The EMBO Journal vol.11 no.12 pp.4507-4517, 1992

Determination of the base recognition positions of zinc fingers from sequence analysis

Grant H.Jacobs

Structural Studies Division, MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

Communicated by A.Klug

The CC/HH zinc finger is a small independently folded DNA recognition motif found in many eukaryotic proteins, which ligates zinc through two cysteine and two histidine ligands. A database of 1340 zinc fingers from 221 proteins has been constructed and a program for analysis of aligned sequences written. This paper describes sequence analysis aimed at determining the amino acid positions that recognize the DNA bases, by comparing two types of sequence variation. Using the idea that long runs of adjacent zinc fingers have arisen from internal gene duplication, the conservation of each position of the finger within the runs was calculated. The conservation of each position of the finger between homologous proteins from different species was also noted. A correlation of the two types of conservation showed clusters of related amino acids. One cluster of three positions was found to be especially variable within long runs, but highly conserved between corresponding fingers of homologous proteins; these positions are predicted to be the base contact positions. They match the amino acid positions that contact the bases in the co-crystal structure determined by Pavletich and Pabo [*Science*, 240, 809-817 (1991)]. An adjacent cluster of four positions on the plot may also be associated with DNA binding. This analysis shows that the base recognition positions can be identified even in the absence of a known structure for a zinc finger. These results are applicable to zinc fingers where the structure of the complex is unknown, in particular suggesting that the individual finger-DNA interaction seen in the Zif268-DNA structure has been conserved in many zinc finger-DNA interactions.

Key words: DNA binding/sequence analysis/zinc fingers

Introduction

The CC/HH zinc finger motif is a small independently folding DNA recognition unit, with a consensus of $^{Y}/_{F}$ X C $X_{2,4}$ C X_3 F X_5 L X_2 H X_{3-5} H (for reviews see Rhodes and Klug, 1988; El-Baradi and Pieler, 1991; Neuhaus and Rhodes, 1991; Kaptein, 1991, 1992). This motif is found in a large number of eukaryotic proteins, many of which have been demonstrated to interact directly with DNA.

The first zinc finger protein to be identified, TFIIIA (Ginsberg *et al.*, 1984), was observed to have nine repeats now known as the 'zinc finger' motif (Böhm and Drescher, 1985; Brown *et al.*, 1985; Miller *et al.*, 1985). It was suggested that this repeat represented a repeated structural unit, each of which bound a single zinc ion, and that TFIIIA had arisen from internal duplication of the gene (Miller *et al.*, 1985). Quantification of the zinc content of the protein and proteolysis studies supported this idea (Miller *et al.*, 1985). Subsequently the genomic sequence of TFIIIA was determined (Tso *et al.*, 1986) which further supported the idea of repeated structural units in that seven of the nine zinc fingers were each encoded by a single exon.

Many footprinting studies of zinc finger-DNA complexes have been done, in particular on the TFIIIA-ICR interaction (e.g. Fairall *et al.*, 1986; Rhodes and Klug, 1986; Vrana *et al.*, 1988; Churchill *et al.*, 1990).

Model structures for the zinc finger motif have been put forward by Berg (1988) and Gibson *et al.* (1988). The structure of a zinc finger was first determined by NMR studies (Párraga *et al.*, 1988; Lee *et al.*, 1989) and resembled more closely the model of Berg. Since then several NMR structures for zinc fingers have been published (e.g. Klevit *et al.*, 1990; Omichinski *et al.*, 1990, 1992; Kochoyan *et al.*, 1991; Neuhaus, *et al.*, 1992).

Recently a co-crystal structure for the complex of a three finger peptide with its DNA site has been published (Pavletich and Pabo, 1991). This structure clearly resembles the model for the interaction proposed by Nardelli *et al.* (1991) and supported by their mutagenesis studies.

Well over 200 zinc finger genes have now been reported, and it has been estimated that there are several hundred zinc finger encoding genes in the human genome (Bellefroid *et al.*, 1989; Crossley and Little, 1991). The large number of zinc finger sequences provides an opportunity to learn more about their structure and function through their sequences. This paper describes sequence analysis aimed at determining the functional (i.e. DNA binding) positions of the zinc finger motif using a database of 1340 fingers from 221 proteins (Jacobs and Michaels, 1990). The analysis presented uses the idea that long runs of adjacent zinc fingers are likely to have evolved by internal duplication. A subset of long runs that are good candidates to have evolved by internal duplication was created, and the conservation of each position noted. Similarly the conservation of each position in corresponding zinc fingers from closely related protein sequences was noted. The most conserved region in these putatively homologous fingers ought to correspond to the DNA recognition surface of the zinc finger. If long runs have, on the whole, evolved either by duplication and divergence or by duplication and conservation of the base recognition of each finger, the positions that contact the bases can be identified as being the most variable or most conserved positions in long runs found within the most conserved region of the homologous fingers. The method of distinguishing neutral and functional variation described here should be applicable to other proteins containing repetitive, independently folding units.

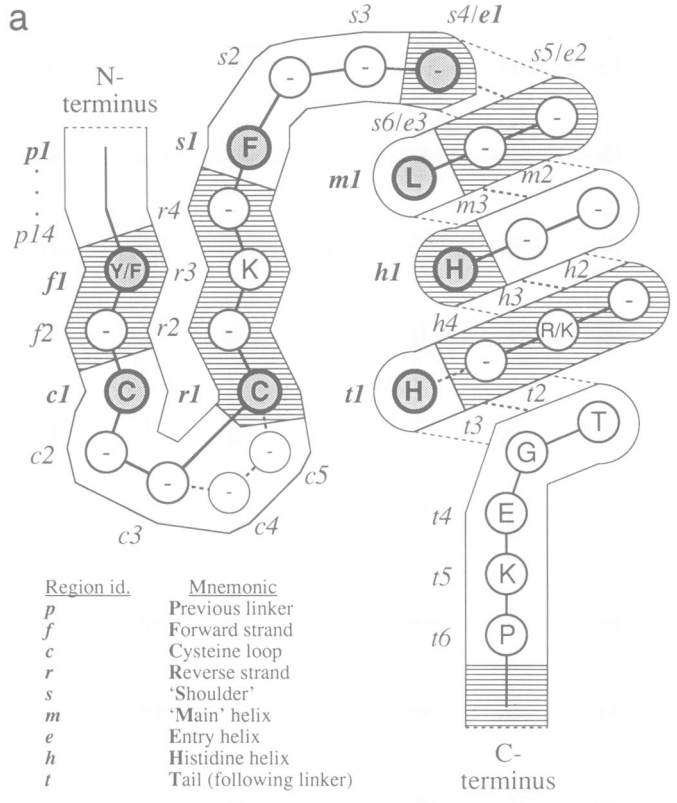
Results

Construction of the zinc finger database

Using the EMBL and GenBank databases, the literature and contributions from authors, a database of 1340 zinc fingers taken from 221 proteins was created. All entries have been checked for duplication, and duplicate entries removed (but annotated for future reference). All zinc fingers of the CC/HH type, including those that closely resemble (but do not match) the general consensus, are held. Each zinc finger was assigned a 'quality level', defined in terms of how well the finger matched the overall consensus definition (see above). The database software (see below) can be used to select all fingers above a given quality level for analysis, avoiding unusual fingers. The collection for analysis was 'frozen' in October 1991 to provide a consistent set to study. This collection contains 1340 fingers from 221 proteins. Only those fingers conforming to a consensus of C X_2 C X_{12} H X_3 H were examined in this work (i.e. fingers with a so-called '2-3 spacing').

