Proc. Natl. Acad. Sci. USA
Vol. 91, pp. 12357–12361, December 1994
Biophysics

DNA recognition code of transcription factors in the helix–turn–helix, probe helix, hormone receptor, and zinc finger families

(DNA–protein interaction/homeodomain/leucine zipper/transcription factor GATA)

MASASHI SUZUKI* AND NAOTO YAGI†

*Medical Research Council Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, United Kingdom; and †Tohoku University, School of Medicine, Seiryo-machi, Sendai, 980-77, Japan

Communicated by Tadamitsu Kishimoto, August 8, 1994

ABSTRACT We have previously reported that in four transcription factor families the DNA-recognition rules can be described as (*i*) chemical rules, which list possible pairings between the 20 amino acid residues and the four DNA bases, and (*ii*) stereochemical rules, which describe the base and amino acid positions in contact. We have incorporated these rules into a computer program and examined the nature of the rules. Here we conclude that the DNA recognition rules are simple, logical, and consistent. The rules are specific enough to predict DNA-binding characteristics from a protein sequence.

A large number of transcription factors, which play dominant roles in transcription regulation by binding to different DNA sequences, have been identified. Since the three-dimensional structure of a protein is uniquely fixed by its amino acid sequence, basic rules are expected, which would predict the DNA-binding specificity from a transcription factor sequence. But, since the initial expectation of such rules (the recognition code) (1), many structural biologists have expressed skepticism about their existence (for example, see ref. 2).

The crystal structures of a number of transcription factor–DNA complexes have been determined (3–27); also a considerable amount of biochemical, genetic, and statistical information about the binding specificity of transcription factors is available (28–34). By using these data, we have devised a method of analyzing the patterns of contacts between DNA bases and amino acid residues (35–40) and have described the DNA-recognition rules of four transcription factor families: the probe helix (PH), which includes homeo and zipper proteins (35, 36); the helix–turn–helix (HTH) (M.S. and M. Gerstein, unpublished results); the zinc finger (ZnF) (37, 38); and the C4 Zn-binding proteins (C4), which include hormone receptors and GATA proteins (38–40). These rules concern contacts from amino acid side chains in a recognition helix to DNA bases in the major groove.

The aim of this paper is to establish a framework of DNA-recognition rules common to the four families and to examine whether, from the nature of the rules, they constitute a recognition code.

Framework of the DNA Recognition Rules

The DNA-recognition rules are of two types, chemical and stereochemical. The chemical rules list possible pairing partners of amino acid side chains and DNA bases through hydrogen bonding or hydrophobic interaction (Fig. 1*a*; ref. 36). The sizes of residues are also important; from a fixed position on an interaction surface, a longer side chain can reach a more distant part of the DNA. The residues are classified roughly into four groups—small, medium, large, and aromatic (Fig. 1*a*; ref. 36). These chemical rules are *general* for any binding motif.

The inclination of the recognition helix in the major groove of DNA is fixed by the structural elements specific to a DNA-binding motif. For instance, a recognition helix of PH has conserved Arg/Lys positions, which bind to DNA phosphates and thereby fix the binding geometry (35, 36). As a consequence, each binding motif uses a set of particular amino acid positions for base recognition. These can be easily summarized into a chart with specifications of the sizes of residues used; each DNA-binding motif has its own *specific* stereochemical chart (Fig. 1 *b–e*). ZnF motifs can be subdivided into two groups (37), but here only the larger group is discussed (A fingers).

Binding Score

We have incorporated the rules into a computer program, which is written in the C programming language and implemented under the Unix operating system. Its core function is to score the match between the given DNA and protein sequences. This binding score is essentially the number of contacts predicted between the two sequences and reflects the binding energy.

To calculate the binding score, points for stereochemical (see the legend to Fig. 1 *b–e*) and chemical (Fig. 1*a*) merits are introduced. The binding score is calculated as the sum over all the contacts of (stereochemical merit point) × (chemical merit point) for each interaction. The chemical merit points given to different base–residue partners are not always the same (Fig. 1*a*). For instance, Arg and Lys could bind by a hydrogen bond to T, G, or A. But in fact they recognize the G base almost exclusively (36), because the G base in a G·C pair is electrically polar (negatively charged), while Arg and Lys have a positive charge. Therefore, binding of Arg or Lys to G should be given more points than to T or A. Similarly, not all the contacts in the stereochemical charts appear to be equally important (refs. 36 and 37; M.S. and M. Gerstein, unpublished results), and this is reflected in differences in the two grades of stereochemical merit points (see contacts marked with diamonds and those not in Fig. 1 *b–e*).

