Published online 30 January 2007 *Nucleic Acids Research, 2007, Vol. 35, No. 4* ***1085–1097***
doi:10.1093/nar/gkl1155

Structure-based prediction of C_2H_2 zinc-finger binding specificity: sensitivity to docking geometry

Trevor W. Siggers and Barry Honig*

Howard Hughes Medical Institute, Center for Computational Biology and Bioinformatics, Department of Biochemistry and Molecular Biophysics, Columbia University, 1130 St. Nicholas Avenue, Room 815, New York, NY 10032, USA

Received November 8, 2006; Revised December 13, 2006; Accepted December 14, 2006

ABSTRACT

Predicting the binding specificity of transcription factors is a critical step in the characterization and computational identification and of *cis*-regulatory elements in genomic sequences. Here we use protein–DNA structures to predict binding specificity and consider the possibility of predicting position weight matrices (PWM) for an entire protein family based on the structures of just a few family members. A particular focus is the sensitivity of prediction accuracy to the docking geometry of the structure used. We investigate this issue with the goal of determining how similar two docking geometries must be for binding specificity predictions to be accurate. Docking similarity is quantified using our recently described interface alignment score (IAS). Using a molecular-mechanics force field, we predict high-affinity nucleotide sequences that bind to the second zinc-finger (ZF) domain from the Zif268 protein, using different C_2H_2 ZF domains as structural templates. We identify a strong relationship between IAS values and prediction accuracy, and define a range of IAS values for which accurate structure-based predictions of binding specificity is to be expected. The implication of our results for large-scale, structure-based prediction of PWMs is discussed.

INTRODUCTION

In eukaryotic gene regulation, highly specific patterns of gene expression are established by *cis*-regulatory elements (CREs)—modular domains of several hundred base pairs containing multiple short DNA sequences (~5–15 bp) that serve as binding sites for transcription factors (TFs) (1) Given a genome, a major challenge is to identify CREs in the sequence and to identify the TFs that bind to them (2) A common situation is that one or more TFs are believed to regulate the expression of a set of genes, and one wants to identify the CREs responsible. Using a description of a TF's binding specificity such as a consensus sequence or position weight matrix (PWM) (3), the genomic regions of interest (e.g. gene upstream and intronic regions) can be searched for binding sites that would indicate a potential CRE. However, since TF-binding sites are commonly short and sequence degenerate, many sites will often be identified while only a few will actually be part of a CRE. This is particularly evident in the large non-coding regions present in higher eukaryotic genomes (4). To confront this problem, additional criteria have been used to refine the initial list of identified sites such as the PWM score (3), clustering of the sites(5–8); models based on previously identified CREs(9–11); or phylogenetic conservation of the sites (12–14). Critical to the success of all these approaches is an accurate description of the TF's binding specificity.

Databases such as TRANSFAC (15) and JASPAR (16) provide data on experimentally determined DNA-binding specificities and provide PWMs for many eukaryotic TFs. Unfortunately, for most TFs there is currently little or no data as to the DNA sequences they bind to. The most recent release of TRANSFAC database (version 7.0–public), the larger of the two databases, has approximately 400 PWMs for all eukaryotic TFs. In comparison, the human genome has approximately 1850 predicted TFs (17); therefore many remain without well-characterized binding specificities. High-throughput experimental approaches, such as SELEX (18), quantitative multiple fluorescence relative affinity (QuMFRA) assay (19), genome-wide location analysis (ChIP-chip) (20), DNA immunoprecipitation with microarray detection (DIP-chip) (21) and protein-binding microarrays (PBM) (22) are increasingly being used to efficiently determine the binding-specificity of individual TFs; however, obtaining high-quality data for many TFs still proves challenging (23). Additional approaches for determining binding specificities that combine computational methods and structural information with sequence data (24) and experimental data (SELEX, phage display) (25) have also been proposed. Recently, various computational

*To whom correspondence should be addressed. Tel: + 1 212 851 4651; Fax: + 1 212 8514 650; Email: bh6@columbia.edu

methods have been reported for using structures of protein–DNA complexes to predict binding specificities directly (26–34), providing a potential complementary strategy to high-throughput determination of TF PWMs.

All purely structure-based methods require either an experimentally determined structure or a computationally derived model of a protein–DNA complex. Given a structure of a TF bound to one or more DNA sequences, a scoring function can be used to evaluate relative affinities. One type of scoring function uses knowledge-based potentials that have been developed through the statistical analysis of many protein–DNA structures. Such potentials have been used to estimate pair-wise amino-acid–nucleotide interaction energies (26–30,35), amino-acid–polynucleotide interaction energies (30) and DNA-deformation energies (36–38). The potentials are based on structural parameters (e.g. amino-acid–nucleotide distances; twist, roll and tilt parameters of base-pair steps) chosen so as to be independent of the residue identities. Therefore, a single template complex can be used to predict affinities for many different protein and DNA sequences by simply changing the identities of the amino acid or the base and re-evaluating the potential. In this case, one generally only needs knowledge of the coordinates of the protein and DNA backbones since, in most current applications, the detailed atomic nature of amino-acid–base interactions are ignored. However, it is possible that specificity will depend in part on side-chain orientation. Indeed, alternate side-chain conformations that affect DNA-binding specificity have been observed in response to both the mutation of neighboring side-chains and the binding of different DNA sequences (39), and likely contribute to the inter-dependent effects on binding affinities that have been observed for amino acid and base residues (19,40).

Another approach to scoring is to build an all-atom model for a set of complexes of interest and to evaluate relative affinities with molecular-mechanics energy functions (31,32,41); physically based, atomic-level energy functions (33,42); or statistically derived atom–atom potentials (42,43). In most recent studies, the protein and DNA sugar–phosphate backbone atoms are held fixed in the original, experimentally derived structure, and alternate DNA sequences are modeled by constructing DNA bases that are co-planar with the original template bases. For each DNA sequence, amino-acid side-chains are then modeled onto the fixed protein backbone by choosing conformations from a rotamer library that minimize the energy function (32,33,42). An alternative method used by Paillard *et al.* (31,41) is to begin with the side-chains in the original template conformations and to perform an energy minimization in the presence of a multi-copy representation of the DNA bases (44,45), where all DNA bases are present at each position simultaneously.

Despite the increasing number of protein–DNA complexes in the Protein Data Bank (PDB), TF sequences still greatly outnumber TF structures. Therefore, for the majority of TFs an experimentally derived structural template containing the TF will not be available. Since many structurally similar proteins dock on the DNA in a similar fashion (46–48), one solution is to use a protein–DNA complex of a structurally homologous protein as a template for a TF for which no DNA-bound complex is available. Recent papers by Morozov *et al.* (42) and Contreras-Moreira and Collado-Vides (35) have used all-atom modeling and knowledge-based approaches, respectively, to predict PWMs using structural homologs as templates. The homology modeling of protein–DNA complexes provides a broadly applicable method for PWM prediction as there are currently enough protein–DNA complexes in the PDB to represent the majority of TF structural families. Integral to a homology modeling approach is the assumption that a protein–DNA complex can serve as the template structure for a structurally homologous protein. This requires that the docking geometry—the spatial orientation of the protein backbone to the DNA molecule—of structural homologs is sufficiently similar that the difference in docking orientation does not affect the outcome of the predictions.

In this article, we test this assumption by using different protein–DNA complexes of structurally homologous C_2H_2 zinc fingers (ZF) as templates in atomic-level modeling predictions of TF-binding specificity. Extensive structure–based predictions of ZF binding specificity are carried out and the sensitivity of prediction accuracy on differences between the model and actual docking geometries is analyzed. Relative docking geometries are quantified in terms of an interface alignment score (IAS) that we recently described (48). The algorithm that calculates IAS values aligns amino acids from the binding interface of two protein–DNA complexes based on the spatial relationships between the amino-acid backbone and DNA base atoms and provides a quantitative measure of the similarity between two protein–DNA interfaces. The IAS has been shown to provide a robust and sensitive measure for comparing the docking geometry of protein–DNA complexes.

