Binding Geometry of α-Helices That Recognize DNA

Masashi Suzuki[1] **and Mark Gerstein**[2]

[1]*MRC Laboratory of Molecular Biology, Cambridge CB2 2QH, United Kingdom;* [2]*Department of Structural Biology, Stanford Medical School, Stanford, California 94305-5400*

ABSTRACT **Many transcription factors have an α-helix that binds to DNA bases in a specific fashion. The DNA-binding geometry of these recognition helices varies substantially. We define a set of parameters to describe the binding geometry of recognition helices and analyze specific stereochemical elements that determine particular geometries. Because the convex surface of the helix must fit into the concave surface of the DNA major groove, the number of degrees of freedom of the recognition helix is reduced from a possible six to a single angle, which we call α. The chemically interacting DNA bases and amino acid residues must lie along a common line and have the same spacing along it. This pairing of base positions with residue positions seems to restrict the binding geometry further to a set of discrete values for α.** © 1995 Wiley-Liss, Inc.

Key words: DNA–protein interaction, crystal structure, transcription factor, gene regulation

INTRODUCTION

Many proteins use an α-helix for sequence-specific recognition of the DNA major groove[1–5] (see also the original papers listed in Table I). In the known DNA complexes with transcription factors that use recognition helices, the DNA is largely kept in the B-form but the inclination of the recognition helix relative to the DNA helix axis varies substantially (Fig. 1). This has been noted by many crystallographers who reported individual structures of DNA–protein complexes, but systematic comparison of the inclination was not carried out.

The aim of this work is twofold. First, we define a set of parameters that can be used for describing the DNA-binding geometry of an α-helix and calculate the parameters using the known crystal structures of transcription factors in complex with DNA. (Obviously for describing the DNA-binding geometry of an α-helix a set of parameters becomes necessary, but to our knowledge no such set has been proposed before this work.)

Second, we analyze the stereochemical elements that fix the DNA-binding geometry of α-helices by using the calculated parameters. In earlier papers we have analyzed patterns of residue-base contacts in complexes of DNA with transcription factors and have found that recognition helices of a particular DNA-binding motif use the same set of residue positions to contact certain base positions.[6,7] This finding suggests that a DNA-binding motif has a unique binding geometry. We further investigate this idea and try to understand the stereochemical basis in terms of the binding parameters.

We do not suggest that recognition helices are the only important aspect of DNA-protein interactions. However, our study does highlight the importance of the binding geometry of the recognition helix in understanding DNA–protein interactions.

MATERIALS AND METHODS

Coordinates

The coordinates of the crystal structures were drawn from the Protein Data Bank.[8] In total, we did calculations on 17 protein structures.

Definition of Axes for DNA and for the Recognition Helix

To calculate the axis of the recognition helix, we fit an appropriate length of ideal α-helix to it. All the recognition helices fit fairly well, and the RMS deviation in doing the fit is between 0.1 and 0.5 Å/atom (for ~10 C_α atoms). Consequently, the axes of the α-helices are well defined.

Because of the greater flexibility of the DNA double helix in comparison with an α-helix, it is much less straightforward to define an axis for DNA. A number of approaches have been tried, (e.g., finding the average screw rotation relating one base to the next, fitting a section of DNA to ideal B-form DNA, and finding the principal axes for a moment of inertia tensor derived from DNA), and some of the recent approaches are quite elaborate.[9–11] However, no method produces completely satisfactory results. Consequently, since it is not clear what we would gain from using an elaborate method, we have deliberately chosen a very simple method. To calculate the DNA axis, we fit the equation of an ideal helix (i.e., the equation $\theta = 2\pi z/p$, where p is the rise per

Received May 4, 1995; revision accepted August 7, 1995.

Address reprint requests to Masashi Suzuki, MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK.

TABLE I. Crystal Structures of DNA-Transcription Factors in Which an α-Helix is Used for DNA Recognition

Name	PDB	Reference
I (one-turn helices)		
TR		
TrpR	1TRO	Otwinowski et al., 1988[18]
TrpR	1TRR	Lawson and Carey, 1993[28]
C6		
Gal4	1D66	Marmorstein et al., 1992[29]
II (two-turn helices)		
HTH		
λR	1LMB	Clarke et al., 1991[30]
λR	—	Jordan and Pabo, 1988[31]
434R	—	Anderson et al., 1987[32]
434R	2OR1	Aggarwal et al., 1988[33]
434R	1RPE	Shimon and Harrison, 1993[51]
434R	1PER	Rodgers and Harrison, 1993[36]
434C	—	Wolberger et al., 1988[34]
434C	3CRO	Mondragón and Harrison, 1991[35]
CAP	1CGP	Schultz et al., 1991[12];
λC	4CRO	Brennan et al., 1990[37]
Hin	1HCR	Feng et al., 1994[38]
Oct1POU	—	Klemm et al., 1994[52]
HNF3	—	Clark et al., 1993[39]
ZnF		
Zif	1ZAA	Pavletich and Pabo, 1991[40]
[all AF]		
TTK	—	Fairall et al., 1993[41]
[all AF]		
GLI	1GLI	Pavletich and Pabo, 1993[42]
[F4-BF, F5-AF]		
p53		
p53	—	Cho et al., 1994[53]
III (three-turn helices)		
PH		
Matα2	—	Wolberger et al., 1991[43]
Engl	1HDD	Kissinger et al., 1990[44]
GCN4	1YSA	Ellenberger et al., 1992[22]
GCN4	—	König and Richmond, 1993[23]
E2	2BOP	Hegde et al., 1992[45]
Oct1homeo	—	Klemm et al., 1994[52]
MX		
Max	—	Ferré-D'Amaré et al., 1993[46]
USF	—	Ferré-D'Amaré et al., 1994[47]
MD		
MyoD	—	Ma et al., 1994[48]
C4		
GlucR	1GLU	Luisi et al., 1991[49]
EstR	—	Schwabe et al., 1993[50]

turn) through the phosphates on each strand. Then we vector average the axes for the two strands.

Another important aspect of the DNA axis calculation is choosing the set of base pairs to use in defining the axis. We could either try to use the few base pairs closest to the recognition helix to define a local axis or try to use all the DNA in the crystal structure to define a global average. As with the calculation method, there is no clearly correct answer, so we have tried to pick the most straightforward approach and use as much of the DNA as possible. We have tried several different definitions of the DNA axes and found that the major features of the α-β plot, which are discussed in this paper, do not change. Consequently, except for two cases, we have used one half-site of the crystal structures to define the axis, unless the DNA is really straight, in which case we use all of the DNA. Since the DNA is

