the native zif268 zinc finger domains.

Optimization of affinity *versus* specificity

We redesigned five amino acid positions in the second zinc finger domain of zif268 in complex with each of the 64 alternative recognition sequences used to assess specificity above. For each of these binding sites, our model selected one of only five distinct protein sequences. However, none of these sequences was found to be specific: the largest difference between the lowest and second-lowest energy complexes for the redesigned domains was 0.35 kcal/mol. This is in contrast to the wild-type protein sequence at this position, which is calculated to have a 2.4 kcal/mol energy gap between the native and the lowest-energy non-native DNA sequence. Interestingly, the calculated binding affinities for all but two of the 64 redesigned complexes are more favorable than for the wild-type complex.

Discussion

Performance and composition of the physical model

We have presented a simple, physically based model that is capable of describing the determinants of affinity and specificity in protein–DNA interfaces. This model reproduces amino acid sequences and conformations at protein–DNA interfaces with high fidelity. Amino acids that are not recovered by our model are generally replaced by conservative substitutions, or by mutations that simultaneously replace an amino acid and a bound water molecule. For fixed protein sequences, our free energy function is capable of discriminating native DNA-binding sites from libraries of decoys. The model also shows some ability to identify preferred binding sites in the homology modeling problem, in which an inferred structure is used as the model for a putative binding interaction.

The energetic components that make up our physical model for protein–DNA interactions are essentially identical with those used to describe the structures of proteins. Future improvements in our model will focus on two differences between the interactions that stabilize proteins and those that dictate affinity and specificity in protein–DNA interfaces: electrostatics and water-mediated interactions. Because of the highly charged nature of the DNA backbone, we tested the effect of including a simple description of electrostatics in our model.[28] Surprisingly, little improvement was gained, and this term was ultimately excluded from the free energy function. Preliminary results with a more accurate electrostatics model are promising, and we are pursuing further this avenue for improvement. Instead, the effect of the electrostatic environment is incorporated through the reference energies for the amino acids, which represent the average free energy of the amino acid at an interface and control the overall amino acid composition. In fact, the most significant differences in the weights determined for protein–DNA interfaces compared to those obtained for protein structure and protein–protein interfaces are for the reference energies. (The ratio of the weights for the Lennard–Jones, solvation, and hydrogen bonding energy terms is 1 : 0.8 : 1.5 for proteins and 1 : 0.6 : 1.8 for protein–DNA interfaces.) Because the polyphosphate backbone of DNA is uniformly negatively charged, the reference energies can account for the average electrostatic environment at protein–DNA interfaces. Water-mediated contacts between amino acid residues and DNA bases are not included in our model. However, it is possible to include potential water molecules in side-chain repacking calculations by expanding current rotamer libraries to include hydrated side-chains.[40] This approach is under investigation.

Implications for the computational design of protein–DNA interfaces

Protein–DNA interfaces are attractive and challenging design targets. Small changes in protein–DNA interactions can have profound biological effects, yet the large number of competing binding sites *in vivo* implies that any successfully designed protein must possess exquisite specificity for its desired target DNA sequence. The capability of our model to both recover native-like protein sequences at DNA interfaces and to discriminate preferred DNA-binding sites for a fixed protein sequence suggests that it should be possible to design novel protein–DNA interfaces and to assess their binding specificity. The model is well suited for current design algorithms: all of the energy terms are pairwise factorable and can be evaluated rapidly. Because water-mediated interactions have been neglected, we expect that interfaces designed using our model will be dominated by direct contacts. Nevertheless, our model holds distinct advantages over recognition codes, which cannot describe conformational rearrangements or detailed atomic packing and are limited to a structural family, and offers greater generality and library complexity than phage display, which is limited by the size of genetically encoded libraries. Additionally, the output of these calculations explicitly includes proposed structural models that can be evaluated both visually and in terms of detailed analysis of the energetic terms at the atomic level. This confers the ability to process and inspect multiple proposed designs for the same DNA interface.

The outlook for the computational design of protein–DNA interfaces is particularly encouraging, because experience shows that the computational design of proteins and protein–protein interfaces is possible even when prediction of the behavior of analogous natural counterparts is not. Proteins with novel topologies have been designed with great accuracy, despite the continued recalcitrance of the high-resolution structure prediction problem. While protein–protein docking is likely to remain a challenging field of research for some time, novel protein–protein interfaces have been designed and characterized.