The C_2H_2 ZF proteins constitute a large family of eukaryotic TFs that have been studied extensively both experimentally (49–51) and computationally (24–26,31–33,42) as a model system for understanding protein–DNA-binding specificity. The proteins in the family are often composed of multiple, structurally homologous, concatenated ZF domains of approximately 30 amino acids that bind successively along the DNA major groove. From the twenty-one available ZF protein structures listed in Table 1, ninety-three individual ZF *domain* templates could be defined (Materials and Methods), representing the largest set of structurally homologous protein–DNA complexes currently available from the PDB. Fourteen of the twenty-one structures listed in the table are based on the well-studied Zif268 protein and include: two wild-type structures at different resolutions; one structure with two wild-type Zif268 proteins bound in tandem; nine structures with mutations in the ZF domain 1 (ZF1) in complex with various off-consensus DNA sites of the form GCGTGGNNN; and two modified Zif268 structures bound as head-to-head dimers. Three of the twenty-one structures are designed ZF proteins engineered to bind novel DNA sequences, and the remaining four structures are of

Table 1. C_2H_2 ZF-DNA PDB files

PDB code	Chains	Description	Res (Å)	Topology[a]
1llm	C,D	Zif268-GCN4 (dimer)	1.5	2_3:3_2
1aay	A	Zif268	1.6	3_2_1
1a1f, 1a1g, 1a1h, 1a1i, 1a1j, 1a1k, 1a1l	A	Zif268 (Fn1 mutants)	1.6	3_2_1
1jk1, 1jk2	A	Zif268 (Fn1 D20A)	1.9	3_2_1
1zaa	C	Zif268	2.1	3_2_1
1mey	C,F	Designed	2.2	3_2_1
1g2d, 1g2f	C,F	Designed	2.2	3_2_1
1p47	A,B	Zif268 tandem	2.2	3_2_1 3_2_1
1f2i	G,H,I,J,K,L	Zif268-extension (dimer)	2.4	2_1:1_2
1ubd	C	YY1 (Yin Yang 1)	2.5	4_3_2_1
2gli	A	GLI (glioblastoma)	2.6	5_4_3_2_1
2drp	A,D	Tramtrack	2.8	2_1
1tf6	A,D	TFIIIA	n/a (NMR)	6_5_4_3_2_1

[a]Topology description for the individual ZF domains; 3_2_1 indicates a polydactyl ZF protein with three ZF domains, 1 refers to the N-terminal ZF domain. Dimerization interfaces between chains are indicated with a colon.

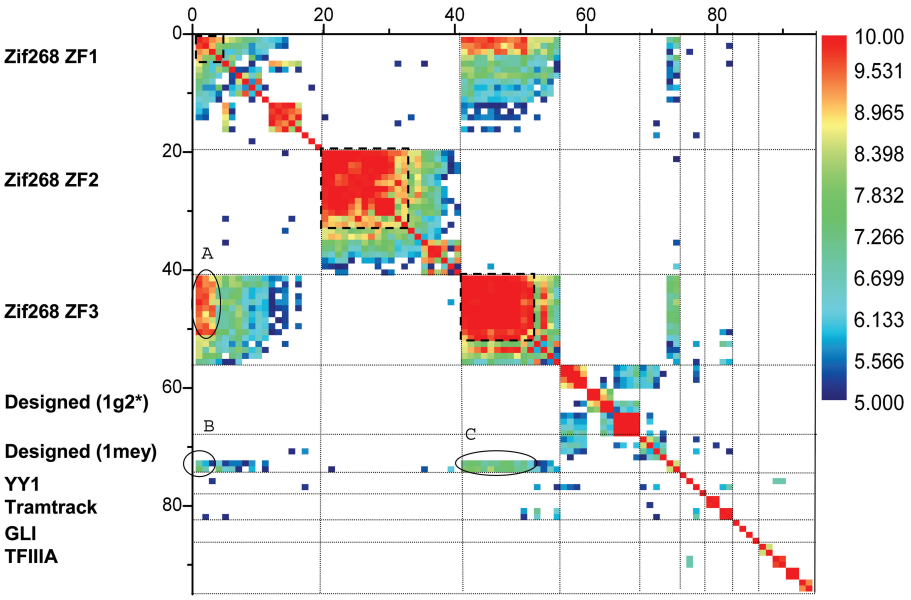

Figure 1. Pair-wise docking comparison of 93 ZF domains. Pair-wise IAS values are shown as a symmetric comparison matrix. Scores below 5.0 are in white (not shown). IAS values between domains with a *wild-type-docking* geometry (see text) in each of the three Zif268 groups are delineated with a heavy dashed line. IAS values between certain ZF domains, or groups of domains, are highlighted: these include wild-type docking domains from the Zif268 ZF1 and ZF3 clusters (region **A**); 1MEY ZF3 and wild-type docking domains from the Zif268 ZF1 (region **B**) and ZF3 (region **C**) clusters. Supplementary Figure S1 contains enlarged versions of the regions along the diagonal and list PDB identifiers and interfacial residue identities for each ZF complex.$units 6,6,5,1

various wild-type ZF proteins containing between two and six ZF domains bound to different DNA sequences.

Pair-wise structural superimpositions of all the ninety-three ZF domains yielded an average backbone heavy-atom RMSD (root mean square deviation) of 0.9 Å (variance 0.12). However, despite the overall structural similarity, significant differences in docking geometries are observed. These differences have an important effect on the ability to predict nucleotide sequences that preferentially bind to a ZF domain. We find a strong relationship between the IAS and prediction accuracy, and define a range of IAS values for which accurate structure-based predictions of PWMs is to be expected. Our results thus provide insight on how template choice can affect atomic-level modeling approaches to binding-specificity predictions and indicate ways in which prediction algorithms can be improved in future work.

MATERIALS AND METHODS

PDB file preparation and definition of orthologous residue positions

The twenty-one C_2H_2 ZF complexes from the PDB (Table 1) were split into individual finger–DNA complexes, yielding ninety-three individual ZF domains in complex with DNA. Some crystals contain multiple structures in the asymmetric unit; each was treated as an

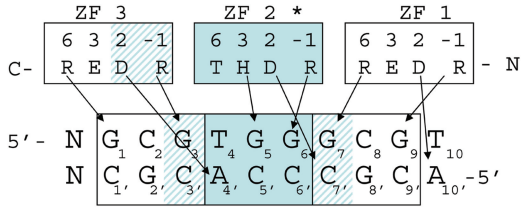

Figure 2. Canonical binding schema for Zif268 bound to its cognate DNA sequence. Arrows indicate hydrogen-bond interactions. Base identities appear as in the PDB file 1aay. Amino acids are numbered according to a canonical ZF numbering scheme (54). Solid shaded regions indicate the amino acids and bases used to calculate the docking-geometry IAS values; the solid *plus* hashed region indicates the interfacial residues that define the 'binding interface' used for modeling Zif268 ZF2 binding specificity.* IAS scores were computed using all amino acid positions −1 through 6.

independent protein–DNA complex. The domain boundaries for individual fingers were taken from the SCOP database (52). ZF domains for which no helical backbone atom was within 16 Å of a DNA purine N9 or pyrimidine N1 atom were not considered.

In order to use different ZF structures as templates, it was necessary to define side-chain and bases in the template structure that align with those being predicted. For the ZF domains, this was accomplished with a structural alignment using the SKA algorithm (53). Bases that correspond to those in the interface with the Zif268 ZF2 domain were identified by manual comparison of each structure. Side-chain and base identities for each template that correspond to those shown in Figure 2 for Zif268 ZF2 are shown in Supplementary Figure S1.

Interface alignment algorithm and docking similarity measures

The *interface alignment* algorithm used in this work is described in detail in Siggers *et al.* (48); however, a brief overview is presented here for clarity. A geometric similarity score $s(i, k; j, l)$ quantifies the spatial relationship of residue i and base k in one complex relative to that of residue j and base l in another. The measure is independent of the amino acid or base identities and is based on three geometric parameters that describe the geometric relationship of the amino-acid backbone with the DNA base atoms. The set of $s(i, k; j, l)$ scores is used to define a residue–residue similarity score, $S(i, j)$, that quantifies the spatial relationship of amino acid i to the local DNA–nucleotides (in a complex m) with that of amino acid j to its local DNA (in complex n). The $S(i, j)$ values, when computed for all pairs of amino acids in two complexes (m and n), define a similarity matrix that is used in a dynamic-programming algorithm to define a geometry-based alignment of the amino acids in the interfaces of the two complexes. The *interface alignment score* (IAS) is the sum of the $S(i, j)$ values for all aligned amino-acid residues. In this work, pair-wise interface alignments between ZF domains (Figure 1) were performed using only the seven amino-acid residues at canonical ZF positions −1 through 6, and bases similar to those at positions 4–6, and 4′–6′ for ZF2 in Figure 2 (i.e. the canonical base-triplet associated with each ZF domain). The computed IAS values were normalized by dividing by seven and resulted in scores ranging from zero to ten. The interface alignment software is freely available at the Honig lab website.

