Article No. jmbi.1998.2421 available online at http://www.idealibrary.com on IDEAL J. Mol. Biol. (1999) **285**, 1917–1934

JMB

Analysis of Zinc Fingers Optimized *via* Phage Display: Evaluating the Utility of a Recognition Code

Scot A. Wolfe, Harvey A. Greisman, Elizabeth I. Ramm and Carl O. Pabo*

Howard Hughes Medical Institute and Department of Biology, Massachusetts Institute of Technology Cambridge, MA 02139, USA

*Corresponding author

Cys_2His_2 zinc finger proteins are composed of modular DNA-binding domains and provide an excellent framework for the design and selection of proteins with novel site specificity. Crystal structures of zinc finger-DNA complexes have shown that many Cys_2His_2 zinc fingers use a conserved docking arrangement that juxtaposes residues at key positions in the "recognition helix" with corresponding base positions in the three to four base-pair subsite. Several groups have proposed that specificity can be explained with a zinc finger-DNA recognition code that correlates specific amino acids at these key positions in the α-helix with specific bases in each position of the corresponding subsite. Here, we explore the utility of such a code through detailed studies of zinc finger variants selected *via* phage display. These proteins provide interesting systems for detailed analysis since they have affinities and specificities for their sites similar to those of naturally occurring DNA-binding proteins. Comparisons are facilitated by the fact that only key DNA-binding residues are varied in each finger while leaving all other regions of the structure unchanged. We study these proteins in detail by (1) selecting their optimal binding sites and comparing these binding sites with sites that might have been predicted from a code; (2) by examining the "evolutionary history" of these proteins during the phage display protocol to look for evidence of context-dependent effects; and (3) by reselecting finger 1 in the presence of the optimized finger 2/finger 3 domains to obtain further data on finger modularity. Our data for optimized fingers and binding sites demonstrate a clear correlation with contacts that would be predicted from a code. However, there are enough examples of context-dependent effects (not explained by any existing code) that selection is the most reliable method for maximizing the affinity and specificity of new zinc finger proteins.

Keywords: phage display; zinc finger; selection; recognition code; DNA-binding domain

Introduction

The Cys_2His_2 zinc fingers constitute one of the most important and versatile families of eukaryotic DNA-binding domains. They occur in numerous transcription factors, and different fingers recognize a diverse set of DNA sites (Pabo & Sauer, 1992; Berg & Shi, 1996). These fingers provide an important model system for studying principles of protein-DNA recognition and offer a useful framework for the selection and design of novel DNA-binding proteins. Structural studies show that each zinc finger module, which contains about 30 amino acid residues, folds to form a compact ββα structure. Four conserved residues in each finger, two cysteine and two histidine residues, are ligands for a central zinc ion that stabilizes this small globular domain. Structural studies of zinc finger-DNA complexes reveal a generally conserved DNA-docking arrangement with the α-helix fitting into the major groove (Pavletich & Pabo, 1991, 1993; Fairall *et al.*, 1993; Elrod-Erickson *et al.*, 1996; Houbaviy *et al.*, 1996; Kim & Berg, 1996; Wuttke

Abbreviations used: GST, glutathionine-*S*-transferase; Ac-BSA, acetylated bovine serum albumin.

0022-2836/99/051917–18 $30.00/0

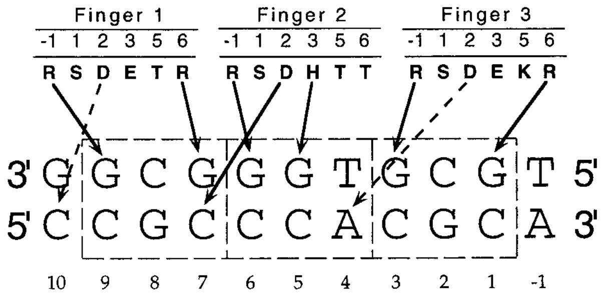


Figure 1. Diagram of the side-chain-base contacts found in the Zif268-DNA complex (Pavletich & Pabo, 1991; Elrod-Erickson *et al.*, 1996). The amino acids present at positions −1, 1, 2, 3, 5 and 6 in the α-helix of each finger are indicated using the single letter notation for amino acid type. Arrows represent observed side-chain-base interactions, and the primary binding site for each finger is indicated with broken boxes. Base contacts with the primary strand of the DNA involve positions −1, 3 and 6 of the helix. Each finger also uses the amino acid of position 2 to contact a flanking base in the secondary strand of the DNA, although the hydrogen-bonding geometry is not ideal for those involving fingers 1 and 3 (indicated with broken lines). Contacts on the primary strand involve a three base-pair subsite, but if we include the contacts from residue 2, each finger recognizes an overlapping four base-pair subsite. The C-terminal finger binds near the 5′ end of the primary strand, and thus the fingers (proceeding in the conventional N → C order) bind "antiparallel" with respect to the conventional (5′ → 3′) direction of the primary DNA strand.

et al., 1997; Nolte *et al.*, 1998). Residues −1, 2, 3 and 6 (numbering with respect to the start of the α-helix) typically make key base contacts that are responsible for defining sequence specificity. Transcription factors using this motif typically contain tandem arrays of these Cys_2His_2 zinc fingers, and often have a set of two, three or four fingers that bind to neighboring subsites on the DNA.

Ever since the Zif268 complex was determined, there has been much discussion about the prospects for a zinc finger-DNA recognition code. Zif268 contains three fingers (Pavletich & Pabo, 1991) and the docking arrangement for each of these fingers is remarkably well conserved. Moreover, residues at specific positions along the α-helix tend to contact a particular position within the corresponding DNA subsite (Figure 1). Related patterns have been observed in other zinc finger-DNA complexes (Pavletich & Pabo, 1993; Fairall *et al.*, 1993; Elrod-Erickson *et al.*, 1996; Houbaviy *et al.*, 1996; Kim & Berg, 1996; Wuttke *et al.*, 1997; Nolte *et al.*, 1998), and this has led to proposals that a "recognition code" may define the specificity of all zinc fingers that use this docking arrangement (Desjarlais & Berg, 1992a,b; Jacobs, 1992; Choo & Klug, 1994b, 1997). Phage display experiments have provided additional data about the preferred side-chain-base interactions of key residues in the α-helix (Rebar & Pabo, 1994; Choo & Klug, 1994b; Jamieson *et al.*, 1994, 1996; Wu *et al.*, 1995), and the proposed "code" has been used, with some success, to design proteins that will recognize desired target sites on double-stranded DNA (Desjarlais & Berg, 1992a, 1994; Jacobs, 1992; Choo *et al.*, 1994; Kim & Berg, 1995; Corbi *et al.*, 1997). At this stage it appears that many contacts can be rationalized in terms of a "recognition code", but there still are important questions about the generality of any such code. In particular, more information is needed on: (1) how well the designed (or selected) zinc fingers can discriminate against closely related DNA sites; (2) whether there is a single unique amino acid sequence that is best for each target site; (3) whether the code really is less reliable (as it now appears) for defining pyrimidine contacts than for purine contacts; and (4) whether the preferred DNA contacts at one position can be influenced by contacts from neighboring residues along the α-helix or by contacts from neighboring fingers.

Our laboratory recently reported a phage display strategy for selecting tandem fingers targeted to novel DNA sites (Greisman & Pabo, 1997). This "sequential" selection strategy optimizes one finger at a time, walking across the target site to generate the new protein. This procedure not only optimizes the contacts for each finger, but also can naturally adjust to incorporate potential context-dependent effects (Figure 2). High concentrations of non-specific DNA were used as a competitor to help ensure that the new proteins would bind with a high degree of specificity and affinity to their target sites. This strategy was used to select zinc finger proteins that would bind to sites normally recognized by p53, by a nuclear steroid receptor and by the TATA-binding protein (Figure 2(e)); we refer to the corresponding optimized zinc finger proteins as $p53_{ZF}$, NRE_{ZF}, and $TATA_{ZF}$. These selected proteins bind their sites with affinities and specificities comparable with that of wild-type Zif268.

Proteins obtained with this sequential selection protocol provide a rich and unbiased source of information about the relevance of a recognition code. Here, we explore this idea from several perspectives. We begin by finding optimal DNA-binding sites for the previously selected set of zinc finger proteins ($p53_{ZF}$, NRE_{ZF} and $TATA_{ZF}$), and we compare these optimal sites with the target sites used in the sequential selection protocol as well as the binding sites that might be predicted from a "recognition code". Further insight is provided by comparing the optimized zinc fingers with other zinc finger proteins. Finally, information on the potential role of context dependence in zinc finger recognition is obtained by (1) examining phage pools present at intermediate stages of the original finger selections (comparing consensus sequences at stages A and B of Figure 2 with consensus sequences at stage C); and by (2) reoptimizing finger 1 within the context of the final optimized versions of fingers 2 and 3 (as illustrated in Figure 2(d)). Each of the experiments provide data that help us evaluate the strengths and limitations of the proposed recognition code and help us compare the effectiveness of this code-based design with a strategy based entirely on phage display methods.

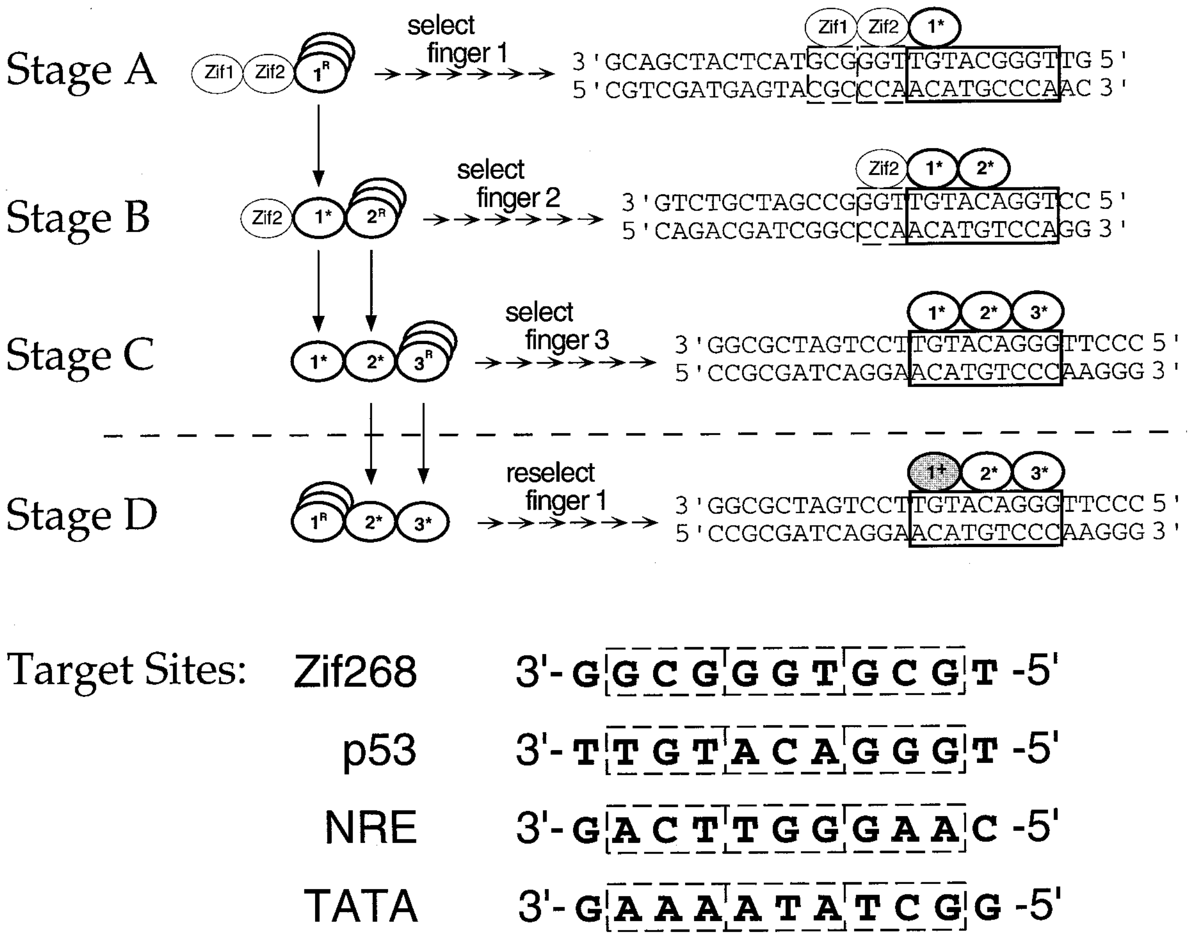

Stage A
Zif1
Zif2
1R
select finger 1
Zif1
Zif2
1*
3 ' GCAGCTACTCATGCGGGTTGTACGGGTTG 5 '
5 ' CGTCGATGAGTACGCCCAACATGCCCAAC 3 '
Stage B
Zif2
1*
2R
select finger 2
Zif2
1*
2*
3 ' GTCTGCTAGCCGGGTTGTACAGGTCC 5 '
5 ' CAGACGATCGGCCCAACATGTCCAGG 3 '
Stage C
1*
2*
3R
select finger 3
1*
2*
3*
3 ' GGCGCTAGTCCTTGTACAGGGTTCCC 5 '
5 ' CCGCGATCAGGAACATGTCCCAAGGG 3 '
Stage D
1R
2*
3*
reselect finger 1
1+
2*
3*
3 ' GGCGCTAGTCCTTGTACAGGGTTCCC 5 '
5 ' CCGCGATCAGGAACATGTCCCAAGGG 3 '
Target Sites:
Zif268
3'- G GCG GGT GCG T -5'
p53
3'- T TGT ACA GGG T -5'
NRE
3'- G ACT TGG GAA C -5'
TATA
3'- G AAA ATA TCG G -5'