The design problem is easier for at least two reasons. First, unlike natural proteins, designed proteins are selected to satisfy simple criteria such as stability and affinity. Consequently, it is possible to "over-design" these proteins, resulting in a successful design even if only part of the predicted stability or affinity is realized. Conversely, the stability of natural proteins is often marginal, and this may limit the ability of current models to identify minimum energy conformations. Second, natural proteins and macromolecular complexes can contain unusual features, such as the outwardly rolled thymine base in the zinc finger example described above. Other examples include bond lengths and angles that deviate from ideal values, side-chains that adopt uncommon (non-rotameric) conformations, and amino acids that appear in unusual environments, such as buried polar residues. All of these features hinder the structural prediction of proteins and protein assemblies. However, design algorithms exclude these complications by construction, and restrict themselves to regions of sequence and conformational space where they perform best, and in which success is most likely.

Specificity *versus* affinity in protein–DNA interfaces

The most striking characteristic of protein–DNA interactions is the exquisite specificity with which proteins recognize their preferred binding sites despite the bewildering number of alternate sites presented *in vivo*. In agreement with experimental evidence,[24,29,30] we observe a trade-off between affinity and specificity. Zinc finger proteins computationally designed for affinity are predicted to bind to their targets more tightly than the wild-type protein binds its target. However, the designed sequences are also predicted to be significantly less specific than the wild-type protein. This suggests that the naturally occurring protein has been, to some extent, optimized for specificity rather than affinity. Experimentally, negative design against non-specific sites is incorporated through selection in the presence of competitor DNA (oligonucleotides *in vitro* or genomic DNA *in vivo*). It is likely that some form of negative design will be required to design DNA-binding proteins with a high level of sequence specificity.[19,31]

Prediction of protein–DNA interactions

The ability of our model to identify preferred DNA-binding sites for a given DNA-binding protein from a library of competitors suggests the potential for *in silico* annotation of DNA-binding proteins. Putative structures of transcription factors can be generated by homology modeling and screened rapidly against large libraries of potential binding sites. This strategy bypasses the time-consuming experimental steps generally required to construct a position-specific scoring matrix for scanning potential binding sites. Although the model is most successful when native protein–DNA structures are used, the zinc finger examples demonstrate that the model has some predictive power even when both the protein and DNA sequences are varied in an interface. The model is not sensitive to the replacement of crystallographically determined side-chain positions by rotameric approximations: native side-chain conformations were not included in any of the rotamer searches. However, the model is sensitive to errors in the relative orientation of the protein and DNA backbones. We speculate that the zif268 mutant proteins for which our model performed poorly involve alternative binding modes or sequence-specific DNA backbone distortions. Further improvements to our model will include the consideration of multiple protein–DNA binding modes, through either the use of multiple crystal structures of alternative complexes, or by combining our model with computational docking protocols. Finally, we note that our model should be complementary to existing profile-based methods for binding site prediction, allowing for the combination of both physically and statistically based methods within a single prediction scheme.

Materials and Methods

Datasets

Crystal structures of protein–DNA complexes were selected from the Nucleic Acid Database.[32] Structures were screened by protein sequence homology to avoid biasing the results towards commonly studied structures. In cases of redundant complexes, the highest-resolution structure was selected. In certain instances, homologous proteins were retained when the DNA sequence or bound conformation differed considerably. Hydrogen atoms were added to the structures, assuming standard bond lengths and angles. Hydroxyl hydrogen positions were determined by optimization of a rotamer-based description of the hydrogen bond network.[17]

