Protein Engineering vol.8 no.4 pp.319–328, 1995

REVIEW

DNA recognition code of transcription factors

Masashi Suzuki[1], Steven E.Brenner, Mark Gerstein[2] and Naoto Yagi[3]

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK, [2]Department of Structural Biology, Stanford Medical School, Stanford, CA 94305-5400, USA and [3]Tohoku University, School of Medicine, Seiryo-machi, Sendai 980-44, Japan

[1]To whom correspondence should be addressed

Key words: DNA binding/DNA–protein interaction/gene expression/molecular recognition

Introduction

Over 35 years have passed since the 'central dogma' of molecular biology (DNA makes RNA makes protein) was proposed (Crick, 1958). Despite its remarkable verification, it is being seen increasingly as limited, for if the whole flow of information in a cell were unidirectional, all cells with the same complement of genetic material would have identical function and morphology. The truth is manifestly otherwise.

A group of proteins, transcription factors, selects the information used in cells by specifically binding to 'regulatory' DNA sequences. Among other effects, this causes the differentiation of cells. These factors act as the final messenger in a transduction pathway of signals which come from outside the cell. Thus, gene expression can be regulated by the environment.

Recognition between a transcription factor and its target DNA is achieved through the physical interaction of the two molecules. Since the structures of both DNA and proteins are determined by their primary sequences, there must be a set of rules to describe DNA–protein interactions entirely on the basis of sequences. The fundamental question is whether these rules are simple and comprehensible, such that the DNA recognition code can be compared with the triplet code which summarizes the rules of how DNA and protein sequences are related in the central dogma.

As we review in this paper, a simple code for DNA recognition by transcription factors does seem to exist. In fact, the recognition rules allow us (i) to predict DNA–protein interactions, (ii) to change the binding specificity of an existing transcription factor, and (iii) probably even to design in a rational way a new protein which binds to a particular DNA sequence. The code has been derived from crystal structures of transcription factor–DNA complexes (Table I) and the vast body of biochemical, genetic and statistical information about the binding specificity of transcription factors.

Most of the transcription factors discussed here use an α-helix, which binds to the DNA major groove, for recognition. Those proteins which have a 'recognition helix' discussed here fall mainly into four families: probe helix (PH), helix–turn–helix (HTH), zinc finger (ZnF) and C4 Zn binding proteins (C4). There is, in addition, one transcription factor family described that uses a β-sheet, the MetJ repressor-like (MR) family. [See Table I for members of these and other families. Note that (i) individual Zn fingers are further subdivided into A and B fingers, AF and BF (Suzuki *et al.*, 1994a), (ii) the PH family includes homeodomain and basic-zipper proteins (Suzuki, 1993) and (iii) the C4 family includes the hormone receptors and the GATA proteins (Suzuki and Chothia, 1994).]

Historical background

The first important step towards the DNA recognition code was achieved by Seeman *et al.* (1976). They noticed that as in some RNA structures, where a third base can bind to the side of a Watson–Crick base pair, a protein side chain can bind to a particular DNA base pair through a bidentate hydrogen bond, thereby discriminating between the DNA base pairs. They modeled two specific amino acid–nucleotide base interactions, Arg–G and Asn/Gln–A, which were later found in many crystal structures.

The next important step was the discovery of DNA binding motifs. As the number of known transcription factors increased, it was recognized that some transcription factors share the same structural framework. The first motif identified was HTH (Sauer *et al.*, 1982). The discovery of several other motifs followed, such as ZnF (Miller *et al.*, 1985) and the basic-domain leucine zipper motif (Landschultz *et al.*, 1988). It was expected that DNA recognition rules would be established rapidly, because to recognize DNA, proteins appeared to use a common structural framework and to vary a few positions to achieve specificity. In this atmosphere, Pabo and Sauer (1984) proposed the term the '[DNA] recognition code'.

Ironically, now that a few dozen structures of DNA–transcription factor complexes are known in atomic detail, the belief in general rules seems to have been largely abandoned (see, for example, Matthews, 1988), although some limited resemblance among DNA binding modes of proteins of the same family is acknowledged (see, for example, Pabo *et al.*, 1990).

Meanwhile, the development of genetic and biochemical techniques, such as footprinting and PCR, enabled other types of approach to the subject. Based on such experiments, Müller-Hill and co-workers argued that a DNA recognition code for HTH proteins does exist (Kisters-Woike *et al.*, 1991; Lehming *et al.*, 1991) but did not explicitly formulate it. Even for ZnF, which has been studied extensively by these types of experiments (Klevit, 1991; Desjarlais and Berg, 1993), Pavletich and Pabo (1993) expressed skepticism in saying that 'it appears quite unlikely that there will be any simple general code'.

One of us noticed that some eukaryotic factors included in homeoproteins and basic-zipper proteins, which were not believed to belong to the same family at that time, actually use very similar α-helices for DNA recognition (Suzuki, 1993). This DNA recognition motif, which has a conserved set of phosphate and base binding positions, is now known as the probe helix (PH). After the framework of the DNA recognition

rules of PH became clear (Suzuki, 1994a), we found that the same principles could be applied to other transcription factor families (Suzuki and Chothia, 1994; Suzuki and Yagi, 1994b; Suzuki *et al.*, 1994a, 1995), including one which uses a β-sheet instead of an α-helix for DNA recognition (Suzuki, 1995a).

DNA recognition code

The major part of the DNA recognition code consists of two types of rule: chemical and stereochemical. The chemical rules are general, while the stereochemical rules are specific to each family of DNA binding proteins.

Chemical rules

The chemical rules are based on the intrinsic chemical ability of a given residue and a base to produce a non-covalent interaction, either through a hydrogen bond or hydrophobic interaction (Figure 1c–f). Such contacts have been noted in the original reports of crystal structures (Table I). Possible pairing partners can be determined (Figure 1a) by examining

a

	small	medium	large	aromatic
A	Cys,Ser, Thr 10	**Asn 15** **Asp 9** His 8	**Gln 15** **Glu 9** Arg,Lys 3 Met 5	Tyr 5 Trp 5
T	**Ala 10** Cys,Ser, Thr	**Val,Ile 12** Asn 10 His 8	**Leu,Met 12** Gln 10 Arg,Lys 5	**Tyr,Phe 12** **Trp 12**
G	Cys,Ser, Thr 10	**His 12** Asn 10	**Arg,Lys 15** Gln 10	Tyr 5
C	**Val 8** Cys,Ser, Thr 10	**Asp 12** Asn 10 His 8 Ile 8	**Glu 12** Gln 10 Leu,Met 8	**Tyr,Phe 8** **Trp 8**

b

aa	C(i)— C(i+1) W(i+1)-W(i)		C(i)— W(i+1) W(i+1)-C(i)	
Asp,Glu	A	A	A	C
	C	A	A	A
	C	C	C	C
Asn,Gln Ser,Cys Thr,Tyr	A	A	A	A
	A	C	A	T
	A	G	A	G
	T	A	T	C
	T	C	G	C
	G	A		
	G	C		
	C	T		
	C	G		
Lys,Arg	G	G	G	G
	T	G	T	G
	A	G	A	G
	G	T		
	G	A		
Leu,Val Ile	T	T		
	T	C		
	C	C		

