BIOINFORMATICS

Vol. 17 no. 7 2001
Pages 622–633

Conformational model for binding site recognition by the E.coli *MetJ transcription factor*

*Rongxiang Liu, Thomas W. Blackwell and David J. States**

Center for Computational Biology and Department of Genetics, Washington University School of Medicine, 700 S. Euclid Ave, St Louis, MO 63110, USA

Received on June 1, 2000; revised on February 27, 2001; accepted on March 4, 2001

ABSTRACT

Motivation: Current methods for identifying sequence specific binding sites in DNA sequence using position specific weight matrices are limited in both sensitivity and specificity. Double strand DNA helix exhibits sequence dependent variations in conformation. Interactions between macromolecules result from complementarity of the two tertiary structures. We hypothesize that this conformational variation plays a role in transcription factor binding site recognition, and that the use of this structure information will improve the predictive power of transcription factor binding site models.

Results: Conformation models for the sequence dependence of DNA helix distortion have been developed. Using our conformational models, we defined a tertiary structure template for the *met* operon repressor MetJ binding site. Both naturally occurring sites and precursor binding sites identified through *in vitro* selection were used as the basis for template definition. The conformational model appears to recognize features of protein binding sites that are distinct from the features recognized by primary sequence based profiles. Combining the conformational model and primary sequence profile yields a hybrid model with improved discriminatory power compared with either the conformational model or sequence profile alone.

Using our hybrid model, we searched the *E.coli* genome. We are able to identify the documented MetJ sites in the promoter regions of *metA*, *metB*, *metC*, *metR* and *metF*. In addition, we find several novel loci with characteristics suggesting that they are functional MetJ repressor binding sites. Novel MetJ binding sites are found upstream of the *metK* gene, as well as upstream of a gene, *abc*, a gene that encodes for a component of a multifunction transporter which may transport amino acids across the membrane. The false positive rate is significantly lower than the sequence profile method.

Availability: The programs of implementation of this algorithm are available upon request. The list of crystal structures used for compiling the mean base step parameters of DNA is available by anonymous ftp at http://stateslab.wustl.edu/pub/helix/StructureList.

Contact: states@ccb.wustl.edu

*To whom correspondence should be addressed.

INTRODUCTION

Regulation of gene expression by sequence specific protein binding to DNA is a central paradigm of molecular biology (Jacob and Monod, 1961; Gilbert and Muller-Hill, 1966, 1967). Sequence specific binding of proteins to DNA guides gene regulation, the control of replication, and many other fundamental biological processes. DNA–protein binding, like any high affinity ligand interaction (e.g. protein–protein interaction), results from complementarity between the three-dimensional structure of the ligand and the binding site. In protein–protein recognition, both sequence (atom contacts) and tertiary structure (molecule docking) of proteins are essential to the specificity of binding. Similarly, we hypothesize that in the DNA site recognition, both primary sequence and three-dimensional conformation of DNA contribute to specifying the binding site.

With the first crystal structure of DNA decamer solved (Drew *et al.*, 1981; Prive *et al.*, 1987), a network of ordered water molecules was observed in both major and minor grooves, and the helix itself was seen to be distorted from the prototypic B-DNA conformation. Both of these features have been seen in subsequent DNA tertiary structure studies, and it has been suggested that the water may mediate the indirect readout of DNA by its recognition proteins (Otwinowski *et al.*, 1988; Wilson *et al.*, 1995; Shakked *et al.*, 1994). Tertiary structure studies of DNA/protein complexes have demonstrated that proteins interact intimately with the sugar–phosphate backbone of the DNA as well as the nucleotide bases (Olson *et al.*, 1998). Sigler has strongly argued that the conformation of the DNA and associated water molecules are used as stereo specific recognition elements by the DNA oligos in TrpR binding site recognition (Luisi and Sigler, 1990). It is further suggested that these phenomena are a general feature of sequence specific DNA protein recognition, and proteins may indirectly read the sequence

of a DNA binding site through its three-dimensional conformation.

With the increasing number of DNA crystal structures determined, general patterns of sequence specific helix distortion have been recognized (Dickerson, 1992; Liu *et al.*, 1998; Lu and Olson, 1999; Allemann and Egli, 1997). In addition, electrophoresis studies have demonstrated sequence dependent helix variations in DNA fragment mobility attribute to helix bending (Diekmann, 1989). These sequence specific structural variations of the DNA helix can be used in the protein recognition of DNA sites, and a wide range of evidence suggests that DNA conformation is an important factor in protein recognizing specific sites on DNA (Harrington, 1992). We have previously shown (Liu *et al.*, 1998) that it is possible to build models for the sequence dependent conformational preferences of double stranded DNA and that these models are able to describe eukaryotic transcription factor binding sites. Here we examine the sequence specific DNA binding interaction of MetJ that repressed gene expression in the bacteria *E.coli*.

The Cambridge Workshop defined a standard set of six parameters to specify a dinucleotide step structure (Dickerson *et al.*, 1989). Several models for sequence dependent helix distortion have been developed (Kopka *et al.*, 1994; Yanagi *et al.*, 1991; Tung and Harvey, 1986). In addition, proteins binding to dsDNA make extensive contacts to the phosphate backbone and interpolate into either the major or minor groove of the DNA helix. Groove width of helix might therefore be an important characteristic used by proteins for the recognition of specific sequences. We have incorporated groove opening of dinucleotide steps into our conformation model for helix structure. Although tables describing DNA helix local structure parameters have been published previously (Bolshoy *et al.*, 1991; Cacchione *et al.*, 1989; Bansal *et al.*, 1995; Olson *et al.*, 1998), a considerable volume of new data is now available. We have used the Nucleic Acid Database (Berman *et al.*, 1992) to retrieve DNA structure data directly and recompiled the mean geometry table of B-DNA.

There is an extensive literature and numerous software tools have been developed to search for transcription factor binding sites on DNA (Quandt *et al.*, 1995; Schug and Overton, 1997; Istrail *et al.*, 1998). However, the sensitivity and specificity of these tools remain less optimal. These methods focus on the consensus sequence of these sites. From data in the literature, very often the transcription factor binding sites are not well conserved in primary sequence (Ghosh, 1992).

To study the role of DNA structure in transcription factor recognition, we choose the well-studied *met* operon repressor MetJ as our model. MetJ is a transcription factor in *E.coli* with a beta ribbon structure (Somers and Phillips, 1992). A crystal structure of transcription factor with its cognate DNA has been solved. Further, *in vitro* binding sites selection studies (SELEX) for this transcription factor are available as well (He *et al.*, 1996).