DNA-based structural superimposition of ZF helices

The DNA sugar and phosphate heavy atoms from the three nucleotide pairs (six nucleotides) used in the IAS calculations were structurally superimposed by minimizing their RMSDs. The transformation used for the superimposition was then applied to the ZF helix atoms.

Modeling side-chains and bases

Conformations of the amino acids and nucleotides being modeled into each template structure were predicted by choosing the lowest-energy set of conformational rotamers using the Monte Carlo procedure outlined in Xiang and Honig (55), with DNA–nucleotide rotamers being treated in an analogous fashion to the amino-acid rotamers. Protein side-chain conformations were taken from a torsion-angle rotamer library derived from the conformer library of Xiang and Honig (55). DNA–nucleotide conformations were obtained from a rotamer library generated by local sampling of the template DNA backbone as described in detail below. Correct base pairing was retained in all cases. The positions of the protein backbone atoms and DNA phosphate atoms were held fixed at all times. Starting from an initial, randomly chosen set of rotamers (i.e. initial conformation set), residue conformations were iteratively minimized by selecting rotamers that yielded the lowest energy for the complex; this procedure continues until no new rotamer leads to a lower energy. From 20 initial conformation sets, the complex with the lowest energy was chosen. Twenty initial conformations was found to be sufficient for the modeling described in this work, additional starting conformations did not results in lower energy predictions. The structure of the low-energy complex was then further refined using a modified version of the *minimize* routine in the TINKER software package (56) that incorporates a sigmoidal dielectric-screening function (see energy description below). Minimization was performed to an RMSD gradient tolerance of 0.0001 kcal/mol/Å. During the course of the minimization, atoms for all residues not being specifically modeled were held fixed. Protein backbone and DNA phosphate atoms for residues being modeled were also held fixed.

An all-atom energy function was used to compute the energy of the complex. The AMBER param98 force-field (57) was used to describe bonded and non-bonded energies; an additional hydrogen-bond term is included to account for the fine geometric dependence of hydrogen bonds. The hydrogen-bond term is a product of four Gaussian functions chosen to approximate the geometric parameter distributions described in

Chen *et al.* (58) and Kortemme *et al.* (59). The function has the form:

$$E_{HB} = -2.0 \times G(\delta_{HA})G(\psi)G(\phi)G(\pi) \quad (1)$$

$$\text{where } G(x) = \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

The value −2.0 defines the strength of an optimal hydrogen bond in units of kcal/mol. The geometric parameters δ_{HA}, ψ, ϕ are as shown in Figure 1 of Kortemme *et al.* (59): δ_{HA}—distance from H (hydrogen) atom to A (acceptor) atom; ψ—angle made by H, A and AA (acceptor antecedent) atoms; ϕ—angle made by D (hydrogen donor), H and A atoms. The π parameter is used only for nitrogen acceptor atoms in DNA bases and describes the smallest acute angle made a line drawn between the nitrogen acceptor and the donor hydrogen atoms and the plane defined by the nucleobase atoms. Parameters used for $G(\delta_{HA})$ were $\mu = 1.95$ Å and $\sigma = 0.25$ Å; above 2.45 Å $G(\delta_{HA})$ was set to 0 (upper cutoff), and below 1.95 Å was set to 1, to account for rotamer sampling error (i.e. close approach). Parameters used for $G(\psi)$ were $\mu = 120^\circ$ and $\sigma = 20^\circ$, with a lower cutoff of 80°. Parameters used for $G(\phi)$ were $\mu = 180^\circ$ and $\sigma = 30^\circ$, with a lower cutoff of 100°. Parameters used for $G(\pi)$ were $\mu = 0^\circ$ and $\sigma = 40^\circ$, with an upper cutoff of 90°.

Solvent screening was described with a sigmoidal distance-dependent dielectric function (dielectric permittivity ranging from 2.0 to 78.3 with a slope of 0.4), and salt effects were approximated by reducing the charge on the phosphate groups to −0.5 e (41,60,61). During the rotamer prediction stage of the iterative procedure, a VDW-softening scheme was used where atom pairs with an interatomic distance between $0.9*R_{min}$ and $1.0*R_{min}$ (R_{min} is the distance at which the VDW function is a minimum), had the inter-atomic distance reset to R_{min}; this distance re-scaling was similarly applied to the coulomb energy term. This VDW-softening scheme was not used in the energy minimization step that yielded the final predicted complex geometry, or when computing a relative binding energy for the complex. The geometry-dependent hydrogen-bond term was not used during the final minimization of each complex, as it was not implemented in TINKER, but was used when computing relative binding energies. The interface modeling software is freely available at the Honig lab website.

DNA–nucleotide rotamers

The DNA–nucleotide rotamer library consists of a set of nucleotide conformations that are generated at runtime by locally sampling around the nucleotide conformations present in the template structure. The nucleotide conformations (rotamers) are generated by applying the *wriggling* procedure described by Cahill *et al.* (62) to four of the six endocyclic DNA backbone torsion angles: $\xi(i-1)$ {C3*-O3*-P-O5*; C3* and O3* from 5′ nucleotide}; α {O3*-P-O5*-C5*; O3* from 5′ nucleotide}; β {P-O5*-C5*-C4*}; γ {O5*-C5*-C4*-C3*}. The wriggling algorithm defines angle perturbations for each dihedral angle so as to keep the overall structural perturbation local. Angle perturbations were between −0.05 and 0.05 radians and conformations were selected every 150 iterations. The wriggling procedure breaks the backbone chain between the O3* atom and the P atom of the 3′ nucleotide which is then closed with a short energy minimization that allows only the nucleotide being sampled to move. While the endocyclic torsion angles δ and ε, and the glycosidic torsion angle (χ), are not explicitly altered, they are perturbed during the chain-closure minimization procedure and are therefore also locally sampled.

The identity of the rotamers can be changed prior to the wriggling procedure by introducing a new planar base moiety and setting the glycosidic angle to that of the template nucleotide. The choice of dihedral angles for the wriggling procedure leaves the phosphate atom positions fixed for each set of rotamers. These phosphate atoms are also held fixed during the chain-closure procedure. Fifty nucleotide rotamers were used in the prediction of each modeled nucleotide. This procedure generates a small ensemble of nucleotide rotamers with conformations close to the template conformation.

RESULTS

Variations in docking geometry

All-on-all alignment of the 93 ZF-DNA interfaces (Figure 1) reveals a range of IAS values from 10 to below 5. As will be discussed below, this corresponds to a fairly wide, and functionally significant, range of docking geometries. IAS values were calculated using the amino acid and base residues represented by the solid-shaded region of Figure 2 and were normalized to range from 10 to 0 (Materials and Methods). The (*m*,*n*) and (*n*,*m*) entries in Figure 1 correspond to the IAS values between ZF domains *m* and *n*. Scores below 5.0 are shown in white. Zif268 and Zif268-derived ZF domains (Table 1) are listed first according to their relation to the three Zif268 domains ZF1-ZF3. For these three Zif268 groups the ZF domain from the high-resolution, wild-type Zif268 structure (hrZif268) (54) is listed first and subsequent domains are organized by decreasing IAS value with this domain. Domains from non-Zif268-derived complexes are listed by protein with the most N-terminal domain listed first. For example, for 'Designed (1g2*)' domains (1g2d and 1g2f, chains C and F) all ZF1 domains comes first, followed by ZF2 and ZF3 domains.

To illustrate the relationship between the IAS measure of docking geometry similarity and a common structural measure such as RMSD we performed a DNA-based structural superimposition of five ZF domains with the hrZif268 ZF2 domain and relate the RMSD of the superimposed ZF helix atoms with corresponding IAS values (see Materials and Methods). Structural superimposition of ZF complexes was performed by minimizing the RMSD between the DNA sugar–phosphate backbone atoms from the three nucleotide pairs nearest each ZF (i.e. nucleotides 4–6, 4′–6′ for ZF2 in Figure 2). This superimposition allows a structural comparison of

Table 2. IAS and RMSD measures of docking similarity between hrZif268 ZF2 and five ZF domains

PDB code[a]	IAS[b]	RMSD ZF helix[c]	RMSD DNA[d]
1zaa_C2	9.2	0.2	0.3
1llm_C2	8.6	1.2	0.6
1f2i_J2	7.8	1.8	0.7
1f2i_K2	5.4	2.4	0.9
1g2d_C1	1.6	1.9	1.1

[a]PDB identifier, chain ID and ZF domain number as in Table 1.
[b]IAS values from an alignment with 1aay, chain A, ZF2 domain (1aay_A2).
[c]RMSD of ZF helix atoms after a DNA-based superimposition of each complex with 1aay_A2.
[d]RMSD of the DNA sugar-phosphate backbone atoms used to perform the DNA-based structural superimposition.

the ZF docking geometry from a common DNA frame of reference. Table 2 lists the pair-wise IAS values, the backbone RMSD of superimposed ZF helices and the RMSD of the DNA sugar-phosphate atoms used for the DNA-based superimposition. An inverse relationship is observed with decreasing IAS values corresponding to increasing RMSD values of both the helix and DNA backbone. The highest IAS value of 9.2, between hrZif268 and ZF2 from 1zaa, corresponds to very low RMSDs of 0.2 Å and 0.3 Å for the superimposed helices and DNA backbones respectively. When the IAS value is 7.8 (1f2i chain J) the helix RMSD is 1.8 Å. Because the IAS is sensitive to deviations in the protein orientation relative to the DNA, as well as structural differences in the DNA bases themselves (i.e. roll, twist, tilt), the IAS differences are not fully captured by either RMSD measures alone. Figure 3 shows ZF helices from three of the five structures (1zaa, 1llm, 1f2i chain K) superimposed with hrZif268 ZF2 helix to further illustrate the degree and nature of docking variation exhibited by different ZF domains and characterized by the corresponding IAS values.