Figure 2. Overview of the sequential selection protocol (stages A, B and C) that successively optimizes fingers 1, 2, and 3 to create a new zinc finger protein (Greisman & Pabo, 1997). Fingers that were present in the phage libraries at each stage of these experiments are indicated on the left-hand side of each panel. Zif1 and Zif2 denote wild-type Zif268 fingers; the superscript R denotes a randomized finger library; and an asterisk denotes the pool of selected sequences. Small horizontal arrows denote the multiple cycles of selection and amplification used when optimizing each finger by phage display. The right-hand side of each panel shows the oligonucleotides used in selections with the p53 target site, and indicates the expected position of the optimized fingers when bound to this site. Vertical arrows indicate that the pool of fingers selected in one stage is incorporated into the phage libraries used in preparing for the next stage of selection. This allowed continued optimization of previous pools of fingers in their new context during the next stage of selection. In stage A, a randomized finger 1 library had been cloned into the pZif12 phagemid display vector, and selections with this library had been performed in parallel at the TATA, p53, and NRE target sites (Greisman & Pabo, 1997). In stage B, the wild-type Zif1 finger had been removed, and a randomized finger 2 cassette ligated to the appropriate vector pools from the previous stage. Fingers were again optimized by phage display. In stage C, the remaining wild-type Zif finger had been removed, a randomized finger 3 cassette was added to the vector pools at each site, and fingers were optimized by phage display. Stage D: in a separate set of experiments involving the p53 and NRE sites, the efficacy of sequential selection was tested by reselecting finger 1 in its final context (i.e. as an N-terminal finger working in conjunction with the final optimized versions of fingers 2 and 3). In this experiment, fingers 2 and 3 represent a single clone from the earlier studies and thus do not undergo any further selection. If sequential selection (involving stages A, B and C) is successful in selecting fingers that are specific in their final context, then we would expect the fingers reselected in stage D (denoted by 1+) to be similar in sequence and affinity to those selected previously (denoted by 1* in stage C). The lower panel shows the DNA-binding sites for Zif268 and the corresponding target sites for the three proteins obtained by sequential selection. Each three base-pair subsite of Zif268, and the anticipated subsites in each target DNA, are indicated by broken boxes. For simplicity, only the primary strand of each duplex site (corresponding with the upper strand in Figure 1) is shown.

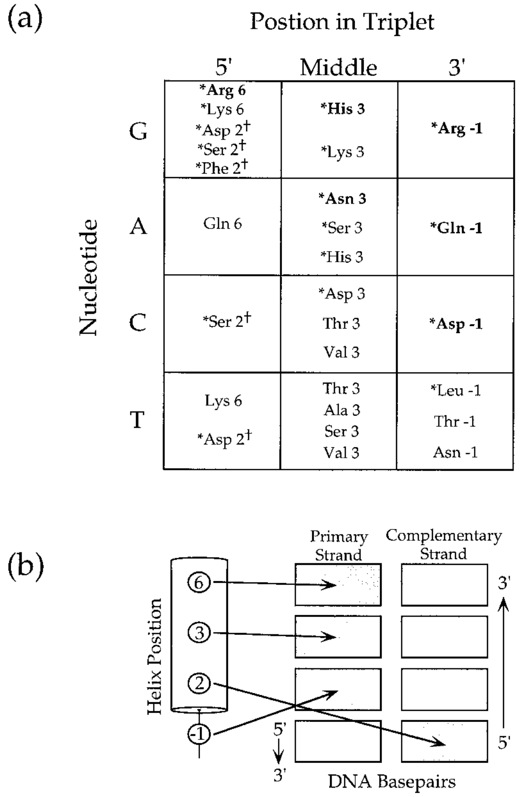
Results and Discussion

Comparing predicted and observed DNA-binding specificities for the NRE_{ZF}, $TATA_{ZF}$, and $p53_{ZF}$ proteins

DNA site selections were performed with the NRE_{ZF}, $TATA_{ZF}$, and $p53_{ZF}$ proteins to find their optimal binding sites and to analyze specificity on a base-by-base level. Since these binding site selection experiments use fully randomized oligonucleotides, they provide an unbiased method for determining the true sequence specificity for each of our selected zinc finger proteins (Greisman & Pabo, 1997). The utility of a recognition code can then be evaluated by examining how well these optimized binding sites match the sites that would be predicted by a code (based, as in Figure 3, on amino acid types present at the DNA recognition positions of each finger). As a control and benchmark for comparison, binding site selections also were performed with Zif268. The results for Zif268 (Figure 4(a)) are consistent with the contacts observed in the crystal structure of this zinc finger-DNA complex (Pavletich & Pabo, 1991; Elrod-Erickson *et al.*, 1996), as well as with earlier data regarding its sequence specificity (Swirnoff & Milbrandt, 1995). Zif268 displays a clear sequence preference at all positions within the core nine base-pair binding site, although the selection is somewhat weaker for the T under finger 2 and for the two bases the at 3′ end of the site (GCGTGGGCG). As observed in the study of specificity by Swirnoff & Milbrandt (1995), Zif268 also displays a modest preference at positions just outside of the canonical nine base-pair binding site and has a consensus sequence of TGCGTGGGCGG.

The NRE_{ZF} protein is interesting because of a striking homology (involving the Tramtrack zinc fingers) that appeared during the course of selection (Greisman & Pabo, 1997). As it happened, fingers 2 and 3 of NRE_{ZF} were selected against DNA subsites that were identical at five of six positions to the Tramtrack site (Fairall *et al.*, 1993). Remarkably, each of the Tramtrack residues that contact these five bases subsequently appeared in the NRE_{ZF} proteins that were generated by sequential selection (Figure 5(b)). A high degree of homology is also observed between the recognition helices of finger 1 of NRE_{ZF} and finger 4 of Gfi-1 (Zweidler-Mckay *et al.*, 1996) and between finger 2 of NRE_{ZF} and finger 2 of YY1 (Houbaviy *et al.*, 1996). As shown in Table 1 (page 1927), the corresponding binding sites also are quite similar (Hyde-DeRuyscher *et al.*, 1995).

DNA site selections with the NRE_{ZF} protein (Figure 4(b)) give a consensus sequence of (3′-NCTNGGGAA-5′) for the preferred binding site. This agrees well with the target site used in the sequential selections (3′-ACTTGGGAA-5′), matching at seven of nine positions. It is also informative to use the proposed recognition code (Figure 3) to



Nucleotide	5'	Middle	3'
G	***Arg 6** *Lys 6 *Asp 2† *Ser 2† *Phe 2†	***His 3** *Lys 3	***Arg -1**
A	Gln 6	***Asn 3** *Ser 3 *His 3	***Gln -1**
C	*Ser 2†	*Asp 3 Thr 3 Val 3	***Asp -1**
T	Lys 6 *Asp 2†	Thr 3 Ala 3 Ser 3 Val 3	*Leu -1 Thr -1 Asn -1

Figure 3. Tentative recognition code, similar to those proposed by Desjarlais & Berg (1992a, 1993) and by Choo & Klug (1994a, 1997) that defines the anticipated specificity of a zinc finger based on the amino acids found at positions −1, 2, 3 and 6 of the recognition helix. (a) Chart summarizing correlations that exist between amino acids located at these positions and the DNA bases that they specify. This code was compiled from contacts observed in crystal structures in which the docking of the fingers was similar to that observed in Zif268. (This docking involves the overall pattern indicated in (b); contacts observed in crystal structures are indicated by asterisks.) Expected contacts (assuming that each finger binds DNA in a canonical manner) are also taken from the sequences of natural and selected fingers in cases where the binding specificity has been defined and more than one example of the contact exists. Amino acids that are most frequently used to define a given base are indicated in bold; † marks contacts involving position 2 from a neighboring C-terminal finger. Note that less than half the positions in the chart involve a one-to-one correspondence between amino acids and bases. There often is ambiguity in that several amino acids may provide alternative ways of recognizing a particular nucleotide and it is not always clear which one is best to use. There are also cases where a given amino acid (such as serine at position 3) can specify more than one DNA base at a given position, and this introduces another level of ambiguity into the recognition process. (b) Cartoon indicating which amino acid positions of the recognition helix and which base positions in the DNA subsite are juxtaposed when fingers dock in the manner observed for Zif268. This overall docking arrangement underlies the recognition code that is summarized in (a).

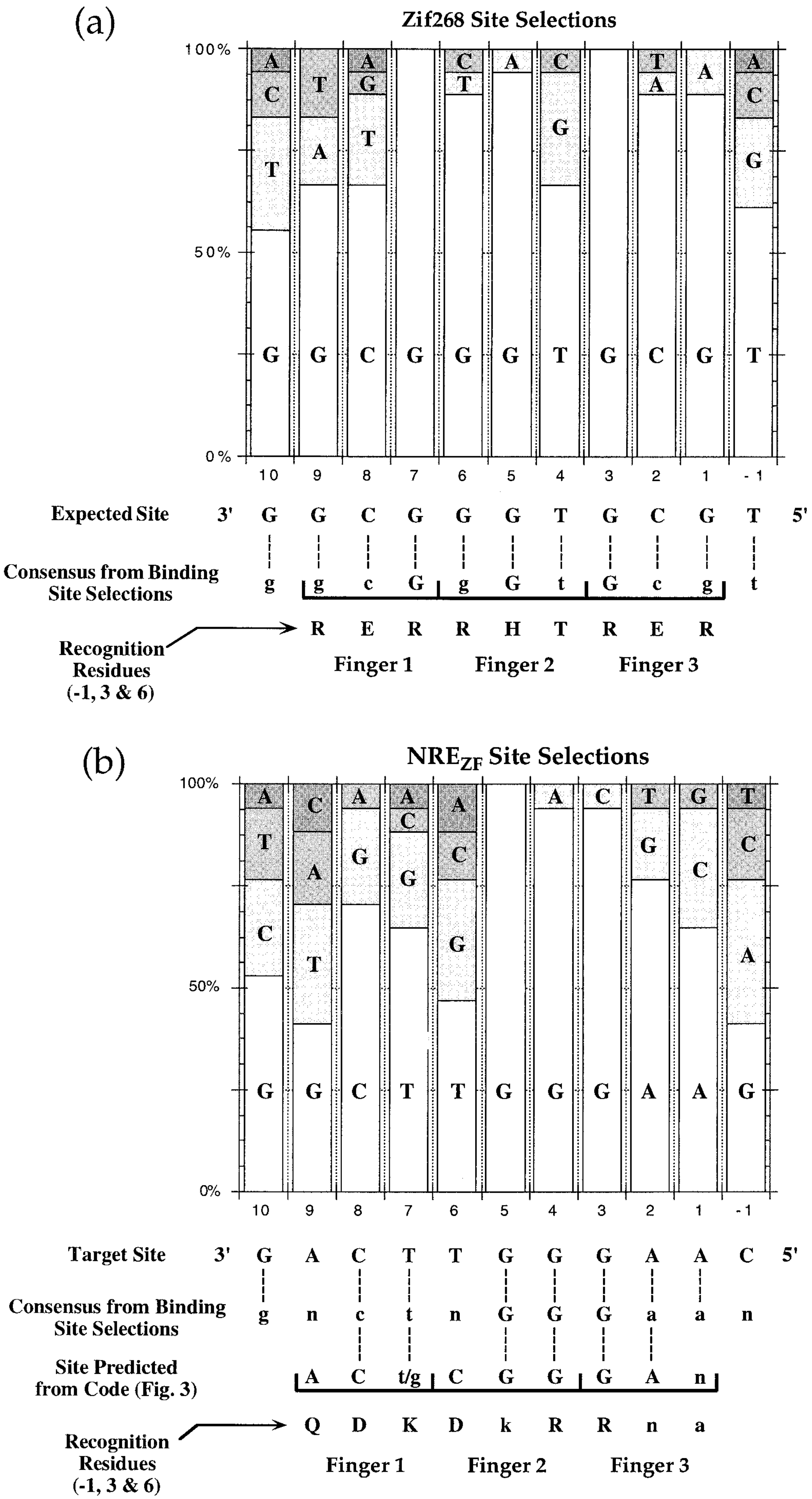


Figure 4(a) and (b) (*legend on page* 1923)

