4564–4576 *Nucleic Acids Research, 2011, Vol. 39, No. 11* *Published online 22 February 2011*
doi:10.1093/nar/gkr048

Extensive protein and DNA backbone sampling improves structure-based specificity prediction for C_2H_2 zinc fingers

Chen Yanover and Philip Bradley*

Program in Computational Biology, Fred Hutchinson Cancer Research Center, Seattle, WA 98109-1024

Received November 2, 2010; Revised January 17, 2011; Accepted January 18, 2011

ABSTRACT

Sequence-specific DNA recognition by gene regulatory proteins is critical for proper cellular functioning. The ability to predict the DNA binding preferences of these regulatory proteins from their amino acid sequence would greatly aid in reconstruction of their regulatory interactions. Structural modeling provides one route to such predictions: by building accurate molecular models of regulatory proteins in complex with candidate binding sites, and estimating their relative binding affinities for these sites using a suitable potential function, it should be possible to construct DNA binding profiles. Here, we present a novel molecular modeling protocol for protein-DNA interfaces that borrows conformational sampling techniques from *de novo* protein structure prediction to generate a diverse ensemble of structural models from small fragments of related and unrelated protein-DNA complexes. The extensive conformational sampling is coupled with sequence space exploration so that binding preferences for the target protein can be inferred from the resulting optimized DNA sequences. We apply the algorithm to predict binding profiles for a benchmark set of eleven C_2H_2 zinc finger transcription factors, five of known and six of unknown structure. The predicted profiles are in good agreement with experimental binding data; furthermore, examination of the modeled structures gives insight into observed binding preferences.

INTRODUCTION

The prediction of macromolecular interactions is a key challenge for computational molecular biology. Given that molecular interactions are determined by the 3D structures and chemical properties of the interacting partners, it seems plausible that such interactions (or at any rate their propensity to occur *in vitro*) could be predicted by structural modeling. Although a general approach to selecting binding partners for any target protein—large-scale docking and binding affinity calculations against all possible cellular partners—remains infeasible, considerable progress has been made in predicting (1,2) and designing (3–5) specific macromolecular interactions. One class of macromolecular interactions that represents a promising target for structure-based prediction consists of those interactions which are mediated by a linear sequence motif (peptide, DNA, RNA) in the partner molecule. For these interactions, the space of possible binding partners can be enumerated concisely, and approximate binding modes can often be inferred from the structures of related complexes. Moreover, these motif-mediated interactions are of central importance in cellular regulation: the interactions of transcription factors, kinases, splicing factors and peptide-recognition modules can all be usefully characterized in terms of linear motif specificity. In this work, we describe a new approach for structure-based prediction of protein-DNA binding specificity, one that can be extended to other classes of motif-mediated interaction. By applying simulation techniques borrowed from *de novo* structure prediction, we demonstrate that modeling of protein-DNA interactions can be improved through the use of large-scale conformational sampling of both partners.

Structural modeling has generated important insights into protein-DNA recognition mechanisms, from studies of the relative contributions of direct and indirect recognition (6–8), and the role of DNA shape (9,10) and interfacial waters (11,12) in sequence-specific recognition, to the validity of the additivity assumption in protein-DNA energetics (13,14). Structural modeling has also been used to predict DNA binding preferences, using a wide range of sampling algorithms and energy functions, including database-derived potentials (15–17), all-atom molecular mechanics force fields (11,18–23), and hybrid scoring

*To whom correspondence should be addressed. Tel: +1 206 667 7041; Fax: +1 206 667 1319; Email: pbradley@fhcrc.org

functions (12,14). These approaches can often generate highly accurate predictions when given an X-ray crystal structure of the target protein in complex with a high affinity binding site. To be widely applicable for functional annotation, however, a method must also be able to generate accurate predictions for proteins whose structures have not been solved (by using the structures of related proteins as modeling templates, for example). This has proven much more difficult, particularly for methods that attempt to build realistic models of interface sidechains through the use of atomistic modeling and high-resolution force fields (14,21). In a recent study of template-based specificity predictions for the transcription factor Zif268 (21), Siggers and Honig found that prediction accuracy was highly sensitive to the structural similarity between the template and target interfaces. They identified a similarity threshold above which accurate template-based predictions could be made. Comparing all available structures for C_2H_2 zinc finger (ZF) family members, they found considerable variation in interface geometry, with the majority of potential target-template pairs having similarity values below their prediction accuracy threshold. For many targets, no suitable templates for specificity prediction were available in the structural database.

We hypothesized that the observed sensitivity of prediction accuracy to target-template similarity was partly due to limited conformational flexibility inherent in traditional template-based approaches. An approach that incorporated large-scale backbone conformational sampling might circumvent this limitation and allow for accurate specificity predictions for a wider range of target proteins. To test this hypothesis, we developed a novel interface fragment assembly protocol in which complete protein-DNA models are constructed from small fragments of related and unrelated protein-DNA complexes. Starting from an initial randomized pool of DNA binding sites, we construct a diverse ensemble of protein-DNA interface fragment assembly models. These models are taken as starting points for all-atom Monte Carlo (MC) refinement simulations that simultaneously explore sequence and structure space. By augmenting the molecular mechanics potential energy function with a DNA-sequence-dependent energy term that captures unbound DNA energies, we can directly infer binding preferences for alternate DNA sequences from their differential sampling rates in these MC simulations.

We applied this protocol to a set of transcription factors in the C_2H_2 ZF family (Figure 1) and demonstrated its efficacy in recapitulating protein-DNA interface structure and predicting ZF binding specificity. The ZF transcription factors constitute the largest family of eukaryotic transcription factors, making up nearly half of all annotated human transcription factors (25). Different family members have widely varying DNA-binding preferences—in contrast to many other TF families in which a core binding motif is conserved—making them challenging targets for template-based prediction. In addition, the C_2H_2 ZF proteins have served as a model system for studying protein-DNA recognition. High-resolution X-ray crystal structures are available for several family members, as are experimental binding data from various sources. ZF DNA recognition codes have been proposed based on examination of available strucural data (26–28), and machine learning algorithms have been developed that leverage the extensive set of binding data to make predictions of DNA binding specificity (29–33). These models provide a valuable external standard against which to judge the results of structural modeling. Finally, ZF proteins have been the subject of extensive protein engineering efforts: designed zinc finger proteins have been used as novel cellular regulators (34) and combined with non-specific nuclease domains to generate targeted, highly specific genome engineering tools (35). Structure-based modeling of ZF binding specificity could further these engineering efforts.

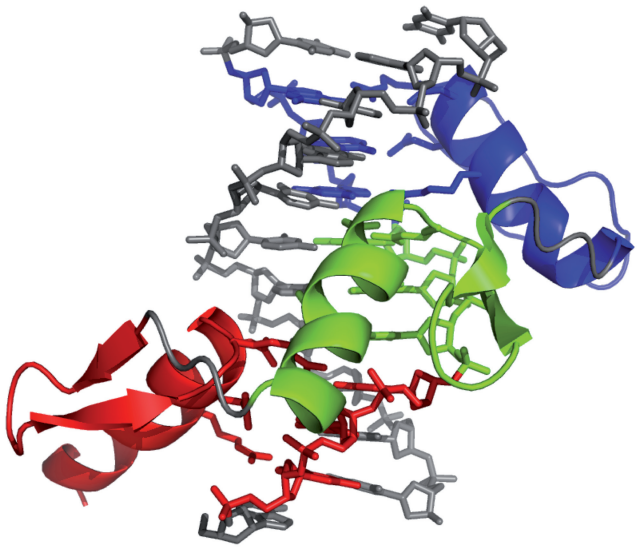

Figure 1. Structure of the C_2H_2 ZF transcription factor Zif268 (24). The three ZF domains are colored blue, green and red (from N- to C-terminus). In the canonical model for ZF–DNA interactions, individual fingers recognize neighboring three base pair sites (here colored to match the corresponding finger). Note that, by convention, the DNA triplets occur in reverse order (red, green, blue) when reading from 5′ to 3′ along the primary strand.

MATERIALS AND METHODS

Modeling preliminaries

All modeling protocols were implemented within the Rosetta (36) molecular modeling package, adapted for modeling of DNA by Havranek *et al.* (20), and will be released for free academic use through the Rosetta commons (http://www.rosettacommons.org). Sampling is conducted in internal coordinates (backbone and sidechain torsion angles, the protein-DNA rigid body orientation, and selected intra-DNA rigid body connections as described below), with Cartesian coordinates updated for evaluation of the energy function and its gradient. Bond lengths and angles are held fixed at ideal values.