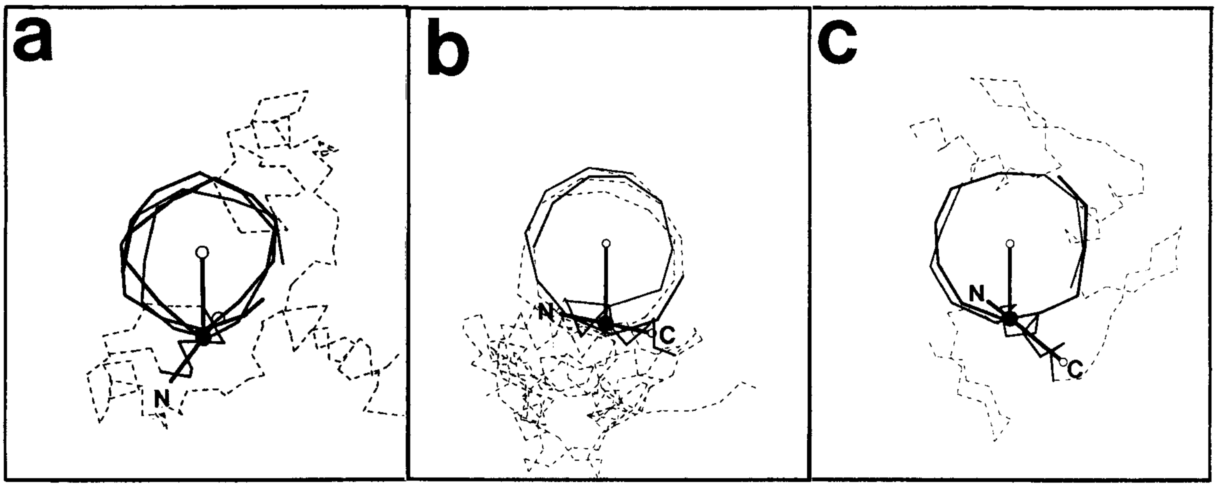

Fig. 1. Examples of the DNA-binding geometry of a recognition helix. Recognition helices are shown binding to DNA. **a:** Gal4 [C6]. **b:** The glucocorticoid receptor [C4]. **c:** Zif268 [AF]. The view is down the local DNA axis, and the recognition helices are always drawn with the N-terminus on the left.

fairly straight over the set of base pairs chosen, the calculated axes agree well with what one would fit by eye, and the RMS deviation between the phosphate atom positions and the ideal helix equation are between 0.5 and 1.5 Å/atom. (Consequently, for those structures for which all the DNA is used in the calculations, we get essentially the same axis if we restrict the calculation to a single half-site.)

The two exceptional cases are GLI and CAP. For the GLI structure, which has three recognition helices, we used the seven base pairs closest to each recognition helix to define the axis. As the DNA in GLI is not strongly deformed, the calculated axes fit the crystal structure fairly well, with RMS deviations less than 0.6 Å/atom. By contrast, the DNA of CAP (1CGP) is dramatically bent: the direction of its axis changes course by ~ 45° over each of the two half-sites.[12] To define the axis for CAP we used the eight base pairs closest to the recognition helix. We felt this was an appropriate balance between getting an accurate local axis and averaging over enough base pairs. Although the plot of CAP in the α-β diagram does not change largely, if we chose a different set of bases or used a different method of calculation for CAP we would have gotten a different axis.

Calculation of the Binding Parameters

To describe the geometry of a recognition helix relative to the DNA major groove precisely, six parameters are needed. We use three cylindrical coordinates, h, d, and θ (Fig. 2a), to describe the position of the center of the helix relative to the DNA, and three angles, α, β, and γ (Fig. 2b), to describe the rotation of the helix around this center. We define the centre of an α-helix as the projection onto the helix axis of the center of the C_α atoms of the residues used for base recognition. As shown in Table II and Figure 3, this center position is at the middle of the second turn for helices that use three turns for base recognition (C4, PH, AF—see Table I and the Materials and Methods section for the naming); at the mid-point between the two turns for helices that use two turns for base recognition [HTH, BF]; and at the middle of the first turn for helices that use only a single turn [C6, TR].

Once the axes of the DNA and the recognition helix are calculated, calculation of parameters for characterizing the binding geometry, α, β, d, and h is straightforward. A line *d* is drawn from the center of the recognition helix to the DNA axis so that it is perpendicular to the axis. The length of this line is parameter d. The distances h_1 and h_2 are calculated by finding, on a plane that contains line *d*, the intersection of the DNA axis with the recognition helix center and each sugar-phosphate backbone. The angles α and β are calculated with the following formulas:

$$\cos\beta = \cos B \csc A$$

and

$$\cos\alpha = \cos A \csc B$$

where B is the angle between the helix axis and the DNA axis and A is the angle between the helix axis and line *d*.

(We will make available electronically supplemental figures and information relevant to our calculations. Send e-mail to mbg@hyper.stanford.edu or use anonymous ftp with the following URL: ftp: hyper.stanford.edu/pub/mbg/DNA/.)

Classification of Transcription Factors

In this paper, as we have done before,[6,7,13] to focus attention on the binding mode of the recognition helices, we classify transcription factors based on the way they bind to DNA rather than on their overall fold. In the crystal structures no more than three turns of an α-helix access bases on the DNA double helix. Therefore, as shown in Table I, recognition helices can be classified into three groups according to the number of turns used for base recognition. The recognition helices can be further classified based on the ways they bind to DNA, i.e., based on the amino acid positions used for base contacting and those for phosphate binding (Table I).

Some particular points regarding the classification are:

1. Zn fingers are divided into two subgroups, A fingers [AF] and B fingers [BF].[14]
2. The C4 family [C4] includes steroid hormone receptors and GATA1.[15]
3. The probe helix [PH] family[16,17] includes zipper proteins and homeo proteins.
4. The recognition helix of the tryptophan repressor binds to DNA in a very different way from that of other HTH proteins,[18] because the two positions that face DNA bases in other HTH proteins are used for binding to the co-factor, tryptophan, and thus this protein is classified into another group [TR], while all the remaining classic HTH proteins are contained in another family [HTH].
5. The factors Max and USF are classified into the same group [MX], while the transcription factor MyoD is classified into another group [MD], as the residue positions used for base recognition in the helix of MyoD are shifted by one helical turn from those in the helix of the MX family.

The recognition helices that use only one turn for base recognition (i.e., those of C6, TR) adopt the "perpendicular fit" (Fig. 3a and text) and thus use main-chain features for contacting the DNA bases. In the crystal structure of Trp repressor [TR] these contacts are intermediated by water molecules,[18] but such water molecules were not detected by a nuclear magnetic resonance (NMR) study[19] and thus these contacts might be made directly.

The classification of the majority of Zn fingers, A fingers [AF], is slightly complicated as the first position used for base recognition is not part of the helix but is N-terminal to it; thus, the AF is intermediate between two and three turns, and for a similar reason p53 is intermediate between one turn and two turns.