Often several different sets of contacts are possible for given protein and DNA sequences. In this case, the pairing with the highest score is chosen. However, it is stereochemically forbidden to make two contacts that cross over each other in the chart. For instance, in Fig. 1*c* aa 5 can contact C3, and aa 8 can contact C2, but not simultaneously. As an example, the binding score of CAP (Fig. 2*h*) is the sum of the products of the chemical and stereochemical points for

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: PH, probe helix; HTH, helix–turn–helix; ZnF, zinc finger; C4, C4 Zn-binding protein.

a chemical

	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**

recognition code

f

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

g

ZnF	W4 aa6 size: l	W3 aa3 size:m(l)	W2 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,

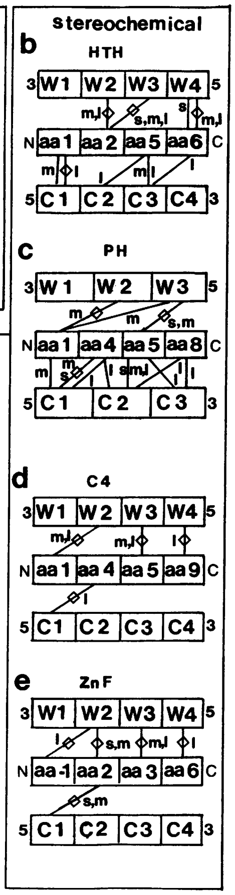

FIG. 1. Chemical (*a*) and stereochemical (*b–e*) rules that make the DNA-recognition code and code tables for C4 (*f*) and ZnF (*g*). (*a*) The chemical merit points are also shown. Residues in boldfaced letters are those important for specificity (specific residues). (*b–e*) Sketches of the DNA major groove with the bases, W1–W4 (top strand) and C1–C4 (bottom strand), to which a recognition helix (in the middle line) binds. The sizes of residues, small (s), medium (m), and large (l), used for the contacts are also shown. Aromatic residues may often be included with the large group. Ten stereochemical merit points are given to the contacts marked with diamonds and five to the other contacts. No stereochemical points are given otherwise. If a hydrophobic interaction takes place to a T base and if one of the two neighboring bases is another T, an additional 3 points is added to the chemical merit point, since this is likely to enhance the hydrophobic environment. The binding specificity of Asn (aa 1) of PH is affected by Asn (aa 2) through side chain–side chain interactions (36); if Asn occupies position 2, Asn (aa 1) interacts with Asn (aa 2) and binds to A (W2), but if not Asn (aa 2) bridges the C1 and W2 bases. For this reason, if position 2 is occupied by Asn, the chemical merit point of Asn (aa 1) to A (W2) is kept at 15; if not, it is decreased to 10 and the residue is allowed to bind to the C1 base at the same time. When a single residue binds to two bases simultaneously, the two contacts are handled independently. This is to simplify the computer program, although the two bases bridged in this way are limited and can be handled as a set (36). The code tables (*f* and *g*) are made by choosing the columns from *a* according to the residue sizes specified in *d* and *e*. The interaction of hydrophobic residues to the C base is weaker and therefore is shown by plain instead of boldfaced letters. Position 3 in ZnF can be occupied by a medium or large residue, but a medium residue is preferable (37); the large residues are shown in the parentheses.

the Arg·G, Arg·G, and Glu·C contacts, respectively—(10 × 15) + (5 × 15) + (10 × 12) = 345.

Consistency and Specificity of the Rules

The DNA recognition rules were originally deduced from 25 crystal structures (3–27) and many other transcription factors whose binding specificity has been characterized by genetic or biochemical experiments (see the references cited in refs. 35–40).

Contacts were predicted by the program for 73 recognition helices: those of 10 PH proteins, 20 HTH proteins, 38 ZnF proteins (specific or very specific A fingers listed in ref. 37), and 5 C4 proteins (selected examples are shown in Fig. 2).