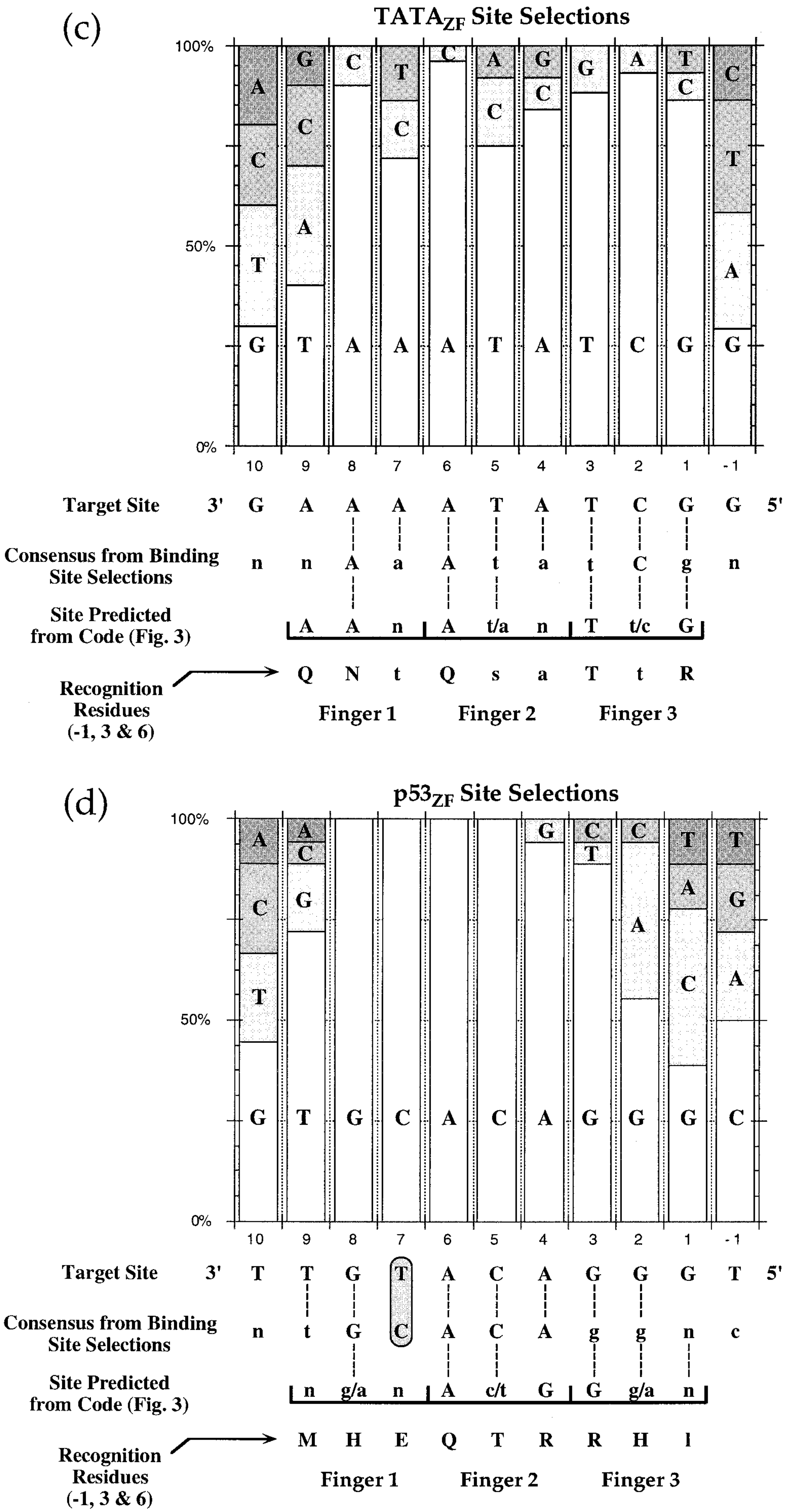


Figure 4(c) and (d) *(legend opposite)*

try to predict the binding site specificity of the NRE$_{ZF}$ protein from its amino acid sequence. This code would predict a binding site of the form (3′-AC(T/G)CGGGAN-5′), and this predicted sequence matches the consensus site at six of nine positions (3′-N**CT**N**GGGA**A-5′). There are two positions where the code would predict a consensus but none was observed, and one position (adenine on the 5′ end) where a consensus was observed even though the alanine at position 6 of finger 3 would not contribute any observed coded contact.

Site selections with the TATA$_{ZF}$ protein (Figure 4(c)) provide another interesting test for the recognition code. This case is intriguing because the target site for the TATA$_{ZF}$ protein is very A+T-rich (Figure 2(e)), while the majority of information about zinc finger recognition is derived from proteins that recognize G+C-rich sites. The consensus sequence obtained from the DNA site selections (3′-NAAATATCG-5′; Figure 4(c)) agrees at eight of nine positions with the expected target site (3′-A**AAATATCG**-5′). As indicated in Figure 4(c), using the recognition code and the sequence of the relevant TATA$_{ZF}$ protein, the code would predict a binding site of the form 3′-AANA(T/A)NT(T/C)G-5′. At six of the nine positions, this matches the consensus sequence obtained *via* site selections. Curiously, one contact that is clearly predicted by the code is not seen in the consensus binding site: as with NRE$_{ZF}$, glutamine at position −1 of the N-terminal finger fails to effectively specify adenine. There also is a clear consensus at one position in the binding site (3′-NAAATATCG-5′) where the corresponding residues (threonine at position 6 of finger 1 or threonine at position 2 of finger 2) would not make any previously observed coded contacts.

The consensus site selected for the p53$_{ZF}$ protein fits these same general trends. Agreement between the consensus sequence and the original target site was very high, with matches at seven out of nine positions. As with the other proteins, using the proposed zinc finger-DNA recognition code (Figure 3) and the sequence of the p53$_{ZF}$ protein, a plausible prediction of the consensus binding site was obtained. In this case the code would predict a binding site of the form 3′-N(G/A)NA(C/T)GG(G/A)N-5′, and this matches the consensus sequence at six of nine positions (3′-T**GCAC**A**GGN**-5′). On the whole, the code is as effective as at other sites, but this case also reveals one outright error in the coded predictions: one would predict that arginine at position 6 of finger 2 would specify guanine (Figure 4(d)), but the DNA site selection shows that adenine is the preferred base at this position. (Note: using a simple code assumes a canonical spacing of the fingers, but sequential selection may allow some variation in spacing and this could be an issue for fingers 2 and 3 of p53$_{ZF}$ (Greisman & Pabo, 1997).)

Perhaps the biggest surprise for the p53$_{ZF}$ site involves the base at position 7: site selections show a clear preference for a C at position 7 of the consensus sequence (3′-TGCACAGGN-5′), even though there was a T at this position in the target site (3′-TG**T**ACAGGG-5′)! This result was so surprising that we measured dissociation constants for target sites that had each of the four possible base-pairs at position 7 of the p53$_{ZF}$ site. We found that the p53$_{ZF}$ protein did have a weak (1.7-fold) preference for C over T but that this protein strongly discriminated against purines at position 7 (binding ~20-fold less well than C).

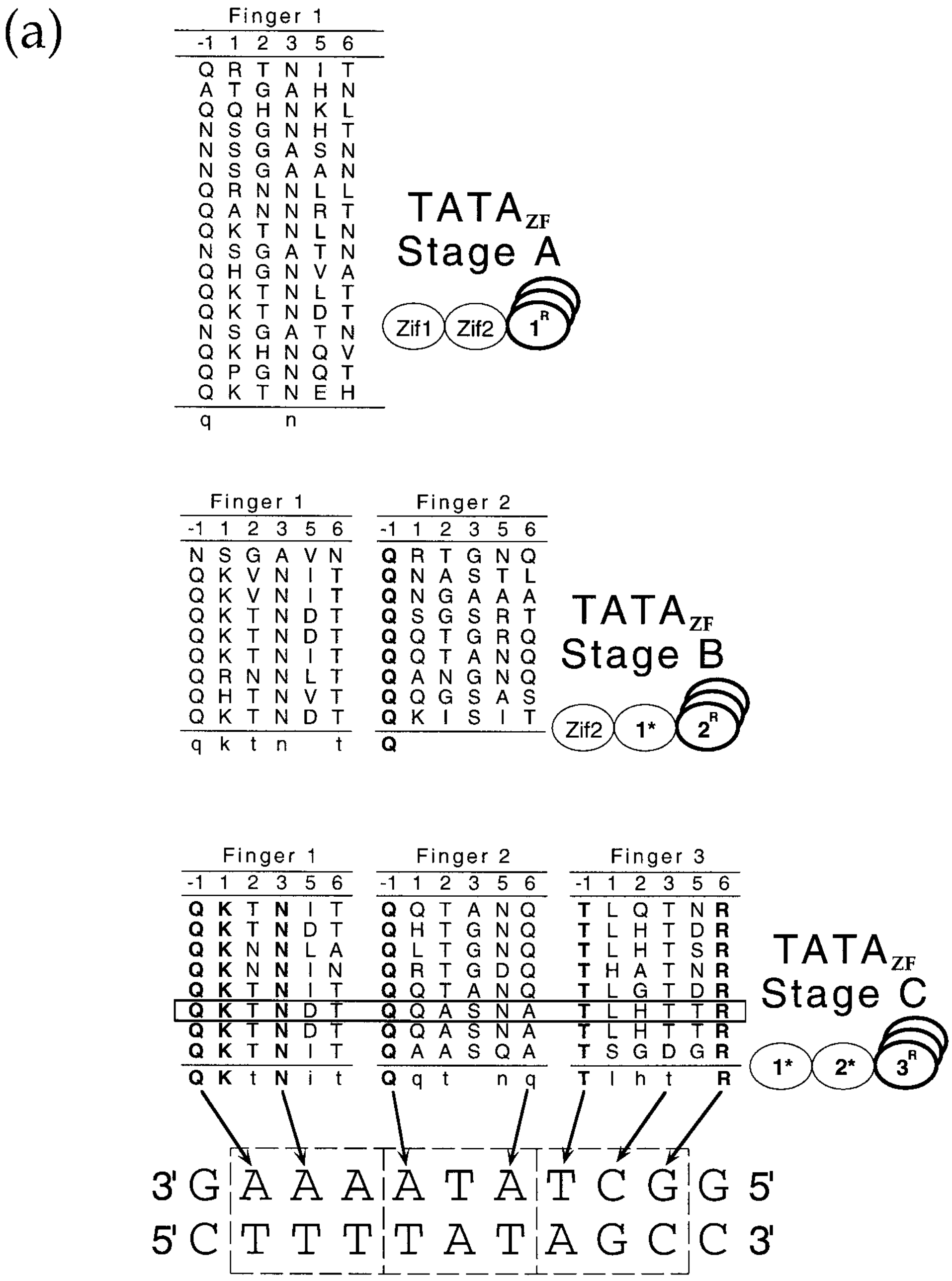
Context-dependent effects and the "evolutionary history" of sequential selection

Although the proposed code (Figure 3) does not yet take context-dependent effects into account, it seems plausible that they may be involved in zinc finger-DNA recognition: the optimal side-chain-

Figure 4. Results of the DNA site selections that were performed with (a) Zif268, (b) NRE$_{ZF}$, (c) TATA$_{ZF}$ and (d) p53$_{ZF}$ proteins. In these experiments 17 to 18 clones of each binding site had been sequenced after the final round of selections, and for each protein these were aligned to give the consensus binding site shown. The data are presented in the form of a histogram, with the most commonly occurring nucleotide for each position shown at the base of the column and the least common (if represented in the aligned sequences) at the top. The sequence of the expected site or the target site used for the selection of the protein (oriented 3′ to 5′) is shown at the bottom of the graph. The consensus binding site from the DNA site selections is listed just below the expected site. A capital letter indicates that a particular base occurred at that position in at least 90 % of the aligned sequences, while a lower case letter denotes that the base was present 50-90 % of the time. For the NRE$_{ZF}$, TATA$_{ZF}$, and p53$_{ZF}$ clones used in the site selections, the binding site that would be predicted based on the recognition code (Figure 3) is listed below the consensus sequence. Broken lines connect those positions where there is a match between the expected site and the consensus sequence, or the predicted site and the consensus sequence. The central nine bases of the binding site are bracketed and for reference the amino acids at positions −1, 3 and 6 of each finger are listed below the DNA base they could contact assuming a canonical docking of the finger with the DNA (Figure 3(b)). The position of the DNA base within the binding site is indicated below each column, using the numbering scheme described in the legend to Figure 1. As explained in Materials and Methods, the TATA$_{ZF}$ consensus sequence (c) represents a composite from three separate experiments in which five base-pair regions had been randomized under each finger. The results are internally consistent for all of the overlapping bases. For the p53$_{ZF}$ selections (d) there was one position (indicated with a shaded oval and discussed in the text) at which the site selections gave a clear consensus for a base (C) different from that which had been present in the original target site (T). (Note: the alignment of recognition residues shown in (d) assumes a canonical spacing for the fingers, but there is a possibility that fingers 2 and 3 of p53$_{ZF}$ bind with some alternate spacing.)

base contacts for a given finger may, to some extent, be influenced by neighboring fingers or subsites. So far the only widely recognized effect of this type involves the residue at position 2, which can contact a base in the subsite of an adjacent finger (Figure 3(b)). Data from the intermediate stages of our sequential selection protocol provide one way of looking for context-dependent effects. We can obtain data about the evolutionary history of our clones, since we actually carried a pool of sequences forward from one step to the next. Thus finger 1 was originally optimized in the context indicated by Figure 2(a).) and then this pool of sequences was carried forward to the next stage, and finger 1 sequences were reoptimized in the context indicated by Figure 2(b). (Likewise, a set of finger 1-finger 2 sequences were carried forward to the next round and reoptimized in the context indicated by Figure 2(c).) Any systematic difference in the finger 1 sequences preferred at stage A and stage B would provide evidence for context-dependent effects in zinc finger recognition.