Alignment analysis software

During the course of the project, a generalized program for the analysis of aligned sequences was written and used to



b

Position	f		c					r				s					m			h				t				
	1	2	1	2	3	4	5	1	2	3	4	1	2	3	4	5	6	1	2	3	1	2	3	4	1	2	3	4	5	6
finger 1	Y	A	C	P	V	E	S	C	D	**R** ¤	R	F	S	**R** Δ	S	D	E	L	T	**R** Δ	**H** ¤	I	R	I	H	T	G	Q	K	P
finger 2	F	Q	C	R	I			C	M	**R** ¤	N	F	S	**R** Δ	S	D	**H** Δ	L	T	T	**H** ¤	I	R ◊	T	H	T	G	E	K	P
finger 3	F	A	C	D	I			C	G	**R** ¤	K	F	A	**R** Δ	**S** ¤	D	E	R	K	**R** Δ	**H** ¤	T	K	I	H	L	R	Q	K	D
buried	•		•					•				•					•			•				•						
structure	B	-	B	-	B					B	-	B	-	B			H	-	H	-	H	-	H	-	H	-	H	-	H	-	H

Fig. 1. Definition of a notation for referring to positions in CC/HH zinc fingers. (**a**) A schematic diagram of the fold of a zinc finger is shown (to obtain an approximation to the real fold, move the strand on the left behind the strand to its right). To the left of each position (circled) is the notation used to refer to that position. The finger is divided into nine regions, eight of which begin with a conserved amino acid position (the exception is the *e* region). Conserved positions are circled in bold. Each region is assigned a single letter region identifier (*p*, *f*, ... *t*). A position is referred to by its region identifier followed by the position within the region. Position 1 of a region is the conserved position the region starts with. Each region identifier has a mnemonic value, as shown. Note that the *e* region overlaps with the *s* region. The *e1* position is the same as the *s4* position. (**b**) The sequences of the three fingers of Zif268 are shown. The Zif268 peptide used in the co-crystal structure of Paveltich and Pabo (1991) contains four amino acids before the first finger (not shown) and ends with the position *t*3 of finger 3. The positional notation (first two rows) is that described in panel a. The observed DNA contacts in the co-crystal structure are highlighted. The positions that are buried inside the core of the fold are marked in the last but one row of the table with a dot (●). The β-sheet (B) and helical amino acids are marked in the final row. Below each finger sequence are marked three classes of positions; positions known to bind the phosphate backbone of DNA (¤), positions that make direct contact(s) with the bases (Δ) and a position that makes an interaction with the following finger (◊).

generate the results shown in this paper. This software is intended for the explorative alignments of a large number of sequences, and is not restricted to zinc fingers. The program generates various outputs including a variety of consensus tables, correlation tables and cluster trees. Its main strength lies in its ability to select subsets of the database. Selection can be based on subsets of proteins, motifs within the proteins, positions within the sequences, types of amino acids held at positions within the sequences and additional numerical characteristics of the sequences (e.g. the number of amino acids between two positions, or overall hydrophobicity) or a combination of these. The user can construct a positional notation for referring to positions of the alignment (for example the one used for zinc fingers in this paper) and present it to the program so that all references to positions can be made via the user's scheme.

Referring to positions in the zinc finger motif

In order to be able to refer to positions in an alignment in a consistent manner, it is useful to devise a notation akin to that used to describe haemoglobin sequences (Perutz *et al.*, 1965). The notational scheme used here is shown in Figure 1a. This notation has the advantage that positional references can withstand the later addition of sequences with different numbers of amino acids between the key conserved positions, and that positions can be referred to independently of their amino acid number in the protein sequence from which they come. The latter greatly facilitates discussion of fingers from different proteins. This notation has been italicized in the text.

Comparing sequences

For all analyses described in the paper, a symmetrical, normalized Dayhoff similarity matrix was used. The original Dayhoff matrix was made symmetrical by averaging the forward and backward mutation rates, then scaled so that all S_{ii} scores were 100, and the lowest S_{ij} was 0 (zero) (see below for definitions of these terms).

A conservation measure was used to compare two or more sequences consistently. This measure is similar to that used in the profile sequence analysis techniques (Gribskov *et al.*, 1987) and measures the degree of 'self-similarity' of a position i.e. how well a position holds the same type of amino acids. The calculation can be described as:

$$C_{total} = \Sigma C_{ij},\ C_{ij} = F_i F_j S_{ij}$$

where C_{total} is the conservation measure for the position concerned; C_{ij} is the conservation of one pair of amino acids, i and j; F_i is the frequency of amino acid i; and S_{ij} is the amino acid similarity score relating amino acids i and j, taken from the amino acid similarity matrix, S, supplied.

This score is the frequency weighted average of comparing every amino acid found at a position against all amino acids found at that position. The S matrix need not be the Dayhoff-based matrix used here: for example, variation could be counted by scoring mismatches rather than matches (i.e. $S_{ii} = 0$, $S_{ij} = 100$).

This measure has several advantages; in particular, the similarities do not depend on arbitrary cutoffs nor on the number of sequences in the sample. Furthermore, two different sets of sequences can be directly compared with this measure, provided they are based on the same similarity matrix. Later the conservation of positions in long runs will be compared with the conservation of the same position between homologous fingers using this measure.

Fingers within a protein are more conserved than fingers from different proteins

Many zinc finger proteins are characterized by containing a region of repeated, adjacent fingers. Many of these long runs have exactly the same repeat unit length and the same number of amino acids between repeated units (i.e. the 'phase' of the repeat is perfectly conserved), suggesting that the run of fingers has arisen from internal duplication from one ancestral finger (see Introduction).

If the runs of zinc fingers did evolve by internal duplication, the similarity scores of the pairwise comparison of fingers found in the same protein would be expected (on average) to be higher than the distribution of scores from comparing fingers found in different proteins. Figure 2 shows a plot of the pairwise comparison of all fingers in the database with a '2-3' spacing (i.e. with two amino acids between the cysteine ligands and three between the histidine ligands). Fingers in the same protein clearly tend to be more similar to each other than those from different proteins: fingers in the same protein are typically 75% similar to one another, whereas fingers from different proteins have a typical similarity of ~65%. This supports the idea that internal duplication has been the main mode of evolution of runs of fingers.

Using internal duplication to locate the DNA base contact positions

Internal duplication is likely to be the main mode of evolution of long runs of fingers and this model of evolution can be used to determine which positions in the repeated unit have a functional role, e.g. which positions bind to DNA or some other factor. This assumes that all repeated units function through the same amino acid positions. More explicitly, it

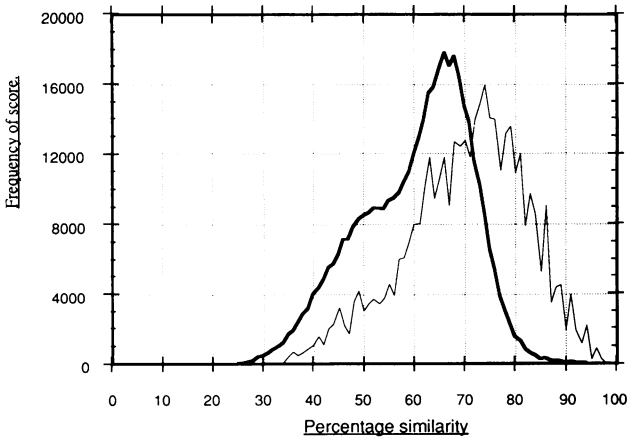


Fig. 2. Scores from comparison of zinc fingers in the same and different proteins. Two distributions are shown. The heavy line is the frequency of the conservation scores resulting from the comparison of two fingers from different proteins. The fine line is the distribution of conservation scores from comparing fingers found in the same protein. All fingers in the database with a 2-3 spacing (915 fingers) were used. The data for comparisons of fingers in the same protein were multiplied by 100 so that the two distributions can be compared. All fingers that are duplicates over the region *f1–t6* were removed from both distributions. The shoulder on the distribution of similarity levels from fingers in different proteins is due to the presence of subclasses of fingers within the 2-3 spacing class; comparisons of fingers from different subclasses yield slightly lower scores (shoulder) than comparisons of fingers in the same subclass (peak). This shoulder is absent when the same plot is performed on fingers in the long run data set, suggesting that the fingers in long runs form one related subclass of fingers.

