Structure-Based Prediction of DNA Target Sites by Regulatory Proteins

Hidetoshi Kono and Akinori Sarai*
Tsukuba Life Science Center, The Institute of Physical & Chemical Research (RIKEN), Ibaraki, Japan

ABSTRACT **Regulatory proteins play a critical role in controlling complex spatial and temporal patterns of gene expression in higher organism, by recognizing multiple DNA sequences and regulating multiple target genes. Increasing amounts of structural data on the protein–DNA complex provides clues for the mechanism of target recognition by regulatory proteins. The analyses of the propensities of base–amino acid interactions observed in those structural data show that there is no one-to-one correspondence in the interaction, but clear preferences exist. On the other hand, the analysis of spatial distribution of amino acids around bases shows that even those amino acids with strong base preference such as Arg with G are distributed in a wide space around bases. Thus, amino acids with many different geometries can form a similar type of interaction with bases. The redundancy and structural flexibility in the interaction suggest that there are no simple rules in the sequence recognition, and its prediction is not straightforward. However, the spatial distributions of amino acids around bases indicate a possibility that the structural data can be used to derive empirical interaction potentials between amino acids and bases. Such information extracted from structural databases has been successfully used to predict amino acid sequences that fold into particular protein structures. We surmised that the structures of protein–DNA complexes could be used to predict DNA target sites for regulatory proteins, because determining DNA sequences that bind to a particular protein structure should be similar to finding amino acid sequences that fold into a particular structure. Here we demonstrate that the structural data can be used to predict DNA target sequences for regulatory proteins. Pairwise potentials that determine the interaction between bases and amino acids were empirically derived from the structural data. These potentials were then used to examine the compatibility between DNA sequences and the protein–DNA complex structure in a combinatorial "threading" procedure. We applied this strategy to the structures of protein–DNA complexes to predict DNA binding sites recognized by regulatory proteins. To test the applicability of this method in target-site prediction, we examined the effects of cognate and noncognate binding, cooperative binding, and DNA deformation on the binding specificity, and predicted binding sites in real promoters and compared with experimental data. These results show that target binding sites for several regulatory proteins are successfully predicted, and our data suggest that this method can serve as a powerful tool for predicting multiple target sites and target genes for regulatory proteins. Proteins 1999;35:114–131.** **© 1999 Wiley-Liss, Inc.**

Key words: **protein–DNA recognition; specific binding; binding site prediction; statistical potential; spatial distribution**

INTRODUCTION

Gene expression in higher organisms is controlled by a wide variety of regulatory proteins, which bind to specific sites of DNA. Rapidly increasing structural data on the protein–DNA complex provide a rich source of information about the interactions between amino acids and base pairs at the atomic level. These studies have revealed several classes of DNA-binding proteins with distinct DNA-binding motifs and a variety of interactions between proteins and DNA. DNA-binding proteins often use certain motifs such as the helix-turn-helix to achieve a sequence-specific fit with DNA, and to interact with base pairs directly. Thus, the direct reading of DNA sequence by amino acids seems to play an important role in the recognition process. Seeman et al. proposed that the formation of "point contacts," particularly double H bonds, between amino acids and base pairs, will be required for amino acids to discriminate different bases.[1] The structural data of protein–DNA complex indeed show some frequently occurring interactions such as Asn–A and Lys–G. However, the structural analysis also shows considerable degrees of redundancy in the specific interactions between amino acids and bases; the same amino acids often interact with different bases and vice versa.[2] Thus, codelike rules are not likely to exist for protein-DNA recognition,[3] although some rules emerged for members of a single structural family or for a group of families that interact in similar ways with DNA.[4,5] Furthermore, protein–DNA binding is usually accompanied by certain conformational changes or deformation of protein[6,7] and DNA.[8–10] Therefore, conformational flexibilities of protein and DNA should be another important factor determining a good structural

*Correspondence to: Akinori Sarai, Tsukuba Life Science Center, The Institute of Physical & Chemical Research (RIKEN), 3–1–1 Koyadai, Tsukuba, Ibaraki 305, Japan. E-mail: sarai@rtc.riken.go.jp

Received 27 August 1998; Accepted 19 November 1998

match on complexation. In the case of c-Myb oncoprotein, the flexibility and stability of its DNA-binding domain have been shown to affect the DNA-binding activity.[11] The flexibility of DNA is sequence dependent,[12] and it can affect the binding affinity with protein as well.[13]

Given the complexity of protein–DNA recognition, how can we explain the specific recognition of particular sequences by regulatory proteins, and how can we predict target sites recognized by proteins? There are several approaches for predicting binding sites by regulatory proteins. First, the most frequently used approach is the "sequence-based" method, which uses a profile of sequence homology from consensus binding sequences for DNA-binding proteins.[14–16] Also, DNA libraries for binding by given sequence variants of DNA-binding proteins has been used to derive binding-site preferences.[17] Second, the "energy-based" approach uses thermodynamic data of protein binding to DNA with systematic mutations.[18–20] This method has been applied to the prediction of target genes of c-Myb oncoprotein.[21] Third, the "structure-based" approach uses information from structural data on the protein–DNA complex. The structures of more than 130 protein–DNA complexes derived from X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy are now registered in the Protein Data Bank (PDB).[22] They contain a large amount of information on the interaction between amino acids and base pairs. The redundancy and conformational variability in the interactions between amino acids and bases imply that for given geometry of a protein complexed with DNA, it is not straightforward to predict the extent of specificity for particular pair of bases and amino acids. That is, simple rule-based methods are not sufficient for the accurate prediction of DNA sequence recognition by proteins. This is similar to the situation in proteins, where complicated interactions among amino acids make rule-based methods impractical for the prediction of structure from sequence. Statistical analysis of pairwise contacts between amino acids found in the structural database of proteins enables one to derive empirical interaction potentials.[23–27] Such information has been used for screening potential sequences that fit with a given structure [three-dimensional–one-dimensional (3D–1D) matching],[28] or for screening structures that fit with a given sequence (threading method).[29,30] These methods have been successfully applied to the prediction of the 3D structure of proteins. We surmise that the distribution of amino acids around bases found in the structures of protein–DNA complex can be used to derive empirical potentials for their interactions, and that the potential can be used to predict DNA target sites for DNA-binding proteins in a combinatorial "threading" procedure, because determining DNA sequences that bind to a particular protein structure should be similar to finding amino acid sequences that fold into a particular structure.

In this article, we apply the above structure-based approach for the prediction of targets of regulatory proteins. We first describe the analyses of the interaction propensities of base and amino acid, based on these structural data, and show that, in spite of the absence of simple codelike correspondence between bases and amino acids, the interaction propensities do exist. Then, we describe the analyses of spatial distributions, including radial and angular distributions, of amino acids around bases, and examine the structural variability within the specific interactions. The present results show remarkable structural flexibility in the specific interactions, with substantial spatial distributions of amino acids with respect to bases, indicating that amino acids with many different geometries can form a similar type of interaction with bases. Such spatial distributions suggest that the structural data can be used to transform the distribution into empirical interaction potentials, which may be used to predict target sites for regulatory proteins. Thus, in the latter half of this article, we describe the derivation of the interaction potential, and its application to the prediction of target binding sites by regulatory proteins. To test the applicability of this method in target-site prediction, we examined the effects of cognate and noncognate binding, cooperative binding, and DNA deformation on the binding specificity, and predicted binding sites in real promoters and compared with experimental data. We demonstrate here that this technique enables us to successfully predict target binding sites for several regulatory proteins, suggesting that this method in combination with other methods can serve as a powerful tool for predicting multiple target sites and target genes for regulatory proteins.

METHODS

Data Set Used for the Analyses of Base–Amino Acid Interaction and for the Derivation of Contact Potentials

We have selected 52 protein–DNA complex structures from PDB by excluding redundant and poor resolution data (resolution ≤ 3.2 Å), shown in Table I. For each complex, we count the number of contacts between amino acids and bases. The contacts were counted if the centers of two atoms were within 3.5Å. Contacts between atoms from the same residues were excluded from the consideration. For the analyses of base–amino acid interaction, we excluded the human immunodeficiency virus type 1 reverse transcriptase (RT) structure (PDB code:1HMI) because of the lack of side-chain coordinates.

Base Propensities of Amino Acids

We define the base propensities of amino acids, P_{ij}, as follows,

$$P_{ij} = \frac{N_{ij}}{\sum_j N_{ij}} \Bigg/ \frac{T_j}{\sum_j T_j}, \tag{1}$$

where N_{ij} is the total number of contacts between base i and amino-acid residue j, and T_j is the total number of residue j in the whole data set. The contacts are classified into four groups: (1) base and amino-acid side-chain; (2) DNA backbone and side-chain; (3) base and protein backbone (main-chain); and (4) DNA backbone and protein

TABLE I. Protein–DNA Complexes Used for the Analyses of Base–Amino Acid Interaction and for the Derivation of Contact Potentials[†]

PDB code	Complex	Resolution (Å)
1AAY	Zif268 (3 zinc fingers)	1.6
1APLd	MATα2	2.7
1BERa	CAP	2.5
1BHMa	Endonuclease BamHI	2.2
1CDW	Human TBP core domain	1.9
1CMA	Met repressor-operator	2.8
1D66a	GAL4	2.7
1DCTa	DNA (C-5) methylase	2.8
1ERI	EcoRI	2.7
1FJLa	Paired homeodomain (PAX)	2.0
1FOSe, f	C-FOS/C-JUN	3.05
1GAT	GATA-1	NMR
1GDTa	Recombinase γ-δ resolvase	3.0
1HCQa	Estrogen receptor	2.4
1HCR	hin recombinase	1.8
1HDDc	Engrailed homeodomain	2.8
1HMI	HIV-1 reverse transcriptase	3.0
1HRY	Human SRY	NMR
1IHF	IHF	2.5
1KLN	DNA polymerase I	3.2
1LATb	Glucocorticoid receptor mutant	1.9
1LCC	lac repressor	NMR
1LMB4	λ repressor	1.8
1MDYa	MyoD bHLH domain	2.8
1MHT	HhaI methyltransferase	2.8
1MSE	Myb	NMR
1NFKa	NF-κB p50	2.3
1OCT	Oct-1 POU homeodomain	3.0
1PARb	Arc repressor	2.6
1PDN	Paired domain (PAX)	2.5
1PER1	434 repressor (OR3)	2.5
1PNR	PurR	2.7
1PUEe	PU.1 ETS-domain	2.1
1PVIb	PvuII	2.8
1PYIa	PPR1	3.2
1RPE1	434 repressor (OR2)	2.5
1RVAa	Eco RV	2.0
1SRS	Serum response factor core	3.2
1STW	ETS1	NMR
1TROa	Trp repressor	1.9
1TSRb	p53 tumor suppressor	2.2
1UBD	Human YYI	2.5
1VAS	Excision repair enzyme	2.75
1YRNa	MATa1/MATα2	2.5
1YSA	GCN4	2.9
1YTF	TFIIA and TBP	2.5
1ZQA	DNA polymerase β (POL B)	2.7
2BOP	Bovine papillomavirus-1 E2	1.7
2DNJ	DNase I	2.0
2DRPa	Tramtrack protein	2.8
2GLI	GLI-DNA complex	2.6
3CRO1	434 Cro (OR1)	2.5

[†] The fourth letter of PDB code shows the chain name.

backbone, and the propensities are calculated for each group. If P_{ij} is smaller than 1, the amino acid is disfavored and if P is greater than 1, the amino acid is favored for the interaction.

