BIOINFORMATICS

Vol. 18 Suppl. 1 2002
Pages S22–S30

DNA sequence and structure: direct and indirect recognition in protein-DNA binding

N. R. Steffen[1], S. D. Murphy[1], L. Tolleri[2], G. W. Hatfield[3] and R. H. Lathrop[1]

[1]Information and Computer Science, University of California, Irvine, Irvine, CA, 92697-3425, USA, [2]Chiron S.p.A., Via Fiorentina, 1, Siena, 53100, Italy and [3]Molecular Biology, Genetics and Biochemistry, University of California, Irvine, Irvine, CA, 92697-4025, USA

ABSTRACT

Motivation: Direct recognition, or direct readout, of DNA bases by a DNA-binding protein involves amino acids that interact directly with features specific to each base. Experimental evidence also shows that in many cases the protein achieves partial sequence specificity by indirect recognition, i.e., by recognizing structural properties of the DNA. (1) Could threading a DNA sequence onto a crystal structure of bound DNA help explain the indirect recognition component of sequence specificity? (2) Might the resulting pure-structure computational motif manifest itself in familiar sequence-based computational motifs?

Results: The starting structure motif was a crystal structure of DNA bound to the integration host factor protein (IHF) of *E. coli*. IHF is known to exhibit both direct and indirect recognition of its binding sites. (1) Threading DNA sequences onto the crystal structure showed statistically significant partial separation of 60 IHF binding sites from random and intragenic sequences and was positively correlated with binding affinity. (2) The crystal structure was shown to be equivalent to a linear Markov network, and so, to a joint probability distribution over sequences, computable in linear time. It was transformed algorithmically into several common pure-sequence representations, including (a) small sets of short exact strings, (b) weight matrices, (c) consensus regular patterns, (d) multiple sequence alignments, and (e) phylogenetic trees. In all cases the pure-sequence motifs retained statistically significant partial separation of the IHF binding sites from random and intragenic sequences. Most exhibited positive correlation with binding affinity. The multiple alignment showed some conserved columns, and the phylogenetic tree partially mixed low-energy sequences with IHF binding sites but separated high-energy sequences. The conclusion is that deformation energy explains part of indirect recognition, which explains part of IHF sequence-specific binding.

Availability: Code and data on request.

Contact: Nick Steffen for code and Lorenzo Tolleri for data. nsteffen@uci.edu, Lorenzo_Tolleri@chiron.it

Keywords: protein-DNA binding sites; sequence motifs or patterns; indirect recognition or readout; integration host factor; IHF.

INTRODUCTION

Sequence implies structure which implies function, and consequently clues to structure, function, and evolution abound in sequence patterns. The sequence patterns considered here govern protein-DNA binding. Some proteins bind more strongly to certain regions of DNA than to other regions, and this is governed by the free binding energy of the protein-DNA interaction.

Sequence motif representations

Common representations for sequence motifs include (a) small sets of short exact strings (e.g., van Helden *et al.*, 1998), (b) weight matrices (e.g., Lawrence *et al.*, 1993), and (c) consensus regular patterns (e.g., Vilo *et al.*, 2000).

Small sets of short exact strings, often those over-represented in the genome, contain strings that occur in protein binding sites or fragments of sites. A weight matrix has rows for bases and columns for sequence motif positions. Here, entries are the probability of observing the base corresponding to the row at the sequence motif position corresponding to the column. A consensus regular pattern (also called a regular expression pattern or a consensus sequence) has a disjunction of bases at each sequence motif position, sometimes augmented by variable-length gaps.

Multiple alignments and phylogenetic trees are also popular tools for depicting related sequences. A multiple alignment shows conservation in its columns, a phylogenetic tree in the way it clusters sequences.

Direct and indirect recognition

Direct recognition occurs when protein amino acid side-chains interact with specific bases in DNA sequences. The pattern of amino acid-nucleotide contacts extracted from a crystal structure can serve as a template for DNA bases

which define sequence motif positions (Benos *et al.*, 2001; Mandel-Gutfreund *et al.*, 2001). The individual entries of a weight matrix can be interpreted as the binding energies of single base-pairs to the protein's DNA-binding surface.

There is mounting evidence that DNA structural properties, beyond direct recognition of individual bases, significantly affect protein-DNA interactions (Baldi and Lathrop, 2001, and references therein). Several well-studied experimental systems exhibit what has come to be called indirect readout or indirect recognition. In this case, the protein appears to recognize structural properties of the double helix (Chen *et al.*, 2001, and references therein). Indirect readout mechanisms include recognition by the protein of structural features in the DNA major and minor grooves, backbone features, intrinsic curvature, hydration shells or spines, and flexibility or deformability.

Some computational methods have successfully identified indirect readout structure motifs. The deformation energy used here (Olson *et al.*, 1998) leads to low-energy movements characteristic of the transition between B-form and A-form DNA, and so to A-form conformational motifs (Lu *et al.*, 2000). Structure motifs based on minor groove opening (Liu *et al.*, 2000) show that including both sequence and structure motifs improves recognition, and strong minor groove base conservation in sequence logos has been shown to imply DNA distortion or base flipping (Schneider, 2001). DNA structural scales combined with Hidden Markov Models have been used to find promoters and discover DNA structural patterns (Baldi *et al.*, 1998). In the work most closely related to the threading aspects of this paper, Sarai and colleagues independently applied structure-based threading methods to both direct (Kono and Sarai, 1999) and indirect (Sarai *et al.*, 2001) recognition of protein-DNA binding sites. They showed that both mechanisms contribute significantly to protein-DNA recognition specificity, that their relative contribution varies, and that their combination increases accuracy.