MetJ is representative of an evolutionarily ancient DNA-binding fold (Aravind and Koonin, 1999; Suzuki, 1995; Raumann *et al.*, 1994). Naturally occurring operators differ from the consensus sequence to a greater extent as the number of metboxes increases. MetJ, while accommodating this sequence variation in natural operators, is very sensitive to particular base changes, even where bases are not directly contacted in the crystal structure of a complex formed between the repressor and consensus operator. The structural determinants of MetJ binding to DNA have been analyzed on the basis of x-ray crystallographic data. It is shown that the DNA binding geometry of the beta-sheet in MetJ and arc can be understood in terms of (i) close fit of the two surfaces and (ii) matching of residue and base positions. Recently, x-ray crystallographic studies have been performed on indications of MetJ and its cognate operator sequence altering bases involved and indirect sequence readout. The overall structure of the mutant complex is very similar to the wild-type complex, but there are small variations in sugar–phosphate backbone conformation and direct contacts to the DNA bases (Garvie and Phillips, 2000). The analysis presented here addresses primarily shape fit as a recognition mechanism.

Here, we examine the helical structure of DNA binding sites of MetJ. Comparisons are made with the DNA helix structure in DNA/protein co-crystal. Substantial similarity is observed between the helical structure of MetJ binding sites and the DNA structure in the co-crystal. For predicting new MetJ binding sites in *E.coli* genome, we find that using either tertiary structure or primary sequence alone is not sufficient to identify the protein binding sites. Combining both structural information and primary sequence information yields a hybrid model that performs much better than using either method alone. We discuss the role of tertiary structure information in site recognition of MetJ. Searching the promoter regions of the *E.coli* genome with our hybrid model, all the documented sites are identified. In addition, three previously undocumented binding sites of MetJ in the promoter region of genes involved in biosynthesis of methionine respectively are identified. Putative binding sites are also found in the promoter region of amino acid transporters.

METHODS

Minor groove opening

In essentially all protein–DNA co-crystals, parts of the protein interpolate into the grooves of the DNA double helix. Both major and minor groove opening widths

Table 1. The parameters used to describe DNA helix structure

Base step	No. of instances	Tilt	Tilt SD	Roll	Roll SD	Twist	Twist SD	Slide	Slide SD	Groove	Groove SD
AA	125	−1.0	3.24	−0.1	3.65	35.9	3.16	−0.64	0.29	11.93	0.95
AC	20	−0.5	4.94	−0.8	5.68	33.4	3.87	−0.55	0.38	13.44	0.99
AG	31	−2.3	3.40	4.3	3.03	30.9	4.72	−0.21	0.58	13.25	1.16
AT	52	0	2.32	−1.2	3.05	32.8	2.67	−0.94	0.21	12.22	0.93
CA	27	0.2	3.15	5.6	3.37	30.9	4.58	0.48	0.50	13.85	0.91
CC	42	2.9	3.52	5.0	4.44	32.6	4.54	0.12	0.52	13.91	0.59
CG	162	0	4.23	2.9	5.63	35.1	4.42	0.09	0.46	14.56	0.80
GA	71	−1.4	2.86	0.5	3.71	39.3	3.02	−0.56	0.41	12.20	1.20
GC	110	0	4.61	−6.0	5.44	37.6	3.80	−0.20	0.38	14.24	1.11
TA	13	0	3.90	1.9	4.20	40.3	5.17	−0.04	0.86	12.51	1.37

The structure of a dinucleotide step is described by these five parameters. Tilt, roll, and twist are angles in degree; slide and groove opening are distance in angstroms. Here, SD is the standard deviation. The number of instances observed for a base step includes the instance of the step plus the instance for its complementary step. Note that the dyad symmetry of the DNA helix constrains the mean tilt of self-complementary base steps to be zero. The table is simplified by omitting complementary steps.

may play a role in binding site recognition. Because these two parameters are closely interdependent, we have incorporated only the minor groove opening in our calculations. For the purposes of these calculations, we define minor groove opening using a set of fixed interatomic distances. The groove opening we used in this study is the contribution of a base step to the groove opening of the DNA helix. It is not equivalent to minor groove width, which is the shortest distance between the two strands. Referring to the nomenclature of el Hassan and Calladine (1998, Figure A1), the minor groove opening at base step i is measured as the distance from phosphate P_{i+1} on the sense strand to phosphate p_{i-1} on the antisense strand. This gives a simple direct measure that is symmetric about the i base step (Figure 1). As shown in Figure 1, this vector is nearly, but not always, perpendicular to the channel of the minor groove.

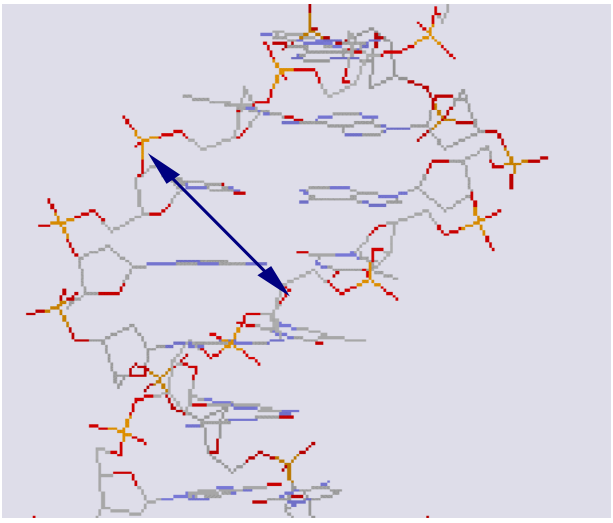

Fig. 1. Definition of the minor groove opening. The groove opening is described using the distance between two phosphates one dinucleotide step away on each strand. The minor groove opening at base step i is measured as the distance from phosphate P_{i+1} on the sense strand to phosphate p_{i-1} on the antisense strand. This gives a simple direct measure that is symmetric about base step i.

Parameter compilation

DNA oligonucleotide crystal structures were retrieved from Nucleic Acid Database (Berman *et al.*, 1992). Since the DNA crystal structures in NDB have increased recently, only structures with high resolution are used in this study. 96 structures lacking bound ligands or sequence mismatches and with resolution not worse than 2.5 Å were retrieved from NDB for analysis. The base step parameters (tilt, roll, twist, and slide) were calculated using curve 5.1 local frame (Lavery and Sklenar, 1988) for the 96 crystal structures. After excluding the base steps with modifications, mean value and standard deviation of each of the five parameters were calculated for the 96 crystal structures (Table 1). A list of these structures is available by anonymous ftp at http://stateslab.wustl.edu/pub/helix/StructureList.

Independence of conformational parameters

Because the DNA helix is a complex structure governed by the interplay of numerous physical–chemical interactions, we tested explicitly whether the five parameters that we have selected were, in fact, independent of one another. The independence of each parameter was evaluated by calculating the correlation coefficient for each pair of parameters. The results are shown in Table 2 below. With two exceptions, only minor correlations are observed

Table 2. The correlation coefficients (R) of each parameter with other parameters

R	Twist	Tilt	Roll	Slide	Groove
Twist	1.000	−0.010	−0.559	0.121	−0.266
Tilt	−0.010	1.000	0.165	−0.148	−0.055
Roll	−0.559	0.165	1.000	0.103	0.167
Slide	0.121	−0.148	0.103	1.000	0.617
Groove	−0.266	−0.055	0.167	0.617	1.000

With two exceptions, only minor correlations are observed between these parameters (R is around 0.1 to 0.2). Roll and twist are anti-correlated with R of −0.559, while slide and minor groove opening are positively correlated with R of 0.617. Since the correlation among pairs of parameters is not strong, we assume for the purposes of subsequent analysis that each of the five parameters is independent.

between these parameters (R is around 0.1–0.2). Roll and twist are anti-correlated with R of −0.559, while slide and minor groove opening are positively correlated with R of 0.617. Since the correlation among pairs of parameters is not strong, we assume for the purposes of subsequent analysis that each of the five parameters is independent.

