DNA Deformation Energy as an Indirect Recognition Mechanism in Protein-DNA Interactions

Kimberly A. Aeling, Nicholas R. Steffen, Matthew Johnson, G. Wesley Hatfield, Richard H. Lathrop, and Donald F. Senear

Abstract—Proteins that bind to specific locations in genomic DNA control many basic cellular functions. Proteins detect their binding sites using both direct and indirect recognition mechanisms. Deformation energy, which models the energy required to bend DNA from its native shape to its shape when bound to a protein, has been shown to be an indirect recognition mechanism for one particular protein, Integration Host Factor (IHF). This work extends the analysis of deformation to two other DNA-binding proteins, CRP and SRF, and two endonucleases, I-CreI and I-PpoI. Known binding sites for all five proteins showed statistically significant differences in mean deformation energy as compared to random sequences. Binding sites for the three DNA-binding proteins and one of the endonucleases had mean deformation energies lower than random sequences. Binding sites for I-PpoI had mean deformation energy higher than random sequences. Classifiers that were trained using the deformation energy at each base pair step showed good cross-validated accuracy when classifying unseen sequences as binders or nonbinders. These results support DNA deformation energy as an indirect recognition mechanism across a wider range of DNA-binding proteins. Deformation energy may also have a predictive capacity for the underlying catalytic mechanism of DNA-binding enzymes.

Index Terms—DNA-protein binding, indirect recognition, indirect readout, DNA bending, perceptron learning, deformation energy.

1 Introduction

Interactions between proteins and DNA govern the development and lives of cells. Proteins bind to specific locations on genomic DNA to control gene expression, replication of DNA, DNA repair, and other vital cellular processes. Proteins bind to DNA using both sequence-specific and structure-specific mechanisms. The recognition of sequence-specific contacts, often called direct readout or direct recognition, is better understood and forms the basis of most characterizations of binding sites. Direct recognition mechanisms involve protein amino acid residues in contact with specific bases in DNA sequences. This mechanism leads to conservation of the bases involved in direct recognition. However, sequence dependence alone does not completely explain specificity in protein-DNA binding. Mutation of bases not in direct contact with the protein can affect binding affinity [1], implying that proteins employ mechanisms other than direct recognition. There is increasing evidence that DNA structural properties significantly affect its interactions with proteins [2]. Recognition of DNA structural properties is referred to as indirect readout or indirect recognition.

DNA structural properties that contribute to indirect readout by proteins include flexibility [3], elasticity [4], bending and kinking [5], major and minor groove widths, and hydration [6]. Hidden Markov Models and DNA structural scales have been combined to identify promoters and DNA structural patterns [7]. Incorporating minor groove width with a sequence component for MetJ binding sites resulted in improved recognition and more accurate predictions of binding affinity than using pure sequence-based methods [8]. B-form to A-form DNA transitions [9] have been characterized using the same deformation energy model used in this work [10], [11]. Sarai et al. applied similar structure-based threading methods to assess both direct and indirect recognition of DNA binding sites for a large number of proteins [12], [13], [14], [15]. Their results indicated that both mechanisms can contribute to specificity, that the relative contributions vary, and that accounting for both direct and indirect recognition aspects improves model accuracy.

The energy required to deform DNA from its native conformation to the conformation in a protein-bound complex provides the basis for a potential recognition mechanism. This energy is modeled here as deformation energy, which is a mean force potential for the energy required to deform DNA. We have analyzed in depth the role of deformation energy in the recognition of specific DNA binding sites by the *E. coli* Integration Host Factor

- *K.A. Aeling and G.W. Hatfield are with the Department of Microbiology and Molecular Genetics, School of Medicine, University of California, Irvine, Irvine, CA 92697-3425. E-mail: {kaeling, gwhatfier}@uci.edu.*
- *N.R. Steffen is with Primarion, 2780 Skypark Drive, Torrance, CA 90505. E-mail: nick.steffen@primarion.com.*
- *M. Johnson and R.H. Lathrop are with the Donald Bren School of Information and Computer Sciences, University of California, Irvine, Irvine, CA 92697-3425. E-mail: {johnsonm, rickl}@uci.edu.*
- *D.F. Senear is with the Department of Molecular Biology and Biochemistry, School of Biological Sciences, University of California, Irvine, Irvine, CA 92697-3425. E-mail: dfsenear@uci.edu.*

Manuscript received 14 Nov. 2005; revised 20 Feb. 2006; accepted 14 Apr. 2006; published online 9 Jan. 2007.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-0131-1105.
Digital Object Identifier no. 10.1109/TCBB.2007.1000.

[16], [17]. The results of these studies are that 1) the distributions of deformation energies for known IHF binding sequences and for random DNA sequences threaded onto the DNA conformation found in the IHF-DNA cocrystallographic complex are distinct, with lower deformation energies on average for known binding sequences; 2) classifiers based on the deformation energies for each base pair step of a sequence discriminate between known binding sequences and random sequences with high accuracy; and 3) the particular sequence in the crystallographic complex has nearly the lowest deformation energy of any sequence, exceeded only by a few, very similar sequences. We concluded, based on these findings, that IHF tends to select binding sites that minimize the deformation energy.

IHF generates a severe distortion in DNA when it binds, bending DNA into a near U-turn [18]. It is perhaps not surprising that deformation energy would play a role in binding site recognition under such circumstances. It seems logical that IHF binding sites should have lower deformation energy on average than nonbinding sites. Other DNA-binding proteins also bend DNA to various degrees when they bind and, even when large-scale bending does not occur, the small-scale structure of the DNA is often still deformed to some degree. Is deformation energy a general property of binding in such cases? Does this help to explain how other proteins recognize their preferred binding sites in DNA?