Integration host factor (IHF) protein

IHF occurs ubiquitously in prokaryotes, and has both architectural and regulatory roles (Rice, 1997). It binds DNA non-specifically as a histone-like multipurpose bender of DNA, plays a role in replication, and is necessary for the formation of recombinogenic complexes. In its regulatory role, which concerns this paper, it binds DNA specifically and serves as a transcription factor. Proteins with these properties have been described in all organisms (Hatfield and Benham, 2002).

Its high specificity of binding is obtained despite a relaxed sequence dependency, which makes IHF binding a good example of indirect recognition (Tolleri, 2002). IHF interacts directly with only three bases in the DNA sequence (Rice *et al.*, 1996). This in part explains why consensus-based prediction methods perform poorly with IHF. Sequences that are known to bind IHF strongly will often, but not always, have a short core sequence motif and an AT-rich region flanking this core. The core motif is highly degenerate (Goodrich *et al.*, 1990; Engelhorn *et al.*, 1995; Ussery *et al.*, 2001).

This paper

This paper asks two questions: (1) Could threading a DNA sequence onto a crystal structure of bound DNA help explain the indirect recognition component of sequence specificity? (2) Might the resulting pure-structure computational motif manifest itself in familiar sequence-based computational motifs? The answer shown here is that: (1) Threading methods can partially explain both binding site recognition and binding strength. (2) DNA structure alone can be used to generate sequence motifs that partially recognize protein-DNA binding sites.

METHODS

When a protein binds to DNA, energy is required to deform the DNA from its native shape to its bound conformation. The hypothesis was that this deformation energy should be correlated with binding specificity and affinity, since a protein would be less likely to bind tightly where a large amount of energy is required than where the DNA readily assumes the bound shape.

This paper used a crystal structure and an objective function to estimate deformation energy. The structure motif was extracted from a crystal structure of a protein-DNA complex. It represented the shape of the DNA in its bound state. The objective function modelled the energy difference between DNA in bound and unbound conformations. DNA sequences were threaded onto the structure motif and scored using the objective function.

The structure motif was transformed automatically into several sequence representations: (a) small sets of short exact strings, (b) weight matrices, (c) consensus regular patterns, (d) multiple sequence alignments, and (e) phylogenetic trees. None had any trained parameters because the original structure motif had none. These were tested computationally on known IHF binding sites, random sequences, and *E. coli* intragenic regions. The process is outlined in Figure 1.

Pure-structure motif

A domino model for DNA bending represents the molecule as a series of rigid rectangular solids, or dominos, each taking the place of a base pair (Calladine and Drew, 1992). The relative position of successive dominos, called a dimer step, is specified by the six dimer step parameters: Shift, Slide, Rise, Tilt, Roll, and Twist.

Figure 2 shows an atomic model of a reference DNA sequence bound to the integration host factor (IHF) protein (Rice *et al.*, 1996). Figure 3 shows the corresponding

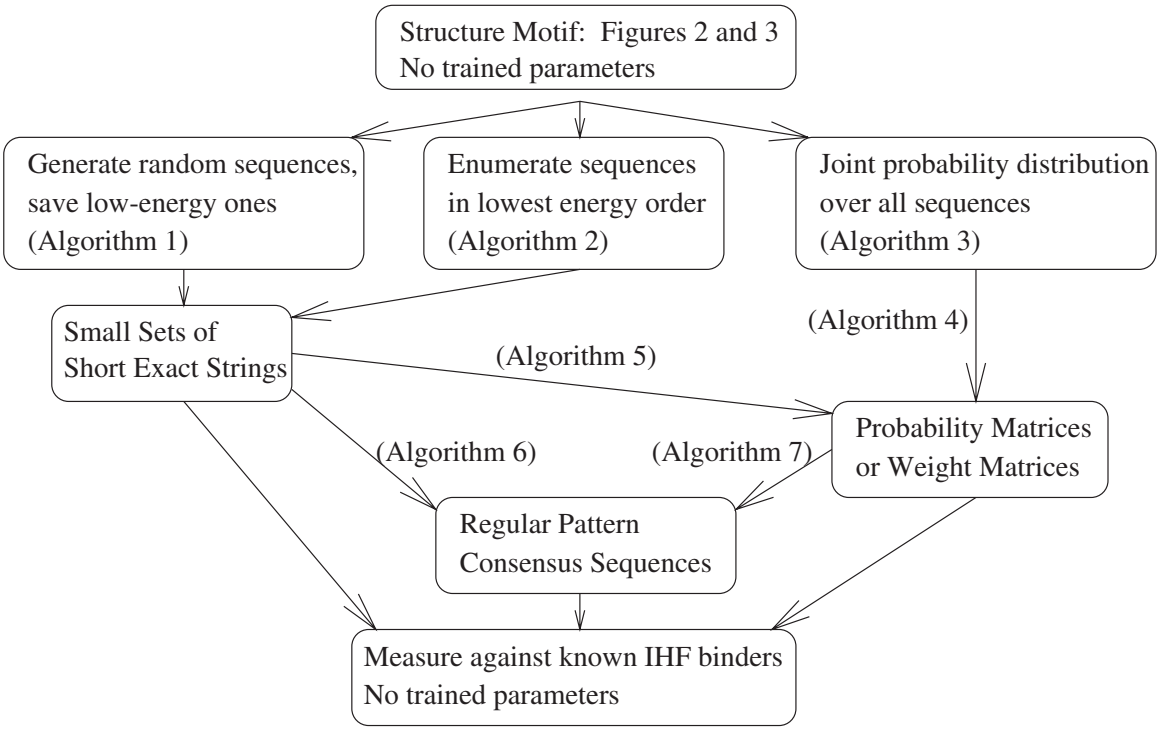


Fig. 1. Schematic outline of the transformations performed in this paper.

domino model, extracted by fitting a plate to the atoms of each base pair. The structure motif is operationalized by the $6 \times (L-1)$ matrix of dimer step parameters for $L-1$ dimers (L base pairs) extracted from the crystal structure.

