20. DNA-Recognition Code Tables for Transcription Factors in the C4 Zinc Binding and C2H2 Zinc Finger Families

By Masashi SUZUKI

MRC Laboratory of Molecular Biology, Hills Road,
Cambridge, CB2 2QH, U.K.

(Communicated by Setsuro EBASHI, M. J. A., June 7, 1994)

Abstract: A simple way of understanding the structure of the DNA recognition code of transcription factors is shown taking two DNA binding motifs as examples (the C4 class zinc binding proteins and the C2H2 zinc fingers). Interaction between DNA and factors of the motifs can be understood as that between two lines, a line connecting residue positions on the surface of a recognition helix and a line connecting base positions in the DNA major groove. The base-residue pairing along the two lines are summarised in the stereochemical charts. The DNA recognition table for each of the two motifs, which can be used for prediction and design of recognition helix-DNA interactions, can be created by choosing residue-base partners from the chemical code table according to the stereochemical chart specific to the motif.

Key words: DNA-protein interaction; DNA recognition code; transcription factor; steroid hormone receptor; GATA; zinc finger.

1. **Introduction.** Understanding the general mechanisms of sequence specific DNA-protein recognition is a major challenge in structural biology. My collaborators and I have devised a method to analyse patterns of amino acid-DNA base contacts and have applied the method to several DNA-binding motifs.¹⁻⁴⁾ The contacts are analysed at two levels: at the "chemical" level, the contacts are understood as amino acid-base partners which can interact with each other through hydrogen bonding or hydrophobic interaction, and at the "stereochemical" level, they are understood in terms of amino acid and base positions available for the contacts. The "chemical" rules are general among all binding motifs, while the "stereochemical" rules are a consequence of a particular binding geometry and thus are specific to a particular DNA-binding motif.

The two types of rules are precise enough to predict the recognised DNA sequence from the protein sequence of a transcription factor and therefore constitute the DNA recognition code; a full statement will be published elsewhere (Suzuki and Yagi, in preparation). In this paper, I describe a simple way of understanding the DNA recognition code using two DNA binding motifs as examples (the C4 class zinc binding proteins^{3),4)} which include steroid hormone receptors and the GATA family, and the C2H2 class zinc fingers^{5),6)}).

2. **DNA-binding geometry.** The amino acid positions of C4 and C2H2 used for base-recognition can be understood by an analysis similar to that used for α -helix interactions.⁷⁾ When an α -helix is cut out through its surface along the N to C axis (Fig. 1b) and is opened flat on a surface (Fig. 1a), the positions which pack against another α -helix to α -helix are found often in the two types of lines⁷⁾ (see Fig. 1a): one type connects, for example 1, 5, 9, (referred as the $i+4n$ type), and the other type connects, for example, -1, 3, 6 (referred as the $i+3n$ type).

These two types of lines are used for binding to DNA bases as well. The C4 family

Fig. 1. Cylindrical (a) and wheel (c, d) projections of an α -helix. An α -helix is cut out through its surface along the N-C axis (b) and opened flat on a surface (a). In (c) and (d) the helix is viewed through the helix axis from the N-terminus.

Fig. 2. Contacts between residue and base positions for C4 Zn binding (a) and C2H2 Zn finger (b) transcription factors. These are sketches of the major groove of DNA (the boxes show base positions) to which a recognition helix binds (the three circles in the central line show amino acid positions). The arrows indicate the contacts. The sizes of the residues are shown with L (large) and M (medium). The W-II base might be recognised by a large residue at position 2 of a Zn finger but a medium residue is preferable (Suzuki *et al.*, in preparation).

uses four positions in its recognition helix for base recognition^{3),4)} and the three out of the four, 1, 5, 9, are found in a line of the $i+4n$ type. The type A C2H2 family (see Suzuki *et al.*, submitted, for the definition of the A/B types of Zn fingers) uses four positions^{5),6)} and the three out of the four positions, -1, 3, 6, are found in a line of the $i+3n$ type. The three positions of each family bind to three consecutive base positions in DNA by simple one residue to one base relationship (see Fig. 2). Therefore, if the fourth position in each helix

which is out of the line is ignored, the rules relating a transcription factor sequence and a DNA sequence become very simple. This does not happen with all DNA-binding motifs. For example, the probe helix family^{1),2)} uses positions 1, 4, 5 and 8 (Fig. 1a). The line connecting these positions becomes an intermediate between the $i+3n$ type and the $i+4n$ type and a simple one base-one residue correlation does not take place.¹⁾

Positions 1, 5, and 9 cluster together on one side of the helix (if the helix is looked along its helix axis, the three positions span 80 degrees), while positions -1, 3 and 6 spread more widely (spanning 120 degrees) (Fig. 1c). Therefore, the latter line is more curved around the helix surface. This might be the reason why position -1 of C2H2 adopts a slightly irregular geometry; precisely speaking, the helix does not start with residue -1 but with the residue after.

