The Role of DNA Deformation Energy at Individual Base Steps for the Identification of DNA-Protein Binding Sites

Nicholas R. Steffen[1] nsteffen@uci.edu
Scott D. Murphy[1] sdmurphy@uci.edu
Richard H. Lathrop[1] rickl@uci.edu
Michael L. Opel[2] mopel@uci.edu
Lorenzo Tolleri[2,3] Lorenzo_Tolleri@chiron.it
G. Wesley Hatfield[2] gwhatfie@.uci.edu

[1] Department of Information and Computer Science, University of California, Irvine, CA 92697, USA
[2] Department of Microbiology and Molecular Genetics, College of Medicine University of California, Irvine, CA 92697, USA

Abstract

We examine the use of deformation propensity at individual base steps for the identification of DNA-protein binding sites. We have previously demonstrated that estimates of the total energy to bend DNA to its bound conformation can partially explain indirect DNA-protein interactions. We now show that the deformation propensities at each base step are not equally informative for classifying a sequence as a binding site, and that applying non-uniform weights to the contribution of each base step to aggregate deformation propensity can greatly improve classification accuracy. We show that a perceptron can be trained to use the deformation propensity at each step in a sequence to generate such weights.

Keywords: DNA-protein binding, indirect recognition, DNA bending, perceptron learning

1 Introduction

DNA-protein recognition mechanisms have been grouped into two classes, direct and indirect. Direct recognition occurs when protein amino acid side chains recognize and form hydrogen bonds with specific base pairs in the major groove of the DNA helix, and provides a high degree of sequence specificity. Indirect recognition occurs when the protein recognizes structural features of the DNA molecule, including minor groove or backbone structural properties, intrinsic DNA curvature, hydration spines or other ordered water structures, salt bridges, and DNA flexibility or deformability [4, 22]. While the structural basis for direct recognition is well understood [2], indirect recognition is much less well understood and has been called the "missing half" of binding site recognition.

While experimental work has been performed to study indirect recognition [4, 22], few computational results have been generated. Hidden Markov models have been combined with DNA structural scales to find promoters and DNA structural patterns [1]. B-form to A-form DNA structural changes [11] have been characterized by the same deformation energy we use here. 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 [10]. Also, Sarai and colleagues have applied structure-based threading methods to both direct and indirect recognition of DNA-protein binding sites for a large number of proteins [9, 16]. They showed that both direct and indirect recognition can contribute to specificity, that their relative contributions vary, and that

[3] Current address: Chiron S.p.A., Via Fiorentina, 1, Siena, 53100, Italy.

combining direct and indirect recognition aspects into their model provides improved accuracy. We have shown that structure can be used to partially separate specific IHF protein-DNA binding sites from random sequences, and we have further demonstrated that pure sequence motifs resulting from transformation of a pure structure representation retain discriminatory ability [19].

IHF is a ubiquitous protein in prokaryotic organisms and employs both direct and indirect recognition. It is involved in both the maintenance of DNA structure and gene regulation [13]. In its architectural role, it binds to DNA as a histone-like protein to organize chromosome structure. In addition to its architectural functions, it uses its DNA bending ability to form recombinogenic DNA complexes, and serves as a transcription factor for global gene regulation. Proteins with these general properties are present in all living cells [7].

IHF prefers specific binding sites over random sites by more than three orders of magnitude. This binding specificity is achieved even though the binding sites have a highly degenerate consensus sequence. This indicates that indirect recognition plays a significant role in specific IHF binding to DNA [14]. Footprint analysis indicates that IHF binding sites are approximately 34 base pairs long. Within this region, the consensus sequence is considerably shorter - on the order of 13 to 15 base pairs [5, 6, 21]. This consensus sequence is degenerate with only nine bases showing significant conservation, and only three of these nine bases are involved in direct recognition [13]. This partially explains why sequence-based methods for identifying binding sites find IHF to be a difficult and challenging problem.

Our previous work has shown that deformation energy, or the propensity for a DNA duplex to assume a deformed structure, can partially discriminate between specific and random IHF binding sites, but that this discriminatory ability is poor [19]. When a protein binds to DNA, the DNA is deformed from its equilibrium shape. This shift to its bound conformation requires energy, which we estimate computationally. We have previously shown that this deformation propensity is correlated with binding affinity [19]. This implies that a DNA sequence that is easier to bend to the bound shape is more likely to be a strong binder than a DNA sequence that is relatively difficult to bend to the bound shape.

In this paper we extend our previous work by reformulating the deformation propensity calculation. In this reformulation the total deformation propensity is calculated as the sum of the weighted deformation propensities at each base step. We show that a perceptron can be trained to generate weights on each step that greatly improve the ability to classify a sequence as a binding site compared with using uniform weights. Our hypothesis here is that the deformation propensity at some steps is more useful than others for the prediction of high affinity binding. Thus, this paper focuses on the deformation propensity at individual steps, rather than on the aggregate deformation propensity.