Interface fragment assembly

The interface fragment assembly protocol combines three types of fragment replacement moves: double-helical fragment moves, in which a continuous stretch of base-paired DNA is replaced by a base paired region of equal length taken from a DNA duplex of known structure (Figure 2A); interface fragment moves, in which the orientation of the protein relative to the DNA is updated based on a template interface (Figure 2B); and protein fragment replacements, in which a continuous stretch of backbone torsion angles are taken from an aligned region of a template structure (Figure 2C). In DNA fragment replacement moves, the internal conformation of the double helix downstream of the fragment replacement (green region in Figure 2A) is unchanged, allowing these moves to preserve base-pairing between strands. The torsion angles in the fragment region are taken from the template, as are the relative orientations of the base-pairs within the fragment. DNA fragment insertion may introduce small breaks in the DNA backbone at the 3′-ends of the fragment region; these are minimized by adjustment of the backbone torsion angles at the fragment junctions after fragment insertion by gradient–descent optimization of a chain closure penalty. Interface fragment moves depend on the definition of two sets of guide atoms, a takeoff set in the DNA and a landing set in the protein. In making the move, the template structure (from which the fragment is taken) is aligned to the model by superimposing the takeoff set of guide atoms; the landing guide atoms in the model are then translated in space to superimpose onto the landing guide atoms in the template, thereby moving the protein relative to the DNA. After the fragment move, the relative orientation of the guide atoms is the same in the template interface and in the model, although the internal conformation of the protein and DNA at the interface will be different. In protein fragment insertions, a stretch of 3 or 9 consecutive residues in the current model is updated by replacing the backbone torsion angles (ϕ, ψ and ω) with corresponding torsion angles taken from a template structure (9-residue fragment insertions are used early in the simulation for internal consistency in early model build-up; less-perturbing 3-residue fragment insertions are used in the later stages to refine the model). Together these three fragment replacement moves provide a mechanism for assembling complete models of protein DNA interfaces guided by the template structures from which the fragments are taken.

To select fragments for a target ZF, we first assemble a list of template fingers from ZFs with solved structures (for benchmarking, we exclude highly similar templates, including the target itself if its structure has been solved). Three- and nine-residue protein fragments are taken from aligned regions of these template fingers; regions not covered by aligned fragments (e.g. if there are unusual insertions or deletions in the beta-strand region) are filled in from Rosetta's *de novo* fragment database. Protein–DNA interface fragments are also taken from these template fingers, using as takeoff guide atoms the C1′ atoms of both strands of the corresponding DNA triplet, and as landing guide atoms the C_α atoms of the canonical helix positions −1 through 6 (template zinc fingers are manually pre-processed to define the mapping between individual fingers and triplets in the binding site). A library of DNA double-helical fragments is selected at the start of each modeling simulation after the DNA sequence has been randomized, thereby guaranteeing that sequence-dependent variations in DNA structure will be sampled. Two, three and five base-pair fragments are selected from a database of solved protein-DNA complexes; fragments are chosen on the basis of DNA sequence similarity (for benchmarking, all ZF structures are excluded from the DNA fragment database). As with protein fragment moves, longer fragments are used earlier in the simulation to build up coherent DNA duplexes, while shorter fragments are used at the end to reduce perturbation of favorable contacts.

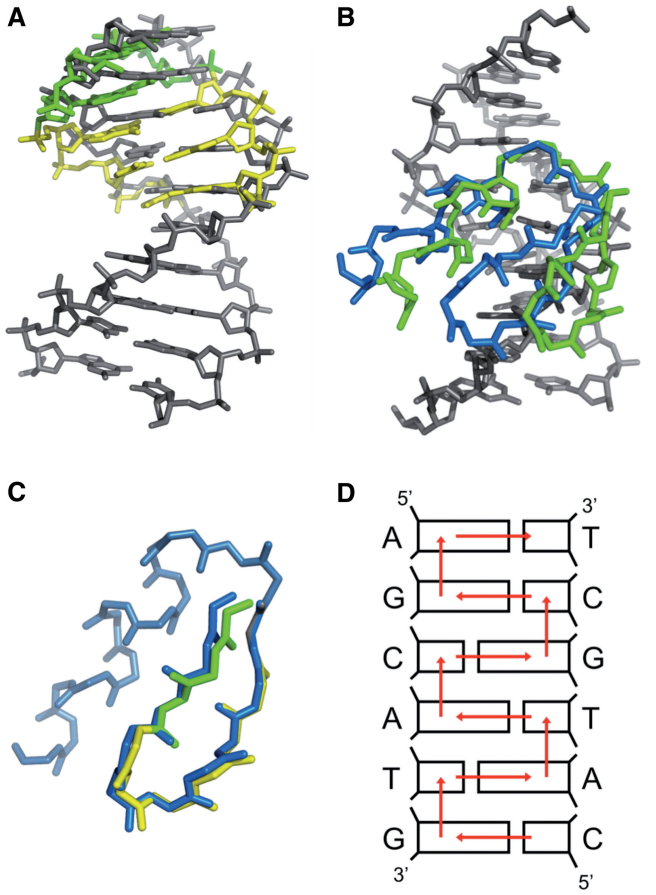

Figure 2. Fragment assembly moves. (**A**) In double-helical DNA fragment insertions, a base-paired region (yellow) taken from a template is inserted into the current model (gray), moving the downstream base pairs (green); the upstream region (gray) remains fixed. (**B**) Interface fragment moves transfer the relative protein–DNA orientation from a template structure without changing protein or DNA internal conformation. (**C**) Protein fragment replacement moves copy consecutive backbone torsion angles from a template into the current model. (**D**) To enable efficient gradient-based minimization of DNA structure in internal coordinates, a kinematic structure (*fold-tree*) is defined on the duplex that creates flexible rigid-body transforms between bases (red arrows) so that downstream base pairing is preserved by internal coordinate changes.

Interface modeling simulations

At the start of each interface fragment assembly simulation, the backbone torsion angles of the protein are initialized to extended values ($\phi = -150$, $\psi = 150$, $\omega = 180$), the DNA conformation is set to standard ideal B-form, and the protein and DNA are separated in space. The simulation proceeds in two phases: a low-resolution phase, in which the protein sidechains are represented by centroid pseudo-atoms and potential energies are calculated using a knowledge-based energy function, and a high-resolution phase, in which all atoms including hydrogens are explicitly modeled and a modified version of Rosetta's standard all-atom potential function is used (see below for details on the energy functions). The low-resolution simulation consists of 7500 MC trials, repeatedly cycling through the three types of fragment assembly moves. The length of the low-resolution simulation was set in order to maximize the quality and conformational diversity of the models, without over-optimizing the knowledge-based potential energy function. The high-resolution refinement simulation consists of 120 Monte Carlo plus Minimization (MCM) trials, in which a perturbation to the system is followed by (i) sidechain optimization at positions whose energy increased and (ii) gradient-based minimization of all flexible degrees of freedom. MCM moves were found to give better conformational sampling in the highly rugged all-atom landscape than standard MC moves (37). The perturbations are of two types: protein backbone moves, in which the backbone torsion angles of a small, randomly selected set of positions are changed (36), and DNA sequence mutation moves, in which a single base-pair in the binding site is mutated. Energetically biased acceptance of these DNA sequence moves gives rise to the sequence preferences seen in the final DNA sequences of the models. A single simulation takes roughly 30 min on a 2GHz processor.

DNA flexibility

Additional flexibility of the DNA duplex was incorporated into Rosetta's rotamer optimization and gradient-based minimization modules. DNA rotamers consisting of small random perturbations to the existing residue were built by applying the *wriggling* procedure (38) to the four backbone torsion angles ζ_{i-1}, α_i, β_i and γ_i, as suggested by Siggers and Honig (21). The wriggling procedure generates modifications to dihedral angles that will minimize downstream coordinate changes. During gradient-based minimization, a kinematic structure is defined on the DNA duplex (Figure 2D) that allows internal coordinate changes to propagate while preserving downstream base-pairing interactions. Using Rosetta's fold-tree internal coordinate framework (39), a ladder of flexible, rigid-body connections is introduced between the bases. With these connections in place, the flexible degrees of freedom for the DNA are the rigid-body transforms between bases and the backbone and sidechain dihedral angles. To allow direct base–base rigid-body connections, it is necessary to introduce chainbreaks between successive DNA residues (for kinematics, the graph of atomic connectivity must be acyclic). These chainbreaks are tethered by addition of a pseudo-energy term to the potential function. A rigid-body linkage is introduced between the central base of the DNA binding site and the closest protein position to define the complete kinematic tree for the system; with this framework, changes to the protein dihedral angles induce conformational updates that propagate outward from the interface.