Comparing DNA structures

The distribution of the conformational parameters is consistent with a Gaussian distribution (Liu *et al.*, 1998, data not shown). We approximate that the probability P of a parameter for one step adopting the value of the same parameter observed in another step is

$$P = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{1}{2}\left(\frac{\mu_1-\mu_2}{\sigma}\right)^2},$$

where $\sigma = (\sigma_1 + \sigma_2)/2$; μ, σ are the mean and standard deviation of the parameters. The overall probability of a dinucleotide step adopting the conformation of another dinucleotide is the joint probability of each of the five parameters adopting the value of the other step. Under the assumption of independence between the parameters, the joint probability is just the product of the probability of each of the five parameters shifts individually.

Combining conformational and primary sequence models

MatInspector, the searching tool used in Transcription Factor Database, works well comparing with other methods (Quandt *et al.*, 1995), and this method is in wide use by the research community. In this study, primary sequence profile search for transcription factor binding sites was performed using MatInspector. In the following calculations, the sequence profile model and tertiary structure model are both trained on SELEX data so that the models are independent of test set of MetJ binding sites.

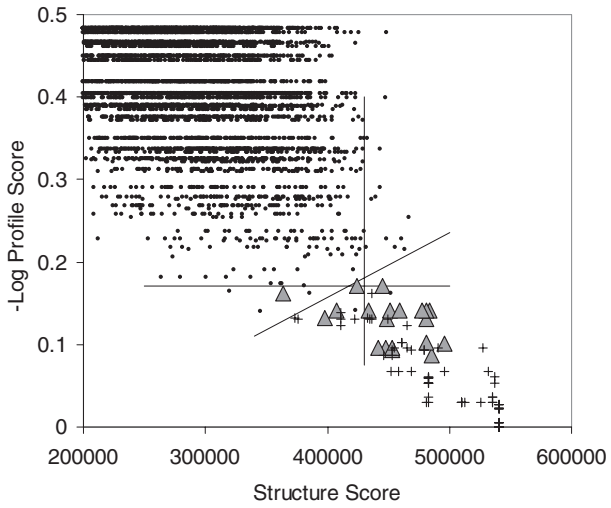

Fig. 2. The scatter plot of structure scores and minus log profile scores for discriminating true binding sites with the MetJ from random sequence fragments. The conformational model was derived from SELEX selected binding sites. The x-axis is the structure similarity score. The y-axis is negative logarithm of the sequence similarity score. In the plot, circles are random sequence fragments, plus signs are SELEX selected binding sites and triangles represent biological binding sites. For the best separation of transcription factor binding sites and random sequence fragments, a diagonal cut discriminates better than either the primary sequence score (horizontal line at 0.17) or conformational score (vertical line at 430 000) alone. This indicates that combining structure information with primary sequence information can improve our ability to recognize functional binding sites.

The structural approach described above compares only the tertiary structure similarity of two DNA sequences. The recognition specific sequence sites by DNA-binding proteins may involve both specific bases (sequence) and three-dimensional conformation of DNA. By combining the primary sequence information and tertiary structure information, we improve our ability to identify transcription factor binding sites. We combined these primary sequence scores with our structural similarity model using a linear discriminant based on manual fitting of a line separating true sites and false sites in the profile–structure scatter plot (Figure 2).

Target regions of the *E.coli* genome

The regions which we searched were extracted from the GenBank *E.coli* whole genome sequence in Fasta format (ecoli.fna) using coding region coordinates given in the GenBank protein table (ecoli.ptt). Both files are dated November 18, 1998. The corresponding GenBank flat file identifies the annotation as ‘version M54’. Each region between two protein coding genes was extended