A recognition helix cannot be totally independent from the protein fold, but there is no simple one-to-one correspondence between the fold and the type of recognition helix.[7,13,20] For example, a homeo protein has the helix-turn-helix [HTH] fold, but its recognition helix is different from those of classic HTH proteins: the residue positions used for base recognition are shifted toward the C-terminus (counted

TABLE II. Sequences of the Recognition Helices*

Name	Sequence												Center
I													
TR	1	2	3	4	5	6	7	8					2.8
TrpR	**I**	**A**	t	I†	T	r	G	s					
C6	1	2	3	4	5	6	7	8					6.2
Gal4	C†	D	I	C†	r	L	**K**	**K**					
p53	1	2	3	4	5	6	7						
p53	C)	P	G	**R**	D	R	**r**						2.5
II													
HTH	1	2	3	4	5	6	7	8	9				3.5
LamR	**q**	**S**	g	V†	G	**A**	L	F	n				
434R	**q**	**Q**	s	I†	E	**Q**	L	E	n				
434C	**q**	**Q**	S	I†	Q	**L**	I	E	A				
CAP	**r**	**E**	T	V†	G	**R**	I	L	k				
LacR	**y**	**Q**	T	V†	S	**R**	V	V	N				
Hin	V	**S**	t	L†	Y	**r**	y	F	P				
Oct1 POU													
	Q	**T**	t	I†	S	**R**	F	E	A				
HNF3	**q**	**N**	S	I†	R	**H**	S	L	s				
AF	1)	2	3	4	5	6	7	8	9	10	11	12	3.5
Zif F1	**R**)	S	**D**	E	L	**T**	**R**	h†	I	r	I	H†	
F2	**R**)	S	**D**	**H**	L	T	T	h†	I	R	T	H†	
F3	**R**)	S	**D**	E	R	K	**R**	H†	T	K	I	H†	
TTK F1	h)	I	**S**	**N**	F	C	**R**	H†	Y	V	T	S	
F2	**R**)	k	**D**	**N**	M	T	A	h†	V	K	I	I	
GLI F5	**D**)	P	**S**	**S**	L	r	**K**	H†	V	K	T	V	
BF	1	2	3	4	5	6	7	8	9	10	11	12	4.5
GLI F4	**A**	**S**	**D**	r	A	**K**	h†	Q	N	R	t	H†	
III													
PH	1	2	3	4	5	6	7	8	9	10	11	12	4.5
E2	**n**	q	V	**K**	**C**	Y	r	**F**	r	V	K	K	
GCN4 (K)	**N**	t	E	**A**	**A**	r	r	**s**	r	A	r	k	
GCN4 (EC)	**N**	t	E	**A**	**A**	r	r	**s**	**R**	A	r	k	
GCN4 (EG)	**n**	t	E	**A**	**A**	R	r	**s**	r	A	r	k	
Matα2	N	w	V	**S**	**N**	R	r	**R**	k	E	k	T	
Engl	**I**	w	F	**Q**	**N**	K	r	A	k	I	k	K	
Antp	**I**	W	F	**Q**	N	R	R	**M**	K	W	K	K	
Oct1 homeo													
	V	W	F	**C**	**N**	R	r	**Q**	K	E	K	R	
Mx	1	2	3	4	5	6	7	8	9	10	11	12	4.5
Max	**H**	n	A	L	**E**	R	K	r	**r**	D	H	I	
USF (1)	**H**	n	E	**V**	**E**	r	r	r	**r**	D	K	I	
USF (2)	H	n	E	V	**E**	r	R	r	**R**	D	K	I	
MD	1	2	3	4	5	6	7	8	9	10	11	12	
MyoD	**r**	K	A	A	**t**	M	r	**E**	r	R	r	L	
C4	1	2	3	4	5	6	7	8	9				4.5
GlucR	G	S	C†	**K**	**V**	F	F	K	**R**				
EstR	**E**	G	C†	**K**	G	F	F	**K**	r				
GATA	**N**	A	C†	G	**L**	Y	Y	K	**L**				

*Residues that bind to DNA bases are shown in bold. Those that bind to DNA phosphates are shown in lower case. Those marked with † are characteristic of the protein fold and are placed opposite the DNA. The centers of the helices are also shown. The residue position 1 of AF is not inside the helix. Protein-DNA contacts in the two GCN4 structures (K, König and Richmond[23]; E, Ellenberger et al.[22]) are slightly different from each other. Also those found in two halves of the Ellenberger structure of GCN4 (EC and EG in the Ellenberger structure) and those of USF (Ferré-D'Amaré et al.[47]; 1 and 2) are slightly different.

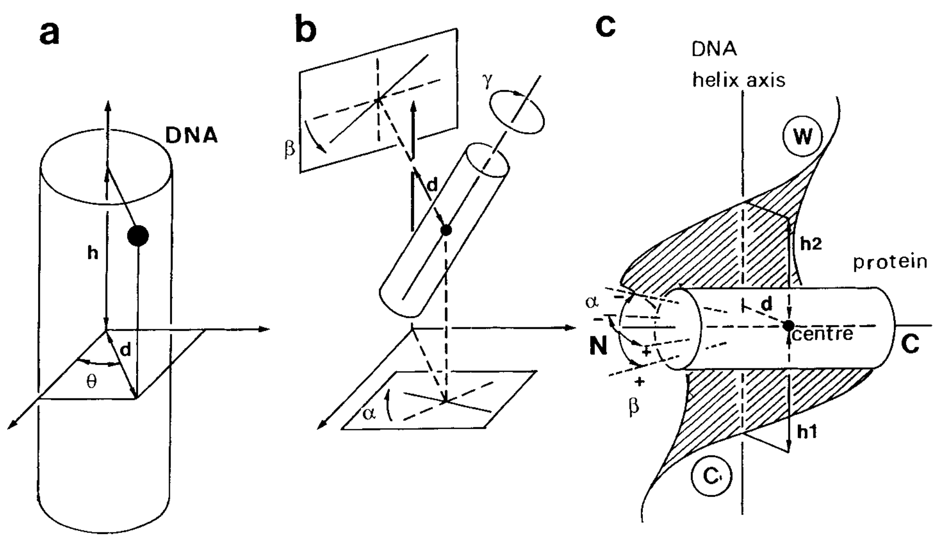

Fig. 2. Definition of the binding parameters. The rigid-body positioning of the recognition helix relative to DNA has at most 6 degrees of freedom. These are expressed in terms of six coordinates (h, d, θ, α, β, γ), which we define here. **a:** Three cylindrical coordinates (h, d, and θ) define the center of the recognition helix. **b:** Three angles (α, β, γ) describe the rotation of the recognition helix around its center. As discussed in the text, the θ and γ angles are not significant for describing the overall fit. **c:** Consequently, we show a close-up of the recognition helix highlighting only the remaining four parameters: d, h, α, and β. The distance d is measured between helix center and DNA axis. It defines a line *d,* which is perpendicular to the DNA axis; h is defined as the ratio h_1/h_2, where h_1 is the distance along the DNA axis from the recognition-helix center to the sugar-phosphate backbone of the Crick (C) strand when one is looking down the line *d,* and h_2 is the analogous distance to that of the Watson (W) strand. α is the angle between the recognition helix axis and the line *d,* when looking down the DNA axis, and β is the angle between the helix axis and the DNA axis, when looking along line *d.*

from the conserved hydrophobic position) and there are additional basic residues at the C-terminus.[16]

RESULTS AND DISCUSSION

Major Features of α-Helix–DNA Interaction

First, we briefly discuss some major features of α-helix–DNA interactions in the crystal structures. An α-helix is essentially straight (e.g., all the recognition helices fit well to a standard helix with RMS deviations less than 0.5 Å/atom; see Materials and Methods), and the number of DNA bases that an α-helix can follow along the curved major groove is limited. Therefore, in the crystal structures a recognition helix accesses only one side of the DNA (Fig. 3c) and binds to *no more than five base pairs.*

In particular, a monomer of GCN4 is a single α-helix, and its recognition helix seems to be free from any strain from the rest of the protein. Also the C-terminus of the helix becomes a "zipper," i.e., a coiled coil, and thus the helix has a curvature. Therefore, it was once predicted that upon binding to two slightly different binding sites the α-helix could deform to adopt the DNA structures.[21] However, after two different DNA–GCN4 complex structures were determined,[22,23] it became clear that it is the DNA but not the α-helix that changes its structure to bind the partner molecule.[23] Indeed, the degree to which the recognition helix of GCN4 is deformed is small (RMS deviation of the recognition helix of GCN4 from a standard helix is 0.13–0.16 Å/atom).