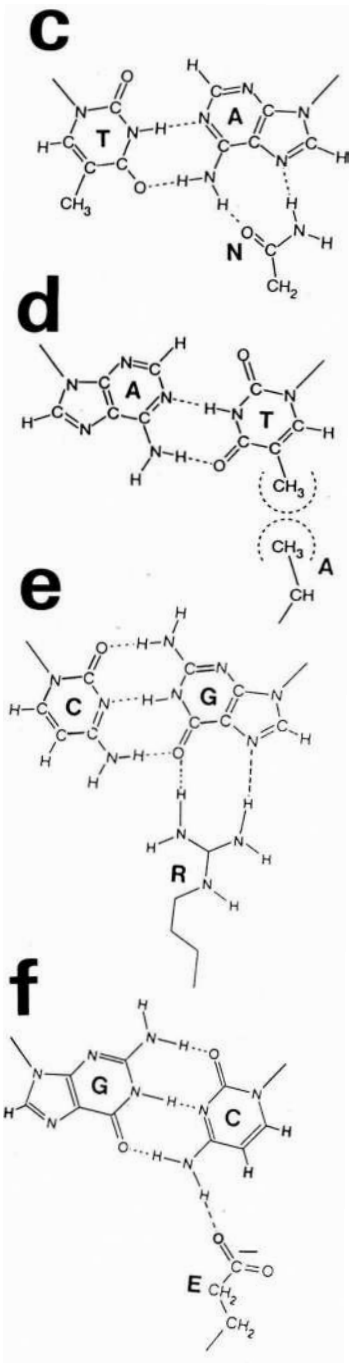

Fig. 1. Chemical code tables (**a** and **b**) and examples of amino acid–DNA base contacts (**c–f**). (a) The table for single amino acid–single base contacts. The 'specific' residue partners (see text) are shown in bold, while non-specific partners are in plain text. Chemical merit points, semi-arbitrary numbers associated with particular contacts, are used to quantify the energy and specificity of a pairing between an amino acid residue and a nucleotide base. For example, the interaction of Arg (to G), which is particularly favorable and specific, with the residue receives 15 merit points, while the interaction between Ser (to any base), which is less specific, is given 10 points. These are combined with stereochemical merit points (Figure 2) to compute a DNA–protein interaction score, as described in the text. (b) Table for the bridging of two bases by single residues: two bases on the same DNA strand (left) and two on different strands (right). (c–f) Base–residue contacts Asn–A, Ala–T, Arg–G and Glu–C are shown. All of these use hydrogen bonds except Ala–T, which involves a hydrophobic interaction.

the chemical features of amino acid residues and DNA bases (Suzuki, 1994a). The binding specificity originates mainly from chemical contacts between amino acid side chains and bases in the major groove. The chemical structure of the protein backbone and of the DNA sugar–phosphate backbone are independent of sequence; consequently, these are not the major source of discrimination (see also the discussion on Gal4 below).

Some residues have a strong binding specificity. For example, Ala, which has only the methyl group in its side chain, can interact strongly only with the T base because it is the only base which has a strongly hydrophobic group on the major groove side. Arg and Lys can bind to A, T and G bases by a hydrogen bond, but in crystal structures these bind to the G base almost exclusively. This is probably because their side chains are positively charged, while the G:C base pair is electrostatically polar and the G base is negatively charged. Others have only weak specificity and are thus less important for the discrimination. For example, Ser has a side chain which can act as either a hydrogen bond donor or a hydrogen bond acceptor, and thus can bind to any base.

It is also possible for some residues to bridge two bases either on the same DNA strand or on different DNA strands [here the two DNA strands are referred to as Watson (W) and Crick (C)]. Possible bridging partners can be listed (Figure 1b) by examining distances between the chemical groups on two bases (Suzuki, 1994a).

Obviously an amino acid residue which contacts a DNA base can be easily replaced by another of a similar size. In this respect, the 20 amino acid residues can be roughly classified into four groups: small, medium, large and aromatic. Aromatic residues have distinctive shapes but are closest in character to the large group. These chemical characteristics, the size and binding specificity of residues, are summarized into the chemical code table (Figure 1a).

Stereochemical rules

Each family of DNA binding proteins has specific DNA

Table I. DNA transcription factor complexes

Name	Resolution (Å)	*R*-factor	DNA (bp)	PDB	Reference
TR					
TrpR	1.9	0.17	19	1TRO	Otwinowski *et al.* (1988)
TrpR	2.4	0.22	19	1TRR	Lawson and Carey (1994)
C6					
Ga14	2.7	0.23	19	1D66	Marmorstein *et al.* (1992)
p53					
p53	2.2	0.21	21	–	Cho *et al.* (1994)
HTH type I					
λR	1.8	0.19	17	1LMB	Clarke *et al.* (1991)
λR	2.5	0.24	20	–	Jordan and Pabo (1988)
λC	3.9	0.50	17	4CRO	Brennan *et al.* (1990)
434R	3.2	0.30	18	–	Anderson *et al.* (1987)
434R	2.5	0.18	20	2OR1	Aggarwal *et al.* (1988)
434R	2.5	0.21	20	1RPE	Shimon and Harrison (1993)
434R	2.5	0.19	20	1PER	Rodgers and Harrison (1993)
434C	3.2	0.27	14	–	Wolberger *et al.* (1988)
434C	2.5	0.22	20	3CRO	Mondragón and Harrison (1991)
CAP	3.0	0.24	31	1CGP	Schultz *et al.* (1991)
LacR	NMR	–	11	1LCC	Chuprina *et al.* (1993)
Hin	1.8	0.23	13	1HCR	Feng *et al.* (1994)
Oct1 POU	3.0	0.24	15	1OCT	Klemm *et al.* (1994)
HTH type II					
HNF3	2.5	0.21	13	–	Clark *et al.* (1993)
ZnF					
Zif (all AF)	2.1	0.18	11	1ZAA	Pavletich and Pabo (1991)
TTK (all AF)	2.8	0.20	19	2DRP	Fairall *et al.* (1993)
GLI (F4–BF, F5–AF)	2.6	0.23	20	1GLI	Pavletich and Pabo (1993)
PH					
Matα2	2.7	0.22	21	–	Wolberger *et al.* (1991)
Eng1	2.8	0.23	21	1HDD	Kissinger *et al.* (1990)
Antp	NMR	–	14	1AHD	Billeter *et al.* (1993)
GCN4	2.9	0.23	20	1YSA	Ellenberger *et al.* (1992)
GCN4	3.0	0.22	19	1DGC	König and Richmond (1993)
E2	1.7	0.20	17	2BOP	Hegde *et al.* (1992)
MX					
Max	2.9	0.23	22	–	Ferré-D'Amaré *et al.* (1993)
USF	2.9	0.24	20	–	Ferré-D'Amaré *et al.* (1994)
MD					
MyoD	2.8	0.22	14	–	Ma *et al.* (1994)
C4					
GlucR	2.9	0.20	19	1GLU	Luisi *et al.* (1991)
EstR	2.4	0.21	18	–	Schwabe *et al.* (1993)
GATA1	NMR	–	16	1GAT	Omichinski *et al.* (1993)
MR					
MetJ	2.8	0.22	19	1CMA	Somers and Philips (1992)
ArcR	2.6	0.23	22	1PAR	Raumann *et al.* (1994)