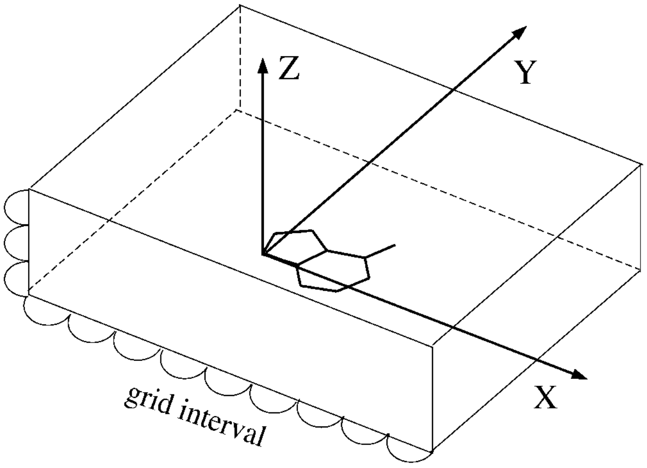


Fig. 1. The definition of the cubic grid. Amino acids within the box are considered for creating potentials.

Radial and Angular Distributions for H-Bondlike Interactions and Hydrophobic Interactions

We have analyzed distributions of distances between H-bond donors or acceptors in proteins and their counterpart atoms in bases. H-bond donors include A N6, G N2, C N4, protein backbone nitrogen atoms, Cys SG, His NE2, His ND1, Lys NZ, Asn ND2, Gln NE2, Arg NE, Arg NH1, Arg NH2, Ser OG, Thr OG1, Tyr OH and Trp NE1, whereas possible acceptors are A N7 and N3, T O4 and O2, G N7, N3 and O6, C O2, protein backbone oxygen atoms and Asp OD1, Asp OD2, Cys SG, Glu OE1, Glu OE2, Gln OE2, His ND1, Met SD, Asn OD1, Gln OE1, Ser OG, Thr OG1, and Tyr OH.

We have also examined radial distributions of hydrophobic amino-acid atoms around bases. The atoms that can participate in hydrophobic interactions include all CB, CG, CG1, CG2, CD1, CD2, CE, and CZ in proteins, and C5M of T.

Angular distributions are also investigated for the above atoms in amino acids against atoms in bases. Angle is defined between BB-B and B ··· X vectors where B is a donor or acceptor in bases, BB is its antecedent atom, and X is a donor or acceptor in proteins. For atoms N7 and N3 of A and G, the centers between C5 and C8, and between C4 and C2 are used as BB, respectively.

Derivation of Statistical Potentials

For a pair of base *a* and amino acid *b* at grid point *s*, the potential is given by the following expressions,[23]

$$\Delta E^{ab}(s) = -RT \ln \left[\frac{f^{ab}(s)}{f(s)}\right] \qquad (2)$$

$$f^{ab}(s) = \frac{1}{1 + m_{ab} w} f(s) + \frac{m_{ab} w}{1 + m_{ab} w} g^{ab}(s), \qquad (3)$$

where m_{ab} is the number of pairs *ab* observed, *w* is the weight given to each observation, *f(s)* is the relative

TABLE II. The Number of Amino Acids and Bases in the Data Set and Contact Numbers Between Amino Acids and Bases

		N_{ij}[a]																			
		DNA–amino acid[b]				Base–s.c.				Base–p.b.				D.b.–s.c.				D.b.–p.b.			
Name	T_j[c]	A	T	G	C	A	T	G	C	A	T	G	C	A	T	G	C	A	T	G	C
	530	5	5	7	6	0	1	1	0	0	1	1	2	4	1	2	2	2	3	4	4
	468	6	7	1	4	3	2	0	0	0	0	0	0	3	4	0	3	0	3	1	1
Ile	500	9	8	13	3	4	3	3	1	0	1	2	0	6	3	7	2	0	3	3	0
Leu	683	3	3	10	4	3	1	2	2	0	0	0	0	0	2	4	2	0	0	4	2
Met	178	3	6	3	3	0	0	0	1	0	4	0	1	2	2	2	1	2	1	1	2
Phe	300	7	11	5	7	4	8	1	3	0	0	0	1	4	6	4	2	0	2	0	2
Trp	79	3	2	7	4	0	0	2	2	0	0	0	0	3	2	5	4	2	0	1	0
Tyr	286	9	17	12	15	2	6	1	2	2	0	1	1	3	14	10	10	4	0	2	3
Arg	569	81	85	93	68	31	31	49	22	6	5	0	0	51	52	45	51	14	9	4	7
Lys	647	39	59	48	40	5	10	28	9	0	2	1	8	32	42	17	21	4	10	4	8
Asn	362	35	28	21	14	24	16	8	7	0	1	1	1	14	14	10	8	2	2	5	3
Asp	366	6	2	4	6	2	0	2	3	0	0	0	1	5	2	2	2	0	1	0	0
Gln	317	27	25	17	10	11	9	7	8	0	1	1	1	14	12	10	2	2	8	3	0
Glu	573	4	6	5	16	0	4	0	10	0	0	1	2	4	0	3	4	0	2	1	2
His	173	11	13	8	4	1	4	3	1	0	1	0	0	10	7	5	2	5	4	0	1
Ser	469	16	32	25	17	2	14	8	2	1	0	1	0	13	17	15	12	5	8	10	5
Thr	416	23	33	22	19	7	9	4	3	1	1	1	0	15	20	18	16	9	9	5	7
Gly	467	10	20	20	14	0	0	0	0	3	7	11	5	0	0	0	0	8	15	11	11
Pro	300	5	3	3	7	3	1	0	0	0	0	0	0	1	2	3	2	1	1	1	6
Cys	136	1	1	0	8	1	0	0	3	0	0	0	1	0	1	0	4	0	0	0	4
M_i[c]		466	474	360	359																

[a] N_{ij} is the number of contacts between amino acid i and base j which are within 3.5 Å.

[b] DNA-amino acid shows the number of contacts between bases (including DNA backbones) and amino acids; base–s.c. is between bases (excluding DNA backbones) and side chains; base–p.b. is between bases and protein backbones; D.b.–s.c. is between DNA backbones and side chains; D.b.–P.b. is between DNA backbones and protein backbones.

[c] T_j and M_i are the total numbers of amino acid j and that of base i in the data set, respectively.

frequency of occurrence of any amino acids at grid point s against any bases, and $g^{ab}(s)$ is the equivalent relative frequency of occurrence of amino acid a against base b. R and T are gas constant and absolute temperature, respectively. Figure 1 shows the definition of coordinate system, where the origin is the N9 atom for A or G and N1 for C or T, X-axis is N9–C4 vector for A or G and N1–C2 vector for C or T. The Y-axis is on the same plane as that defined by a purine or a pyrimidine and perpendicular to the X-axis, in a right-hand reference frame. We considered amino acids that are within a box shown in Figure 1. The box was divided into cubic boxes, where the potential is defined. We examined various sizes of cubic boxes with grid interval from 1 to 6 Å. The size of the whole box was also varied independently between $|x| = |y| = 7.5$ Å, $|z| = 2$Å and $|x| = |y| = 15$ Å, $|z| = 8$Å.

Energy Calculations

The sum of the potentials for a sequence with a given length is defined as the energy for the sequences. The length of DNA sequence depends on the interface size of the binding proteins. The energies for the sequences in the crystal structures were characterized by their Z-scores against random sequences. The Z-score is defined as $(X - m)/\sigma$, where X is the potential sum for a sequence, m is the mean of X, and σ is the standard deviation. Z-scores for random sequences usually range from -4 to 4.

RESULTS AND DISCUSSION

Interaction Preference of Amino Acids with DNA

We have analyzed 52 protein–DNA complex structures shown in Table I, counting the number of contacts between different parts of protein and DNA. We classified these contacts of each amino acid against DNA into four groups according to whether protein backbones or amino-acid side-chains interacting with DNA bases or backbones, and the results are shown in Table II. We calculated propensities for base–side-chain and DNA backbone–side-chain.

Base propensities of amino acids are shown in Fig. 2a–d. These propensities indicate that if the value is greater than 1.0, the amino acid tends to be favored for a given base. Comparison of these figures shows that purines (A and G) have strong preferences for only a few amino acids, whereas pyridines (T and C) show a large variation in the side-chain preference. For A, Asn is the most favorable amino acid, followed by Arg and Gln. These three amino acids also have favorable interactions with T. In addition, Ser, Thr, His, Tyr, and Phe often interact with T. Guanine has the distinct preference for Arg and Lys compared with the other amino acids. Cytosine tends to prefer Arg, Lys, Asn, Gln, Trp, Cys, and Glu, but its preference is not distinctive except Arg. When the amino acids that have more than 10 contacts with bases are considered, only charged or polar amino acids remain for G and C, and they interact with the bases almost always from the major

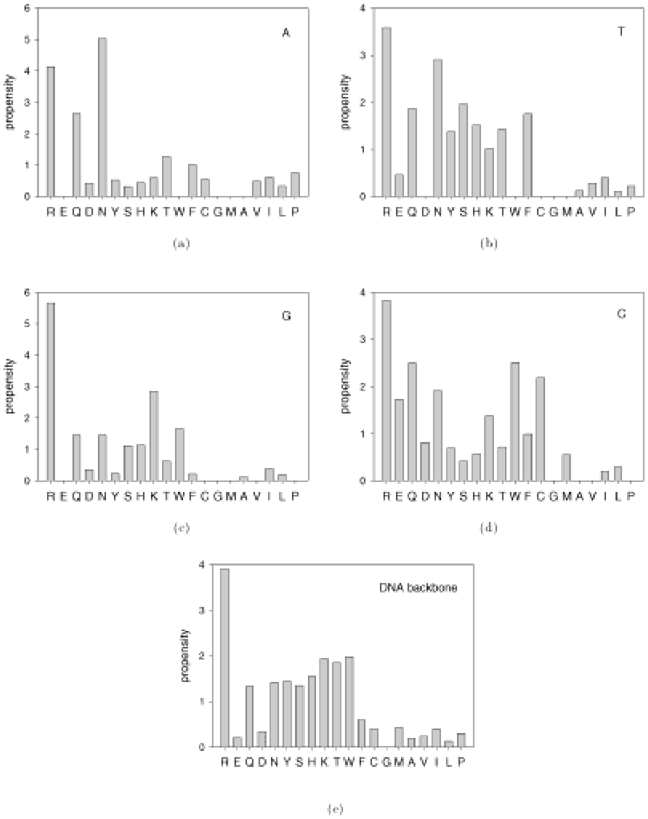


Fig. 2. Base propensities, A (**a**), T (**b**), G (**c**), and C (**d**) of side-chains and DNA backbone propensity of side-chains (**e**) are plotted against every amino acid, shown by a one-letter code. Amino acids are ordered from left to right according to their values of hydration energy in the native state of proteins.[32]

groove. On the other hand, for A, Arg and Thr are often observed in the minor groove. For T, Phe, Arg, and Lys preferentially interact with base from the minor groove. These base propensities of amino acids basically support the classical view,[1] which is based on the donor and acceptor patterns of bases, i.e., in the major groove, one donor (N6) and acceptor (N7) of A, two acceptors (O6 and N7) of G, one acceptor (O4) of T, and one donor (N4) of C; in the minor groove, one acceptor (N3) of A, one acceptor (N3) and donor (N2) of G, and one acceptor (O2) each of C and T. Arg and Lys are often around G because they can form bidentate contacts with O6 and N7 of G. Also, Asn and Gln can interact similarly with A by forming bidentate contacts with N7 and N6 of A. It is interesting to note that the propensities of A, T, and C for Lys are rather low, whereas all the bases favor Arg in spite of resemblance in their functional groups. Also, the preference of Gln for A is lower than Asn, even though their only difference is the length of side-chain.

Mandel-Gutfreund et al.[2] analyzed 28 protein–DNA complexes in terms of H bonds, and found that the interactions between amino-acid side-chains and DNA backbone constitute about half of the interactions. With a larger data set, we also found that the contacts between the amino-acid side-chains and DNA backbone show a similar trend. The DNA backbone (phosphate plus sugar ring of DNA) propensity of the side-chain is shown in Fig. 2e. As expected, the DNA backbone favorably interacts with amino acids with positive charge. The DNA backbone has the strongest propensity toward Arg, whereas showing lower propensity toward Lys in spite of the similar character in their functional group. Also, the backbones show favorable interactions with polar amino acids such as Ser, Thr, His, Asn, and Gln. On the other hand, the backbones exhibit low propensity toward those amino acids with negative charge such as Asp and Glu. Likewise, the backbones rarely interact with hydrophobic amino acids except for Trp.