Calculating relative binding affinities

In making predictions of DNA interaction specificity, our goal is to calculate relative binding affinities of the target protein for a large set of alternative binding sequences. This is done by allowing transitions between DNA sequences during the simulations via the sequence mutation MC moves; the relative frequencies with which different sequences are sampled in the final models are taken as indicators of their relative binding affinities. Here we are implicitly using a thermodynamic cycle of the form given in Figure 3 for the special case of two sequences (11). To calculate $\Delta\Delta G$, the difference between the two binding affinities, we can instead compute the difference between the ΔG^{mut} values for mutating from sequence 1 to sequence 2 in the bound and unbound states. In principle, these two ΔG values could be estimated from transition probabilities between the two sequences in MC simulations of the bound and unbound states. One challenge is that these simulations would be inherently biased by the differing internal energies of the base-pairs and base-steps: G–C base pairs have greater electrostatic interaction energies due to an additional Watson–Crick hydrogen bond; pyrimidine–purine base steps have weaker stacking interactions than purine–pyrimidine steps due to the geometry of B-form DNA (40), etc. This would in turn lead to highly biased and inefficient exploration of DNA sequence space, although the difference in sampling frequencies between the bound and unbound simulations could in principle serve in estimating relative affinities, given sufficient sampling. Our solution is to introduce a DNA-sequence-dependent correction term (E_{DNA}) to the potential energy function that captures the unbound energy of a given DNA sequence so as to balance sequence sampling in the unbound simulations. Adding this term to the energies of the bound and unbound systems in the cycle in Figure 3 doesn't change the ΔG's of binding, but it does change the ΔG's of mutation; in particular the unbound ΔG of mutation goes to 0.0, allowing us to calculate $\Delta\Delta G$ of binding by estimating the ΔG of mutation in the bound state. To capture base-pair and base-step energy

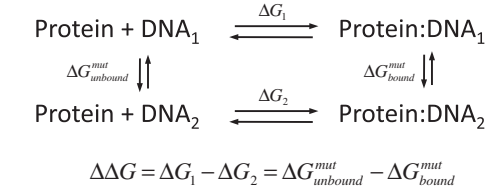

Figure 3. Thermodynamic cycle used to calculate relative binding affinities for two DNA sequences.

differences, we parameterize E_{DNA} as a sum of base-step coefficients. The E_{DNA} terms for the 10 base-steps were fit in an iterative fashion by performing 1000 unbound DNA simulations—each with the same number of MC sequence mutation moves as a bound simulation—calculating frequencies of the different base-steps, updating the E_{DNA} terms, and re-running the unbound simulations until the base-step sampling frequencies converged to approximately equal levels. A similar approach has been used by Endres and Wingreen (41) to estimate unbound energies, although their unbound energy contribution is based on base-step energies rather than sampling frequencies.

Binding specificity predictions for zinc finger proteins

To generate a binding specificity profile for a C_2H_2 ZF, we first parsed the protein sequence into individual fingers using the Pfam (43) `zf-C2H2` profile hidden Markov model. Binding simulations were conducted as described above for each of the fingers individually. In each binding simulation, we consider the DNA binding site to consist of a 5 base-pair region centered on the canonical triplet. The complete DNA molecule consists of the 5 base-pair binding site together with an additional G:C base pair on either side to provide structural context. The DNA sequence of the binding site is randomized at the start of each independent simulation and optimized during the all-atom MC simulation through the energetically biased acceptance of DNA mutation moves. Due to the limited number of mutation moves attempted during each simulation, the raw sequence preferences in the final DNA sequences are rather weak. As an estimate of the true binding preferences we boost the raw profile by taking all frequencies to the sixth power and renormalizing (this corresponds to a linear rescaling of energies, under the mapping between probabilities and energies given by the Boltzmann distribution). Note that this boosting procedure doesn't change the ordering of the bases; instead it is designed to provide an estimate of what the fully converged sequence preferences would be. The choice of exponent is somewhat arbitrary, and was based on inspection of frequency profiles from a limited number of very long MC simulations. To facilitate comparison with experimental binding data, we computed a position-specific frequency matrix (PFM) for the complete protein by combining the binding profiles for the individual fingers as indicated in Figure 4. When combining single-finger PFMs, internal fingers contribute only the three core triplet columns, while terminal fingers contribute additional context on either side (Figure 4). To construct a simple position-specific scoring matrix (PSSM) from this PFM, we take the logarithm of the frequencies after dividing by a uniform background of 0.25.

Potential energy functions

The potential energy function for the low-resolution phase of the interface fragment assembly protocol has three components: a DNA internal energy term, a protein internal energy term and a protein-DNA interaction term. To calculate the internal energy of the DNA, we use a modified version of the database-derived potential

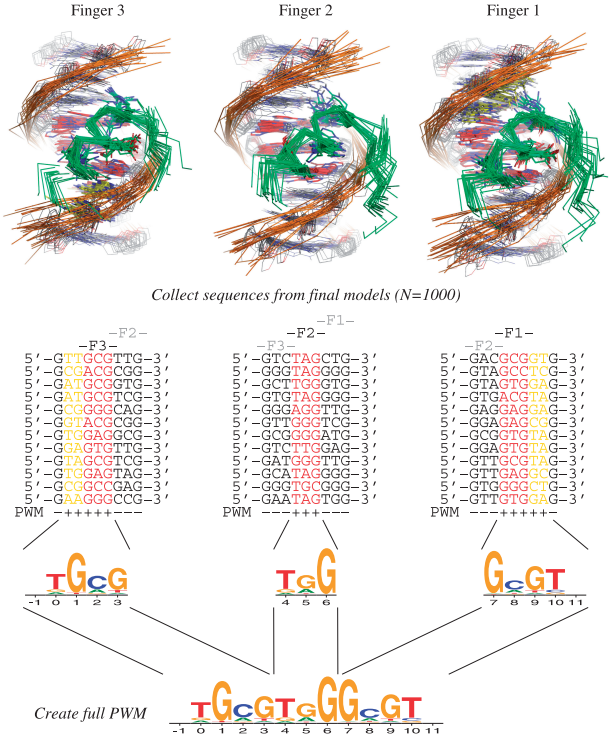

Figure 4. Binding specificity predictions. To generate a PFM for a poly-ZF protein, we perform binding simulations on individual ZF domains and combine the results into a single specificity profile. Simulation results for the 3-finger ZF protein Zif268 are shown. At the top, a subset of the final protein–DNA interface models are superimposed. Green ribbons are used to depict the protein, with key specificity determining sidechains shown; the DNA is portrayed in stick representation with a phosphate backbone ribbon. Carbons in the core triplet binding site are colored red. Carbons in neighboring bases that contribute to the final combined PFM are colored yellow. DNA sequence preferences calculated from the final models in 1000 independent simulations are used to construct single-finger PFMs (middle), which are combined into a binding profile for the complete protein (bottom). PFMs are depicted as sequence logos using the program WebLogo (42); structure images were generated with the PyMOL molecular graphics program.

introduced by Olson *et al.* (44), and extended to base-pairs by Morozov *et al.* (14). In these potentials, the internal DNA energy is a harmonic function of canonical base-step and base-pair parameters (roll, tilt, shift, etc); the force constants for the potential are derived from statistical analysis of base-pair and base-step geometry in crystal structure DNA. We modified the original base-step energy function to score base-to-base ($i \rightarrow i+1$) interactions rather than base pair–base pair interactions to give a more fine-grained potential (double-helical fragment insertions can introduce strand-specific strained geometry at fragment junctions). The protein internal energy is calculated using Rosetta's standard low-resolution potential function, which has terms capturing van der Waals interactions, residue environment and residue-pair preferences, and backbone torsion strain (45). To calculate the protein–DNA interaction energy, we parameterized a knowledge-based interaction potential modeled on Rosetta's protein potential that includes a residue-environment term capturing each amino acid's

propensity to occur at the protein–DNA interface, and an amino acid–base interaction term derived from frequencies of protein–DNA contacts. Details of the parameterization can be found in the Supplementary Data.

