Pacific Symposium on Biocomputing 6:139-150 (2001)

A STRUCTURE-BASED APPROACH FOR PREDICTION OF PROTEIN BINDING SITES IN GENE UPSTREAM REGIONS

Y. MANDEL-GUTFREUND, A. BARON, H. MARGALIT
Department of Molecular Genetics and Biotechnology
The Hebrew University - Hadassah Medical School
POB 12272, Jerusalem 91120 Israel

hanah@md2.huji.ac.il ; yael@gorby.ucsc.edu

The challenge of identifying DNA regulatory sequences based on sequence information only has been emphasized in view of the fast accumulation of new genes in the databases. While most predictive algorithms are based on multiple alignments of already known binding sites, here we examine the usefulness of a novel approach that is based on structural information of the protein-DNA complex. It has already been shown that specific recognition between a protein and its DNA target is achieved by stereo-chemical complementarity between the protein amino acids and the DNA bases. The proposed computational scheme uses crystallographic information to define the set of amino acid-base contacts between the proteins of a given DNA-binding protein family and their DNA targets. The compatibility of a given protein to bind to putative regulatory DNA sequences is then evaluated by knowledge-based parameters for amino acid-base interactions. By this procedure gene upstream regions may be screened for potential binding sites for regulatory proteins. Predictions are demonstrated for the *E. coli* cyclic AMP receptor protein (CRP) which recognizes the DNA via the helix-turn-helix motif, and for various Zif268-like proteins which belong to the Cys_2His_2 zinc finger family. The advantages and limitations of this approach are discussed.

Introduction

Sequences upstream transcription start positions play a major role in the regulation of gene expression. They are recognized by regulatory proteins which act upon binding as transcription repressors or activators, controlling the rate of transcription initiation. The identification of such sequences upstream from a specific gene is therefore essential for understanding its transcription regulation. Traditionally, the identification of DNA regulatory sequences and of the base pairs that play a role in specific binding has been carried out by a variety of experimental methods. These include mutation analysis and direct binding measurements (e.g. Takeda *et al.*, 1989), selection experiments by phage display libraries (e.g. Choo & Klug, 1994), and co-crystalization of the protein-DNA complex (e.g. Kim & Burley, 1994). Presently, with the accumulation of many new gene sequences due to the large-scale genomic sequencing projects, we are faced by the challenge of predicting the gene regulatory sequences based on sequence data alone. For this we need a computational tool that will screen the upstream region of the gene and identify potential binding sites for regulatory proteins.

Various strategies have been employed in the development of search procedures for DNA binding sites (e.g. Thieffry *et al.*, 1998; Goodrich *et al.*, 1990; O'Neill,

1989; Berg & von Hippel, 1988; Schneider *et al.,* 1986). Most of these approaches require a large ensemble of experimentally determined DNA binding sequences of a certain protein, and essentially define the site by a matrix that represents the frequency of bases at each position. This approach is limited, however, to regulatory proteins for which binding sites have been already identified experimentally. An alternative approach would be to base the prediction on the mode of binding between the protein and the DNA, and search for DNA sequences that are preferred for this binding mode. Such an approach would be most suitable for searching potential binding sites of regulatory proteins that belong to a well defined DNA-binding protein family.

Two key components are required in order to apply such an approach:
1) Knowledge on the protein and DNA residue positions that are involved in binding, preferably from a crystal structure of the complex. 2) A method to evaluate the compatibility of different DNA bases and amino acids to interact (Suzuki *et al.,* 1994; Suzuki & Yagi, 1994, Kono & Sarai, 1999). Here we present a computational scheme that uses knowledge-based parameters for amino acid-base interactions based on data from crystal structures of protein-DNA complexes (Mandel Gutfreund & Margalit, 1998). By applying these parameters to specified binding models, a score that reflects the compatibility between a protein sequence and a DNA site can be evaluated. The applicability of this scheme is demonstrated for two examples of binding sites that are recognized by members of two distinct families of DNA-binding proteins: 1) The *E. coli* CRP which recognizes the DNA via the helix-turn-helix motif. 2) Various Zif268-like proteins which belong to the Cys_2His_2 zinc finger family, but differ in the residues that are involved in binding. We show that the current procedure succeeds fairly well in identifying the experimentally determined binding sites.