Spatial Distributions of Amino Acids Around Bases

We have analyzed the structures of the protein–DNA complex to examine how amino acids are distributed in space around each base. Figure 3 shows some amino-acid distributions around bases. Arg is often observed around A, and the distribution of its functional group is more spread around A than around G (Fig. 3a and d). It is interesting that Arg has strong preference in the minor groove of A, interacting with the N3 atom. On the other hand, Lys rarely appears around A (Figs. 2a and 3b). As shown in Fig. 3d, Arg clusters in the major groove of G, mostly forming double H bonds with O6 and N7 atoms of G. This figure also shows that a wide range of Cα positions of Arg is possible to form similar H-bond interactions. Also, Lys clusters in the major groove of G to form similar H bonds with various conformations (Fig. 3e).

Adenine and guanine are both purines, but the rank orders of amino acid in base propensities are different. Adenine has strong preference with Asn, forming double H bonds with N7 and N6 (Fig. 3c). However, Asn also shows other types of interactions with one H bond and interactions in the minor groove (H bonds with N3) of A. It is interesting that amino-acid distributions around A are different from those around A of adenylate. Asn often appears around A of DNA, whereas not around that of adenylate.[31] This may be because of the difference in the manner of interaction. In protein–DNA complexes, A rarely stacks into the protein interior like that of adenylate because of the steric hindrance.

Ser and Thr have an OH group, which can be an H-bond donor as well as an acceptor, e.g., it serves as a donor for N7 of G (Fig. 3f) and O4 of T (Fig. 3g), or as an acceptor for O6 of G (Fig. 3f). Thr also has hydrophobic interactions between its CH$_3$ group and C5M of T (Fig. 3g). Negatively charged Glu clusters around the potential H-bonding atoms (N4) of C, as shown in Fig. 3j.

As to hydrophobic interactions, T is expected to have hydrophobic interactions around C5M in the major groove. In fact, we did find contacts with hydrophobic amino acids (Fig. 3g–i), but their frequencies are rather low (Fig. 2b). Almost all of Phe residues involved in the interaction with T are in the minor groove (Fig. 3h), whereas Tyr are in the major groove (Fig. 3h). Tyr residues are also found stacked on T (Fig. 3h). There is no clustering of hydrophobic residues in the region around H-bond forming atoms. The lack of strong preference against hydrophobic side-chains may be because of the nature of hydrophobic interactions, which is less specific than H bonds in terms of the counterpart atoms. Therefore, these results imply that electrostatic interactions and H bonds between bases and polar or charged amino acids are mainly used to achieve specific interactions. We found that the base propensities of amino acids correlate well with the free energy of amino-acid hydration. This can be seen in Figure 2, where amino acids are ordered according to their values of hydration energy in the native state of proteins.[32] This energy also represents a kind of polarity of amino acids. This result means that the number of contacts between bases and side-chains correlates with the amino-acid tendency for exposure to the solvent.

Radial and Angular Distributions for H-Bond and Hydrophobic Interactions

We have investigated radial and angular distributions of atoms of amino acids around potential H-bond forming and hydrophobic atoms of bases. Some of the radial distributions are shown in Figure 4 (histograms for G N7, G O6, C N4, and T C5M). The radial distributions of H-bond forming atoms of amino acid around their counterparts of bases have a peak at 2.6–3.0 Å, which is similar to the case of H bonds in proteins.[33,34] As a control, the distributions of non-H-bond atoms show no such peaks. Figure 4d shows a histogram of distribution of amino-acid hydrocarbons around C5M (to which a methyl group is attached) of T. There is a peak near 3.7 Å and a plateau in a range up to 5 Å. This is due to the van der Waals contacts and hydrophobic interactions around the methyl group (see Fig. 3g–i), since the radii of carbon atoms are 1.8–2.0 Å. The distribution of the same atoms around O4 of T shows no such features.

Figure 5 shows the angular distributions of atoms around potential H-bond-forming and hydrophobic atoms of bases. Earlier studies found that H bonds involving side-chains[35] and N-H ··· O=C H bonds in small crystal structures[36] tend to coordinate with the lone pairs of acceptor. A bond perfectly lined up with the lone pair would give an angle 120° at the acceptor if its electron shells have sp^2 trigonal planar hybridization. Actually, the angular distribution around G O6 has a clear peak at 130° (Fig. 5b). H bonds in proteins usually favor a linear configuration with a major peak at 180°. However, some donors or acceptors already form H bonds in base pairing. That is why we observed a peak in a direction of remaining lone pairs (Fig. 5b). Also, N7 of G has a peak at 160°, which reflects the direction of a lone pair (Fig. 5a). In the same way, N4 of C shows a peak at 120°. When the distance cutoff for H bond is shortened from 3.5 to 3.0 Å, peaks of the distributions became clearer. This may be because of the effect of excluded atoms of proteins by the 3 Å cutoff, which form H bonds with other bases and tend to obscure the distribution. The angular distribution of hydrocarbon atoms around C5M of T clearly shows a different pattern from those of H-bond donors and acceptors. There is a broad distribution between 90° and 180°, indicating that hydrophobic interactions have no directional preference (Fig. 5d).

Structure-Based Potentials from Protein–DNA Complexes

The results in the preceding sections show that there are significant redundancy and remarkable structural flexibility in the interactions between amino acids and bases. The same amino acid can interact not only with different bases, but also with the same bases with different geometries. The coexistence of interaction propensity and variability indicates that the sequence recognition by regulatory proteins must be not only specific enough to discriminate specific sites from nonspecific sequences, but also flexible enough to recognize a set of multiple target sites with certain sequence variation to control multiple genes cooperatively. The spatial distributions of base–amino acid interactions suggest that the structural data can be used to transform the distribution into empirical interaction potentials, which may be used to predict target sites for regulatory proteins. Below, we describe the derivation of the interaction potentials.

We considered a total of 4,621 amino acid–base pairwise distributions for A, 4,991 for T, 3,968 for G, and 3,459 for C from the 52 protein–DNA complex structures to determine amino acid–base potentials. A set of pairwise potentials between DNA bases and Cα atoms of all amino acids was then empirically determined by a statistical analysis of the protein–DNA complex structures using a modification of the method of Sippl.[23] For any given amino acid–base pair, these potentials represent the mean force between the base and the flexible amino-acid side-chains, and they are related to the probability that a particular interaction exists in native structures. For example, consider the distribution of Asn around A, as shown in Fig. 3c. Based on such distributions, the potentials of amino acid–base pairs can be calculated using Cα positions of amino acids and projected onto the plane of base (see Fig. 6a for Asn-A and Fig.6b for Asp-A as a comparison). In Fig. 6a and b, the blue regions show that Asn frequently appears at coordinates (0, −6) and (9, −6) in the minor groove as well as (0, 9), (3, 9) and (6, 9) in the major groove, interacting with the N3 and N6 atoms of A. By contrast, the distribution for Asp does not show a strong positional preference around A, showing that although the side-chains of Asn and Asp may be similar, their occupancy patterns around A are dissimilar.

Assessment of Binding Site Prediction by Regulatory Proteins

Our ability to predict DNA binding sequences from patterns of atomic interaction between bases and amino acids is to be assessed by analyzing the energies of 14 crystal and one NMR protein–DNA structures exhibiting distinct characteristics (Table III). Random DNA sequences of certain length (see legend of Fig. 8) were

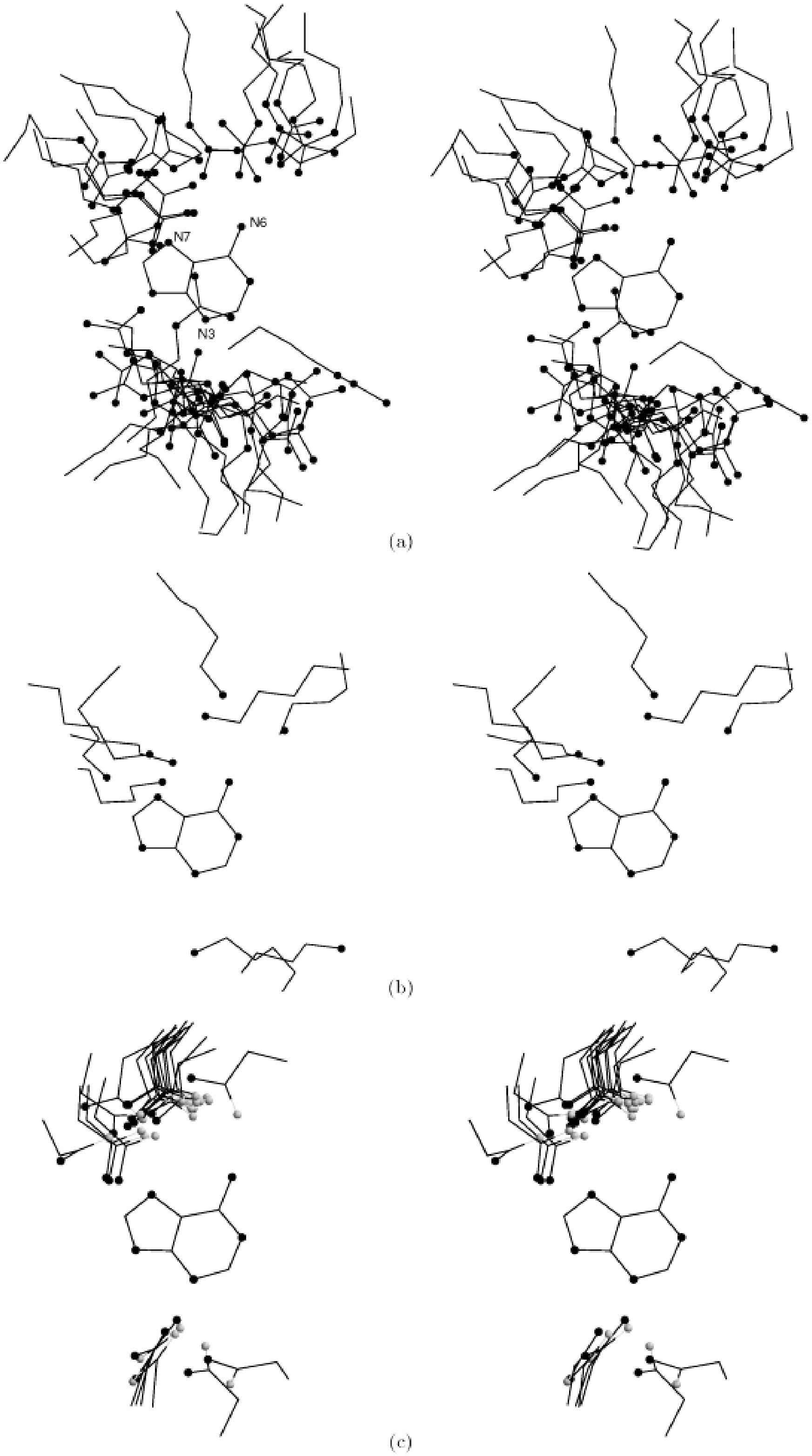


Fig. 3. Spatial distributions of amino acids around bases. Only base and side-chain are shown. Nitrogen and oxygen atoms are shown by filled circles and gray circles, respectively. **a:** Arg distribution around A. **b:** Lys distribution around A. **c:** Asn distribution around A. **d:** Arg distribution around G. **e:** Lys distribution around G. **f:** Ser distribution around G. **g:** Ser and Thr (in thick line) distributions around T. **h:** Phe (in thick line) and Tyr distributions around T. **i:** Leu, Ile, Val, Ala, and Phe distributions around T. **j:** Glu distribution around C.