High IAS scores within the three Zif268 groups in Figure 1 indicate that many of the ZF domains exhibit a conserved docking geometry. However, docking-geometry variation is clearly evident within the three groups, particularly for Zif268 ZF1. Within each group, ZF domains that have an IAS greater than 9.0 with the domain from the hrZif268 structure will be referred to as exhibiting a *wild-type* docking geometry. Pair-wise comparisons between wild-type-docking Zif268 domains are enclosed by the heavy dashed line. In all three Zif268 groups, domains from proteins involved in dimeric complexes (see Table 1) exhibit non-wild-type docking. The designed Zif–GCN4 complex (1llm) has two ZF domains linked to a helical region derived from GCN4; dimerization occurs via a leucine-zipper interface between the two GCN4-type helices. The modified Zif268 protein (1f2i) has an N-terminal peptide extension that makes contact with a hydrophobic patch on the neighboring ZF domain and mediates dimerization. While not strictly a dimer, the two tandemly bound ZF proteins, 1p47 chains A and B, do interact with each other and we observe that the three ZF domains from chain B are all

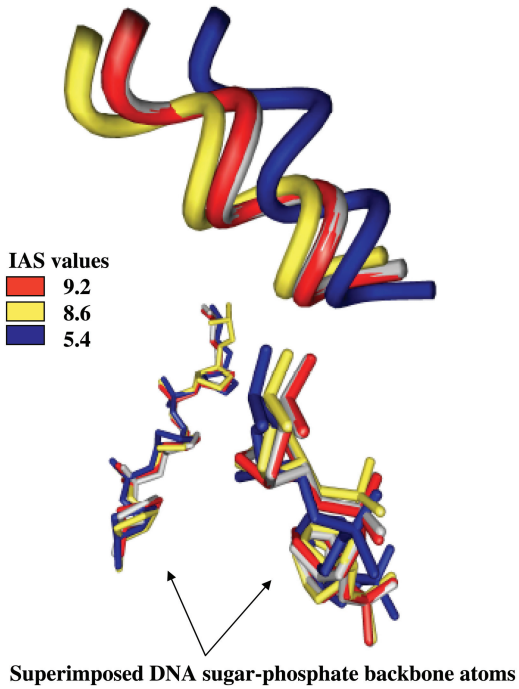

Figure 3. DNA-based structural superimposition of four ZF helices. Worm representation of ZF helix residues (canonical numbering 1–11) are shown for four ZF2 domains: two wild-type Zif268 proteins, 1aay/hrZif268 (gray) and 1zaa (red); and two modified Zif268 proteins, 1llm chain C (yellow) and 1f2i chain K (blue). Colors correspond to IAS color scale used in Figure 1. DNA sugar-phosphate heavy atoms used for the structural superimposition are shown in stick form. IAS values for comparisons of hrZif268 ZF2 (gray) with the three other ZF domains are shown and colored accordingly.

bound in a non-wild-type fashion. The ZF1 cluster also contains many sequence variants where either the DNA or the protein sequences have been mutated (Table 1: 1a1*, 1jk*; Supplementary Figure S1). With the exception of 1jk1, these domains also exhibit non-wild-type docking. Thus, interactions between apposing ZF proteins and sequence mutations within a ZF complex (protein or DNA) can both alter the docking geometry.

The majority of interface alignments between different ZF domains both from the same protein and from different proteins have IAS values below 5.0 (white space). This suggests considerable docking variation within the C_2H_2 ZF family. It is of interest, however, that docking similarities are observed between different ZF domains both within the same ZF proteins and from different ZF proteins (e.g. Figure 1, regions A, B and C); these similarities are discussed in more detail below. In the following section, we investigate the effect that the observed variation in ZF docking geometry can have on TF-binding-specificity predictions.

Effects of template-docking geometry on prediction of binding-specificity

Binding-specificity predictions were performed for the ZF2 domain from Zif268 using different templates as a basis for the predictions. For each template, IAS values

obtained from an interface alignment with the Zif268 ZF2 (hrZif268) complex were calculated. The scores were derived from all amino-acid–base pairs within the 'binding interface' defined in Figure 2 (solid-shaded region). Prediction accuracy was determined by comparing predictions to the results of Bulyk *et al*. (51) who used a protein-binding microarray technology to determine binding affinities between Zif268 (ZF1-ZF3) and all sixty-four GCGNNNGCG sequence variants of the Zif268 consensus-binding site shown in Figure 2. Since the three central bases are recognized by the ZF2 domain, the affinity measurements provide a comprehensive data set against which ZF2 binding-specificity predictions can be compared (Table 3).

IAS values (IAS_{Zif268}) were calculated between all ninety-three template structures and the ZF2 domain from the hrZif268 structure. Based on the IAS values, twenty-three ZF domains (including hrZif268 ZF2 itself) spanning a range of IAS values were chosen as templates for the prediction of nucleotide sequences that preferentially bind to Zif268 ZF2. For each template, we generated sixty-four models of a complex between the modeled structure of Zif268 ZF2 and DNA, each corresponding to a different tri-nucleotide sequence at nucleotide positions orthologous to 4–6 (4′–6′) in Figure 2. For each template, amino acids at the canonical positions −1, 2, 3 and 6 were changed to Arg, Asp, His and Arg, respectively, so as to correspond with those present in the Zif268 ZF2 domain. Nucleotides as positions 3 and 7 in Figure 2 that flank the tri-nucleotide sequence were built as Gua to better agree with the GCGNNNGCG consensus sequence. Finally, for templates with an adjacent C-terminal ZF domain, the amino acids at the canonical −1 and 2 positions were built as Arg and Asp, respectively, to correspond with the Zif268 ZF3 domain. Previously it had been suggested that the Asp(2) residue of Zif268 ZF3 domain might contribute to the selectivity of nucleotides at the 4′ position (Figure 2) (54). However, our predictions were largely insensitive to the treatment of these additional residues and their inclusion did not appreciably affect the results (data not shown).

For each of the sixty-four complexes generated per template, side-chain and base conformations were predicted as discussed in the Materials and Methods section. Relative binding energies for each of the sixty-four nucleotide sequences were calculated as the difference between the energy of the complex and the energy of an unbound DNA molecule generated by removing the protein from the complex and then energy minimizing the DNA structure using the protocol described in the Materials and Methods section. Protein and DNA sequence differences outside of the interfacial residues explicitly being modeled can contribute to the calculated binding energies and therefore present a potential source of prediction variability between different template complexes. However, test predictions performed for several templates in which the energetic contribution from these non-interfacial residues (protein and DNA) was ignored did not appreciably affect the predicted relative binding energies (data not shown).

Table 3. Highest affinity GCGNNNGCG sequences bound by Zif268

Experiment[a]	Exp. rel.ΔG[b]	Prediction[c]	Calc. rel.ΔG[d]
TGG	0.0	TGG	0.0
TAG	0.5	GGG	3.6
GGG	1.3	TAG	4.0
CGG	1.5	CGG	6.6
AGG	1.7	GAG	8.4
TTG	1.9	TGA	8.6
GAG	1.9	TTG	9.7
		AGG	10.5
		GGA	12.3
		CAG	12.8

[a]NNN base triplet identities from the seven highest-affinity GCGNNNGCG sequences identified by Bulyk *et al*. (51).
[b]Measured relative binding free energies (kT) of each sequence for Zif268 (51).
[c]Highest affinity predicted sequences listed according to calculated relative binding energies. hrZF268 was used as template structure for all predictions.
[d]Calculated relative binding affinities (kT) for each predicted sequence.