Methods

Ranking putative DNA binding sites

The binding mode of a regulatory protein with its DNA binding site is determined according to the crystal structure of the complex. We extract from the co-crystal data the pairs of amino acid-base that are in contact. The binding model is then defined by the corresponding positions in the protein and DNA sequences that participate in these contacts. Given the amino acids that are present at these defined protein positions, and according to the binding model, the goal is to fit DNA sequences that would be most compatible for binding. These would be sequences composed of certain bases at the defined positions that would fit the amino acids at the corresponding protein positions. To search for such sequences, scores that reflect the compatibility between the various combinations of amino acid-base are required (see below). Given that such scores are available, a score for the compatibility between a DNA sequence and a protein sequence is obtained by

summing up the individual scores of pairs of amino acid-base according to the binding model. A long DNA sequence can be searched for preferred binding sites by calculating the scores in overlapping windows of length L (defined by the binding model) along the sequence. The highest scoring windows are expected to be the favorable DNA binding sites.

Amino acid-base scoring matrix

It is well known that specific recognition between a regulatory protein and its DNA target is achieved by structural complementarity between the interacting elements, and by specific interactions between the protein amino acids and DNA bases. The latter involve mostly hydrogen bonds and hydrophobic interactions. Previously we have shown clear preferences of certain amino acids to interact with certain bases (Mandel-Gutfreund & Margalit, 1995). The distribution of pairs of amino acid-base that we extracted from the structures was used by us to derive a scoring matrix for amino acid-base interaction (Mandel Gutfreund & Margalit, 1998). The scores are based on the frequencies of pairs of amino acid-base that are in contact in 53 crystallographically solved protein-DNA complexes, and were derived by calculating the likelihood ratio between the frequency of pairs of amino acid-base in the data and the theoretical probability of obtaining such pairs. The scoring matrix that is presented in Table 1 differs slightly from the original one (Mandel Gutfreund & Margalit, 1998); in addition to classical hydrogen bonds and hydrophobic interactions that were used in the original table, it is based also on pairs of amino acid-base that interact via $CH^{\cdots}O$ hydrogen bonds. The latter were shown by us recently to contribute significantly to specific amino acid-base recognition (Mandel Gutfreund *et al.*, 1998).

Results and Discussion

Prediction of binding sites for Cys_2His_2 proteins

The set of contacts involved in the interactions between the Cys_2His_2 Zif268-like proteins and the DNA is defined by a consensus binding model of Zif268 protein and its DNA binding site. This binding model is based on structural data and experimental binding data (Choo & Klug, 1997) (Figure 1). The inherent assumption is that the same set of contacts between the protein and the DNA is kept upon substitution of the sequences. For the Zif268 protein this simplistic assumption has been confirmed experimentally by several selection studies using protein and DNA variants (Reviewed in Choo & Klug, 1997).

Table 1: Scoring Matrix
(log odds of observed/expected frequency of base-amino acid pairs
in crystallographically solved complexes)

	G	A	T	C
GLY	-4.07	-4.07	-4.07	-4.07
ALA	-4.07	-4.07	0.53	-3.85
VAL	-4.07	-4.07	-0.31	-3.71
ILE	-4.07	-4.07	0.52	-3.58
LEU	-4.07	-4.07	-1.07	-4.07
PHE	-4.07	-4.07	-0.94	-0.25
TRP	-2.10	-4.07	-2.10	-4.07
TYR	-3.00	-3.00	0.40	0.00
MET	-2.71	-0.41	0.29	-0.41
CYS	-2.37	-0.06	-2.37	-0.06
THR	-3.59	-0.19	0.10	-1.29
SER	0.29	-0.81	0.98	-0.41
GLN	-0.23	1.03	1.28	-3.22
ASN	0.35	1.80	1.04	1.16
GLU	-4.07	-1.37	0.01	0.93
ASP	-4.07	-3.50	-3.50	0.88
HIS	1.43	0.33	0.73	-2.67
ARG	2.60	0.21	1.12	-4.07
LYS	2.02	-0.21	0.08	-4.07
PRO	-4.07	-4.07	-0.43	-3.42