Table I. Summary of proteins with 10 or more zinc finger motifs: decision to use in 'long runs' study.

Protein name and zinc finger database index number[j]		No. of fingers in protein	Motifs used for long run study (inclusive)	Number of linkers whose length ≠ 5[a,i]	Number of fingers with different spacings[b]	Alternation of fingers present?[c]	Number of 'odd' zinc ligands present[d]
zfpt$m	17	12	1–12				
HPF1$h	65	15	2–15				1 (1)
PF2$h[g]	66	20	2–20				1 (1,missing)
HPF4$h[e]	67	16	2–16				2 (1,11)
HPF9$h[h]	68	11	1–10	1 (11)	1 (11)		3 (1,3,10)
HPFp5$h	70	13	1–12				1 (13)
HPFp7$h	72	14	1–14	1 (before 1)			1 (5)
Kox1$h	79	11	2–11		1 (1)		2 (1,11)
HF-10$h	115	10	1–10	1 (10)			
zfp-35$m	133	18	2–16	1 (1)	1 (17)		
zfp-36$h	134	13	1–10		1 (11)	?	
mfg2$m	137	14	3–14				3 (1,2)
mfg3$m	138	10	1–10				
mkr3$m	155	15	2–14	1 (15)			2 (1,15)
mkr4$m[f]	157	14	1–3, 5–14				4 (4)
H-plk$h[h]	161	16	1–12		2 (13,15)		5 (1,2,3,12)
XlcGF46a	181	10	1–10	1 (10)			
XlcGF48b	182	12	1–12	1 (12)			
XlcGF57a	187	12	1–11	1 (12)		?	1 (12)
XlcGF58a	188	15	1–15	1 (15)			1 (10)
XlcOF-6	195	16	2–16	2 (1,16)	1 (1)		2 (2,13)
ZNF41$h	219	19	2–19		1 (1)		3 (1,3)
HTF6$h	221	23	1–23			?	4 (1,3,4,5)
ZFH-2	16	17		15	14		2 (5,12)
znf7$h	52	16		4 (1,7,14,16)	1 (2)		2 (1,2)
Evi-1$m	55	10		3 (1,6,7)	3 (1,4,9)		2 (1,7)
Evi-1$h	56	10		3 (1,6,7)	3 (1,4,9)		2 (1,7)
mkr5$m	158	12			2 (3,10)	Yes	2 (12)
SuHW$d	162	13		8	5		2 (1,5)
XFG5-2$x	167	10		1 (1)			3 (1,2)
XlcGF20a	174	18		6 (6–11)	1 (14)		5 (6,9)
XlcGF26a	175	12			1 (11)	Yes	
XlcGF66a	191	13		2 (2,13)	4 (4,8,10,12)	Yes	3 (1,5)
XlcOF-7a	197	22		1 (22)	3 (2,10,17)	?	1 (2)
XlcOF-8d	199	15		1 (9)			
XlcOF-22	204	13			3 (2,10,12)	?	2 (1)
XlcOF-28	206	13		1 (9)			
Xfin$xl	208	37		5			2 (1,37)
ZNF6	209	13		3	5	Yes	5 (3,7)
Zfa$m	210	13		3	5	Yes	1 (3)
Zfb	211	13		5	5	Yes	1 (3)
Zfy-2$md	212	13		3	5	Yes	
Zfy-2$mm	213	13		3	6	Yes	
Zfy-1$m	214	13		3	5	Yes	
Zfy-1$h	215	13		3	5	Yes	
Zfx$h	216	13		3	5	Yes	
Zfx-1$h	217	13		3	5	Yes	
Zfx$m	218	13		3	5	Yes	
MZF1$h	220	13		2 (4,13)			
Total proteins		49	23	33	28	18	31
motifs		703	305	92	99		71

[a] All linkers whose length ≠ 5 are counted. Linker length is measured as the number of amino acids between the *t1* position of the first finger and the *f1* position of the following finger, exclusive.
[b] The number of spacings that are not 2-3 are recorded.
[c] Only those proteins with obvious alternating fingers are recorded. Subtle multi-finger duplications are likely to still be present.
[d] This includes the relatively common *t1* = C and less common *r1* = Y. Furthermore, the first and last fingers of long runs are frequently 'near-miss' duplications, i.e. duplications that preserve the phase of the repeats, but do not show some of the conserved positions characteristic of zinc finger motifs.
[e] The first finger of HPF4 was published from the *c1* position of first finger, hence the different linker length reported before the first finger.
[f] For mkr4, an unusual, perfectly 'in phase', repeat was observed and has been recorded as finger number 4. This was removed from this study as indicated.
[g] In HPF2, finger 1, the *t1* zinc ligand is 'missing'.

[h] HPF9 and H-plk are related and are likely to be alternative transcripts from the same gene.
[i] If an unusual linker is noted for the last motif in a protein, it means that this linker is either incompletely sequenced, or the protein's C-terminus occurs within five amino acids of the last finger. In the cases that the last finger has been published with the sequence incomplete in the region after *t1*, these have been accepted.
[j] For reasons of space the references to the sequences have not been listed, nor any alternative names they are known by in the literature.

is required that the sequence duplications represent duplication of a structural unit, conserving the fold of the unit, in which each duplicated unit interacts with the DNA bases in the same orientation, through the same amino acid positions. In a sense the analysis also tests this idea: if different positions make contacts to the bases in different fingers in a particular run of fingers, the results for that run will be the average of the different recognition schemes, and no correlation will be observed. The presence of a strong correlation will confirm that the same single-finger−DNA interaction is being employed in all (or most) of the fingers examined. Note that the arrangement of fingers along the DNA can still differ whilst retaining the same single-finger−DNA interaction for each finger; this work does not depend on a particular arrangement of fingers on the DNA.

It seems1 reasonable to assume that all fingers take on the same fold. All the fingers examined strongly conserve all core amino acids, as well as the number of amino acids between each strongly conserved position. In addition, zinc is a strong determinant of the folding of zinc finger peptides (see for example Lee *et al.*, 1991), suggesting that the geometry around the zinc ion is conserved. This in turn will determine much of the overall fold of the finger. All NMR structures to date show the same overall fold; the main differences are in the *h* region. When this region has a different number of amino acids, the fold within this region differs.

As the fingers duplicated the base recognition positions could have either (i) diverged (e.g. the new fingers have adapted to binding different DNA sequences), or (ii) remained conserved to retain the same function (e.g. bind the same sequence). With respect to DNA binding the first model implies that the duplicated fingers go on to recognize a different base sequence from their ancestor. The second model implies that a duplicated finger retains the same specificity as its ancestor.