In most examples, the predicted contacts are essentially the same as those observed or predicted in earlier work. Thus the rules can consistently explain the amino acid–base contacts. However, this does not necessarily suggest that the rules can explain how factors discriminate between the target and other DNA sequences; if many other DNA sequences were recognized by a factor in similar ways, the factor could not choose the correct site. We now examine this aspect (specificity) of the rules in two ways.

We first compare the binding score given to the real binding site with those for sites consisting of all other possible base combinations (Fig. 3). HTH, C4, and ZnF recognize four base pairs, which have 256 possible combinations. PH recognizes three base pairs, and the number of combinations is 64. In our calculation, the real binding sequence is usually found among a small number of DNA sequences that score the highest (Fig. 3); the rules are sufficiently specific to exclude the rest of the DNA sequences, which score less. To evaluate the specificity of the rules, we introduce the specificity index, which is defined as $(100 - n - \frac{m}{2})\%$, where n is the percentage of the DNA sequences that score higher than the real binding sequence and m is that of the DNA sequences that score the same as the real binding sequence. If a factor has two natural binding sequences—sequence i, which scores higher than sequence j—n is defined as the percentage that scores higher than i, and m is defined as the percentage that scores between i and j. The average indices calculated for the factors are 93% (PH) (96% if Max is excluded, which is further discussed in M.S. and M. Gerstein, unpublished results), 99% (C4), 96% (ZnF), and 92% (HTH).

As a second test we now examine the DNA sequence of a region regulated by a transcription factor *in vivo*. When the binding score is calculated for every four base pairs along the DNA, shifting one base pair at a time, the highest score is given for the experimentally identified binding site (Fig. 4). Since DNA has two strands, the score must be calculated along each of the two strands.

The above two tests have shown that the rules are highly specific. In the crystal structures, some additional contacts are seen from outside a recognition helix, but the binding specificity of a recognition helix seems to be essentially sufficient to specify uniquely the DNA-binding sites.

Spacing Type

An α-helix can bind to no more than five base pairs because of the curvature of the DNA major groove; it can access only one side of the DNA (44). To recognize more than five base pairs, two or more helices are used in combination, essentially by either relating the two with a twofold symmetry axis or repeating them in tandem. The classic HTH proteins and zipper proteins of the PH family use "symmetrical" arrangements (denoted here as S), while ZnF proteins use a "tandem" arrangement (denoted here as T). C4 proteins use both types of arrangements (45).

Symmetrical arrangements can be characterized by whether the C terminus (denoted with the "+" sign) or the N terminus (denoted with the "−" sign) is closer to the dyad axis and the number of bases along the DNA between the two binding sites (for example, S +6 for the HTH protein CAP). By knowing the spacing type, the plot of the binding score can be improved. When the binding scores of the two DNA strands for CAP binding are shifted by six base pairs and added to each other, the