Table 4. Dependence of Zif268 ZF2 binding-specificity predictions on template docking geometry

PDB code[a]	IAS_{Zif268}[b]	Top_{seq}[c]	Top_{rank}[d]	Top 3[e]	Top 6[e]	Top 7[e]
1aay_A2 (hrZif268)	10.0	TGG	1	3	7	8
1jk1_A2	10.0	TGG	1	3	7	59
1p47_A2	10.0	TGG	1	3	8	55
1a1k_A2	9.9	TGG	1	4	7	9
1a1l_A2	9.9	TGG	1	3	7	12
1a1f_A2	9.5	TGG	1	3	9	13
1a1i_A2	9.5	TGG	1	6	7	16
1zaa_C2	9.2	TGG	1	4	8	9
1p47_B2	8.6	AAG	3	8	9	19
1llm_C2	8.6	AGG	2	6	15	18
1f2i_J2	7.8	GGG	4	7	7	12
1f2i_H2	7.3	GGG	10	21	14	21
1f2i_L2	6.8	GGG	9	16	16	21
1f2i_G2	5.5	AGG	4	7	7	11
1f2i_K2	5.4	GGG	4	6	6	16
1f2i_K1	4.5	GAT	18	18	32	52
1jk1_A1	4.0	GGA	8	11	11	57
1p47_B1	3.3	GAG	25	25	33	40
1aay_A3	3.0	AAG	4	9	9	16
1aay_A1	2.7	GAT	8	8	22	27
1g2f_F2	2.1	GGG	14	31	31	57
1g2f_C2	1.9	TGG	1	33	32	33
1g2d_C1	1.6	GGG	13	30	47	57

[a]PDB identifier, chain ID and ZF domain number as in Table 1.
[b]IAS value between each template and 1aay_A2 (hrZif268 ZF2).
[c]Top predicted sequence NNN (i.e. GCGNNNGCG).
[d]Predicted rank of the consensus TGG (i.e. GCGTGGGCG) sequence.
[e]Indicate how far down the list of ranked predicted sequences you need to go to include the top N binding sequences determined by Bulyk *et al*. (51). For example, a 7 in the top 6 column indicates that the 6 highest-affinity experimentally determined sequences are present within the top 7 predicted sequences.

hrZif268 ZF2 template yields accurate predictions. Table 3 lists the seven high-affinity binding sequences identified by Bulyk *et al*. and the ten highest-affinity predicted sequences using hrZif268 ZF2 as the template for itself. As can be seen from the table, the high-affinity TGG sequence is correctly predicted as

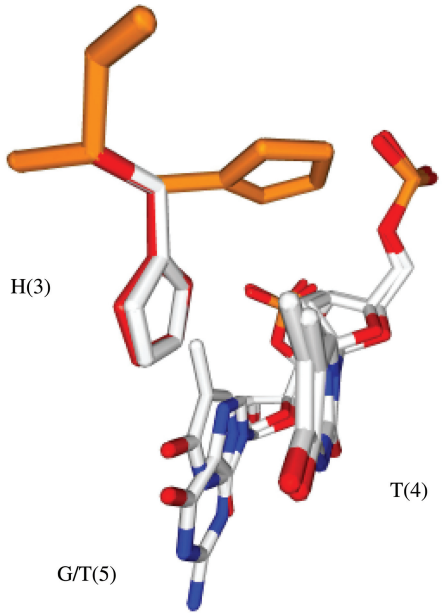

Figure 4. Native and predicted His(3) side-chain conformations. Side-chain conformations for the His residue at canonical ZF position 3 are shown from hrZif268 ZF2 (1aay His149; white), and from the complexes modeled with the TGG sequence (red) and TTG sequence (brown) using hrZif268 ZF2 as a template. DNA bases are shown in CPK coloring and correspond to the modeled TG (TGG) and TT (TTG) bases. Residue numbering as in Figure 1.

the strongest binder, the TGG, TAG, and GGG sequences are correctly identified as the three highest-affinity sequences, and all seven of the highest-affinity experimentally determined binding sequences are present in the top-eight predicted sequences. Noteworthy is the correct prediction of the TTG sequence as a high-affinity sequence. In the complex modeled with the TTG sequence, the His(3) side-chain adopts a conformation considerably different from those predicted for all other sequences. His(3) rotates out of the major groove and forms a hydrogen bond with a DNA-backbone phosphate group (Figure 4). Predicting this alternate side-chain conformation is required to correctly identify the TTG sequence as a high-affinity site demonstrating both the importance of allowing side-chain flexibility in the modeling process and that a template structure can be used to effectively model bound complexes where the side-chain conformations are different than those in the original template. The striking agreement of the high-affinity predicted sequences with all seven high-affinity sequences demonstrates that our atomic-level modeling approach can yield highly accurate predictions given an appropriate template structure. Furthermore, it suggests that the docking geometry of the hrZif268 ZF2 template bound to the TGG sequence is a reasonable representation of the docking geometry when bound to the alternate high-affinity sequences.

We found that the use of the geometry-dependent hydrogen bond improved the accuracy of the binding-specificity predictions. In the case of the Zif268 ZF2 domain discussed here, we found that not including the hydrogen bonding term did not greatly affect predictions when the 1.6 Å Zif268 structure (1aay) was used as a template. However, when the 2.1 Å Zif268 structure (1zaa) was used TGG was no longer predicted to be the top sequence, and recovery of the top 3 and top 7 experimental sequences (Table 4, columns 5 and 7) changed from 4 and 9, to 5 and 12, respectively. Thus, the constraints imposed by requiring that good hydrogen bonds be formed appear to reduce inaccuracies due to errors in the structures that are used.

Allowing some flexibility in the DNA structure by using DNA rotamers rather than requiring that all the bases be co-planar with the bases of the original template also improved prediction accuracy. For example, when the modeled bases were assumed to be co-planar with those in the 1.6 Å Zif268 template structure (1aay), the AGG sequence (see Table 3) was no longer ranked in the top eight. When the same assumption was applied to the 2.16 Å Zif268 template (1zaa) the effects were even larger; TGG was no longer ranked number 1 and recovery of the top 3 and top 7 experimental sequences (Table 4, columns 5 and 7) changed from 4 and 9, to 9 and 20, respectively. Our results clearly illustrate the importance of allowing maximum flexibility for both the DNA and the protein side-chains, thus allowing for the energetic optimization the interface that is formed between the two macromolecules.

Relating prediction accuracy to IAS_{Zif268} scores. Table 4 lists the binding-site predictions for Zif268 obtained from each of the 23 templates. The templates are ordered according to their IAS_{Zif268} scores. For each template, we report the top nucleotide sequence that is predicted and, in addition, the number of nucleotide sequences that must be predicted in order to recover the top 1, top 3, top 6 or top 7 experimentally determined sequences, respectively. Thus for example, using hrZif268 ZF2 as a template (top row of Table 4), the top experimental sequence (TGG) is recovered as the one with the best score, the top 3 experimental sequences are included in the top 3 predicted sequences, 7 predicted sequences are required to recover the top 6 measured sequences, and 8 predicted sequence are required to recover the top 7 experimental sequences.

As can be seen in the table, TGG is correctly predicted as the highest-ranking sequence using all templates with IAS_{Zif268} scores above 9.0, while only one template (1g2f_C2) with a score below 9.0 identified TGG as the top sequence. However in the 1g2f_C2 template, the predicted side-chain base contacts for the complex modeled with the TGG sequence were different from those observed in the hrZif268 ZF2 complex so that the agreement is likely to be fortuitous. Consistent recovery of the top 3 and top 6 experimental sequences within the top 6 and 10 predicted sequences, respectively, is observed for templates with IAS_{Zif268} scores 9.0 and above. Accurate recovery of all top 7 sequences appears more problematic and this appears to be primarily due to poor prediction of the TTG sequence; however, reasonable recovery within the top 16 predicted sequences is observed for most templates with IAS_{Zif268} scores above 9.0. These results demonstrate that templates with docking

geometries more closely resembling the hrZif268 ZF2 structure yield consistently more accurate results. Finally, it is worth noting that the hrZif268 ZF2 template itself resulted in the most accurate predictions, demonstrating that the high-resolution structure of the Zif268 ZF2 domain does in fact provide the best template geometry.