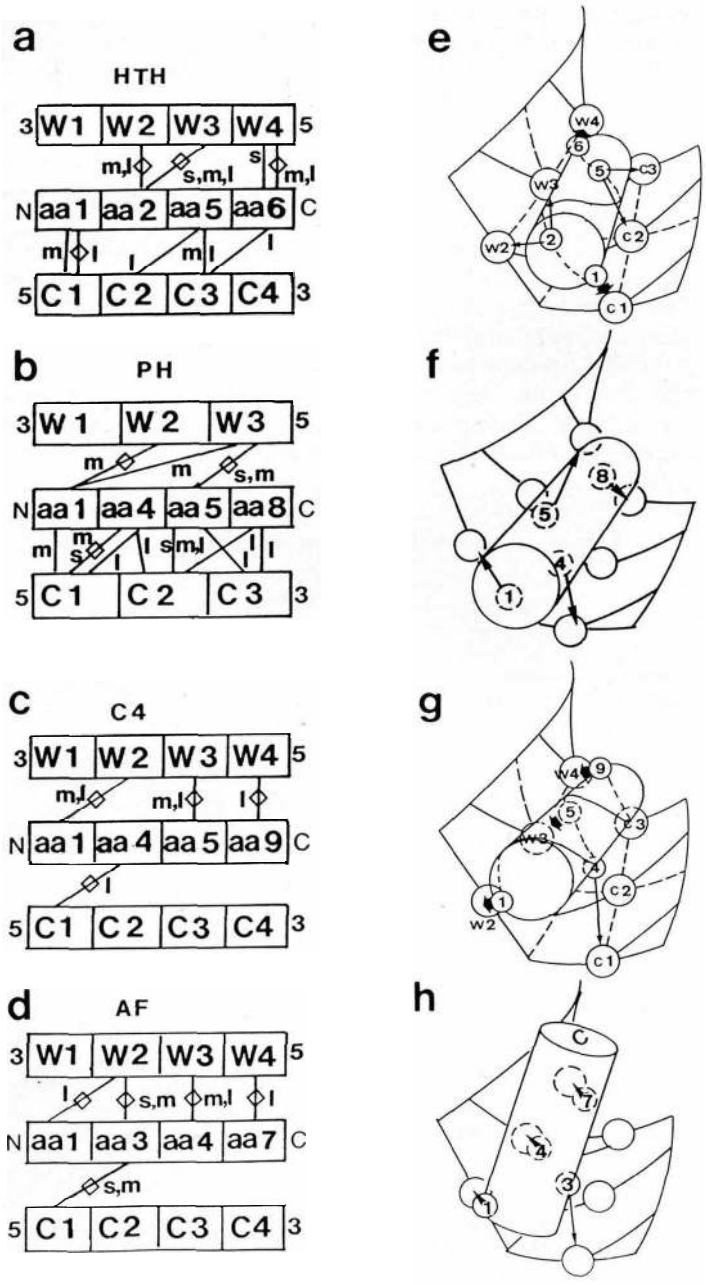

Fig. 2. Stereochemical charts (**a–d**) and base contacts (**e–h**) of HTH (a and e), PH (b and f), C4 (c and g) and AF (d and h) families, as deduced from molecular structures determined by NMR and crystallography. (a–d) Sketches of the DNA major groove with the bases W1–W4 (top) and C1–C4 (bottom), to which a recognition helix (in the central line) binds. The sizes of residues (small, s; medium, m; large, l) used for the contacts are also shown. In many cases more than one contact is possible. The optimal contacts are noted by a diamond; other potential contacts are indicated by a line. For quantitating the quality of an interaction (see text), 10 stereochemical merit points are given to the contacts marked with diamonds, while five are given to the other contacts. No stereochemical points are allotted otherwise. (e–h) The helix–groove geometry that generates the stereochemical charts depends upon patterns of interaction between residues and bases.

binding geometry (see DNA binding geometry below). As a consequence, proteins of the same family share the same pattern of contacting amino acid and base positions (Suzuki and Yagi, 1994b). The pattern can be deduced from crystal and NMR structures of DNA–protein complexes and is summarized in a stereochemical chart (Figure 2). The pattern can be improved further by using genetic and biochemical experimental data. A stereochemical chart is essentially a sketch of a recognition helix binding to the DNA major groove. Different transcription factor families adopt different binding geometries and therefore have different stereochemical charts.

In addition, to specify the residue base pairs, a stereochemical chart must include the sizes of residues in contact with DNA bases. Thus, it indicates which positions in the transcription factor specifically contact bases and shows what residue sizes are compatible with these positions. From a fixed position on the interaction surface, a long side chain can reach further into the DNA major groove, while at another position which is very close to the DNA a small residue can easily fit in but a bulky residue may not.

The stereochemical charts of the HTH, PH, AF and C4 families have been deduced. Stereochemical rules will be determined in the near future for other families, such as Myb-LexA [the protein structures have been determined by Ogata *et al.* (1994) and Fogh *et al.* (1994); we find that the two structures are very similar and their DNA binding specificity can be explained by the same stereochemical chart; Suzuki, 1995b)], LysR (its DNA binding domain has been crystallized; Tyrrell *et al.*, 1994; see also a review of the family by Schell, 1993), OmpR (its DNA binding domain has been crystallized; Kondo *et al.*, 1994), HMG (its structure has been determined in the absence of DNA; Read *et al.*, 1993; Weir *et al.*, 1993; Jones *et al.*, 1994) and HU (its structure has been determined in the absence of DNA; Tanaka *et al.*, 1984; White *et al.*, 1989; Reisman *et al.*, 1993).

Specificity of the rules

To understand the nature of the chemical and stereochemical rules further, and to test them, they were incorporated into a computer program (Suzuki and Yagi, 1994a,b). The core function of the program is to score the match between given DNA and protein sequences. This binding score is essentially the number of contacts predicted between the two sequences and thus reflects the binding energy. To calculate the binding score, points of chemical (Figure 1a) and stereochemical (Figure 2) merit were introduced. The binding score is calculated by summing over all the contacts the stereochemical merit value multiplied by the chemical merit value.

The system was tested by finding the best binding score between a given transcription factor sequence and every DNA sequence of the length (3 or 4 bp) recognized by the factor. The *in vivo* binding sequence was usually found from among a small number of DNA sequences which scored the highest (Figure 3a–c). To evaluate the specificity of the rules, a specificity index was introduced which is defined as $100 - n - (m/2)$, where n is the percentage of DNA sequences which score higher than the real binding sequence, and m is the percentage of DNA sequences which score the same as the real binding sequence (Suzuki and Yagi, 1994b). The average specificity index (which corresponds to the 'success' rate of prediction) calculated is: for PH, 96; for C4, 99; for AF, 96; and for HTH, 92. Thus, while the system does not always select the actual binding sequence as being the single optimal sequence, it does select the actual sequence as being one of the best.

Therefore, when the system was tested to find a binding site in a region of DNA known to bind the transcription factors, it had little difficulty selecting the correct position: the highest score is given to the experimentally identified binding site (Figure 3d–f). The rules are specific enough to