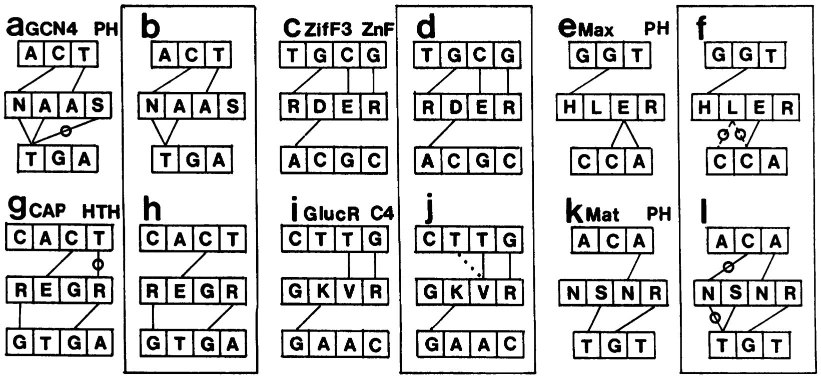

FIG. 2. Comparison between contacts observed in the crystal structures (*a*, *c*, *e*, *g*, *i*, and *k*) and computer-predicted contacts (*b*, *d*, *f*, *h*, *j*, and *l*). The figures are drawn in the same way as in Fig. 1. The dotted line (· · · ·) in *j* shows an additional predicted hydrophobic interaction to the neighboring T base. A pair of two dashed lines (– – –) in *f* show two alternative contacts with the same score. The contacts that are predicted but not observed and those observed but not predicted are marked with circles. The side chain of Asn (aa 1) in Matα2 (*k*) is not well described in the original report of the crystal structure (4). The residue is predicted to contact the C (C1) and T (W2) bases (*l*). Leu (aa 4) of Max is predicted to make contacts with C (C1) or C (C2) (*f*). The figures of the original report (5) show that this leucine does seem close to C (C1), but the coordinates have not been published and the paper does not mention this contact.

new plot shows a clearer peak (Fig. 4*e*). Thus, a weaker binding specificity of a HTH recognition helix (see the previous section) is compensated by combining two such helices.

The spacing type of the majority of ZnF proteins is T −1 [i.e., two neighboring fingers share one base pair (−1) in a tandem (T) arrangement (37)]. A single finger appears to be incapable of discriminating between DNA sequences, but the combination of two or three fingers does seem to be sufficient (see figure 9 of ref. 37). This can explain why fingers are always found in a repeat.

The two experimentally identified ADR1 (ZnF)-binding sites in its regulatory DNA region are predicted successfully (Fig. 4*c*). The two sites are likely to be recognized by a symmetrical dimer of ADR1 molecules, each of which has two ZnF motifs in tandem (T −1), with the superspacing type of S +6 (Fig. 4*c*). Therefore, the communication between DNA and proteins can be described with increasing accuracy, from the chemical, the stereochemical, the spacing to the superspacing levels.

Prediction and Design

Our computer program successfully identifies the binding sites of transcription factors whose binding specificities have been characterized experimentally. Therefore, it may be natural to expect that it can (*i*) predict the yet unknown binding specificity of a protein sequence and (*ii*) design a factor that would recognize a particular DNA sequence.

In the ZnF and C4 families, a simple table relating DNA and protein sequences can be produced (Fig. 1 *f* and *g*; ref. 38). Three residues of C4—1, 5, and 9—bind to the three consecutive bases W2–W4, by a simple one residue–one base relationship, while ZnF positions −1, 3, and 6 bind to W2–W4. Therefore, by choosing specific partner residues in the correct columns from Fig. 1*a* according to the amino acid sizes shown in Fig. 1 *d* and *e*, recognition tables for the three positions from two types can be constructed (see ref. 38 for further discussion).