Now assume that one of these two processes has tended to dominate, i.e. zinc finger proteins have, in general, been adapted to bind either non-repetitive (i) or repetitive (ii) DNA sequences. Although in principle (and practice) both can take place within a single protein, the analysis will proceed on the grounds that typically one of the two has dominated. This subsequently proves to be the case. In the first situation the most variable positions within any one long run would be expected to be the base recognition positions. Conversely, the latter model would be supported if the most conserved positions proved to the base recognition positions.

To summarize, we are looking for evidence of evolution of functional positions by duplication followed by either divergence or conservation. The underlying assumptions of this work are that: (i) internal duplication is the main mode of evolution of these runs of fingers; (ii) all (or most) zinc fingers have the same fold; (iii) all (or most) zinc fingers interact with their ligands in the same manner; and (iv) DNA binding is site-specific in these runs of fingers.

The work described here was initially done less rigorously prior to the publication by Pavletich and Pabo (1991) of the zinc finger−DNA co-crystal structure, and has since been repeated as described below.

Conservation of positions in long runs

Using software written for the analysis of aligned sequences (G.Jacobs, unpublished work), a subset of sequences consisting of runs of 10 or more fingers, with each zinc finger having the same number of amino acids between the zinc ligands (i.e. the same 'spacing') and the same number of amino acids (five) between each finger ('linker length') was selected. Other 'bad' features, such as unusual amino acids in the zinc ligand positions and obvious alternation patterns (such as those in the ZFY and ZFX proteins), were avoided. Alternating patterns were avoided because higher order duplications of independent fingers cannot be distinguished from duplication of higher order structures composed of more than one finger; I wanted to select runs of fingers in which the fingers were likely to be structurally independent units. This resulted in a subset of 305 fingers from 23 proteins being selected from an original set of 703 fingers from 49 proteins that contained 10 or more fingers. The number of fingers in the runs vary from 10 to 22 fingers. These runs are referred to as 'long runs'. Table I shows a summary of the selection process and the fingers used.

Long runs are needed so that there are enough data for each individual run of fingers to be informative (i.e. so that measures of conservation and variability at each position over the run are meaningful). All the fingers used for the tests have a 2-3 spacing (i.e. $CX_2CX_{12}HX_3H$). This has the advantage that there is no need to consider different spacing groups, which may take on different folds (at least around the *c* and *r* regions).

When the scores from comparison of fingers within a long run and the comparison of fingers from different long runs were plotted, it was found that fingers within a run were more similar to one another than those in different runs (data not shown), as was previously found for the overall dataset (see Figure 2).

As an initial step towards analysing these sequences, the long run subset was summarized by taking a simple two-level ($50-<80\%$ and $\geq 80\%$) consensus of each of the chosen runs. The results are summarized in Figure 3a. Three solvent-exposed positions, *s3*, *s6* and *m3*, were found to be especially variable (see also Figure 1).

The positions that are well conserved over all the runs (*c1*, *r1*, *s1*, *m1*, *h1* and *t1*) include all the positions now known to be necessary to create the fold of the structure, and one position now known to interact with the phosphate backbone of DNA (*r3*). Note that positions *c3*, *r4*, *s4*, etc., can be distinguished in that although they tend to be conserved in most proteins, when they are conserved they are frequently conserved at the lower level of $50-<80\%$ (Figure 3a). The remaining core position, *f1*, is less well conserved due to the frequent exchange of Y and F, which appear to be essentially equivalent at this position.