In order to gain another perspective on the relationship between IAS score and prediction accuracy, we organized the templates into five groups and, for each, the three highest-affinity predicted sequences from each template were submitted together to WebLogo (63,64) to generate a DNA sequence logo representing the binding-specificity predicted using that group of templates. Groups one through three (Figure 5B–D) were defined based on their IAS_{Zif268} values and consists of the ZF2 domain templates from Zif268 and Zif268-derived proteins (Table 1). Group four (Figure 5E) consists of the four ZF1 domains from Zif268-derived proteins. Group five (Figure 5F) consists of the two ZF2 domains from the designed TATA-binding ZF 1g2f.

As can be seen from Figure 5, the predicted specificity for the first group of templates (Figure 5B; $9.0 < IAS_{Zif268}$) is in excellent agreement with experiment (Figure 5A). Slightly poorer agreement is exhibited for the second template group (Figure 5C; $8.0 < IAS_{Zif268} < 9.0$); however, there were only two templates in this group, so the statistics are not good. For the third group of templates (Figure 5D; $5.0 < IAS_{Zif268} < 8.0$), there is excellent agreement with the experimental logo at base positions 2 and 3; however, we observe no selectivity for thymine at base position 1 (i.e. Thy(4) in the canonical ZF2 numbering scheme, Figure 2). This group of templates consists of five ZF2 domains from different chains of the 1f2i complex (Table 1). The intermediate IAS_{Zif268} values for these templates are almost entirely due to an altered geometric relationship between helix position 6 and the bases at position 1. The Thr side-chain modeled at position 6 interacts both with the base at position 1 as well as the His(3) side-chain, which in the hrZif268 complex makes strong van der Waals interactions with a Thy at base position 1. This altered template geometry results in higher relative binding energies when a Thy base is present at position 1, as compared to predictions using the hrZif268 template, which result in a reduced selectivity for Thy at this position. For template groups four and five (Figure 5E and F; $AS_{Zif268} < 5.0$), we observe poor agreement with the experimental logo and qualitatively different specificity predictions at most base positions. These incorrect predictions are the result of incorrect side-chain predictions resulting from the use of an incorrect docking geometry.

Factors that affect docking geometry

The preceding results demonstrate that IAS_{Zif268} values correlate well with prediction accuracy. This suggests that the IAS calculated between two ZF domains provides a good indicator of whether one structure can be used as a template to model the binding specificity of the other. Perhaps the most striking feature of the pair-wise docking

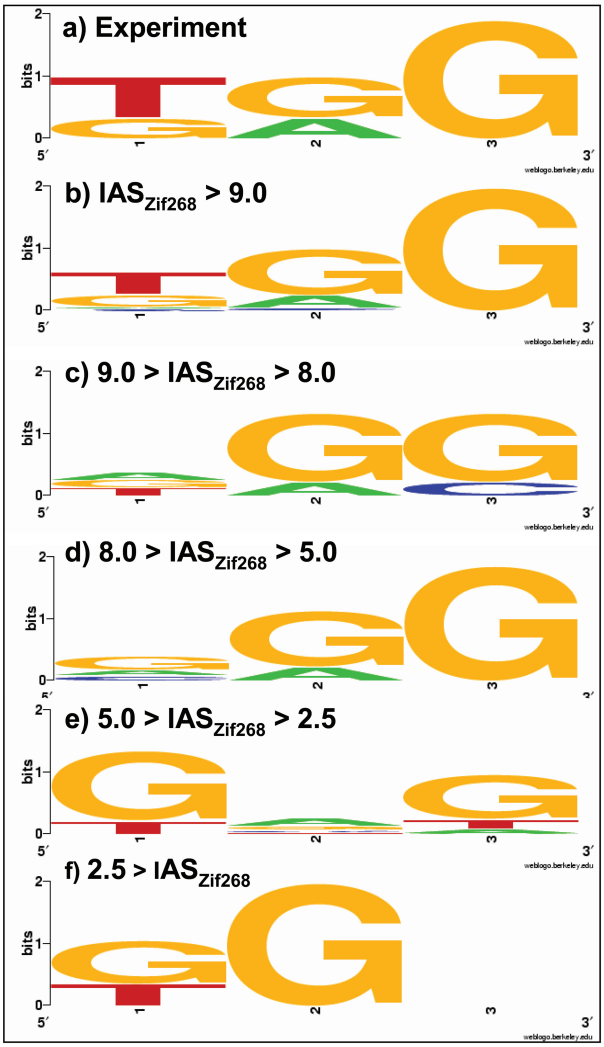

Figure 5. Predicted and experimental binding-specificity logos for the three-base-pair sequence recognized by the Zif268 ZF2 domain. (**A**) Logo generated from the three highest-affinity experimental sequences (51) (**B**) to (**F**) Logos generated for the five template groups using the three highest-affinity predicted sequences from each template within the group. The range of template IAS_{Zif268} scores for each group is shown. Logos were generated using the WebLogo (64).

comparisons in Figure 1 is the low-docking similarity between most ZF domains indicated by the large number of IAS scores below 5.0 (shown in white). Indeed as discussed above, there is even considerable variation in docking geometry within Zif268 groups. These results suggest that the majority of ZF domains would not provide good template structures for modeling the binding specificity of different ZF domains. For example, if based on Table 4 and Figure 5B we use an IAS cutoff score of 9.0 to define an appropriate template, only the Zif268 ZF1 and ZF3 domains would be sufficiently similar to effectively model one another (IAS values above 9 in Figure 1 region A). Furthermore, at an IAS cutoff of 9, many pairs of ZF domains *within* the large Zif268 groups are not similar enough in docking geometry to provide accurate specificity predictions. How then does one know in advance whether a known structure can be used as a

template for the prediction of the binding specificity for a protein of unknown structure?

In the case of Zif268 groups, we have identified two factors that affect docking geometry. First, dimerization appears to induce deviations from the wild-type docking geometry. For example, with the exception of ZF2 from 1llm chain D, all ZF domains in the dimeric ZF protein structures 1f2i and 1llm exhibit non-wild-type docking (Figure 1, Supplementary Figure S1). Perhaps surprisingly, dimerization affects *not* only the ZF domains that participate in the dimer interface but the ZF domains which are adjacent to the interacting domains are also bound in a non-wild-type manner. Wolfe *et al.* (65) noted that the DNA molecules in the 1llm and 1f2i complexes were not as significantly under-wound as in the wild-type Zif268 complex, which would explain why all ZF domains were affected. The three ZF domains from chain B of the 1p47 tandem dimer complex (Table 1) similarly exhibit non-wild-type docking. The observation that only ZF domains from chain B were affected may be an example of asymmetric binding of homodimeric protein–DNA complexes (29), where monomers are observed to bind in similar but different conformations.

Docking geometries can also be affected by specific hydrogen-bond interactions. In the wild-type Zif268 ZF3 structure, the Arg(−1) side-chain makes two bidentate hydrogen-bond interactions: one with the Gua(3) base and one with the Asp(2) side-chain (numbering as in Figure 2). The Arg(−1):Asp(2) interaction is believed to help orient the Arg sidechain so as to interact favorably with Gua(3) (39). A further bidentate hydrogen bond is made between the Arg(6) and Gua(1) residues. Despite the considerable variability in docking geometry for different ZF domains, in some cases high IAS values were observed between different ZF domains both from within a single protein and from different proteins. The three highest scoring such similarities (Figure 1; regions A, B and C) correlated with the presence of this set of three bidendate hydrogen-binding interactions. Region A identifies the high IAS values between the wild-type-docking Zif268 ZF1 and ZF3 domains. Regions B and C identify the strong similarity between the wild-type-docking Zif268 ZF1 and ZF3 domains, and the ZF3 domain from 1mey. Apart from the Zif268 ZF1 and ZF3 domains, 1mey ZF3 domains are the only other ZF domains with Arg residues at both canonical position—1 and 6 that make bidentate interactions with each Gua base of a GXG base triplet (Figure S1, RDHR:GAG).

We suggest that the high pair-wise IAS values observed in regions A, B and C are a result of a common docking geometry required to correctly form the set of three, stereo-specific, bidentate hydrogen bonds. This is supported by the observation that mutations that disrupt one of these bidentate interactions perturb the ZF docking geometry. Eight of nine ZF1 domains in the Zif268 ZF1 group (seven 1a1* complexes; 1jk2 complex; Table 1 and Supplementary Figure S1) that have mutations that change at least one of Arg(−1), Asp(2) or Gua(3) exhibit a non-wild-type docking geometry. Interestingly, in contrast to the effect of dimerization discussed previously, the altered docking geometry of these ZF1 domains does not appear to affect the neighboring ZF domains which all dock to the DNA in a wild-type fashion.