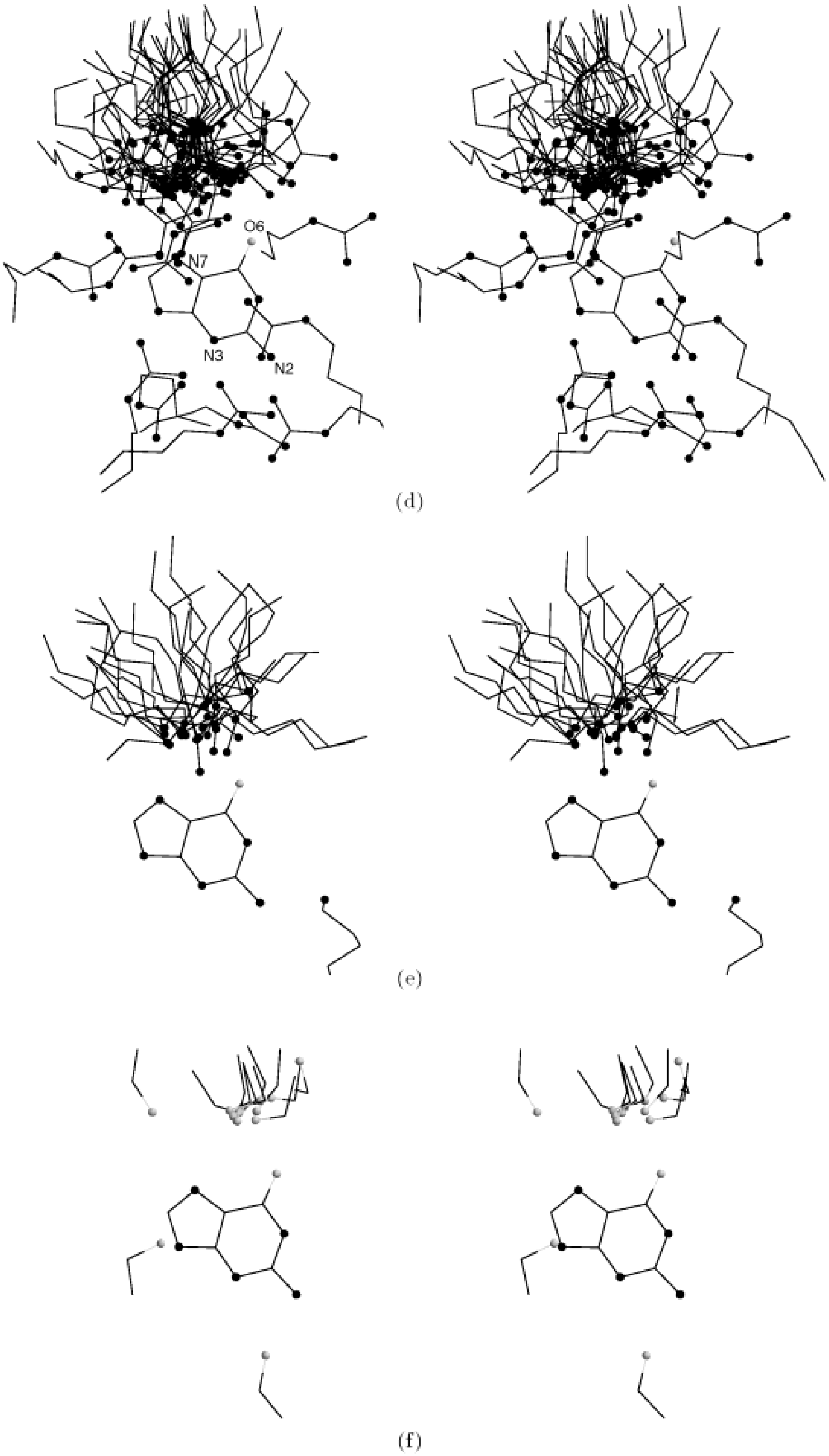


Figure 3. (Continued.)

threaded in a template of protein–DNA geometry, and the energies for the sequences were calculated. The energies for the sequences in the crystal/NMR structures are characterized by their Z-scores compared with 50,000 random sequences. The random sequences are composed of equal amount of A, T, G, and C. We also tested various base

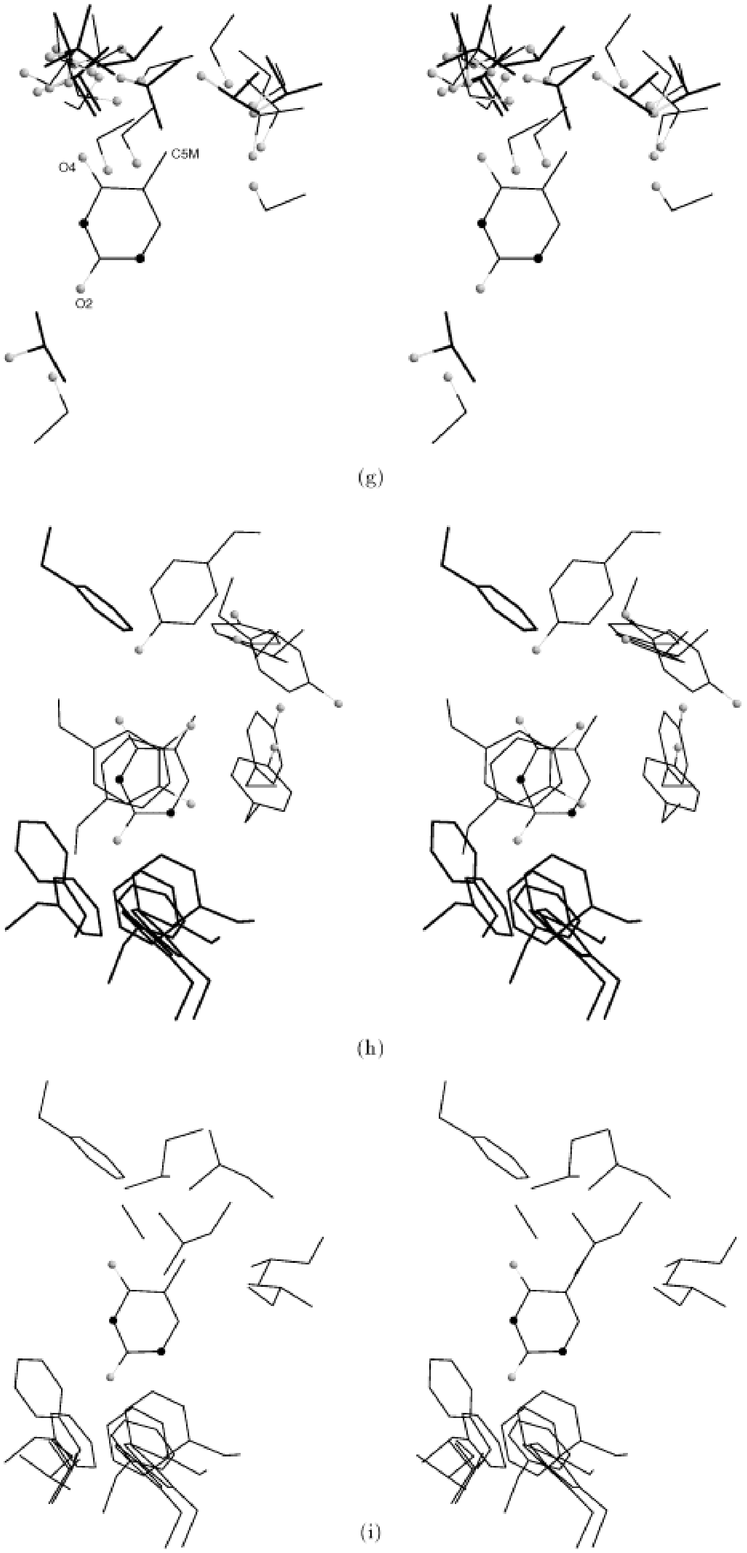


Figure 3. (Continued.)

compositions according the actual sequences in the regulatory regions of DNA, but it did not affect the results significantly. Thus, we used the random sequences with uniform base composition throughout.

To search for the optimal grid size and box size to obtain the best performance in binding site prediction, we independently changed the grid interval from 1 to 6 Å and box size from $|x| = |y| = 7.5$ Å and $|z| = 2$ Å to $|x| = |y| = 15$ Å and

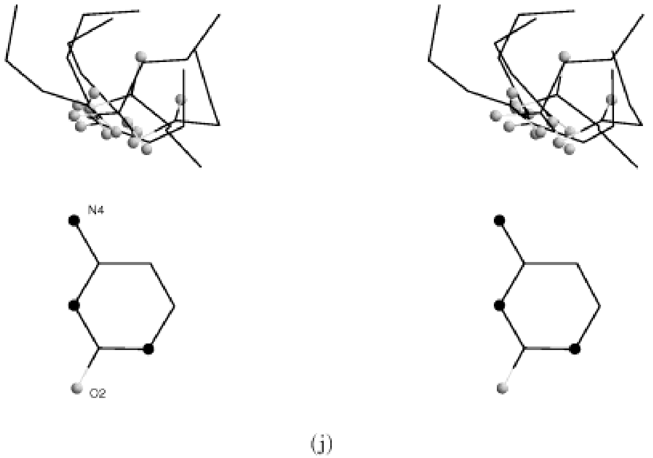


Figure 3. (Continued.)

$|z| = 8$ Å (Fig. 1). Figure 7 shows the best average Z-score at each grid interval for the proteins in Table III except 1LAT, which is noncognate binding. The cubic boxes with grid interval of 3.0 Å and the whole box size of $|x| = |y| = 13.5$ Å, $z = 6$ Å yielded the best average Z-score. The potentials for grid intervals smaller than 3.0 Å were erratic because of the shortage of data compared with the number of grid points. On the other hand, the potentials for larger grid intervals were too coarse to be used for the prediction. Thus, the grid interval of 3.0 Å was used in this study and yielded the average Z-score of -2.8 (Fig. 7).

If the potentials were derived from a database including the target complex itself, they would be unduly biased, and the Z-score for the sequence in the crystal would become highly favorable because the energy potentials, being derived from a rather small data set, would tend to overestimate the Z-score. Nevertheless, when potentials are derived in a more proper way from a database excluding the target complex, the recognition sequences are still readily detected, yielding quite favorable Z-scores (several examples are shown in Fig. 8a). Thus, even with only 52