A modified version of Rosetta's all-atom potential energy function (46) was used to calculate all-atom energies during the high-resolution phase of the fragment assembly protocol. Three changes were made to the potential. As described above, a E_{DNA} was added to capture the energy of the DNA molecule in the unbound state. Rosetta's database-derived, residue-pair potential (the *fa_pair* term) was replaced with a weak, short-ranged explicit electrostatics term. In this term, a simple, linearly increasing distance-dependent dielectric ($\epsilon = 20r$) was used to model solvent screening effects, with all interactions truncated at 5.5 Å, thereby preserving the short-ranged nature of the all-atom potential. Incorporation of the explicit electrostatics term in addition to Rosetta's orientation-dependent hydrogen bonding potential (47) helps to prevent unfavorable short-range electrostatic interactions, modulates the interaction strength of charged and polar hydrogen bonds, and rewards electrostatic interactions with the phosphate backbone. The third modification was to the Lazaridis–Karplus (LK) implicit solvent model (48) used in Rosetta. In this model, the interaction energy for two atoms depends only on their distance and atom types, not on their relative orientation. This neglects the fact that interactions with solvent are anisotropic: polar atoms typically have preferred hydrogen-bonding directions, for example. We found that an isotropic solvation model did a poor job of capturing the many stacking interactions seen at protein–DNA interfaces (and the stacking of the DNA bases themselves). In these stacking interactions, packing around polar atoms is typically out of the plane of their preferred hydrogen bonding interactions, so that these may still be satisfied by forming hydrogen bonds within the macromolecular system or with solvent. In the classic LK model these packing interactions are as unfavorable as directly occluding the hydrogen bonding groups. We created a simple, orientation-dependent variant of the LK model in which the isotropic solvation energy for bringing an atom near a polar atom is multiplied by a scaling factor based on the distance between the desolvating atom and the nearest optimal water location (details can be found in the Supplementary Data). With this modification, the stacking interactions found at protein–DNA interfaces are not considered to be as unfavorable as interactions that directly prevent polar atoms from forming hydrogen bonds with water.

Assessment of prediction accuracy

To assess the performance of the interface fragment assembly protocol, we assembled a benchmark of C_2H_2 ZFs with available experimental binding data (Figure 7). We chose ZF proteins of known structure with 2–4 ZFs, and added to these a set of yeast 2-finger ZF transcription factors of unknown structure whose binding specificity had been profiled by protein binding microarrays (49).

We used two measures to assess the accuracy of ZF structure predictions. To measure similarity in protein–DNA interface orientations, we computed the Interface Alignment Score (IAS) of Siggers *et al.* (50), which quantifies and compares the spatial orientation between the backbones of interface amino acids and their neighboring bases. As a measure of interface sidechain accuracy and successful recovery of specificity-determining contacts, we calculated the fraction of native protein–DNA hydrogen bonds also present in the model, restricting to hydrogen bonds involving major groove atoms in the binding site.

To assess the accuracy of binding specificity predictions, we used a simple metric that counted the number of positions at which the preferred base in prediction and experiment agree. We also implemented a simplified variant of the BLiC score, a recently introduced similarity measure for comparison of PFM columns (51). The BLiC score is based on the Jensen–Shannon divergence, a standard measure of the distance between two probability distributions. The BLiC score has the attractive feature that similarity between two perfectly flat columns is rewarded less than similarity between two information-rich columns (51); in addition, we have found that BLiC scores above zero generally correspond to similar PFM columns, allowing easy assessment of the numerical values. We calculated the similarity score between two PFM columns P and Q using the equation

$$BLiC = JSD(P + Q, B) - JSD(P, Q), \quad (1)$$

where $JSD(X, Y)$ equals the Jensen–Shannon divergence between the probability distributions X and Y; B is the uniform distribution (0.25, 0.25, 0.25, 0.25); $P+Q$ is the average of the two distributions P and Q. For each experimental binding profile column, we assign P-values to observed BLiC similarity scores for that column based on the distribution of BLiC scores seen when comparing 10 000 random PFM columns to that experimental column. Random columns were built by generating four samples from a uniform distribution and normalizing their sum to 1.0.

RESULTS

We have developed a structure-based approach for predicting the specificity of protein–DNA interactions. In this approach, a large number of independent MC folding and binding simulations are conducted, simultaneously sampling the conformation of the protein and DNA as well as the DNA binding site sequence. The binding site sequence is randomized at the start of each independent simulation, and the MC moves by which we explore sequence space are random; nonetheless, sequence preferences emerge in the final models due to the energetically biased acceptance of these sequence moves. By incorporating a DNA-sequence-dependent energy term into our potential energy function that explicitly balances transition probabilities in the unbound state, we are able to use these sequence preferences to estimate relative binding affinities. For comparison with

experimental binding profiles, we construct a PFM from the base frequencies in the final models, although we note that this represents only one projection of the full diversity of optimized binding site sequences. As a test of our approach, we have conducted binding simulations for a collection of naturally occurring and engineered DNA binding proteins in the C_2H_2 ZF family.

Fragment assembly of unbound DNA

We first asked whether the DNA fragment assembly protocol is able to generate acceptable models of unbound DNA duplexes. Recall that the double-helical fragments that make up our DNA fragment library are taken from crystal structures of protein–DNA complexes, in which the DNA is often deformed by interactions with the protein. We selected two high-resolution unbound DNA crystal structures [1d49 (52) and 7bna (53)], containing 10 and 12 base pairs, respectively. For each target we chose double-helical fragments from our library based on sequence similarity to the DNA sequence just as in the bound simulations. We then generated 1000 all-atom models by low-resolution fragment assembly followed by high-resolution refinement. The results are depicted in Figure 5: the final models are similar to the corresponding crystal structures, as judged by RMSD (Figure 5A) and by visual inspection of the low-energy models (Figure 5B). These similarity values are within the range seen in molecular dynamics simulations of unbound DNA duplexes (54) (note that these fragment-rebuilding simulations have no input knowledge of the native structure).

Structure prediction for C_2H_2 zinc fingers

We expected that accurate prediction of binding specificity would depend on accurate recapitulation of bound conformations. To test the ability of the fragment assembly protocol to predict protein–DNA interface structures we conducted modeling simulations on a subset of the C_2H_2 ZFs in our benchmark set for which crystal structures were available (15 individual ZFs in 5 proteins). For each target, we generated 1000 models using the fragment assembly protocol while holding the DNA sequence fixed at the crystal structure sequence. Since no sequence sampling is conducted, the accuracy of the final models depends only on the potential energy function and the conformational sampling algorithm, allowing us to assess just these components of the full protocol. We calculated the similarity of each model to the native structure using the IAS (50), which we found to be superior to RMSD as a correlate of binding sequence prediction. To assess whether the fragment assembly procedure was able to improve upon the input templates, we calculated the similarity to the native of all structures used as fragment sources for each target. For 13 of the 15 cases, the fragment assembly protocol was able to generate models that were closer to the native structure than any template (detailed results for all targets are given in Supplementary Figures S3–S5). In order to make a fair comparison, we selected only the lowest energy models, taking a number equal to the number of fragment templates. The left panel of Figure 6 compares the median similarity score for these models (green bars) to the median similarity score of the input templates (blue bars): in 13 of 15 cases, the low-energy models are more similar to the native structure than the input templates (the difference is negligible for the remaining 2 cases). We also compared the similarity score of the models to that of the input template with highest sequence identity (purple bars; sequence identity was calculated over the entire length of the finger); again the fragment-assembly models showed greater similarity to the native.

For a subset of the targets, we conducted all-against-all fixed-backbone homology modeling simulations, using each target as a template for all other targets (81 target-template combinations). At the start of each simulation, the protein and DNA sequences of the template finger were mutated to match those of the target. The interface sidechains were then optimized using a MCM protocol that included rotameric sampling of protein

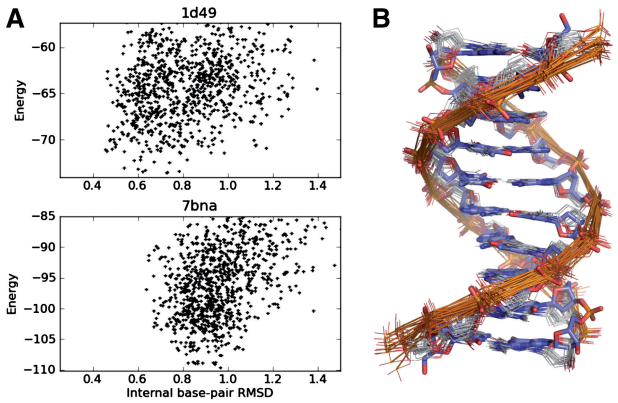

Figure 5. Fragment assembly of unbound DNA structures. (**A**) Scatter plots of base RMSD to native (excluding terminal base pairs) versus all-atom energy for models built by double-helical fragment assembly followed by all-atom refinement. One energy unit is quivalent to ~1.3 k cal/mol. (**B**) Superposition of the 1d49 crystal structure model (in stick representation, carbon atoms are purple) and the 25 lowest-energy fragment assembly models (in wireframe, with gray carbon atoms).