DISCUSSION

The increasing number of protein–DNA complexes whose three-dimensional structures are available offers the possibility of using structural information to identify CREs selected by a particular TF. A number of workers have already shown that it is possible to reproduce known specificity patterns, i.e. in the form of PWMs or top-ranked sequences, starting with the structure of a known complex and changing the identity of bases in the interface (30,32,33,35,41,42). The results shown in Table 3 confirm that three-dimensional structures can be used as a basis for calculating the DNA-binding specificity of TFs. In addition, they extend previous studies in the sense that the calculations are successful in identifying a series of high-affinity binding sequences. Paillard *et al.* (31) also reported good agreement with the high-affinity sequences identified by Bulyk *et al.* (51); however the TTG sequence was not identified, perhaps due to the fact that they did not account for alternate side-chain conformations. Endres *et al.* (32) predicted a PWM for the full-length Zif268 protein in excellent agreement with experimental data; however, a detailed comparison with the Bulyk *et al.* data (i.e. the ranking of the seven top sequences from Table 3) was not carried out. TGG was correctly identified as the highest-affinity site by Havranek *et al.* (33) using an atomic-level modeling approach and by Liu *et al.* (30) using a knowledge-based potential. However, Havranek *et al.* did not report results for other than the top-scoring sequence while of the high-affinity sequences reported by Liu *et al.* only TTG matched one of the top seven from Table 3. The close agreement we report here between our top predictions and all seven of the top experimentally determined sequences illustrates that an all-atom modeling approach that incorporates side-chain flexibility can yield predictions in excellent agreement with experiment for a series of high-affinity sequences. The importance of accounting for side-chain flexibility is demonstrated by the fact that the identification of the high-affinity TTG sequence (Table 3) required the prediction that His(3) rotates out of the major groove (Figure 4).

The applicability of structure-based predictions would be dramatically enhanced if a particular structure could be used as a template for specificity predictions on a much larger set of structural homologs. That is, given a structure for one member of a TF family it would be of great value to be able to predict high-affinity binding sequences to other family members. Recent articles by Morozov *et al.* (42) and Contreras-Moreira and Collado-Vides (35) have attacked this problem using all-atom modeling and knowledge-based approaches, respectively, to predict PWMs using structural homologs as templates. Although some results were in good agreement with experimentally determined PWMs, this was not uniformly the case. Both groups discuss the need to choose an appropriate template structure to achieve accurate results

and propose the use of interfacial-residue sequence identify as a guide in selecting an appropriate template. Contreras-Moreira and Collado-Vides introduce two RSMD-based measures of interface similarity between structurally superimposed template complexes (one for protein and one for DNA) and demonstrate a strong correlation between protein-sequence identity and their similarity measures. However, we have seen here that sequence identity does not in itself guarantee strong similarity in docking geometry since effects such as domain–domain dimerization can effect this geometry, even for identical sequences.

In this article, we have carried out a systematic study of the effect of docking geometry on the accuracy of binding-specificity predictions. Our recently developed *interface alignment score* (IAS) (48) is used as a measure of the similarity in the docking geometry of different protein–DNA complexes. Binding-specificity predictions for the Zif268 ZF2 domain showed excellent agreement with experiment when the protein–DNA interface was very similar ($9.0 < IAS_{Zif268}$) to that seen in the high-resolution wild-type structure (Table 3 and 4, Figure 5). The most accurate predictions, highlighted in Table 3, were achieved when the hrZif268 ZF2 structure was used as the template for itself. As can be seen in Table 4, the accuracy of TF-binding-site predictions is extremely sensitive to the docking geometry of the template that is used. The good results obtained when templates with near wild-type geometries were used begin to degrade with IAS_{Zif268} scores below ~9.0, and below a score of 5.0 the predictions are incorrect (Table 4, Figure 5). A pair-wise IAS value above 9.0 appears to provide a reasonably high-confidence prediction that two structures will result in similar binding-specificity predictions.

Based on the results of Figure 1 which show that the majority of pairs of ZF domains have IAS values below 5.0. It appears that many ZF domains could not be used as templates to model the binding specificity of another domain. This problem is likely to be quite general and indeed we have seen in our previous study that there are often significant variations in docking geometry within a protein family (48). In some cases, as has been suggested previously (35,42), it may be possible to select an appropriate template structure by maximizing sequence identity over the interfacial amino acids. Our observations that a common set of Arg-Gua and Asp-Arg bidentate hydrogen-bond interactions correlated with high-docking similarity suggests that conserved stereo-specific hydrogen-bond interactions, might be particularly informative for template selection. More generally, structural and computational studies of individual protein families aimed at understanding the sequence determinants of docking geometry variations would be of particular value.

An alternate approach is to move beyond the docking geometries provided by existing template complexes. One potential strategy is to start with a bound complex and to sample docking space by altering the DNA conformation while keeping the protein backbone fixed. In this work, we applied the local conformational sampling *wriggling* algorithm (61) to the DNA backbone (keeping the phosphate atoms fixed) to produce small variations in the DNA–nucleotide positions so as to facilitate the search for optimal side-chain–base interactions (Materials and Methods). Extending this local sampling approach to include all endocyclic torsion angles and sugar-pucker conformations would provide a means to locally sample docking geometry around a given template structure. A recent article by Rohs *et al.* (66) introducing a promising new Monte Carlo sampling algorithm which can efficiently and accurately simulate DNA conformations provides another promising approach. Recent progress in protein–DNA docking that incorporates DNA flexibility offers an alternate approach by generating templates *de novo* (67), and could be used in conjunction with local DNA sampling to further sample docking space. Approaches such as these could alleviate any inherent limitations due to the docking geometries of available template structures. We note that our *interface alignment* algorithm would provide a particularly useful method for analyzing the sampling characteristics of these algorithms and could be used to characterize both the degree of docking geometry sampled and how these docking geometries relate to those of known structures.

The phenomenon that different members within a protein family exhibit variation in their docking geometries is not restricted to the C_2H_2 ZF proteins. Previously, we performed a pair-wise comparison of the docking geometry for the Homeodomain recognition helices (48) and similarly found that many comparisons had IAS values below 9.0. Indeed, it is likely that docking variation, at the level described here as functionally relevant for atomic-level modeling, will be a characteristic of many other protein families as well. Therefore, future approaches that address template-docking variation, such as refined template selection or relaxing the rigid-template approach and sampling docking space, will allow us to more effectively use the available template structures to predict binding specificity of whole TF families. This, in turn, will provide an invaluable tool for addressing the current paucity of TF-binding-specificity data.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Cinque Soto and Antonina Silkov for helpful discussions; Donald Petrey for considerable help with use of the TROLL structural-modeling C++ libraries; and Remo Rohs for critical reading of the manuscript.

Conflict of interest statement. None declared.

REFERENCES

1. Levine,M. and Davidson,E.H. (2005) Gene regulatory networks for development. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 4936–4942.
2. Collins,F.S., Green,E.D., Guttmacher,A.E. and Guyer,M.S. (2003) A vision for the future of genomics research. *Nature*, **422**, 835–847.
3. Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.
4. Bulyk,M.L. (2003) Computational prediction of transcription-factor binding site locations. *Genome Biol.*, **5**, 201.