3. **Code tables.** To create the DNA-recognition code, it is essential to understand the chemical rules. In my previous paper,¹⁾ I produced a table which summarises the binding specificity of amino acid residues. Here I rearrange the table according to the base partners [A, T, G, C, on the rows] and the sizes [small, medium, large, aromatic on the columns] (Fig. 3). There are two types of base-residue partners: specific (Fig. 3a) and non-specific (Fig. 3b). For instance, Arg could bind to three bases, A, T, G, by a H-bond but in fact almost exclusively binds to the G base.¹⁾ This is probably because the G base is negatively charged, while Arg has a positive charge. Therefore Arg-G pairing is important (or specific), while Arg-A and Arg-T pairing is less so. Also some residues such as Ser can bind to any base and are non-specific.

The sizes of residues used for each contact to a DNA base are another consequence of the binding geometry and are described as part of the stereochemical rules.¹⁾ These have been already identified for C4 (ref. 3) and for C2H2 (Suzuki *et al.*, submitted); shown in Fig. 2 with M (medium) and L (large). Therefore, by choosing the columns of the right amino acid sizes from Fig. 3a (for example, the column of L in Fig. 3a is chosen for recognition of the first base (W-I) in the C4-binding site) the simplified version of the

a. Specific partners.

Base	aa (small)	aa (medium)	aa (large)	aa (aromatic)
A		Asn, Asp	Gln, Glu	
T	Ala	Val, Ile	Leu, Met	Tyr, Phe, Trp
G		His	Arg, Lys	
C	(Val)	Asp, (Ile)	Glu, (Leu, Met)	(Tyr, Phe, Trp)

b. Non-specific partners.

Base	aa (small)	aa (medium)	aa (large)	aa (aromatic)
A	Ser, Cys, Thr		Met, Arg, Lys	Tyr, Trp
T	Ser, Cys, Thr	Asn, His	Gln, Arg, Lys	
G	Ser, Cys, Thr	Asn	Gln	Tyr
C	Ser, Cys, Thr	Asn, His	Gln	

Fig. 3. Chemical code tables of protein-DNA interaction. The chemical rules, which list amino acid-base partners for contacts, can be summarised into the shape of a table. Recognition of the C base by hydrophobic residues is possible but may be less specific than that of the T base. See text and also ref. 1 for full understanding of the tables.

a. C4.

	Base: WI aa:9 size:L	Base:W II aa:5 size:M, L	Base:W III aa:1 Size:M, L
A	Gln, Glu	Asn, Gln, Asp, Glu	Asn, Gln, Asp, Glu
T	Leu, Met	Val, Ile, Leu, Met	Val, Ile, Leu, Met
G	Arg, Lys	His, Arg, Lys	His, Arg, Lys
C	Glu (Ile)	Asp, Glu (Leu, Met, Ile)	Asp, Glu (Leu, Met, Ile)

b. Zn finger type A.

	Base: W I aa:6 size:L	Base:W II aa:3 size:M[L]	Base:W III aa:-1 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. DNA-recognition code tables for the C4 Zn binding (a) and C2H2 Zn finger (b) families. The code tables are made by choosing the appropriate columns from Fig. 3a according to the amino acid sizes shown in Fig. 2.

DNA-recognition code tables for these DNA-binding motifs are completed (Fig. 4).

These code tables can be used for predicting the binding specificity from a protein sequence and for designing new factors which would bind to a particular sequence of DNA. For instance, in C4, to recognise the A base at W-I, position 9 of the helix should be occupied by the large residue Gln and if a C4 factor has Gln at 9, it likely binds to A^(W-I). It should be noted that these tables are a simplified version and for full understanding of the code, a slightly more complicated discussion is necessary (Suzuki and Yagi, in preparation).

Acknowledgements. I thank Dr. C. Chothia for stimulative discussion and Dr. N. Yagi and Mr. S. E. Brenner for critical reading of the manuscript.

References

- 1) Suzuki, M. (1994): Structure, 2, 317-326.
- 2) — (1993): EMBO J., 12, 3221-3226.
- 3) Suzuki, M., and Chothia, L. (1994): Proc. Japan Acad., 70B, 58-61.
- 4) Suzuki, M., and Yagi, N. (1994): ibid., 70B, 62-66.
- 5) Pavletich, N. P., and Pabo (1991): Science, 252, 809-817.
- 6) Fairall, L. *et al.* (1993): Nature, 366, 483-487.
- 7) Chothia, C., Levitt, M., and Richardson, D. (1977): Proc. Natl. Acad. Sci. U.S.A., 74, 4130-4134.