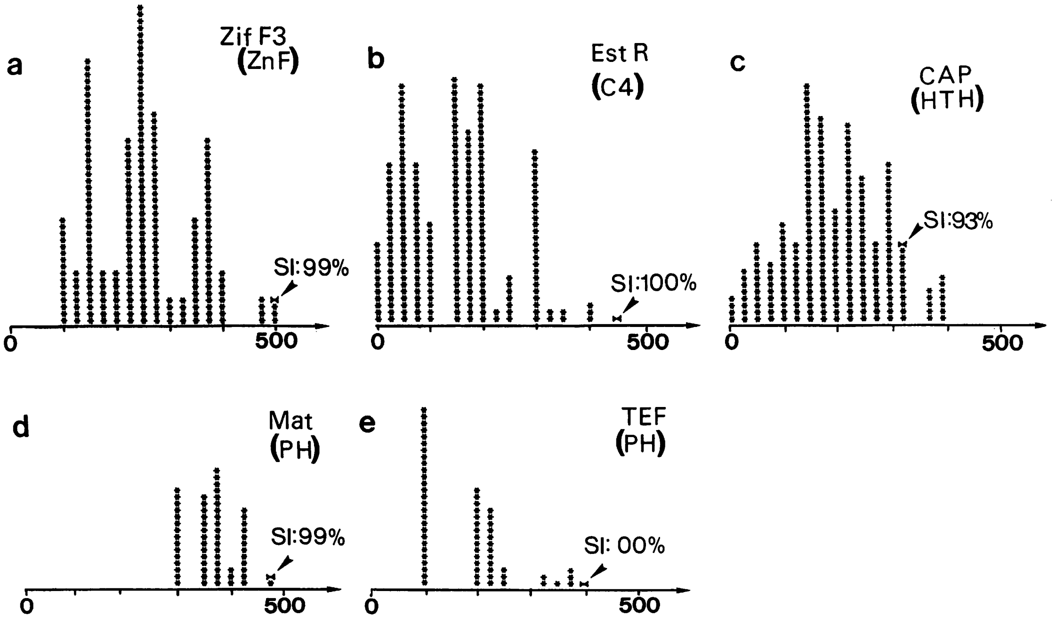

FIG. 3. Distribution of the binding scores for Zif268 finger 3 (ZnF) (*a*), estrogen receptor (C4) (*b*), CAP (HTH) (*c*), Matα2 (PH) (*d*), and TEF (PH) (*e*). The scores given to the real binding sites (marked with arrowheads) are compared with those given to the rest of all the possible combinations of DNA bases. The abscissas show the binding score, while the ordinates show the number of DNA sequences with that score. The specificity index (SI) is also shown. Note that TEF has Asn (aa 1) and Asn (aa 2) but Matα2 has Asn only at position 1 (see legend to Fig. 1).