2 Method

A set of 177 sequences known to bind to IHF was identified [20]. Test sets were constructed using this set or subsets of these sequences. Control sets were generated from random sequences containing the same GC content (50.8%) as *E. coli*. These sequences were threaded onto the structural motif defined by the co-crystal structure of an IHF protein-DNA complex and scored using an objective function. A perceptron was trained on the set of random sequences (negative examples) and known IHF binding sequences (positive examples) using the deformation propensity at each base step as the features. This produced a set of weights on the deformation propensity at each step. The average deformation propensity of each step in the set of specific binding sequences and the set of random sequences were calculated. The information content of each base in a set of binding sequences was determined and the correlation with deformation propensity was calculated. The average deformation propensity and perceptron scores over a sliding window 34 base pairs long across a set of aligned IHF binding sequences and associated flanking regions was calculated. This window was moved from an initial position 100 bp before the binding site to an ending position 100 bp after the binding site.

2.1 Structural Model and Deformation Propensity

We have shown elsewhere how to use the crystal structure of a DNA-IHF complex as the basis for a DNA structural motif [19]. Using this method, the motif is represented as the conformation of the DNA when bound to a protein, and an objective function is used to estimate the energy required to bend DNA from its native conformation to the bound shape. More specifically, DNA sequences are threaded onto the structural motif and the objective function is used to estimate the energy difference between the bound and unbound shapes for those sequences. Figure 1 shows a ribbon model of the DNA-IHF complex (a nick in the DNA crystal structure was repaired by the crystallographer, Phoebe Rice - personal communication). DNA has a persistence length of about 150 base pairs, but upon IHF binding, it is bent at an angle of 160 to 180 degrees over a span of only the 34 base pairs that comprise the binding sequence. Even among the relatively straight segments flanking the sharp bends, produced by proline intercalations into the DNA helix, the DNA undergoes significant deformations from its equilibrium shape.

Figure 1: Ribbon model of IHF bound to DNA.

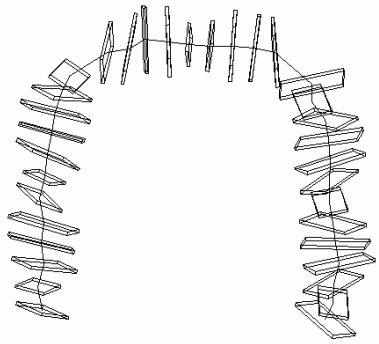

Figure 2: Domino model of Figure 1.

Figure 2 shows a domino model of the bound DNA. The base pairs have been modeled as rigid rectangular slabs, or dominoes, that provide the best least-squares fit to the base atoms [3]. Two adjacent dominoes form a base step. The relative displacement of one domino from another is fully defined by six parameters - three translations (slide, shift, and rise) and three rotations (roll, tilt, and twist). Thus, the full structure of the bound DNA is defined by these six parameters at each base step - in this specific case requiring six parameters at 33 steps. Moving each base step from its equilibrium position requires energy that is a function of the amount that it is moved and the type of nucleotides that compose the step. Olson *et al.* modeled these deformations as a spring using a harmonic function [12]. This requires an equilibrium value and force constants for each pair of the six step displacement parameters for each pair of nucleotides. They empirically derived these values using an inverse harmonic analysis of the base displacement parameters in crystal structures of 70 DNA-protein complexes including some homologous structures and the IHF/DNA complex.

The propensity for DNA to bend from its equilibrium shape to the protein-bound shape is estimated by summing the deformation energies of each base step in the region of interest. The deformation propensity $\Delta E(i,\, x,\, y)$ estimates 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 propensity $\Delta E_{total}(S)$ estimates the energy difference for sequence S between its equilibrium and bound conformations, i.e. its deformation propensity:

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

$$\Delta E(i,x,y) = \frac{1}{2}\sum_{j=1}^{6}\sum_{k=1}^{6} f_{jk}\Delta\theta_j\Delta\theta_k(i,x,y) \tag{2}$$

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

where S is a sequence, L is the length of sequence S, $\theta_j(i)$ is base step parameter j at base step i, $\theta_j^0(x,y)$ is the equilibrium value of base step parameter j for nucleotide constituents x and y, and f_{jk} are the coefficients impeding deformation.

If the contributions at each base step are weighted non-uniformly, equation (1) must be modified to include the weights at each step:

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

where W_i is the weight applied to the deformation propensity at base step i.

It is important to note that the computation of deformation propensity using this formulation requires only knowledge of the nucleotides at each base step and the relative positions of each domino. Because the structure is known from the crystallographic data, the energy is a function only of the sequence. We can thus thread any sequence onto the structure and compute its deformation propensity.

