Evaluation of a Modular Strategy for the Construction of Novel Polydactyl Zinc Finger DNA-Binding Proteins†

David J. Segal,*,‡ Roger R. Beerli,§ Pilar Blancafort, Birgit Dreier,|| Karin Effertz, Adrian Huber, Beate Koksch,⊥ Caren V. Lund, Laurent Magnenat, David Valente, and Carlos F. Barbas, III*

The Skaggs Institute for Chemical Biology and the Department of Molecular Biology, The Scripps Research Institute, La Jolla, California 92037

Received September 6, 2002; Revised Manuscript Received December 3, 2002

ABSTRACT: In previous studies, we have developed a technology for the rapid construction of novel DNA-binding proteins with the potential to recognize any unique site in a given genome. This technology relies on the modular assembly of modified zinc finger DNA-binding domains, each of which recognizes a three bp subsite of DNA. A complete set of 64 domains would provide comprehensive recognition of any desired DNA sequence, and new proteins could be assembled by any laboratory in a matter of hours. However, a critical parameter for this approach is the extent to which each domain functions as an independent, modular unit, without influence or dependence on its neighboring domains. We therefore examined the detailed binding behavior of several modularly assembled polydactyl zinc finger proteins. We first demonstrated that 80 modularly assembled 3-finger proteins can recognize their DNA target with very high specificity using a multitarget ELISA-based specificity assay. A more detailed analysis of DNA binding specificity for eight 3-finger proteins and two 6-finger proteins was performed using a target site selection assay. Results showed that the specificity of these proteins was as good or better than that of zinc finger proteins constructed using methods that allow for interdependency. In some cases, near perfect specificity was achieved. Complications due to target site overlap were found to be restricted to only one particular amino acid interaction (involving an aspartate in position 2 of the α-helix) that occurs in a minority of cases. As this is the first report of target site selection for designed, well characterized 6-finger proteins, unique insights are discussed concerning the relationship of protein length and specificity. These results have important implications for the design of proteins that can recognize extended DNA sequences, as well as provide insights into the general rules of recognition for naturally occurring zinc finger proteins.

In recent years, advances in the area of protein engineering and in our understanding of protein−DNA interactions have enabled the creation of novel DNA-binding proteins that are capable of recognizing virtually any desired DNA sequence (*1*−*3*). Such proteins have enabled the development of artificial transcription factors, which have been shown to up- or down-regulate a growing list of specific endogenous genes (*4*−*6*). Successful transgenic plants (*7*, *8*) and preclinical studies (*9*) have validated the utility of novel DNA-binding proteins to produce targeted gene regulators and therapeutics. New sequence-specific tools such as targeted endonucleases and integrases are nearing functional readiness (*10*, *11*).

The technology that has made these advances possible is based on the DNA-recognition properties of one particular class of DNA-binding domains, the Cys_2-His_2 zinc finger (Figure 1). This domain is the most common DNA-binding motif found in eukaryotes and is by far the most prevalent type of domain found in the human genome, with 4500 examples identified (*12*). Each 30-amino acid domain contains a single amphipathic α-helix stabilized by zinc ligation to two β-strands (Figure 1B). Sequence-specific recognition is provided by contact of amino acids of the N-terminal portion of the α-helix with base edges of predominantly one strand in the major grove of the DNA (Figure 1C). Among naturally occurring zinc finger domains, DNA interactions can be grouped as canonical and non-canonical types (*13*). Two examples of proteins with canonical type DNA recognition are the transcription factors Zif268 (*14*, *15*) and Sp1 (16). In these proteins, each domain recognizes essentially a three nucleotide subsite. Amino acids in positions −1, 3, and 6 (numbered with respect to the start of the α-helix) contact the 3′, middle, and 5′ nucleotides,

† This study was supported in part by National Institutes of Health grants AI41944, CA86258, and DK61803 and funding from the Torrey Mesa Research Institute to C.F.B. Postdoctoral fellowships were received by A.H. and L.M. from the Swiss National Science Foundation.

* Corresponding authors. C.F.B.: Address, The Scripps Research Institute, BCC-550, North Torrey Pines Road, La Jolla, CA 92037; phone, (858) 784-9098; fax, (858) 784-2583; e-mail, carlos@scripps.edu. D.J.S.: Address, Department of Pharmacology and Toxicology, University of Arizona, College of Pharmacy, 1703 E Mable Ave, Tucson, AZ 85721; phone, (520) 626-8782; fax, (520) 626-2466; e-mail, segal@pharmacy.arizona.edu.

‡ Current address: Department of Pharmacology and Toxicology, University of Arizona, Tucson, AZ 85721.

§ Current address: Cytos Biotechnology, Wagistrasse 25, CH-8952 Zurich-Schlieren, Switzerland.

|| Current address: Department of Biochemistry, 44K38, University Zurich-Irchel, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland.