To obtain data about the evolutionary history of our proteins, we sequenced numerous clones from the intermediate stages of the selections that had yielded the TATA$_{ZF}$, NRE$_{ZF}$ and p53$_{ZF}$ proteins. Sequences from the intermediate stages of the TATA$_{ZF}$ selection (Figure 5(a)) reveal that fingers 1 and 2 continue to undergo optimization throughout the protocol, but there is no evidence for con-



Finger 1					
-1	1	2	3	5	6
Q	R	T	N	I	T
A	T	G	A	H	N
Q	Q	H	N	K	L
N	S	G	N	H	T
N	S	G	A	S	N
N	S	G	A	A	N
Q	R	N	N	L	L
Q	A	N	N	R	T
Q	K	T	N	L	N
N	S	G	A	T	N
Q	H	G	N	V	A
Q	K	T	N	L	T
Q	K	T	N	D	T
N	S	G	A	T	N
Q	K	H	N	Q	V
Q	P	G	N	Q	T
Q	K	T	N	E	H
q			n		

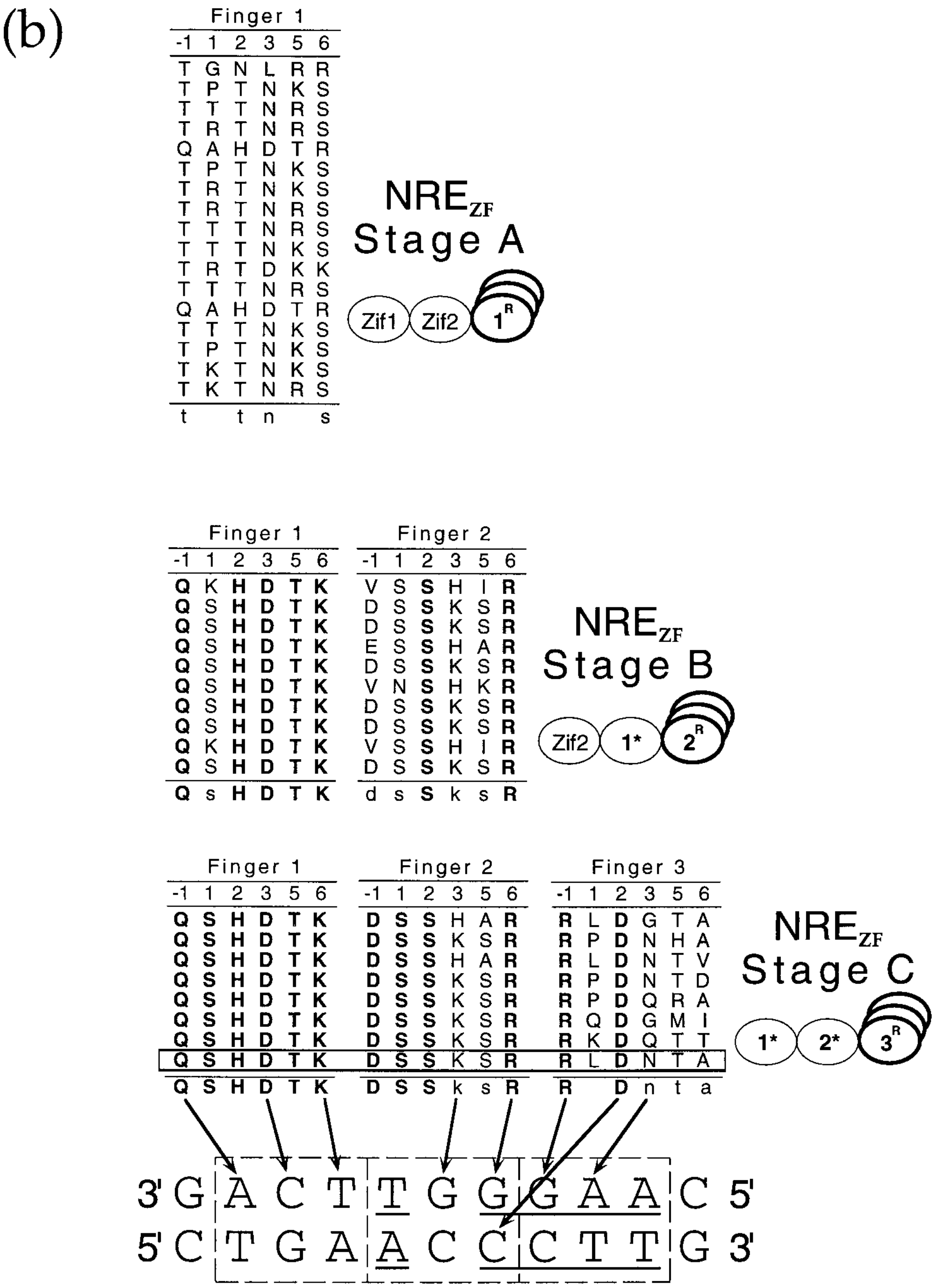
Finger 1						Finger 2					
-1	1	2	3	5	6	-1	1	2	3	5	6
N	S	G	A	V	N	**Q**	R	T	G	N	Q
Q	K	V	N	I	T	**Q**	N	A	S	T	L
Q	K	V	N	I	T	**Q**	N	G	A	A	A
Q	K	T	N	D	T	**Q**	S	G	S	R	T
Q	K	T	N	D	T	**Q**	Q	T	G	R	Q
Q	K	T	N	I	T	**Q**	Q	T	A	N	Q
Q	R	N	N	L	T	**Q**	A	N	G	N	Q
Q	H	T	N	V	T	**Q**	Q	G	S	A	S
Q	K	T	N	D	T	**Q**	K	I	S	I	T
q	k	t	n		t	**Q**					

Finger 1						Finger 2						Finger 3					
-1	1	2	3	5	6	-1	1	2	3	5	6	-1	1	2	3	5	6
Q	**K**	T	**N**	I	T	**Q**	Q	T	A	N	Q	**T**	L	Q	T	N	**R**
Q	**K**	T	**N**	D	T	**Q**	H	T	G	N	Q	**T**	L	H	T	D	**R**
Q	**K**	N	**N**	L	A	**Q**	L	T	G	N	Q	**T**	L	H	T	S	**R**
Q	**K**	N	**N**	I	N	**Q**	R	T	G	D	Q	**T**	H	A	T	N	**R**
Q	**K**	T	**N**	I	T	**Q**	Q	T	A	N	Q	**T**	L	G	T	D	**R**
Q	**K**	T	**N**	D	T	**Q**	Q	A	S	N	A	**T**	L	H	T	T	**R**
Q	**K**	T	**N**	D	T	**Q**	Q	A	S	N	A	**T**	L	H	T	T	**R**
Q	**K**	T	**N**	I	T	**Q**	A	A	S	Q	A	**T**	S	G	D	G	**R**
Q	**K**	t	**N**	i	t	**Q**	q	t		n	q	**T**	l	h	t		**R**

Figure 5(a) (*legend on page* 1926)

text-dependent effects. Thus the finger 1 pool after stage A (top panel of Figure 5(a)) displays a consensus at key α-helical positions that is consistent with that found in the final protein. Finger 1 undergoes further selection in stages B and C to reach a consensus at all six positions. (Of course, stochastic processes may play some role in these selections. If several fingers in a previously selected pool are about equally effective, the eventual "winner" may depend primarily on who is paired with the best finger in the next round.) Finger 2 displays similar behavior, with continued optimization in the next round but no evidence for context-dependent effects. Comparing the consensus of the $TATA_{ZF}$ clones for fingers 1 and 2 with other proteins that recognize the same sites in different contexts (Gfi-1, finger 5; YY1, finger 4; and CF2-II, finger 4; Table 1 (page 1927)) shows that the corresponding fingers have identical, or very similar, amino acid residues at positions −1, 3, and 6 (Gogos *et al.*, 1992; Hyde-DeRuyscher *et al.*, 1995; Zweidler-Mckay *et al.*, 1996). In this case it appears that context-dependent effects are minimal and that these fingers can bind and function as relatively independent modular units.

In contrast, sequences from the intermediate stages of the NRE_{ZF} selection provide evidence for some context-dependent effects and also raise



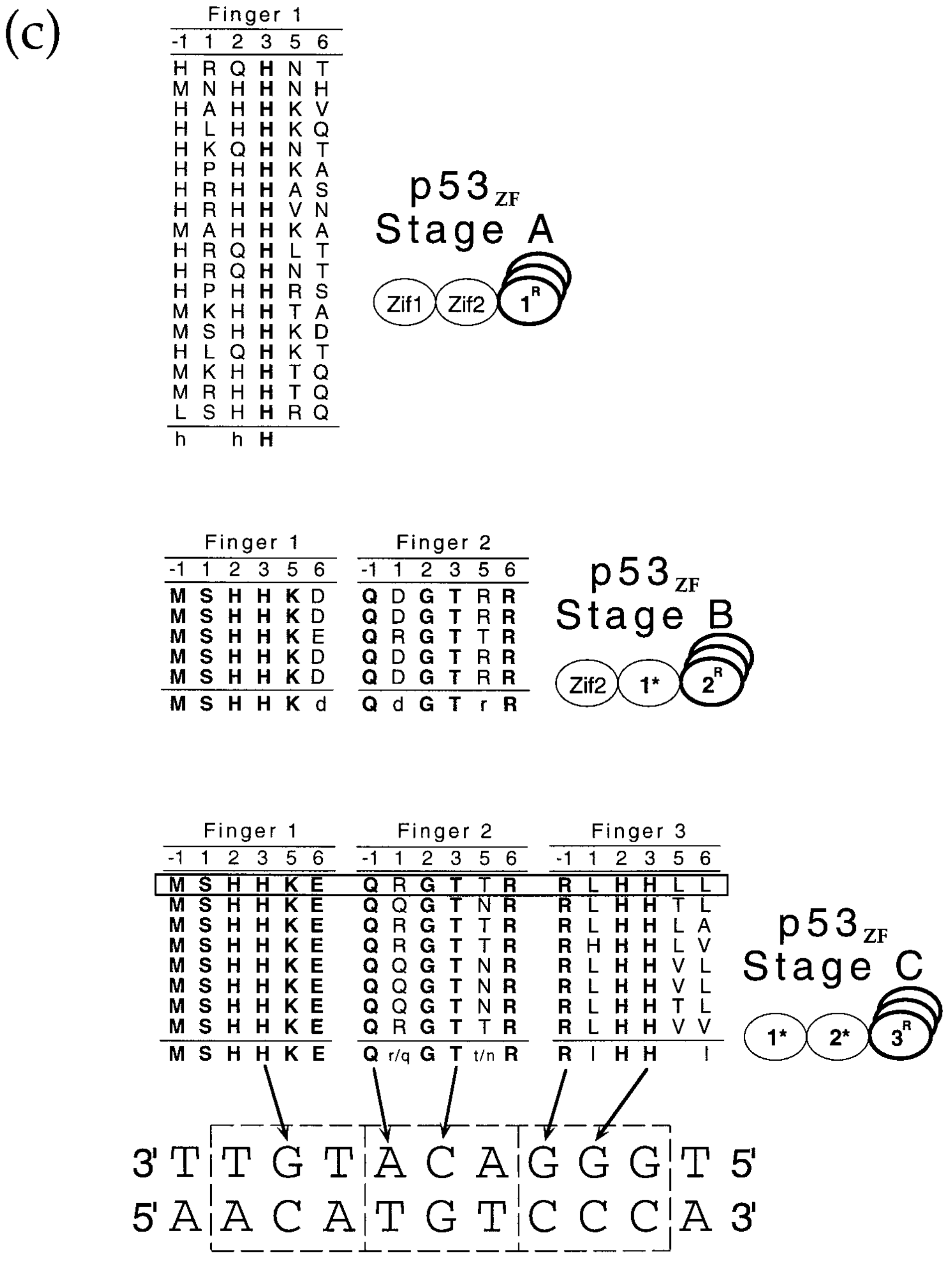
Finger 1					
-1	1	2	3	5	6
T	G	N	L	R	R
T	P	T	N	K	S
T	T	T	N	R	S
T	R	T	N	R	S
Q	A	H	D	T	R
T	P	T	N	K	S
T	R	T	N	K	S
T	R	T	N	R	S
T	T	T	N	R	S
T	T	T	N	K	S
T	R	T	D	K	K
T	T	T	N	R	S
Q	A	H	D	T	R
T	T	T	N	K	S
T	P	T	N	K	S
T	K	T	N	K	S
T	K	T	N	R	S
t		t	n		s