Published objective function

Deformation of a dimer step requires energy that is a function of the nucleotides which compose it and the amount by which it is deformed. Olson *et al.* (1998) model this as a spring, using a harmonic function. This requires an equilibrium position and force constant for each pair of nucleotides and each pair of dimer step parameters. The necessary constants are estimated from the Boltzmann transform of a frequency analysis of the six dimer step parameters in crystal structures of proteins bound to DNA. Numeric values in this paper are as in (Olson *et al.*, 1998).

The objective function used here sums dimer energy over the dimer steps in the motif. A dimer step corresponds to adjacent dominos in Figure 3. The dimer energy $\Delta E(i, x, y)$ at dimer step i with flanking nucleotides x and y corresponds to the energy required to displace x and y from equilibrium to positions in the structure motif at step i. The total deformation energy $\Delta E_{total}(S)$ corresponds to the energy difference between sequence S in bound and unbound conformations.

Fig. 2. IHF bound to DNA. A crystallographic nick in the DNA has been smoothed (Phoebe Rice, personal communication).

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

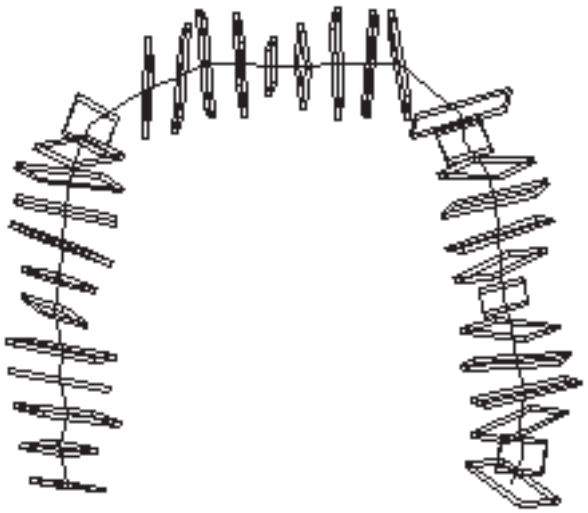


Fig. 3. Domino model from Figure 2.

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

$$\Delta\theta_j(i, x, y) = \theta_j(i) - \theta_j^0(x, y) \qquad (3)$$

where S is a sequence of length L, $\theta_j(i)$ is dimer step parameter j from dimer step i (i.e., the jth value of Shift, Slide, etc., from the structure motif at i), $\theta_j^0(x, y)$ is the equilibrium value of the dimer step parameter j for flanking bases x and y, and the f_{jk} are elastic constants impeding deformations (Olson *et al.*, 1998).

Structure = Markov network

Here we show that the pure-structure motif shown in Figure 3 is a Markov network (Pearl, 1988). Each domino corresponds to a node representing a random variable $X_i \in \{A, C, G, T\}$. Each dimer step $\theta(i)$ corresponds to an edge linking the nodes $\langle X_i, X_{i+1}\rangle$ on either side of the step. The graph cliques $\mathbf{c}_i = \langle X_i, X_{i+1}\rangle$ are the pairs of nodes flanking each dimer step.

Boltzmann's transform, $P \propto \exp(-E/k_BT)$, yields energies as the negative logarithm of unnormalized probabilities. Equation 1 yields an energy in units of k_BT, so the unnormalized probability of observing bases x and y at dimer i is $\exp(-\Delta E(i, x, y))$. Thus, $\exp(-\Delta E(i, x_i, x_{i+1}))$ is a compatibility function $g_i(\mathbf{c}_i)$ measuring the relative compatibility of any two bases in clique $\mathbf{c}_i$.

Consequently, a Gibbs potential yields a complete and consistent joint probability distribution over all bases at each node Pearl (1988, p. 105), i.e., over the space of all sequences on the structure motif. Because the underlying graph is linear, the normalizing constant, here α, as well as the marginal probability W_{ij} of observing base i at node j, can be computed in time linear in the length of the structure motif (Pearl, 1988).

Equation 5 relates $g_i(\mathbf{c}_i)$, a compatibility function on clique $\mathbf{c}_i$, to ΔE. Equation 6 gives the Gibbs potential by which the g_i yield a joint probability distribution over sequences. Equation 7 gives the normalizing constant (or partition function) as the sum over all sequences of the product over all cliques. Equation 8 distributes the sums over the products. The recursive sum $f_i(j)$ is the part of equation 8 to the right of x_i assuming $x_i = j$, and $b_i(j)$ is the same to the left of x_i. The weight matrix entry $W_{i,j}$ is the marginal probability that $S_j = i$.

$$g_i(\mathbf{c}_i) = g_i(x_i, x_{i+1}) \qquad (4)$$

$$= \exp(-\Delta E(i, x_i, x_{i+1})) \qquad (5)$$

$$P(S) = \alpha^{-1} \prod_{i=1}^{L-1} g_i(S_i, S_{i+1}) \qquad (6)$$

$$\alpha = \sum_{x_1,\ldots,x_L} \prod_i g_i(\mathbf{c}_i) \qquad (7)$$

$$= \sum_{x_1} \sum_{x_2} g_1(x_1, x_2) \sum_{x_3} g_2(x_2, x_3) \ldots$$

$$\ldots \sum_{x_L} g_{L-1}(x_{L-1}, x_L) \qquad (8)$$

$$= \sum_{x_1} f_1(x_1) = \sum_{x_L} b_L(x_L) \qquad (9)$$

$$W_{i,j} = b_j(i) f_j(i)/\alpha \qquad (10)$$

$$f_i(x) = \begin{cases} \sum_{x_{i+1}} g_i(x, x_{i+1}) f_{i+1}(x_{i+1}), \\ \quad \text{if } i < L \\ 1, \text{ otherwise} \end{cases} \qquad (11)$$

$$b_i(x) = \begin{cases} \sum_{x_{i-1}} b_{i-1}(x_{i-1}) g_{i-1}(x_{i-1}, x), \\ \quad \text{if } i > 1, \\ 1 \text{ otherwise} \end{cases} \qquad (12)$$

Sequence sets and scoring

All sequences used here are of length $L = 34$.