To address these questions, we identified all DNA-binding proteins that offered the opportunity for analysis of deformation energy, similar as for IHF. Each protein was required to meet all of the following criteria:

1. the crystal structure of the protein bound to its DNA substrate has been solved,
2. the protein consensus binding site exhibits degeneracy,
3. there exists a large set of known specific DNA-binding sites for the protein,
4. protein binding may make partial use of indirect protein recognition, and
5. there is a significant deformation of the otherwise B-form DNA upon protein binding.

In addition to IHF, four DNA-binding proteins met these criteria: CRP [19], [20], I-CreI [21], I-PpoI [21], and SRF [22]. CRP (Cyclic AMP Receptor Protein) is a transcriptional activator in *E. coli*. I-CreI and I-PpoI are homing endonucleases found in prokaryotic organisms. SRF (Serum Response Factor) is a transcription factor in humans that binds to the DNA major groove. In this work, we have examined the role of deformation energy in site-specific binding by these five proteins.

For all five proteins, we found the mean deformation energy of DNA sequences known to be protein binding sites was significantly different from random sequences. In four of the five cases, the mean deformation energy was lower for known binding sequences than for random sequences. In one case, the reverse was true: Known binding sites for I-PpoI had mean deformation energy higher than random sequences. This unanticipated finding may be a result of the catalytic mechanism employed by I-P-poI. Classifiers trained using the deformation energy at each base pair step as training features showed good cross-validated accuracy when classifying previously unseen sequences as binders or nonbinders. These results support deformation energy as the basis for an indirect recognition mechanism across a wider range of DNA-binding proteins.

2 Methods

All sequences known to bind to IHF [23], CRP [19], [20], I-CreI [21] , I-PpoI [21], and SRF [22] at high affinities were identified from the available literature (see supplemental data). For the DNA-binding proteins, these are known binding sites. For the two homing endonucleases, these are catalytically active sites of hydrolysis. An important difference between these two is that, whereas binding sites are expected to have been optimized for binding affinity, a thermodynamic property that is directly related to deformation energy, catalytic sites are expected to have been optimized for hydrolysis. This is a kinetic property which reflects both the binding affinity and the turnover rate. The extent to which this is related to deformation should depend on the mechanism of catalytic rate enhancement.

Positive (functional) test sets were constructed using known binding sequences. Sets of negative (control) sequences were generated randomly for each protein, using the same distribution of base frequencies as in its native genome. The number of random sequences generated for each protein was nine times larger than the number of known binding sites. All sequences were threaded onto the structural motif defined by the cocrystal structure of their respective protein-DNA complex. The resulting deformation energy was calculated as described below. Perceptrons were trained on the set of random sequences (negative examples) and known binding sequences (positive examples) using the individual deformation energies for each base pair step in the binding site as training features. The sequence having the globally lowest deformation energy when threaded onto its crystal structure was computed and compared with the DNA sequence in the crystal structure. The distribution of deformation energies of the entire ensemble of control sequences was compared to the distribution of deformation energies for all known sites.

2.1 Structural Model and Deformation Energy

The crystal structure of a protein-DNA complex is the basis for a DNA structural motif [17]. The structural motif is represented as the conformation of the DNA when bound to a protein. An objective function is used to model the energy required to distort DNA from its native conformation to the conformation found in the bound complex [16]. More specifically, DNA sequences are threaded onto the structural motif and the objective function is used to model the energy difference between the bound and unbound shapes for those sequences.

Fig. 1a, Fig. 1c, Fig. 1e, Fig. 1g, and Fig. 1i show ribbon models of the protein-DNA complexes used here. The DNA in each complex is bent to varying extents. Fig. 1b, Fig. 1d, Fig. 1f, Fig. 1h, and Fig. 1j show models of the bound DNA in each complex in which the base pairs have been modeled as rigid rectangular slabs, or dominoes, that provide the best least-squares fit to the base atoms [11]. Each pair of