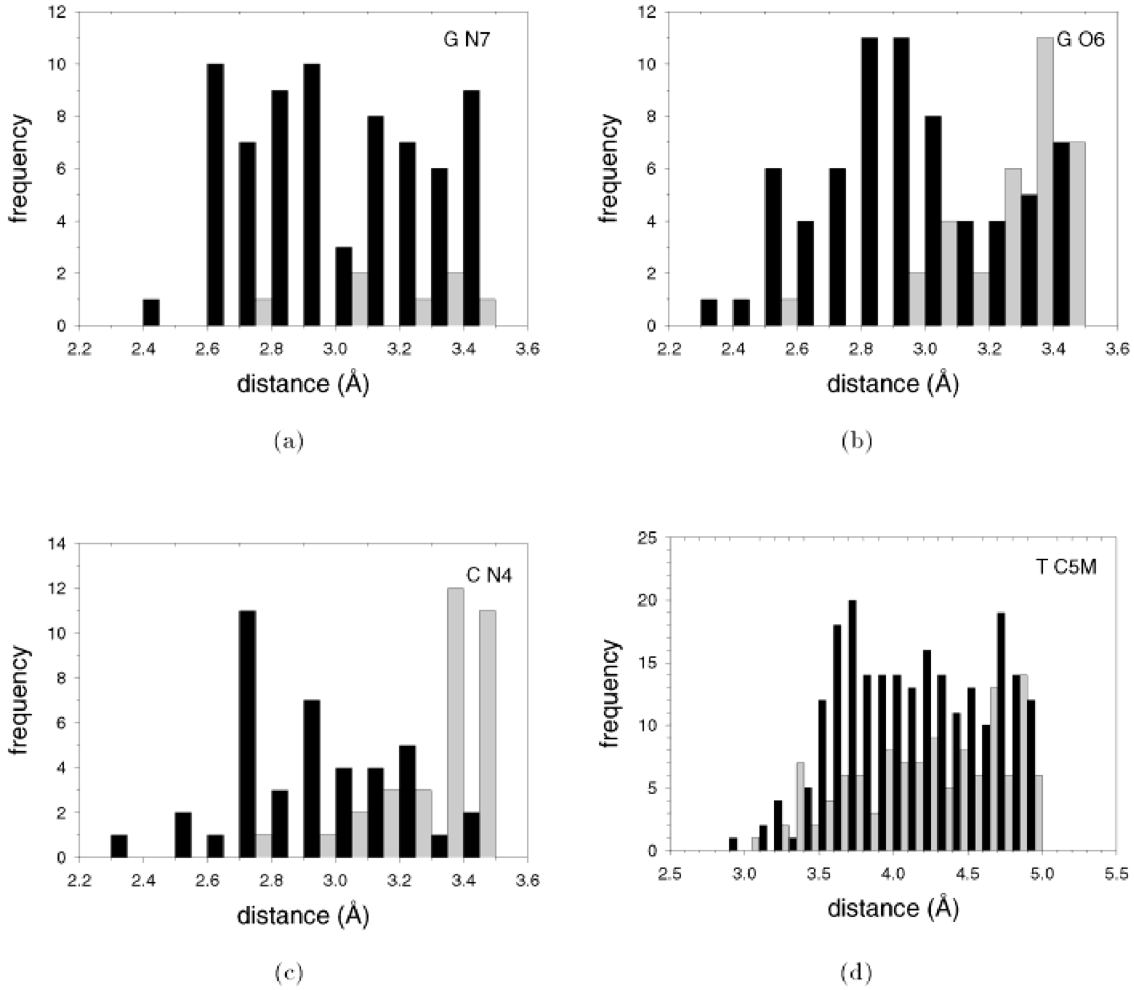
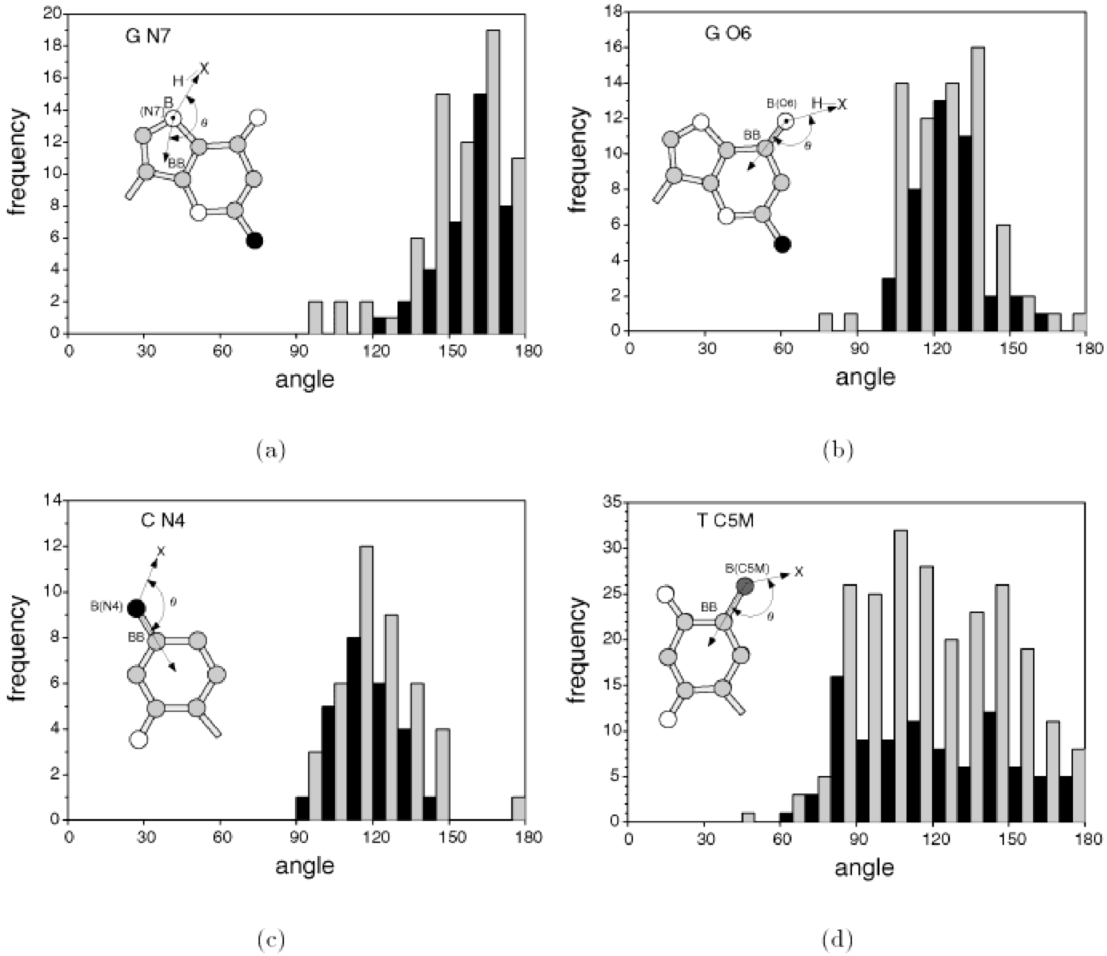


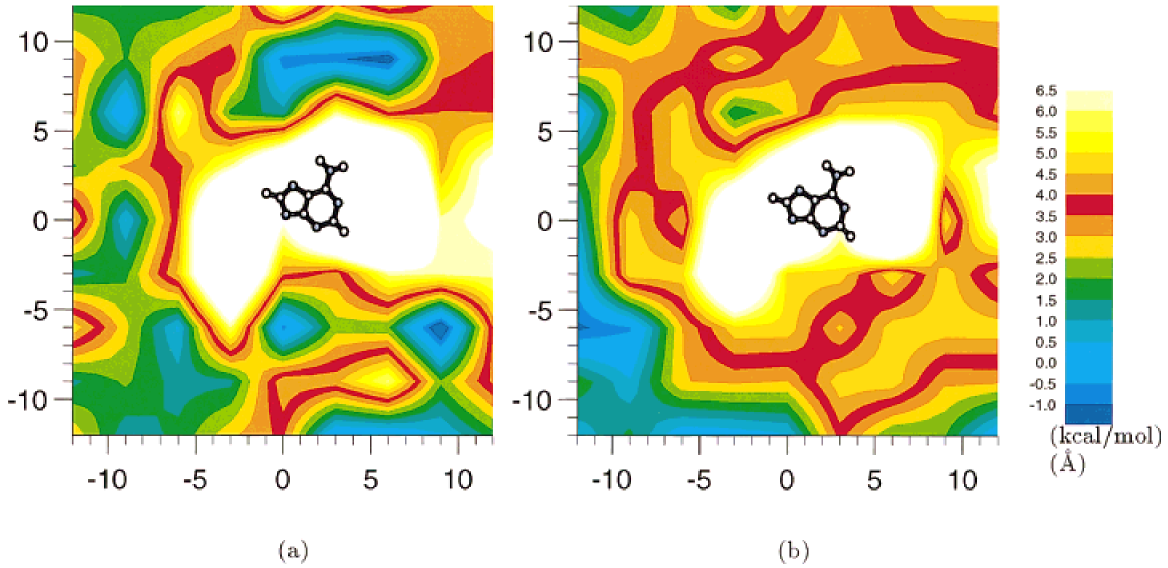
Fig. 4. Radial distributions of atoms in amino acids around H-bond-forming and hydrophobic atoms of bases. **a:** Distribution around N7 of G. **b:** Distribution around O6 of G. **c:** Distribution around N4 of C. **d:** Distribution around C5M of T. Gray bars in (a), (b), and (c) show the reference distribution of atoms that do not form H bonds. In (d), the distribution around O4 of T is plotted as a reference (gray bars).



(a) (b)

(c) (d)

Fig. 5. Angular distributions of atoms in amino acids around H-bond-forming and hydrophobic atoms of bases. **a:** Distribution around N7 of G. **b:** Distribution around O6 of G. **c:** Distribution around N4 of C. **d:** Distribution around C5M of T. Black and gray bars in (a), (b), and (c) show distributions of atoms that are within 3.0 and 3.5A, respectively, from the target atoms. In (d), black and gray bars are distributions of atoms that are within 4.0 and 5.0 A, respectively. Angle definition is shown as inset. Donor atoms of H-bond are shown by filled circles: N2 of G and N4 of C. Acceptor atoms are shown by open circles: O4 and O2 of T and O6, N7 and N3 of G, and O2 of C.



(a) (b)

Fig. 6. **a:** The potential map of Asn–A projected onto the purine plane. The map was created based on the Cα distributions of amino acids within a box of $|x| = |y| = 13.5$ Å, $|z| = 6$ Å with a grid interval of 3 Å. Cα atoms are often located in the blue-colored regions. Here, the temperature was set to 293K so that $RT = 0.582$ kcal/mol and w was an arbitrary number and set to 1/20 in Equation 1. **b:** The potential map of Asp–A.

TABLE III. Protein–DNA Complexes Used for Test†

PDB code	Protein name	DNA conformation	No. of contacts	No. of water-mediated contacts
1CGP	Catabolite gene activator[58]	B, bend by 90°	6 (0)	N.A.
1OCT	Oct-1 POU homeodomain[59]	B, bend by 3°	8 (0)	N.A.
1FJL	Paired homeodomain (S50Q)[55]	B, bend by 21°	7 (0)	5 (0)
1PDN	Paired domain[60]	B, bend by 20°	5 (5)	0 (0)
1TF3	TFIIIA (3 zinc fingers)[56]	intermediate A and B	15 (0)	N.A.
1AAY	Zif268 (3 zinc fingers)[46]	intermediate A and B	18 (0)	9 (0)
1MEY	Designed zinc finger[50]	B (major groove widened)	17 (0)	2 (0)
1APL	MATα2[45]	B	7 (3)	N.A.
1YRN	MATa1/MATα2[44]	B, bend by 60° a1	7 (0)	0 (0)
		α2	10 (4)	3 (0)
1PYI	PPR1[61]	B	4 (5)	0 (0)
1HCQ	Estrogen receptor[62]	B	8 (0)	N.A.
1GLU	Glucocorticoid receptor[38]	B (major groove widened)	4 (0)	0 (0)
1LAT	Glucocorticoid receptor mutant[37]	B	4 (0)	2 (0)
1NFK	NF-κB p50[41]	B	10 (2)	1 (0)
1SVC	NF-κB p50[40]	B, bend by 15°	12 (0)	2 (1)

†Contacts between amino acids and bases are counted within a distance of 3.5 Å (for water-mediated interactions, 3.0 Å). Contacts between atoms from the same residues and bases were excluded from the consideration. The numbers in parentheses denote the number of contacts between backbone of protein and bases per monomer.