TEST SET (one set): A set of 60 IHF binding sites was assembled from the literature. Of these, 28 were strong or medium binders and of good measurement quality.

SMALL SETS OF SHORT EXACT STRINGS (two sets): A set of 10 000 random sequences (GC content 50.8%, as in the *E. coli* genome) was generated and sorted by ΔE_{total}; its lowest-energy 100 were used as one set. The 100 globally lowest-energy sequences were enumerated and used as the second set.

CONTROL SETS (two sets): One was a new set of 10 000 random sequences. The second was 10 000 sequences randomly selected from *E. coli* intragenic (coding) regions.

SCORING: The score of a sequence against a small set of short exact strings is its average Hamming distance over every short string in the small set. Its score against a weight matrix is the sum over each column of the negative logarithm of the probability corresponding to the base in that position of the sequence. Its score against a consensus regular pattern is the sum over matched disjuncts in the pattern of the negative logarithm of the quantity four divided by the size of the disjunct.

ALGORITHMS

This section gives the algorithms shown in Figure 1. L, GC, and TH are fixed in advance, based on prior knowledge, and not treated as trained parameters. $X \oplus Y$ is X concatenated to Y. Bases A, C, G, T are identified with 1, 2, 3, 4. G(I, J, K) = g_i(J, K) = exp($-\Delta E$(I, J, K)).

Algorithm 1

Generate N random sequences of GC content desired, return the M of lowest ΔE_{total}.

Algorithm 2

Each successive call to `Next_Best_Seq()` returns the globally next-best sequence. Qs is a 4×(L-1) array of initially empty priority queues with smallest value on top. The queues hold state between calls. Queue Qs[i,j] holds those subsequences of length $L + 1 - j$ that begin with base i. The algorithm proceeds by dynamic programming.

```
PROCEDURE Next_Best_Seq (L, Qs) {
 IF (Qs are all empty)
  FOR I FROM 1 TO 4
   FOR J FROM 1 TO 4
    Push(I⊕J,ΔE(L-1,I,J),Qs[I,L-1]);
 Next_Best_Aux (L, Qs, 1);
 I = argmin_{i=1}^{4} Key(Top(Qs[i, 1]));
 Return(Pop(Qs[I, 1]));}
PROCEDURE Next_Best_Aux (L, Qs, K) {
 IF (K < L - 2)
  Next_Best_Aux(L, Qs, K + 1);
 FOR I FROM 1 TO 4 {
  E0 = Key(Top(Qs[I, K + 1]));
  S = Pop(Qs[I, K + 1]);
  IF (S is not NULL)
   FOR J FROM 1 TO 4 {
    E = ΔE(K,J,I) + E0;
    Push(J⊕S, E, Q[J, K]);}}}
```

Algorithm 3

`Joint_Probability()` implements equation 6, `Norm()` implements equation 9, and `Forward()` implements equation 11.

```
PROCEDURE Joint_Probability (S, L) {
 RETURN(exp(-ΔE_total(S))/Norm(L));}
PROCEDURE Norm (L) {
 F = Forward(L);
 Return(∑_{i=1}^{4} F[1, i]);}
PROCEDURE Forward (L) {
 F = new ARRAY [L, 4];
 FOR J FROM 1 TO 4 F[L, J] = 1;
 FOR I FROM L - 1 DOWNTO 1
  FOR J FROM 1 TO 4
    F[I,J] = ∑_{k=1}^{4} G(I,J,k)*F[I+1,k];
 RETURN(F);}
```

Algorithm 4

`Weight_Matrix()` implements equation 10 and `Backward()` implements equation 12.

```
PROCEDURE Weight_Matrix (L) {
 W = new ARRAY [4, L];
 A = Norm(L);
 F = Forward(L);
 B = Backward(L);
 FOR I FROM 1 TO 4
  FOR J FROM 1 TO L
   W[I,J] = .01+.96*B[J,I]*F[J,I]/A;
 RETURN(W);}
PROCEDURE Backward (L) {
 B = new ARRAY [L, 4];
 FOR J FROM 1 TO 4 B[1, J] = 1;
 FOR I FROM 2 TO L
  FOR J FROM 1 TO 4
    B[I,J] = ∑_{k=1}^{4} B[I-1,k]*G(I-1,k,J);
 RETURN(B);}
```

Algorithm 5

$W_{i,j} = (C_{i,j} + 1)/\sum_{k=1}^{4}(C_{k,j} + 1)$, where W is the returned weight matrix and $C_{i,j}$ is the number of counts of base i at position j.

Algorithm 6

Run EMBOSS (Rice *et al.*, 2000).

Algorithm 7

The algorithm collects all letters in W that exceed a specified probability threshold TH.

```
PROCEDURE Reg_Pat (W, TH, L) {
 R = new ARRAY [L];
 FOR I FROM 1 TO L {
  FOR J FROM 1 TO 4
   IF (W[J, I] > TH) Push(J, R[I]);
  IF (R[I] is null) R[I] = "N";
  ELSE R[I] = IUPAC_Code(R[I]);}
 RETURN(R);}
```

RESULTS

Figure 4 shows histograms of deformation energies for the sequences considered. The test set of IHF binding

Table 1. Sequences and ΔE_{total}. Random mean = 215.54, random standard deviation = 21.7 (a) Reference sequence of Figure 2. (b) The lowest-energy 10 of 10 000 random sequences by Algorithm 1. (c) The globally lowest-energy 10 sequences by Algorithm 2. (d) The strongest 10 IHF binding sites.