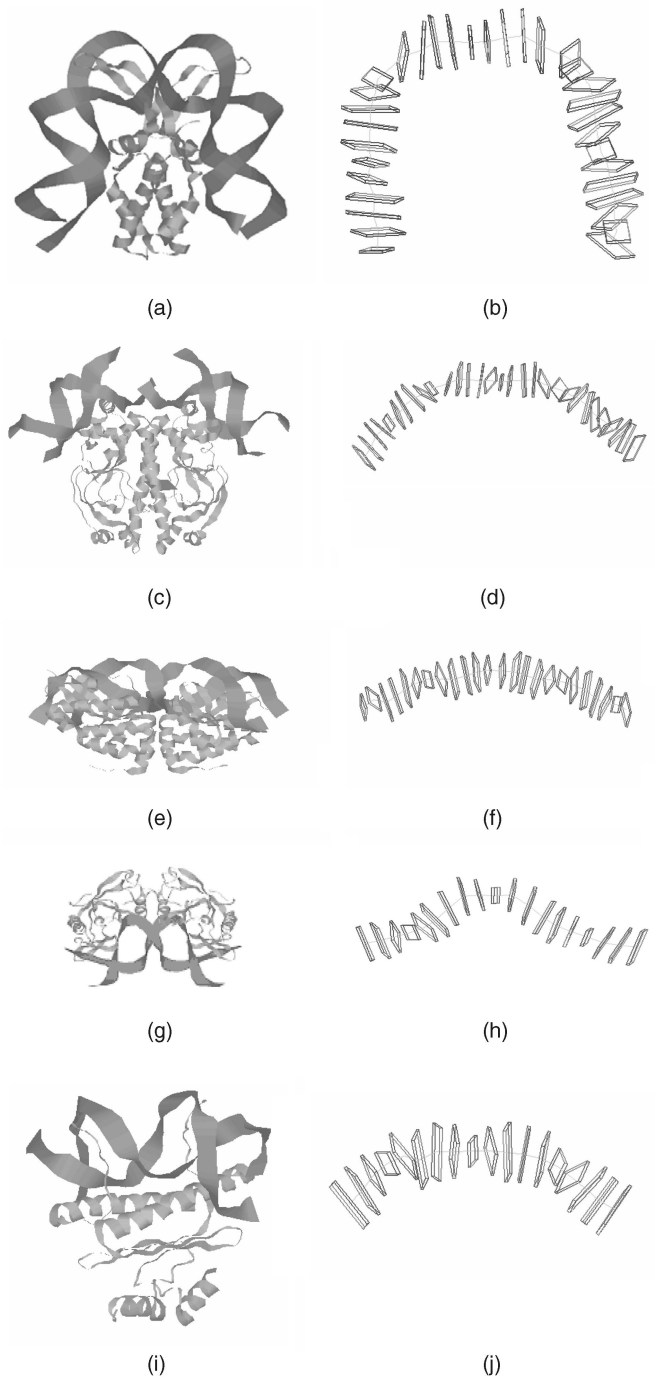

Fig. 1. Protein-DNA complexes shown as ribbon models and the corresponding domino model of the DNA for five protein-DNA complexes examined in this paper. (a) IHF-DNA ribbon model, (b) corresponding domino model. (c) CRP-DNA ribbon model, (d) corresponding domino model. (e) I-CreI-DNA ribbon model, (f) corresponding domino model. (g) I-PpoI-DNA ribbon model, (h) corresponding domino model. (i) SRF-DNA ribbon model, (j) corresponding domino model.

adjacent dominoes forms one base pair step. The relative displacement of one domino from another is fully defined by six parameters—three are translations (slide, shift, and rise) and three are rotations (roll, tilt, and twist). The structure of DNA in a bound complex is therefore defined fully by these six parameters for all of the base pair steps. Deforming any base pair step from its preferred, or equilibrium, conformation requires an input of energy. The energy required is a function of the degree (or extent) of deformation and the neighboring nucleotides that compose the base pair step. Olson et al. modeled these deformations as springs using a harmonic function [10]. An equilibrium value and a force constant were calculated for each of the six displacement parameters for each pair of nucleotides.

The energy required for DNA to distort from its equilibrium conformation to the conformation when bound by protein is modeled by summing the individual deformation energies [10] of each base pair step in the region of interest. The local deformation energy $\Delta\mathrm{E}(i,x,y)$ models the energy needed to move nucleotides x and y from their equilibrium positions to the positions in the bound structure at Step i. The total deformation energy $\Delta\mathrm{E}_{total}(S)$ models the energy difference of sequence S between its native and bound conformations:

$$\Delta E_{total}(S) = \sum_{i=1}^{L-1} \Delta E(i,x,y), \tag{1}$$

$$\Delta E(i,x,y) = \frac{1}{2}\sum_{j=1}^{6}\sum_{k=1}^{6} fjk\Delta\theta j\Delta\theta k(i,x,y), \tag{2}$$

$$\Delta\theta j(i,x,y) = \theta j(i) - \theta_j^0(x,y), \tag{3}$$

where S is a DNA sequence, L is the length of sequence S, $\theta j(i)$ is base pair step parameter j at base pair step i, $\theta_j^0(x,y)$ is the equilibrium value of base pair step parameter j for nucleotide constituents x and y, and f_{jk} is a coefficient that relates deformation to energy. The coefficients were empirically determined and tabulated by Olson et al. [10].

It is important to note that the computation of deformation energy using this formulation has no free parameters. It depends only on the sequence of the nucleotides and the relative positions of the dominos at each base pair step. Given a structure, the deformation energy is a function only of the sequence threaded onto that structure. Any sequence can be threaded onto the crystal structure and its deformation energy computed.

2.2 Sequence Sets

For each protein of interest, the distribution of deformation energies was calculated for two sets of sequences: a set of sequences known to either bind specifically to the protein or to be cleaved specifically by the enzymes (see supplemental data), and a set of random sequences of the same length as the binding sequences. Random sequences are mostly assumed not to be binding sites. The random sequences were constructed with bases drawn from a population with the same base frequencies as the native genome of the protein (Table 1). The number of random sequences was chosen as nine times the number of binding sequences to facilitate 10-fold cross-validation. The deformation energies calculated were binned and plotted as a histogram. For each protein, the mean deformation energy of binding sequences was compared to that of random sequences. Statistically significant differences in populations were determined using a comparison of means from two sample tests.

TABLE 1
Protein Properties

Protein	PDB entry	Native organism	Binding site length, bp	# of known binding sites	Bending characteristics
IHF	1IHF	*Escherichia coli*	34	43	Proline intercalation & bending. Total bend 160-180 degrees.
CRP	1J59	*Escherichia coli*	22	42	Kinked. Total bending about 90 degrees.
I-CreI	1G9Y	*Chlamydomona s. reinhardtii*	24	21	Moderately curved.
I-PpoI	1A74	*Physarum polycephalum*	19	61	Kinked. Total bending about 90 degrees.
SRF	1SRS	*Homo sapiens*	16	54	Moderately curved.

2.3 Perceptron Learning

For each protein, a perceptron was trained using the functional set of known binding sequences as positive examples and the control set of random sequences as negative examples [16]. The features were the deformation energy at each base pair step (2). The perceptron learned a separate weight for the contribution to the deformation energy for each base pair step and a threshold for the weighted sum of deformation energies to separate the known binding sites from the random sequences. The perceptron score was calculated using (4) with the weights generated by the perceptron. To prevent memorization and better estimate the accuracy on previously unseen sequences, 10-fold cross-validation was performed in all tests and the average classification accuracy was reported. Each classifier created on each fold of the training data had its weight vector normalized to facilitate comparison between scores from different folds. The weight vectors were normalized by dividing each component by the magnitude of the vector.