⊥ Current address: Department of Organic Chemistry, University Leipzig, Johannisallee 29, 04103 Leipzig, Germany.

10.1021/bi026806o CCC: $25.00
Published on Web 01/28/2003

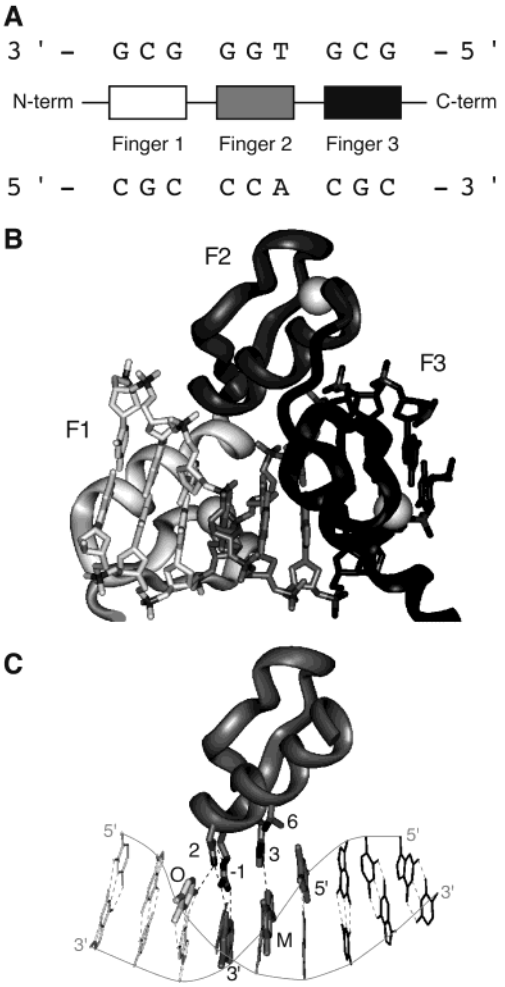


FIGURE 1: Representations of zinc finger−DNA interactions, based on the structure of Zif268 (14). (A) Diagram showing the antiparallel orientation of a 3-finger protein to its DNA target. The target sequence is shown as the top strand. (B) A structural representation of a 3-finger protein bound to nine bp of DNA. The protein and DNA are shaded as in (A). Zinc ions are shown as spheres. (C) The DNA-contacting residues of finger 2 and the bases typically contacted in the major grove. The residues are numbered (−1, 2, 3, 6) with respect to the α-helix. The 5′ ("5′"), middle ("M"), and 3′ ("3′") nucleotides that comprise the binding triplet for that domain are on one strand of the DNA. The nucleotide typically involved in target site overlap interactions ("O") is on the opposite strand.

respectively. Positions −2, 1, and 5 are often involved in direct or water-mediated contacts to the phosphate backbone. Position 4 is typically a leucine residue that packs in the hydrophobic core of the domain. Position 2 has been shown to interact with other helix residues and with bases depending on the protein and DNA sequences.

In previous work, we have used combinatorial mutagenesis and selection methods to modify the binding specificity of naturally occurring zinc finger domains (*17*−*19*). Starting with a canonical-type 3-finger protein, amino acids in positions −2 through 6 of the central domain were randomized. Proteins that could specifically recognize a new three-nucleotide subsite were selected by phage display, then optimized by site-directed mutagenesis. We have reported domains that bind with high affinity and specificity to the 16 members of the 5′-GNN-3′ set of DNA triplets and 14 of the 16 5′-ANN-3′ sequences. The selection of domains recognizing 5′-CNN-3′ and 5′-TNN-3′ sequences is in progress. These accomplishments bring us within reach of the ability to specifically recognize any of the 64 possible three-nucleotide subsites. Zinc finger domains are useful for the construction of new DNA-binding proteins because they are organized in tandem arrays, allowing recognition of extended, nonpalindromic DNA sequences. Consequently, we have assembled our optimized domains into 6-finger proteins, which have the theoretical capacity to recognize an 18-bp target site (*4*, *17*, *20*, *21*). A site of this length has the potential to be unique in the human genome, as well as all other known genomes. The published 5′(G/A)NN-3′ domains (*17*−*19*) allow for the rapid construction of more than one billion unique proteins, potentially capable of targeting one unique site for every 32 base pairs of DNA.