2.2 Structural Parameters, Information Content, and Deformation Propensity at Individual Base Steps

The six structural parameters at each base step were extracted from the DNA/IHF complex. Using these parameters, the average deformation propensity at each base step was calculated for the random sequences and the known high affinity binding sites as described in Section 2.1. The information content of each position in the set of binding sites was calculated according to Schneider and Stephens [18]:

$$R(l) = 2 - H(l) \tag{5}$$

$$H(l) = -\sum_{b=a}^{t} f(b,l) \log_2 f(b,l) \tag{6}$$

where *R(l)* is the information content at position l, 2 is the maximum uncertainty, *H(l)* is the uncertainty at position l, b is one of the bases (a,c,g,t), and *f(b,l)* is the frequency of base b at position l in the set of sequences.

2.3 Perceptron Learning

A set of 79 known IHF binding sites with high binding affinity occurring in *E. coli* and related organisms was collected from the literature [20] and their deformation energies were estimated. A set of 1,000 random sequences of length 34 with the same GC content as *E. coli* (50.8%) was generated and their deformation energies were estimated. It is possible that the set of random sequences contained one or more specific binding sequences. A perceptron was trained using the test set of known binding sequences as positive examples and the random sequences as negative examples. The 33 features were the deformation propensity at each of the 33 base steps. The perceptron learned weights for each feature and a threshold for the weighted deformation propensity in order to provide the maximum separation between positive examples (binding sites) and negative examples (random sequences). The perceptron score was calculated using Equation (4) with the weights generated by the perceptron. To prevent memorization and better estimate the generality, leave-one-out cross validation was performed in all tests.

2.4 Proximity Scan

In another approach to determine whether binding sites can be distinguished from non-binding sites based on their aggregate deformation propensity, the deformation propensity of 34 base pair segments of regions containing binding sites was calculated. To do this, 28 sequences containing *E. coli* binding sites were aligned using the Smith-Waterman algorithm [17]. The 100 base pairs flanking these binding

sites were retained, yielding a set of sequences spanning 234 base pairs centered on the binding site. For each sequence, the deformation propensity for a 34 base pair segment was calculated at all displacements from 100 base pairs before the binding site to 100 base pairs after the binding site. The deformation propensity for each offset was averaged over the 28 sequences in the set, yielding 200 average deformation propensity values as a function of the offset from the binding site. This calculation was repeated using the perceptron weights to non-uniformly weight the deformation propensity contributions of each base step.

3 Results

Here we demonstrate that non-uniformly weighting the contributions of each base step improves our ability to distinguish specific, high affinity IHF binding sites from non-specific, low affinity IHF binding sites when compared with uniform weighting. We also show that deformation propensity, though a function of sequence, may identify aspects of sequence not found by conventional sequence analysis tools.