The major interaction positions in DNA are eight bases, C1–C4 and W2–W5, on one side of the DNA (Fig. 4a). In this paper we use the Watson(W)-Crick(C) notation for the two DNA strands. The DNA strand that runs from 5′ to 3′, when the recognition helix follows the DNA from N to C, is called the Crick strand, and the other, the Watson strand. By combining the name of the DNA strand and the base pair number, the bases are named C1, C2, W1, W2, etc. (C1 is the partner of W1). The base pair number increases along the N-C direction of the recognition helix.

In the crystal structures no more than *three turns of an α-helix* access bases on the DNA double helix. This is because the pitch of an α-helix is 5.4Å and the DNA-facing side of three turns spans 10.8Å (Fig. 3c), while the diameter of bases around the DNA helix axis is approximately 10Å. (A fourth turn, however, may be used for binding to phosphates.) Therefore, as shown in Table I, recognition helices can be classified into *three groups* according to the number of turns used for base recognition.

To use all the three turns for base recognition, an α-helix adopts the "parallel" fit (Fig. 3c), while to use only one turn, it adopts the "perpendicular" fit (Fig. 3a). Thus the number of the turns itself reflects the binding geometry. In this paper we use a further classification for recognition helices based on the ways they bind to DNA, i.e., based on the amino acid positions used for base contacting and

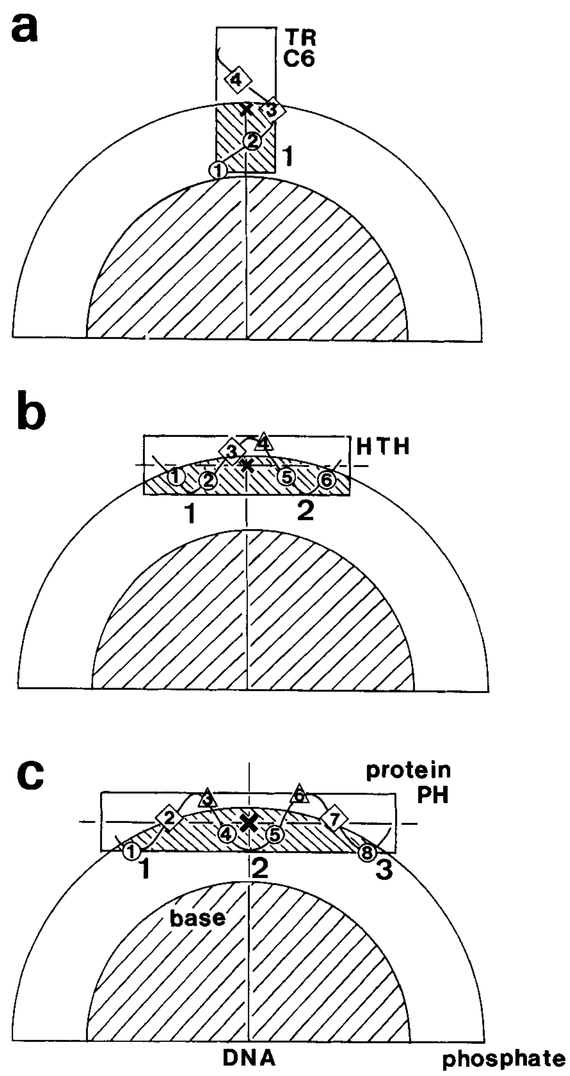

Fig. 3. The different length of recognition helices. Recognition helices of one turn [TR/C6] (**a**), two turns [HTH] (**b**), and three turns [PH] (**c**) are drawn schematically, looking down the DNA axis. Only the half of the DNA that faces the protein is shown. The residues shown with circles bind to DNA bases; those with diamonds bind to DNA phosphates; and those with triangles face away from the DNA. The centers of the recognition helices are marked with an x. The numbers 1, 2, and 3 show the first, second, and third turns, respectively.

those for phosphate binding (Table I; see also Materials and Methods).

Definition of Binding Parameters

To describe the geometry of a recognition helix relative to the DNA major groove precisely, six parameters are needed. We use three cylindrical coordinates, h, d, and θ (Fig. 2a), to describe the position of the center of the helix relative to the DNA, and three angles, α, β, and γ (Fig. 2b), to describe the rotation of the helix around this centre. In discussing the overall geometry of a recognition helix we find that four of the six parameters are not important or have a limited range of values. The parameters θ and γ are taken into account by the formalism that will be described later in this paper. These are rotations around the axes of the DNA or of the α-helix, and because of the helical-symmetric characters of the two molecules, they cause no large change in the overall fit. In other words, we consider both the α-helix and the DNA cylindrically symmetric.

From the crystal structures, we have calculated values for the parameters d and h (Fig. 5a). The parameter d is the length of the shortest path from the center of the α-helix to the DNA axis. It has a nearly constant value (Fig. 5a) of 8.7 ± 0.8 Å, which is slightly smaller than the radius of the sugar-phosphate backbones around the DNA axis. The parameter h is defined as the ratio h_1/h_2, where h_1 is the distance along the DNA axis from the helix center to the sugar-phosphate backbone of the Crick strand, and h_2 is the analogous distance to the Watson strand (Fig. 2c). This parameter also does not change much (0.99 ± 0.18).

The d and h parameters are nearly constant because the center of a recognition helix must always be an appropriate distance from the bottom of the DNA major groove, lest the positions used for base recognition, which are N-terminal or C-terminal to the center, move too close to or too far from the DNA base pairs. Also the helix center must keep a similar distance to the two sugar-phosphate backbones, lest the helix collides with the backbones.

Fitting of Surfaces

In what follows we concentrate on the two remaining parameters, α and β. As shown in Figure 2, α is the angle between the recognition helix axis and the shortest path *d* from the helix center to the DNA axis, when looking down the DNA axis, and β is the angle between the recognition helix axis and the DNA axis, when looking down the shortest path *d*. It is apparent from Figure 5b that the α, β values for the recognition helices of each DNA binding motif are clustered together (e.g., see the clustering of values for the HTH proteins, 3–7 in Fig. 5b, AF, 9–12, and PH, 14–18). Thus, each DNA-binding motif has its own particular binding geometry.