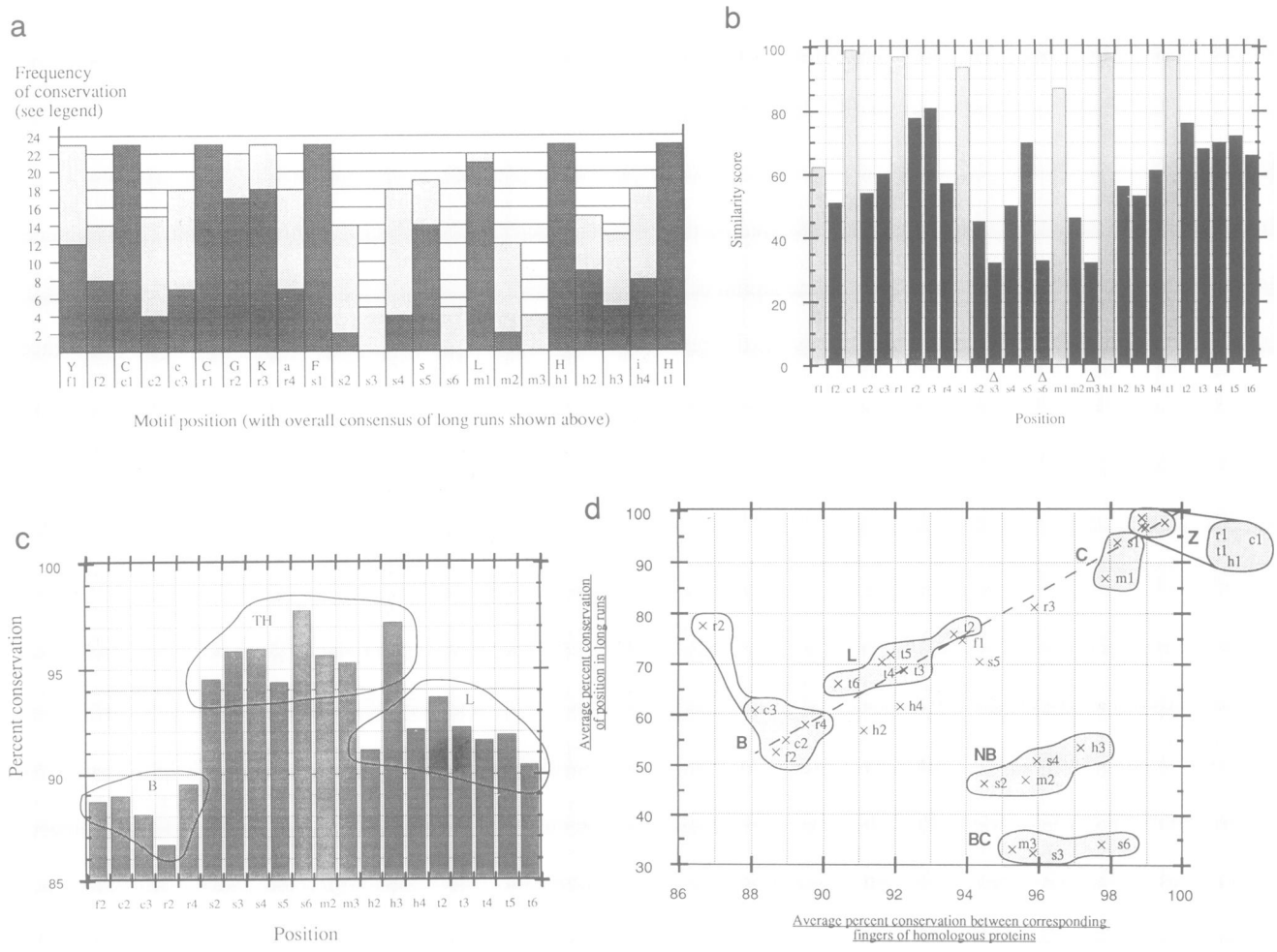


Fig. 3. (**a**) Conservation of positions in long runs of zinc fingers. Shown is a stacked bar chart of a count of the number of runs in which the most common amino acid at each position is conserved at one of two different levels. The light bars cover the range $\geq 50 - < 80\%$, the dark bars $\geq 80 - 100\%$. There are 23 runs in total. For clarity only the positions in the 'body' of the finger are shown. The positions are labelled as described in Figure 1a. Above the position labels is the consensus of all fingers in the long runs (lowercase, $\geq 50 - 80\%$; uppercase, $\geq 80\%$). For example, examining the *c3* position; 11 of the runs hold a single amino acid conserved between 50 and 80% of the time, eight runs hold a single amino acid in >80% of their fingers and five runs had no single amino acid conserved more than 50%. (**b**) Numerical conservation of positions in long runs of zinc fingers. The conservation of each position in the long runs, using the calculation described in the text, is shown for each position of the zinc finger. Note that the trend is essentially the same as in panel a. The base contact positions are highlighted with a triangle above the position notation. (**c**) Conservation of positions between homologous fingers with a 2-3 spacing. For each exposed position, the average conservation score from the comparison of homologous fingers is shown. The most variable position is *r2* and the most conserved, *s6*. Three levels are marked out ($<90\%$, $90 - 94\%$, $\geq 95\%$), which divide the finger into three distinct patches on the surface [β-sheet (B), tip and helix (TH) and linker (L)]. Note that the 'linker' group (L) includes one helical position, *h2*. 33 groups of homologous fingers were used to create these data. A similar trend is observed in the equivalent plot from 75 groups of homologous fingers from the five main spacing classes (2-3, 2-4, 2-5, 4-3 and 4-4, data not shown). (**d**) Correlation of conservation of positions within long runs and conservation of positions in homologous fingers. For each position two scores were calculated, firstly the average conservation of that position in each of the long runs (vertical axis, see panel b), and the average conservation of that position in each of the groups of homologous finger proteins (horizontal axis, see panel c). Each position is labelled to the right of where it occurs in the plot. The labels and positions of the *r1*, *t1*, *c1* and *h1* positions are shown in the enlarged region to the right of the plot. Note that the vertical and horizontal axes start from 30% and 86%, respectively, not 0%. The main trend starts from 50% conservation within long runs. No surface positions are well conserved in long runs, with the exception of *r3*. The dashed line is not a regression line—it is to illustrate the main trend. Amino acids that fall close together in space appear close together in the correlation plot. These spatially related amino acids have been marked in groups and named as follows: BC, base contact positions; NB, neighbours to base contact positions; Z, zinc ligands; C, core amino acids (except *f1*); L, linker amino acids; B, β-sheet positions

Interestingly, the *s5* position could be included as a potential well conserved position—possibly pointing to a structural role such as stabilizing the α-helix or the tip (this conservation could also be interpreted as a result of a conserved contact with DNA, but this appears unlikely from examination of the Zif268 co-crystal structure). One way in which *s5* could stabilize the α-helix would be to cap the N-terminus of the helix (see Richardson and Richardson, 1988; Serrano and Fersht, 1989; Bell *et al.*, 1992), but this would require that the structure of the tip alter so that the helix begins at the *s6* position. Serine, the most frequent amino acid found at this position (see Table III), is a favoured N-terminal capping amino acid. Desjarlais and Berg (1992a) suggest that *s5* =S could hydrogen bond to *s3*=Q, forming a buttress interaction analogous to that seen between *s5* and *s3* in the ZIF268−DNA structure.

Further plots were done using the numerical measure of conservation described above, with essentially identical outcome (Figure 3b). Several other amino acid similarity matrices were tried (identity, minimum genetic code distance

and structure–function) all with consistent results (data not shown). This indicates that the results are a property of the sequences and not the method of comparison.

Conservation of positions in corresponding fingers from homologous proteins

The conservation analysis of the long runs still leaves open the identity of the base contact positions because we cannot discriminate whether the positions have merely drifted during the course of duplication due to the lack of selective pressure (in which case the most conserved exposed positions should be the base recognition positions) or have diverged in order to recognize different base sequences (in which case the most variable positions would be the base recognition positions). This can be resolved by comparing corresponding zinc fingers from proteins whose finger regions are highly similar (e.g. >90% identical). The majority of very closely related proteins from different organisms would be expected to bind the same DNA sequence. In fact, many of the groups of potentially homologous finger regions have some members that are known to bind the same site (see Table II). The base recognition positions should fall onto the face (i.e. surface patch of the three dimensional structure) that is most conserved between homologous fingers.

Taking zinc finger proteins from different species which are very highly conserved over the whole of their zinc finger domains (Table II), I compared each position in the first finger of each group of homologues, then each position of the second finger and so on. Only fingers with a '2-3' spacing were used, to avoid different structures in the *c* and *h* regions. Figure 3c shows a bar chart of the average conservation of each surface position of the zinc finger motif between homologous fingers. Examining the conservation scores and the closeness of the positions on the structure of the zinc finger, the distribution can be broken into three groups. The β-sheet region (B) shows the lowest overall conservation. Here different amino acids have been tolerated in homologous fingers; this region would be expected to face away from the DNA. The tip and most of the α-helix (TH) show the highest conservation, suggesting that this region faces the DNA. Finally the linker (L) shows intermediate conservation. Since these proteins are very similar, only a relatively small range of conservation scores is observed. The core positions (not shown) are all highly conserved, as would be expected.

Correlating conservation within a long run against conservation between homologous fingers

We wish to locate positions that are on the surface of the fingers that are well conserved in homologous fingers, but are either well conserved or poorly conserved within long runs. No surface positions were found to be well conserved within long runs, suggesting that the main mode of evolution has been duplication followed by divergence. The most variable positions in the long runs fall into the most conserved region of fingers from homologous proteins, suggesting that these positions are the base recognition positions.

Figure 3d shows a correlation plot of the average conservation score from comparing fingers within a long run against the average conservation from comparing homologous fingers. Three positions (*s3*, *s6* and *m3*) fall markedly off the main trend of the distribution. These three positions are clearly the candidate base recognition positions, and indeed include all the conserved contacts with the bases in the Zif268–DNA structure. Four nearby positions from the tip region fall nearby in the graph, suggesting that their choice of amino acids has also been constrained during evolution. All other positions (except *r2*) fall along the main trend of the distribution.

Examining the structure of the zinc finger, positions clustered in the plot were found clustered in the structure. These clusters include the core positions [zinc ligands (Z) and the hydrophobic positions in the tip of the finger (C)], the linker region (L) and the positions on the surface of the β-sheet (B). The *r2* position, also on the surface of the β-sheet, lies away from the main trend of the graph.

Table II.

Related proteins (common name)	Are any homologues known to bind same sequence?	Proteins per group	Fingers in protein[g]
Snail		3	4–5
TFIIIA	Yes	3	9
p43	Yes	2	9
'Ig' fingers[a]	Yes	10	2–5
'bZIP' fingers[b]	Yes	2	1
Evi	Yes	2	10
Wilm's Tumour (WT)	Yes	2	4
Egr-1[c]	Yes	3	3
Egr-2		2	3
Egr-3		2	2–3
Egr-4		2	3
GLI[d]	Yes	5	2–5
hunchback (hb)		2	6
SWI 5[e]		2	3
Zfx/Zfy[f]		5	13

[a] The 'Ig fingers' are the proteins known to bind the immunoglobin enhancers. These include KBP1, MBP1, PRDII, HIV-EP1 and others.
[b] The 'bZIP' fingers are single zinc fingers found at the very N-terminus of some bZIP (basic leucine zipper) proteins. The bZIP motif is found in a large family of related transcription factors.
[c] The egr proteins have been divided into their subclasses to avoid the possibility that the different subgroups bind different sequences.
[d] The GLI group include ciD (*cubitus interruptus* dominant).
[e] SWI 5 has a single homologue, ACE2.
[f] As the relationship of the Zfx and Zfy proteins is complex, they have been pooled into one group.
[g] Where a range of fingers for the protein is shown, some of the sequences are incomplete. For reasons of space, all the proteins used have not been listed, and the sequence references have not been cited. Interested readers are welcome to contact the author for details.