If the contributions at each base pair step are weighted nonuniformly, (1) must be modified to include the weights at each step:

$$\Delta WE_{total}(S) = \sum_{i=1}^{L-1} wi \cdot \Delta E(i, x, y), \quad (4)$$

where w_i is the weight applied to the deformation energy at base pair step i and is computed by the perceptron.

In addition to the cross-validated accuracy, a receiver operating characteristic (ROC) curve was generated for each protein's classifier. The area under the curve was calculated for each protein.

2.4 Lowest Deformation Energy Sequence

For the bound DNA structure in each protein-DNA complex, the sequence with the globally lowest deformation energy when threaded onto that structure was computed using a dynamic programming algorithm [17]. The lowest deformation energy and the deformation energy of the crystal sequence, represented as z scores based on the statistics of the deformation energy of the random sequences, are shown in Table 2. Also shown are the crystal sequence and lowest energy sequence, with matching bases highlighted.

3 Results

For each protein, the deformation energy was calculated for a set of known binding sequences and a control set of random sequences. Fig. 2 shows the distribution of deformation energy for these sets. For all of the proteins, the mean deformation energies for binding sequences and for random sequences were statistically significantly different. In four of the five cases, the average deformation energy was lower for known specific binding sequences than for the random sequences. However, the average deformation energy of sequences specifically cleaved by I-PpoI was higher than for random sequences.

Classifiers were built to predict whether a sequence was a binding site or not for each of the proteins examined in this paper. The distribution of perceptron scores for the binding sequences and random sequence for each protein are shown in Fig. 3. The accuracy of these classifiers is summarized in Table 3. Three of the classifiers, those for IHF, I-PpoI, and SRF, showed cross-validated accuracy greater than 0.95, with a correspondingly large area under the ROC curves (data not shown and Table 3), indicating a relatively high quality classifier. The classifiers for CRP and I-CreI had cross-validated accuracies of approximately 0.90.

The sequence with the lowest energy when threaded onto the DNA structure in each protein-DNA complex was calculated and compared with the sequence in the crystal structure in Table 2. The z score from the random sequences' deformation energy is also shown. The z score is the number of standard deviations from the mean for a

TABLE 2
Sequences with Lowest Deformation Energy when Threaded onto the Crystal Structures

Protein	Sequence	Deformation energy as z score relative to random sequences
IHF		
Crystal	GCCAAAAAAGCATTGCTTATCAATTTGTTGCACC	-2.8
Lowest DE	CCGAAAAAACCATTGCTTACGAACGCATTGGCCG	-4.7
	* ***** ********* ** *** *	
CRP		
Crystal	AAGTGTGACATATGTCACACTT	-1.9
Lowest DE	GATCACGAACGACCTCGTGTAC	-3.5
	* ** * **	
I-CreI		
Crystal	CGAAACTGTCTCACGACGTTTTGC	-2.3
Lowest DE	CGATAACATTATGCGGTGGTACCG	-3.3
	*** * * ** * *	
I-PpoI		
Crystal	GACTCTCTTAAGGTAGCAA	1.3
Lowest DE	CGGGGTGCCCGGGATGATT	-4.0
	* ** *	
SRF		
Crystal	CTTCCTAATTAGGCCAT	-2.8
Lowest DE	TACCCGAACGCCGCCAT	-6.1
	** ** *****	

Asterisks mark matches between the crystal sequence and the sequence with the lowest deformation energy for each protein. The right column gives the deformation energy for the crystal and lowest energy sequences, shown as a z score relative to the deformation energy for random sequences threaded onto the same structure.

given quantity. Measurements were normalized to z scores for convenient comparisons across measurements.

4 Discussion

For each of the proteins studied here, the distribution of deformation energies for specific sites overlaps the distribution of deformation energies of random sequences. However, in all cases, there is also a statistically significant difference in the mean value of the two distributions. There are several possible causes for this. First, for a given protein, all sequences are threaded onto the same structure. It is likely that, as the DNA sequence changes, its conformation varies in the protein-bound complex. The actual deformation will be less than as calculated because chemical interactions minimize the energy. This results in inaccuracies in calculating the deformation energy for a given sequence. This could also explain why the crystal sequences tend to have lower deformation energy than the average binding sequence. Second, the proteins studied here all use direct recognition mechanisms to identify their binding sites in addition to indirect mechanisms that rely on structure. Both sequence and structure contribute in varying amounts to recognition. Third, other indirect recognition mechanisms may be at work which are not considered in the calculation of deformation energy. For example, the IHF-DNA complex contains ordered water, suggesting a hydration spine. Ordered water seems to play a role in I-CreI-DNA binding as well.

All of the DNA-binding nonenzyme proteins studied (IHF, CRP, and SRF) showed a preference for DNA binding sites that had a low deformation energy. However, different results were obtained for the DNA-binding enzymes. Although I-P-poI and I-CreI are both homing endonucleases, the mean deformation energy of I-P-poI binding sites was higher than random, while the opposite was true for I-CreI. In considering this result, it is important to keep in mind that the DNA sites for the enzymes are actually substrates; hence, they are selected based on hydrolysis rate rather than on binding affinity. In addition, while I-PpoI and I-Cre-I are both homing endonucleases, they are also members of different enzyme families with distinct structural and catalytic properties. Considered in this light, these differences may reflect the different catalytic mechanisms that are employed by these two endonucleases [24], [25], [26].