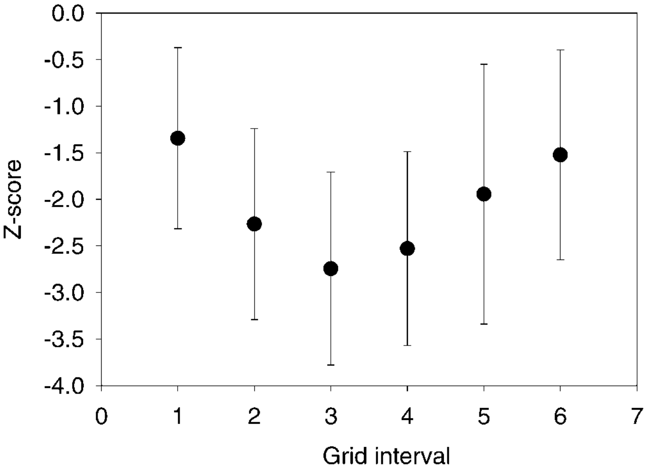


Fig. 7. Grid-interval dependence of averaged Z-score for proteins in Table III except for 1LAT. Z-scores of the co-crystallized DNA sequences are calculated against 50,000 random DNA sequences. The whole box size, in which amino acids are considered to produce potentials, was varied, and the best averaged Z-score was selected at every grid interval.

structures in the data set, the derived potentials will be useful for detecting probable binding sequences in the genome. One reason for the success and its high sensitivity may be that protein–DNA recognition really involves only a limited number of interaction types between bases and amino acids, and the available structures provide a sufficient sample of these.

It should be noted that the present method requires only Cα atom positions of proteins. Thus, high-resolution structures are not essential for detecting binding sites by the present method. In fact, we could detect the real binding site using the NMR structure of TFIIIA (see 1TF3 in Fig. 8a. The advantage of using Cα positions of amino acids in deriving the potentials is that we can collect sufficient numbers of samples from the database. The distribution of Cα is likely to reflect difference in amino acid character as shown in Figure 3. It can also include the effect of multiple side-chain conformations (see later discussion). On the other hand, the precision of the prediction may be sacrificed in this low-resolution approach. However, inclusion of more detailed structural information (e.g., orientation of Cα–Cβ vectors) would require larger numbers of structural data. We must await more accumulation of the data in future for further refinement of the prediction accuracy.

Specificity Difference in Cognate and Noncognate Binding

An interesting example for the structure-specificity relationship is the cognate and noncognate complex structures of nuclear receptor. Gewirth and Sigler[37] solved the crystal structure of an estrogen receptor (ER)-like DNA-binding domain (a glucocorticoid receptor (GR) DNA-binding domain altered by mutation) bound to the wrong type of half-site (a glucocorticoid response element; GRE) and revealed an interface resembling the specific interfaces of the GR or ER bound to their correct response elements. The subtle difference in binding specificity can be tested by the present analysis. When this noncognate complex (1LAT, Fig. 8b left) was used as a template, the GRE site was not detected ($Z = -0.1$). On the other hand, when a specific complex structure between GR and GRE (1GLU)[38] was used, a more favorable Z-score ($Z = -1.9$, Fig. 8b right) was obtained. Thus, this result indicates that our method can detect a subtle difference in binding specificity for target sites with some sequence variations.

The structural database also contains some structures of the same protein bound to different DNA sequences (e.g., the transcription factor, NF-κB). These structures were used to test whether subtle differences in specificity could be detected by analysis of energy potentials. One example is NF-κB, which is a transcription factor of great importance in cellular signal transduction, particularly in the immune system.[39] Müller et al.[40] determined that the NF-κB p50 homodimer binds to a duplex oligonucleotide

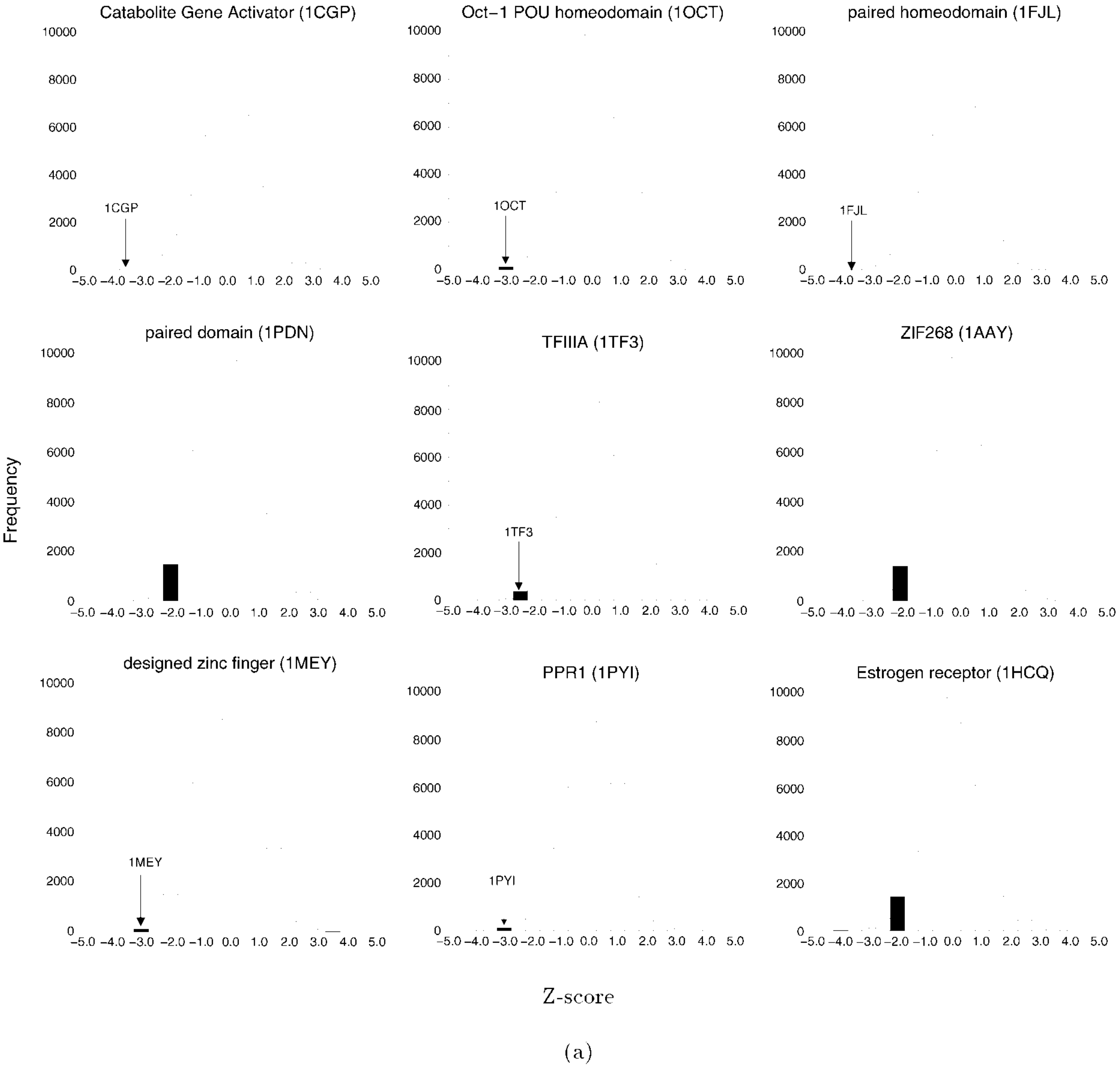
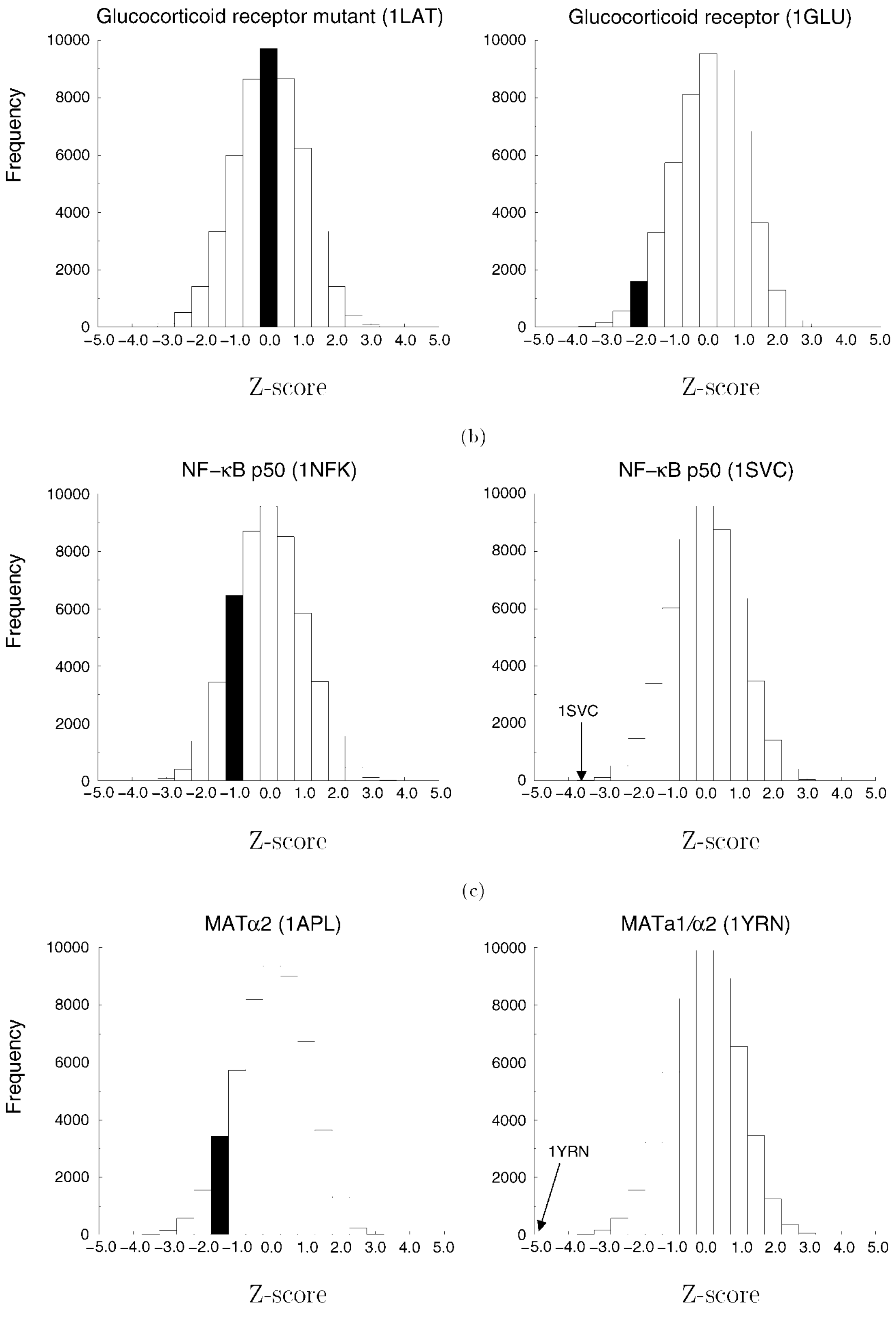


Fig. 8. Histograms of Z-scores for 50,000 random DNA sequences. In each histogram, the positions of the co-crystallized DNA sequences with protein are shown as filled bars and arrows. In each case, the database includes 52 complex-structures, except where the test complex was itself in the database, in which case it has been excluded. The considered DNA sites for wild type are as follows: **a:** 5′-GTCACACTTT-3′ for 1CGP: 5′-ATGCAAAT-3′ for 1OCT: 5′-TAATCTGATTA- 3′ for 1FJL; 5′-CGTCACGGTTGA-3′ for 1PDN; 5′-GGATGGGAGACC-3′ for 1TF3; 5′-GCGTGGGCGT-3′ for 1AAY; 5′-TGAGGCAGAACT-3′ for 1MEY; 5′-CGGCAATTGCCG-3′ for 1PYI; 5′-AGGTCACAGTGACCT-3′ for 1HCQ. **b:** 5′-CAGAACATG-3′ for 1LAT; 5′-CAGAACATC-3′ for 1GLU. **c:** 5′-GGGAATTCCC-3′ for 1NFK; 5′-GGGGAATCCCC-3′ for 1SVC. **d:** 5′-CATGTAATT-3′ for 1APL; 5′-TGTAATTTATTACATC-3′ for 1YRN.

with an 11-bp consensus recognition site located in the major histocompatibility complex class I enhancer; whereas Ghosh et al. [41] used a 10-bp idealized motif related, but not identical, to the natural sites. These structures are very similar, but differences in the details of DNA recognition between the two structures arise because of a fundamental difference between the DNA sequences used. It has been known that NF-κB binds 30 times more strongly to the former (odd-numbered motif) than to the latter (even-numbered motif).[42] Thus, we calculated the binding specificity of NF-κB to these sites using the corresponding structures. We found a difference in specificity, reflected by the higher Z-score for 1SVC (odd-numbered motif; Fig. 8c right) compared with 1NFK (even-numbered motif; Fig. 8c left). Although the specificity does not always have the relationship with the affinity (thermodynamic preference), these results strongly suggest that the present method can detect subtle differences in the specificity, which is attributed to the subtle structural differences.