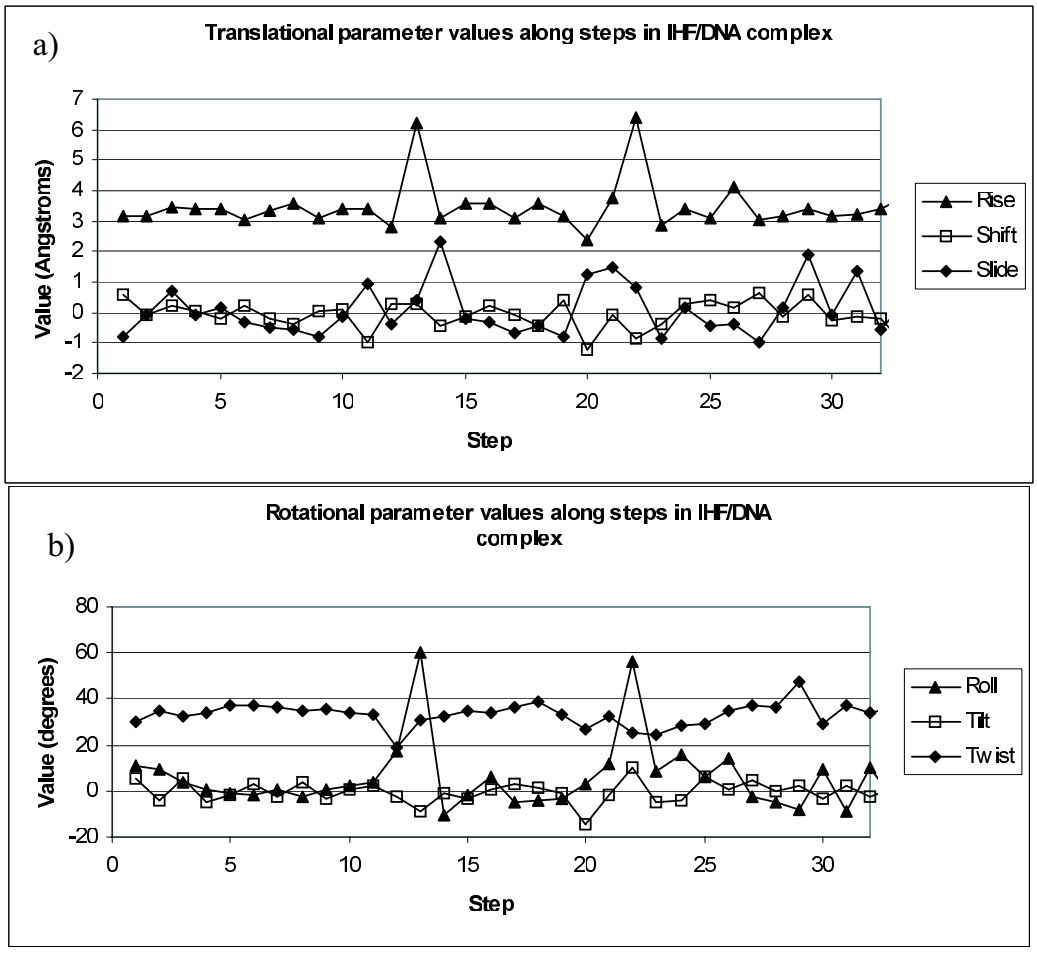

Figure 3: a) Values of slide, shift, and rise for base steps along the DNA-IHF complex. b) Values of roll, tilt, and twist for base steps along the DNA-IHF complex.

Figure 3 shows the values of the six structural parameters at each step in the reference DNA sequence. The proline intercalation sites are steps 13 and 22. The direct recognition sites are bases 21, 22, and 29. The consensus sequence includes bases 18 through 30.