I-PpoI is a member of the His-Cys box family of homing endonucleases. It has a relatively low affinity for its cleavage sites, i.e., a K_d of approximately 15–20 nM at physiological salt concentration, but a fast turnover rate of 0.11 s^{-1} [25]. I-PpoI binding generates a severe bend in the DNA that is localized particularly to the sites of phosphodiester hydrolysis. The bend serves to widen the minor groove, making the scissile phosphates more accessible for cleavage. Perhaps more significant, the phosphates are highly distorted, mimicking the pentavalent transition state and aligned properly for SN2 displacement by an activated water molecule in the active site of the enzyme. A metal ion in the active site of the enzyme binds two oxygens of the scissile phosphate as ligands. While this contributes to distortion of the phosphate, the metal ion is not catalytic. In this mechanism, binding energy has been used to distort the DNA substrate and drive the conformation of the scissile phosphates toward that of the transition state of the hydrolysis reaction. This is a common mechanism of enzymatic rate enhancement, one that results in highly efficient catalysis in this case, with a kcat/Km of 1e-8 $M^{-1}s^{-1}$, or near the diffusion controlled limit. For such a mechanism, a "stiff" DNA substrate might be particularly

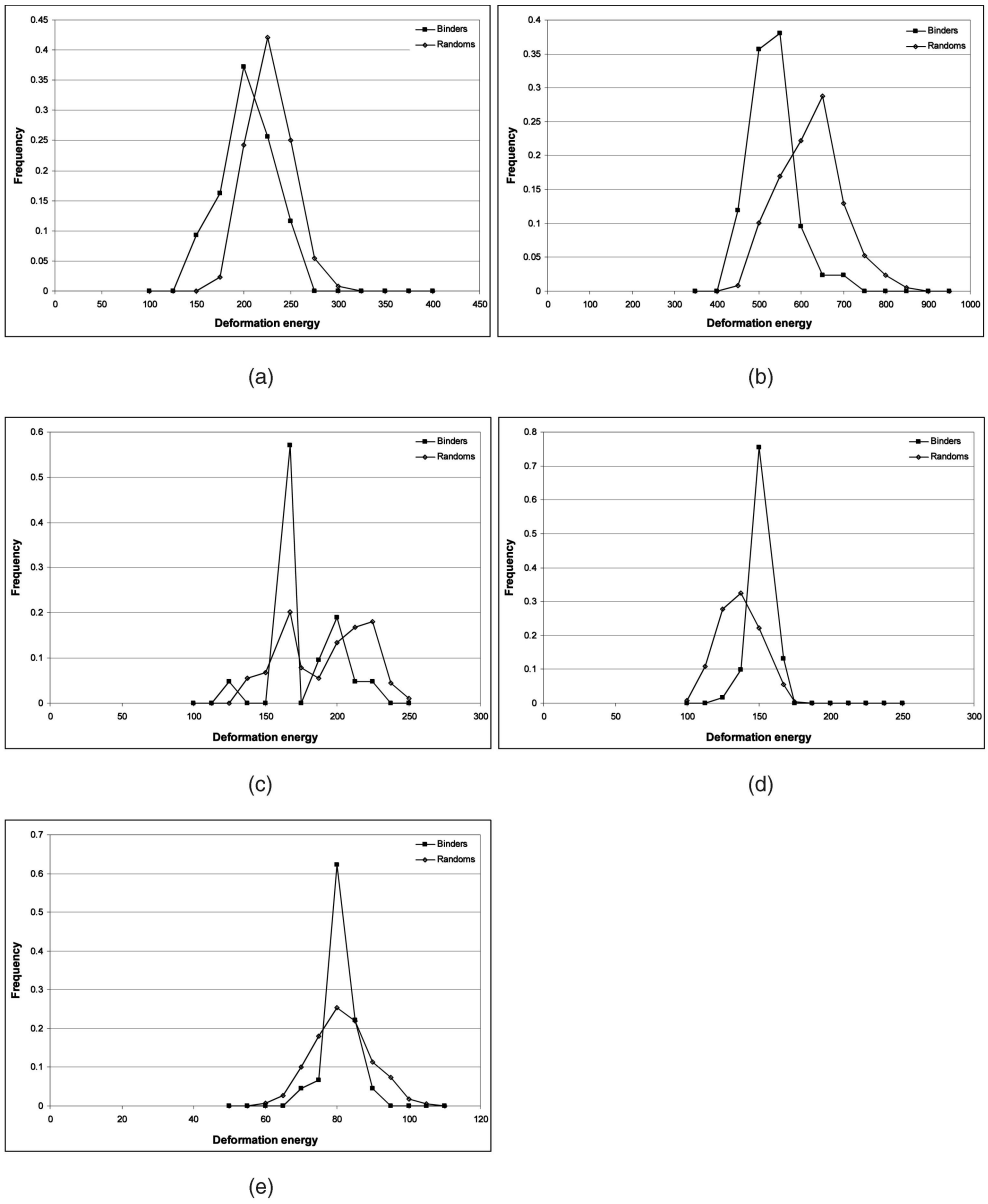

Fig. 2. Distribution of deformation energies of known binding sequences and random sequences drawn from the same base frequencies as each protein's native organism. (a) IHF, $p < \text{le} - 7$. (b) CRP, $p < \text{le} - 22$. (c) I-CreI, $p < 0.01$. (d) I-PpoI, $p < \text{le} - 53$. (e) SRF, $p < \text{le} - 33$. p is the significant difference from random.

advantageous to the process of converting binding energy into local distortion of the active site and, consequently, rate enhancement.

In contrast, I-CreI is a member of the LAGLIDADG family of homing endonucleases, named for the sequence motif used to structure the DNA substrate and a catalytic metal ion in the enzyme active site. I-CreI has an almost opposite balance of kinetic parameters in comparison to I-PpoI. It features high affinity for its cleavage sites, e.g., K_d of 0.1 nM, and substantially lower turnover rate of 5e-4 s^{-1} [24]. The scissile phosphates of the two strands in undistorted B-form DNA are positioned almost optimally across the minor groove for catalysis mediated by a single catalytic metal ion in the active site of the enzyme. While I-CreI does introduce a bend, the deformation it generates in its DNA substrate is neither as severe as I-PpoI nor as localized to the scissile phosphates. With such a mechanism, a "flexible" DNA substrate might be more advantageous to I-CreI, just as for site-specific binding proteins.