The zinc finger domains used to construct polydactyl proteins were initially selected and optimized as the finger-2 domain (F2) of a 3-finger protein (*17*−*19*). The binding specificity of each domain was determined in this "F2 context" using a stringent multitarget ELISA assay. One goal of the current study was to determine if the domains maintain their exquisite specificity when repositioned at the finger 1 or 3 positions and when they are incorporated into polydactyl 6-finger proteins. We also previously examined the potential of three different frameworks (the non-DNA-contacting regions of zinc finger domains) for arranging the domains into multifinger proteins (*20*). The F2 domains were linked in tandem (F2-backbone) or just the DNA-contacting residues of the domain were transplanted to the framework of the 3-finger proteins Zif268 or Sp1C (a consensus framework based on the Sp1 protein (*22*)). Proteins with an Sp1C-backbone were generally found to have a higher affinity than those with the other two. In a published example, the affinity of the 6-finger protein E2C improved 50-fold by displaying the same DNA-contacting residues in an Sp1C- rather than a F2-backbone (*20*). However, increased affinity often correlates with decreased specificity. Therefore, another goal of the current study was to investigate if the use of a F2, Zif, and Sp1C backbone affected specificity.

Finally, others in the field have observed that some domains in fact recognize a four-nucleotide subsite, with the fourth nucleotide overlapping the first nucleotide of the next site (*5*, *6*, *23*−*27*). This concern, referred to as target site overlap, would limit our ability to assemble our domains in any desired order. To address this concern, other groups have developed randomization and selection strategies in which two or more domains are modified simultaneously (*28*, *29*), or each domain is selected sequentially in the "context" of a previously selected domain (*30*). Construction of new DNA-binding proteins by these procedures is laborious because new and/or multiple randomized libraries must be screened for each DNA target sequence. In contrast, our approach enables the rapid construction of multidomain proteins but requires that each domain be modular and independent. Therefore, we were interested in examining the extent to which target site overlap affects domain modularity and binding specificity of polydactyl proteins assembled using our methodology.

Our studies of a large number of modularly assembled proteins demonstrates that our zinc finger domains generally maintain their specificity regardless of their new position. Effects due to target site overlap were evident but typically limited to predictable cases. In 3-finger proteins, specificity

was found to be as good or better than for proteins constructed by other methods. The recognition patterns of the 6-finger proteins were more complex. Potential explanations, such as framework restrictions and increased affinity, are discussed. Overall, these results validate our modular assembly strategy as a robust method for the generation of new high-affinity, site-specific DNA-binding proteins.

METHODS AND MATERIALS

Assembly of 3- and 6-Finger Proteins. Proteins were assembled from oligonucleotides using domain sequences and methods previously described. Genes for polydactyl proteins were cloned into a modified pMAL-c2 bacterial expression vector (New England Biolabs). Expressed proteins contained a maltose-binding protein (MBP) purification tag at the N-terminus and an hemophilus influenza hemaglutinin (HA) epitope tag at the C-terminus.

Multitarget Specificity Assays. These assays were performed as described (*19*). Essentially, freeze/thaw extracts containing the overexpressed maltose-binding protein zinc-finger fusion proteins were prepared from IPTG-induced cultures using the Protein Fusion and Purification System (New England Biolabs) in Zinc Buffer A (ZBA; 10 mM Tris, pH7.5/90 mM KCl/1 mM $MgCl_2$/90 μM $ZnCl_2$). Streptavidin (0.2 μg) was applied to a 96-well ELISA plate, followed by the indicated DNA targets (0.025 μg). Biotinylated hairpin oligonucleotides containing the indicated target sequences were immobilized on streptavidin-coated 96-well ELISA plates. Target hairpin oligonucleotides had the sequence 5′-Biotin-GGA$N^{1\prime}N^{1\prime}N^{1\prime}N^{2\prime}N^{2\prime}N^{2\prime}N^{3\prime}N^{3\prime}N^{3\prime}$GGG TTTT CCC $N^3N^3N^3N^2N^2N^2N^1N^1N^1$TCC-3′, where $N^1N^1N^1$ was the 3-nucleotide finger-1 target sequence and $N^{1\prime}N^{1\prime}N^{1\prime}$ its complement. The plates were blocked with ZBA/3% BSA. Eight 2-fold serial dilutions of the extracts were applied in 1× Binding Buffer (ZBA/1% BSA/5 mM DTT/0.12 μg/μl sheared herring sperm DNA), and bound protein was detected by mAb mouse antimaltose binding protein (Sigma) and mAb goat-antimouse IgG conjugated to alkaline phosphatase (Sigma). Alkaline phosphatase substrate (Sigma) was applied, and the OD_{405} was quantitated with SOFTmax 2.35 (Molecular Devices). All titration data were background subtracted from ELISA wells containing extract but no oligonucleotide.

CAST Assays. Fusion proteins were purified over amylose resin to >90% homogeneity using the Protein Fusion and Purification System (New England Biolabs) according to the manufacturer's recommendations, except that ZBA/5 mM DTT was used as the column buffer. Proteins were eluted with 10 mM maltose, concentrated, and stored in ZBA containing 50% glycerol/5 mM DTT at −20°C. Protein purity and concentration were determined from Coomassie blue-stained SDS−PAGE gels by comparison to BSA standards.