Helices of three turns [PH (14–18), C4 (19)] have $\alpha \approx 0$, the parallel fitting, while helices of one turn [TR (1), C6 (2)] have large values of α, the perpendicular fitting. Helices of two turns [HTH (3–7), AF (9–12), BF (13)] occupy the region of intermediate α (Fig. 5b).

A remarkable feature of the α-β plot is that the values are distributed essentially along two lines: $\beta = 0.67\alpha + 25$ and $\beta = -0.67\alpha + 25$. Thus the binding geometry of a recognition helix essentially has a single degree of freedom and can be charac-

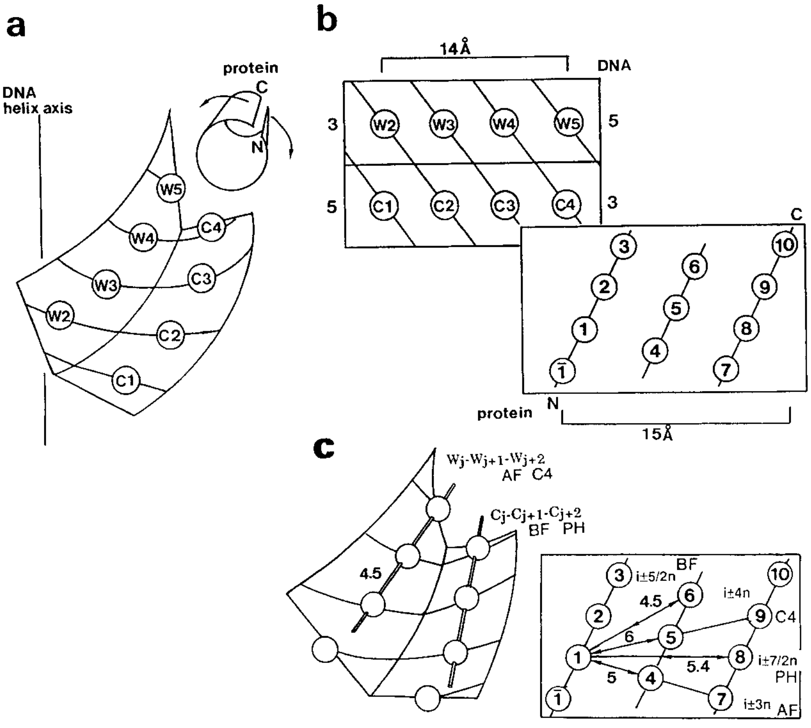

Fig. 4. Residue and base positions used for DNA-protein contacts. When a recognition helix interacts with DNA, a well-defined set of nucleotide bases and amino acid residues are involved in the interaction. These are shown schematically here. **a**: A recognition helix in the DNA major groove. Note that the base pairs are approximately perpendicular to the DNA axis, while the major groove is tilted about 65° from this axis. **b**: For recognizing eight base positions in the five base pairs (C1–C4 and W2–W5), up to nine residue positions in three turns in a recognition helix can be used (1–9). Note that the recognition helix is opened and seen from its *inside*. **c**: To a rough approximation, the bases on DNA can be arranged along various possible lines, which we call base lines, and the residues in the protein can likewise be arranged along residue lines. When the recognition helix interacts with DNA, it is possible to match up particular residue lines and base lines. The correspondence between residue lines and base lines found in the crystal structures is shown in the table below and those of group 1 are shown in this subfigure.

base lines	positions in the recognition helix forming the residue line		spacing between residue positions	structure
group 1	1, 5, 9	i±4n	4.5–6 Å	C4
	1, 4, 7	i±3n		AF
	1, 4/5, 8	i±7/2n		PH
	1, 2/5, 6	i±5/2n		BF
group 2	1, 6	i±5n	9–12 Å	HTH
	1, 9	i±8n		MX
	1, 8	i±7n		MD, PH
group 3	1, 2	mainchain i±1n	3.5 Å	TR, C6

Usually more than one way can be found to choose and connect residue positions in a recognition helix. First, if it is possible to form a residue line connecting three positions, which binds a base line connecting three base positions, this residue line is identified as representing the helix. If not, the line connecting two residues of the largest separation is chosen. In some cases it is not easy to determine whether the line type is i ± Nn or i ± Nn/2. Since the first residue of PH (aa1) binds to C1 and W1, two binding modes are possible. If aa1 binds to C1, aa8 binds to C3, and aa4 or aa5 binds to C2, then there is a line connecting the three residue positions i ± 7/2n, and the PH has group 1 binding. However, if aa1 binds only to W2, the line type is i ± 7n (group 2).

terized by the single parameter α (in addition to a sign).

It seems possible to rationalize the relationship between α and β in terms of the shape of the major groove. When α is 0, the recognition helix must become parallel to the major groove, lest the ends of the helix hit the sugar-phosphate backbones. This fixes β at around 25°. (The major groove is tilted about 65° from the DNA helix axis.) For appreciable positive and negative values of α, the recognition helix has more freedom to move in the β direction. However, to keep a tight fit between the recognition helix surface and the major groove surface, the range of β values is again sharply limited for a given α value, that is, one does not observe a recognition helix standing up in the midst of the groove without contacting either of the two sugar-phosphate backbones (i.e., α << 0°, β = 25°).

a

No	Name	PDB	α(°)	β(°)	d(Å)	h(h1/h2)
TR						
1	TrpR	1TRO	-55	-2	7.7	1.20
C6						
2	Gal4	1D66	73	-34	9.0	1.00
HTH						
3	λR	1LMB	-22	6	8.0	1.01
4	434R	2OR1	-24	12	8.3	1.16
5	434C	3CRO	-25	15	8.3	1.23
6	CAP	1CGP	-27	17	8.8	0.80
7	Hin	1HCR	-18	-7	9.4	1.36
AF						
9	ZifF1	1ZAA	-39	54	9.7	0.87
10	ZifF2	1ZAA	-48	57	9.3	0.97
11	ZifF3	1ZAA	-41	49	9.3	0.94
12	GLIF5	---	-46	52	7.8	0.82
BF						
13	GLIF4	---	-37	49	7.3	1.09
PH						
14	Engl	1HDD	-6	26	9.8	0.77
16	E2	1BOP	-12	29	9.2	0.88
17	GCN4	1YSA	-22	20	9.0	0.78
18	GCN4	1DGC	-7	23	8.1	0.98
C4						
19	GlucR	1GRD	-12	11	10.5	1.17
	Average		---	---	8.7	0.99
	S.D.		---	---	±0.8	±0.18