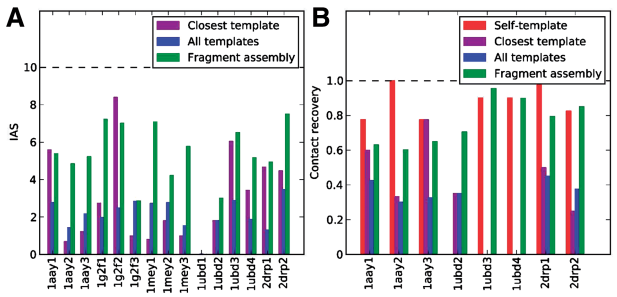

Figure 6. Structure prediction for C_2H_2 ZF proteins. (**A**) Median IAS similarities to the target for the fingers used as fragment templates (blue), the lowest-energy fragment assembly models (green), and the single template with the highest sequence identity to the target (purple). (**B**) Fraction of native contacts recovered by the fragment assembly models (green), and by fixed-backbone sidechain prediction simulations starting from either the native backbone (red), all fragment template backbones (blue), or the backbone of the template with highest sequence identity to the target (purple).

sidechains and DNA residues. This protocol was repeated 10 times to give 10 template-based models. For each of these template-based models, and for each of the fragment-assembly structure prediction models, we counted the number of correctly predicted protein–DNA hydrogen bonds involving major groove atoms in the binding site (these protein–DNA contacts would be expected to contribute strongly to binding specificity). Supplementary Figure S6 shows cumulative histograms of these recovered contacts for all 81 target-template combinations. The results are summarized in the right panel of Figure 6, which shows the total fraction of native contacts recovered by the fragment assembly models and the template-based models. Not surprisingly, fixed-backbone simulations starting from the target backbone itself ('self-template') showed the highest recovery of native contacts; the fragment assembly models were close behind, out-performing both the median template and the template with the highest sequence identity (note that the target and all highly sequence-similar structures are excluded from fragment selection).

Binding specificity prediction

Having demonstrated that the fragment assembly protocol is capable of recapitulating native interface structures and key specificity-determining contacts when given the correct DNA binding site, we turned to the more challenging problem of predicting binding specificity *de novo*. Our benchmark set consisted of a total of 27 individual ZFs in 11 proteins—five of known structure and six of unknown structure. For each individual finger we generated 1000 models using the interface fragment assembly protocol, starting each simulation with a randomized binding site sequence and allowing the DNA sequence to evolve throughout the simulation by MC sequence exploration. PFMs were constructed from the DNA sequences in the final models, and these PFMs for individual fingers were combined to yield full PFMs for the 11 target proteins (Figure 4). Figure 7 shows the predicted PFMs depicted as sequence logos (42) beneath the corresponding experimental binding data for the 11 benchmark proteins. Overall, the agreement between the binding specificity predictions and the experimental data is good. Focusing on the core triplet positions in each binding site (81 positions total), the preferred base in prediction and experiment match in 79% of cases. As a more refined measure of profile similarity, we calculated the BLiC score (51) for all triplet positions, as well as an associated *P*-value. We found that 75% of the positions had a positive BLiC similarity score, and 76% had an associated BLiC *P*-value <0.05. We conclude that binding preferences at 75–80% of the positions are well predicted by the interface fragment assembly protocol. Note that for the 6 proteins of unknown structure, these simulations constitute a prediction for how the protein binds to the DNA: by aligning the predicted binding specificity to the experimental binding specificity in Figure 7, we are implicitly predicting the mapping between the individual fingers and the experimental profile.

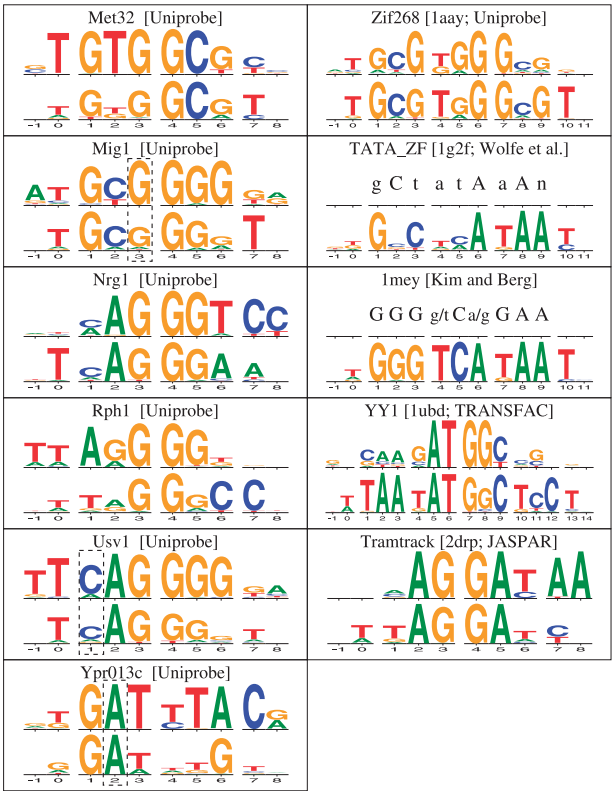

Figure 7. Binding specificity predictions. For each benchmark protein, the experimental binding profile is shown above the structure-based specificity prediction. Experimental data sources are indicated in brackets. PFM columns are numbered so that columns 1-3 correspond to the last finger, columns 4-6 correspond to the second-to-last finger, and so on (see Figure 4). For the three boxed columns, structural determinants of binding preferences are illustrated in Figure 10.

To help assess the contribution of backbone flexibility in the fragment assembly protocol, we conducted a series of fixed-backbone, template-based binding specificity prediction simulations. As in the structure prediction comparison, we performed an all-against-all analysis of 9 individual ZFs, using each finger as a template for fixed-backbone specificity predictions targeted at the other fingers and itself. The results are given in Figure 8. For 6 of the 9 targets, the fragment assembly predictions (green bars) are better than any of the fixed-backbone predictions, even those based on the crystal structure of the target itself ('self-template'), suggesting the importance of backbone flexibility in assessing the energetic cost of mutations away from the crystal structure DNA sequence. If we exclude the self-template predictions, the fragment assembly results are better in 7 of the cases (recall that we exclude the structure being predicted as well as any highly sequence-similar structures from fragment selection). None of the predictions are successful for the remaining two fingers, fingers 1 and 4 from YY1, which represent challenging targets: both are outside the highly specific portion of the binding motif [Figure 7 PFM columns 1–3 (F4) and 10–12 (F1)]; in the YY1 crystal structure, finger 1 lifts off the DNA and does not make any hydrogen bonds or other obvious specificity determining contacts to the major groove.

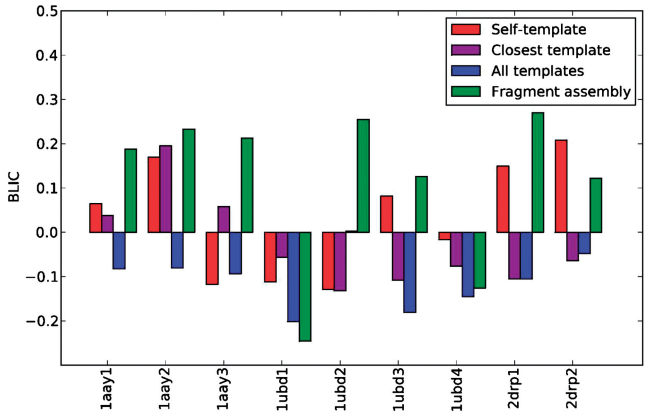

Figure 8. Comparison to fixed-backbone specificity predictions. BLiC scores for the fragment assembly specificity predictions (green bars) are compared to scores of fixed-backbone, template-based predictions started from the target structure itself (red bars) or the template with highest sequence identity to the target (purple bars), as well as the median BLiC score for all non-self template-based predictions (blue bars).