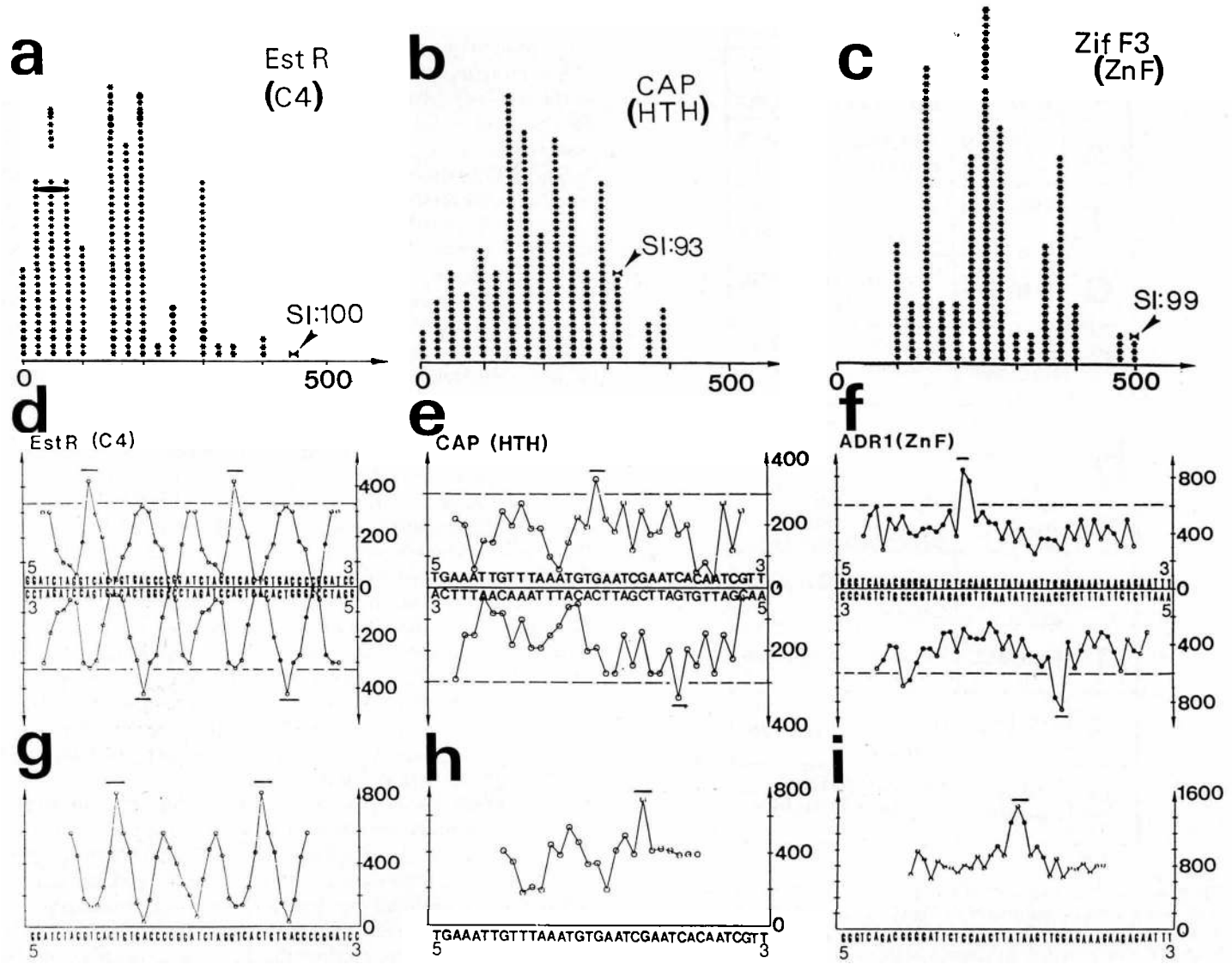

Fig. 3. Prediction of the binding sites for factors: C4, estrogen receptor (**a**, **d** and **g**); HTH, CAP (**b**, **e** and **h**); and AF, ZifF3 (**c**) and ADR1 (**f** and **i**). (a–c) The scores given to the real binding sites (marked with arrows) are compared with those given to the rest of all the possible combinations of DNA bases. The abscissae show the binding score, while the ordinates show the number of DNA sequences with that score. The specificity indices (SI) are also shown. (d–f) The binding score is calculated at every 4 bp shifting 1 bp along the DNA strand each time. The DNA sequences were taken from Deeley and Yanofsky (1992), Seiler-Tuyns *et al.* (1986) and Thukral *et al.* (1991). The experimentally identified binding sites are marked with bars. The dotted lines show the cut-off levels which separate real peaks from the background. (g–i) The binding scores of the two DNA strands are added together according to the spacing types, thus yielding enhanced discrimination of the actual binding site.

predict the DNA target of a transcription factor and thus may well be used to design a factor which would recognize a particular DNA sequence.

Further complications in DNA–protein interactions have been reported, such as water-mediated contacts (see the discussions in Feng *et al.*, 1994; Suzuki, 1994a) and contacts from outside recognition helices (see, for example, Clarke *et al.*, 1991). However, the chemical and stereochemical rules can explain the DNA binding specificity of most of the well-characterized transcription factors; thus, direct contacts from recognition helices to bases in the DNA major groove seem to be the main source of the specificity. The Trp repressor has been reported to bind to the DNA through water molecules (Otwinowski *et al.*, 1988), but similar contacts to the same DNA bases seem possible without the water molecules directly from the recognition helix (Zhang *et al.*, 1994).

The TATA-box binding protein distorts DNA largely when it binds (J.L.Kim *et al.*, 1993; Y.Kim *et al.*, 1993). The fitting of the two molecules is achieved by van der Waals contacts rather than hydrogen bonding or hydrophobic interaction. Further study is necessary to understand this binding specificity.

Recognition code table

A table which relates the amino acid sequence of a recognition helix (or sheet) with the DNA base sequence it binds can be constructed by combining the chemical code and a stereochemical chart (Suzuki, 1994b; Suzuki and Yagi, 1994b). The table can be made by picking acceptable pairs of amino acids and nucleotide bases from the chemical code table following specification of the amino acid sizes and contacts in a stereochemical chart. The resultant combined tables for C4 and for ZnF (AF) are shown in Figure 4a and b respectively. These tables can be used to predict the DNA binding specificity from a transcription factor sequence and also to design a new

a

C4	W4 aa9 size: l	W3 aa 5 size:m,l	W2 aa1 size:m,l
A	Gln,Glu	Asn,Asp, Gln,Glu	Asn,Asp, Gln,Glu
T	Leu, Met	Val, Ile, Leu,Met	Val, Ile, Leu,Met
G	Arg, Lys	His,Arg, Lys	His, Arg, Lys
C	Glu Leu, Met	Asp,Glu Leu,Met, Ile	Asp,Glu Leu,Met, Ile

b

AF	W4 aa 7 size: l	W3 aa4 size:m(l)	W2 aa l size:l
A	Gln,Glu	Asn,Asp (Gln,Glu)	Gln,Glu
T	Leu,Met	Val, Ile (Leu,Met)	Leu,Met
G	Arg,Lys	His (Arg,Lys)	Arg,Lys
C	Glu Leu,Met	Asp (Glu) Leu, Met, Ile	Glu Leu,Met,

Fig. 4. Recognition code tables of C4 (**a**) and AF (**b**). The code tables are made by choosing the columns from Figure 1a according to the residue sizes specified in Figure 2c and d. The interaction of hydrophobic residues with the C base is weaker; therefore it is shown in plain characters instead of bold. Position 4 in AF can be occupied by medium or large residues, but a medium residue is preferable; the large residues are shown in parentheses.

transcription factor which could bind to a particular DNA sequence.

Müller-Hill and co-workers have been studying the binding specificity of HTH proteins (Jansen *et al.*, 1987; Kisters-Woike *et al.*, 1991; Lehming *et al.*, 1991) and basic zipper proteins (J.Kim *et al.*, 1993; Suckow *et al.*, 1993, 1994) systematically. Code-oriented mutagenesis experiments have been carried out on ZnF (Desjarlais and Berg, 1993, 1994; Choo and Klug, 1994a,b; Reber and Pabo, 1994). Similar but less intensive studies have been published on homeoproteins (Hanes and Brendt, 1991; Treisman *et al.*, 1992; Dear *et al.*, 1993). These results coincide well with the recognition rules discussed above [for example, compare figure 1 of Suzuki and Yagi (1994a) with Figure 2 of Choo and Klug (1994b)].

DNA binding geometry

Different binding geometries, each of which is specific to a transcription factor family, are the bases for the sets of stereochemical rules. Two structural factors, tight fitting of the DNA and protein surfaces and matching of the residue and base positions, determine the geometries.

Binding geometry of a recognition helix

Like protein–protein interactions (Janin and Chothia, 1990), DNA–protein interactions require tight fitting of the two surfaces. They involve 20–80 chemical contacts between the two molecules and the burial of ~2800 Å^2 of accessible surface area (Hegde *et al.*, 1992).

The DNA major groove is more than wide enough to accommodate an α-helix, so more than one close-fitting binding geometry is possible (Figure 5f). The binding geometry is restricted further by the requirements of sequence-specific DNA–protein interactions: it seems that at least three contacts are needed between a recognition helix and DNA bases to confirm the sequence specificity. The requirement for matching residue and base positions can be understood as the pairing of a line connecting residue positions and another line connecting base positions (Suzuki, 1994b; M.Suzuki and M.Gerstein, manuscript submitted).