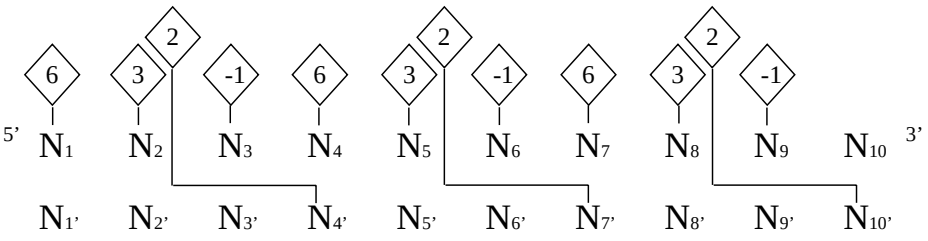

Figure 1 A consensus model for the Zif268-like zinc finger protein binding to DNA, based on the crystal structure of the Zif268-DNA complex (Elrod-Erickson *et al.,* 1996) and experimental studies (Choo & Klug, 1997).

As illustrated, three consecutive zinc fingers bind in an "anti parallel" modular fashion to a 10 base pair DNA site, each finger binds four base pairs. In each finger, positions -1, 3 and 6 with respect to the α helix that binds the DNA, contact three adjacent bases on one DNA strand, while position 2 binds to a base in the following sub-site but on the complementary DNA strand.

Nine promoter regions regulated by five different proteins which belong to Zif268-like (Cys_2His_2) zinc finger protein family are analyzed. The regulatory regions of these genes are defined as annotated in GenBank. When the promoter region has not been specifically defined, we either use a region of 500 base pairs upstream from annotated mRNA start sites, or otherwise use a region of 1000 base pairs upstream from annotated open reading frames (ORFs). The amino acids that are involved in binding according to the model are determined by sequence alignment of each of the proteins with Zif268.

An example of the scores computed for binding of the Sp1 zinc finger protein to all overlapping windows of length 10 spanning the *dhfr* promoter region is presented in Figure 2. In this example, all four experimentally defined binding sites (Kandonaga *et al.,* 1986) are ranked by the scoring scheme as the highest among all other overlapping 10-mers in this promoter region.

Table 2 summarizes the calculated scores for 21 experimentally identified protein binding sites within the nine different promoter regions. As shown in the last column of the table, the scores obtained by the experimentally defined sites are ranked among the highest possible scores in a given promoter region. It is important to note that several sites along a promoter region may obtain the same score. Thus, the rank does not simply represent the number of other sites which are scored higher, but indicates the relative position of this score among the other scores. Except for the third Krox20 binding site in the Hox-B2 promoter and the second SWI-5 binding site in the HO promoter, all other binding sites are ranked among the highest 2% scores. In 7 out of the 9 promoters at least one of the binding sites is predicted with the highest or second highest score.

Table 2: Ranking the Experimentally Defined Binding Sites of Zif268 by the Computed Score

Protein	**Promoter**	**L** [1]	**Sequence** [2]	**Score** [3]	**Rank** [4]
	mouse dhfr	304	GGG GCG GGG C	13.62	1
	mouse dhfr		GGG GCG GAG C	12.52	2
	HIV-LTR	536	GAGGCG TGG C	12.73	1
	HIV-LTR		GGG GAG TGG C	11.53	2
Sp1	*HIV-LTR*		TGG GCG GGA C	9.45	3
	human MtII-A	300	GGG GCG GGG C	13.62	1
	SV40	561	GGG GCG GAG A	13.21	1
	SV40		GGG GCG GGA C	11.39	2
	SV40		GGG GCG GGA T	10.29	5
	SV40		TGG GCG GAG T	9.48	7
	SV40		TGG GCG GAA C	8.35	9
	mouse zif	1000	GCG GGG GCG A	10.40	5
Zif268	*mouse zif*		GCG GGT GAG C	6.62	18
	mouse junD	944	GCG GGG GCG G	15.34	1
	mouse junD		GCG GGG GCC G	8.68	11
	mouse Hox-B2	537	GCG TGG GTG G	13.73	1
Krox20	*mouse Hox-B2*		GAG GGG GAG G	10.74	4
	mouse Hox-B2		CCG TGG GAG T	1.31	56
EKLF	*mouse β-globin*	121	AGG GTG TGG C	6.82	1
SWI-5	*HO*	1191	ATG GCG TGG C	9.50	8
	HO		ATA GCA TGC T	-4.15	264