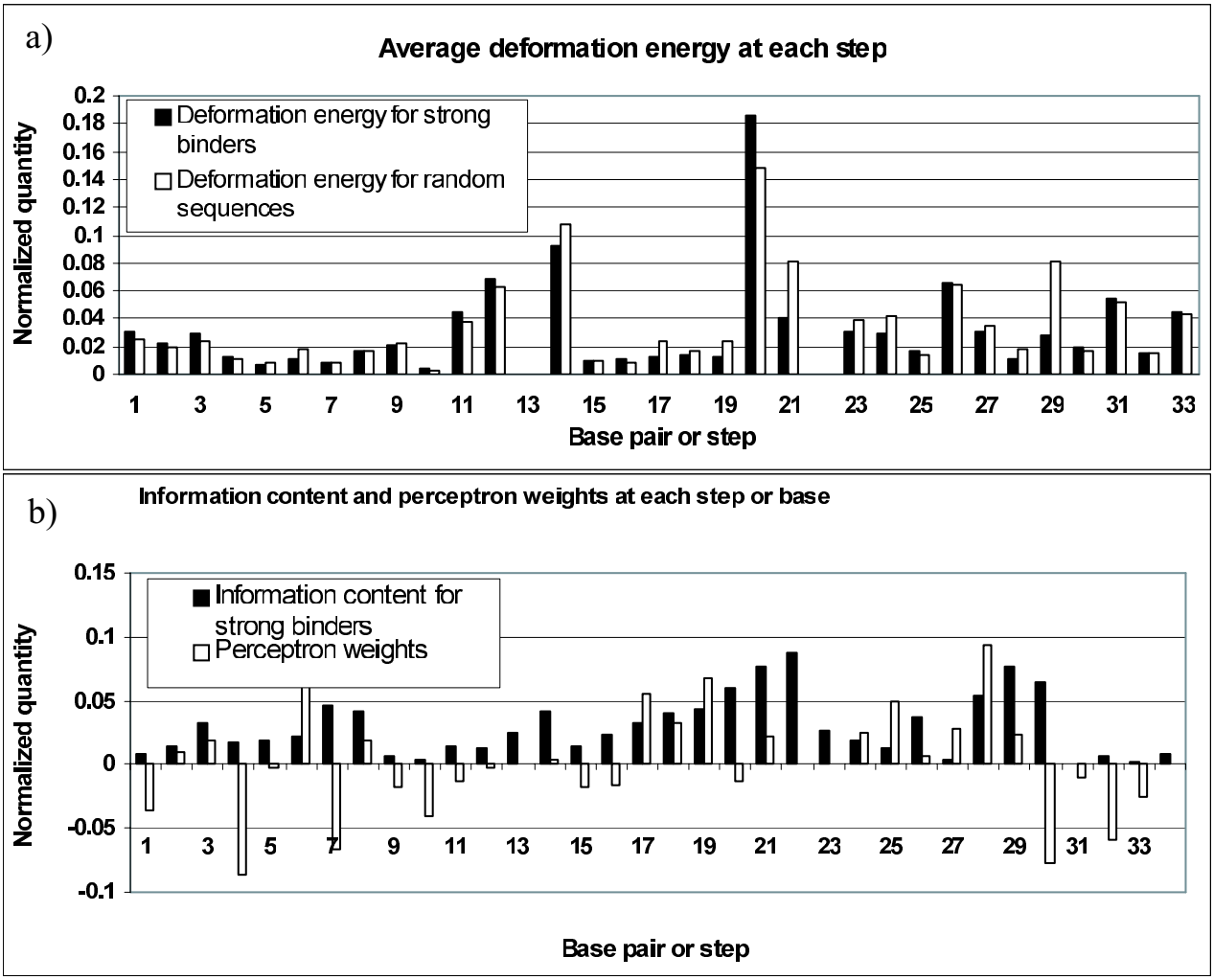

Figure 4: a) Average deformation propensity for specific and random IHF binding sequences, b) information content, and perceptron weights at each position (base step or base) in the DNA bound to IHF. Deformation propensities, information content and perceptron weights have been normalized so that their total over all base steps (deformation propensity, perceptron weights) or bases (information content) is 1.0.

The data in Figure 4 show the average deformation propensity at each base step for the specific binding sites and random sequences, the weights generated by the perceptron as described in Section 2.3, and the information content of each site in the set of known binding sequences as described in Section 2.2. The proline intercalation sites are steps 13 and 22. The direct recognition sites are bases 21, 22, and 29. The consensus sequence includes bases 18 through 30. For the specific binding sites, there is very little correlation between the deformation propensity at a base step and the information content of the sequence. This correlation was performed twice, using the bases on each side of the steps because a base step involves the bases on either side of the step. The correlation coefficient between the deformation propensity and the information content at bases 1 through 33 is 0.12, and between the deformation propensity and the information content at bases 2 through 34 is 0.15. However, the sign on the perceptron weight at a given step agrees with the sign of the difference between the deformation propensities of the specific and random binding sequences at 25 of 31 steps (p <0.0005 assuming a binomial distribution).