Randomized libraries of double-stranded DNA were created by PCR amplification of 150 pmole of a library oligonucleotide, 5′-GAGCTCATGGAAGTACCATAG−$(N)_{10,12,or21}$−GAACGTCGATCACTCGAG-3′, with the primers 5′-GAGCTCATGGAAGTACCATAG-3′ and 5′-CTCGAGTGATCGACGTTC-3′ (10 cycles; 15 s at 94 °C, 15 s at 70 °C, and 60 s at 72 °C). Libraries were trace labeled by inclusion of 10 μCi [α^{32}P]-dATP in the PCR reaction. Proteins were incubated with 1 pM DNA library in 1× Binding Buffer/10% glycerol for 1 h at room temperature, then separated on a 5% polyacrlyamide gel in 0.5× TBE buffer. Imaging of dried gels was performed using a PhosphorImager and ImageQuant software (Molecular Dynamics). The mobility of faint protein/DNA complexes was determined from positive controls in early rounds. Complexes were eluted from excised gel fragments in elution buffer (0.1% SDS/0.5M NH_3OAc/10mM MgOAc) overnight at 37 °C and then reamplified by 15 cycles of PCR as described above.

Protein concentration was approximately 1 or 0.1 μM (for 3- or 6-finger proteins, respectively) in the first round, then decreased in subsequent rounds as protein/DNA complexes became visible. CAST selections were repeated until 50% of the input library formed protein/DNA complexes (typically 5−12 rounds). For sequence determination, amplified DNA was cloned without restriction digest into pCR2.1-TOPO (Invitrogen) by topoisomerase-mediated ligation. Data for the 6-finger E2C(S) protein are a composite of two sets of oligonucleotides, one in which the first 9 bp (Half-Site 1, HS1) of the target site was fixed (12 bp randomized) and another in which HS2 was fixed (12 bp randomized). Data for the 6-finger Aart(S) protein are from one oligonucleotide pool with 21 bp randomized. Data for all 3-finger proteins were based on an oligonucleotide pool with 10 bp randomized.

RESULTS AND DISCUSSION

Multitarget ELISA Specificity Assays. To assess the validity of our modular approach, we first performed a cursory analysis on a large sample of proteins. Eighty 3-finger proteins were chosen randomly from the hundreds of multifinger proteins assembled in our laboratory. The proteins contained domains recognizing not only 5′-GNN-3′ type sequences but also 5′-ANN-3′ and 5′-TNN-3′ sequences. As a reference, the protein Zif268 was also included (Figure 2, #51). They were divided into eight sets of 10 proteins, and their relative affinity for the 10 DNA-target sites in their set was measured in a multitarget ELISA assay (Figure 2). The intention was to determine the extent to which proteins generated by the modular approach could bind their cognate (intended) target and to assess the specificity of that interaction.

The primary result was that all of the 80 proteins tested were able to bind their cognate target DNA. Most proteins also displayed excellent specificity for their cognate target, with little or no affinity for any of the other targets in the set. In only five cases (proteins 13, 19, 49, 67, and 76) did a protein bind a noncognate target with an affinity at or above 75% of the maximum binding signal. Protein 13 actually preferred binding targets 15 and 20 over its cognate target. There is no obvious explanation for why the five proteins showed increased affinity for some of the noncognate targets. An alignment of the bound cognate and noncognate target sites (not shown) often revealed a match of 5−6 bp between the 9-bp sites. However, such matches also exist between other targets for which there was no cross-reaction. More to the point, none of the proteins corresponding to the bound, noncognate targets cross-reacted with any other target in the set (that is, protein 76 bound target 73, but protein 73 did