The perceptron results are displayed graphically in Fig. 3. This shows the distribution of perceptron scores (weighted deformation energy) for the binding sequences and for random sequences for each protein. The perceptron has increased the separation between the distribution of scores between binding and random sequences in all cases, though, for CRP and I-CreI, the separation is significantly lower than for the other three proteins. This is surprising for CRP, which had a large separation between the deformation energy of the known binding sites and random sequences.

Each feature used for the perceptrons is based on a specific dinucleotide combination at a specific location of the overall sequence, which is projected to a scalar (the deformation energy at that base pair step) by a parameter-free calculation

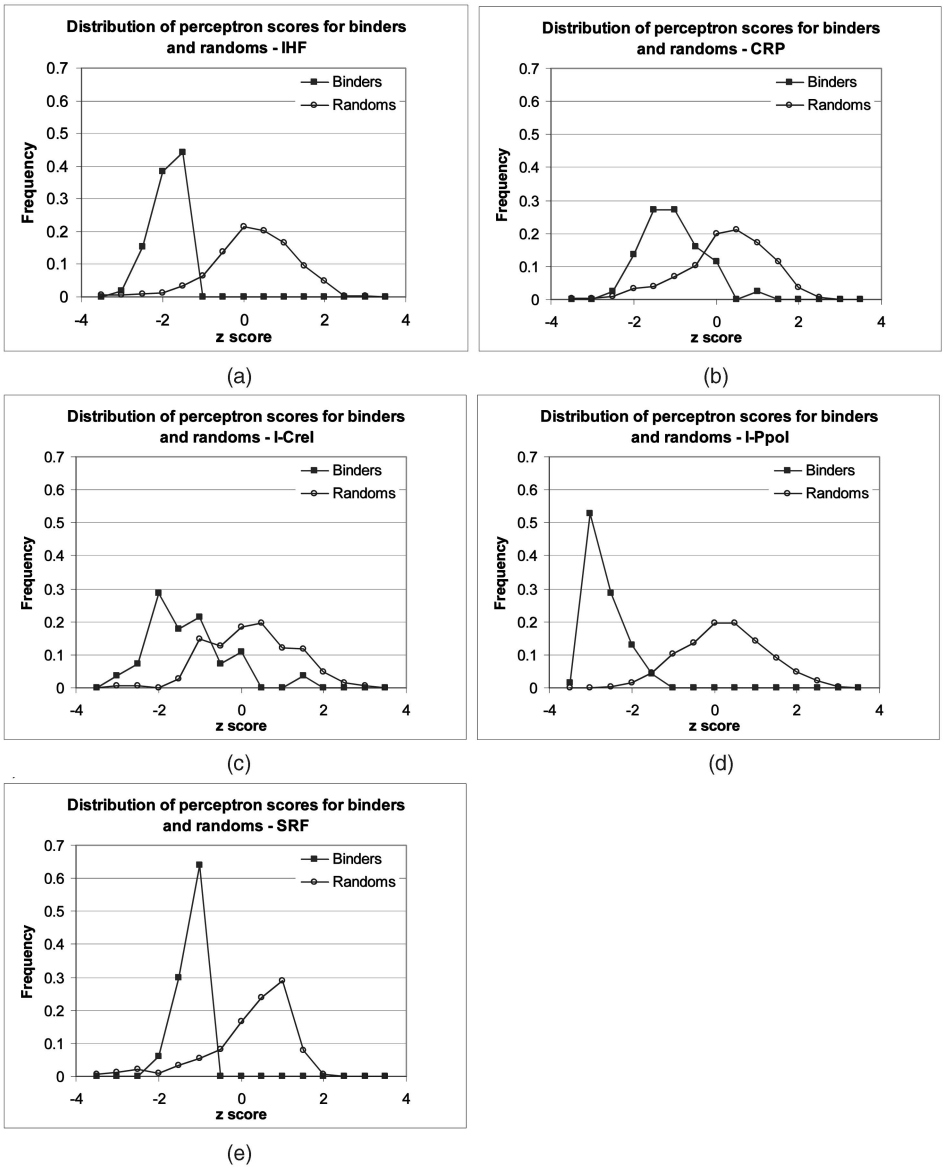

Fig. 3. Distribution of perceptron scores for binding sequences and random sequences for (a) IHF, (b) CRP, (c) I-CreI, (d) I-PpoI, and (e) SRF. Scores are expressed as z scores relative to the scores of the random sequences. All histograms are normalized to have unit area.

using only the bound DNA structure. Thus, for each position in the sequence, a feature will always be one of 16 values (one for each possible dinucleotide combination, but different values at each position) and, so, is an encoding of the dinucleotide at that position, in a sense. Perhaps the perceptron is simply focusing on a few key positions in the binding sequence that are highly conserved in all of the known examples? This seems unlikely because the resulting values are not linearly separable in the general case. For example, if deformation energy assigned similar values to both a conserved and a nonconserved dinucleotide at a given position or mapped a conserved dinucleotide to the middle of the distribution of values from nonconserved dinucleotides, then no possible weight setting could accept the conserved one and exclude the nonconserved ones. The only case in which a conserved dinucleotide is linearly separable from the nonconserved dinucleotides occurs when deformation energy maps the conserved dinucleotide to one or the other tail of the distribution of 16 dinucleotide deformation energy values at that position. In that case, however, parameter-free structural considerations identify the conserved dinucleotide as the optimal (tail-end) sequence at that position, so perhaps it was conserved because it is structurally appropriate. Alternatively, the perceptron weights could aggregate many weakly conserved positions by the optimal weighted sum