The free energy function

During these calculations the free energies of all-atom models for all sampled structures were evaluated using a nine-term function that is described in detail in the Supplementary Material. Briefly, these terms are: a Lennard–Jones potential used to model attractive and repulsive van der Waals atomic forces, an implicit solvation term based on the model developed by Lazaridis & Karplus,[33] an orientation-dependent hydrogen bonding term derived from a statistical analysis of high-resolution protein structures,[34] a pair interaction term that crudely models electrostatic interactions between amino acid side-chains, a backbone torsional term that accounts for differences in the local structure propensities of the amino acids, and 20 reference energies that control the overall amino acid composition. The free energy function is essentially the same as that used previously for protein design calculations,[35] with the exceptions of the values of the 20 reference energies, the derived weights for each term and an augmentation of the hydrogen bonding term. The protein atom types used for the Lennard–Jones, solvation, and hydrogen bonding terms are the same as those defined previously. The van der Waals radii of all protein heavy-atoms were determined based on observed interatomic distances between atom types in high-resolution structures. Atom types for the DNA, with the exception of P, were assigned from the set of existing protein atom types. Assignments for each DNA atom were to the protein atom type with the most similar chemical environment hybridization state (see Supplementary Material). The van der Waals radius for the P atom type was taken from CHARMM27, and the solvation parameters were copied from the S atom type. Deprotonated nitrogen atoms were included as hydrogen bond acceptors with the same distance and angular energy dependence as *sp*2 hybridized oxygen atoms. Rather than defining an acceptor base atom (such as a carbonyl C atom for a carbonyl O acceptor) for the ring nitrogen atom, a virtual atom was used whose position was the average of the two heavy-atoms bonded to the ring nitrogen atom, and the minimum of the angular component of the hydrogen bond potential was shifted to 180°. This angular term enforces linearity of the hydrogen bond from the virtual base through the acceptor to the shared hydrogen atom, and imposes a penalty for out-of-plane hydrogen bonding geometries. The parameters for the hydrogen bonding potential were derived from protein structures and are consistent with the results of quantum mechanical calculations on small molecule models;[36] a potential derived directly from a dataset of protein–nucleic acid complexes showed similar levels of recovery of native amino acid residues at protein–RNA interfaces.[37]

Parameterizing the energy function on protein–DNA interfaces

The weights for the various components of the free energy function were optimized to recover native protein sequences for positions at the interface with DNA. Interface positions are defined as all those within 6.0 Å of any DNA atom. Each energy component in the free energy function was evaluated for all rotamers of all amino acids, using a backbone-dependent library.[20] All other amino acids in the interface were held in their native conformation. A variable-metric optimization method was used to determine weights for the energy components of the free energy function that maximized the recovery of native amino acid sequences.[38] The dataset was divided into five subsets of equal size. Families of related proteins, as determined by sequence homology, were segregated to single subsets. Each subset was evaluated using weights determined from the other four. Though the weights were nearly identical for each subset, this procedure facilitated the independent testing of each one.

Protein–DNA interface redesign

Total protein–DNA interface redesign calculations were performed using the same definition for the interface given above. Sequence recovery of the naturally occurring amino acids at the DNA interface for each complex of the dataset using independently derived parameter weights was used to assess the accuracy of the force-field. In these calculations, side-chains at each sequence position along the DNA interface were substituted one-by-one by all amino acids in all the rotamer conformations in a backbone-dependent library.[20]

Native side-chain conformation recovery

The ability of the force-field to recover native side-chain conformations for complexes from the dataset was also assessed. In these experiments, all of the interface side-chains for each complex were repacked simultaneously. Side-chain dihedral angles were deemed to be correct if they were within 40° of the crystallographically determined values.

Binding site prediction

The specificity of DNA-binding proteins was calculated by evaluating the free energy function for the protein in complex with native DNA binding site and with a library of alternate DNA binding sites. Alternate DNA molecules were constructed by keeping the conformations of nucleotides that were the same as the native DNA fixed, and building alternate bases onto the native phosphate backbone using standard bonds and angle, and preserving the χ dihedral angle from the native DNA. The free energy of each structure was evaluated after optimization of side-chain conformations for all amino acid residues in the interface using a Monte Carlo-based rotamer search.[16]

Availability

The computational module was implemented as a protein–nucleic acid interaction module in the ROSETTA software package, and will be available at no cost to academic users† in a manner similar to previously released modules.