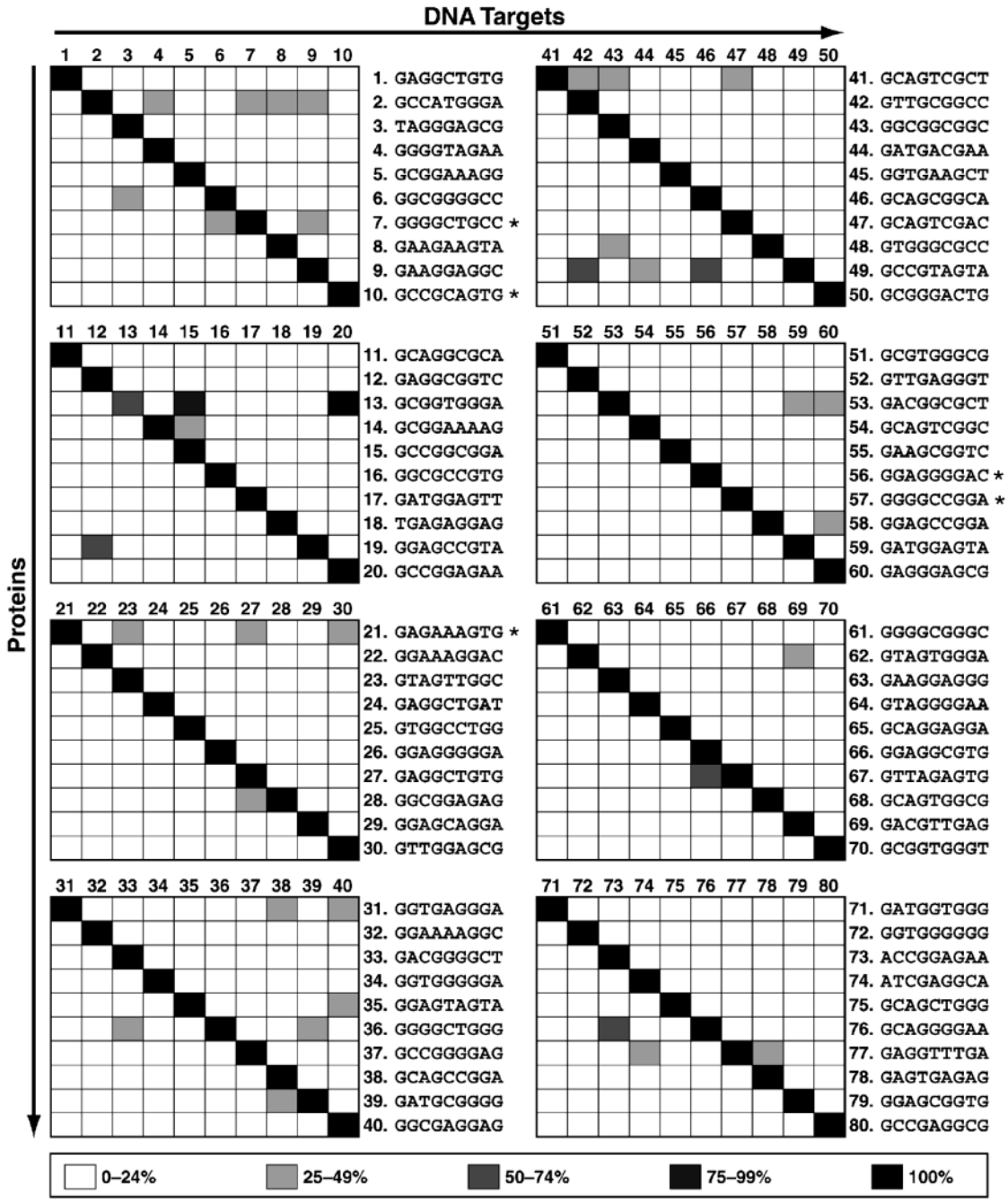
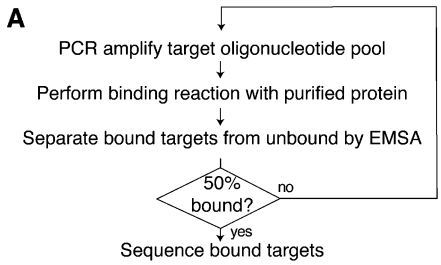


FIGURE 2: Specificity of 80 proteins based on the multitarget ELISA assay. Eight sets of 10 3-finger proteins were tested for binding to ten DNA targets. The numbered list to the right of each set correspond to both the intended recognition sequence of the proteins and the sequences of the DNA targets. Proteins used for CAST analysis are indicated by an asterisk (*). The maximum binding signal for each protein was normalized to be 100%. Shading indicates the normalized signal intensity according to the scale at the bottom. Experiments were performed in duplicates. The standard deviation of the measurements was typically less than 25% (not shown).

not bind target 76 nor any other noncognate target). From this it can be concluded that the observed promiscuity is a property of these particular proteins and not related to general factors such as the number of matches (within limits) or the number of guanines in the target sequences.

Target Site Selection Experiments. The multitarget ELISA specificity study found only 5 of 80 proteins (6.25%) to have extraordinary promiscuity, and only one (1.25%) to have inappropriate specificity. Although these results suggest that more than 90% of proteins created by the modular approach bind their cognate target with very high specificity, it should be noted that the 10 DNA targets in each set represent only 0.003% of all possible 9-bp targets. To provide a more detailed analysis of binding specificity, a cyclical amplification and selection of targets (CAST) assay was performed (*31*). CAST is a common and accurate method for determining the preferred binding site(s) for DNA-binding proteins and has been used to examine the specificity of naturally occurring zinc finger proteins such as Zif268 (32) and Sp1 (*33*−*35*), as well as several created by selection or design (*36*−*40*). In the current study, a cycle commenced with an in vitro binding reaction containing purified protein and a pool of randomized DNA targets (see Methods and Materials and Figure 3A). The bound targets were separated from unbound by a gel electrophoresis mobility shift assay (EMSA). The DNA targets had been designed with primer