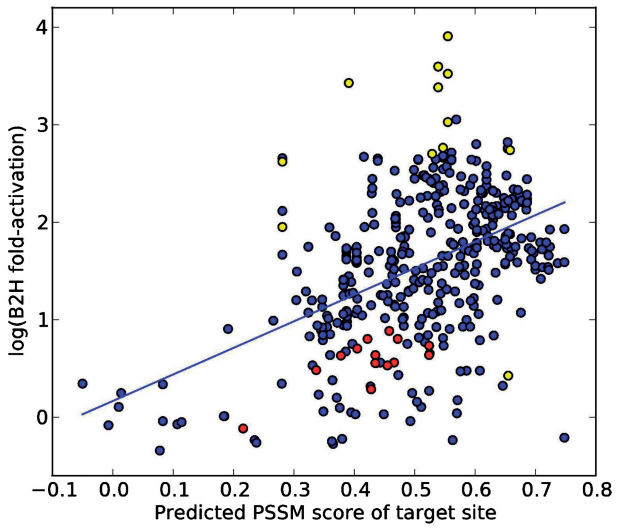

Figure 9. Strength of target-site match to predicted binding profile correlates with *in vivo* activity (fold-activation of a reporter gene in a bacterial 2-hybrid assay, see text) for 401 designed ZF proteins. For each protein, we converted the structure-based PFM into a PSSM and plotted the PSSM score of the selection target site against the experimentally assayed binding activity as reported in the ZiFDB (56). Yellow circles indicate proteins with large experimental error; red circles represent a subset of proteins for which the experimental activity is likely to be an underestimate (see text).

We also compared our structure-based approach with three previously described and publicly accessible algorithms for predicting ZF–DNA interactions: a structure-based approach incorporating family-specific amino acid–nucleotide interaction preferences learned from experimental binding data ['Kaplan05' (30)]; Zinc Finger Binding site database (ZIFIBI), which uses a hidden Markov model to generate binding site predictions (31); and a recent machine learning approach that incorporates data on binding and non-binding DNA sites through the use of a support vector machine ['Persikov09' (33)]. Given that experimental binding data for the proteins of known structure were likely used to train one or more of these methods, we restricted our comparison to the six ZF proteins without solved structures whose specificities were recently profiled by protein binding microarrays (49). As it was not straightforward to generate full PFMs for each algorithm, we focused on the simple metric that counts the number of positions at which the preferred base in prediction and experiment agree (details of the comparison can be found in the Supplementary Data). With this metric, over the set of 6 ZFs of unknown structure (36 positions), our method recovered 86% of the positions, Persikov09 also recovered 86%, and ZIFIBI recovered 75%. The Kaplan05 web server was not able to locate all 12 of the ZF domains; restricting to the subset of 9 ZF domains found, we recovered 89% of the 27 positions and Kaplan05 recovered 78%. It should be emphasized that these comparisons involve a small number of positions, and have correspondingly large statistical uncertainties; nonetheless, the success of our high-resolution structural approach suggests that it can make non-trivial predictions of ZF–DNA interactions.

Specificity predictions for OPEN zinc-finger arrays

The engineered zinc finger protein $TATA_{ZF}$ (55) was the least successful target in our binding specificity prediction benchmark. To assess whether engineered zinc finger proteins are inherently more challenging than naturally occurring zinc fingers, we conducted binding specificity predictions on 401 3-finger ZF proteins generated by the OPEN (Oligomerized Pool ENgineering) platform (35) and available for download from the ZiFDB database (56). For each protein, we downloaded the amino acid sequence, the 9 base pair target site for which the protein was selected, and a quantitative *in vivo* measure of binding activity. Comparing the structure-based predictions to the target sites, we found that the predicted optimal base matched the target site base at 80% of the positions (2889 of the 3609 positions, full results are given in the Supplementary Table S1; note that the target site is not necessarily the optimal binding site for each finger). This level of accuracy agrees well with the results reported above on a smaller set consisting primarily of naturally occurring ZFs. We then asked whether we could predict the experimentally measured level of activity seen for each protein [fold activation over background of a reporter gene in a bacterial 2-hybrid (B2H) assay, 'B2H fold-activation']. We calculated a predicted binding score for each target site by computing its match to a PSSM derived from the final sequences in the fragment assembly simulations. Figure 9 shows a scatter plot of the predicted binding score (x-axis) versus the logarithm of the experimentally determined B2H fold-activation. Although there is considerable scatter, a correlation between the predicted and experimentally observed activities can be clearly seen (linear regression fit: $R^2 = 0.23$, P-value $= 6.3e-24$). Shown in yellow are a subset of ZFs with high experimental error (SD > 5), which show larger deviations from the best-fit line. In

red are a subset of ZFs whose target sites have higher baseline levels of promoter activity, which may lower the apparent fold-activation [Supplementary Data for Maeder *et al.* (35)]; indeed, these ZFs all fall below the best-fit line. Although these correlations between structure-based predictions and an *in vivo* biological readout are encouraging, there is likely room for improvement: using a linear model whose inputs are the experimentally determined K_D values for the component ZF modules, Sander *et al.* (57) were able to achieve significantly more accurate ($R^2 = 0.64$) predictions of B2H fold-activation for a set of 53 modularly-assembled three-finger ZF proteins.

In our initial examination of the modeling simulations described above, there were several ZFs with high levels of *in vivo* activity whose target sites nonetheless scored quite poorly based on the structural simulations. Visual examination of the predicted and experimental sites revealed that the target sites would score more highly if shifted in the 3′ direction by a single base. Closer examination of the Supplementary Data for Maeder *et al.* (35) revealed that these outliers were the results of a selection for a single site (VF3540R), and that the authors had also concluded that the ZFs in question bind to a shifted site. The target site as reported in the ZiFDB was not updated, however, making this an interesting 'blind' test of the utility of structure-based binding prediction.

DISCUSSION

We have described a fragment-assembly protocol for predicting protein–DNA interface structures and binding preferences. In this protocol, diverse models of protein–DNA complexes are assembled from small pieces of related and unrelated protein–DNA structures. These models are taken as starting points for an all-atom, MC refinement procedure that combines sequence mutation moves in the DNA binding site with conformational perturbations in order to simultaneously explore sequence and structure space. Starting from an initial pool of random binding sites, position-specific preferences emerge in the final models through energetically-biased acceptance of the DNA sequence moves. By incorporating an unbound-state energy term into the potential function, we are able to infer relative binding affinities from the sampling frequencies of different DNA sequences in these models.

We applied this interface fragment assembly protocol to make binding specificity predictions for a benchmark of 11 ZF proteins, 5 of known structure. Overall, these structure-based predictions agreed reasonably well with available experimental binding data, as judged by visual and quantitative comparison of preferred bases and sequence profiles (Figure 7). This agreement is encouraging in light of the fact that each ZF is simulated individually, without the structural context provided by its immediate neighbors; this suggests a significant degree of modularity in binding. By examining the modeled structures, we can form hypotheses about the structural determinants of binding specificity. Figure 10 provides three examples in which nucleotide preferences in the final DNA sequences can be explained in terms of structural features. Panels (A) and (B) illustrate well-known amino acid-base preferences: Asn for A and Arg for G. These preferences agree with previously described recognition codes for ZF-DNA interactions (28). Figure 10C provides an example in which the simple logic of a recognition code is broken: Gln at position 6 had been proposed to recognize A at triplet position 1, but in both Usv1 and Nrg1 the experimentally observed preference is for C. Examination of low-energy structural models with C at position 1 shows that rather than forming a canonical bidentate interaction (which would be possible with an A), the Gln forms a pair of hydrogen bonds that bridge the C at position 1 and a T on the complementary strand that pairs with the A at position 2. This A at position 2 is in turn determined by Asn at helix position 3—as in Figure 10A—providing an example of higher-order correlation between DNA-contacting residues.

Examining the predicted and experimental profiles more closely, a number of trends can be detected. First, there is a tendency for the structure-based simulations to

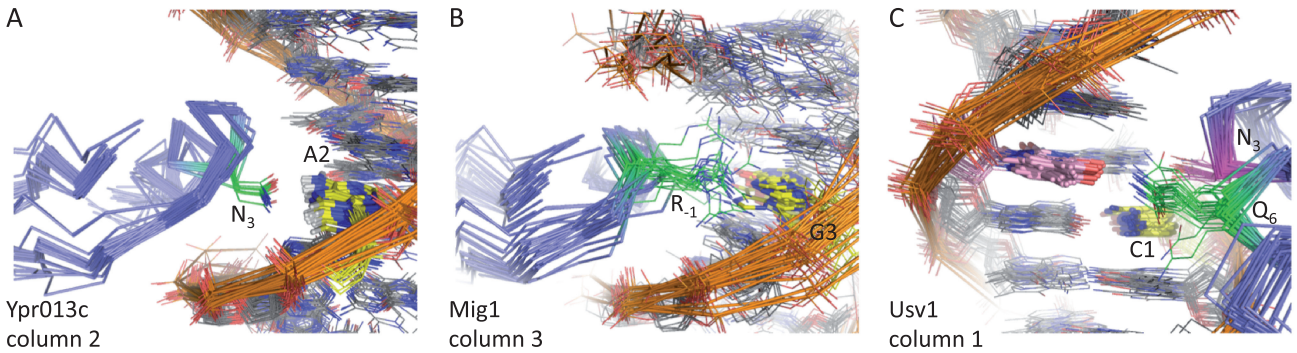

Figure 10. Structural basis of binding specificity: determinants of nucleotide preferences in the fragment assembly models are analyzed at 3 sites in the benchmark set. The base at the site of interest is colored yellow; interacting sidechains are colored green; additional interacting bases and sidechains shown in pink and purple. (**A**) Asn at helix position 3 ('N_3') makes a bidentate hydrogen bond with A at position 2 ('A2') in the Ypr013c site. (**B**) Arg at position −1 can form a bidentate hydrogen bond with G at position 3 in the Mig1 site. (**C**) Correlation between helix positions breaks the simple logic of a ZF recognition code: Gln at helix position 6 forms a pair of hydrogen bonds with C at position 1 and with the T paired with an A at position 2, when Asn is also present at helix position 3. Gln at helix position 6 had been proposed to specify A rather than C at position 1 (28).