Binding geometry is determined by three types of residue position found around a recognition helix: (i) those which contact DNA bases, (ii) those which contact DNA phosphates, and (iii) those which interact with the rest of the protein (Figure 5b). Many residue positions which are routinely used for identifying DNA binding motifs, such as the hydrophobic position in HTH and the Cys and His residues in C4 and ZnF, fall into group (iii). Residues in group (iii) do not interact with DNA, so they must be placed facing away from the DNA, thus limiting the rotation of the helix. The way these three types are arranged into a single recognition helix is specific to each family of DNA binding proteins; this is the reason why each family has its specific binding geometry and thus its specific stereochemical chart.

If the DNA major groove was filled with water up to the height of the sugar–phosphate backbones, a recognition helix binding to the DNA is found half 'sunk' in the 'sea' (Figure 5b). Type (ii) residues are found on the watermark around the helix; type (i) residues are found in the 'wet' area; type (iii) residues are found in the 'dry' area. The shape of the watermark can be examined by cutting the helix and opening it flat (Figure 5c and d; note that the α-helix surface is seen from inside the helix). The binding geometry can be predicted from such a 'watermark' plot. The helical wheel projection, which is often used for similar arguments, is not as useful unless the helix binds parallel to the DNA major groove.

As an α-helix is essentially straight, it cannot follow many DNA bases around the curved major groove [see, for example, the discussion on the segmented helix in Suzuki *et al.* (1994b)]. Consequently, a single recognition helix can access only one side of the DNA (Figure 5b) and can bind to five consecutive base pairs or less, usually 4 bp (Figure 5a). No more than three turns of a straight α-helix can be involved in the recognition because the pitch of an α-helix is 5.4 Å, so the DNA-facing side of three turns spans 10.8 Å and the diameter of bases around the DNA helix axis is ~10 Å (Figure 5b).

Binding geometry of a β-sheet

The two antiparallel strands in a β-sheet are twisted and thus the sheet has a curvature (Chothia, 1984): at one face the surface of a β-sheet is concave and at another face it is convex. Therefore, two ways of fitting a β-sheet into the DNA major groove are possible (Suzuki, 1995a): facing either its concave (the convergent fitting) or its convex (the divergent fitting) surface to the DNA. Since the local DNA major groove is deepest in the middle, the divergent fitting is appropriate and

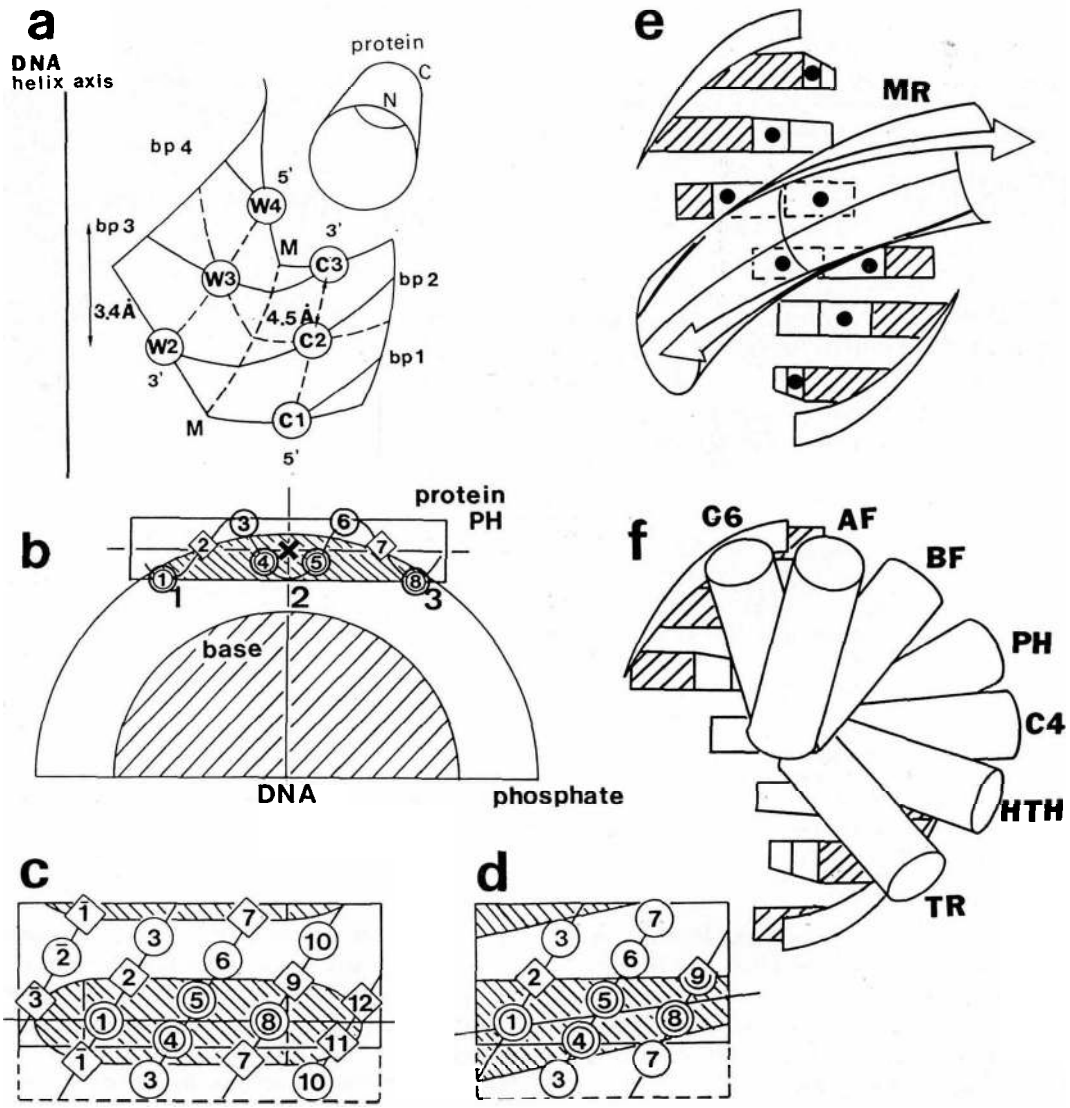

Fig. 5. DNA binding geometry of an α-helix and a β-sheet. (**a**) A DNA major groove is shown near a recognition helix. These bases on the Watson strand (W2–W4) and on the Crick strand (C1–C3) can make contacts with the helix. Note that this means that two pairs of bases and two unpaired bases make contacts. Because of the curvature of the DNA and the rigidity of the helix, it is more difficult for positions W1 or C4 to contact the helix. (**b**) A recognition helix (PH) of three turns drawn along the DNA helix. Only the half of the DNA facing the protein is displayed. The residues shown with double circles bind to DNA bases, those shown with diamonds bind to DNA phosphates, and those with single circles face away from the DNA. The center of the helix is marked with an ×. The numbers 1–3 show the first, second and third turns, respectively. (**c** and **d**) Watermark plots of PH (c) and C4 (d). These plots are produced by classifying the residue positions as in (b), and then cutting the helix open along one line (near residues 4 and 1) and viewing the flattened surface from the side that has been inside the helix. The lines centering the contacting surfaces are shown. Notice that in PH (c), which binds parallel to the major groove (f), the watermark is uniformly shaped and the center runs parallel to the helix axis. C4 (d), in contrast, has a recognition helix which binds perpendicularly to the DNA helix axis, to make enough base–residue contact so that the helix must be angled into the major groove. This results in a wedge-shaped watermark, off the helix axis. (**e**) A β-sheet of the MR family binds to the DNA major groove in the divergent fitting mode. In this mode, the β-sheet 'dives' into the groove (described by a line) in the center. Bases of the major contacting positions are marked (●). (**f**) A recognition helix can be positioned in the major groove in a variety of different ways, all of which preserve close contacts. The geometries of seven different families of transcription factors (C6, AF, BF, PH, C4, HTH and TR) are shown.

is actually used for binding by the MetJ family (Figure 5e). Because of the curved nature of a β-sheet, it can fit into the DNA major groove better than an α-helix does (see also Phillips, 1994), and follows 6 bp contacting four to six bases of the eight base positions closest to the sheet (C1–C4 and W2–W6). Important distances for the match here are those between the i-th and $i + 2$-th residues on a β-strand (13.2 Å) and between the j-th and $j + 3$-th bases on the DNA (13.5 Å).