† www.bakerlab.org

Acknowledgements

We thank members of the Baker laboratory for many helpful discussions. J.J.H. is a fellow of the Jane Coffin Childs Memorial Fund for Medical Research. C.M.D. is fellow of the National Science Foundation and has been supported also by a National Institutes of Health Interdisciplinary Genomic Sciences training grant. We thank Barry Stoddard and Gabriele Varani for their encouragement and insights, and Barry Stoddard for detailed discussions of the Dre protein–DNA interface. This work was supported also by the Howard Hughes Medical Institute.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2004.09.029

References

1. Segal, D. J., Goncalves, J., Eberhardy, S., Swan, C. H., Torbett, B. E., Li, X. L. & Barbas, C. F. (2004). Attenuation of HIV-1 replication in primary human cells with a designed zinc finger transcription factor. *J. Biol. Chem.* **279**, 14509–14519.
2. Tan, S. Y., Guschin, D., Davalos, A., Lee, Y. L., Snowden, A. W., Jouvenot, Y. *et al.* (2003). Zinc-finger protein-targeted gene regulation: genomewide single-gene specificity. *Proc. Natl Acad. Sci. USA*, **100**, 11997–12002.
3. Ordiz, M. I., Barbas, C. F. & Beachy, R. N. (2002). Regulation of transgene expression in plants with polydactyl zinc finger transcription factors. *Proc. Natl Acad. Sci. USA*, **99**, 13290–13295.
4. Kim, J. S. & Pabo, C. O. (1998). Getting a handhold on DNA: design of poly-zinc finger proteins with femtomolar dissociation constants. *Proc. Natl Acad. Sci. USA*, **95**, 2812–2817.
5. Pabo, C. O., Peisach, E. & Grant, R. A. (2001). Design and selection of novel Cys(2)His(2) zinc finger proteins. *Annu. Rev. Biochem.* **70**, 313–340.
6. Jeltsch, A., Wenz, C., Wende, W., Selent, U. & Pingoud, A. (1996). Engineering novel restriction endonucleases: principles and applications. *Trends Biotech.* **14**, 235–238.
7. Bulyk, M. L. (2004). Computational prediction of transcription-factor binding site locations. *Genome Biol.* **5**, 201.
8. Choo, Y. & Klug, A. (1997). Physical basis of a protein–DNA recognition code. *Curr. Opin. Struct. Biol.* **7**, 117–125.
9. Mandel-Gutfreund, Y. & Margalit, H. (1998). Quantitative parameters for amino acid–base interaction: implications for prediction of protein–DNA binding sites. *Nucl. Acids Res.* **26**, 2306–2312.
10. Benos, P. V., Lapedes, A. S. & Stormo, G. D. (2002). Is there a code for protein–DNA recognition? Probab(ilistical)ly. *BioEssays*, **24**, 466–475.
11. Pabo, C. O. & 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.
12. Miller, J. C. & 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.
13. Man, T. K. & Stormo, G. D. (2001). Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. *Nucl. Acids Res.* **29**, 2471–2478.
14. Bulyk, M. L., Johnson, P. L. F. & Church, G. M. (2002). Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucl. Acids Res.* **30**, 1255–1261.
15. Benos, P. V., Bulyk, M. L. & Stormo, G. D. (2002). Additivity in protein–DNA interactions: how good an approximation is it? *Nucl. Acids Res.* **30**, 4442–4451.
16. Kuhlman, B. & Baker, D. (2000). Native protein sequences are close to optimal for their structures. *Proc. Natl Acad. Sci. USA*, **97**, 10383–10388.
17. Kortemme, T. & Baker, D. (2002). A simple physical model for binding energy hot spots in protein–protein complexes. *Proc. Natl Acad. Sci. USA*, **99**, 14116–14121.
18. Chevalier, B. S., Kortemme, T., Chadsey, M. S., Baker, D., Monnat, R. J. & Stoddard, B. L. (2002). Design, activity, and structure of a highly specific artificial endonuclease. *Mol. Cell*, **10**, 895–905.
19. Havranek, J. J. & Harbury, P. B. (2003). Automated design of specificity in molecular recognition. *Nature Struct. Biol.* **10**, 45–52.
20. Dunbrack, R. L. & Karplus, M. (1993). Backbone-dependent rotamer library for proteins—application to side-chain prediction. *J. Mol. Biol.* **230**, 543–574.
21. Lovell, S. C., Word, J. M., Richardson, J. S. & Richardson, D. C. (2000). The penultimate rotamer library. *Proteins: Struct. Funct. Genet.* **40**, 389–408.
22. Luscombe, N. M., Laskowski, R. A. & Thornton, J. M. (2001). Amino acid–base interactions: a three-dimensional analysis of protein–DNA interactions at an atomic level. *Nucl. Acids Res.* **29**, 2860–2874.
23. Lesser, D. R., Kurpiewski, M. R. & Jenjacobson, L. (1990). The energetic basis of specificity in the EcoRI endonuclease DNA interaction. *Science*, **250**, 776–786.
24. Segal, D. J., Dreier, B., Beerli, R. R. & Barbas, C. F. (1999). Toward controlling gene expression at will: selection and design of zinc finger domains recognizing each of the 5′-GNN-3′ DNA target sequences. *Proc. Natl Acad. Sci. USA*, **96**, 2758–2763.
25. Dreier, B., Segal, D. J. & Barbas, C. F. (2000). Insights into the molecular recognition of the 5′-GNN-3′ family of DNA sequences by zinc finger domains. *J. Mol. Biol.* **303**, 489–502.
26. Bulyk, M. L., Huang, X. H., Choo, Y. & Church, G. M. (2001). Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. *Proc. Natl Acad. Sci. USA*, **98**, 7158–7163.
27. Elrod-Erickson, M., Rould, M. A., Nekludova, L. & Pabo, C. O. (1996). Zif268 protein–DNA complex refined at 1.6 Å: a model system for understanding zinc finger–DNA interactions. *Structure*, **4**, 1171–1180.
28. Warshel, A. & Russell, S. T. (1984). Calculations of electrostatic interactions in biological-systems and in solutions. *Quart. Rev. Biophys.* **17**, 283–422.
29. Dreier, B., Beerli, R. R., Segal, D. J., Flippin, J. D. & Barbas, C. F. (2001). Development of zinc finger domains for recognition of the 5′-ANN-3′ family of DNA sequences and their use in the construction of artificial transcription factors. *J. Biol. Chem.* **276**, 29466–29478.
30. Hurt, J. A., Thibodeau, S. A., Hirsh, A. S., Pabo, C. O.