B

```
atag GGA GGG GAC Agaa
atag GGA GGG GAC Ggaa
atag GGA GGG GAC Ggaa
atag GGA GGG GAC Ggaa
atag GGA GGG GAC Ggaa
atag GGA GGG GAT Cgaa
cata gGA GGG GAC ATga
tcAA GGA GGG GAc tatg
tcAG GGA GGG GAc tatg
tcAG GGA GGG GAc tatg
tcAG GGA GGG GAc tatg
tcGG GGA GGG GAc tatg
tcGG GGA GGG GAc tatg
tcGG GGA GGG GAc tatg
ttcA GGA GGG GAC ctat
```

FIGURE 3: Overview of the CAST assay. (A) A flow diagram describing the steps of the CAST assay. (B) Raw data from the CAST analysis of B3-HS2(S). Randomized regions are in capital letters, flanking regions are in lower case. Nucleotides not matching the expected target site are underlined.

sites flanking the randomized region, therefore allowing the bound targets to be amplified by PCR and used as input in subsequent cycles. CAST was performed for 5−12 cycles until 50% of the input DNA formed DNA/protein complexes, after which members of the pool were sequenced (as an example, Figure 3B). In general, the quality of the data improved only slightly with more rounds (data not shown).

CAST data were collected for 10 proteins, eight 3-finger, and two 6-finger proteins (Figure 4). The 6-finger protein E2C was assayed, as were the two 3-finger proteins used to construct it, E2C-HS1 and E2C-HS2 (20). For E2C-HS1, F2-, Zif-, and Sp1C-framework versions were analyzed (designated E2C-HS1(F2), (Z), and (S), respectively, in Figure 4). For all other proteins, only the Sp1C-backbone was used. The 6-finger Aart protein, composed of domains recognizing 5′-ANN-3′ and 5′-TNN-3′ type sequences (*17*), was also assayed. Although this protein had an affinity of 7.5 pM, its component 3-finger proteins had affinities below detection and were not analyzed. The remaining 3-finger proteins provide additional examples of domains that recognize 5′-GNN-3′ and 5′-ANN-3′ type sequences. Some domains appear in two or more proteins in different positions and contexts (i.e., different neighboring domains and DNA sequences).