Finger 1						Finger 2					
-1	1	2	3	5	6	-1	1	2	3	5	6
Q	K	**H**	**D**	**T**	**K**	V	S	**S**	H	I	**R**
Q	S	**H**	**D**	**T**	**K**	D	S	**S**	K	S	**R**
Q	S	**H**	**D**	**T**	**K**	D	S	**S**	K	S	**R**
Q	S	**H**	**D**	**T**	**K**	E	S	**S**	H	A	**R**
Q	S	**H**	**D**	**T**	**K**	D	S	**S**	K	S	**R**
Q	S	**H**	**D**	**T**	**K**	V	N	**S**	H	K	**R**
Q	S	**H**	**D**	**T**	**K**	D	S	**S**	K	S	**R**
Q	S	**H**	**D**	**T**	**K**	D	S	**S**	K	S	**R**
Q	K	**H**	**D**	**T**	**K**	V	S	**S**	H	I	**R**
Q	S	**H**	**D**	**T**	**K**	D	S	**S**	K	S	**R**
Q	s	**H**	**D**	**T**	**K**	d	s	**S**	k	s	**R**

Finger 1						Finger 2						Finger 3					
-1	1	2	3	5	6	-1	1	2	3	5	6	-1	1	2	3	5	6
Q	**S**	**H**	**D**	**T**	**K**	**D**	**S**	**S**	H	A	**R**	**R**	L	**D**	G	T	A
Q	**S**	**H**	**D**	**T**	**K**	**D**	**S**	**S**	K	S	**R**	**R**	P	**D**	N	H	A
Q	**S**	**H**	**D**	**T**	**K**	**D**	**S**	**S**	H	A	**R**	**R**	L	**D**	N	T	V
Q	**S**	**H**	**D**	**T**	**K**	**D**	**S**	**S**	K	S	**R**	**R**	P	**D**	N	T	D
Q	**S**	**H**	**D**	**T**	**K**	**D**	**S**	**S**	K	S	**R**	**R**	P	**D**	Q	R	A
Q	**S**	**H**	**D**	**T**	**K**	**D**	**S**	**S**	K	S	**R**	**R**	Q	**D**	G	M	I
Q	**S**	**H**	**D**	**T**	**K**	**D**	**S**	**S**	K	S	**R**	**R**	K	**D**	Q	T	T
Q	**S**	**H**	**D**	**T**	**K**	**D**	**S**	**S**	K	S	**R**	**R**	L	**D**	N	T	A
Q	**S**	**H**	**D**	**T**	**K**	**D**	**S**	**S**	k	s	**R**	**R**		**D**	n	t	a

Figure 5(b) (*legend overleaf*)

interesting questions about the recognition code. The NRE$_{ZF}$ finger 1 sequences have a clear consensus after the first stage of the selection (upper panel in Figure 5(b)), but this sequence (with threonine, asparagine and serine at positions −1, 3, and 6) is radically different from the final sequences



Finger 1 -1	1	2	3	5	6
H	R	Q	**H**	N	T
M	N	H	**H**	N	H
H	A	H	**H**	K	V
H	L	H	**H**	K	Q
H	K	Q	**H**	N	T
H	P	H	**H**	K	A
H	R	H	**H**	A	S
H	R	H	**H**	V	N
M	A	H	**H**	K	A
H	R	Q	**H**	L	T
H	R	Q	**H**	N	T
H	P	H	**H**	R	S
M	K	H	**H**	T	A
M	S	H	**H**	K	D
H	L	Q	**H**	K	T
M	K	H	**H**	T	Q
M	R	H	**H**	T	Q
L	S	H	**H**	R	Q
h		h	**H**		

Finger 1 -1	1	2	3	5	6	Finger 2 -1	1	2	3	5	6
M	**S**	**H**	**H**	**K**	D	**Q**	D	**G**	**T**	R	**R**
M	**S**	**H**	**H**	**K**	D	**Q**	D	**G**	**T**	R	**R**
M	**S**	**H**	**H**	**K**	E	**Q**	R	**G**	**T**	T	**R**
M	**S**	**H**	**H**	**K**	D	**Q**	D	**G**	**T**	R	**R**
M	**S**	**H**	**H**	**K**	D	**Q**	D	**G**	**T**	R	**R**
M	**S**	**H**	**H**	**K**	d	**Q**	d	**G**	**T**	r	**R**

Finger 1 -1	1	2	3	5	6	Finger 2 -1	1	2	3	5	6	Finger 3 -1	1	2	3	5	6
M	**S**	**H**	**H**	**K**	**E**	**Q**	R	**G**	**T**	T	**R**	**R**	L	**H**	**H**	L	L
M	**S**	**H**	**H**	**K**	**E**	**Q**	Q	**G**	**T**	N	**R**	**R**	L	**H**	**H**	T	L
M	**S**	**H**	**H**	**K**	**E**	**Q**	R	**G**	**T**	T	**R**	**R**	L	**H**	**H**	L	A
M	**S**	**H**	**H**	**K**	**E**	**Q**	R	**G**	**T**	T	**R**	**R**	H	**H**	**H**	L	V
M	**S**	**H**	**H**	**K**	**E**	**Q**	Q	**G**	**T**	N	**R**	**R**	L	**H**	**H**	V	L
M	**S**	**H**	**H**	**K**	**E**	**Q**	Q	**G**	**T**	N	**R**	**R**	L	**H**	**H**	V	L
M	**S**	**H**	**H**	**K**	**E**	**Q**	Q	**G**	**T**	N	**R**	**R**	L	**H**	**H**	T	L
M	**S**	**H**	**H**	**K**	**E**	**Q**	R	**G**	**T**	T	**R**	**R**	L	**H**	**H**	V	V
M	**S**	**H**	**H**	**K**	**E**	**Q**	r/q	**G**	**T**	t/n	**R**	**R**	l	**H**	**H**		l

Figure 5. Amino acid sequences that show the evolutionary history of the phage pools (corresponding with stages A, B and C of the selections as indicated in Figure 2 and reported by Greisman & Pabo (1997)). Sequences are presented for phage selected at the TATA$_{ZF}$ (a), NRE$_{ZF}$ (b) and p53$_{ZF}$ (c) target sites. The amino acid position within the recognition helix of each finger is indicated at the top of each column, and the consensus sequence for that column is indicated at the bottom. A bold uppercase letter indicates absolute conservation of an amino acid, while a lowercase letter indicates the residue is present in at least 50 % of the clones. The corresponding selection stage (as described in the legend to Figure 2) for each set of sequences is indicated at the right-hand side. The DNA target site used in selecting the protein is shown below the final set of sequences. As before, plausible base contacts predicted from the code (and assuming a canonical three base-pair spacing) are indicated with arrows. As mentioned earlier, the NRE$_{ZF}$ binding site under finger 2 and 3 matches the Tramtrack site at 5 of 6 positions (underlined bases), and residues that make base specific contacts in the Tramtrack complex were also recovered in these selections (Greisman & Pabo, 1997). In each case, the specific clone that was used for biochemical studies has been boxed. At stage B of the p53$_{ZF}$ selections we found several clones that still had both of the wild-type Zif fingers still attached (as in stage A), but such sequences were eliminated from the current analysis and are not shown in this Figure.

Table 1. Homology in amino acid sequence and DNA site specificity between selected and naturally occurring zinc fingers

Protein	Finger no.	Triplet recognized (3′ → 5′)	Helix sequence (−1, 1, 2, 3, 5 & 6)
NRE_{ZF}	1	ACT	**Q**SH**D**T**K**
Gfi-1	4	ACT	**Q**KS**D**K**K**
NRE_{ZF}	2	TGG	D**SSK**S**R**
YY1	2	CGG	E**SSK**K**R**
Tramtrack	1	TAG	HI**S**NC**R**
NRE_{ZF}	3	GAA	**R**?**DNTA**
Tramtrack	2	GAA	**R**K**DNTA**
$TATA_{ZF}$	1	AAA	**Q**K**TNIT**
Gfi-1	5	AAA	**Q**SS**NIT**
YY1	4	aAA	**Q**S**TN**KS
$TATA_{ZF}$	2	ATA	**Q**QT?N**Q**
CF2-II	4	ATA	**Q**SNTK**Q**
$p53_{ZF}$	3	GGG	**R**LH**H**?**L**
GKLF	3	G(g/a)(g/a)	**R**SD**H**A**L**

(with glutamine, aspartic acid and lysine at corresponding positions). As stage B is completed, the consensus sequence for finger 1 changes dramatically and comes to resemble that found in the final optimized protein. These changes in the consensus sequence between stage A and stage B of the NRE_{ZF} selections provide strong presumptive evidence for context-dependent effects in zinc finger recognition. In this regard, it also is interesting that a natural homolog exists that resembles the final optimized finger: finger 4 of Gfi-1 recognizes the same triplet as finger 1 of the NRE_{ZF} protein, and it has the same amino acid residues at positions −1, 3 and 6 (Zweidler-Mckay *et al.*, 1996; Table 1). It is also intriguing that the final finger 1 sequence for NRE_{ZF} makes plausible coded contacts, while the consensus sequence after stage A cannot be rationalized from the known code (Figure 3).

Examining the evolutionary history of the $p53_{ZF}$ clones also reveals changes in the amino acid consensus of the fingers during the course of sequential selections (Figure 5(c)), and again suggests that some context-dependent effects may be involved. Positions −1 and 6 of finger 1 are especially interesting, since there initially is a consensus (stage A) for histidine at position −1 and no meaningful consensus at position 6. However, after finger 2 is added (stage B) we find that finger 1 consistently has a methionine at position −1 and an acidic residue (glutamic or aspartic acid) at position 6. Curiously, none of these particular preferences, as seen either at stage A or stage B, can be explained by the existing code. As in the NRE_{ZF} selection, these results suggest that context-dependent effects may influence recognition, but the nature of these effects is not understood, and they are not accounted for in any proposed zinc finger recognition code. At a practical level, our sequential selection strategy seems to readily sort out context-dependent effects by reoptimizing fingers at later stages of the protocol. Reoptimization can readily occur in our protocol because we always carry a small pool of clones forward from one stage to the next. (These issues raise some interesting theoretical questions about sampling efficiency and library size, but the success of our method suggests that these are not serious practical problems. Furthermore, the optimal sequences for finger 1 of NRE_{ZF} and $p53_{ZF}$ were confirmed by the reselection experiments described below.)

Reselection of finger 1 for the NRE_{ZF} and $p53_{ZF}$ proteins

The role of context dependence and questions about sampling efficiency also were addressed by reselecting finger 1 of the NRE_{ZF} protein and finger 1 of the $p53_{ZF}$ protein in arrangements where they were coupled to the final optimized versions of fingers 2 and 3 (as indicated schematically in Figure 2(d)). We focused on finger 1, since it (of all the fingers) is originally selected in the least relevant context, and we studied NRE_{ZF} and $p53_{ZF}$ since the changes observed between stages A and B (discussed above) suggested that context-dependent effects may have influenced the initial results. The reselection of finger 1 from $p53_{ZF}$ also allows us to reexamine the problem of distinguishing between T and C at base position 7 of the $p53_{ZF}$ site. (As noted earlier, the target was 3′-TGTA-CAGGG-5′, but the $p53_{ZF}$ protein has a slight preference for C at this position.) Finally, reoptimizing offers a useful perspective on the overall efficacy of our selection scheme and on the quality of the proteins obtained *via* sequential optimization.

In these experiments, finger 1 of the $p53_{ZF}$ protein and finger 1 of the NRE_{ZF} protein were reselected by phage display under buffer conditions identical with those used previously, and a high concentration of non-specific competitor (1.5 mg/ml calf thymus DNA) was used to help ensure the specificity of the selected proteins. These experiments used a single, previously selected sequence for the finger 2/finger 3 region of each protein, but a new finger 1 library was constructed and codons corresponding to positions −1, 1, 2, 3, 5 and 6 were randomized. After eight rounds of selection and amplification, the retention efficiencies for the NRE_{ZF} and $p53_{ZF}$ phage pools plateaued, but sequencing of these pools did not reveal a clear consensus for the $p53_{ZF}$ sequences. Additional rounds of selection were performed with these phage pools using a semi-specific competitor DNA. (These oligonucleotides contained the finger 2/finger 3 recognition region coupled with non-cognate DNA sequences in the region where finger 1 would bind.) After four additional rounds of selection and amplification under these conditions, the phage retention efficiencies plateaued again, and the final pools were sequenced.