3.1 Perceptron Learning Results

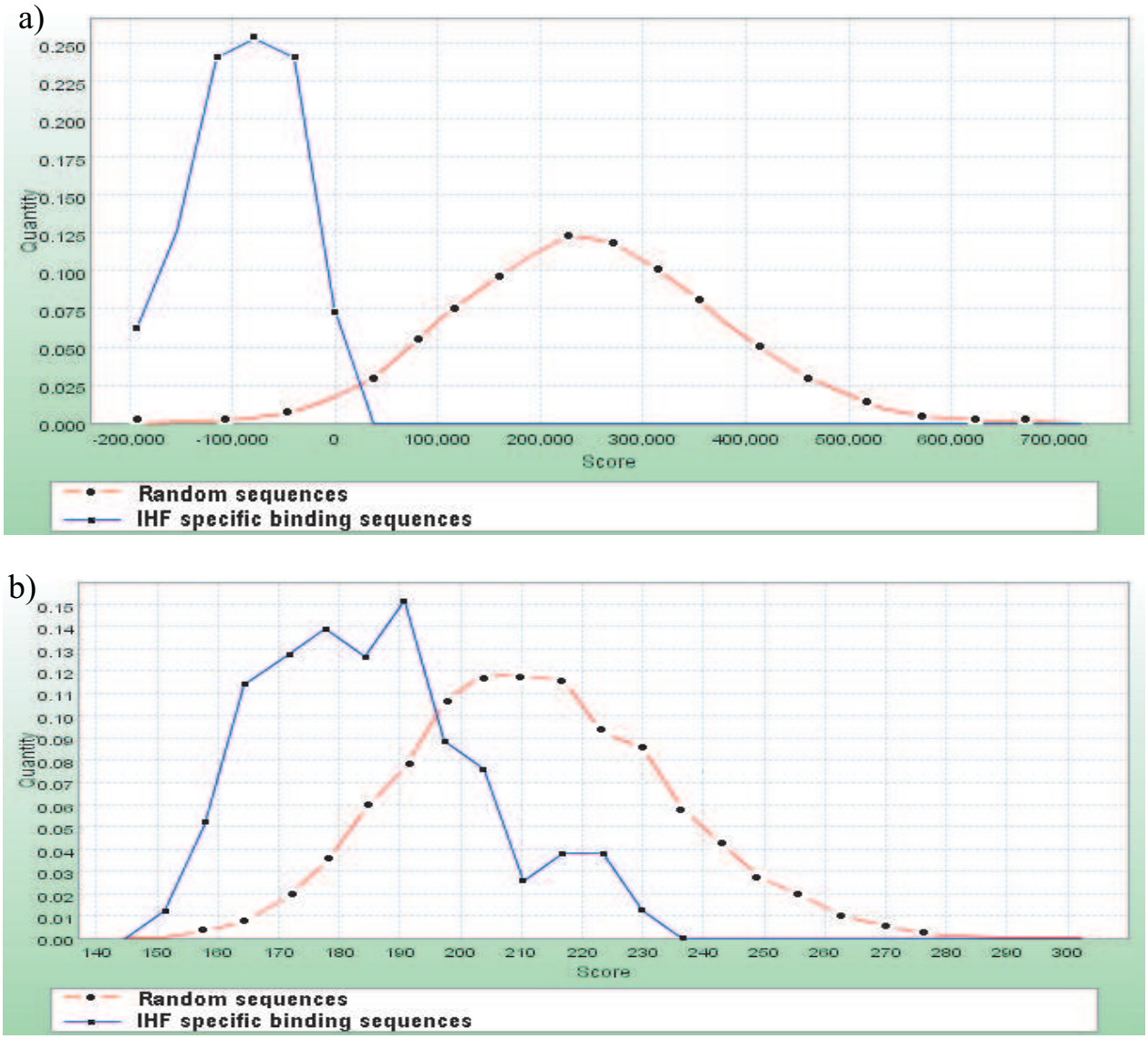

Figure 5: a) Perceptron scores for 79 known binding sites and 1,000 random sequences, b) Distribution of deformation propensity for 79 known binding sites and 1,000 random sequences. The area under all curves has been normalized to 1.

The distribution of deformation propensity, generated using uniform weights on each step's contribution, for the same 79 known binding sites and 1,000 random sequences is shown in Figure 5a. The cross-validated data in Figure 5b show that a perceptron can learn weights for the deformation propensity at each base step that improve the separation of known specific binding sites from random sequences. The perceptron score represents the deformation propensity with learned non-uniform weights applied to the deformation propensity at each base step. There is a small overlap between the two distributions, which contains sequences that either were misclassified by the perceptron or are IHF binding sites that occur by chance among the random sequences. 68 of the 79 known binding sequences were correctly classified by the perceptron, and all but 15 of the 1,000 random sequences were classified as non-binding sites. Thus, the overall classification error rate was 2.6%.