Effect of Cooperativity on Binding Specificity

Transcription factors usually bind to their target sites in cooperation with other factors, by which a combination of different factors allows a complex mode of gene regulation. Thus, the cooperative binding should play an important role in protein-DNA recognition. An example of coopera-



(b)

(c)

(d)

Figure 8. (Continued.)

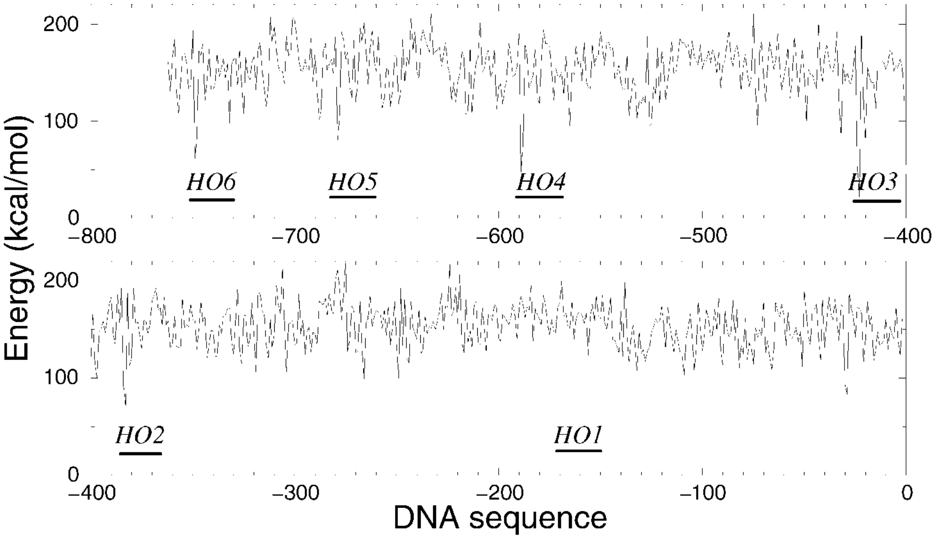


Fig. 9. Detection of the binding sites by a MATa1/α2 complex. The empirical potential was applied to the upstream of *S. cerevisiae* site-specific endonuclease (*HO*) gene (accession no. M14678). In the upstream region of the *HO* gene, six consensus patterns (including reverse sequences) exist (*HO1— HO6*),[48] and all of them but *HO1* were detected with the present method: *HO6*–736 **TGT**ATTCATTCA**CATC** (reverse), *HO5*–669 **TGT**CTTCAACTG**CATC** (reverse), *HO4*–576 **TGT**ATTTAGTTA**CATC** (reverse), *HO3*–411 **TGT**TATTATTTA**CATC**, *HO2*–371 **TGT**TCACATTAA**CATC**, *HO1*–150 **T**T**T**AGAACGCTT**CATC** (reverse). Invariant and highly conserved bases are highlighted in boldface.[44]

tive binding is found in homeodomains. In yeast, MATa1 and MATα2, homeodomain proteins, form a heterodimer that binds to DNA and represses transcription in a cell-specific manner. Individually, the α2 and a1 proteins have only modest affinity for DNA; nonetheless, the a1/α2 heterodimer binds to DNA with higher specificity and affinity.[43] We therefore tested the ability of our method to detect changes in binding specificity manifested in this cooperativity. Discrimination of target DNA by MATa1/α2–DNA (PDB code: 1YRN)[44] was highly significant (Z = −5.2, Fig. 8d right), whereas discrimination by MATα2–DNA (PDB code: 1APL)[45] was significantly lower (Z = −1.5, Fig. 8d left). The cooperativity was revealed when α2 specificity was selectively estimated within the MATa1/α2–DNA complex, where substantially greater discrimination was observed (Z = −4.3) than was seen with 1APL. The enhanced specificity may be caused by structural changes on binding of the two proteins to DNA. Thus, the present method can a detect subtle difference in specificity caused by structural changes induced by cooperative binding.

Effect of DNA Deformation

Deformations of DNA structures are often observed in protein–DNA complexes (Table III), and they likely contribute to the specificity of the protein–DNA recognition.[8–10] As one example, we consider the structure of the Zif268 zinc finger–DNA complex at 1.6 Å resolution (PDB code: 1AAY),[46] where the DNA takes an intermediate conformation between the canonical A-DNA and B-DNA and root mean square deviations from the A-DNA and B-DNA are 4.2 and 2.9 Å, respectively. To examine the effect of DNA deformation on binding, two complexes are modeled by replacing the native DNA with either canonical A- or B-DNA using root mean square fitting between them. The B = DNA complex shows selectivity lower than the crystal complex (Z = −1.4 and −1.9, respectively), but the selectivity was greatly increased if Zif268 was included in the database (Z = −2.0 and −5.0, respectively). Specificity of the A-DNA complex was much lower (Z = 0.8), even when the potential derivation included Zif268, indicating that the B conformation is more likely to be recognized by Zif268. These results suggest that the conformational changes of DNA may make a significant contribution to the specificity.

Role of Water Molecules

Water molecules are often observed in the protein–DNA complex, and their importance in both the specificity and affinity of protein–DNA interactions has been suggested.[47] Table I shows the number of direct contacts between amino acid and base as well as that of water-mediated contact. Zif268–DNA complex (PDB code: 1AAY),[46] which is not a very high Z-score (Fig. 8a, has nine water-mediated contacts. It is possible that these water molecules contribute to enhancing the Zif268 specificity. The effect of water molecules in the derivation of our potential is difficult at present because of the limited number of structural data containing water molecules. However, additional high-resolution structural data will be required to quantitate the contribution of water molecules to the specificity.

Target Site Detection in Real Promoter Sequences

Consideration of MATa1/α2 also enables us to test the capacity of this method to discriminate targets within real promoter sequences. The sum of the potential energies was calculated for every 16 base pairs along the DNA by shifting one base pair at a time using a cocrystal structure as a template (PDB code: 1YRN). MATa1/α2 regulates transcriptional repression of the *HO* gene by binding to the upstream region of the gene.[48] Among the six consensus sites in the upstream region (Fig. 9, *HO1–HO6*), fragments containing either site 6 (nucleotides −715 to −761) or site 3 (nucleotides −397 to −444) confer regulation by a1-α2 proteins.[49] Our calculation resolved binding sites for *HO2–HO6*, but not the binding site for *HO1*. In fact, deletion analysis has revealed that site 1 was not sufficient for regulation by a1-α2 proteins.[49] This demonstrates that the

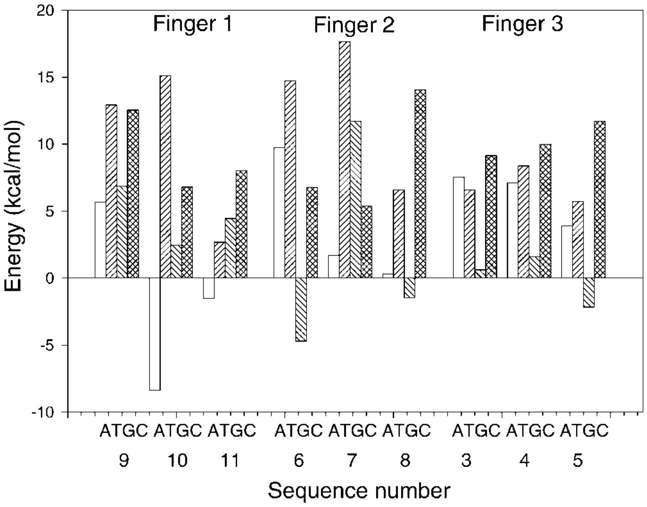


Fig. 10. Calculated base preference for the three zinc-finger positions of 1MEY. (A/G)AA, G(A/C)(G/A), and GGG are preferred by Finger 1, Finger 2, and Finger 3, respectively, in terms of the derived potentials. DNA sequence number shown at the bottom of the figure is taken from the crystal structure.

method may be one of the useful tools to detect candidates of binding site in practice.

Design of DNA-Binding Protein

The rational design of sequence-specific DNA-binding proteins will provide reagents for both biological research and gene therapy. In particular, zinc-finger proteins appear to provide the most versatile framework for design. Kim and Berg[50] reported the crystal structure of a complex made up of a newly designed protein comprised of three consensus-sequence-based zinc-finger domains and an oligonucleotide corresponding to a favorable DNA-binding site (1MEY; Table III). Each of the three zinc-fingers respectively recognizes three bases: GAA, (G/T)C(G/A), and GGG.[51,52] The calculated specificity for the crystal structure with the corresponding DNA sequence, GAA, GCA, and GAG, was quite high (Fig. 8a, 1MEY), but the most preferred DNA sequences according to the calculations were (A/G)AA, G(A/C)(G/A), and GGG (shown in Fig. 10), which is in agreement with the experimental results. These results suggest that our potential function can be applied to the design of DNA-binding proteins to have either increased or altered binding specificity.

Comparison with Other Methods

The prediction of binding specificity using a knowledge-based approach has been attempted by deriving amino acid–base score[53] or quantitative parameters.[54] These methods succeeded in the prediction of specificity for a given distinct framework in which specific pairs of amino acid and base are known to interact. However, it is not clear whether these methods can be generalized, because any amino acids not in the given framework were not considered. Amino acids at other sites may play an important role in recognition, e.g., through water-mediated interactions from various directions.[46,55] In addition, amino acids in the protein–DNA interface can take multiple conformations, as shown in the TFIIIA complex.[56,57] In the structure, Lys26, Lys29, and Lys92, located at the −1, 3, and 6 position, respectively (numbering with respect to the start of each recognition α helix), take multiconformations to form H bonds with different bases. Zinc-finger domains are often perceived as independent modular units, each recognizing three to four base pairs by way of specific and discrete interactions with a limited set of amino-acid side-chains. However, the NMR structure of TFIIIA has demonstrated that such models are oversimplified. Our method considers all the contributions of amino acids in the given box for each base (Fig. 1). It also includes implicitly the flexibility of side-chains. Thus, our method could still predict the binding site correctly for TFIIIA (1TF3 in Fig. 8a).

CONCLUSION

The present analyses on the interactions between amino-acid side-chain and bases indicate that there is no distinct codelike correspondence between amino acids and bases, but clear preferences exist. On the other hand, the analysis of spatial distribution of amino acids around bases shows that even those amino acids with strong base preference are distributed in a wide space around bases, allowing amino acids with many different geometries to form a similar type of interaction with bases. Such conformational flexibility may allow regulatory proteins to recognize a set of multiple target sites with certain sequence variation, and control multiple genes cooperatively. The spatial distributions of base–amino acid interactions have suggested that the structural data can be used to transform the distribution into empirical interaction potentials, which may be used to predict target sites for regulatory proteins. We have demonstrated that the energy potentials extracted from the distributions of Cα atoms around DNA bases of the known protein–DNA complex structures are sufficiently sensitive to detect the DNA binding sites of regulatory proteins. Moreover, this method can also be applied to proteins of unknown structure having substantial sequence similarity to known proteins, on the basis of which structures can be modeled and binding sites can be predicted. Anticipated increases in the amount of structural data should enable further refinement of the potential, either to improve its precision or to permit inclusion of further details, and make this a powerful tool for predicting multiple target sites and genes for regulatory proteins.