Both the $p53_{ZF}$ and NRE_{ZF} selections reached a consensus after these four additional rounds of selection (Figure 6). In both cases, it turns out that the primary consensus sequence for the reselected finger had changes near the N terminus of the recognition helix (positions −1, 1 and 2) while retaining the original residues at positions 3 and 6. Thus it appears as if this end of the finger (which is at one extreme end of the binding site) was most influenced by context-dependent effects in the original selection protocol. For both NRE_{ZF} and $p53_{ZF}$, reselecting finger 1 gives a shorter amino acid at position −1. (The primary NRE_{ZF} consensus now has asparagine instead of glutamine, and the $p53_{ZF}$ consensus now has threonine instead of methionine.) At position 1 of each protein there tends to be a longer, positively charged amino acid (lysine or arginine) replacing the serine obtained with the original sequential selection protocol. These residues may make phosphate contacts as observed in the Tramtrack structure (Fairall *et al.*, 1992). Comparing sequences of the NRE_{ZF} clones also revealed a secondary subpopulation with a consensus sequence that was identical at four of the six positions with that obtained using the original sequential selection protocol.

Specific and non-specific dissociation constants were determined to directly assess the effect of these changes in sequence. For these studies we selected clones (boxed sequences in Figure 6) that most closely matched the new consensus for the $p53_{ZF}$ protein and the new primary and secondary consensus sequences for the NRE_{ZF} protein. These proteins were recloned into an appropriate vector, expressed, and purified along with corresponding proteins from the original selections (using the NRE_{ZF} and $p53_{ZF}$ clones indicated by boxes in Figure 5(b) and 5(c)). Dissociation constants for these five proteins were determined by electrophoretic mobility shift analysis (Table 2). Among the NRE_{ZF} clones, the protein with the new consensus consistently bound the DNA site with twofold

(a)

Clones from NRE_{ZF} Finger 1 Reselection

	Helix Position					
CLONE #	-1	1	2	3	5	6
1	N	R	T	D	G	K
9	N	R	T	D	G	K
3	N	R	T	D	Q	K
11	N	R	T	D	G	K
14	N	R	T	D	G	K
6	N	K	T	D	G	K
7	N	K	T	D	G	K
8	N	K	T	D	G	K
10	N	K	T	D	G	K
15	N	K	T	D	G	K
4	Q	S	G	D	Q	K
12	Q	S	G	D	A	K
13	Q	S	G	D	A	K
16	Q	S	G	D	V	K
2	R	K	D	D	T	K
5	R	K	D	D	T	K
Primary Consensus	n	+	t	D	g	K
Secondary Consensus	q	s	g	D	?	K
Previous Consensus	Q	S	H	D	T	K

(b)

Clones from $p53_{ZF}$ Finger 1 Reselection

	Helix Position					
CLONE #	-1	1	2	3	5	6
1	T	R	Q	H	S	E
3	T	K	Q	H	K	E
4	T	Q	Q	H	R	E
5	T	Q	Q	H	R	E
9	T	A	Q	H	R	E
11	T	S	Q	H	A	E
7	T	R	Q	H	E	D
12	T	R	Q	H	E	D
13	T	R	G	H	E	D
6	N	Q	Q	H	K	E
2	R	Q	G	A	S	E
8	G	K	S	Q	E	D
10	G	K	S	Q	E	D
Consensus	t	+	q	h	?	e
Previous Consensus	M	S	H	H	K	E

Figure 6. Finger 1 amino acid sequences for the (a) NRE_{ZF} and (b) $p53_{ZF}$ proteins following 12 rounds of reselection by phage display. As indicated in Figure 2(d), the randomized finger 1 was reselected in a construct containing previously optimized finger 2 and finger 3 sequences. Residues selected at each of the six randomized positions are shown. For both NRE_{ZF} and $p53_{ZF}$, the primary consensus sequence obtained after reselection differs from the previous consensus sequence (given at the bottom of each Table) at positions −1, 1 and 2. The final NRE_{ZF} pool also had sequences conforming to a secondary consensus that was more closely related to the consensus obtained in the original sequential selections (Greisman & Pabo, 1997). Capital letters in the consensus indicate positions that were absolutely conserved in the final phage pools. Lowercase letters denote those positions where a particular amino acid occurs 50-95 % of the time; + sign indicates a preference for a positively charged amino acid. Boxes highlight those clones for which the K_d was measured.

Table 2. Dissociation constants of zinc finger clones selected by phage display

Clone	K_d (pM)	K_d^{NS} (μM)	K_d^{NS}/K_d
NRE_{ZF} original	27.6 (±13.5)	1.98 (±0.88)	71,700
NRE_{ZF} clone 6	14.1 (±6.6)	1.28 (±0.15)	90,800
NRE_{ZF} clone 12	21.6 (±6.9)	1.62 (±0.51)	75,000
$p53_{ZF}$ original	18.5 (±10.6)	0.59 (±0.16)	31,900
$p53_{ZF}$ clone 3	2.1 (±0.9)	0.65 (±0.18)	309,500
Zif268	10 (±6)[a]	0.32 (±0.3)[a]	31,000[a]
$TATA_{ZF}$	120 (±70)[a]	3.0 (±0.3)[a]	25,000[a]

[a] Data reported by Greisman & Pabo (1997).

higher affinity than the original NRE_{ZF} protein. The dissociation constants for the other two NRE_{ZF} proteins (i.e. the original sequence from the sequential selections and the secondary consensus from the new selections) were essentially identical, as might be expected given the similarities in their sequences. The new $p53_{ZF}$ protein bound 8.8-fold more tightly than the original $p53_{ZF}$ protein, and this difference clearly must result from the amino acid changes at positions −1, 1 and 2 of finger 1. To analyze the specificity of binding, non-specific dissociation constants for all of these proteins were estimated in competition experiments using calf thymus DNA (Table 2). Within experimental error, the non-specific binding constants for all three of the NRE_{ZF} clones were identical. This also was true for the two $p53_{ZF}$ clones, indicating that reselection had increased affinity for the target site without increasing non-specific binding. Thus, the new $p53_{ZF}$ protein had both a higher affinity (K_d) and a higher ratio of specific to non-specific binding (K_d^{NS}/K_d increases). Our data indicate that this reselected protein now binds more tightly and specifically than wild-type Zif268 (Table 2).

Summary and Conclusions

There are many issues one might address in evaluating the idea of a recognition code that can be used for zinc fingers that bind DNA similarly to those found in Zif268. Specifically, we are interested in (1) evaluating the accuracy and utility of existing codes by comparing these with the specificities of natural and selected proteins; (2) improving the code by cataloguing more contacts and accounting for higher-order correlations; and (3) thinking about the evolutionary and structural basis for the proposed code. In short, we want to understand how well predictions from a code correlate with results from selections, and to understand both the practical and theoretical significance of any proposed code. Although we cannot provide definitive answers at this stage, our data from these experiments are relevant to each of these key issues.

The first set of experiments in this paper involved finding the optimal binding sites for a set of proteins that previously had been obtained using our sequential selection protocol (Greisman & Pabo, 1997). These experiments serve to rigorously test our zinc finger selection method and we find that our novel zinc finger proteins have the desired sequence specificity for most of the basepairs in their binding sites. Comparing target sites with consensus sequences obtained from site selection shows agreement, respectively, at seven out of nine ($p53_{ZF}$), seven out of nine (NRE_{ZF}) and eight out of nine ($TATA_{ZF}$) base positions for an overall "score" of 82%. Overall, our sequential selection protocol seems to be very effective in generating zinc finger proteins that recognize the desired target sites. We note that most of the errors (4/5) involve "errors of omission", i.e. positions where no clear consensus is obtained in the binding site selection data. (Position 7 of the $p53_{ZF}$ site was the only case where there is a clear consensus for a base differing from that in the target site.) We also note that our zinc finger proteins tend to be somewhat less effective in specifying bases at the very termini of the binding site. This may be related, as suggested by Choo (1998), to "fraying" of side-chain-base contacts at the very end of the complex.

Aligning the protein sequences with their optimal binding sites also lets us evaluate the effectiveness of the proposed zinc finger DNA recognition code (Figure 3). In general, we find a very good correlation between (1) bases that are observed in the consensus sequence and (2) contacts that would be predicted from a recognition code. Analysis of the data for the various sites shows matches at six out of nine positions for $p53_{ZF}$, six out of nine positions for NRE_{ZF}, and six out of nine positions for $TATA_{ZF}$, giving an overall score of 67%. This is a very significant correlation, and it clearly shows that the code contains a considerable amount of useful information. However, it still is interesting that the score (67%) for a code/consensus site correlation is somewhat lower than the score (82%) for a target site/consensus site correlation. Thus it appears that the code contains a significant amount of information and often can give a correct prediction, but it also appears that selection still has a meaningful advantage when trying to make the very best protein. (Note that our scoring system is rather generous as we always give full credit at positions where the preferred base is one of two alternatives suggested by the code.)

Any systematic analysis of recognition, whether based on ideas about a code or involving proteins obtained *via* phage display, must ultimately consider how well the specificity of the protein for the DNA is defined at each position of the binding site. In this sense, the percentile scoring systems used above are too simple: even at positions where there is a match involving the intended target site, the consensus binding site and the site predicted from a code, one will still be interested in the energetic effectiveness of the discrimination. Our data certainly are consistent with the notion that discrimination is more effective at some positions than others, as suggested by the histograms in Figure 4. However, more precise statistics, with a much larger pool of sequences, or direct K_d measurements with a series of altered sites would be needed to rigorously analyze the specificity at each position. We note that the current codes do not even address this point about relative energetic preference. When several amino acids are listed as alternatives at a given position, one naturally wonders which will be best and whether the choice depends on the context.

Situations where there is some obvious mismatch between the target site, the consensus binding site, and the sequence predicted by the code are the most serious, and yet most straightforward type of problem. One striking example involves the adenine at position 4 of the $p53_{ZF}$ target site (Figure 4(d)). The consensus sequence from binding site selections also has adenine at this position, showing that the previously selected zinc finger proteins can effectively specify this region of the target site. However, if we use a conventional spacing in aligning the $p53_{ZF}$ zinc fingers with this binding site (i.e. using a three base-pair spacing between neighboring fingers) we find that the arginine at position 6 in finger 2 should be paired with the adenine at position 4 in the target site. This clearly defies simple expectations, and it seems obvious that (1) the arginine must make non-canonical contacts (perhaps making some bridging contacts that involve neighboring bases or the phosphate backbone), and/or that (2) the $p53_{ZF}$ fingers may be shifted to have an alternative spacing along the DNA (Greisman & Pabo, 1997). Either possibility involves difficulties in the straightforward application of a recognition code.

Another very intriguing case, which raises fascinating questions about site specificity and energetics of base discrimination, involves position 7 of the $p53_{ZF}$ binding site. The target site used to select $p53_{ZF}$ had a T at this position, but the consensus binding site has a C at position 7. (The recognition code gives no information on what base should be preferred by a protein that has glutamic acid at position 6 of the recognition helix.) Results of the binding site selections, with a fully conserved C at this position, were surprising enough that K_d measurements were done with oligonucleotides containing A, C, G, or T at this position. We find that the $p53_{ZF}$ zinc finger has an appropriate two-fold preference for C over T, but discriminates significantly against A or G. This observation fits with the general notion that zinc fingers may be less effective at specifying pyrimidine than purine bases. It also raises interesting questions about the limits of recognition and specificity. Will it be hard, at this level of detail, to distinguish certain closely related sites, especially when pyrimidine bases are involved? Given these data, it seems quite plausible that we may have selected the best protein for this site (i.e. the target site containing the T) but that it just happens to bind somewhat better to the alternative site (containing C at position 7)! In particular, we note that the reselection of finger 1 of $p53_{ZF}$ gave no changes in the consensus at positions 3 and 6 of the α-helix, suggesting that we may, at least in the current structural context, have done as well as possible for this base. It remains possible, at least in principle, that we might have achieved more effective discrimination if we had randomized additional positions in the finger (perhaps adjusting the structural framework to facilitate recognition) or if we had used a specific counterselection step with the alternative site in an attempt to improve discrimination. However, this result still raises profound questions about the ability of a code to translate from a space of targeted DNA sequences to a space of amino acid sequences. At this stage it appears there are limits on the degree of specificity that can be achieved at a given position.