3.2 Proximity Scan Results

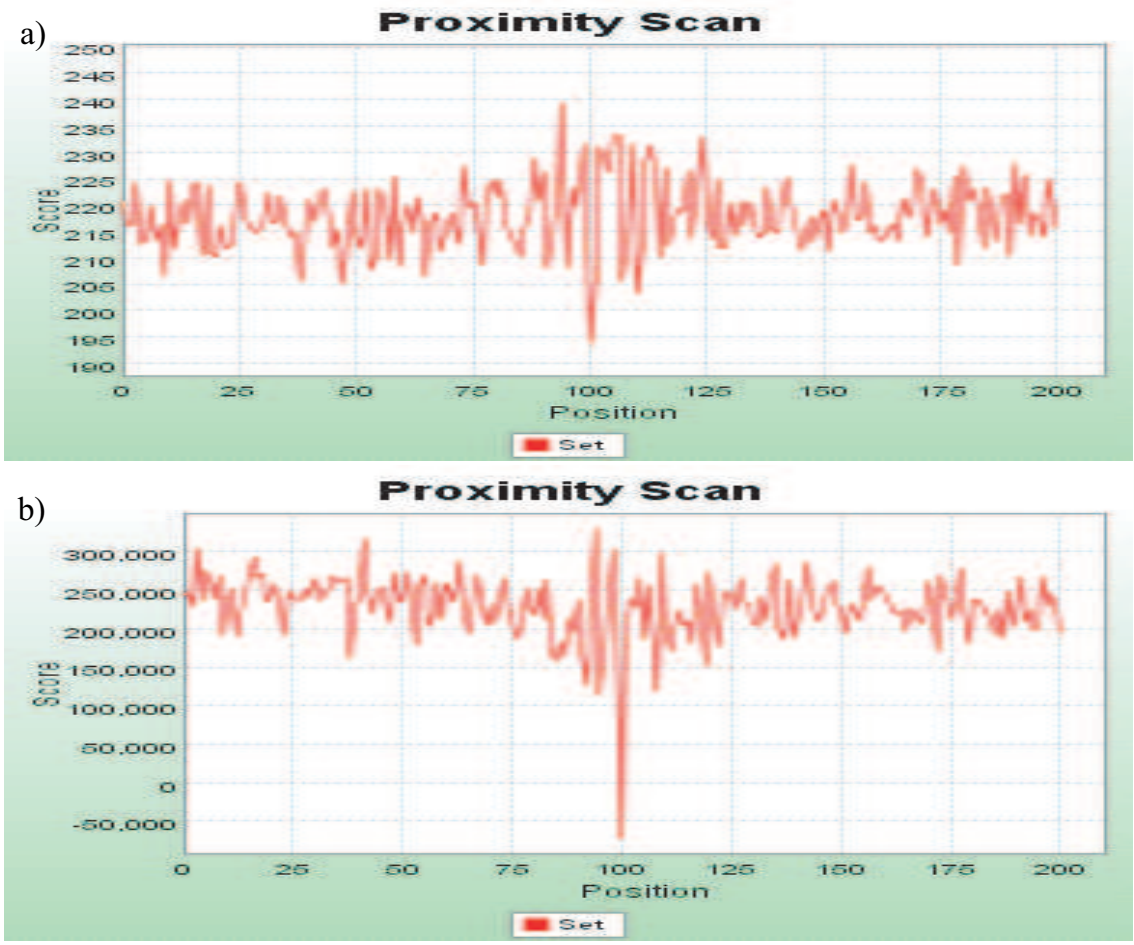

Figure 6: Deformation propensity as a function of the displacement from the specific binding sites. Deformation propensity is averaged over 28 aligned known binding sites. a) Uniform weights on deformation propensity at each step. b) Perceptron weights applied to deformation propensity at each step.

The deformation propensity of DNA segments of length 34, averaged over 28 known binding sites in *E. coli.* is shown in Figure 6. The lowest deformation propensity occurs when the DNA segment coincides with the binding site. The difference from the mean of the score at the binding site location was 3.8 σ when uniform weights were and 8.8 σ for the perceptron weighted score.

4 Discussion

This paper describes a computational method for the identification of DNA-protein binding sites based on the propensity for protein-induced DNA duplex deformation at each base step in a DNA sequence. This method estimates the amount of energy required to deform each base step in a DNA sequence to match a known conformation. It then learns weights that are applied to these step-wise deformation propensities. The sum of the weighted deformation propensities can be used to distinguish, with good accuracy, specific, high affinity binding sites from random sequences.

The data in Figure 4 show that the differences in deformation propensity for IHF binding to random or known binding sites is statistically significant. Unweighted deformation propensity cannot be used in isolation to accurately classify a sequence as a binding site or not as indicated by the large overlap between the deformation propensity of specific binding sites and the deformation propensity

of random sequences. However, the deformation propensity at all steps is not equally informative. Indeed, we show that classification accuracy can be significantly improved by applying non-uniform weights to the deformation propensity at each base step. In this work, we used perceptron learning to generate these weights. The weight at a given base step tends to be negative when the deformation propensity for the strong binding sites is larger than the deformation propensity at that step for the non-binders, and negative for the other case. This maximizes differences between the contributions to the total weighted deformation propensity of the random and known binding sequences. Larger weights at the more informative steps lead to much better separation between specific binding sites and random sequences than is achieved when the contribution of each base step is equally weighted (Fig. 5). Here the correlation between the deformation propensity and information content for the high affinity binding sequences was quite low, perhaps indicating that indirect recognition picks up features other than those contributed by sequence alone (Fig. 6).

Since high affinity protein-DNA interactions require both direct and indirect recognition, deformation propensity alone may not always be sufficient to separate binding from non-binding sites (Figure 4). This is to be expected in the case of IHF, which is known to use both recognition methods to find its binding sites. For example, Saecker and Record [15] have shown that the disruption of protein surface salt bridges can function to lower the energy costs of DNA wrapping. Presumably this results in a trading off of the cost of DNA deformation against the benefit of more energetically favorable contacts. Also, Holbrook *et al.* [8] have demonstrated two distinct binding modes for IHF which they interpret as specific and non-specific. In the specific binding mode, we suggest that the effects of stabilizing interactions, facilitated by DNA wrapping, are traded off against the costs of DNA deformation. In the non-specific mode, they show that DNA doesn't bend around IHF. In this case, we expect binding specificity to be low because the effects of these stabilizing interactions are not realized. Thus, deformation propensity is not expected to directly predict binding affinity. Nevertheless, it is clear that deformation propensity can be used as a computational tool to discriminate specific from nonspecific IHF binding sites with more accuracy than sequence based algorithms.

In summary, we have presented a computational method for identifying protein binding sites in DNA. This augments current methods that rely on sequence information for identifying sites of many important proteins (for example, see [10]). This may be especially important in proteins that bind DNA with highly degenerate sequence specificity, such as transcription factors in eukaryotic cells. Furthermore, many of these proteins are known to distort the structure of DNA at their binding sites. If indeed indirect recognition is the missing half of DNA-protein binding motifs, methods such as those described here may help to provide a more complete approach for the identification of all DNA-protein binding sites.

Acknowledgments

This work was supported in part by the National Institutes of Health (GM 55073 to GWH), the University of California Biotechnology Research and Education Program (2002-08 to RHL and GWH), and the UCI Institute of Genomics and Bioinformatics. NRS is a predoctoral fellow and MLO is a postdoctoral fellow supported by a National Institutes of Health training grant (LM 07443). Phoebe Rice kindly provided the repaired crystal structure. We are grateful to Wilma Olson and Victor Zhurkin for helpful discussions, and to Don Senear for valuable insights.