(a) Reference DNA sequence: ΔE_{total}

```
GCCAAAAAAGCATTGCTTATCAATTTGTTGCACC 155.64
```

(b) Lowest-energy 10 of 10 000 random sequences:

```
CCGGAAACCGGATCATGAGCGAGTCCGGTGGGCG 143.18
CTGGATTAGATTGCCCAACCGAGCCCACTCTACC 152.14
CCCGTGTTGTCGCGCCTGTCAAGCGCACTAGCGA 152.17
TGAGCTTTTGGACGGACTTGCGTAATGTGAGCCG 155.11
ATTGTTAAGATATCAAGATACGACACAGCCTCGG 155.34
TGTGTTAACGCGAGCTGTGTGAATACCTTCGCAA 156.14
AGATCAGCACTTGCACAACCCGGCGCGGCATAGC 157.51
GTCCGGAACGCTGCAGTAATGGGTCTTCTGGGGC 157.78
GTGGTGTTTCCGGGGATTTACATCGGACGACAGC 157.97
GCCGGTAGTATAAGCCCGTGTAGTCTAATATGGC 159.13
```

(c) Globally lowest-energy 10 sequences:

```
CCGAAAAAACCATTGCTTACGAACGCATTGGCCG 114.58
CCGAAAAAACCATTGCTTACGAACGCGTTGGCCG 114.66
CCGAAAAAACCATTGCTTACGAGCGCATTGGCCG 114.77
CCGAAAAAACCATTGCAAACGAACGCATTGGCCG 114.82
CCGAAAAAACCATTGCTTACGAGCGCGTTGGCCG 114.85
CCGAAAAAACCATTGCAAACGAACGCGTTGGCCG 114.90
CCCGCAAAACCATTGCTTACGAACGCATTGGCCG 114.92
CCCGCAAAACCATTGCTTACGAACGCGTTGGCCG 115.00
CCGAAAAAACCATTGCAAACGAGCGCATTGGCCG 115.01
CCGAAAAAACCATTGCAAACGAGCGCGTTGGCCG 115.09
```

(d) Strongest 10 IHF binding sites:

```
ACAATATTTTTCTTTTAAATCAATGGGATAGCGA 199.66
ATTTAAATAACAATAAAAATCATAAACTTAAAAT 191.17
TAGTCGATCGTTAAGCGATTCAGCACCTTACCTC 201.98
CGATAATTACTCATAAAAATCATCATATTAGAAA 177.38
TTAAGAAAATTTATACAAATCAGCAATATACCCA 169.59
AATTAAATACCTTTAAATATCAACAAGTTAAAGT 185.72
AAAAATTAGAAACACATTGAAAACAATACCTTGA 208.02
GATAAAATCCATTTTAATTTCAGTCATTTAATAA 228.21
GATAAGAATATATTAAATATCAGTGAGTTAATAA 184.83
TAATAATTCAAGGTTAAAATCAATAACTTATTCT 218.63
```

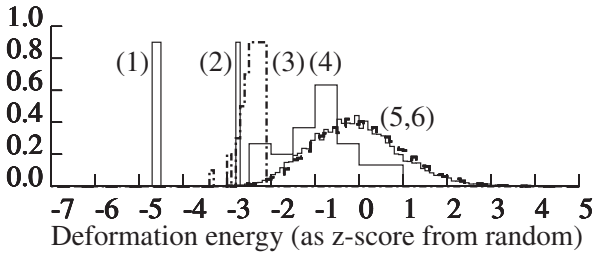


Fig. 4. Histograms of deformation energies for the sequence sets considered. (1) 100 globally lowest, $m = 115$. (2) reference DNA sequence, $m = 156$. (3) 100 lowest random (dash-dot), $m = 165$. (4) test set of 60 IHF binding sites, $m = 195$, $p < 5 \times 10^{-18}$. (5) 10 000 *E. coli* intragenic regions, $m = 213$. (6) 10 000 random sequences (dashed), $m = 215.54$, $s = 21.7$. Sets 5 and 6 are visually indistinguishable. x-axis is z-score from random mean and standard deviation (tick=1σ). y-axis is frequency scaled to unit area under curve (tick=0.2; y clipped at 0.9). m is distribution mean; s is standard deviation; p is significance of difference from random.