Our data also provide evidence for context-dependent effects in zinc finger recognition that are not accounted for by any simple recognition code. Thus the proposed code (Figure 3) involves a correspondence between particular positions in the finger and particular base positions in the subsite; it implicitly assumes that the contacts will not be influenced by neighboring fingers or subsites. (The only generally recognized exception, which is illustrated in Figure 3(b), involves a potential contact from position 2 of a finger to the binding site of its N-terminal neighbor.) However, our evolutionary history of the sequential selections suggests that there are other context-dependent effects in zinc finger-DNA recognition. Given that we carry a pool of fingers forward from one round to the next and that we reselect after adding the next random finger, it is not surprising that the finger 1 sequences become less diverse at later stages. However, it is surprising to see a tentative consensus develop at stage A, but then change during stage B. This clearly occurs in our selections, indicating that a relatively rare subpopulation of fingers from stage A (which presumably is less common because it binds less tightly or less specifically in the initial context) can come to predominate at stages B and C. This effect is quite readily seen in examining the NRE_{ZF} selections where the finger 1 sequence t?tn?s, predominates at stage A but changes to QsHDTK during stage B. Although further experiments are needed to determine the basis for these differences, it seems that changes in the structural context (such as the differences between having finger 1 as a C-terminal module at stage A and a central module at stage B) might be responsible.

Our experiments involving the reselection of finger 1, in the context of the final finger 2 and finger

3 sequences, also provide important information about context dependence and the recognition code. Thus, in the original selections of finger 1 for the $p53_{ZF}$ and NRE_{ZF} binding sites, the sequence of this finger reached a strong consensus at the end of stage B, when it occupied the center position (Figure 5(b) and (c)). The reselection of this finger in its final, proper context (Figure 2, stage D) yielded other sequences for residues at positions −1, 1 and 2 of the recognition helix (Figures 6(a) and (b)). We presume that these differences involve some type of context-dependent effect, and it is possible that the allowed docking modes for an N-terminal finger (Figure 2, stage D) are slightly different from those for a central finger (Figure 2, stage B). As we think about a recognition code, it also is fascinating to see how these reselections tend to give correlated changes at positions −1, 1 and 2. There appear to be modest differences in energy between the old and new sequences, but (1) the emergence of a new consensus and (2) the fact that there are correlated changes at several positions provide evidence for context-dependent effects. (It seems plausible that correlated changes at this end of the finger might be mediated directly *via* side-chain-side-chain effects or indirectly *via* subtle changes in the docking.)

One also can see indications of higher-order interactions, not predicted by the code, when comparing sequences of various fingers and subsites. For example, we note that the consensus sequence of NRE_{ZF} finger 2 (with D, S, K and R at positions −1, 2, 3 and 6) is very similar to a finger (DSNR) selected by Rebar & Pabo (1994; Figure 7(a)). Although the only significant difference between these fingers involves position 3, they recognize remarkably different DNA sites (3′-TGG-5′ and 3′-CAG-5′ respectively). This difference in specificity is probably due to the large size of the amino acid at position 3 (lysine) in NRE_{ZF} finger 2 (although it also may be affected by having somewhat different docking arrangements for a terminal finger and a central finger). We note that finger 2 of YY1 has lysine at a corresponding position (Table 1) and the lysine, due to its length, alters the orientation of the recognition helix in the major groove such that the amino acid at position −1 cannot make a canonical contact with the 3′ base in this triplet. With a canonical docking arrangement, we tend to expect a short residue (certainly not a lysine) at position 3 of the finger, and it appears that using larger residues may affect what contacts are possible from neighboring positions of the α-helix. This highlights another weakness of the code: the length/size of an amino acid at one position will influence the type of amino acid that can be effective at other positions of the helix due to the rigidity of this scaffold (Choo & Klug, 1997).

Another interesting example involves comparison between $p53_{ZF}$ finger 2 (with Q, G, T and R at positions −1, 2, 3 and 6) and the QGSR finger selected by Rebar & Pabo (1994; Figure 7(b)). The only differences involve a conservative threonine to serine change at position 3 and a substitution of serine for arginine at position 1. Nevertheless, these fingers specify a different DNA base at the 5′ end of their site. One would naturally anticipate that arginine at position 6 would specify guanine at this position, but the $p53_{ZF}$ finger prefers an adenine. This clearly violates expectations for a code (arginine ↔ guanine interactions are prototypical zinc finger contacts!), and we infer that there must be some alternative contact or docking arrangement that accounts for the difference in specificity, perhaps a bridging contact from the arginine, a different spacing of fingers along the site, or a critical contact from the histidine at position 2 of the flanking finger.

Given this set of experimental approaches, the proteins obtained by sequential selection (Greisman & Pabo, 1997) have provided a rich source of data about zinc finger-DNA recognition, providing new perspectives on ideas about a recognition code. The DNA site selection experiments, by selecting optimal binding sites for our zinc finger proteins, prove that the zinc finger proteins obtained with the sequential selection protocol are highly specific for their target sites. We also see a clear correlation between the optimal binding sites and those that would be predicted from a code based on the amino acid sequence of each protein, but our data suggest that selection still is significantly more powerful than design. Analyzing sequences of the zinc fingers present at intermediate stages of the selection process and data obtained *via* the reselection of finger 1 both provide strong evidence for context-dependent effects. In summary, it appears that the code contains useful information about DNA recognition by this family of zinc finger proteins, but that the proposed codes cannot account for all of the correlated effects. Selection still remains the most reliable method for obtaining zinc finger proteins with optimal specificity for a given site.

Materials & Methods

Expression of proteins

All proteins were expressed as glutathione-*S*-transferase (GST)-fusions in the plasmid pGEX-2T (Pharmacia). The NRE_{ZF}, $p53_{ZF}$ and $TATA_{ZF}$ clones that were used in the current studies correspond with the boxed sequences in Figure 5, and these are the same clones that were studied Greisman & Pabo (1997). Boxes in Figure 6 (which shows sequences obtained after reselecting finger 1) also highlight the clones that were studied in more detail. Residues 333-414 of Zif268 (Christy *et al.*, 1988) were used in the DNA site selection studies and gel-shift assays of this protein. After inducing protein expression and lysing the cells by sonication, GST fusion proteins were purified by batch extraction using GST-resin and the standard procedure provided by Pharmacia. Following elution of the proteins with 10 mM glutathione, thrombin (SIGMA) cleavage was used to remove the GST. Once cleavage had proceeded to completion, as monitored by SDS/polyacrylamide gel electrophoresis, thrombin was inhibited with 1 mM phenylmethylsulfonyl fluoride (PMSF) and the protein solution was aliquoted and stored at −80 °C.

DNA site selections

DNA site selections on Zif268 and the NRE$_{ZF}$ and p53$_{ZF}$ clones were performed with an oligonucleotide containing 20 random bases flanked by primer binding sites. These were synthesized using standard phosphoramidite chemistry (Applied Biosystems 382 DNA synthesizer) and the sequences were as follows: template 5′-GCCGAATTCAGCTTACCAGAN$_{20}$TCGGAACTCGAGTTGGCC-3′; primer A, 5′-GCCGAATTCAGCTTACCAG -3′; and primer B, 5′-GGCCAACTCGAGTTCCG-3′. Restriction sites (underlined) were included to facilitate cloning. Following deprotection, oligonucleotides were purified by denaturing PAGE. About one nanomole of double-stranded DNA was generated for these selection studies by using the template with 20 random bases, primer B, and Klenow DNA polymerase (6 units, Boehringer Mannheim). This duplex was body-labeled by using [α-^{32}P]dCTP instead of dCTP in the extension reaction. The full-length labeled duplex was then purified by non-denaturing PAGE. Site selections with this duplex DNA were performed by using gel electrophoresis to isolate DNA fragments that could form a stable complex with the protein. Approximately 20 pmol of labeled duplex was combined with various concentrations of each protein, using enough to form a detectable protein-DNA complex. Samples were incubated for one hour at room temperature in a site selection buffer (1 × SSB) containing 15 mM Hepes (pH 7.5), 50 mM KCl, 50 mM potassium glutamate, 50 mM potassium acetate, 5 mM MgCl$_2$, 20 μM ZnSO$_4$, 100 μg/ml acetylated bovine serum albumin (Ac-BSA); 0.1 % (w/v) NP-40, 5 % (v/v) glycerol. Samples were then loaded and electrophoresed on a non-denaturing 10 % gel (0.5 × TBE). After electrophoresing for sufficient time to separate free DNA from the protein-DNA complex, the gel was dried and exposed to film. Bands which correspond with the appropriately shifted protein-DNA complex were excised, and the

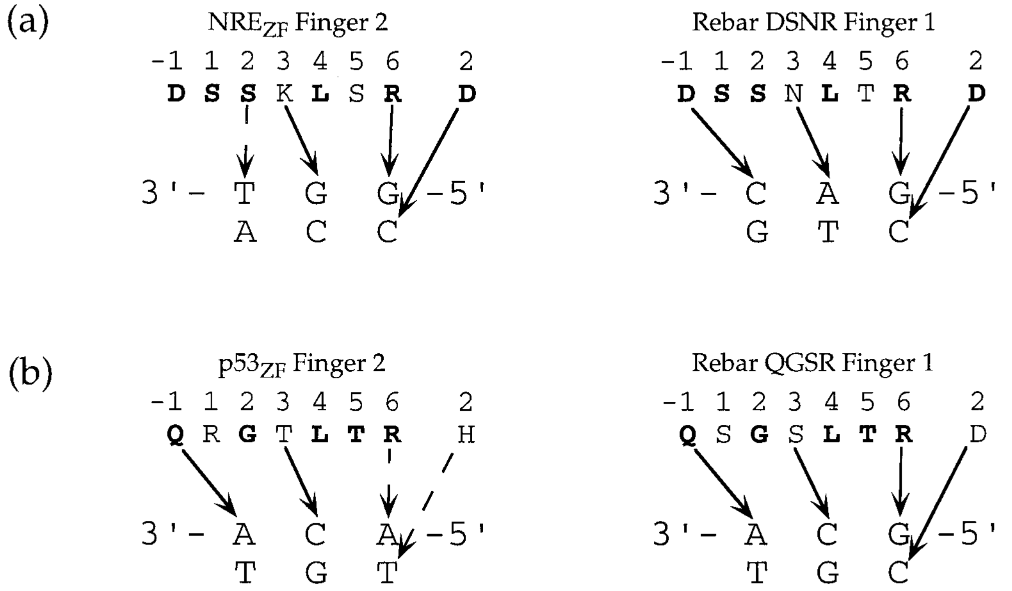


Figure 7. Comparison of fingers with similar amino acid sequences that recognize different DNA triplets. (a) Comparison of DNA sites for NRE$_{ZF}$ finger 2 and DSNR finger 1 from a phage display study by Rebar & Pabo (1994). Key residues in each finger are indicated with single letter designations for the amino acids, and the position with respect to the start of the α-helix is indicated above each amino acid. The amino acid found at position 2 of the neighboring finger on the C-terminal side is indicated to the right of each sequence, and it may also influence the specificity of the protein for these three base-pairs. Amino acids that are common to both fingers (in this case positions −1, 1, 2, 4, 6 and position 2 from the neighboring finger) are indicated in bold. Continuous arrows denote amino acids that appear to specify the sequence preference of these fingers, with proposed contacts based on the recognition code, on recent site selections with the DSNR protein (E.I.R., S.A.W. & C.O.P., unpublished results), on the crystal structure of the DSNR-DNA complex (Elrod-Erickson *et al.*, 1998), and on the crystal structure of the Tramtrack complex (Fairall *et al.*, 1993). Even though these fingers share common amino acids at position −1, 1 and 2, they prefer a different 3′ base (T for NRE$_{ZF}$ and C for DSNR). As discussed in the text, this may be the result of the bulky lysine side-chain at position 3 of NRE$_{ZF}$ finger 2. In the structure of YY1 (Houbaviy *et al.*, 1996), lysine at this position appears to change the orientation of the recognition helix so that the amino acids at position −1 and 2 can no longer contact the 3′ base of the subsite. Because lysine could disrupt the corresponding contact for the NRE$_{ZF}$ finger, we have used a broken arrow in marking the potential contact from position 2 of the NRE$_{ZF}$ finger. (b) Comparison of DNA sites for p53$_{ZF}$ finger 2 and QGSR finger 1 from a selection study by Rebar & Pabo (1994). Alignment of the p53$_{ZF}$ residues with the DNA site assumes a normal three base-pair spacing of fingers, but leads to surprising inferences about contacts with adenine at the 3′ position of this triplet. According to the canonical pattern of contacts, the specificity for this base-pair should be defined by the amino acids at position 6 of this finger and position 2 of the neighboring finger. Although arginine 6 would be expected to specify a guanine at this position, it is conceivable that the spacing of the fingers changes slightly in this region allowing an alternate readout of the DNA. (Sequential selection puts no direct constraint on the spacing of the fingers.)

DNA extracted overnight in 1 ml of a buffer containing 0.5 M NH_4OAc, 10 mM $MgCl_2$, 1 mM EDTA, and 0.1 % (w/v) SDS. The selected DNA was then ethanol precipitated twice and resuspended in 50 μl doubly distilled water. This sample (20 μl) was used for PCR amplification using cloned Pfu (Stratagene), with [α-^{32}P]dCTP to body-label the resulting oligonucleotides. After gel purification, this DNA was used for the next round of site selection, and the entire process was repeated six more times. The final DNA pools were cloned into pBluescript II SK+ (Stratagene). These plasmids were transformed into *Escherichia coli* XL1-Blue (Stratagene), and individual clones were sequenced as described above. DNA sequences from these clones (17 to 18 per protein) were aligned by visual inspection.