[1]The length in base pairs of the region analyzed. [2]The sequence of the experimentally defined binding site. [3]Scores for the binding sites, based on a binding template and on the matrix in table 1. [4]The rank of the known site among the scores of all other possible 10-mers in the promoter region.

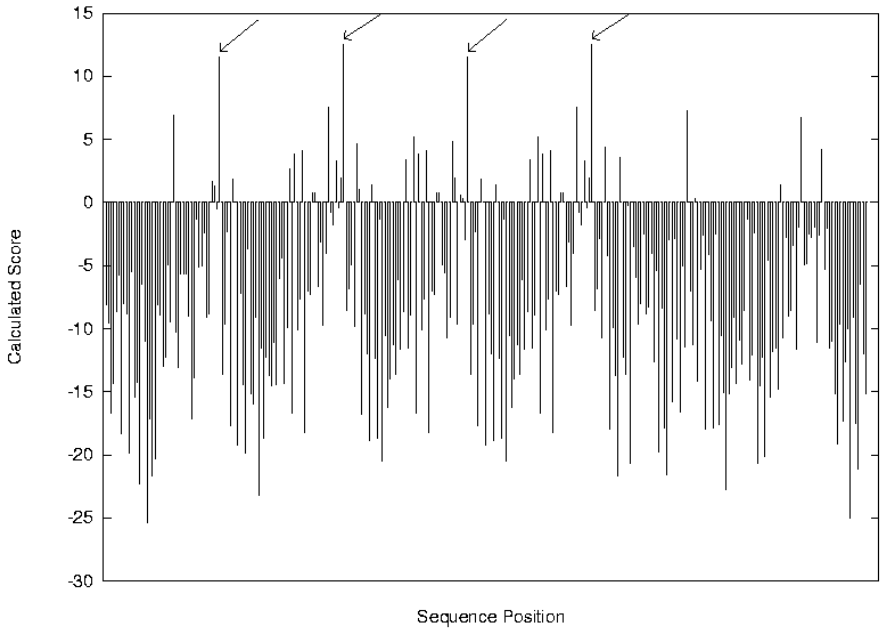

Figure 2 Prediction of Sp1 binding sites in the *dhfr* promoter. Calculated scores for overlapping windows of length 10 along the promoter sequence are presented. The binding sites that were identified experimentally (Kandonaga *et al.*, 1986) are indicated by arrows.

Prediction of CRP binding sites

CRP is known to regulate many genes in *E.coli*. Many of its binding sites have been identified experimentally by various methods, such as DNAase I footprinting, hydroxyl radical footprinting, genetic deletion and mutation experiments (Barber & Zhurkin, 1990). The binding model for CRP (Figure 3) is defined based on the 2.5 Å crystal structure of the CRP/DNA complex (Parkinson *et al.*, 1996). Only specific interactions between the protein side chains and the DNA base edges are considered, including CH···O interactions. As suggested by Parkinson *et al. (*1996), we consider the interactions of the dimer with the two DNA half sites as perfectly symmetric, where the contacts are based on the left half site. Contacts with the central region of the binding site are not included. CRP acts either as a repressor or activator and it is possible that it uses alternative recognition models for binding in each case. However, since the basic recognition motif is present in all sites, it is conceivable that the mode of binding in all cases is similar. Also, experimental results regarding the effect of mutations on the DNA binding specificity of various sites (e.g. Ebright

et al., 1984) are consistent with the binding model suggested by the crystal structure (Parkinson *et al.*, 1996). Therefore, as above, we assume that the same binding framework is used by the protein in binding its different binding sites.