Discussion

Interpretation of the main trend of the correlation plot

By examining the correlation plot in Figure 3d, several clusters of positions have been identified. Positions that are strongly conserved in both long runs and between homologous fingers are to be found in the top right-hand corner of the plot. Positions that are variable in both data sets are in the bottom left corner. Positions that have duplicated and diverged during the evolution of the long runs are to be found in the lower portion of the graph, whereas those positions which have tended to remain conserved after duplication are found in the upper region of the graph. The clusters in the correlation plot are also shown in an atomic

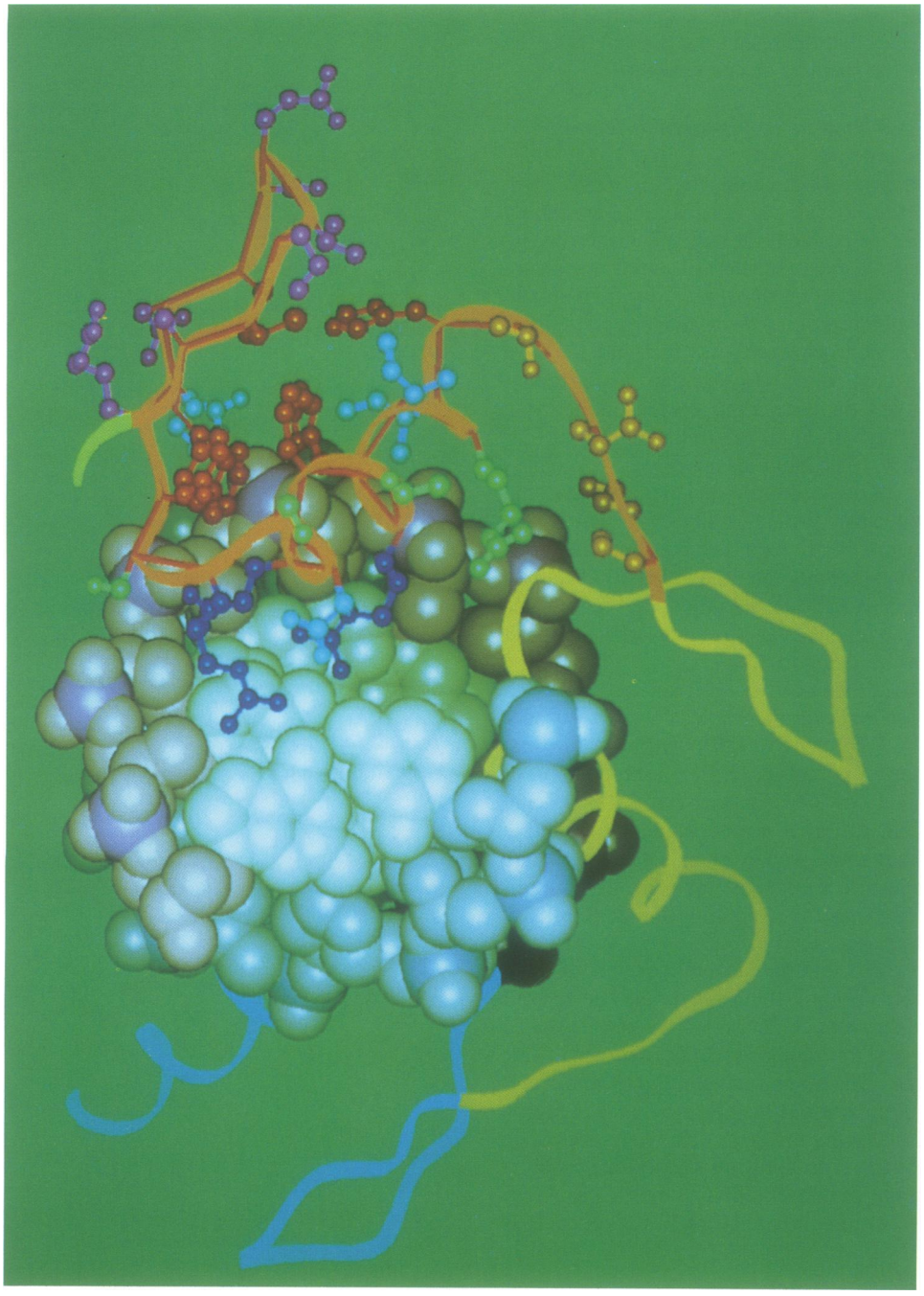


Fig. 4. Spatial relationship of the clustered positions in the zinc finger–DNA interaction. Shown is a view down the long axis of the DNA duplex of the Zif268–DNA co-crystal structure. A ribbon traces the main chain of the three fingers in the complex; the first finger is shown with a brown ribbon and the ribbon for the second finger is yellow. The first finger is detailed as a ball and stick drawing, with amino acids belonging to each of the clusters from the plot of Figure 3d shown in a different colour. Using the names for the clusters from the legend of Figure 3d, the colours for each cluster are: BC, deep blue; NB, light green; Z and C, red; L, yellow; B, purple. All non-clustered positions are shown in light blue. The α-helix is angled $\sim 45°$ away from the plane of the paper, that is the N-terminus of the helix is closer to the viewer than the C-terminus. The ribbon of the second finger is shown in order to illustrate the proximity of the *h3* position and the tip of the following finger.

model of the interaction of the first finger of the Zif268 three-finger peptide with DNA (Figure 4).

Amino acids close together on the surface of the structure were found to fall close together in the correlation plot. Positions involved in either the maintenance of the fold of the zinc finger or in interactions with invariant structural features of DNA would be expected to be strongly conserved both within long runs and between homologous fingers. All core positions and one surface position fall into this category. The four zinc ligands are the most strongly conserved, then the hydrophobic positions in the tip (*s1* and *m1*). Nearby is *r3*, a surface position on the reverse strand, strongly conserved as lysine (or occasionally arginine). Lysine is rarely conserved to help maintain the fold of a protein, due to its long flexible side chain. This position is an excellent candidate for a position that is functionally required, regardless of the DNA sequence recognized. It is characteristic of DNA binding proteins that a few such basic

Table III. Consensus amino acids at selected positions of the zinc fingers in long runs

Cluster	Position	75–<100%	50–<75%	20–<50%	10–<20%	5–<10%
NB	*s2*			S	N T	R K I
NB	*s4*			K S	R	
NB	*m2*			T	I V	N R L
NB	*h3*			R K	I	L
B	*r2*		G			
(None)	*r3*	K				
(None)	*s5*		S			T

For clarity the rarer amino acids are not shown.

positions are conserved in order to 'anchor' or orientate the DNA binding module with respect to the phosphate backbone. In the Zif268–DNA structure, this position is found contacting a phosphate in each of the three fingers.