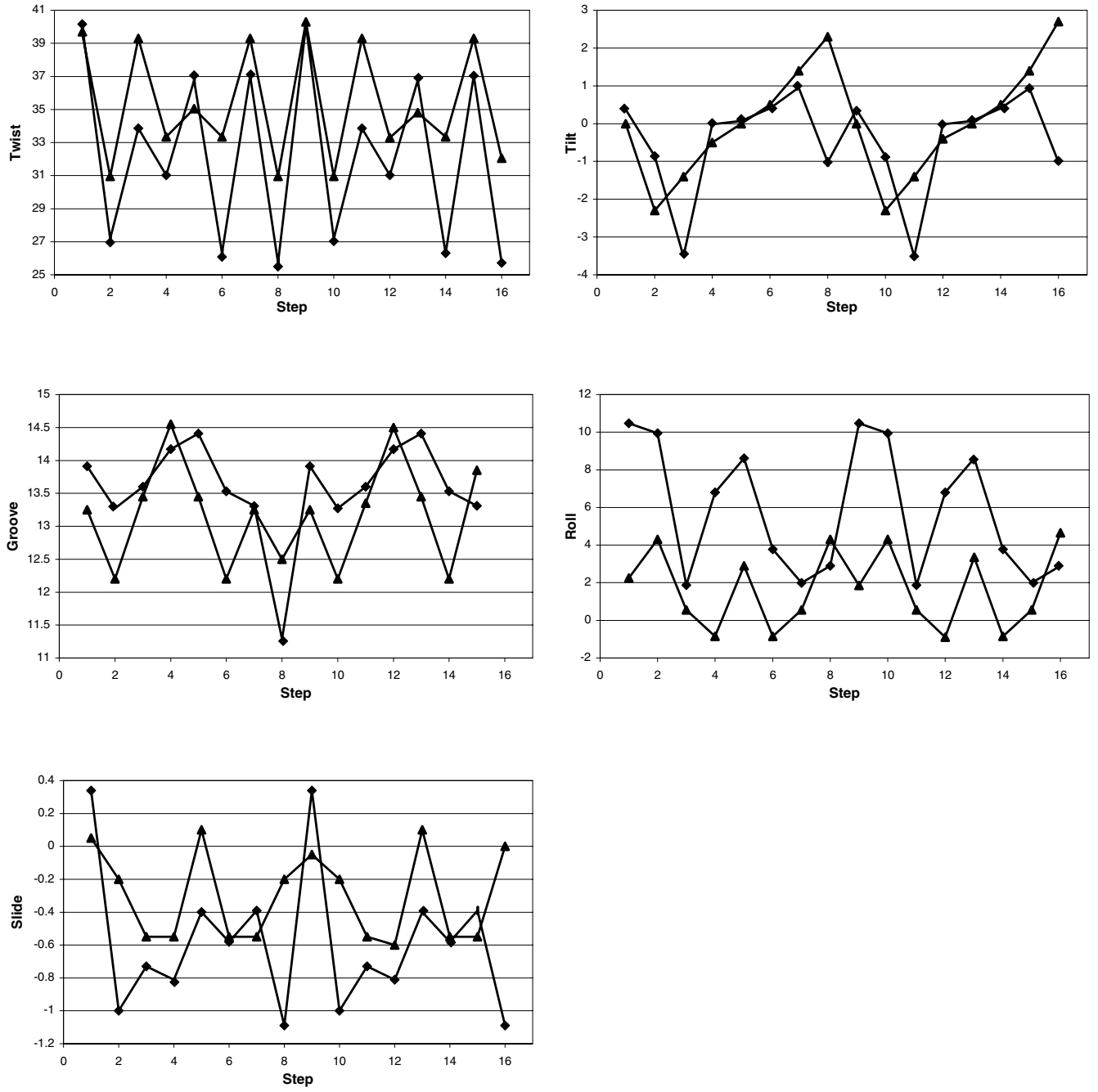

Fig. 3. The comparison of common structure of DNA found among transcription factor binding sites and structure of DNA in the protein/DNA crystal. The diamond lines show the conformation observed in the crystal structure; triangle lines are the mean structure observed by applying our conformational model to known binding sites for MetJ. The twist and tilt parameters exhibit the greatest variations and have greatest similarity. The groove and slide have strong similarity as well. Other two parameters (rise and shift) have least sequence dependent structural variations. They are not included in our analysis and not shown in the figure. The consensus sequence of the MetJ binding sites is TAGACGTCTAGACGTCT.

by 100 nucleotides on either end; transcript termination regions between two convergently transcribed genes were excluded from the search. Functional RNA genes are omitted from ecoli.ptt. Consequently they were included in our search, unless the neighboring protein coding genes are convergently transcribed.

RESULTS

Comparison of sequence derived conformational models with crystallographic structure

The DNA conformation seen in the bound complex of MetJ with its cognate recognition site is distorted (Somers and Phillips, 1992). The energetics of binding to the cognate recognition site relative to the energetics of binding to non-specific DNA determines the sequence specificity of DNA binding by transcription factors. There is an energetic cost associated with distortion of the DNA helix, and sites that are easy to distort into the final conformation (crystal) should be preferred energetically. To assess whether the conformation of known DNA binding sites for MetJ is close to the DNA conformation found in the MetJ/DNA complex, we compared the mean of the predicted structural parameters derived from known MetJ binding sites with DNA structure seen in the crystal of MetJ/DNA complex. As demonstrated in Figure 3, the preferred helix conformation for MetJ sites is substantially similar to the bound conformation.

Comparisons between the mean structure of MetJ known binding sites and the structure of DNA in protein/DNA complexes were done for each of the five parameters. The similarity between the predicted mean structure and the observed conformation in the complex is obvious for each of the five parameters. Since twist and tilt are the two parameters with the largest variations, they encode more information than do the other parameters (Table 1). The similarity of the preferred DNA structure for MetJ binding sequences to the DNA conformation in the MetJ/DNA crystal suggests that this repressor makes use of helical distortion in recognition of its specific binding sites. The roll has very similar pattern between crystal structure and mean structure of MetJ binding sites, though the absolute values are smaller in free DNA. Since the protein generally bend DNA helix after bound. Bending of DNA results in dramatically increasing roll. So the difference in roll is not unexpected.

Correlation of tertiary structure and primary sequence similarity

We have previously demonstrated sequence dependent biases in DNA conformation (Liu *et al.*, 1998), but comparing the tertiary structure similarity is not simply another way of comparing primary sequence similarity. To evaluate the independence of tertiary structure information and primary sequence information, the correlation coefficient of structure similarity scores versus profile similarity scores was computed for 10 000 random sequence fragments. The correlation coefficient between the tertiary structure score and the primary sequence score was 0.46 for MetJ. This confirms that the conformational structure score does not measure features of the binding site that are apparent in the primary sequence weight matrix. This is further illustrated graphically in Figure 2.

Discriminating functional sites

We next assessed the ability of our conformational model and the primary sequence profiles to recognize functional repressor binding sites. Both binding sites upstream of five documented MetJ regulated genes (identified by MatInspector with score >0.675) and sites derived from *in vitro* selection experiments (He *et al.*, 1996) were examined. The primary sequence profile scores (of MatInspector) are calculated as a product over sites while our conformational score is calculated as a sum over log likelihoods. To convert these two scores to a comparable functional form, the logarithm of the profile score was used so that both scores were sums over independent sites and could be analyzed with normal statistics. The scatter plot for the distribution of scores calculated with the MetJ model of both functional sites and random sequences are shown (Figure 2). As the figure demonstrates, a linear discriminant combining both structure and sequence information (diagonal cut) improves our ability to discriminate functional sites from random sequence based on either model alone (horizontal and vertical cuts). The scores of biological binding sites fall between the random sequence fragments and sites selected by SELEX, which makes sense since biological systems usually do not select tightest binding sites.

ROC curves

The scatter plot (Figure 2) demonstrates that combining sequence and structure scores improves discrimination compared to using either method alone at a single score threshold. Receiver Operating Characteristic (ROC) curves provide another general method for assessing the performance of discriminant functions (Swets, 1988). Here we use ROC curve analysis to demonstrate that improved discrimination is achieved over a range of thresholds. ROC curves were calculated for the profile score, structure score, or combined score (Figure 4). 20 oligonucleotides with MatInspector score >0.675 from the intergenic regions upstream of the five documented MetJ regulated genes were treated as true positives. Randomly generated oligomers were assumed negative.

The ROC curve for the combined score has a higher rate of true positive detection over a range of false positive rates when compared with curves for either the primary sequence model or conformation model alone. The ROC curve analysis for MetJ binding sites supports the conclusion that combining tertiary structure and primary sequence profile scores is better than either method alone. Another popular method of searching transcription factor binding pattern is using regular expression. The pattern for MetJ in TFD (Ghosh, 1992) is AGACRTCYAGACGTMT. This pattern matches to only one of the 13 MetJ binding

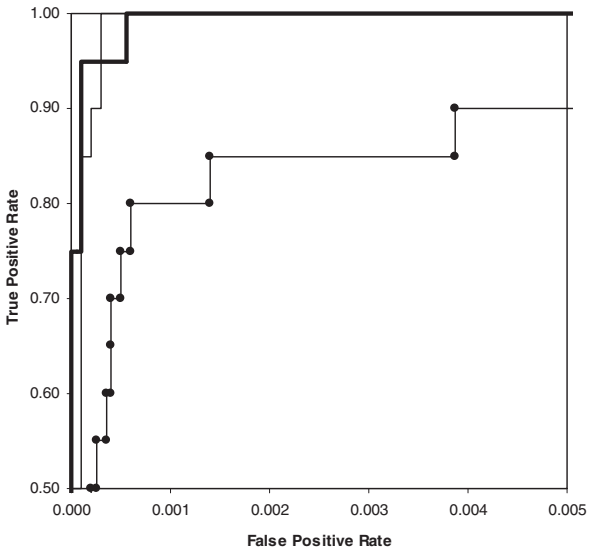

Fig. 4. The ROC curves of searching transcription factor binding sites. The heavy line represents the method of combining both sequence and structure information; the light line represents using only sequence information (matrix similarity); the line with solid circles represents using structure information only (structure similarity). The combined model has better discriminant power than either sequence profile or structure model alone.