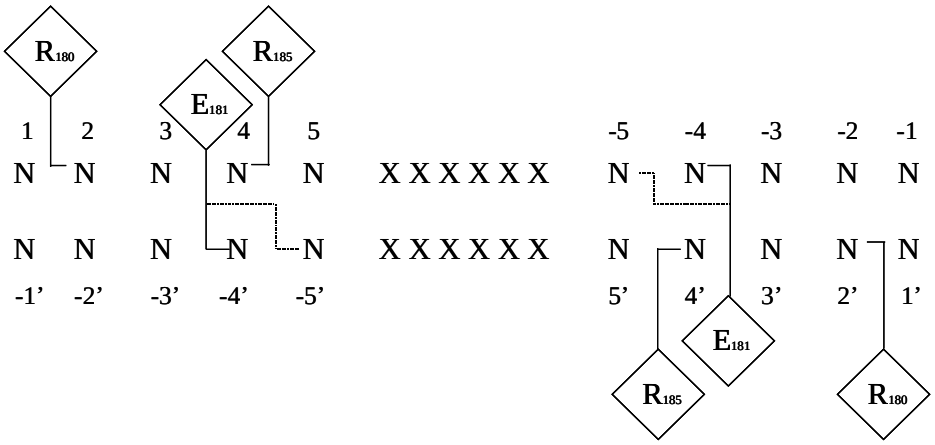

Figure 3 A consensus scheme of the CRP-DNA complex. As illustrated, the interactions considered for the consensus pattern are symmetric, based on the contacts observed in the crystal structure between the monomer A and the left DNA half site (Parkinson *et al.*, 1996). The bases which are part of the consensus site are denoted with an N and numbered from 5' to 3' (bases on the complementary strand are primed). The six central base pairs are assigned with an X. The interactions in each half site involve Arg 180 and the second base in the consensus box, Glu 181 and positions 4 and 5 of the complementary strand, and Arg 185 with and fourth position. The contact between Glu 181 and position 5 which involves a CH···O interaction is designated by a dotted line.

15 regulatory regions of *E.coli* genes regulated by CRP are analyzed. Preferred binding sites of *E.coli* CRP are searched for in upstream regions of these genes. The regulatory regions are defined as in Thieffry *et al.* (1998), covering 400 base pairs upstream and 50 base pairs downstream from annotated ORFs. Table 3 summarizes the scores for 20 known CRP binding sites in the 15 promoter regions. The predictions are less successful than in the previous example, but still, the scores of more than 50% of the sites studied fall within the five highest ranks. 15 out of the 20 sites are ranked within the highest 2% scores.

The advantage of this procedure for identification of protein binding sites in DNA regulatory regions is that it does not need an ensemble of binding sites of a given protein for prediction of new binding sites. If the protein belongs to a family whose binding mode is known, a single representative complex of the family in which the set of contacts is defined provides the sufficient information. Thus, it can be used for the prediction of binding sites for newly identified proteins that are clustered to a defined family. In principle, like the structural genomics initiative

Table 3: Ranking of CRP Binding Sites in *E.coli* Promoters by the Computed Score