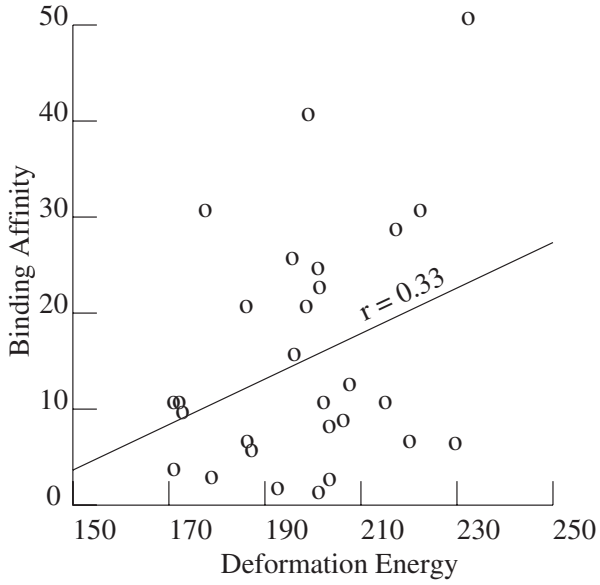
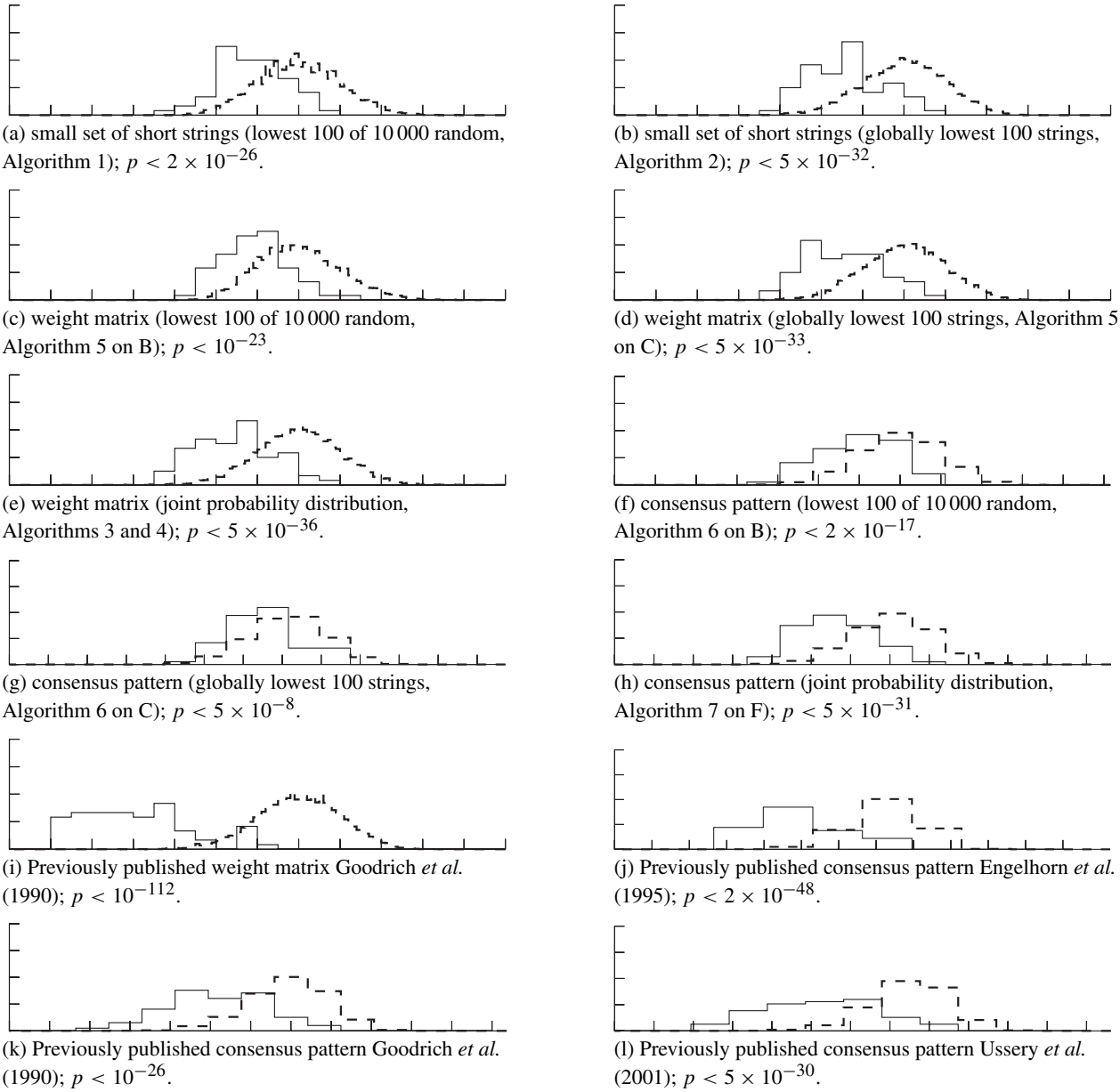


Fig. 5. IHF binding affinity vs. deformation energies. The 28 binding sites included were of medium or strong binding affinity and good measurement quality (Tolleri, 2002).

proteins showed statistically significant separation from random and *E. coli* intra-genic sequences ($p < 5 \times 10^{-18}$). Figure 5 shows the relationship between IHF binding affinity and deformation energy. Only the 28 medium or strong binders of good measurement quality were used (Tolleri, 2002). Deformation energy was positively correlated with binding affinity ($r = 0.33$).

Table 1 shows the lowest-energy 10 out of 10 000 random sequences by Algorithm 1, and the 10 globally lowest-energy sequences by Algorithm 2.

Figure 6 shows each representation's separation of 60 known IHF binding sites from unrelated (random or intragenic) sequences. Every case showed statistically significant separation ($p < 5 \times 10^{-8}$ or better). Generally, weight matrices produced the largest separation and consensus patterns the smallest. Sequence motifs derived from the globally lowest-energy sequences produced greater separation than those from the lowest random sequences. The previously published weight matrix, extracted from IHF binding sites, produced the largest separation.



(a) small set of short strings (lowest 100 of 10 000 random, Algorithm 1); $p < 2 \times 10^{-26}$.

(b) small set of short strings (globally lowest 100 strings, Algorithm 2); $p < 5 \times 10^{-32}$.

(c) weight matrix (lowest 100 of 10 000 random, Algorithm 5 on B); $p < 10^{-23}$.

(d) weight matrix (globally lowest 100 strings, Algorithm 5 on C); $p < 5 \times 10^{-33}$.

(e) weight matrix (joint probability distribution, Algorithms 3 and 4); $p < 5 \times 10^{-36}$.

(f) consensus pattern (lowest 100 of 10 000 random, Algorithm 6 on B); $p < 2 \times 10^{-17}$.