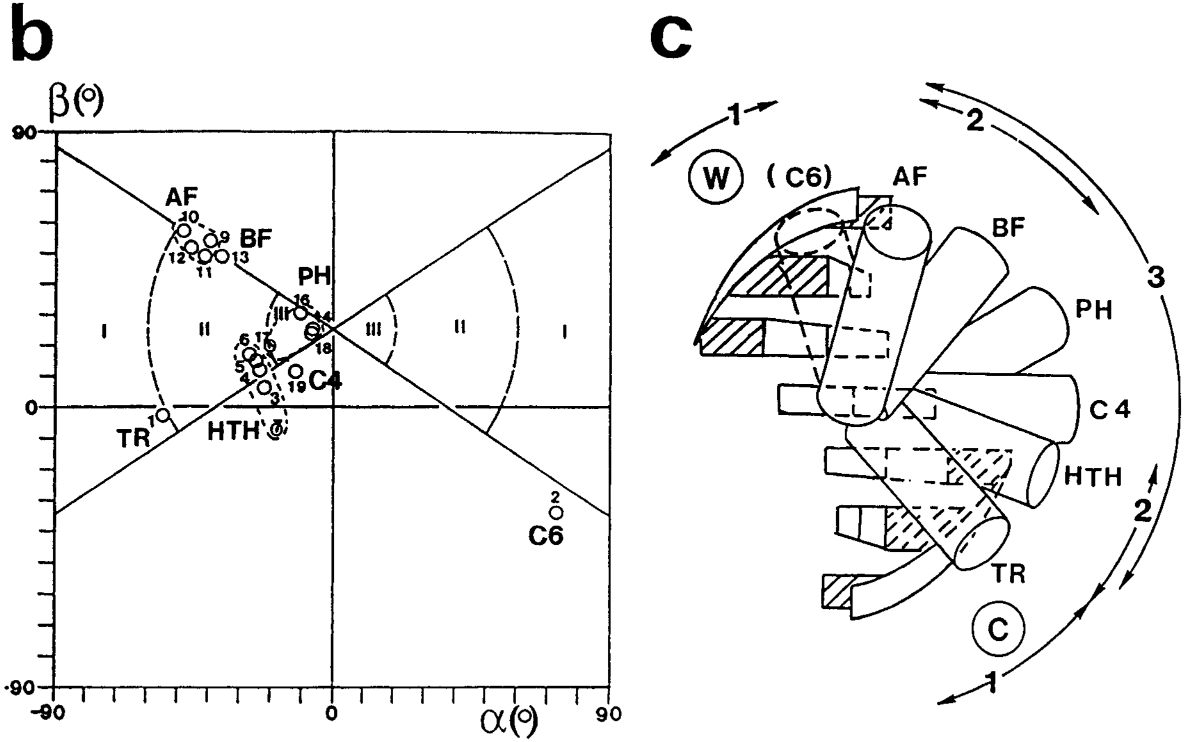

Fig. 5. Binding parameters calculated from the crystal structures. **a:** Using the definitions in Figure 4, we calculated values for the α, β, h, and d parameters for recognition helices listed in Table I. We were obviously only able to do calculations on the 17 proteins listed in the table with publicity available coordinates. This calculation involved 40 distinct recognition helix for a given protein (e.g., many of the structures are dimers), we averaged the values. It is clear that the parameters d and h do not vary appreciably among the different recognition helices. Consequently, in parts b–d, we focus our attention on α and β. **b:** A plot of β versus α. The numbers 1–19, indicating the protein names, are the same as in a. The numbers I–III show the regions occupied by helices of one, two, and three turns, respectively. Although helices with different numbers of turns are not necessarily separated from each other completely on the plot, those with more turns tend to be located nearer the center. To a good approximation β is related to α by the straight-line equations: $\beta = \pm 0.67\alpha + 25°$. **c:** Consider a recognition helix fixed at its N-terminus to the center of DNA base pair and rotated along the surface of the major groove. As shown in the figure, as the helix is rotated, it will pass through the geometries of all the known types of recognition helices (e.g., HTH, C4, PH, etc.). Furthermore, as it is rotated, the number of α-helical turns contacting DNA bases will vary between 1 and 3, and this number is indicated in the figure as well. As shown in the figure, the recognition helix of C6 is found in the left half of the diagram, while those of TR and HTH, for example, are found in the right half. The left half of the diagram appears to be twofold symmetric to the right half. However, to keep the Watson (W)-Crick (C) notation of the DNA strands and N-C direction of the helix the same, the left half of the diagram cannot be directly compared with right half when the N-terminus of the helix is held fixed. The left half can only be used when the C-terminus of the helix is fixed to the central base pair.

To understand the nature of the relationship between α and β further, it is useful to introduce another coordinate system, in which one of the termini of the recognition helix is fixed instead of its center (Fig. 5c). This coordinate system is, of course, equivalent to and interchangeable with the coordinate system discussed at length above. If the first residue position used for base-recognition at the N-terminus is fixed near a particular DNA base pair (precisely speaking, it can be fixed onto a base on either the Watson strand or the Crick strand, but this distinction seems unnecessary here), the left half (α ≤ 0°) of the α-β plot can be described (Fig. 5c). To follow the DNA surface, the α and β angles become dependent on each other. The C-terminus of the helix can move from the "parallel" orientation [PH, C4] to the "perpendicular" orientation [TR] and then back to the "parallel" orientation. This movement corresponds to the trace in half of the α-β plot on the left side. Similarly, if the residue position used for base recognition at the C-terminus is fixed near a particular DNA base pair, the other half of the α-β plot is traced (see C6 in Fig. 5c).

The Watson-Crick notation of the two DNA strands is dependent on the N-C direction of the recognition helix. At the α angle of ±90°, the N-C direction becomes reversed and thus the Watson strand becomes the Crick strand and vice versa. Thus the two lines in the α-β plot are part of a closed curve, shaped like a bowtie or the infinity sign (∞).

It seems now clear that like protein-protein interaction,[24] DNA-protein interaction involves the *fitting of two surfaces* (see the following section).

Three Functional Types of Residue Positions

The fact that each DNA-binding motif has its own particular binding geometry can be understood by analyzing the functional types of positions around the recognition helices. These positions in the crystal structures can be classified into three types: (A) those that contact DNA bases, (B) those that contact DNA phosphates, and (C) those which are "exposed" and can interact with the rest of the protein.

Many residue positions that are routinely used for identifying DNA-binding motifs, such as the hydrophobic positions in HTH proteins and the Cys and His residues in zinc fingers, are characteristics of the protein fold and are of type (C). These residues can still be important for the binding geometry, since they must be placed on the far side of the DNA, limiting the rotation of the helix. The way in which these three types of residue positions are combined into a single helix is specific to each kind of recognition helix (Table II), and this seems to be the reason why each kind has its specific binding geometry.