The remaining core position, *f1*, which is strongly conserved as either F or Y (and sometimes H), is located on the plot away from the other hydrophobic core positions (*s1*, *m1*). This is likely to be due to F and Y being essentially equivalent at this position. In the long runs, while one of F or Y predominates in any one run, the other is almost always present. Likewise, this position is frequently substituted in homologous fingers. The amino acid similarity in the Dayhoff-based matrix used, however, scores these as being more distant than essentially equivalent amino acids.

The linker positions all fall in the middle of the main trend of the graph. They do not appear to be as 'important' as the core positions either within long runs or between homologous fingers. Given that many of the sequences with long runs were selected from cDNA libraries using a probe corresponding to a linker sequence of HTGEKPYXC (where X stands for any amino acid), it is perhaps surprising that the linker region is not more highly conserved. The linker positions are likely to be a little more variable than expected, since the 'linker' regions after the last fingers in a run have been included in the analysis (strictly speaking, linkers are regions *between* fingers). This also explains the trend of the linker positions to become more variable as you go along the linker (from *t1* to *t6*); the last 'linker' might be expected to be more variable further away from the zinc finger region. The few linker regions involved are unlikely to alter the results.

All exposed positions on the β-sheet are clustered together in Figure 3d, with the exception of *r2*, which is strongly conserved within long runs but is the most variable position in homologous fingers. Glycine is the most common amino acid at this position in long runs. In the Zif268 structure the ϕ angle of this amino acid in fingers 2 and 3 (but not finger 1) is positive, an angle only easily adopted by glycine. Homologous fingers show a wide variety of amino acids at this position, incompatible with holding a positive ϕ angle, suggesting that some homologous fingers have a different turn structure which allows a negative ϕ angle at *r2* which can still orientate the cysteine residues for zinc ligation. This would account for the high variability of *r2* in homologous fingers and its position in the top left corner of the plot. In finger 1 of the Zif268–DNA structure (which has four amino acids in the *c* region), *r2*=D 'buttresses' *r3*=R with a hydrogen bond to *r3*'s Nη1 atom. Position *r3* in turn contacts the phosphate backbone of the DNA molecule via its Hϵ nitrogen. As *r2* is on the opposite side of the β-strand to *r3*, the *r2* side chain bends completely back on itself to achieve this interaction. The other fingers show no interactions at this position.

Interpretation of the potential DNA binding positions

The analysis presented here shows that three positions (*s3*, *s6* and *m3*) have diverged during the duplication of the long runs, but are strongly conserved in homologous fingers. These positions are found in the tip and N-terminal region of the helix and are clearly candidates for the base recognition positions. These three positions are now known to be the only three positions used to make contact with the bases in the Zif268–DNA structure (Pavletich and Pabo, 1991).

Four positions from the tip region were found to fall close together in the graph (*s2*, *s4*, *m2* and *h3*). A variety of reasons can be proposed to explain this finding. Their substitution may be restricted due to being on the surface of the motif that interacts with the DNA, even though they may not make direct contact. They might make stabilizing interactions with a base contact amino acid (e.g. the *s3*=R–*s5*=D interaction observed in the Zif268–DNA complex). Alternatively, they could contact the bases in some, but not all, proteins and therefore be partially conserved. Interactions with bound or trapped waters might be made. The exact nature of these positions and the interactions they make may be dependent on the base sequence being recognized (which will affect the detailed orientation of the finger with respect to DNA).

The *s2* position favours S or T in fingers in the long runs (see Table III). At this position in the Zif268–DNA structure, fingers 1 and 2 are found to be making contacts with bound waters. In particular the *s2* position of finger 1 might make water-mediated interactions with two DNA phosphate backbone groups.

The *m2* and *s4* positions' association with the DNA contact positions is interesting in that their side chains point in the opposite direction to the base contact positions in the zinc finger structure. In all three fingers of Zif268, *s4* is conserved as serine, which is a common amino acid at this position in zinc fingers. It is interesting that three quite different interactions are made in the three fingers of the complex. In finger 1, a possible contact with a thymidine base is made as well as with some bound waters (including an indirect interaction with the DNA backbone). In finger 2, an indirect, i.e. via water, contact to a guanine base is seen with *s4*, and in finger 3 a direct interaction with a phosphate is seen. If the *m2* position (lysine) of finger 1 was reoriented it could make contact to the DNA's phosphate backbone, but no actual interaction is seen in the co-crystal

structure. The *m2* position in finger 2 contacts bound waters and no interactions are observed in finger 3.

The *h3* position is not involved in DNA recognition in the Zif268–DNA structure, but is 'in line' with the other DNA binding positions in the zinc finger structure (it is present on the same face of the α-helix as the other base contact positions). It is the last position on the α-helix that could be in line with the DNA binding positions, given that the helix ends at the *t2* position. However, in the co-crystal structure it appears to be too far away to contact the bases directly. The *h3* position of finger 2 (arginine in the Zif268–DNA structure) makes a side chain–carboxyl interaction with the next finger, as previously reported. No equivalent interaction could be observed in finger 1, although if the *h3* side chain was reoriented it could make such an interaction. A conserved feature such as this finger–finger contact would be expected to result in the position being conserved over the long runs, but this was not found to be the case. Two reasons might account for the *h3* position being variable in long runs. Firstly, if the finger was not followed by another finger in the same groove in the protein–DNA interaction (either because it was the last finger in the run or because the next finger did not interact with the next three bases of the major groove), this position might be allowed to vary. Alternatively, several very different amino acids could be acceptable at this position. Three classes of amino acids: are held at this position in the long runs (see Table III): R or K, hydrophobic amino acids, and occasionally S or T. In the co-crystal structure this position is below the linker (see Figure 4). A hydrophobic amino acid (or the hydrophobic portion of the K or R side chains) could fill the space below the linker. Serine or threonine may be small enough to share this space with some water molecules, with which they could interact.

Although a precise role for each of these four positions cannot be assigned as yet, a possible unifying theme (with the exception of *h3*) might be that they make contact with bound water molecules and occasionally directly to the DNA itself. If the *s3*, *s6* and *m3* positions comprise a simple recognition system, the *s2*, *s4* and *m2* positions are much more subtle, and may be 'context dependent': their actual interactions may depend on the base sequences contacted by the finger and the presence of other amino acids in the protein sequence. In the Zif268–DNA structure, fingers 1 and 3 make different interactions, despite binding the same base sequence with the same amino acids, suggesting that a possible subtle deformation from the natural complex is present. The crystal structure of other zinc finger–DNA complexes will prove interesting to illustrate the role of these positions.

This analysis shows that *s3*, *s6* and *m3* can be identified as the most probable base recognition positions even in the absence of any structure for the zinc finger (for simplicity, the above discussion assumes a model of the finger in the absence of DNA). Since all the positions that are conserved in both tests are either (i) already known to be the zinc ligands, (ii) hydrophobic or (iii) strongly conserved as basic; these would respectively be core amino acids (ii) and a position that recognizes a structural feature of the DNA (iii). This result would suggest that the DNA binding positions have varied to adapt to different binding subsites. The base recognition positions would then be expected to be variable within a long run, but conserved between corresponding fingers of homologous proteins. Only these three positions are strongly conserved within homologous pairs of fingers and divergent within long runs. These positions would be expected to be on the surface of the structure since they are rarely hydrophobic, and all conserved (hydrophobic and zinc ligand) positions are already accounted for.