(g) consensus pattern (globally lowest 100 strings, Algorithm 6 on C); $p < 5 \times 10^{-8}$.

(h) consensus pattern (joint probability distribution, Algorithm 7 on F); $p < 5 \times 10^{-31}$.

(i) Previously published weight matrix Goodrich *et al.* (1990); $p < 10^{-112}$.

(j) Previously published consensus pattern Engelhorn *et al.* (1995); $p < 2 \times 10^{-48}$.

(k) Previously published consensus pattern Goodrich *et al.* (1990); $p < 10^{-26}$.

(l) Previously published consensus pattern Ussery *et al.* (2001); $p < 5 \times 10^{-30}$.

Fig. 6. Histograms of distribution separations. Sequence motif scores on test set (solid) and 10 000 random (dashed; *E. coli* intragenic results are similar). Labels and axes are as in Figure 4.

Table 2 shows the correlation between binding affinity for the 28 strong and medium IHF binding sites in the test set and their score against all sequence motifs in Figure 6. The only negative correlations arose from consensus sequences, which performed worst in Figure 6.

Table 3 shows a multiple alignment of consensus regular patterns built by EMBOSS (Rice *et al.*, 2000) from Table 1(B,C) extended to 100 sequences each, and by Algorithm 7 from the joint probability distribution extracted from the Markov network. Some columns, e.g., the A-tract and RTTR sequence motif, are well conserved across all patterns. In contrast, the ATCA motif, including two of the three bases IHF uses for direct recognition, is conserved only across the sequence-derived patterns.

Figure 7 shows a phylogenetic tree built by PHYLIP (Felsenstein, 1993) from the 10 strongest binding IHF sites, plus the 10 lowest-energy and 10 highest-energy of 10 000 random sequences. There is some mixing between IHF binding sites (IB2, 4, 6) and low-energy random sequences (RL3, 4), but high-energy random sequences are separated from IHF binding sites by at least four links (IB7 to RH5, 7).

Table 2. Correlation of sequence representation scores with IHF binding affinity. Binding affinities correspond to Figure 5. Sets correspond to Figure 6 as indicated; e.g., (A) corresponds to Figure 6(A).

(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)	(I)	(J)	(K)	(L)
0.20	0.06	0.25	0.06	0.12	0.06	−0.11	0.06	0.38	−0.07	0.07	0.05

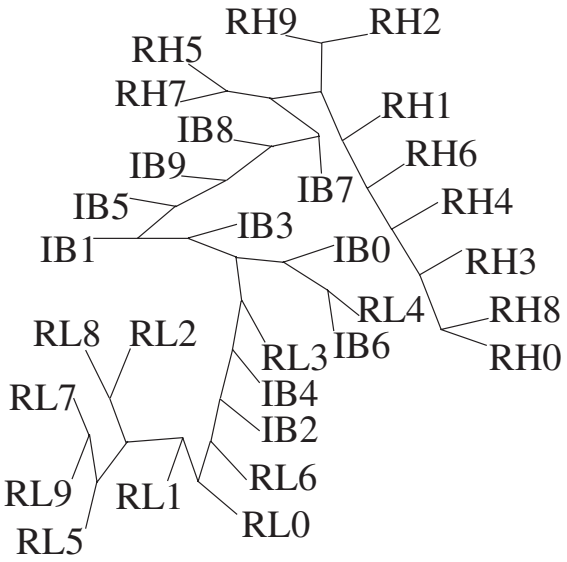


Fig. 7. Unrooted phylogenetic tree (Felsenstein, 1993) of 10 lowest-energy random sequences (RL), 10 strongest binding IHF binding sites (IB), and 10 highest-energy random sequences (RH).

Table 3. Consensus regular patterns. (A) Built by EMBOSS (Rice *et al.*, 2000) from lowest-energy 100 out of 10 000 random sequences of Algorithm 1. (B) Built by EMBOSS from globally lowest-energy 100 sequences of Algorithm 2. (C) Built by Algorithm 7 from the joint probability distribution extracted from the Markov network. (D) The reference DNA sequence of Figure 2 (Rice *et al.*, 1996). (E) Built by EMBOSS from the test set of 60 IHF binding sites. (F) Published consensus pattern (Goodrich *et al.*, 1990). (G) Published consensus pattern (reverse complement) (Ussery *et al.*, 2001). (H) Published consensus pattern (reverse complement) (Engelhorn *et al.*, 1995). IUPAC: R={A,G}, S={C,G}, W={A,T}, Y={C,T}, N=any. "*"=base used in direct recognition. "#"=proline intercalation site.

```
GGAGTAAATCCGTCGGATCCGAGCGCGTTATCGN (A)
CCGAAAAAACCACGGCTTATTCGCGCATTGGCCG (B)
CCSRAAAAACCATTGCWWACGARCGCRTTGGCCG (C)
GCCAAAAAAGCATTGCTTATCAATTTGTTGCACC (D)
TTATAAAAATTTTTAAAATCAATAAGTTACAAA (E)
   ATNAWNTYNAWTWAWWWCAANAAGTTR     (F)
             WAAATCAANAAGTTR       (G)
              WATCAANNNNTTR        (H)
          #      **#     *
```

DISCUSSION AND CONCLUSIONS

Sequence motif representations of a crystal structure partially separated known binding sequences from random and intragenic sequences. Most exhibited positive correlation with binding affinity. Conservation was observed in some columns of a multiple alignment. A phylogenetic tree separated high-energy random sequences from, but partially mixed low-energy sequences with, known IHF binding sites. In summary: Carry-over occurred from a DNA crystal structure to the purely sequence-based representations used by common popular pattern-based inference methods.