The DNA minor groove is narrower; unless the DNA is greatly distorted, a β-sheet cannot go down to the bottom of the groove because it is blocked by the sugar–phosphate backbone. The local sugar–phosphate backbones are closest to the protein in the middle, and thus the convergent fitting is predicted to be used [see the discussion on DNA binding by the HU family in White *et al.* (1989)].

Recognition helix and protein folding

The same type of protein fold can adopt more than one type of recognition helix, and the same type of recognition helix can be incorporated into more than one type of protein fold (Suzuki, 1995b; Suzuki *et al.*, 1995). Historically, 'DNA

Table II. Recognition types of transcription factor

Spacer type	Number of bases in the direct binding site			Combination of different helices	6 (β-sheet)
	3 (α-helix)	4 (α-helix)	5 (α-helix)		
$S^C + 1$	PH(HDZip,bZip) [4]				
$S^C + 2$	PH(bZip,bHLH) [5]	C4(ThyR) [6]		HTH(4)–PH(3) [5.5 Oct1]	
$S^N + 2$		HTH(RafR) [6]			
$S^N + 3$		HTH(EbgR,MalR) [7]			
$S^N + 4$		HTH(LacR,GalR) [8]			
$S^C + 5$		C4(EstR,GlcR) [9]			
$S^C + 6$		HTH(CAP,434C,434R,16–3R) [10]			
$S^N + 6$	C6(PPR1) [9]	HTH(DeoR) [10]			
$S^C + 7$		HTH(λR,λC) [11]			
$S^C + 8$	TR [11]				
$S^N + 8$		HTH(CytR) [12]			
$S^N + 10$	C6(PUT3) [13]	HTH(P22C,P22R) [14]			
$S^N + 11$	C6(Gal4) [14]				
ST + 2					MR(MetJ) [8]
ST + 3					MR(MntR) [9]
ST + 4					MR(TraY) [10]
ST + 5					MR(ArcR) [11]
T − 1		AF–AF [3]	BF–BF [4]		
T + 1				BF(5)–AF(4) [5.5]	
T + 3		C4(RXR)–C4(RAR), C4(RXR)–C4(COUP), C4(RXR)–C4(PPAR), C4(RXR)–C4(RXR) [7]			
T + 5		C4(RXR)–C4(VDR) [9]			
T + 6		C4(RXR)–C4(ThyR) [10]			
T + 7		C4(RXR)–C4(RAR) [11]			

The 'spacer number' is defined as the number of base pairs between two direct binding sites. The numbers of base pairs between two centers of recognition elements (the 'spacing' number) are shown in square parentheses. The table shows type and examples: S^C, symmetric, C-terminus central; S^N, symmetric, N-terminus central; ST, symmetric and tandem; T, tandem.

binding' motifs were identified using conserved amino acid positions, i.e. by their fold. Amino acid residues which contact DNA bases are not conserved among a family; by changing these residues the protein family discriminates between DNA sequences. Therefore, when a term such as 'helix–turn–helix' is used it is not clear whether the word implies a particular protein fold or a particular type of recognition helix often found in classic prokaryotic HTH proteins. This complication has created much confusion. Moreover, it should be remembered that a protein which has the HTH protein fold is not necessarily involved in DNA binding at all (Brennan and Matthews, 1989).

The homeodomain has the HTH protein fold, but the residue positions used for contacting DNA bases in the homeo recognition helix is shifted by one turn towards the C-terminus (counted from the conserved hydrophobic position), and it has an extended C-terminus which has many basic residues to bind to DNA phosphates. As a consequence, the homeo recognition helix does not use the HTH-type DNA binding geometry, but in fact uses the PH type, the type to which basic zipper proteins belong (Suzuki, 1993). Indeed, some homeoproteins have a zipper following the recognition helix (HDzip proteins; see Sessa *et al.*, 1993).

To understand the DNA recognition rules, the type of a recognition helix is more important than the whole protein fold. Thus in Table I we show the classification of transcription factors according to the types of recognition helix.

Recognition rules at higher levels

The number of DNA base pairs which can be contacted by a recognition element (either a sheet or a helix) is limited. To recognize more DNA bases, two or more elements are used in combination, essentially by either relating the two with a 2-fold symmetry axis or repeating them in tandem.

The classic HTH proteins and zipper proteins of the PH family use 'symmetrical' arrangements (denoted S), while ZnF proteins (AF and BF) use a 'tandem' arrangement (denoted T). Different C4 proteins use both types of arrangement (Umesono *et al.*, 1991). The number of spacer base pairs between the two symmetrical binding sites is different in classic HTH proteins. This is one of the main reasons why DNA–HTH protein complexes, as a whole, look very different from each other [see figure 7.27 of Branden and Tooze (1991)], even when the mode of helix–DNA interaction is the same (Suzuki *et al.*, 1995). To connect two direct binding sites, structural features of the spacer DNA may play important roles [see the discussion on the spacer DNA in the 434 operators by Drew *et al.* (1990); Koudelka, 1993; also compare the crystal structures of the 434 repressor binding to different operators in Aggarwal *et al.* (1988); Rodgers and Harrison, 1993; Shimon and Harrison, 1993)].

Symmetrical arrangements can be characterized by (i) whether the C-termini (S^C) or the N-termini (S^N) are closer to the dyad axis and (ii) the number of 'spacer bases' between the two binding sites (Table II). By combining the number of base pairs in the direct binding site (four for HTH) and the spacing type ($S^C + 6$ for CAP), the recognition type of a transcription factor can be described as $4(S^C + 6{:}2)$ for CAP (Note that 2 is the number of recognition helices).

Knowledge of the spacing type permits the improved calculation of specificity. For example, when the binding scores calculated for CAP monomer binding to each DNA strand are shifted by 6 bp and added together, the new plot that results shows a clearer peak (Figure 3h).

Gal4 and related proteins of the C6 family share the same recognition helix and thus the same local binding specificity. The recognition helix binds to the DNA, facing its C-terminus towards the DNA. Most of the important contacts in this binding geometry are made from the CO groups at the C-terminal edge of the helix (Marmorstein *et al.*, 1992); it would be very difficult to modify such binding specificity by changing the side chains of residues. Instead, the C6 family discriminates between DNA sequences by changing the number of spacer DNA base pairs (Reece and Ptashne, 1993; Suzuki and Yagi, 1995a). For such discrimination the spacing type is essential.

The spacing type of the A-type ZnF is T −1, i.e. two neighboring fingers share 1 bp (−1) in a tandem (T) arrangement. A single finger appears to be incapable of discriminating between DNA sequences, but the combination of two or three fingers seems to be sufficient [see figure 9 of Suzuki *et al.* (1994a)]. This explains why fingers are always found in a repeat. The two ADR1 (AF) binding sites in its regulatory DNA region are likely to be recognized by a symmetrical dimer of ADR1 molecules, each of which has two ZnF in tandem (T −1), with the superspacing type $S^C + 6$ [Figure 3i; its recognition type is 4(T–1:2)(S^C + 6:2)]. Thus, the communication between DNA and proteins can be described with increasing specificity, from the chemical, to the stereochemical, to the spacing to the superspacing levels.