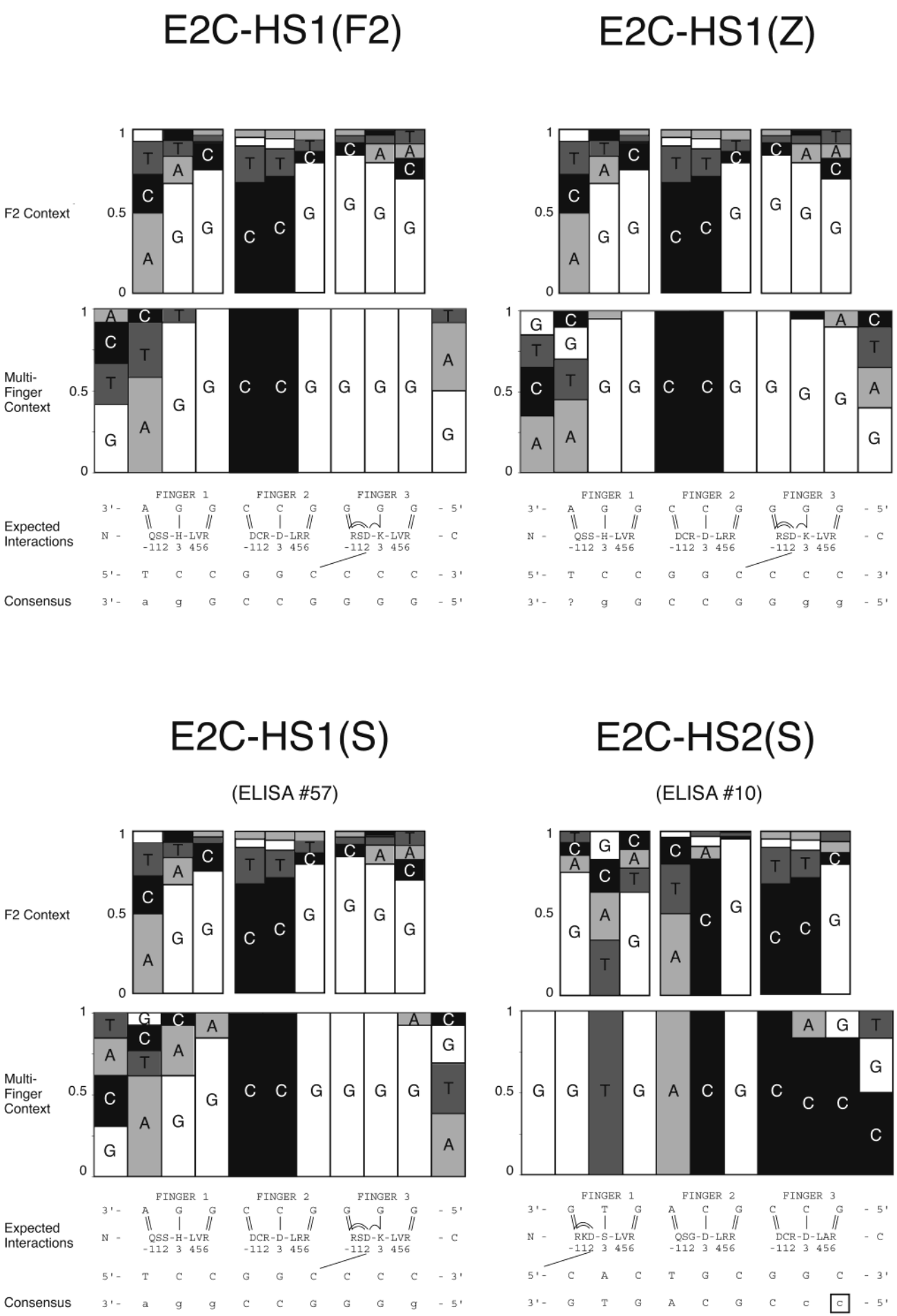
General Aspects of Specificity. Overall, the CAST analysis demonstrates that the modular approach can create proteins that bind with excellent specificity (Figure 4). This more detailed analysis fully supports conclusions of the broad-based multitarget ELISA study (Figure 2). The specificity of the 3-finger proteins tested here is as good or better than that of proteins produced by other methods such as sequential selection (*39*), bipartite library selection (*29*), zinc finger recognition codes (*36*, *37*, *41*), or other combinations of rational design and selection approaches (*40*). Specificity degenerates most frequently at the ends of the protein, consistent with observations by others (*42*). This is likely due to "breathing" between the terminal DNA-contacting residues and the ends of the oligonucleotide target. In some cases, such as HDII-HS2(S) and B3-HS1(S), only a single, terminal nucleotide was incorrectly specified in just one of the 10 or 15 target sequences recovered from CAST.

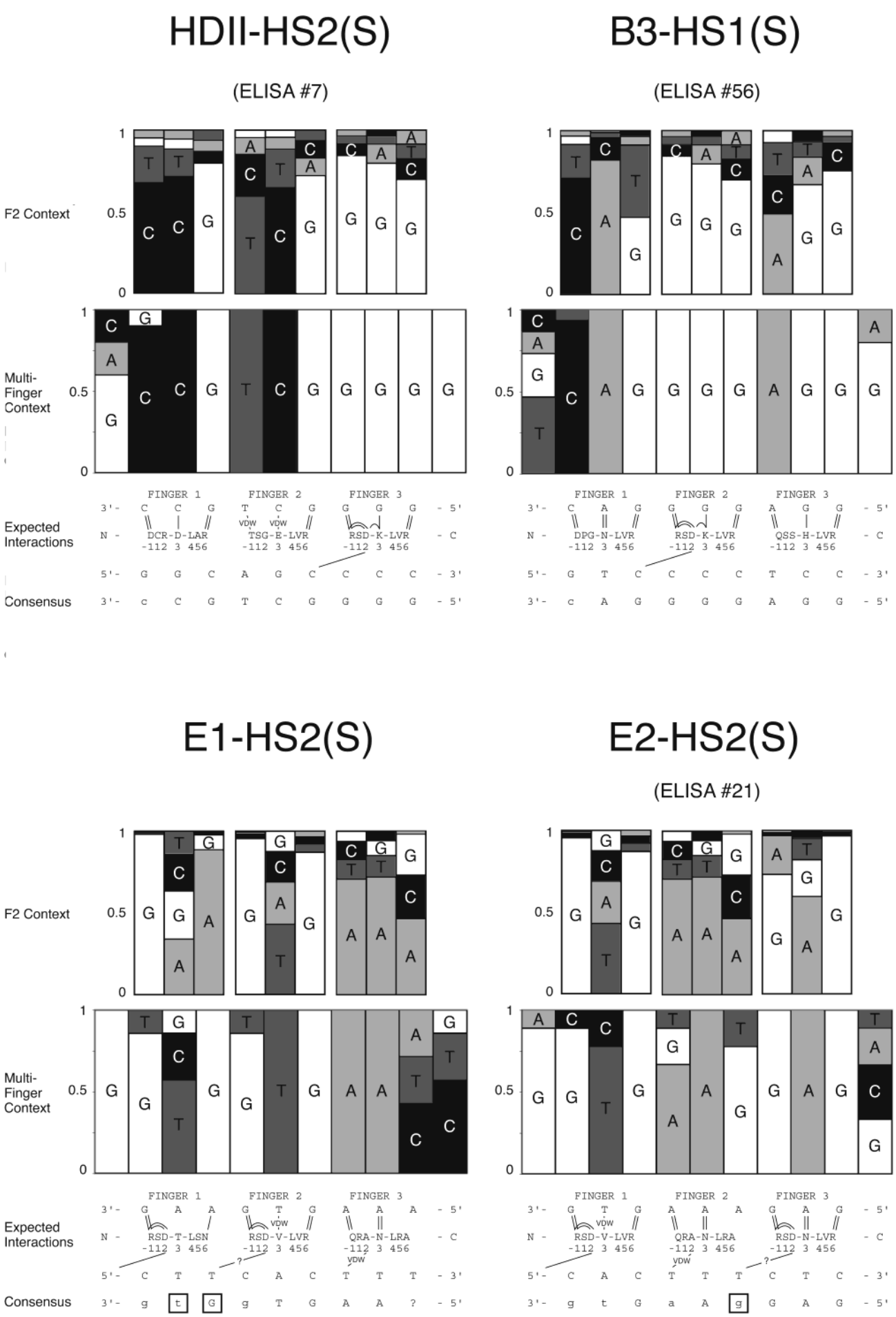
Other proteins displayed varying degrees of specificity. Examples can be found of poor specificity, nonspecificity, and even inappropriate specificity (denoted in the consensus sequence as lowercase letters, question marks, and boxes, respectively). In most cases the observed specificity can be understood in terms of the expected interactions (or lack of interaction) combined with a dominating target site overlap effect. Several exceptions are discussed below.