Implications for understanding zinc finger–DNA interactions

The strong correlation observed in the plot of Figure 3d indicates that the same single-finger–DNA interaction is employed by all (or at least the majority) of the fingers examined. This in turn suggests that most, if not all, zinc fingers with a 2-3 spacing will make the same overall single-finger–DNA interaction as observed in the Zif268 structure (Pavletich and Pabo, 1991). This is likely to apply to (at least some) zinc fingers with different spacings, as the spacing of the region interacting with DNA is conserved and this region's structure is conserved in NMR and crystal structures of fingers with different numbers of amino acids in the *c* and *h* regions. That is, the spacing of the *c* and *h* regions does not appear to affect the structure of the tip of the finger. Using these observations unknown zinc-finger–DNA interactions can be modelled using the general orientation observed in the Zif268 structure.

The *s3*, *s6* and *m3* positions appear to be the base recognition positions in many zinc fingers, suggesting that the idea of a 'code for recognition' based on primarily these positions, and perhaps occasionally some of the nearby positions, may be a realistic notion. The single amino acid–single base interactions observed in the Zif268 structure hint that other one-to-one relationships may be found for the *s3*, *s6* and *m3* positions and their target bases. On the other hand, the amino acids held at the *s2*, *s4*, *m2* (and possibly *h3*) positions and their interactions (with both DNA and other parts of the protein) may depend on the bases being contacted and the side chains around them, as has already been seen in some early work (Desjarlais and Berg, 1992a,b; Pavletich and Pabo, 1991). So although it might be possible to derive a relatively simple 'code' for the *s3*, *s6* and *m3* positions, the complete code may have to work in a background of more complex interactions at the *s2*, *s4*, *m2* and *h3* positions.

Acknowledgements

The author would especially like to thank Andrew McLachlan for numerous discussions and many valuable ideas on this project. Aaron Klug is thanked for discussions, as are the other members of the LMB who work on zinc fingers. Jade Li is thanked for her constructive comments on an early version of this manuscript. Tom Pieler is thanked for the contribution of his sequences which were the original foundation of the database (Nietfeld *et al.*, 1989). Nikola Pavletich is thanked for making available the co-ordinates of the Zif268–DNA complex structure. The manuscript was improved by the thoughtful and positive criticism of one of the referees. The author was supported by a New Zealand Medical Research Council (now Health Research Council) Post-Graduate Scholarship and an ORS award during early period of this work and more recently by the UK's MRC.

References

Bell,J.A., Becktel,W.J., Sauer,U., Baase,W.A. and Matthews,B.W. (1992) *Biochemistry*, **31**, 3590–3596.

Bellefroid,E.J., Lecocq,P.J., Benhida,A., Poncelet,D.A., Belayew,A. and Martial,J.A. (1989) *DNA*, **8**, 377–387.

Berg,J.M. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 99–102.

Böhm,S. and Drescher,B. (1985) *Stud. Biophys.*, **107**, 237–247.
Brown,R.S., Sander,C. and Argos,P. (1985) *FEBS Lett.*, **186**, 271–274.
Churchill,M.E.A., Tullius,T.D. and Klug,A. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 5528–5532.
Crossley,P.H. and Little,P.F.R. (1991) *Proc. Natl. Acad. Sci. USA*, **88**, 7923–7927
Desjarlais,J.R. and Berg,J.M. (1992a) *Proteins*, **12**, 101–104.
Desjarlais,J.R. and Berg,J.M. (1992b) *Proteins*, **13**, 272.
El-Baradi,T. and Pieler,T. (1991) *Mech. Dev.*, **35**, 155–169.
Fairall,L., Rhodes,D. and Klug,A. (1986) *J. Mol. Biol.*, **192**, 577–591.
Gibson,T.J., Postma,J.P.M., Brown,R.S. and Argos,P. (1988) *Protein Engng*, **2**, 209–218.
Ginsberg,A.M., King,B.O. and Roeder,R.G. (1984) *Cell*, **39**, 479–489.
Gribskov,M., McLachlan,A.D. and Eisenberg,D. (1987) *Proc. Natl. Acad. Sci. USA*, **84**, 4355–4358.
Jacobs,G. and Michaels,G. (1990) *New Biol.*, **2**, 583–584
Kaptein,R. (1991) *Curr. Opinon Struct. Biol.*, **1**, 63–70.
Kaptein,R. (1992) *Curr. Opinon Struct. Biol.*, **2**, 109–115.
Klevit,R.E., Herriott,J.R. and Horvath,S.J. (1990) *Proteins*, **7**, 215–226.
Kochoyan,M., Havel,T.F., Nguyen,D.T., Dahl,C.E., Keutmann,H.T. and Weiss,M.A. (1991) *Biochemistry*, **30**, 3371–3386
Lee,M.S., Gippert,G.P., Soman,K.V., Case,D.A. and Wright,P.E. (1989) *Science*, **245**, 635–637.
Lee,M.S., Gottesfeld,J.M. and Wright,P.E. (1991) *FEBS Lett.*, **279**(2), 289–294.
Miller,J., McLachlan,A.D. and Klug,A. (1985) *EMBO J.*, **4**, 1609–1614.
Nardelli,J., Gibson,T.J., Vesque,C. and Charnay,P. (1991) *Nature*, **349**, 175–178.
Neuhaus,D. and Rhodes,D. (1991) *Curr. Biol.*, **1**, 268–270.
Neuhaus,D., Nakaseko,Y., Nagai,K. and Klug,A. (1990) *FEBS Lett.*, **262**, 179–184.
Neuhaus,D., Nakaseko,Y., Schwabe,J.W.R. and Klug,A. (1992) *J. Mol. Biol.*, in press.
Nietfeld,W., El-Baradi,T., Mentzel,H., Pieler,T., Köster,M., Pöting,A. and Knöchel,W. (1989) *J. Mol. Biol.*, **208**, 639–659.
Omichinski,J.G., Clore,G.M., Apella,E., Sakaguchi,K. and Gronenborn,A.M. (1990) *Biochemistry*, **29**, 9324–9334.
Omichinski,J.G., Clore,G.M., Robien,M., Sakaguchi,K., Apella,E. and Gronenborn,A.M. (1992) *Biochemistry*, **31**, 3907–3917.
Párraga,G., Horvath,S.J., Eisen,A., Taylor,W.E., Hood,L., Young,E.T. and Klevit,R.E. (1988) *Science*, **241**, 1489–1492.
Pavletich,N.P. and Pabo,C.O. (1991) *Science*, **252**, 809–817.
Perutz,M.F., Kendrew,J.C. and Watson,H.C. (1965) *J. Mol. Biol.*, **13**, 669–678.
Rhodes,D. and Klug,A. (1986) *Cell*, **46**, 123–132.
Rhodes,D. and Klug,A. (1988) Eckstein,F. and Lilley,D.M.J. (eds), *Nucleic Acids and Molecular Biology*. Springer-Verlag, Berlin, Heidelberg, vol 2, pp. 149–166.
Richardson,J.S. and Richardson,D.C. (1988) *Science*, **240**, 1648–1652.
Serrano,L. and Fersht,A.R. (1989) *Nature*, **342**, 296–299.
Tso,J.Y., Van Den Berg,D.J. and Korn,L.J. (1986) *Nucleic Acids Res.*, **14**, 2187–2200.
Vrana,K.E., Churchill,M.E.A., Tullius,T.D. and Brown,D.D. (1988) *Mol. Cell. Biol.*, **8**, 1684–1696.

Received on June 10, 1992; revised on August 20, 1992