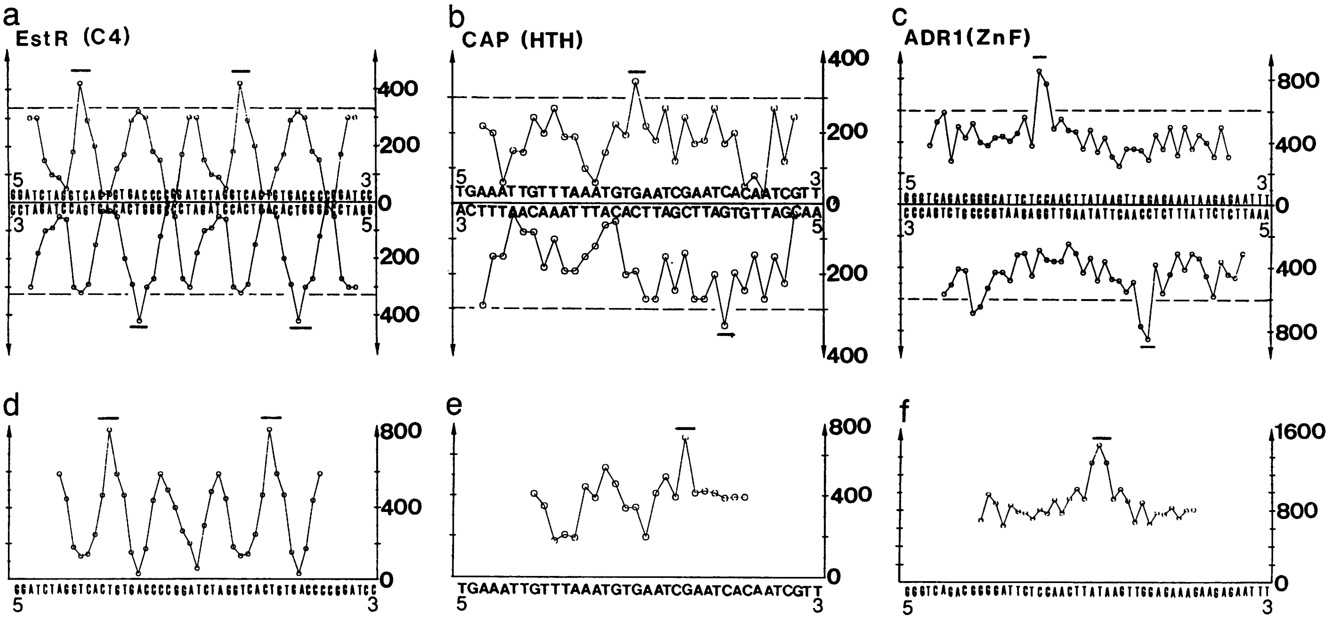

FIG. 4. Prediction of the binding sites for factors: estrogen receptor (C4) (*a* and *d*), CAP (HTH) (*b* and *e*), and ADR1 (ZnF) (*c* and *f*). (*a–c*) The binding score is calculated at every four base pairs by shifting one base pair along the DNA strand at a time. The DNA sequences were taken from refs. 41–43. The experimentally identified binding sites are marked with bars. The dotted lines show the cut-off levels, which separate real peaks from the background. (*d–f*) The binding scores to the two DNA strands are added to each other according to the spacing types. Note that a new peak for a dimer turns up in the center of two monomer binding sites on different DNA strands. The spacing types of symmetrical arrangements identified are S +1, PH (HDZip, bZip); S +2, PH (bZip, bHLH) and C4 (ThyR); S −2, HTH (RafR); S −3, HTH (EbgR, MalR) S −4, HTH (LacR, GalR); S +5, C4 (EstR, GlcR; although the three base pairs at the center of the binding sites are often described as the spacer, because these sequences vary, here the five base pairs that are not contacted by the recognition helices are defined as the spacer); S −6, HTH (DeoR) and C6 (PPR1); S +6, HTH (CAP, 434C, 434R, 16-3R); S +7, HTH (λR, λC); S +8, TrpR; S +8, HTH (CytR); S −10, HTH (P22C, P22R, LexA) and C6 (PUT3), S −11, C6 (Gal4). The spacing types of tandem arrangements identified are T −1, ZnF(A)–ZnF(A); T O,ZnF(B)–ZnF(B); T +1, ZnF(B)–ZnF(A); T +3, C4(RXR)–C4(RAR), C4(RXR)–C4(COUP), C4(RXR)–C4(PPAR), C4(RXR)–C4(RXR); T +5, C4(RXR)–C4(VDR); T +6, C4(RXR)–C4(ThyR); T +7, C4(RXR)–C4(RAR).

The rules will be further improved as information becomes available. For example, in this study, changes in the DNA structure upon binding proteins and the sequence-dependent differences in the DNA structures are ignored. However, the framework and the major features of the rules are unlikely to change. We have shown that the DNA-recognition rules for well-characterized factors in the four families are simple, logical, consistent, and specific. We therefore believe that these rules constitute the DNA-recognition code.

We thank Drs. C. Chothia, J. Finch, and A. Klug and Mr. S. E. Brenner for their critical reading of the paper.