DNA site selections for the $TATA_{ZF}$ protein were carried out as described above, except that the optimal binding site was derived as a composite from three sets of overlapping oligonucleotides (which randomize the 5′ end, the middle, and the 3′ end of the site). The target site that had been used for the initial $TATA_{ZF}$ protein selections (Greisman & Pabo, 1997) and the randomized template for our subsite selections were as follows: $TATA_{ZF}$ site GGCTATAAAAG; Template 1, 5′-GCCGAATTCAGCTTACACTCNNNNNNNNTAAAAGCTGAGACTCGAGTGCGC; Template 2, 5′-GCCGAATTCGACTATACCCTNGGCNNNNNAAGNTCAGGCCTCGAGGTCGC; and Template 3, 5′-GCCGAATTCGCAATTCCAGTGGCTATNNNNNNNNATGGACCTCGAGCTGGC. (Underlining emphasizes region corresponding with the original nine base-pair target site.) These three templates were processed as described above to generate labeled duplex DNA, and site selections were performed in parallel, using each template in a separate binding reaction. (Using different flanking sequences for PCR prevented cross-contamination of the samples.) After four rounds of selection, individual clones were sequenced (more than 20 per template), and a composite consensus DNA site was generated (Figure 4(c)). Good agreement at all of the overlapping bases (i.e. at positions 3, 4, 6 and 7 of the site) supported the validity of this method.

Reselection of finger 1 from the NRE_{ZF} and $p53_{ZF}$ proteins

Individual clones from the original $p53_{ZF}$ and NRE_{ZF} selections were chosen for the reselection of finger 1. (These correspond with the clones that had been expressed for binding studies (Greisman & Pabo, 1997), and are indicated by boxes in Figure 5(b) and (c). Positions −1, 1, 2, 3, 5 and 6 from the recognition helix of finger 1 were randomized with 16 of the possible 20 amino acid residues (Cys, Phe, Tyr and Trp were excluded) using the subset of codons provided by the coding scheme (A, C, G) N (C, G). Most aspects of library construction and phage display were carried out as described (Greisman & Pabo, 1997); the library sizes for the NRE_{ZF} and $p53_{ZF}$ pools were 6.1×10^8 and 9.8×10^8, respectively. Due to a slight bias that we observed against G in the randomized bases of the library, this number of clones will sample most, but not all, of the available sequence space. After a series of control studies, we decided to perform the selections outside of the anaerobic chamber and without taking any special precautions to maintain anaerobic conditions. Initial binding reactions for these selections were done in 1 × PSB (25 mM potassium phosphate (pH 7.8), 60 mM potassium glutamate, 60 mM potassium acetate, 2 mM $MgCl_2$, 20 μM $ZnSO_4$, 100 μg/ml Ac-BSA, 5 % (v/v) glycerol, 0.5 % (v/v) Triton X-100, 1.5 mg/ml sheared calf thymus DNA). In later rounds, this binding buffer was supplemented with salt and/or additional non-specific DNA to increase the selection stringency. Due to the high background retention of the randomized $p53_{ZF}$ pool, its binding buffer was supplemented with 50 mM NaCl at round two, 100 mM NaCl at round three, and with 100 mM NaCl and an additional 1.5 mg/ml calf thymus DNA for rounds four to eight. The per cent phage retained for both the $p53_{ZF}$ and NRE_{ZF} pools plateaued at ~1 % by round eight, but sequencing revealed that the $p53_{ZF}$ selections had not reached a consensus. Both the NRE_{ZF} and $p53_{ZF}$ proteins were subjected to an additional four rounds of selection in 1 × PSB supplemented with 100 mM NaCl, 1.5 mg/ml calf thymus DNA and 6 μM semi-specific duplex DNA competitor. The semi-specific competitor is an oligonucleotide which, at each position under the finger 1 binding site, contains every base except that present in the target site. Thus the $p53_{ZF}$ competitor contains sequences of the form: 5′-CCCTTGGGACA(A,C,G)(A,C,T)(A,C,G)(A,C,G)CCTGATCGCGGTTCGCG-3′ (where underlining indicates the position of the nine base-pair target site). After the final round of selections, phagemids from each pool were sequenced as described below.

Sequencing of phage pools

After a series of selection and amplification cycles, phage pools from the final round of the selection were used to infect *E. coli* XL1-Blue (Stratagene) at a multiplicity of infection ⩽1. Individual clones were isolated and phagemid was purified from 5 ml cultures grown in 2 × YT (Qiagen). Phagemids were then sequenced by the Sanger dideoxy method using [α-^{35}S]dATP or by PCR using dye terminator chemistry (Biopolymers Laboratory, MIT Center for Cancer Research).

Dissociation constant determination

Dissociation constants were determined as described (Greisman & Pabo, 1997) with the following exceptions: (1) mobility shift assays were performed in 0.5 × TBE; (2) labeled oligonucleotides had the following sequences: (NRE_{ZF}) 5′-GATCCCCGTCAAGGGTTCAGTCCGGAATT; ($p53_{ZF}$) 5′-GATCCCCCTTGGGACATGTTCCTGGAATT; and (Zif268) 5′-GGCCGCGGGGCTATAGCGTGGGCGTACGAATT.

Acknowledgments

We thank J. Pomerantz, B. Wang, E. Rebar, and E. Fraenkel for their valuable advice and discussions. H.A.G. was supported by a predoctoral fellowship from the Howard Hughes Medical Institute and by the Massachusetts Institute of Technology fellowship from the Centocor Corporation; C.O.P., S.A.W. and E.I.R. were supported by the Howard Hughes Medical Institute.

References

Berg, J. M. & Shi, Y. (1996). The galvanization of biology: a growing appreciation for the roles of zinc. *Science*, **271**, 1081-1085.

Choo, Y. (1998). End effects in DNA recognition by zinc finger arrays. *Nucl. Acids Res.* **26**, 554-557.

Choo, Y. & Klug, A. (1994a). Selection of DNA binding sites for zinc fingers using rationally randomized DNA reveals coded interactions. *Proc. Natl Acad. Sci. USA*, **91**, 11168-11172.

Choo, Y. & Klug, A. (1994b). Toward a code for the interactions of zinc fingers with DNA: selection of randomized fingers displayed on phage. *Proc. Natl Acad. Sci. USA*, **91**, 11163-11167.

Choo, Y. & Klug, A. (1997). Physical basis of a protein-DNA recognition code. *Curr. Opin. Struct. Biol.* **7**, 117-125.

Choo, Y., Sanchez-Garcia, I. & Klug, A. (1994). *In vivo* repression by a site-specific DNA-binding protein designed against an oncogenic sequence. *Nature*, **372**, 642-645.

Christy, B. A., Lau, L. F. & Nathans, D. (1988). A gene activated in mouse 3T3 cells by serum growth factors encodes a protein with "zinc finger" sequences. *Proc. Natl Acad. Sci. USA*, **85**, 7857-7861.

Corbi, N., Perez, M., Maione, R. & Passananti, C. (1997). Synthesis of a new zinc finger peptide; comparison of its "code" deduced and "CASTing" derived binding sites. *FEBS Letters*, **417**, 71-74.

Desjarlais, J. R. & Berg, J. M. (1992a). Redesigning the DNA-binding specificity of a zinc finger protein: a data base-guided approach. *Proteins: Struct. Funct. Genet.* **12**, 101-104.

Desjarlais, J. R. & Berg, J. M. (1992b). Toward rules relating zinc finger protein sequences and DNA binding site preferences. *Proc. Natl Acad. Sci. USA*, **89**, 7345-7349.

Desjarlais, J. R. & Berg, J. M. (1993). Use of a zinc-finger consensus sequence framework and specificity rules to design specific DNA binding proteins. *Proc. Natl Acad. Sci. USA*, **90**, 2256-2260.

Desjarlais, J. R. & Berg, J. M. (1994). Length-encoded multiplex binding site determination: application to zinc finger proteins. *Proc. Natl Acad. Sci. USA*, **91**, 11099-11103.

Elrod-Erickson, M., Rould, M. A., Nekludova, L. & Pabo, C. O. (1996). Zif268 protein-DNA complex refined at 1.6 Å: a model system for understanding zinc finger-DNA interactions. *Structure*, **4**, 1171-1180.

Elrod-Erickson, M., Benson, T. E. & Pabo, C. O. (1998). High-resolution structures of variant Zif268-DNA complexes: implications for understanding zinc finger-DNA recognition. *Structure*, **6**, 451-464.

Fairall, L., Harrison, S. D., Travers, A. A. & Rhodes, D. (1992). Sequence-specific DNA binding by a two zinc-finger peptide from the *Drosophila melanogaster* Tramtrack protein. *J. Mol. Biol.* **226**, 349-366.

Fairall, L., Schwabe, J. W. R., Chapman, L., Finch, J. T. & Rhodes, D. (1993). The crystal structure of a two zinc-finger peptide reveals an extension to the rules for zinc finger/DNA recognition. *Nature*, **366**, 483-487.

Gogos, J. A., Hsu, T., Bolton, J. & Kafatos, F. C. (1992). Sequence discrimination by alternatively spliced isoforms of a DNA binding zinc finger domain. *Science*, **257**, 1951-1955.

Greisman, H. A. & Pabo, C. O. (1997). Sequential optimization strategy yields high-affinity zinc finger proteins for diverse DNA target sites. *Science*, **275**, 657-661.

Houbaviy, H. B., Usheva, A., Shenk, T. & Burley, S. K. (1996). Cocrystal structure of YY1 bound to the adeno-associated virus P5 initiator. *Proc. Natl Acad. Sci. USA*, **93**, 13577-13582.

Hyde-DeRuyscher, R. P., Jennings, E. & Shenk, T. (1995). DNA binding sites for the transcriptional activator/repressor YY1. *Nucl. Acids Res.* **23**, 4457-4465.

Jacobs, G. H. (1992). Determination of the base recognition positions of zinc fingers from sequence analysis. *EMBO J.* **11**, 4507-4517.

Jamieson, A. C., Kim, S. H. & Wells, J. A. (1994). *In vitro* selection of zinc fingers with altered DNA-binding specificity. *Biochemistry*, **33**, 5689-5695.

Jamieson, A. C., Wang, H. & Kim, S. H. (1996). A zinc finger directory for high-affinity DNA recognition. *Proc. Natl Acad. Sci. USA*, **93**, 12834-12839.

Kim, C. A. & Berg, J. M. (1995). Serine at position 2 in the DNA recognition helix of a Cys2-His2 zinc finger peptide is not, in general, responsible for base recognition. *J. Mol. Biol.* **252**, 1-5.

Kim, C. A. & Berg, J. M. (1996). A 2.2 Å resolution crystal structure of a designed zinc finger protein bound to DNA. *Nature Struct. Biol.* **3**, 940-945.

Nolte, R. T., Conlin, R. M., Harrison, S. C. & Brown, R. S. (1998). Differing roles for zinc fingers in DNA recognition: structure of a six-finger transcription factor IIIA complex. *Proc. Natl Acad. Sci. USA*, **95**, 2938-2943.

Pabo, C. & Sauer, R. T. (1992). Transcription factors: structural families and principles of DNA recognition. *Annu. Rev. Biochem.* **61**, 1053-1095.

Pavletich, N. P. & Pabo, C. O. (1991). Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 Å. *Science*, **252**, 809-817.

Pavletich, N. P. & Pabo, C. O. (1993). Crystal structure of a five-finger GLI-DNA complex: new perspectives on Zn fingers. *Science*, **261**, 1701-1707.

Rebar, E. J. & Pabo, C. O. (1994). Zinc finger phage: affinity selection of fingers with new DNA-binding specificities. *Science*, **263**, 671-673.

Swirnoff, A. H. & Milbrandt, J. (1995). DNA-binding specificity of NGFI-A and related zinc finger transcription factors. *Mol. Cell. Biol.* **15**, 2275-2287.

Wu, H., Yang, W. P. & Barbas, C. F. (1995). Building zinc fingers by selection: toward a therapeutic application. *Proc. Natl Acad. Sci. USA*, **92**, 344-348.

Wuttke, D. S., Foster, M. P., Case, D. A., Gottesfeld, J. M. & Wright, P. E. (1997). Solution structure of the first three zinc fingers of TFIIIA bound to the cognate DNA sequence: determinants of affinity and sequence specificity. *J. Mol. Biol.* **273**, 183-206.

Zweidler-Mckay, P. A., Grimes, H. L., Flubacher, M. M. & Tsichlis, P. N. (1996). Gfi-1 encodes a nuclear zinc finger protein that binds DNA and functions as a transcriptional repressor. *Mol. Cell. Biol.* **16**, 4024-4034.

Edited by P. E. Wright

(Received 7 August 1998; received in revised form 13 November 1998; accepted 17 November 1998)