If the DNA major groove were filled with water up to the height of the sugar-phosphate backbones, a recognition helix binding to DNA would be half "sunk" into the "sea" (Fig. 3a). The type (B) residues are found on the watermark around the helix, while the type (A) residues are on the "wet" area and the type (C) residues are on the "dry" area.[16]

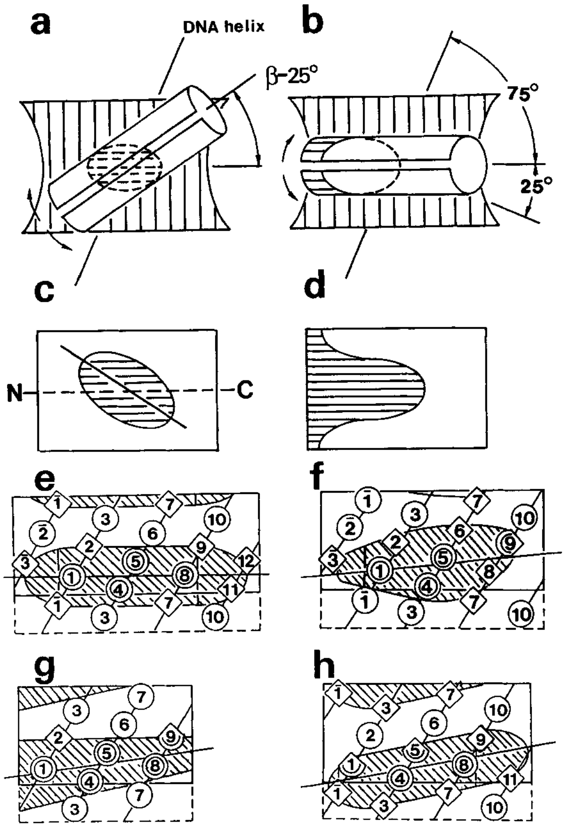

Fig. 6. Watermark analysis of recognition helices. **a–d:** If the DNA major groove were filled with water, as shown in a and b, the residue positions used for base recognition would be found in the corresponding "wet" parts shown in c and d. The shape of the part (i.e., its watermark) reflects the binding geometry: a and c show β = 60°, α = 0; and b and d show β = 25°, α = 25°. (Note that these are not usual values for recognition helices.) **e–h:** "Watermark" plots of PH (e), MX (f), C4 (g), and MD (h) are shown. The residues contacting DNA bases are shown in double circles; those contacting phosphates in diamonds; those binding to a phosphate and a base at the same time in half-double-circles, half-diamonds. Lines fit through the middle of the wet parts are also shown. For example, Fisher et al.[54] classified positions around the recognition helix of a MX transcription factor, Myc, into types (A)–(C) on the basis of carefully designed biochemical experiments. Later two crystal structures[46,47] showed that the conclusions of Fisher et al.[54] are essentially correct. The binding geometry of an MX recognition helix (the coordinates have not been published) can be predicted by comparing its watermark plot with those of other well-characterized helices.

The shape of the watermark can be examined closely by cutting the helix and opening it flat (see Figs. 3, 6 and note that the α-helix surface is seen from inside the helix). Two features of the watermark shape are important (Fig. 6): the angle of a line that fits through the middle of the wet part relative to the α-helix axis (Fig. 6a,c) (this tilt angle corresponds to angle β − 25°) and whether the part becomes wider toward the N-terminus or toward the C-terminus (this corresponds to the α angle, Fig. 6b,d). Thus one can relate a watermark plot to the binding geometry. If the residue positions in the recognition helix that contact the DNA bases are characterized by biochemical or genetic experiments, a watermark plot may be used to understand the binding geometry (Fig. 6). The helical wheel projection that is often used for purposes similar to the watermark plot is not so useful, unless the helix binds parallel to the DNA major groove ($\alpha = 0^\circ$, $\beta = 25$).

Sequence-specific DNA-protein interaction is achieved through contacts between DNA bases and amino acid residues at the (A) positions. The distance between two residues used for base recognition must be similar to the distance between the two bases they contact. Thus, if three (A) residues are arranged along a line (a "residue line") and three bases are arranged along another line (a "base line"), then it may be appropriate for the residue line to bind to the base line (the base lines and residue lines found in the crystal structures are summarized in Figure 3c; see also discussion on DNA-α helix interaction in ref. 25 and on DNA-β sheet interaction in ref. 26). This resembles another situation of protein-protein interaction in which two residue lines can be used for analysis.[27]

ACKNOWLEDGMENTS

We thank Dr. C. Chothia for his critical reading of the manuscript; and Drs. M. Levitt and A. Lesk for assistance with the calculations. M.G. acknowledges support from a Damon Runyon-Walter Winchell fellowship (DRG-1272).

NOTE ADDED IN PROOF

After this paper was prepared, some more coordinates of DNA–protein complexes were published. The parameters of these structures will be discussed elsewhere (see M. Suzuki and N. Yagi, Proc. Japan Acad., in press; see also a review by M. Suzuki, D. Loaks, and N. Yugi in *Advances in Biophysics,* in press).

REFERENCES

1. Harrison, S.C. A structural taxonomy of DNA-binding domains. Nature 353:715–719, 1991.
2. Branden, C. Tooze, J. "Introduction to Protein Structure." New York: Garland, 1991.
3. Perutz, M. "Protein Structure." New York: Freeman and Co., 1992.
4. Steiz, T.A. "Structural Studies of Protein–Nucleic Acid Interaction." Cambridge: Cambridge University Press, 1993.
5. Neidle, S. "DNA Structure and Recognition." Oxford: IRL Press, 1994.
6. 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 91:12357–12361, 1994.
7. Suzuki, M., Brenner, S.E., Gerstein, M., Yagi, N. DNA recognition code of transcription factors. Protein Eng. 8:319–328, 1995.
8. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Jr, Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M. The Protein Data Bank: A computer-based archival file for macromolecular structure. J. Mol. Biol. 112:535–542, 1977.
9. Babcock, M.S., Pednault, E.P.D., W.K. Olson, W.K. Nucleic acid structure analysis: Mathematics for local Cartesian and helical structure parameters that are truly comparable between structures. J. Mol. Biol. 237:125–156, 1994.
10. Yanagi, K., Prive, G.G., Dickerson, R.E. Analysis of local helical geometry in three B-DNA decamers and eight dodecamers. J. Mol. Biol. 217:201–214, 1991.
11. Fratini, A.V., Kopka, M., Drew, H., Dickerson, R.E. Reversible bending and helix geometry in a B-DNA dodecamer: CGCGAATT^{Br}CGCG. J. Biol. Chem. 257:14686–14707, 1982.
12. Shultz, S.C., Shields, G.C., Steitz, T.A. Crystal structure of a CAP-DNA complex: The DNA is bent by 90°. Science 253:1001–1007, 1991.
13. Suzuki, M. Transcription factors, Myb and LexA, discriminate between DNA sequences by the same recognition mechanism. Proc. Jpn. Acad. B71:27–31, 1995.
14. Suzuki, M., Gerstein, M., Yagi, N., Stereochemical basis of DNA recognition by Zn fingers. Nucleic Acids Res. 22: 3397–3405, 1994.
15. Suzuki, M., Chothia, L. DNA recognition rules for steroid hormone receptors and GATA1: (1) chemical and stereochemical rules. Proc. Jpn. Acad. B70:58–61, 1994.
16. Suzuki, M. Common features in DNA recognition helices of eukaryotic transcription factors. EMBO J. 12:3221–3226, 1993.
17. Suzuki, M. A framework for the DNA–protein recognition code of the probe helix in the transcription factors: The chemical and stereochemical rules. Structure 2:317–327, 1994.
18. Otwinowski, Z., Schevitz, R.W., Zhang, R.-G., Lawson, C.L., Joachimak, A., Marmorstein, R.Q., Luisi, B.F., Sigler, P.B. Crystal structure of trp repressor/operator complex at atomic resolution. Nature 335:321–329, 1988.
19. Zhang, H., Zhao, D., Revington, M., Lee, W., Jia, X., Arrowsmith, C., Jardetzky, O. The solution structures of the trp repressor–operator DNA complex. J. Mol. Biol. 238: 592–614, 1994.
20. Suzuki, M., Yagi, N., Gerstein, M. DNA recognition and superstructure formation by HTH proteins. Protein Eng. 8:329–338, 1995.
21. Hu, J.C. and Sauer, R.T. The basic region-leucine zipper family of DNA-binding proteins. In: "Nucleic Acids and Molecular Biology," Vol. 6. Eckstein, F. and Lilley, D.M.J., eds. Berlin: Springer-Verlag, 1992:82–101.
22. Ellenberger, T.E., Brandl, C.S., Struhl, K., Harrison, S.C. The GCN4 basic region leucine zipper binds DNA as a dimer of uninterrupted α helices: Crystal structure of the protein–DNA complex. Cell 71:1223–1237, 1992.
23. König, P., Richmond, T. The X-ray structure of the GCN4-bZip bound to ATF/CREB site DNA shows the complex depends on DNA flexibility. J. Mol. Biol. 233:139–154, 1993.
24. Janin, J., Chothia, C. The structure of protein–protein recognition sites. J. Biol. Chem. 265:16027–16030, 1990.
25. Suzuki, M. DNA-recognition code tables for transcription factors in the C4 zinc binding and C2H2 zinc finger families. Proc. Jpn. Acad. B70:96–99, 1994.
26. Suzuki, M. DNA-recognition by a β-sheet. Protein Eng. 8:1–4, 1995.
27. Chothia, C. Principles that determine the structure of proteins. Annu. Rev. Biochem. 53:537–572, 1984.
28. Lawson, C.L., Carey, J. Tandem binding in crystals of a *trp* repressor/operator half-site complex. Nature 366:178–182, 1994.
29. Marmorstein, R., Carey, M., Ptashne, M., Harrison, S.C. DNA recognition by GAL4: Structure of a protein-DNA complex. Nature 356:408–414, 1992.
30. Clarke, N.D., Beamer, L.J., Goldberg, H.R., Berkower, C., Pabo, C.O. The DNA binding arm of λ repressor: Critical contacts from a flexible region. Science 254:267–270, 1991.
31. Jordan, S.R., Pabo, C.O. Structure of the lambda complex at 2.5Å resolution: Details of the repressor–operator interactions. Science 242:893–899, 1988.
32. Anderson, J.E., Ptashne, M., Harrison, S.C. Structure of the repressor–operator complex of bacteriophage 434. Nature 326:846–852, 1987.