1. Pabo, C. O. & Sauer, R. T. (1984) *Annu. Rev. Biochem.* **53,** 293–321.
2. Matthews, B. W. (1988) *Nature (London)* **335,** 294–295.
3. Pabo, C. O., Aggarwal, A. K., Jordan, S. R., Beamer, L. J., Obeysekare, U. R. & Harrison, S. C. (1990) *Science* **247,** 1210–1213.
4. Wolberger, C., Vershon, A. K., Liu, B., Johnson, A. D. & Pabo, C. O. (1991) *Cell* **67,** 517–528.
5. Ferré-D'Amaré, A. R., Prendergast, G. C., Ziff, E. B. & Burley, S. K. (1993) *Nature (London)* **363,** 38–45.
6. Ellenberger, T. E., Brandl, C. S., Struhl, K. & Harrison, S. C. (1992) *Cell* **71,** 1223–1237.
7. König, P. & Richmond, T. (1993) *J. Mol. Biol.* **233,** 139–154.
8. Ferré-D'Amaré, A. R., Pognonec, P., Roeder, R. G. & Burley, S. K. (1994) *EMBO J.* **13,** 180–189.
9. Clarke, N. D., Beamer, L. J., Goldberg, H. R., Berkower, C. & Pabo, C. O. (1991) *Science* **254,** 267–270.
10. Schwabe, J. W., Chapman, L., Finch, J. T. & Rhodes, D. (1993) *Cell* **75,** 567–578.
11. Omichinski, J. G., Clore, G. M., Schaad, O., Felsenfeld, G., Trainor, C., Appella, E., Stah, S. J. & Gronenborn, A. M. (1993) *Science* **261,** 438–446.
12. Kissinger, C. R., Liu, B., Martin-Blanco, E., Kornberg, T. B. & Pabo, C. O. (1990) *Cell* **63,** 579–590.
13. Hegde, R. S., Grossman, S. R., Laimins, L. A. & Sigler, P. B. (1992) *Nature (London)* **359,** 505–512.
14. Jordan, S. R. & Pabo, C. O. (1988) *Science* **242,** 893–899.
15. Anderson, J. E., Ptashne, M. & Harrison, S. C. (1987) *Nature (London)* **326,** 846–852.
16. Aggarwal, A. K., Rodgers, D. W., Drottar, M., Ptashne, M. & Harrison, S. C. (1988) *Science* **242,** 899–907.
17. Wolberger, C., Dong, Y., Ptashne, M. & Harrison, S. C. (1988) *Nature (London)* **335,** 789–795.
18. Mondragon, A. & Harrison, S. C. (1991) *J. Mol. Biol.* **219,** 321–334.
19. Rodegers, D. W. & Harrison, S. C. (1993) *Structure* **1,** 227–240.
20. Shultz, S. C., Shields, G. C. & Steitz, T. A. (1991) *Science* **253,** 1001–1007.
21. Brennan, R. G., Roderick, S. L., Takeda, Y. & Matthews, B. W. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 8165–8169.
22. Feng, J.-A., Johnson, R.-C. & Dickerson, R. E. (1994) *Science* **263,** 348–355.
23. Clark, M. L., Halay, E. D., Lai, E. & Barley, S. K. (1993) *Nature (London)* **364,** 412–420.
24. Pavletich, N. P. & Pabo, C. O. (1991) *Science* **252,** 809–817.
25. Fairall, L., Schwabe, J., Chapman, L., Finch, J. T. & Rhodes, D. (1993) *Nature (London)* **366,** 483–487.
26. Pavletich, N. P. & Pabo, C. O. (1993) *Science* **261,** 1701–1707.
27. Luisi, B. F., Xu, X. W., Otwinowski, Z., Freedman, L. P., Yamamoto, K. R. & Sigler, P. B. (1991) *Nature (London)* **352,** 497–505.
28. Seeman, N. C., Rosenberg, J. M. & Rich, A. (1976) *Proc. Natl. Acad. Sci. USA* **73,** 804–808.
29. Lehming, N., Sartorius, J., Kisters-Woike, B., von Wilcken-Bergmann, B. & Müller-Hill, B. (1991) in *Nucleic Acids and Molecular Biology*, eds. Eckstein, F. & Lilley, D. M. J. (Springer, Heidelberg), Vol. 5, pp. 114–125.
30. Kisters-Woike, B., Lehming, N., Sartorius, J., von Wilcken-

Bergmann, B. & Müller-Hill, B. (1991) *Eur. J. Biochem.* **198**, 411–419.

31. Desjarlais, J. R. & Berg, J. M. (1993) *Proc. Natl. Acad. Sci. USA* **90**, 2256–2260.
32. Klevit, R. E. (1991) *Science* **253**, 1367–1393.
33. Suckow, M., von Wilcken-Bergmann, B. & Müller-Hill, B. (1993) *EMBO J.* **12**, 1193–1200.
34. Treissman, J., Harris, E., Wilson, D. & Desplan, C. (1992) *BioEssays* **14**, 145–150.
35. Suzuki, M. (1993) *EMBO J.* **12**, 3221–3226.
36. Suzuki, M. (1994) *Structure* **2**, 317–326.
37. Suzuki, M., Gerstein, M. & Yagi, N. (1994) *Nucleic Acids Res.* **22**, 3397–3405.
38. Suzuki, M. (1994) *Proc. Jpn. Acad.* **B70**, 96–99.
39. Suzuki, M. & Chothia, L. (1994) *Proc. Jpn. Acad.* **B70**, 58–61.
40. Suzuki, M. & Yagi, N. (1994) *Proc. Jpn. Acad.* **B70**, 62–66.
41. Deeley, M. & Yanofsky, C. (1992) *J. Bacteriol.* **151**, 942–951.
42. Seiler-Tuyns, A., Walker, P., Martinez, E., Mérillat, A.-M., Givel, F. & Wahli, W. (1986) *Nucleic Acids Res.* **14**, 8755–8770.
43. Thukral, S. K., Eisen, A. & Young, E. T. (1991) *Mol. Cell. Biol.* **11**, 1566–1577.
44. Suzuki, M., Neuhaus, D., Gerstein, M. & Aimoto, S. (1994) *Protein Eng.* **7**, 461–470.
45. Umesono, K., Murakami, K. K., Thompson, C. C. & Evans, R. M. (1991) *Cell* **65**, 1255–1267.