TABLE 3
Classifier Accuracy and Area under ROC Curve for Classifiers Trained on Feature Sets from Five Proteins

Protein	Cross-validated accuracy	Area under ROC curve
IHF	0.956	0.960
CRP	0.908	0.874
I-CreI	0.890	0.837
I-PpoI	0.979	0.997
SRF	0.963	0.989

of small distributional shifts at each position, similarly to how an extended weight matrix aggregates many weakly conserved columns for a degenerate sequence-based motif. Previously, we have shown that there can be substantial carry-over from structure-based features to many common sequence-based representations [17], including weight matrices. In other work, structure-based features yielded better predictors of function than did sequence-based features [27]. The results here also support the idea of informative structure-based features. The dividing line between sequence and structure is not clearly drawn. Indeed, sequence indirectly encodes all of indirect recognition by encoding structures that affect function. The key question is the level of abstraction upon which to base predictions and explanations.

Table 2 shows the sequences that have the lowest deformation energy when threaded onto each of the five protein-DNA co-crystal structures. IHF and SRF both show significantly more matches than would be expected by chance ($p < 2e-6$ and $p < 2e-3$, respectively). This indicates some correspondence between structure and sequence. For the other three proteins, the number of matches is well within the expected range. This is also true when the sequence with the largest deformation energy is calculated for I-PpoI, which seems to prefer relatively stiff DNA sequences.

In summary, this work examined whether deformation energy helps to discriminate preferred binding sites from other sites in DNA. Five DNA-binding proteins were studied. In each case, the binding sites for these proteins showed statistically significant differences in mean deformation energy when compared with random sequences. These results support DNA deformation energy as an indirect recognition mechanism across a wider range of DNA-binding proteins.

ACKNOWLEDGMENTS

This work was supported in part by a grant from the US National Institutes of Health (GM68903 to G.W. Hatfield). K.A. Aeling is the recipient of a UCI BIT Program Predoctoral Fellowship (NIH T15 LM-07443). N.R. Steffen was the recipient of a UCI BIT Program Predoctoral Fellowship (NIH T15 LM-07443). Phoebe Rice kindly supplied the smoothed atomic model of bound IHF. The authors thank Wilma Olson and Victor Zhurkin for discussion about their potential. Domino models made using X3DNA by X.-J. Lu. Molecular visualizations made using RASMOL by R. Sayle. Code and supplemental data are available at http://www.igb.uci.edu.

REFERENCES

[1] B.R. Szymczyna and C.H. Arrowsmith, "DNA Binding Specificity Studies of Four ETS Proteins Support an Indirect Read-Out Mechanism of Protein-DNA Recognition," *J. Biological Chemistry*, vol. 275, pp. 28363-28370, 2000.

[2] P.F. Baldi and R.H. Lathrop, "DNA Structure, Protein-DNA Interactions, and DNA-Protein Expression," *Session Introduction at Pacific Symp. Biocomputing*, 2001.

[3] M.E. Hogan and R.H. Austin, "Importance of DNA Stiffness in Protein-DNA Binding Specificity," *Nature*, vol. 329, pp. 263-266, 1987.

[4] M.M. Gromiha, "Influence of DNA Stiffness in Protein-DNA Recognition," *J. Biotechnology*, vol. 117, pp. 137-145, 2005.

[5] R.E. Harrington and I. Winicov, "New Concepts in Protein-DNA Recognition: Sequence-Directed DNA Bending and Flexibility," *Programming Nucleic Acid Research in Molecular Biology*, vol. 47, pp. 195-270, 1994.

[6] S. Chen, A. Gunasekera, X. Zhang, T.A. Kunkel, R.H. Ebright, and H.M. Berman, "Indirect Readout of DNA Sequence at the Primary-Kink Site in the CAP-DNA Complex: Alteration of DNA Binding Specificity through Alteration of DNA Kinking," *J. Molecular Biology*, vol. 314, pp. 75-82, 2001.

[7] P. Baldi, Y. Chauvin, S. Brunak, J. Gorodkin, and A.G. Pedersen, "Computational Applications of DNA Structural Scales," *Proc. Int'l Conf. Intelligent Systems in Molecular Biology*, vol. 6, pp. 35-42, 1998.

[8] R. Liu, T.W. Blackwell, and D.J. States, "Conformational Model for Binding Site Recognition by the E.Coli MetJ Transcription Factor," *Bioinformatics*, vol. 17, pp. 622-633, 2001.

[9] X.J. Lu, Z. Shakked, and W.K. Olson, "A-Form Conformational Motifs in Ligand-Bound DNA Structures," *J. Molecular Biology*, vol. 300, pp. 819-840, 2000.

[10] W.K. Olson, A.A. Gorin, X.J. Lu, L.M. Hock, and V.B. Zhurkin, "DNA Sequence-Dependent Deformability Deduced from Protein-DNA Crystal Complexes," *Proc. Nat'l Academy of Sciences USA*, vol. 95, pp. 11163-11168, 1998.

[11] C.R. Calladine and H.R. Drew, "Principles of Sequence-Dependent Flexure of DNA," *J. Molecular Biology*, vol. 192, pp. 907-918, 1986.

[12] H. Kono and A. Sarai, "Structure-Based Prediction of DNA Target Sites by Regulatory Proteins," *Proteins*, vol. 35, pp. 114-131, 1999.

[13] M. Michael Gromiha, J.G. Siebers, S. Selvaraj, H. Kono, and A. Sarai, "Intermolecular and Intramolecular Readout Mechanisms in Protein-DNA Recognition," *J. Molecular Biology*, vol. 337, pp. 285-294, 2004.