Not all known IHF binding sites have a low deformation energy. It is possible that not all IHF/DNA complexes have the same conformation, that the direct recognition energetics contribute more than the indirect recogntion energetics, or that other indirect effects are operative. For example, IHF binding sites often exhibit an A-tract or AT-rich region, which may have an associated hydration spine that contributes additional indirect recognition.

Implications of these results are: (a) Some current sequence motifs probably implicitly recognize structural properties of DNA. Structure can be important even when sequence is relatively conserved. (b) Higher-order sequence models of structure-based motifs should be useful, since structure-based signals involve two or more bases. (c) DNA deformation energy is a useful computational tool. (d) Structure might usefully inform sequence motif methods, e.g., the joint probability distribution might produce better prior distributions, initial weights, or starting positions, than would random values.

Deformation energy alone is not diagnostic for binding sites, but this would be expected because IHF is known to bind using both direct and indirect recognition mechanisms. It does appear to explain some aspects of IHF binding, as shown by the partial separation of IHF binding sites from random sequences and the positive correlation of binding affinity with deformation energy. The use of structure motifs in addition to sequence motifs may improve binding site recognition, particularly when sequence is not well conserved in some parts of the binding site.

ACKNOWLEDGEMENTS

Phoebe Rice kindly supplied the smoothed atomic model of bound IHF. Thanks to Wilma Olson and Victor Zhurkin for discussion about their potential. Thanks to Craig Benham, Steve Hampson, the IBaM seminar, and the blind referees, for helpful comments on the paper. Domino model made using X3DNA by X.-J.Lu. Figures 2 and 3 visualization from RASMOL by R.Sayle. Figure 7 visualization redrawn from TREEVIEW by R.Page.

REFERENCES

Baldi,P.F., Chauvin,Y., Brunak,S., Gorodkin,J. and Pederson,A.G. (1998) Computational applications of DNA structural scales. In Glasgow,J. *et al.*, (eds), *Intelligent Systems for Molecular Biology*. AAAI Press, Menlo Park, CA, pp. 35–42.

Baldi,P.F. and Lathrop,R.H. (2001) DNA structure, protein-DNA interactions, and DNA-protein expression. In Altman,R. *et al.*, (eds), *Pacific Symposium on Biocomputing*. World Scientific, Singapore, pp. 101–102.

Benos,P.V., Lapedes,A.S., Fields,D.S. and Stormo,G.D. (2001) SAMIE: Statistical algorithm for modeling interaction energies. In Altman,R. *et al.*, (eds), *Pacific Symposium on Biocomputing*. World Scientific, Singapore, pp. 115–126.

Calladine,C.R. and Drew,H.R. (1992) *Understanding DNA: The Molecule and How it Works*. Academic Press, London.

Chen,S., Gunasekera,A., Zhang,X., Kunkel,T.A., Ebright,R.H. and Berman,H.M. (2001) 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.

Engelhorn,M., Boccard,F., Murtin,C., Prentki,P. and Geiselmann,J. (1995) *In vivo* interaction of the *E. coli* integration host factor with its specific binding sites. *Nucleic Acids Res.*, **23**, 2959–2965.

Felsenstein,J. (1993) PHYLIP (Phylogeny inference package) version 3.5c.. Distributed by the author., Department of Genetics, University of Washington, Seattle.

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

Hatfield,G.W. and Benham,C.J. (2002) DNA topology-mediated control of global gene expression in *E. coli*. *Ann. Rev. Genet.* (in press).

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

Lawrence,C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.F. and Wootton,J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.

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

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

Mandel-Gutfreund,Y., Baron,A. and Margalit,H. (2001) A structure-based approach for prediction of protein binding sites ingene upstream regions. In Altman,R. *et al.*, (eds), *Pacific Symposium on Biocomputing*. World Scientific, Singapore, pp. 139–150.

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**, 11 163–11 168.

Pearl,J. (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA.

Rice,P.A. (1997) Making DNA do a U-turn: IHF and related proteins. *Curr. Opin. Struct. Biol.*, **7**, 86–93.

Rice,P.A., Yang,S.W., Mizuuchi,K. and Nash,H.A. (1996) Crystal structure of an IHF-DNA complex: A protein-induced DNA U-turn. *Cell*, **87**, 1295–1306.

Rice,P.A., Longden,I. and Bleasby,A. (2000) EMBOSS: The European molecular biology open software suite. *Trends Genet.*, **16**, 276–277.

Sarai,A., Selvaraj,S., Gromiha,M.M., Siebers,J.G., Prabakaran,P. and Kono,H. (2001) Target prediction of transcription factors: Refinement of structure-based method. In Matsuda,H. *et al.*, (eds), *Genome Informatics 2001*, Genome Informatics Series 12, Universal Academy Press, Tokyo, pp. 384–385.

Schneider,T.D. (2001) Strong minor groove base conservation in sequence logos implies DNA distortion or base flipping during replication and transcription initiation. *Nucleic Acids Res.*, **29**, 4881–4891.

Tolleri,L. (2002) An interdisciplinary approach employing computational, biochemical, and genomic methods to examine the effects of chromosome structure on the regulation of gene expression, PhD Thesis, Università degli Studi di Pavia e Firenze.

Ussery,D., Larsen,T.S., Wilkes,K.T., Friis,C., Worning,P., Krogh,A. and Brunak,S. (2001) Genome organisation and chromatin structure in *E. coli*. *Biochimie*, **83**, 201–212.

van Helden,J., André,B. and Collado-Vides,J. (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.*, **281**, 827–842.

Vilo,J., Brazma,A., Jonassen,I., Robinson,A. and Ukkonen,E. (2000) Mining for putative regulatory elements in the yeast genome using gene expression data. In Altman,R. *et al.*, (eds), *Intelligent Systems for Molecular Biology*. AAAI Press, Menlo Park, CA, pp. 384–394.