Gene	Sequence [1]	Position [2]	Score [3]	Rank [4]
lacZ	G**GATA**ACAATT**TCAC**A	-28	8.50	10
	T**GTGA**GCGGAT**AACA**A	-35	5.64	27
	T**GTGA**GTTAGC**TCAC**T	-107	12.28	1
malE	T**GTGA**TCTCTG**TTAC**A	-116	8.97	6
malK	C**GAGG**ATGAGA**ACAC**A	-122	11.82	1
	C**TCGG**TTTAGT**TCAC**C	-156	11.72	2
malT	T**GTGA**CACAGT**GCAA**A	-139	6.72	12
ompA	C**CTGA**CGGAGT**TCAC**A	-172	5.61	21
araC	A**GTGT**CTATAA**TCAC**G	-175	10.90	1
araE	T**GGAA**TATCCA**TCAC**A	-128	8.97	3
crp	T**GCAA**AGGACG**TCAC**A	-133	8.97	6
	G**GCGA**CCTGGG**TCAT**G	-235	9.89	3
deoC	T**GTGA**TGTGTA**TCGA**A	-93	10.80	4
	T**TTGA**ACCAGA**TCGC**A	-146	10.80	4
exuT	G**GTGA**GAGCCA**TCAC**A	-136	12.28	2
fur	T**GTAA**GCTGTG**CCAC**G	-155	9.89	3
galE	T**TTAT**TCCATG**TCAC**G	-75	6.11	13
glpD	T**GTTA**TACATA**TCAC**T	-113	8.50	6
melR	C**GTGC**TCCCAC**TCGC**A	-71	8.20	14
tnaA	T**GTGA**TTCGAT**TCAC**A	-371	12.28	1

[1]The sequence of a known binding site. Bases in the core of the binding site are indicated in bold. [2]Position relative to the ORF of the gene. [3,4] See Table 2.

that attempts to provide solved structures that represent all possible folds, a spectrum of co-crystals that cover all protein-DNA binding modes may be achieved also in the future. In such a case, the approach presented here has clear advantage on identification of binding motifs based on binding data.

The computational procedure essentially resembles the threading approach that is used to evaluate sequence-structure fit in proteins. There, a sequence is threaded through a three-dimensional template of a protein based on a known protein structure, and the sequence-structure fit is evaluated by statistical pairwise potentials. Here, the DNA sequence is "threaded" through the amino acids that are involved in binding, according to a binding model defined by the protein-DNA complex, and the compatibility between the two is evaluated by knowledge-based parameters for amino acid-base interactions. The potential of this approach was also demonstrated by Kono & Sarai (1999), who applied it similarly to the problem of protein-DNA binding, using a different set of knowledge-based parameters for amino acid-base interactions.

From the results it is clear that the computational scheme does not always predict the known site at the first rank. There are several possible reasons for that: 1) The quantitative parameters were derived from the pair interactions in a variety of protein-DNA complexes, and reflect the likelihood of interaction in general. Consequently, possible position dependent effects that are specific to each binding motif are masked. For example, in the zif268-like zinc fingers steric constraints that are position dependent are probably imposed by the specific orientation of the protein binding element relative to the DNA (Choo & Klug, 1997). Conceivably, incorporation of position dependent effects that are specific to each binding motif together with the knowledge-based parameters may yield better predictions (Suzuki & Yagi, 1994; Choo & Klug, 1997). 2) There are other factors that affect binding, such as the sequence context of the binding sites, coupled interactions where one amino acid is assisted by another in contacting the DNA (e.g. Aggarwal *et al.*, 1988; Elrod-Erickson *et al.*, 1996), and the structure of the DNA binding site (e.g. Kim & Burley, 1994; Parkinson *et al.*, 1996; Rice *et al.*, 1996). Incorporation of these considerations in future predictive schemes are expected to improve the predictions. However, although the computational procedure does not always succeed in identifying exclusively the binding site of a given protein, it is powerful in reducing the number of potential binding sites and eliminating the ones that are less favorable. Its success in identifying binding sites of proteins from two entirely different protein families suggests that it may be useful for identifying binding sites of proteins from various protein families in which the framework of the amino acid - base interactions is defined.

Acknowledgment

This study was supported by the Israeli National Science Foundation. We thank Yael Altuvia for her help in preparing this manuscript.

References

Aggarwal, A.K., Rodgers, D.W., Drottar, M., Ptashne, M. & Harrison, S.C. (1988). Recognition of a DNA operator by the repressor of phage 434: a view at high resolution. *Science,* **242,** 899-907.

Barber, A.M., Zhurkin, V.B. (1990). CAP binding sites reveal pyrimidine-purine pattern characteristic of DNA bending. *J Biomol Struct Dyn* **8,** 213-232.