ACKNOWLEDGMENTS

The authors thank Dr. C.O. Pabo for sending us the coordinates of 2GLI and Dr. R.L. Jernigan for useful comments on the manuscript. H.K. is supported by the special postdoctoral researchers program from RIKEN.

REFERENCES

1. Seeman NC, Rosenberg JM, Rich A. Sequence-specific recognition of double helical nucleic acids by proteins. Proc Natl Acad Sci USA 1976;73:804–808.

2. Mandel-Gutfreund Y, Schueler O, Margalit H. Comprehensive analysis of hydrogen bonds in regulatory protein-DNA complexes: in search of common principles. J Mol Biol 1995;253:370–382.
3. Matthews BW. Protein–DNA interaction. No code for recognition. Nature 1988;335:294–295.
4. Suzuki, M. A framework for the DNA–protein recognition code of the probe helix in transcription factors: the chemical and stereochemical rules. Structure 1994;2:317–326.
5. Choo Y, Klug A. Selection of binding sites for zinc fingers using rationally randomised DNA reveals coded interactions. Proc Natl Acad Sci USA 1994;91:11168–11172.
6. Spolar RS, Record MTJ. Coupling of local folding to site-specific binding of proteins to DNA. *Science* 1994;263:777–784.
7. Lefstin JA, Yamamoto KR. Allosteric effects of DNA on transcriptional regulators. Nature 1998;392:885–888.
8. Dickerson RE, Chiu TK. Helix bending as a factor in protein/DNA recognition. Biopolymers 1997;44:361–403.
9. Olson WK, Gorin AA, Lu XJ, Hock LM, Zhurkin VB. DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. Proc Natl Acad Sci USA 1998;95:11163–11168.
10. Nekludova L, Pabo CO. Distinctive DNA conformation with enlarged major groove is found in Zn-finger–DNA and other protein-DNA complexes. Proc Natl Acad Sci USA 1994;91:6948–6952.
11. Ogata K, Kanei-Ishii C, Sasaki M, et al. The cavity in the hydrophobic core of Myb DNA-binding domain is reserved for DNA recognition and trans-activation. Nat Struct Biol 1996;2:178–187.
12. Sarai A, Mazur J, Nussinov R, Jernigan RL. Sequence dependence of DNA conformational flexibility. Biochemistry 1989;28:7842–7849.
13. Koudelka GB, Harrison SC, Ptashne M. Effect of non-contacted bases on the affinity of 434 operator for 434 repressor and Cro. Nature 1987;326:886–888.
14. Schneider TD, Stormo GD, Gold L. Information content of binding sites on nucleotide sequences. J Mol Biol 1986;188:415–431.
15. Berg OG, von Hippel PH. Selection of DNA binding sites by regulatory proteins—statistical-mechanical theory and application to operators and promoters. J Mol Biol 1987;193:723–750.
16. Berg OG, von Hippel PH. Selection of DNA binding sites by regulatory proteins. II. The binding specificity of cyclic AMP receptor protein to recognition sites. J Mol Biol 1988;200:709–723.
17. Lustig B, Jernigan RL. Consistencies of individual DNA base-amino acid interactions in structures and sequences. Nucleic Acids Res 1995;23:4707–4711.
18. Takeda Y, Sarai A, Rivera VM. Analysis of the sequence-specific interactions between Cro repressor and operator DNA by systematic base substitution experiments. Proc Natl Acad Sci USA 1989;86:439–443.
19. Sarai A, Takeda Y. λ repressor recognizes the approximately 2-fold symmetric half-operator sequences asymmetrically. Proc Natl Acad Sci USA 1989;86:6513–6517.
20. Tanikawa J, Yasukawa T, Enari M, et al. Recognition of specific DNA sequences by the c-Myb proto-oncogen product: Role of three repeat units in the DNA-binding domain. Proc Natl Acad Sci USA 1993;90:9320–9324.
21. Deng QL, Ishii S, Sarai A. Binding site analysis of c-Myb: screening of potential binding sites by using the mutation matrix derived from systematic binding affinity measurements. Nucleic Acids Res 1996;24:766–774.
22. Bernstein FC, Koetzle TF, Williams GJB, et al. The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol 1977;112:535–542.
23. Sippl M. Calculation of conformational ensembles for potentials of mean force: an approach to the knowledge-based prediction of local structures in globular proteins. J Mol Biol 1990;213:859–883.
24. Miyazawa S, Jernigan RL. Estimation of effective inter-residue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 1985;18:534–552.
25. Bryant SH, Lawrence CE. An empirical energy function for threading protein sequence through the folding motif. Proteins 1993;16:92–112.
26. Jones DT, Thornton JM. Potential energy functions for threading. Curr Opin Struct Biol 1996;6:210–216.
27. Miyazawa S, Jernigan RL. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. J Mol Biol 1996;256: 623–644.
28. Bowie JU, Lüthy R, Eisenberg D.A method to identify protein sequences that fold into a known three-dimensional structure. Science 1991;253:164–170.
29. Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. Nature 1992;358:86–89.
30. Ouzounis C, Sander C, Scharf M, Schneider R. Prediction of protein structure by evaluation of sequence-structure fitness. J Mol Biol 1993;232:805–825.
31. Moodie SL, Mitchell JBO, Thornton JM. Protein recognition of adenylate: an example of a fuzzy recognition template. J Mol Biol 1996;263:486–500.
32. Oobatake M, Ooi T. Hydration and heat stability effects on protein unfolding. Prog Biophys Mol Biol 1993;59:237–284.
33. Baker EN, Hubbard RE. Hydrogen bonding in globular proteins. Prog Biophys Mol Biol 1984;44:97–179.
34. McDonald IK, Thornton JM. Satisfying hydrogen bonding potential in proteins. J Mol Biol 1994;233:777–793.
35. Ippolito JA, Alexander RS, Christianson DW. Hydrogen bond stereochemistry in protein structure and function. J Mol Biol 1990;215:457–471.
36. Taylor R, Kennard O, Versichel W. Geometry of the N-H ··· O=C hydrogen bond. 1. Lone-pair directionality. J Am Chem Soc 1983;105:5761–5766.
37. Gewirth DT, Sigler PB. The basis for half-site specificity explored through a non-cognate steroid receptor–DNA complex. Nat Struct Biol 1995;2:386–394.
38. Luisi BF, Xu WX, Otwinowski Z, Freedman LP, Yamamoto KR, Sigler PB. Crystallographic analysis of the interaction of the glucocorticoid receptor with DNA. Nature 1991;352:497–505.
39. Kuriyan J, Thanos D. Structure of NF-κB transcription factor: a holistic interaction with DNA. Structure 1995;3:135–141.
40. Müller CW, Rey FA, Sodeoka M, Verdine GL, Harrison SC. Structure of the NF-κB p50 homodimer bound to DNA. Nature 1995;373:311–317.
41. Ghosh G, Duyne GV, Ghosh S, Sigler PB. Structure of NF-κB p50 homodimer bound to a κB site. Nature 1995;373:303–310.
42. Chytil M, Verdine GL. The Rel family of eukaryotic transcription factors. Curr Opin Struct Biol 1996;6:91–100.
43. Goutte C, Johnson AD. Recognition of a DNA operator by a dimer composed of two different homeodomain proteins. EMBO J 1994; 13:1434–1442.
44. Li T, Stark MR, Johnson AD, Wolberger C. Crystal structure of the MATa1/MATα2 homeodomain heterodimer bound to DNA. Science 1995;270:262–269.
45. Wolberger C, Vershon AK, Liu B, Johnson A, Pabo CO. Crystal structure of a MATα2 homeodomain-operator complex suggests a general model for homeodomain–DNA interactions. Cell 1991;67: 517–528.
46. Erickson EM, Rould MA, Mekludova L, Pabo C. Zif268 protein-DNA complex refined at 1.6 Å: a model system for understanding zinc finger-DNA interactions. Structure 1996;4:1171–1180.
47. Schwabe JWR. The role of water in protein–DNA interactions. Curr Opin Struct Biol 1997;7:126–134.
48. Miller AM, MacKay VL, Nasmyth KA. Identification and comparison of two sequence elements that confer cell-type specific transcription in yeast. Nature 1985;314:598–603.
49. Russell DW, Jensen R, Zoller MJ, et al. Structure of the *Saccharomyces cerevisiae HO* gene and analysis of its upstream regulatory region. Mol Cell Biol 1986;6:4281–4294, 1986.
50. Kim CA, Berg JM. A 2.2 Å resolution crystal structure of a designed zinc finger protein bound to DNA. Nat Struct Biol 1996;3:940–945.
51. Desjarlais JR, Berg JM. Length-encoded multiplex binding site determination: application to zinc finger proteins. Proc Natl Acad Sci USA 1994;91:11099–11103.
52. Kim CA, Berg JM. Serine at position 2 in the DNA recognition helix of a Cys$_2$-His$_2$ zinc finger peptide is not, in general, responsible for base recognition. J Mol Biol 1995;252:1–5.
53. Suzuki M, Yagi N. DNA recognition code of transcription factors in the helix-turn-helix, probe helix, hormone receptor, and zinc finger families. Proc Natl Acad Sci USA 1994;91:12357–12361.
54. Mandel-Gutfreund Y, Margalit H. Quantitative parameters for amino acid-base interaction: implications for prediction of protein-DNA binding sites. Nucleic Acids Res 1998;26:2306–2312.

55. Wilson DS, Guenther B, Desplan CJK. High resolution crystal structure of a paired (Pax) class cooperative homeodomain dimer on DNA. Cell 1995;82:709–719.
56. Wuttke DS, Foster MP, Case DA, Gottesfeld JM, Wright PE. Solution structure of the first three zinc fingers of TFIIIA bound to the cognate DNA sequence: determinants of affinity and sequence specificity. J Mol Biol 1997;273:183–206.
57. Foster MP, Wuttke DS, Radhakrishan I, Case DA, Gottesfeld JM, Wright PE. Domain packing and dynamics in the DNA complex of the N-terminal zinc fingers of TFIIIA. Nat Struct Biol 1997;4:605–608.
58. Schultz SC, Shields GC, Steitz TA. Crystal structure of a CAP–DNA complex: the DNA is bent by 90°. Science 1991;253:1001–1007.
59. Klemm JD, Rould MA, Aurora R, Herr W, Pabo CO. Crystal structure of the Oct-1 POU domain bound to an octamer site: DNA recognition with tethered DNA-binding modules. Cell 1994;77:21–32.
60. Xu W, Rould MA, Jun S, Desplan C, Pabo CO. Crystal structure of a paired domain-DNA complex at 2.5 Å resolution reveals structural basis for Pax developmental mutations. Cell 1995;80:639–650.
61. Marmorstein R, Harrison SC. Crystal structure of a PPR1-DNA complex: DNA recognition by proteins containing a Zn_2Cys_6 binuclear cluster. Genes Dev 1994;8:2504–2512.
62. Schwabe JW, Chapman L, Finch JT, Rhodes D. The crystal structure of the estrogen receptor DNA-binding domain bound to DNA: how receptors discriminate between their response elements. Cell 1993;75:567–578.

INTRODUCTION

METHODS

Fig.
1.
TABLE
I
RESULTS
AND DISCUSSION
TABLE
II
Fig.
2.
Fig.
3.
Figure
3.
Figure
3.
Figure
3.
Fig.
4.
Fig.
5.
Fig.
6.
Fig.
7.
TABLE
III
Fig.
8.
Figure
8.
Fig.
9.
CONCLUSION

ACKNOWLEDGMENTS

REFERENCES

Fig.
10.