Table 3. Function of genes with predicted/documented MetJ binding sites

Combined model	Sequence model	Comments
metB, *metF*, *metA*, *metC*, *metR*	*metB*, *metF*, *metA*, *metC*, *metR*	Documented MetJ binding sites, methionine biosynthesis and regulatory proteins
metK		Converts methionine to *S*-adenosylmethionine
abc	*abc*	Contains one ATP-binding domain, possible transporter
ykfD	*ykfD*	Putative amino acid transporter
ahpC, dmsB, atpA, hemG, mgtA	*dcd, emrY, rfaK, atpA, hemG, rrfD*	Other genes with known function (see text)
yafD, ybdH, ycbK, yhaQ	*yacH, yafU, ybdL, ybhB, ycbK, yccD, ycjF, yeeU yfiA, yqgD, yigG*	Open reading frames without known functions

Both combined model and sequence model can identify all the known binding sites for MetJ. But the sequence model has 14 more hits in the unrelated gene category and unknown gene category. The combined model likely has much lower false positive rate. The searching results using conformation model alone are not shown in this table because the number of hits is too large for the table.

sites with no false matches and to none of the 10 000 random oligos. It has only one data point (0, 0.08). If we plotted it on the ROC curve, it would be in the lower left corner on the *y*-axis. The regular expression method is probably not very useful because it only predicts very few of the true sites.

A third measure of discrimination performance is the *Z*-score (the number of standard deviations from the mean score) for true positives. Using the combined scoring method, the *Z*-score increases by 0.5–0.9 standard deviations for documented MetJ binding sites. As is shown in Figure 2, some of the true MetJ binding sites would be missed by using either sequence similarity scores or structure similarity scores only. The results of our analysis show that structure information is used by the DNA-binding protein MetJ in the recognition of its cognate DNA binding sites.

Searching the *E.coli* genome for putative MetJ sites

Using the combined conformational structure scores and sequence profile scores described above, we searched *E.coli* promoters, which are upstream regions of CDS (coding sequence) in the *E.coli* genome for MetJ binding sites. A score cutoff, 1630, was set at the lowest score for a documented MetJ binding site using the hybrid model. The sequence profile score cutoff is also set to the lowest score for documented MetJ sites. Table 3 gives brief functional characterizations for the downstream genes, paraphrased from descriptions in the GenBank feature table. Table 4 gives details for 21 high-scoring predicted MetJ sites in 17 *E.coli* intergenic regions.

These 17 intergenic regions include all five loci with experimentally verfied MetJ binding sites, plus 12 novel predictions. The five experimentally verified MetJ binding sites are those upstream of the *metC*, *metR*, *metB*, *metF* and *metA* genes (Phillips *et al.*, 1989). In addition, MetJ binding sites were identified upstream of *metK* involved in the regulation of methionine metabolism and *abc*, a gene encodes a component of transporter which may transport amino acids across cell membrane. The genes which would be regulated by the novel sites are generally involved with NADH and regulating cellular potential (*ahpC* is C22 subunit of alkyl hydroperoxide reductase; *dmsB* is B subunit of anaerobic dimethyl sulfoxide reductase; *atpA* is membrane ATP synthase subunit; *hemG* is involved in heme biosynthesis, required for aerobic respiration; *mgtA* is Mg^{2+} transport and ATPase). Our prediction is that transcription from these loci will be de-repressed under conditions of low methionine.

Using sequence similarity information (sequence model) alone to search the promoter regions, if we set the

Table 4. Predicted MetJ binding sites in the *E.coli* genome

Flanking genes[(a)]	MatInspector score	Helix structure score	Combined score	Strand	Midpoint position in the genome[(b)]	Sequence of binding sites
yaeD | abc	0.76	433 183.6	3043.2	−	222 736	TTAGACGTCT GGATGCCTTA
yafC | D	0.67	478 807.3	2129.5	+	231 111	ATAGAGGTTT CAAAGTCAAA
trs5_1 | ykfD	0.77	491 498.9	5508.5	+	274 523	TTGGATGTTT AGATGTCCAT
trs5_1 | ykfD	0.74	486 487.6	4491.1	−	274 523	ATGGACATCT AAACATCCAA
trs5_1 | ykfD	0.71	433 805.3	1647.0	+	274 531	TTAGATGTCC ATACGTTTAG
ybdH | L	0.80	494 832.4	6431.1	+	632 735	TTAGACATCT AAACGTCTTG
ybdH | L	0.79	509 332.4	6713.9	−	632 735	CAAGACGTTT AGATGTCTAA
dsbG | ahpC	0.69	455 730.8	1875.4	+	638 166	GAGGAAGTAT AGATGTCCTT[(d)]
dmsA > B	0.64	496 691.5	1846.4	−	942 558	ACGGACGTTG AGTGGTCAGT[(c)]
ycbB > K	0.71	463 446.0	2761.5	+	982 211	TGAGTCATCT TGACGTCTGC
ycbB > K	0.71	477 956.2	3307.1	−	982 211	GCAGACGTCA AGATGACTCA
yqgD | metK	0.71	433 511.5	1636.0	+	3084 588	ATAGCCATCC AGATGTTAAT[(c)]
b3007 | metC	0.77	481 096.1	5117.4	+	3150 225	TTAGACATCC AGACGTATAA
b3007 | metC	0.78	453 188.1	4337.2	−	3150 225	TTATACGTCT GGATGTCTAA
yhaQ > P	0.69	463 345.1	2161.7	−	3256 803	CACGATGTCG AAACGTCCGT
atpH > A	0.71	432 546.0	1599.7	+	3917 509	TAAGACTGCA AGACGTCTGC[(c)]
atpH > A	0.71	487 260.9	3656.9	−	3917 509	GCAGACGTCT TGCAGTCTTA[(c)]
metE | R	0.73	448 012.8	2760.7	−	4010 486	TTAGCCGTCC AGATGTTTAC
trkH > hemG	0.73	476 692.6	3839.0	+	4032 134	TTGGTCGTCT CGAGGTCTTT
trkH > hemG	0.75	461 581.3	3834.6	−	4032 134	AAAGACCTCG AGACGACCAA
metJ | B	0.76	485 338.9	5004.2	+	4126 165	AAAGAAGTTT AGATGTCCAG
metJ | B	0.71	483 529.3	3516.6	−	4126 165	CTGGACATCT AAACTTCTTT
metJ | B	0.70	481 033.2	3126.9	+	4126 173	TTAGATGTCC AGATGTATTG
metJ | B	0.71	458 993.7	2594.1	−	4126 181	ATGGACGTCA ATACATCTGG
metJ | B	0.76	452 888.2	3784.1	+	4126 181	CCAGATGTAT TGACGTCCAT
metL > F	0.70	451 328.2	2010.0	+	4130 131	CTTTACATCT GGACGTCTAA
metL > F	0.77	495 599.4	5662.7	−	4130 131	TTAGACGTCC AGATGTAAAG
metL > F	0.70	480 794.4	3117.9	+	4130 139	CTGGACGTCT AAACGGATAG
metL > F	0.77	441 781.6	3639.1	−	4130 139	CTATCCGTTT AGACGTCCAG
yjaB | metA	0.70	477 022.3	2976.1	+	4211 824	CTGGATGTCT AAACGTATAA
yjaB | metA	0.78	447 508.2	4123.6	−	4211 824	TTATACGTTT AGACATCCAG
treR | mgtA	0.68	459 137.6	1699.0	+	4465 105	CGTGACGTTT TAACGTCCCT