A protein of the MR family produces a recognition sheet on the interaction surface of the two monomers, and thus each recognition site on DNA is essentially 2-fold symmetrical. Such protein dimers interact further to recognize tandem repeats of the binding sites (Phillips, 1994; Suzuki and Yagi, 1995b). Because of the 2-fold symmetry in each dimer, the two dimers are related by another 2-fold symmetry (ST). The recognition type of the MR is described as 6(ST + 2:*n*).

New spacing rules will be described on the basis of structural (Clore *et al.*, 1994; Gronenborn and Clore, 1994) and biochemical (Alberti *et al.*, 1993) studies of the oligomerization of transcription factors.

Acknowledgements

We dedicate this paper to Max Perutz on the occasion of his 80th birthday. We thank Drs C.Chothia and H.Nakamura for their continuous encouragement. M.G. acknowledges support from a Damon Runyon-Walter Winchell fellowship (DRG-1272). S.E.B. acknowledges support from Herchel Smith Harvard scholarship and St John's College Benefactors scholarship.

References

Aggarwal,A.K., Rodgers,D.W., Drottar,M., Ptashne,M. and Harrison,S.C. (1988) *Science*, **242**, 899–907.
Alberti,S., Oehler,S., von Wilcken-Bergmann,B. and Müller-Hill,B. (1993) *EMBO J.*, **12**, 3227–3236.
Anderson,J.E., Ptashne,M. and Harrison,S.C. (1987) *Nature*, **326**, 846–852.
Billeter,M., Quian,Y.Q., Otting,G., Müller,M., Gehring,W. and Wüthrich,K. (1993) *J. Mol. Biol.*, **234**, 1084–1094.
Branden,C. and Tooze,J. (1991) *Introduction to Protein Structure*. Garland Publishing, New York.
Brennan,R.G. and Matthews,B.W. (1989) *J. Biol. Chem.*, **264**, 1903–1906.
Brennan,R.G., Roderick,S.L., Takeda,Y. and Matthews,B.W. (1990) *Proc. Natl Acad. Sci. USA*, **87**, 8165–8169.
Cho,Y., Gorina,S., Jeffrey,P.D. and Pavletich,N.P. (1994) *Science*, **265**, 346–355.
Choo,Y. and Klug,A. (1994a) *Proc. Natl Acad. Sci. USA*, **91**, 11163–11167.
Choo,Y. and Klug,A. (1994b) *Proc. Natl Acad. Sci. USA*, **91**, 11168–11172.
Chothia,C. (1984) *Annu. Rev. Biochem.*, **53**, 537–572.
Chuprina,V.P., Rullmann,J.A.C., Lamerichs,R.M.J.N., Van Boom,J.H., Boelens,R. and Kaptein,R. (1993) *J. Mol. Biol.*, **234**, 446–462.
Clark,M.L., Halay,E.D., Lai,E. and Barley,S.K. (1993) *Nature*, **364**, 412–420.
Clarke,N.D., Beamer,L.J., Goldberg,H.R., Berkower,C. and Pabo,C.O. (1991) *Science*, **254**, 267–270.
Clore,M.G., Omicinski,J.G., Sakaguchi,K., Zambrano,N., Sakamoto,H., Appella,E. and Gronenborn,A.M. (1994) *Science*, **265**, 386–391.
Crick,F.H.C. (1958) *Symp. Soc. Exp. Biol.*, **12**, 548–555.
Dear,T.N., Sanchez-Garicia,I. and Rabbitts,T.H. (1993) *Proc. Natl Acad. Sci. USA*, **90**, 4431–4435.
Deeley,M. and Yanofsky,C. (1992) *J. Bacteriol.*, **151**, 942–951.
Desjarlais,J.R. and Berg,J.M. (1993) *Proc. Natl Acad. Sci. USA*, **90**, 2256–2260.
Desjarlais,J.R. and Berg,J.M. (1994) *Proc. Natl Acad. Sci. USA*, **91**, 11099–11103.
Drew,H.R., McCall,M.J. and Calladine,C.R. (1990) In Cozzarelli,N. and Wang,J.C. (eds), *DNA Topology and its Biological Effects*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 1–56.
Ellenberger,T.E., Brandl,C.S., Struhl,K. and Harrison,S.C. (1992) *Cell*, **71**, 1223–1237.
Fairall,L., Schwabe,J., Chapman,L., Finch,J.T. and Rhodes,D. (1993) *Nature*, **366**, 483–487.
Feng,J.-A., Johnson,R.-C. and Dickerson,R.E. (1994) *Science*, **263**, 348–355.
Ferré-D'Amaré,A.R., Prendergast,G.C., Ziff,E.B. and Burley,S.K. (1993) *Nature*, **363**, 38–45.
Ferré-D'Amaré,A.R., Pognonec,P., Roeder,R.G. and Burley,S.K. (1994) *EMBO J.*, **13**, 180–189.
Fogh,R.H., Ottleben,G., Rüterjans,H., Schnarr,M., Boelens,R. and Kaptein,R. (1994) *EMBO J.*, **13**, 3936–3944.
Gronenborn,A.M. and Clore,G.M. (1994) *Science*, **263**, 536.
Hanes,S.D. and Brendt,R. (1991) *Science*, **251**, 426–430.
Hegde,R.S., Grossman,S.R., Laimins,L.A. and Sigler,P.B. (1992) *Nature*, **359**, 505–512.
Janin,J. and Chothia,C. (1990) *J. Biol. Chem.*, **265**, 16027–16030.
Jansen,C., Gronenborn,A.M. and Clore,G.M. (1987) *Biochem. J.*, **246**, 227–232.
Jones,D.M.N., Seales,M.A., Shaw,G.L., Churchill,M.E.A., Ner,S.S., Keeler,J., Travers,A.A. and Neuhaus,D. (1994) *Structure*, **2**, 609–627.
Jordan,S.R. and Pabo,C.O. (1988) *Science*, **242**, 893–899.
Kim,Y., Geiger,J.H., Hahn,S. and Sigler,P.B. (1993a) *Nature*, **365**, 512–520.
Kim,J.L., Nikolov,D.B. and Burley,S.K. (1993b) *Nature*, **365**, 520–527.
Kim,J., Tzamarisa,D., Ellenberger,T., Harrison,S.C. and Struhl,K. (1993c) *Proc. Natl Acad. Sci. USA*, **90**, 4513–4517.
Kissinger,C.R., Liu,B., Martin-Blanco,E., Kornberg,T.B. and Pabo,C.O. (1990) *Cell*, **63**, 579–590.
Kisters-Woike,B., Lehming,N., Sartorius,J., von Wilcken-Bergmann,B. and Müller-Hill,B. (1991) *Eur. J. Biochem.*, **198**, 411–419.
Klemm,J.D., Rould,M.A., Aurora,R., Herr,W. and Pabo,C.O. (1994) *Cell*, **77**, 21–32.
Klevit,R.E. (1991) *Science*, **253**, 1367–1393.
Kondo,H., Miyaji,T., Susuki,M., Tate,S., Mizuno,T., Nishimura,Y. and Tamaka,I. (1994) *J. Mol. Biol.*, **235**, 780–782.
König,P. and Richmond,T. (1993) *J. Mol. Biol.*, **233**, 139–154.
Koudelka,G.B. (1993) In Eckstein,F. and Lilley,D.M.J. (eds), *Nucleic Acids and Molecular Biology 7*. Springer-Verlag, Berlin, Germany, pp. 16–27.
Landschultz,W.H., Johnson,P.F. and McKnight,S.L. (1988) *Science*, **240**, 1759–1764.
Lawson,C.L. and Carey,J. (1994) *Nature*, **366**, 178–182.
Lehming,N., Sartorius,J., Kisters-Woike,B., von Wilcken-Bergmann,B. and Müller-Hill,B. (1991) In Ecstein,F. and Lilley,D.M.J. (eds), *Nucleic Acids and Molecular Biology 5*. Springer-Verlag, Berlin, Germany, pp. 114–125.
Luisi,B.F., Xu,X.W., Otwinowski,Z., Freedman,L.P., Yamamoto,K.R. and Sigler,P.B. (1991) *Nature*, **352**, 497–505.
Ma,P.C., Rould,M.A., Weintraub,H. and Pabo,C.O. (1994) *Cell*, **77**, 451–459.
Marmorstein,R., Carey,M., Ptashne,M. and Harrison,S.C. (1992) *Nature*, **356**, 409–414.
Matthews,B.W. (1988) *Nature*, **335**, 294–295.
Miller,J., McLachlan,A.D. and Klug,A. (1985) *EMBO J.*, **4**, 1609–1614.
Mondragón,A. and Harrison,S.C. (1991) *J. Mol. Biol.*, **219**, 321–334.
Ogata,K., Morikawa,S., Nakamura,H., Serikawa,A., Inoue,T., Kanai,H., Sarai,A., Ishii,S. and Nishimura,Y. (1994) *Cell*, **79**, 639–648.
Omichinski,J.G., Clore,G.M., Schaad,O., Felsenfeld,G., Trainor,C., Appella,E., Stah,S.J. and Gronenborn,A.M. (1993) *Science*, **261**, 438–446.
Otwinowski,Z., Schevitz,R.W., Zhang,R.-G., Lawson,C.L., Loachimiak,A., Marmorstein,R.Q., Luisi,B.F. and Sigler,P.B. (1988) *Nature*, **335**, 321–329.
Pabo,C.O. and Sauer,R.T. (1984) *Annu. Rev. Biochem.*, **53**, 293–321.
Pabo,C.O., Aggarwal,A.K., Jordan,S.R., Beamer,L.J., Obeysekare,U.R. and Harrison,S.C. (1990) *Science*, **247**, 1210–1213.
Pavletich,N.P. and Pabo,C.O. (1991) *Science*, **252**, 809–817.
Pavletich,N.P. and Pabo,C.O. (1993) *Science*, **261**, 1701–1707.
Phillips,S.E.V. (1994) *Annu. Rev. Biophys. Biomol. Struct.*, **23**, 671–701.