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

34. Wolberger, C., Dong, Y., Ptashne, M., Harrison, S.C. Structure of a phage 434 Cro/DNA complex. Nature 335: 789–795, 1988.

35. Mondragón, A., Harrison, S.C. The phage 434 Cro/OR1 complex at 2.5Å resolution. J. Mol. Biol. 219:321–334, 1991.

36. Rodgers, D.W., Harrison, S.C. The complex between phage 434 repressor DNA-binding domain and operator site Or3: Structural differences between consensus and non-consensus half-sites. Structure 1:227–240, 1993.

37. Brennan, R.G., Roderick, S.L., Takeda, Y., Matthews, B.W. Protein–DNA conformational changes in the crystal structure of a λ cro-operator complex. Proc. Natl. Acad. Sci. USA 87:8165–8169, 1990.

38. Feng, J.-A., Johnson, R.-C. Dickerson, R.E. Hin recombinase bound to DNA: The origin of specificity in major and minor groove interactions. Science 263:348–355, 1994.

39. Clark, M.L., Halay, E.D., Lai, E., Barley, S.K. Co-crystal structure of the HNF-3/fork head DNA-recognition motif resembles histone H5. Nature 364:412–420, 1993.

40. Pavletich, N.P., Pabo, C.O. Zinc finger-DNA recognition: Crystal structure of a Zif268-DNA complex at 2.1Å. Science 252:809–817, 1991.

41. Fairall, L., Schwabe, J., Chapman, L., Finch, J.T., Rhodes, D. The crystal structure of a two zinc-finger peptide from the *Drosophila* protein Tramtrack complexed with DNA residues; an extension to the rules for zinc finger/DNA recognition. Nature 366:483–487, 1993.

42. Pavletich, N.P., Pabo, C.O. Crystal structure of a five-finger GLI–DNA complex: New perspectives on Zn fingers. Science 261:1701–1707, 1993.

43. Wolberger, C., Vershon, A.K., Liu, B., Johnson, A.D., Pabo, C.O. Crystal structures of a Matα2 homoeodomain–operator complex suggests a general model for homoeodomain-DNA interactions. Cell 67:517–528, 1991.

44. Kissinger, C.R., Liu, B., Martin-Blanco, E., Kornberg, T.B., Pabo, C.O. Crystal structure of an engrailed homoeodomain–DNA complex at 2.8Å resolution: A framework for understanding homoeodomain–DNA interactions. Cell 63:579–590, 1990.

45. Hegde, R.S., Grossman, S.R., Laimins, L.A., Sigler, P.B. Crystal structure at 1.7Å of the bovine papillomavirus-1 E2 DNA-binding domain bound to its DNA target. Nature 359:505–512, 1992.

46. Ferré-D'Amaré, A.R., Prendergast, G.C., Ziff, E.B., Burley, S.K. Recognition by Max of its cognate DNA through a dimeric b/HLH/z domain. Nature 363:38–45, 1993.

47. Ferré-D'Amaré, A.R., Pognonec, P., Roeder, R.G., Burley, S.K. Structure and function of the b/HLH/Z domain of USF. EMBO J. 13:180–189, 1994.

48. Ma, P.C., Rould, M.A., Weintraub, H., Pabo, C.O. Crystal structure of MyoD bHLH domain–DNA complex: Perspectives on DNA recognition and implications for transcription activation. Cell 77:451–459, 1994.

49. Luisi, B.F., Xu, X.W., Otwinowski, Z., Freedman, L.P., Yamamoto, K.R., Sigler, P.B. Crystallographic analysis of the interaction of the glucocorticoid receptor with DNA. Nature 352:497–505, 1991.

50. Schwabe, J.W., Chapman, L., Finch, J.T., Rhodes, D. The crystal structure of the complex between the oestrogen receptor DNA-binding domain and DNA at 2.4Å: How receptors discriminate between their response elements. Cell 75:567–578, 1993.

51. Shimon, L.J.W., Harrison, S.C. The phage 434 OR2/R1-69 complex at 2.5Å resolution. J. Mol. Biol. 232:826–838, 1993.

52. Klemm, J.D., Rould, M.A., Aurora, R., Her, W., Pabo, C.O. Crystal structure of the Oct-1 POU domain bound to an octamer site: DNA recognition with tethered DNA-binding modules. Cell 77:21–32, 1994.

53. Cho, Y., Gorina, S., Jeffrey, P.D., Pavletich, N.P. Crystal structure of a p53 tumor supressor–DNA complex: Understanding tumorigenic mutations. Science 265:346–355, 1994.

54. Fisher, D.E., Parent, L.A., Sharp, P.A. High affinity DNA-binding Myc analogs: Recognition by an α-helix. Cell 72: 467–476, 1993.