Berg, O.G., von Hippel, P.H. (1988). Selection of DNA binding sites by regulatory proteins. II. The binding specificity of cyclic AMP receptor protein to recognition sites. *J Mol Biol* **200,** 709-723.

Choo, Y., Klug, A. (1994). Selection of DNA binding sites for zinc fingers using rationally randomized DNA reveals coded interactions. *Proc Natl Acad Sci U S A* **91,** 11168-11172.

Choo, Y., Klug, A. (1997). Physical basis of a protein-DNA recognition code. *Current Opinion in structural biology* **7,** 117-125.

Ebright, R.H., Cossart, P., Gicquel Sanzey, B. & Beckwith, J. (1984). Mutations that alter the DNA sequence specificity of the catabolite gene activator protein of E. coli. *Nature* **311,** 232-235.

Elrod-Erickson, M., Rould, M.A., Nekludova, L. & Pabo, C.O. (1996). zif268 protein-DNA complex refined at 1.6 A: a model system for understanding zinc finger-DNA interactions. *Structure* **4,** 1171-1180.

Goodrich, J.A., Schwartz, M.L. & McClure, W.R. (1990). Searching for and predicting the activity of sites for DNA binding proteins: compilation and analysis of the binding sites for Escherichia coli integration host factor (IHF). *Nucleic Acids Res* **18,** 4993-5000.

Kandonaga, J.T., Jones, K.A. & Tjian, R. (1986). promoter-specific activation of RNA polymerase II transcription by sp1. *TIBS* **11,** 20-23.

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

Kim, J.L., Burley, S.K. (1994). 1.9 A resolution refined structure of TBP recognizing the minor groove of TATAAAAG. *Nat Struct Biol* **1,** 638-653.

Mandel-Gutfreund, Y., Schueler, O. & Margalit, H. (1995). Comprehensive analysis of hydrogen bonds in regulatory protein DNA-complexes: in search of common principles. *J. Mol. Biol.* **253,** 370-382.

Mandel-Gutfreund, Y., Margalit, H. (1998). Quantitative parameters for amino acid-base interaction: implications for prediction of protein-DNA binding sites. *Nucleic Acids Res* **26,** 2306-2312.

Mandel Gutfreund, Y., Margalit, H., Jernigan, R.L. & Zhurkin, V.B. (1998). A role for CH...O interactions in protein-DNA recognition. *J Mol Biol* **277,** 1129-1140.

O'Neill, M.C. (1989). Consensus methods for finding and ranking DNA binding sites. Application to Escherichia coli promoters. *J Mol Biol* **207,** 301-310.

Parkinson, G., Wilson, C., Gunasekera, A., Ebright, Y.W., Ebright, R.E. & Berman, H.M. (1996). Structure of the CAP-DNA complex at 2.5 angstroms resolution: a complete picture of the protein-DNA interface. *J Mol Biol* **260,** 395-408.

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

Schneider, T.D., Stormo, G.D., Gold, L. & Ehrenfeucht, A. (1986). Information content of binding sites on nucleotide sequences. *J Mol Biol* **188,** 415-431.

Suzuki, M., Gerstein, M. & Yagi, N. (1994). Stereochemical basis of DNA recognition by Zn fingers. *Nucleic Acids Res* **22,** 3397-3405.

Suzuki, M., Yagi, N. (1994). DNA recognition code of transcription factors in the helix-turn-helix, probe helix, hormone receptor, and zinc finger families. *Proc Natl Acad Sci U S A* **91,** 12357-12361.

Takeda, Y., Sarai, A. & Rivera, V.M. (1989). Analysis of the sequence-specific interactions between Cro repressor and operator DNA by systematic base substitution experiments. *Proc Natl Acad Sci U S A* **86,** 439-443.

Thieffry, D., Salgado, H., Huerta, A.M. & Collado Vides, J. (1998). Prediction of transcriptional regulatory sites in the complete genome sequence of Escherichia coli K-12. *Bioinformatics* **14,** 391-400.