The table shows the scores of predicted binding sites for which a linear combination of sequence and structure based scores exceeds a threshold of 1630. This threshold is set at the lowest score found for a documented MetJ binding site. Four sites overlap with another by one 8-mer repeat unit, thus 17 intergenic regions are represented.

[(a)] '|' indicates divergently transcribed genes; '>' indicates tandemly transcribed genes.

[(b)] the coordinate shown is for the nucleotide to the right of the midpoint position of the site in the GenBank *E.coli* whole genome sequence as of November 18, 1998. For orientation '+', this is nucleotide 11; for orientation '−', it is the reverse complement of nucleotide 10 as shown.

[(c)] binding site occurs in upstream coding region.

[(d)] binding site overlaps start codon.

threshold to recover all the known MetJ binding sites, we got five more hits than the combined model (Table 3). The five additional genes have no apparent relation to methionine synthesis. They are therefore assumed to be false positives. Our combined model has the obvious advantage over the sequence similarity model. When we searched a random sequence data set the size of the *E.coli* genome with the MatInspector model for MetJ, the number of random positives exceeds the number of biological sites at a score threshold of 0.75. In genes with an established role in methionine biosynthesis, half of the MetJ sites identified using our combined model could not be distinguished from random positives using primary sequence score alone (MatInspector score < 0.75). For comparison, we also searched the promoter regions using DNA helical structural similarity alone (conformation model). If the threshold is set as all the known MetJ sites be recovered, we got 275 hits. So the conformation model does not work very well by itself.

Using Z-scores, we can establish a cutoff above which random hits would not be expected in searching a database as large as the *E.coli* upstream sequence set of coding region. At this score threshold, we identified two hits upstream of the hypothetical open reading frames *ykfD*

and *ybdH*. *ykfD* is a gene which encodes a putative amino acid transporter. It has multiple MetJ binding sites of very high score. *E.coli* has at least two transport systems for methionine uptake, a high-affinity system and one or more low affinity systems. These systems are multi-protein complexes (Greene, 1996) and their components have not been completely defined. *ykfD* might well be one of the methionine transport system components. The function of the product of *ybdH* is unknown, and has not been studied so it is too early to state whether *ybdH* is truly regulated by MetJ.

At the score threshold above which one false positive may be expected at random, four additional hits were found. Two of these are upstream of *abc* and *ycbK*. The *abc* is a known gene which encodes part of amino acids transporter. *ycbK* is another hypothetical gene that has not been studied extensively. Two additional hits are at *hemG* and *atpA* promoter region. These two genes are not connected with MetJ directly, although we cannot rule out the possibility that heme biosynthesis and ATP synthesis are coupled to methionine biosynthesis and these represent true positives.

At a score where several false positives are expected, matches were observed upstream of the genes *metR* and *metK*, and upstream of two hypothetical genes, open reading frames, *yafD* and *yhaQ*. Since the function of the latter two gene products is not known, we do not know whether they are related with MetJ. At lower scores, additional hits were observed upstream of *ahpC*, *mgtA* and *dmsB*. These may be false positives.

DISCUSSION

The effort to understand the determinants of DNA helical structure started with the first high-resolution crystal structure of a DNA oligonucleotide (Drew *et al.*, 1981). As the DNA structure data accumulates, several groups have compiled the average DNA base step parameters (Bolshoy *et al.*, 1991; Cacchione *et al.*, 1989; Bansal *et al.*, 1995; Olson *et al.*, 1998). Those parameters are either compiled too early and we have much more crystal structures now, or in Olson's case, only complexes of DNA and protein were considered. It is widely believed that proteins distort the DNA conformation after binding. So selecting high quality unbound DNA crystal structures for our study is more appropriate for accessing the structure of free DNA in a physiological condition. Therefore we compiled our own DNA base step parameter table using high resolution (2.5 Å or better) unbound DNA crystal structures from NDB. On comparison with the previous compiled base step parameters, our parameters are closest to the set compiled by Olson *et al.*, except their roll and twist values are generally larger than ours (Olson *et al.*, 1998). That could be due to the effects of protein binding and subsequent distortion. At the time when Bansal's set of parameters were compiled (Bansal *et al.*, 1995), few DNA structures were available, and the CA steps were dominated by the BII form of DNA, perhaps accounting for very high twist and big roll in this data set. But it has been known that the BI form of DNA structure is much more common and probably the form exists *in vivo* (Hartmann *et al.*, 1993). We therefore compiled our parameters exclusively on crystal structures in which the DNA was the BI conformation.

The base step parameters like twist, roll, tilt, and slide have been given a standard definition (Dickerson *et al.*, 1989). But the definition for groove width has not been agreed on. Both the shortest phosphate P–P distance across the groove (Kopka *et al.*, 1985) or the shortest O4′–O4′ distance across the groove (Kim *et al.*, 1993) have been used. These are approximations of actual groove width. Later studies covered that the definitions are structure dependent (Suzuki and Yagi, 1996). In other words, the definitions are often unstable, since fine changes in DNA structures can easily alter which pair of phosphorus atoms represents the closest distance at a given point along the structure (Stofer and Lavery, 1994). To eliminate this drawback, Lavery fits the two strands of DNA to smooth curves passing the phosphates and the shortest distance between the two curves are defined as the groove width (Stofer and Lavery, 1994). This definition is the most appropriate but not easily related to atomic coordinates. More importantly, all the above definitions lack the association with base steps. The Calladine definition (el Hassan and Calladine, 1998, Equation A1) is the average of two consecutive measurements. This definition is the least of the contributions to a base step, and the sequence dependent variations are averaged out. We introduce a new term to measure the contribution of a base step to the groove width, called groove opening. It is not equivalent to groove width, instead, measuring how open a base step is in the minor groove. We found the groove opening of base steps shows sequence dependent variations and contributes to our analysis.