5. Frith,M.C., Hansen,U. and Weng,Z. (2001) Detection of cis-element clusters in higher eukaryotic DNA. *Bioinformatics*, **17**, 878–889.
6. Markstein,M., Markstein,P., Markstein,V. and Levine,M.S. (2002) Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 763–768.
7. Berman,B.P., Nibu,Y., Pfeiffer,B.D., Tomancak,P., Celniker,S.E., Levine,M., Rubin,G.M. and Eisen,M.B. (2002) Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 757–762.
8. Wildonger,J., Sosinsky,A., Honig,B. and Mann,R.S. (2005) Lozenge directly activates argos and klumpfuss to regulate programmed cell death. *Genes Dev.*, **19**, 1034–1039.
9. Wasserman,W.W. and Fickett,J.W. (1998) Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.*, **278**, 167–181.
10. Krivan,W. and Wasserman,W.W. (2001) A predictive model for regulatory sequences directing liver-specific transcription. *Genome Res.*, **11**, 1559–1566.
11. Halfon,M.S., Grad,Y., Church,G.M. and Michelson,A.M. (2002) Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model. *Genome Res.*, **12**, 1019–1028.
12. Loots,G.G., Ovcharenko,I., Pachter,L., Dubchak,I. and Rubin,E.M. (2002) rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res.*, **12**, 832–839.
13. Blanchette,M., Schwikowski,B. and Tompa,M. (2002) Algorithms for phylogenetic footprinting. *J. Comput. Biol.*, **9**, 211–223.
14. Wasserman,W.W., Palumbo,M., Thompson,W., Fickett,J.W. and Lawrence,C.E. (2000) Human-mouse genome comparisons to locate regulatory sites. *Nat Genet.*, **26**, 225–228.
15. Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
16. Sandelin,A., Alkema,W., Engstrom,P., Wasserman,W.W. and Lenhard,B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32**, D91–94.
17. Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
18. Oliphant,A.R., Brandl,C.J. and Struhl,K. (1989) Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast GCN4 protein. *Mol. Cell. Biol.*, **9**, 2944–2949.
19. 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.*, **29**, 2471–2478.
20. Ren,B., Robert,F., Wyrick,J.J., Aparicio,O., Jennings,E.G., Simon,I., Zeitlinger,J., Schreiber,J., Hannett,N. *et al.* (2000) Genome-wide location and function of DNA binding proteins. *Science*, **290**, 2306–2309.
21. Liu,X., Noll,D.M., Lieb,J.D. and Clarke,N.D. (2005) DIP-chip: rapid and accurate determination of DNA-binding specificity. *Genome Res.*, **15**, 421–427.
22. Bulyk,M.L., Gentalen,E., Lockhart,D.J. and Church,G.M. (1999) Quantifying DNA-protein interactions by double-stranded DNA arrays. *Nat. Biotechnol.*, **17**, 573–577.
23. Harbison,C.T., Gordon,D.B., Lee,T.I., Rinaldi,N.J., Macisaac,K.D., Danford,T.W., Hannett,N.M., Tagne,J.B., Reynolds,D.B. *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.
24. Kaplan,T., Friedman,N. and Margalit,H. (2005) Ab initio prediction of transcription factor targets using structural knowledge. *PLoS Comput. Biol.*, **1**, e1.
25. 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.
26. 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.
27. Kono,H. and Sarai,A. (1999) Structure-based prediction of DNA target sites by regulatory proteins. *Proteins*, **35**, 114–131.
28. Mandel-Gutfreund,Y., Baron,A. and Margalit,H. (2001) A structure-based approach for prediction of protein binding sites in gene upstream regions. *Pac. Symp. Biocomput.*, 139–150.
29. Selvaraj,S., Kono,H. and Sarai,A. (2002) Specificity of protein-DNA recognition revealed by structure-based potentials: symmetric/asymmetric and cognate/non-cognate binding. *J. Mol. Biol.*, **322**, 907–915.
30. Liu,Z., Mao,F., Guo,J.T., Yan,B., Wang,P., Qu,Y. and Xu,Y. (2005) Quantitative evaluation of protein-DNA interactions using an optimized knowledge-based potential. *Nucleic Acids Res.*, **33**, 546–558.
31. Paillard,G., Deremble,C. and Lavery,R. (2004) Looking into DNA recognition: zinc finger binding specificity. *Nucleic Acids Res.*, **32**, 6673–6682.
32. 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.
33. 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.
34. Morozov,A.V., Kortemme,T., Tsemekhman,K. and Baker,D. (2004) Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 6946–6951.
35. Contreras-Moreira,B. and Collado-Vides,J. (2006) Comparative footprinting of DNA-binding proteins. *Bioinformatics*, **22**, e74–80.
36. Olson,W.K., Gorin,A.A., Lu,X.J., Hock,L.M. and Zhurkin,V.B. (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 11163–11168.
37. Thayer,K.M. and Beveridge,D.L. (2002) Hidden Markov models from molecular dynamics simulations on DNA. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 8642–8647.
38. Steffen,N.R., Murphy,S.D., Lathrop,R.H., Opel,M.L., Tolleri,L. and Hatfield,G.W. (2002) The role of DNA deformation energy at individual base steps for the identification of DNA-protein binding sites. *Genome Inform. Ser Workshop Genome Inform.*, **13**, 153–162.
39. 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.
40. Bulyk,M.L., Johnson,P.L. and Church,G.M. (2002) Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res.*, **30**, 1255–1261.
41. Paillard,G. and Lavery,R. (2004) Analyzing protein-DNA recognition mechanisms. *Structure (Camb.)*, **12**, 113–122.
42. Morozov,A.V., Havranek,J.J., Baker,D. and Siggia,E.D. (2005) Protein-DNA binding specificity predictions with structural models. *Nucleic Acids Res.*, **33**, 5781–5798.
43. Zhang,C., Liu,S., Zhu,Q. and Zhou,Y. (2005) A knowledge-based energy function for protein-ligand, protein-protein, and protein-DNA complexes. *J. Med. Chem.*, **48**, 2325–2335.
44. Lafontaine,I. and Lavery,R. (2000) ADAPT: a molecular mechanics approach for studying the structural properties of long DNA sequences. *Biopolymers*, **56**, 292–310.
45. Lafontaine,I. and Lavery,R. (2000) Optimization of nucleic acid sequences. *Biophys. J.*, **79**, 680–685.
46. Garvie,C.W. and Wolberger,C. (2001) Recognition of specific DNA sequences. *Mol. Cell.*, **8**, 937–946.
47. 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.
48. Siggers,T.W., Silkov,A. and Honig,B. (2005) Structural alignment of protein–DNA interfaces: insights into the determinants of binding specificity. *J. Mol. Biol.*, **345**, 1027–1045.
49. Choo,Y. and Klug,A. (1997) Physical basis of a protein-DNA recognition code. *Curr. Opin. Struct. Biol.*, **7**, 117–125.

50. Wolfe,S.A., Nekludova,L. and Pabo,C.O. (2000) DNA recognition by Cys2His2 zinc finger proteins. *Annu. Rev. Biophys. Biomol. Struct.*, **29**, 183–212.
51. Bulyk,M.L., Huang,X., Choo,Y. and Church,G.M. (2001) Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 7158–7163.
52. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
53. Petrey,D. and Honig,B. (2003) GRASP2: visualization, surface properties, and electrostatics of macromolecular structures and sequences. *Methods Enzymol.*, **374**, 492–509.
54. Elrod-Erickson,M., Rould,M.A., Nekludova,L. and Pabo,C.O. (1996) Zif268 protein-DNA complex refined at 1.6A: a model system for understanding zinc finger-DNA interactions. *Structure*, **4**, 1171–1180.
55. Xiang,Z. and Honig,B. (2001) Extending the accuracy limits of prediction for side-chain conformations. *J. Mol. Biol.*, **311**, 421–430.
56. Ponder,J.W. and Richards,J. (1987) An efficient Newton-like method for molecular mechanics energy minimization of large molecules. *J. Comput. Chem.*, **8**, 1016–1025.
57. Cheatham,T.E.,III, Cieplak,P. and Kollman,P.A. (1999) A modified version of the Cornell et al. force field with improved sugar pucker phases and helical repeat. *J. Biomol. Struct. Dyn.*, **16**, 845–862.
58. Chen,Y., Kortemme,T., Robertson,T., Baker,D. and Varani,G. (2004) A new hydrogen-bonding potential for the design of protein-RNA interactions predicts specific contacts and discriminates decoys. *Nucleic Acids Res.*, **32**, 5147–5162.
59. 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.
60. Hingerty,B.E., Ritchie,R.H., Ferrell,T.L. and Turner,J.E. (1985) Dielectric effects in biopolymers: the theory of ionic saturation revisited. *Biopolymers*, **24**, 427–439.
61. Lavery,R., Zakrzewska,K. and Sklenar,H. (1995) JUMNA (JUnction Minimization of Nucleic Acids). *Comput. Phys. Commun.*, **91**, 135–158.
62. Cahill,S., Cahill,M. and Cahill,K. (2003) On the kinematics of protein folding. *J. Comput. Chem.*, **24**, 1364–1370.
63. Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.
64. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
65. Wolfe,S.A., Grant,R.A. and Pabo,C.O. (2003) Structure of a designed dimeric zinc finger protein bound to DNA. *Biochemistry*, **42**, 13401–13409.
66. Rohs,R., Sklenar,H. and Shakked,Z. (2005) Structural and energetic origins of sequence-specific DNA bending: Monte Carlo simulations of papillomavirus E2-DNA binding sites. *Structure (Camb.)*, **13**, 1499–1509.
67. van Dijk,M., van Dijk,A.D., Hsu,V., Boelens,R. and Bonvin,A.M. (2006) Information-driven protein-DNA docking using HADDOCK: it is a matter of flexibility. *Nucleic Acids Res.*, **34**, 3317–3325.