over-predict T at the first position of the triplet (Rph1 column 1; $TATA_{ZF}$ columns 4 and 7; 1mey column 7; YY1 columns 1, 4, and 10; Tramtrack column 1). This contrasts with a tendency to under-predict T at the second and third positions. Examination of low-energy models revealed that the ZF often approaches the phosphate backbone more closely in simulations than in the experimentally determined structures, which allows the methyl group of a thymine at triplet position 1 to form favorable packing interactions with a conserved aromatic residue in the core of the finger. Adding neighboring fingers might correct this tendency through direct or linker-mediated interactions; it is also possible that the solvation parameters for the phosphate oxygens could be adjusted to preserve the hydration levels seen in native structures. A second interesting trend can be seen at position 0 (the position one base upstream of the first triplet): all seven of the experimental profiles from the Uniprobe protein-binding microarray database have a T at position 0, as do 14 of the 15 predicted profiles. This position is not typically directly contacted by the protein in models or ZF crystal structures. In models, there is a tendency toward large propeller twisting when a T occurs at this position, suggesting that the apparent experimental preference for T may be due to effects on DNA conformation rather than direct protein–DNA interactions.

There are several limitations to the approach as currently implemented. Modeling the binding preferences of individual fingers in isolation is likely an important source of disagreement between experiment and prediction. We chose to model only single fingers for reasons of computational efficiency—smaller systems are much faster to simulate—as well as ease of interpretation, however we are currently investigating simulations in which two or more fingers are modeled simultaneously. As with all structure-based methods, prediction accuracy is determined in part by the quality of the underlying potential energy function. In this work we have used Rosetta's standard all-atom potential function, with modifications for working with protein–DNA complexes. Although Rosetta's force field has been validated in a variety of modeling and design applications, it is likely that it can be improved for modeling protein–DNA interactions. As an example, our model for the unbound-state energy of a DNA sequence depends only on the base-step composition of that sequence. It is likely that there are higher-order sequence-dependencies, for example A-tract conformational preferences (10), that can only be captured by considering longer-range interactions. A final limitation concerns not the modeling simulations themselves, but the post-processing to generate PFMs. This process implicitly assumes that binding preferences are position-independent, which is unlikely to be true in general (58). It should be possible to analyze the final sequences without making this assumption, particularly in cases where high-throughput experimental binding data from protein binding microarrays (59) or high-throughput sequencing of *in vitro* selected sequences (60) permit direct comparison of more subtle sequence preferences.

Notwithstanding these limitations, we expect that the interface fragment assembly protocol, and its application to the study of ZF–DNA interactions, will have a range of applications. The protocol itself can be extended to other families with multiple template structures. Indeed, analysis of protein–DNA interface variation within three other families of DNA-binding proteins (homeodomain, b-ZIP and b-HLH proteins, Supplementary Figure S2) indicates that interface geometry is as conserved in these families as in the ZF family, suggesting that the interface fragment assembly approach may yield useful predictions for them as well. The protocol can also be applied to aid in understanding the binding preferences of ZF proteins that contain more than four fingers, whose interactions with DNA are significantly more complex. In the case of the insulator protein CTCF, for example, the core binding site consists of only 12 base pairs (61), although the protein contains 11 zinc fingers (which could in principle recognize more than 30 base pairs). Similar complexity is seen for other ZF proteins, for example NRSF (62) and Blimp-1 (63). Structural modeling offers a promising avenue for unraveling these complexities.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank the members of the Rosetta development community for their many contributions to the software used in this research. We also gratefully acknowledge superlative computing support from FHCRC PHS IT, with special thanks to Jeffrey Katcher, Carl Benson and Dirk Petersen.

FUNDING

National Institutes of Health (RO1GM088277 to P.B.); New Development funding from the Fred Hutchinson Cancer Research Center to (P.B.). Funding for open access charge: FHCRC new development funding.

Conflict of interest statement. None declared.

REFERENCES

1. Janin,J., Henrick,K., Moult,J., Eyck,L.T., Sternberg,M.J., Vajda,S., Vakser,I. and Wodak,S.J. (2003) CAPRI: a Critical Assessment of PRedicted Interactions. *Proteins*, **52**, 2–9.
2. Janin,J. and Wodak,S. (2007) The third CAPRI assessment meeting Toronto, Canada, April 20-21, 2007. *Structure*, **15**, 755–759.
3. Chevalier,B.S., Kortemme,T., Chadsey,M.S., Baker,D., Monnat,R.J. and Stoddard,B.L. (2002) Design, activity, and structure of a highly specific artificial endonuclease. *Mol. Cell*, **10**, 895–905.
4. Joachimiak,L.A., Kortemme,T., Stoddard,B.L. and Baker,D. (2006) Computational design of a new hydrogen bond network and at least a 300-fold specificity switch at a protein-protein interface. *J. Mol. Biol.*, **361**, 195–208.
5. Grigoryan,G., Reinke,A.W. and Keating,A.E. (2009) Design of protein-interaction specificity gives selective bZIP-binding peptides. *Nature*, **458**, 859–864.
6. Steffen,N.R., Murphy,S.D., Tolleri,L., Hatfield,G.W. and Lathrop,R.H. (2002) DNA sequence and structure: direct and