MetJ was chosen for this analysis because x-ray crystal structure of the bound protein MetJ/DNA complex has been solved and because a collection of SELEX selected binding sequences have been published (He *et al.*, 1996). Detailed examination of structural features of free MetJ binding sites demonstrates that the mean preferred structure of the transcription factor binding sites is similar to the DNA conformation observed in the co-crystal structure of DNA in MetJ/DNA complex. This suggests that the energy of DNA conformational distortion during complex formation is minimal for these sites; consistent with the hypothesis that structural information plays a role in protein (transcription factors) recognition of DNA binding sites. Though the rolls are generally smaller in free DNA than in the protein-bound DNA, the two curves have

the same pattern (Figure 3). The similarity in roll pattern does contribute prediction power to our analysis. The differences in roll values between free DNA and bound DNA are caused by protein binding. It is well known that protein bends DNA after binding. And bending will dramatically increase the roll of DNA helix.

Combining sequence and structure similarity for searching MetJ binding sites in *E.coli* promoter regions, we have much better performance than using either sequence similarity or structure similarity alone. The sequence model gives five more likely false positives than the combined model. But the conformation model alone has the worst performance, with many more likely false positives. The performance of the conformation model is worse than the sequence model is probably expected. Much more sequence data are available to train the sequence model, and sequence profile based models are well established and highly refined. In contrast, models based on DNA helix conformation have only recently been developed and was parameterized using very limited data. It is likely that the performance of conformation based models will be enhanced giving more structure data available for training purposes.

The combined sequence/structure model presented here appears to offer superior performance in the recognition of MetJ binding sites compared with either primary sequence based profiles or the tertiary structure model alone. In addition to finding all of the documented MetJ binding sites, we found *metK* to be involved in methionine biosynthesis which have putative MetJ binding sites. We suggest that, like the other components of the methionine biosynthesis pathway, this gene is also regulated by MetJ. In addition, we have identified putative MetJ binding sites upstream of several genes coding for transporters or putative transporters. Complex membrane transport systems for the amino acid and intermediates in their biosynthesis pathways are present in *E.coli*, and the components of these transporter systems have not been fully characterized. We suggest that these genes might also be co-regulated by MetJ and may play a role in methionine metabolism.

The five documented loci are also listed in a survey study along with a sixth binding site upstream of *metE* (Phillips *et al.*, 1989). But a more thorough research presents transcript mapping and footprinting data which suggests that this sixth site is not used *in vivo* (Cai *et al.*, 1989). *MetE* and *metR* are divergently transcribed from a common intergenic region. The *metE* promoter and the first of two *metR* promoters substantially overlap. Their transcription start sites are 29 nucleotides apart. The demonstrated MetJ protein binding site covers the transcriptional start site for *metE* and the −35 region of the (first) *metR* promoter. The enzyme MetE is repressed almost 10-fold by methionine, while the transcriptional activator MetR is repressed only 3-fold (Wu *et al.*, 1993). The sixth MetJ binding site, immediately upstream of the *metE* start codon, can indeed be footprinted *in vitro*, but this requires a 10-fold higher concentration of MetJ protein (and its co-factor *S*-adenosylmethionine) than footprinting the shared binding site which covers the two promoters and the *metE* transcription start.

Similar organization is seen at the *metJ/metB* intergenic region. Again, an enzyme (MetB) and a transcriptional regulator (MetJ) are divergently transcribed from a common intergenic region. The two genes have closely spaced promoters, both of which are repressed by a single MetJ binding array. The enzyme (MetB) has only one promoter and transcription start site, and is strongly repressed. The regulator (MetJ) has three promoters, of which only the first is repressed by MetJ. In view of these examples, it will not be surprising to find both flanking genes repressed by a shared MetJ binding site, at some of the predicted loci which show two divergent transcripts.

This study is focused on MetJ binding sites due to MetJ being a very well studied transcription factor. Much information can be used for comparison or validation. But MetJ is not the only transcription factor which uses structural information in specifying its binding sites. Another excellent example is *trp* operon repressor (TrpR). It has been suggested that the Trp repressor uses structural information in recognizing its cognate binding site (Luisi and Sigler, 1990). Applying our structure based approach to binding site definition also improved performance for TrpR. The homeobox proteins are other proteins which have been studied and suggested that DNA structure plays a role in site recognition (Wilson *et al.*, 1995). Furthermore, we have also analyzed the ROC curves for an additional 20 transcription factors from both prokaryotes and eukaryotes. The results showed that it is a general phenomenon of structure similarity information being used by transcription factors.

ACKNOWLEDGEMENTS

This work was supported in part through grants from the National Institutes of Health (HG-R01-01391), Department of Energy (DE-FG02-94ER61910) and Merck Foundation for Genome Research (grant #225).

REFERENCES

Allemann,R.K. and Egli,M. (1997) DNA recognition and bending. *Chem. Biol.*, **4**, 643–650.

Aravind,L. and Koonin,E.V. (1999) DNA-binding proteins and evolution of transcription regulation in the archaea. *Nucleic Acids Res.*, **27**, 4658–4670.

Bansal,M., Bhattacharyya,D. and Ravi,B. (1995) NUPARM and NUCGEN: software for analysis and generation of sequence dependent nucleic acid structures. *CABIOS*, **11**, 281–287.

Berman,H.M., Olson,W.K., Beveridge,D.L., Westbrook,J., Gelbin,A. and Schneider,B. (1992) The nucleic acid database. *Biophys. J.*, **63**, 751–759.

Bolshoy,A., McNamara,P., Harrington,R.E. and Trifonov,E.N. (1991) Curved DNA without A-A: experimental estimation of all 16 DNA wedge angles. *Proc. Natl Acad. Sci. USA*, **88**, 2312–2316.

Cacchione,S., De Santi,P., Foti,D.P., Palleschi,A. and Savino,M. (1989) Periodical polydeoxynucleotides and DNA curvature. *Biochemistry*, **28**, 8706–8713.

Cai,X.Y., Maxon,M.E., Redfield,B., Glass,R., Brot,N. and Weissbach,H. (1989) Methionine synthesis in *Escherichia coli*: effect of the MetR protein on metE and metH expression. *Proc. Natl Acad. Sci. USA*, **86**, 4407–4411.

Dickerson,R.E. (1992) DNA structure from A to Z. *Meth. Enzymol.*, **111**, 67–110.

Dickerson,R.E., Bansal,M., Calladine,C.R., Hunter,W.N., Lavery,R., Nelson,H.C., Olson,W.M., Saenger,W., Shakked,Z., Sklenar,H., Soumpasis,D.M., Tung,C.S., Kitzing,E., Wang,A.H.J. and Zhurkin,V.B. (1989) Definitions and nomenclature of nucleic acid structure parameters. *J. Mol. Biol.*, **205**, 787–791.

Diekmann,S. (1989) The migration anomaly of DNA fragments in polyacrylamide gels allows the detection of small sequence-specific DNA structure variations. *Electrophoresis*, **10**, 354–359.

Drew,H.R., Wing,R.M., Takano,T., Broka,C., Tanaka,S., Itakura,K. and Dickerson,R.E. (1981) Structure of a B-DNA dodecamer: conformation and dynamics. *Proc. Natl Acad. Sci. USA*, **78**, 2179–2183.

el Hassan,M.A. and Calladine,C.R. (1998) Two distinct modes of protein-induced bending in DNA. *J. Mol. Biol.*, **282**, 331–343.