Raumann,B.E., Rould,M.A., Pabo,C.O. and Sauer,R.T. (1994) *Nature*, **367**, 754–757.
Read,C.M., Cary,P.D., Crane-Robinson,C., Driscoll,P.C. and Norman,D.G. (1993) *Nucleic Acids Res.*, **21**, 3427–3436.
Rebar,E.J. and Pabo,C.O. (1994) *Science*, **263**, 671–673.
Reece,R.J. and Ptashne,M. (1993) *Science*, **261**, 909–910.
Reisman,J.M., Hsu,V.L., Jariel-Encontre,I., Lecou,C., Sayre,M.H., Kearns,D.R. and Parello,J. (1993) *Eur. J. Biochem.*, **213**, 865–873.
Rodgers,D.W. and Harrison,S.C. (1993) *Structure*, **1**, 227–240.
Sauer,R., Yocum,R.R., Doolittle,R.F., Lewis,M. and Pabo,C. (1982) *Nature*, **298**, 447–451.
Schell,M.A. (1993) *Annu. Rev. Microbiol.*, **47**, 597–626.
Schultz,S.C., Shields,G.C. and Steitz,T.A. (1991) *Science*, **253**, 1001–1007.
Schwabe,J.W., Chapman,L., Finch,J.T. and Rhodes,D. (1993) *Cell*, **75**, 567–578.
Seeman,N.C., Rosenberg,J.M. and Rich,A. (1976) *Proc. Natl Acad. Sci. USA*, **73**, 804–808.
Seiler-Tuyns,A., Walker,P., Martinez,E., Mérillat,A.-M., Givel,F. and Wahli,W. (1986) *Nucleic Acids Res.*, **14**, 8755–8770.
Sessa,G., Morelli,G. and Ruberti,L. (1993) *EMBO J.*, **12**, 3507–3517.
Shimon,L.J.W. and Harrison,S.C. (1993) *J. Mol. Biol.*, **232**, 826–838.
Somers,W.S. and Philips,S.E. (1992) *Nature*, **359**, 387–393.
Suckow,M., von Wilcken-Bergmann,B. and Müller-Hill,B. (1993) *EMBO J.*, **12**, 1193–1200.
Suckow,M., Madan,A., Kisters-Woike,B., von Wilcken-Bergmann,B. and Müller-Hill,B. (1994) *Nucleic Acids Res.*, **22**, 2198–2208.
Suzuki,M. (1993) *EMBO J.*, **12**, 3221–3226.
Suzuki,M. (1994a) *Structure*, **2**, 317–327.
Suzuki,M. (1994b) *Proc. Jap. Acad.*, **B70**, 96–99.
Suzuki,M. (1995a) *Protein Engng*, **8**, 1–4.
Suzuki,M. (1995b) *Proc. Jap. Acad.*, **B71**, 217–231.
Suzuki,M. and Chothia,L. (1994) *Proc. Jap. Acad.*, **B70**, 58–61.
Suzuki,M. and Yagi,N. (1994a) *Proc. Jap. Acad.*, **B70**, 62–66.
Suzuki,M. and Yagi,N. (1994b) *Proc. Natl Acad. Sci. USA*, **91**, 12357–12361.
Suzuki,M. and Yagi,N. (1995a) *Proc. Jap. Acad.*, **B71**, 817–891.
Suzuki,M. and Yagi,N. (1995b) *Proc. Jap. Acad.*, **B71**, 108–113.
Suzuki,M., Gerstein,M. and Yagi,N. (1994a) *Nucleic Acids Res.*, **22**, 3387–3405.
Suzuki,M., Neuhaus,D., Gerstein,M. and Aimoto,S. (1994b) *Protein Engng*, **7**, 461–470.
Suzuki,M., Yagi,N. and Gerstein,M. (1995) *Protein Engng*, **8**, 329–338.
Tanaka,I., Appelt,K., Dijk,J. and White,S.W. (1984) *Nature*, **310**, 376–381.
Thukral,S.K., Eisen,A. and Young,E.T. (1991) *Mol. Cell. Biol.*, **11**, 1566–1577.
Treisman,J., Harris,E., Wilson,D. and Desplan,C. (1992) *BioEssays*, **14**, 145–150.
Tyrrell,R., Davies,G.J., Wilson,K.S. and Wilkinson,A.J. (1994) *J. Mol. Biol.*, **235**, 1159–1161.
Umesono,K., Murakami,K.K., Thompson,C.C. and Evans,R.M. (1991) *Cell*, **65**, 1255–1267.
Weir,H.M., Kraulis,P.J., Hill,C.S., Raine,A.R.C., Laue,E.D. and Thomas,J.O. (1993) *EMBO J.*, **12**, 1311–1319.
White,S.W., Appelt,K., Wilson,K.S. and Tanaka,I. (1989) *Proteins*, **5**, 281–288.
Wolberger,C., Dong,Y., Ptashne,M. and Harrison,S.C. (1988) *Nature*, **335**, 789–795.
Wolberger,C., Vershon,A.K., Liu,B., Johnson,A.D. and Pabo,C.O. (1991) *Cell*, **67**, 517–528.
Zhang,H., Zhao,D., Revington,M., Lee,W., Jia,X., Arrowsmith,C. and Jardetzky,O. (1994) *J. Mol. Biol.*, **238**, 592–614.

Received September 6, 1994; revised November 8, 1994; accepted January 17, 1995