indirect recognition in protein-DNA binding. *Bioinformatics*, **18(Suppl. 1)**, S22–S30.
7. Paillard,G. and Lavery,R. (2004) Analyzing protein-DNA recognition mechanisms. *Structure*, **12**, 113–122.
8. Gromiha,M., Siebers,J.G., Selvaraj,S., Kono,H. and Sarai,A. (2004) Intermolecular and intramolecular readout mechanisms in protein-DNA recognition. *J. Mol. Biol.*, **337**, 285–294.
9. Joshi,R., Passner,J.M., Rohs,R., Jain,R., Sosinsky,A., Crickmore,M.A., Jacob,V., Aggarwal,A.K., Honig,B. and Mann,R.S. (2007) Functional specificity of a Hox protein mediated by the recognition of minor groove structure. *Cell*, **131**, 530–43.
10. Rohs,R., West,S.M., Sosinsky,A., Liu,P., Mann,R.S. and Honig,B. (2009) The role of DNA shape in protein-DNA recognition. *Nature*, **461**, 1248–53.
11. Liu,L.A. and Bader,J.S. (2007) Ab initio prediction of transcription factor binding sites. *Pac. Symp. Biocomput.*, 484–495.
12. Temiz,N.A. and Camacho,C.J. (2009) Experimentally based contact energies decode interactions responsible for protein-DNA affinity and the role of molecular waters at the binding interface. *Nucleic Acids Res.*, **37**, 4076–4088.
13. O'Flanagan,R.A., Paillard,G., Lavery,R. and Sengupta,A.M. (2005) Non-additivity in protein-DNA binding. *Bioinformatics*, **21**, 2254–2263.
14. 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.
15. Kono,H. and Sarai,A. (1999) Structure-based prediction of DNA target sites by regulatory proteins. *Proteins*, **35**, 114–131.
16. 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.
17. Contreras-Moreira,B. and Collado-Vides,J. (2006) Comparative footprinting of DNA-binding proteins. *Bioinformatics*, **22**, e74–e80.
18. Lafontaine,I. and Lavery,R. (2000) ADAPT: a molecular mechanics approach for studying the structural properties of long DNA sequences. *Biopolymers*, **56**, 292–310.
19. 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.
20. Havranek,J.J., Duarte,C.M. and Baker,D. (2004) A simple physical model for the prediction and design of protein-DNA interactions. *J. Mol. Biol.*, **344**, 59–70.
21. Siggers,T.W. and Honig,B. (2007) Structure-based prediction of C2H2 zinc-finger binding specificity: sensitivity to docking geometry. *Nucleic Acids Res.*, **35**, 1085–1097.
22. Donald,J.E., Chen,W.W. and Shakhnovich,E.I. (2007) Energetics of protein-DNA interactions. *Nucleic Acids Res.*, **35**, 1039–1047.
23. Jamal Rahi,S., Virnau,P., Mirny,L.A. and Kardar,M. (2008) Predicting transcription factor specificity with all-atom models. *Nucleic Acids Res.*, **36**, 6209–6217.
24. Elrod-Erickson,M., Rould,M.A., Nekludova,L. and Pabo,C.O. (1996) Zif268 protein-DNA complex refined at 1.6 A: a model system for understanding zinc finger-DNA interactions. *Structure*, **4**, 1171–1180.
25. Emerson,R.O. and Thomas,J.H. (2009) Adaptive evolution in zinc finger transcription factors. *PLoS Genet.*, **5**, e1000325.
26. Desjarlais,J.R. and Berg,J.M. (1992) Toward rules relating zinc finger protein sequences and DNA binding site preferences. *Proc. Natl Acad. Sci. USA*, **89**, 7345–7349.
27. Choo,Y. and Klug,A. (1994) Toward a code for the interactions of zinc fingers with DNA: selection of randomized fingers displayed on phage. *Proc. Natl Acad. Sci. USA*, **91**, 11163–11167.
28. Wolfe,S.A., Greisman,H.A., Ramm,E.I. and Pabo,C.O. (1999) Analysis of zinc fingers optimized via phage display: evaluating the utility of a recognition code. *J. Mol. Biol.*, **285**, 1917–1934.
29. 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.
30. Kaplan,T., Friedman,N. and Margalit,H. (2005) Ab initio prediction of transcription factor targets using structural knowledge. *PLoS Comput. Biol.*, **1**, e1.
31. Cho,S.Y., Chung,M., Park,M., Park,S. and Lee,Y.S. (2008) ZIFIBI: prediction of DNA binding sites for zinc finger proteins. *Biochem. Biophys. Res. Commun.*, **369**, 845–848.
32. Liu,J. and Stormo,G.D. (2008) Context-dependent DNA recognition code for C2H2 zinc-finger transcription factors. *Bioinformatics*, **24**, 1850–1857.
33. Persikov,A.V., Osada,R. and Singh,M. (2009) Predicting DNA recognition by Cys2His2 zinc finger proteins. *Bioinformatics*, **25**, 22–29.
34. Ordiz,M.I., Barbas,C.F. III and Beachy,R.N. (2002) Regulation of transgene expression in plants with polydactyl zinc finger transcription factors. *Proc. Natl Acad. Sci. USA*, **99**, 13290–13295.
35. Maeder,M.L., Thibodeau-Beganny,S., Sander,J.D., Voytas,D.F. and Joung,J.K. (2009) Oligomerized pool engineering (OPEN): an 'open-source' protocol for making customized zinc-finger arrays. *Nat. Protoc.*, **4**, 1471–1501.
36. Rohl,C.A., Strauss,C.E., Misura,K.M. and Baker,D. (2004) Protein structure prediction using Rosetta. *Methods Enzymol.*, **383**, 66–93.
37. Li,Z. and Scheraga,H.A. (1987) Monte Carlo-minimization approach to the multiple-minima problem in protein folding. *Proc. Natl Acad. Sci. USA*, **84**, 6611–6615.
38. Cahill,S., Cahill,M. and Cahill,K. (2003) On the kinematics of protein folding. *J. Comput. Chem.*, **24**, 1364–1370.
39. Wang,C., Bradley,P. and Baker,D. (2007) Protein-protein docking with backbone flexibility. *J. Mol. Biol.*, **373**, 503–519.
40. Dickerson,R.E. (1998) DNA bending: the prevalence of kinkiness and the virtues of normality. *Nucleic Acids Res.*, **26**, 1906–1926.
41. Endres,R.G. and Wingreen,N.S. (2006) Weight matrices for protein-DNA binding sites from a single co-crystal structure. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, **73(6 Pt 1)**, 061921.
42. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
43. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.
44. Olson,W.K., Gorin,A.A., Lu,X.J., Hock,L.M. and Zhurkin,V.B. (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl Acad. Sci. USA*, **95**, 11163–11168.
45. Simons,K.T., Ruczinski,I., Kooperberg,C., Fox,B.A., Bystroff,C. and Baker,D. (1999) Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins*, **34**, 82–95.
46. Kuhlman,B., Dantas,G., Ireton,G.C., Varani,G., Stoddard,B.L. and Baker,D. (2003) Design of a novel globular protein fold with atomic-level accuracy. *Science*, **302**, 1364–1368.
47. 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.
48. Lazaridis,T. and Karplus,M. (1999) Effective energy function for proteins in solution. *Proteins*, **35**, 133–152.
49. Zhu,C., Byers,K.J., McCord,R.P., Shi,Z., Berger,M.F., Newburger,D.E., Saulrieta,K., Smith,Z., Shah,M.V., Radhakrishnan,M. *et al.* (2009) High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Res.*, **19**, 556–566.
50. 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.
51. Habib,N., Kaplan,T., Margalit,H. and Friedman,N. (2008) A novel Bayesian DNA motif comparison method for clustering and retrieval. *PLoS Comput. Biol.*, **4**, e1000010.
52. Quintana,J.R., Grzeskowiak,K., Yanagi,K. and Dickerson,R.E. (1992) Structure of a B-DNA decamer with a central T-A step: C-G-A-T-T-A-A-T-C-G. *J. Mol. Biol.*, **225**, 379–395.
53. Holbrook,S.R., Dickerson,R.E. and Kim,S.H. (1985) Anisotropic thermal-parameter refinement of the DNA dodecamer

cgcgaattcgcg by the segmented rigid-body method. *Acta Crystallographica Sec. B-Struct. Sci.*, **41**, 255–262.
54. Reddy,S.Y., Leclerc,F. and Karplus,M. (2003) DNA polymorphism: a comparison of force fields for nucleic acids. *Biophys. J.*, **84**, 1421–1449.
55. Wolfe,S.A., Grant,R.A., Elrod-Erickson,M. and Pabo,C.O. (2001) Beyond the 'recognition code': structures of two Cys2His2 zinc finger/TATA box complexes. *Structure*, **9**, 717–723.
56. Fu,F., Sander,J.D., Maeder,M., Thibodeau-Beganny,S., Joung,J.K., Dobbs,D., Miller,L. and Voytas,D.F. (2009) Zinc Finger Database (ZiFDB): a repository for information on C2H2 zinc fingers and engineered zinc-finger arrays. *Nucleic Acids Res.*, **37**, D279–D283.
57. Sander,J.D., Zaback,P., Joung,J.K., Voytas,D.F. and Dobbs,D. (2009) An affinity-based scoring scheme for predicting DNA-binding activities of modularly assembled zinc-finger proteins. *Nucleic Acids Res.*, **37**, 506–515.
58. Benos,P.V., Bulyk,M.L. and Stormo,G.D. (2002) Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res.*, **30**, 4442–4451.
59. Philippakis,A.A., Qureshi,A.M., Berger,M.F. and Bulyk,M.L. (2008) Design of compact, universal DNA microarrays for protein binding microarray experiments. *J. Comput. Biol.*, **15**, 655–665.
60. Jolma,A., Kivioja,T., Toivonen,J., Cheng,L., Wei,G., Enge,M., Taipale,M., Vaquerizas,J.M., Yan,J., Sillanpaa,M.J. *et al.* (2010) Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res.*, **20**, 861–873.
61. Renda,M., Baglivo,I., Burgess-Beusse,B., Esposito,S., Fattorusso,R., Felsenfeld,G. and Pedone,P.V. (2007) Critical DNA binding interactions of the insulator protein CTCF: a small number of zinc fingers mediate strong binding, and a single finger-DNA interaction controls binding at imprinted loci. *J. Biol. Chem.*, **282**, 33336–33345.
62. Johnson,D.S., Mortazavi,A., Myers,R.M. and Wold,B. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.
63. Doody,G.M., Care,M.A., Burgoyne,N.J., Bradford,J.R., Bota,M., Bonifer,C., Westhead,D.R. and Tooze,R.M. (2010) An extended set of PRDM1/BLIMP1 target genes links binding motif type to dynamic repression. *Nucleic Acids Res.*, **38**, 5336–5350.