References

[1] Baldi, P., Chauvin, Y., Brunak, S., Gorodkin, J., and Pederson, A.G., Computational applications of DNA structural scales, *Intelligent Systems for Molecular Biology*, AAAI Press, Menlo Park, CA, USA, 35–42, 1998.

[2] Branden, C. and Tooze, J., *Introduction to Protein Structure*, Garland Publishing, Inc., 1991.

[3] Calladine, C.R. and Drew, H.R., Principles of sequence-dependent flexure of DNA, *J. Mol. Biol.*, 192:907–918, 1986.

[4] Chen, S., Gunasekera, A., Zhang, X., Kunkel, T.A., Ebright, R.H., and Berman, H.M., 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. Mol. Biol.*, 314:75–82, 2001.

[5] Engelhorn, M., Boccard, F., Martin, C., Pretki, P., and Geiselmann, J., In vivo interaction of the *E. coli* integration host factor with its specific binding sites, *Nucleic Acids Res.*, 23:2959–2965, 1995.

[6] Goodrich, J.A., Schwartz, M.L., and McClure, W.R., Searching for and predicting the activity of sites for DNA binding proteins, *Nucleic Acids Res.*, 18:4993–5000, 1990.

[7] Hatfield, G.W and Benham, C.J., DNA topology-mediated control of global gene expression in *E. coli*, *Ann, Rev. Genet*, (in press).

[8] Holbrook, J.A., Tsodikov O.V., Saecker, R.M., and Record, M.T. Jr., Specific and non-specific interactions of integration host factor with DNA: Thermodynamic evidence for disruption of multiple IHF surface salt- bridges coupled to DNA binding, *J Mol.Biol.*, 310:379–401, 2001.

[9] Kono, H. and Sarai, A., Structure-based prediction of DNA target sites by regulatory proteins, *Proteins: Structure, Function, and Genetics*, 35:114–131, 1999.

[10] Liu, R., Blackwell, T.W., and States, D.J., Conformational model for binding site recognition by the *E. coli* MetJ transcription factor, *Bioinformatics*, 17(7):622–633, 2001.

[11] Lu, X.-J., Shakked, Z., and Olson, W.K., A-form conformational motifs in ligand-bound DNA structures, *J. Mol. Biol.*, 300:819–840, 2000.

[12] Olson, W. K., Gorin, A., Lu, X., Hock, L., and Zhurkin, V., DNA sequence-dependent deformability deduced from protein-DNA crystal complexes, *Proc. Natl. Acad. Sci. USA*, 95:11163–11168, 1998.

[13] Rice, P.A., Making DNA do a U-turn: IHF and related proteins, *Curr. Op. Struct. Biol.*, 7:86–93, 1997.

[14] Rice, P.A., Yang, S.-W., Mizuuchi, K., and Nash, H.A., Crystal structure of and IHF-DNA complex: A Protein-induced DNA U-turn, *Cell*, 887:1295–1306, 1996.

[15] Saecker, R.M. and Record, M.T., Protein surface salt bridges and paths for DNA wrapping, *Curr. Opin. Struct. Biol.*, 12:311–319, 2002.

[16] Sarai, A., Selvaraj, S., Gromiha, M.M., Siebers, J.-G., Prabakan, P., and Kono, H., Target prediction of transcription factors: Refinement of structure-based method, *Genome Informatics*, 12:384–385, 2001.

[17] Smith, T.F. and Waterman, M.S., Identification of common molecular subsequences, *J. Mol. Biol.*, 147(1):195–197, 1981.

[18] Schneider, T.D. and Stephens, R.M., Sequence logos: A new way to display consensus sequences, *Nucleic Acids Res.*, 18:6097–6100, 1990.

[19] Steffen, N.R., Murphy, S.D., Tolleri, L., Hatfield, G.W., and Lathrop, R.H., DNA sequence and structure: Direct and indirect recognition in protein-DNA binding, *Bioinformatics*, 18:(Suppl 1):S22–S30, 2002.

[20] Tolleri, L., An interdisciplinary approach employing computational, biochemical, and genomic methods to examine the effects of chromosome structure on the regulation of gene expression, *Ph.D. Thesis, Universita degli Studi di Pavia e Firenze*, 2002.

[21] Ussery, D., Larsen, T.S., Wilkes, K.T., Friis, C., Worning, P., Krogh, A., and Brunak, S., Genome organization and chromatin structure in *Escherichia coli*, *Biochimie*, 83:201–212, 2001.

[22] Wenz, C., Jeltsch, A., and Pingoud, A., Probing the indirect readout of the restriction enzyme *Eco*RV. Mutational analysis of contacts to the DNA backbone, *Journal of Biological Chemistry*, 271(10):5565–5573, 1996.