& Joung, J. K. (2003). Highly specific zinc finger proteins obtained by directed domain shuffling and cell-based selection. *Proc. Natl Acad. Sci. USA*, **100**, 12271–12276.

31. Hecht, M. H., Richardson, J. S., Richardson, D. C. & Ogden, R. C. (1990). *De novo* design, expression, and characterization of felix—a 4-helix bundle protein of native-like sequence. *Science*, **249**, 884–891.
32. Berman, H. M., Westbrook, J., Feng, Z. K., Iype, L., Schneider, B. & Zardecki, C. (2002). The Nucleic acid database. *Nucl. Acids Res.* **58**, 889–898.
33. Lazaridis, T. & Karplus, M. (1999). Effective energy function for proteins in solution. *Proteins: Struct. Funct. Genet.* **35**, 133–152.
34. Kortemme, T., Morozov, A. V. & 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.
35. Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L. & Baker, D. (2003). Design of a novel globular protein fold with atomic-level accuracy. *Science*, **302**, 1364–1368.
36. Morozov, A. V., Kortemme, T., Tsemekhman, K. & Baker, D. (2004). Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations. *Proc. Natl Acad. Sci. USA*, **101**, 6946–6951.
37. Yu, C., Kortemme, T., Robertson, T., Baker, D. & Varani, G. (2004). A new hydrogen-bonding potential for the design of protein–RNA interactions predicts specific contacts and discriminates decoys. *Nucl. Acids Res.*, in press.
38. Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (1992). *Numerical Recipes in C: The Art of Scientific Computing* (2nd edit.). Cambridge University Press, New York.
39. Foloppe, N. & MacKerell, A. D. (2000). All-atom empirical force field for nucleic acids: I. Parameter optimization based on small molecule and condensed phase macromolecular target data. *J. Comput. Chem.* **21**, 86–104.
40. Jiang, L., Kuhlman, B., Kortemme, T. & Baker, D. (2004). A solvated rotamer approach to modeling water mediated hydrogen bonds at protein–protein interfaces. *Proteins: Struct. Funct. BioInform.*, in press.

Edited by F. E. Cohen

(Received 8 March 2004; received in revised form 28 July 2004; accepted 10 September 2004)