[14] A. Sarai, S. Selvaraj, M.M. Gromiha, J.G. Siebers, P. Prabakaran, and H. Kono, "Target Prediction of Transcription Factors: Refinement of Structure-Based Method," *Proc. 12th Int'l Conf. Genome Informatics*, 2001.

[15] S. Selvaraj, H. Kono, and A. Sarai, "Specificity of Protein-DNA Recognition Revealed by Structure-Based Potentials: Symmetric/Asymmetric and Cognate/Non-Cognate Binding," *J. Molecular Biology*, vol. 322, pp. 907-915, 2002.

[16] N.R. Steffen, S.D. Murphy, R.H. Lathrop, M.L. Opel, L. Tolleri, and G.W. Hatfield, "The Role of DNA Deformation Energy at Individual Base Steps for the Identification of DNA-Protein Binding Sites," *Genome Information Series Proc. Workshop Genome Information*, vol. 13, pp. 153-162, 2002.

[17] N.R. Steffen, S.D. Murphy, L. Tolleri, G.W. Hatfield, and R.H. Lathrop, "DNA Sequence and Structure: Direct and Indirect Recognition in Protein-DNA Binding," *Bioinformatics*, vol. 18, supplement 1, pp. S22-30, 2002.

[18] P.A. Rice, "Making DNA Do a U-Turn: IHF and Related Proteins," *Current Opinions in Structural Biology*, vol. 7, pp. 86-93, 1997.

[19] G. Parkinson, C. Wilson, A. Gunasekera, Y.W. Ebright, R.E. Ebright, and H.M. Berman, "Structure of the CAP-DNA Complex at 2.5 Angstroms Resolution: A Complete Picture of the Protein-DNA Interface," *J. Molecular Biology*, vol. 260, pp. 395-408, 1996.

[20] S.C. Schultz, G.C. Shields, and T.A. Steitz, "Crystal Structure of a CAP-DNA Complex: The DNA Is Bent by 90 Degrees," *Science*, vol. 253, pp. 1001-1007, 1991.

[21] G.M. Argast, K.M. Stephens, M.J. Emond, and R.J. Monnat Jr., "I-PpoI and I-CreI Homing Site Sequence Degeneracy Determined by Random Mutagenesis and Sequential in Vitro Enrichment," *J. Molecular Biology*, vol. 280, pp. 345-353, 1998.

[22] R. Pollock and R. Treisman, "A Sensitive Method for the Determination of Protein-DNA Binding Specificities," *Nucleic Acids Research*, vol. 18, pp. 6197-6204, 1990.

[23] L. Tolleri, "An Interdisciplinary Approach Employing Computational, Biochemical, and Genomic Methods to Examine the Effects of Chromosome Structure on the Regulation of Gene Expression," Universita degli Studi di Pavia e Firenze, Italy, 2002.

[24] B. Chevalier, D. Sussman, C. Otis, A.J. Noel, M. Turmel, C. Lemieux, K. Stephens, R.J. Monnat Jr., and B.L. Stoddard, "Metal-Dependent DNA Cleavage Mechanism of the I-CreI LAGLIDADG Homing Endonuclease," *Biochemistry*, vol. 43, pp. 14015-14026, 2004.

[25] B.L. Stoddard, "Homing Endonuclease Structure and Function," *Quarterly Rev. Biophysics*, vol. 38, pp. 49-95, 2006.

[26] E.A. Galburt, M.S. Chadsey, M.S. Jurica, B.S. Chevalier, D. Erho, W. Tang, R.J. Monnat Jr., and B.L. Stoddard, "Conformational Changes and Cleavage by the Homing Endonuclease I-PpoI: A Critical Role for a Leucine Residue in the Active Site," *J. Molecular Biology,* vol. 300, pp. 877-887, 2000.
[27] S.A. Danziger, S.J. Swamidass, J. Zeng, L.R. Dearth, Q. Lu, J.H. Chen, J. Cheng, V.P. Hoang, H. Saigo, R. Luo, P. Baldi, R.K. Brachmann, and R.H. Lathrop, "Functional Census of Mutation Sequence Spaces: The Example of p53 Cancer Rescue Mutants," *IEEE/ACM Trans. Computational Biology and Bioinformatics,* to appear.

Kimberly A. Aeling received the BS degree in molecular biology and biochemistry with a minor in information and computer science in 2002 from the University of California, Irvine. She is currently pursuing the PhD degree in microbiology and molecular genetics at the University of California, Irvine. She is the recipient of an NLM training fellowship in biomedical informatics.

Nicholas R. Steffen received the BS degree in engineering from San Diego State University and the PhD degree in information and computer science from the University of California, Irvine.

Matthew Johnson received the BS degree in astrophysics from the University of California, Los Angeles, in 2000. He then spent time in industry as a consultant before returning to academia. He is currently pursuing the PhD degree in information and computer science, with a concentration in artificial intelligence and computational neuroscience. He was awarded the Dean's Fellowship award for 2003-2005 and currently has a fellowship from 2005-2007 as an ARCS Foundation Scholar.

G. Wesley Hatfield received the BS degree in analytical biology from the University of California, Santa Barbara, and the PhD degree in biophysical chemistry from Purdue University. He is currently a professor in the Department of Microbiology & Molecular Genetics at the University of California, Irvine's (UCI) School of Medicine. He is the associate director of UCI's Institute for Genomics and Bioinformatics and the director of IGB's Computational Biology Research Laboratory.

Richard H. Lathrop received the BA degree in mathematics from Reed College in 1978 and the PhD degree in artificial intelligence from the Massachusetts Institute of Technology in 1990. He is currently an associate professor in the School of Information and Computer Science at the University of California, Irvine.

Donald F. Senear received the PhD degree from the University of Washington. He is currently a professor in the Department of Molecular Biology & Biochemistry at the University of California, Irvine. He is a member of the the Biophysical Society and the Protein Society.