Garvie,C.W. and Phillips,S.E. (2000) Direct and indirect readout in mutant met repressor–operator complexes. *Structure Fold Des.*, **8**, 905–914.

Ghosh,D. (1992) TFD: the Transcription Factors Database. *Nucleic Acids Res.*, **20**, 2091–2093.

Gilbert,W. and Muller-Hill,B. (1966) Isolation of the lac repressor. *Proc. Natl Acad. Sci. USA*, **56**, 1891–1898.

Gilbert,W. and Muller-Hill,B. (1967) The lac operator is DNA. *Proc. Natl Acad. Sci. USA*, **58**, 2415–2421.

Greene,R.C. (1996) Biosynthesis of methionine. *Escherichia coli and Salmonella: Cellular and Molecular Biology*, 2nd edn, ASM Press, Washington, DC, pp. 542–560.

Harrington,R.E. (1992) DNA curving and bending in protein–DNA recognition. *Mol. Microbiol.*, **6**, 2549–2555.

Hartmann,B., Piazzola,D. and Lavery,R. (1993) BI–BII transitions in B-DNA. *Nucleic Acids Res.*, **21**, 561–568.

He,Y.Y., Stockley,P.G. and Gold,L. (1996) *In vitro* evolution of the DNA binding sites of *Escherichia coli* methionine repressor, MetJ. *J. Mol. Biol.*, **255**, 55–66.

Istrail,S., Pevzner,P. and Waterman,M. (1998) *Proceedings of the Second Annual International Conference on Computational Molecular Biology*. ACM Press, New York.

Jacob,F. and Monod,J. (1961) Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.*, **3**, 318–356.

Kim,Y., Geiger,J.H., Hahn,S. and Sigler,P.B. (1993) Crystal structure of a yeast TBP/TATA-box complex. *Nature*, **465**, 512–520.

Kopka,M.L., Goodsell,D.S., Baikalov,I., Grzeskowiak,K., Cascio,D. and Dickerson,R.E. (1994) Crystal structure of a covalent DNA-drug adduct: anthramycin bound to C-C-A-A-C-G-T-T-G-G and a molecular explanation of specificity. *Biochemistry*, **33**, 13593–13610.

Kopka,M.L., Yoon,C., Goodsell,D., Pjura,P. and Dickerson,R.E. (1985) Binding of an antitumor drug to DNA, Netropsin and C-G-C-G-A-A-T-T-BrC-G-C-G. *J. Mol. Biol.*, **183**, 553–563.

Lavery,R. and Sklenar,H. (1988) The definition of generalized helicoidal parameters and of axis curvature for irregular nucleic acids. *J. Biomol. Struct. Dyn.*, **6**, 63–91.

Liu,R., Blackwell,T.W. and States,D.J. (1998) A structure based similarity measure for nucleic acid sequence comparison. *Proceedings of Second Annual International Conference on Computational Molecular Biology*. ACM Press, New York, pp. 173–181.

Lu,X.J. and Olson,W.K. (1999) Resolving the discrepancies among nucleic acid conformational analyses. *J. Mol. Biol.*, **285**, 1563–1575.

Luisi,B.F. and Sigler,P.B. (1990) The stereochemistry and biochemistry of the trp repressor–operator complex. *Biochim. Biophys. Acta*, **1048**, 113–126.

Olson,W.K., Gorin,A.A., Lu,X.J., Hock,L.M. and Zhurkin,V.B. (1998) DNA sequence-dependent deformability deduced from protein–DNA crystal complexes. *Proc. Natl Acad. Sci. USA*, **95**, 11163–11168.

Otwinowski,Z., Schevitz,R.W., Zhang,R.G., Lawson,C.L., Joachimiak,A., Marmorstein,R.Q., Luisi,B.F. and Sigler,P.B. (1988) Crystal structure of trp repressor/operator complex at atomic resolution. *Nature*, **335**, 321–329.

Phillips,S.E., Manfield,I., Parsons,I., Davidson,B.E., Rafferty,J.B., Somers,W.S., Margarita,D., Cohen,G.N., Saint-Girons,I. and Stockley,P.G. (1989) Cooperative tandem binding of met repressor of *Escherichia coli*. *Nature*, **341**, 711–715.

Pittard,A.J. (1996) Biosynthesis of the aromatic acids. *Escherichia coli and Salmonella: Cellular and Molecular Biology*, 2nd edn, ASM Press, Washington, DC, pp. 458–484.

Prive,G.G., Heinemann,U., Chandrasegaran,S., Kan,L.S., Kopka,M.L. and Dickerson,R.E. (1987) Helix geometry, hydration, and G.A mismatch in a B-DNA decamer. *Science*, **238**, 498–504.

Quandt,K., Frech,K., Karas,H., Wingender,E. and Werner,T. (1995) MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res.*, **23**, 4878–4884.

Raumann,B.E., Rould,M.A., Pabo,C.O. and Sauer,R.T. (1994) DNA recognition by beta-sheets in the Arc repressor–operator crystal structure. *Nature*, **367**, 754–757.

Schug,J. and Overton,G.C. (1997) Modeling transcription factor binding sites with Gibbs Sampling and Minimum Description Length encoding. *ISMB*, **5**, 168–271.

Shakked,Z., Guzikevich-Guerstein,G., Frolow,F., Rabinovich,D., Joachimiak,A. and Sigler,P.B. (1994) Determinants of repressor/operator recognition from the structure of the trp operator binding site. *Nature*, **368**, 469–473.

Somers,W.S. and Phillips,S.E. (1992) Crystal structure of the met repressor–operator complex at 2.8 A resolution reveals DNA recognition by beta-strands. *Nature*, **359**, 387–393.

Stofer,E. and Lavery,R. (1994) Measuring the geometry of DNA grooves. *Biopolymers*, **34**, 337–346.

Suzuki,M. (1995) DNA recognition by a beta-sheet. *Protein Eng.*, **8**, 1–4.

Suzuki,M. and Yagi,N. (1996) An in-the groove view of DNA structures in complexes with proteins. *J. Mol. Biol.*, **255**, 677–687.

Swets,J.A. (1988) Measuring the accuracy of diagnostic systems. *Science*, **240**, 1285–1293.

Tung,C.S. and Harvey,S.C. (1986) Computer graphics program to reveal the dependence of the gross three-dimensional structure of the B-DNA double helix on primary structure. *Nucleic Acids Res.*, **14**, 381–387.

Wilson,D.S., Guenther,B., Desplan,C. and Kuriyan,J. (1995) High resolution crystal structure of a paired (Pax) class cooperative homeodomain dimer on DNA. *Cell*, **82**, 709–719.

Wu,W.F., Urbanowski,M.L. and Stauffer,G.V. (1993) MetJ-mediated regulation of the *Salmonella typhimurium* metE and metR genes occurs through a common operator region. *FEMS Microbiol. Lett.*, **108**, 145–150.

Yanagi,K., Prive,G.G. and Dickerson,R.E. (1991) Analysis of local helix geometry in three B-DNA decamers and eight dodecamers. *J. Mol. Biol.*, **217**, 201–214.