Target Site Overlap. Structural and biochemical analysis of the protein Zif268 found that aspartate in position 2 (Asp^2) of one α-helix can hydrogen bond to a nucleotide on the less-heavily contacted strand in the binding site of a neighboring domain (*14*, *23*, *26*). The hydrogen bond required an extracyclic amine group on the contacted nucleotide (either C or A), thereby influencing the 5′ nucleotide in the neighboring site to be G or T. This type of phenomenon, known as target site overlap, has led to the suggestion that zinc finger domains may more generally recognize a 4-bp site. Indeed, recent structural data demonstrate that some domains in canonical, Zif-backbone proteins can recognize a 4- or even 5-bp site (*25*). The implications suggest dire consequences for our modular approach based on a 3-bp site.

The CAST data generally support target site overlap by Asp^2. When Asp^2 occurs in the finger 1 position, as in E2C-HS2(S), E1-HS2(S), and E2-HS2(S), the neighboring nucleotide is specified as G. Interestingly, T was not specified. The overlap effect is less dramatic for the 6-finger proteins, but that may be due to increased breathing at the ends of the longer protein. Internally, the effects of Asp^2 can be seen in cases where the neighboring domain does a poor job of specifying its 5′ nucleotide. For example, Ala^6 in finger 2 of E2-HS2(S) was not expected to contact its 5′ nucleotide (*17*). Asp^2 in finger 3 specifies the nucleotide to be G or T. This domain previously demonstrated cross-reactivity to 5′ G (*17*), and the additional contact in the current context further enforces the cross-reaction. Similarly, Asn^6 in finger 1 of E1-HS2(S) was expected to contact N7 of either A or G (*17*). Asp^2 in finger 2 ensures specificity of G. The interactions in the 6-finger Aart(S) are less clear. Asp^2 in finger 6 seems to specify G or T in the finger-5 subsite, but the effect of Asp^2 in finger 5 is more ambiguous.

CAST data did not reveal strong evidence for target site overlap by an amino acid in position 2 other than Asp^2. Ser^2 (in finger 1 of the three E2C-HS1 proteins studied) and Gly^2 (in finger 1 of B3-HS1(S)) do not specify any particular neighboring nucleotide. G is partially specified as the neighboring nucleotide when Arg^2 appears in finger 1 of HDII-HS2(S); however, the neighboring nucleotide is mis-specified as A when Arg^2 appears in finger 3 of E2C(S). Similarly, A is strongly specified as the neighboring nucleotide when Ala^2 appears in finger 4 of Aart(S); however, the neighboring nucleotide is mis-specified as G when Ala^2 appears in finger 3 of Aart(S). Lys^2 in finger 2 of Aart(S) could potentially be responsible for the partial mis-specification of a neighboring C, but that would require further investigation.


E2C-HS1(F2)
E2C-HS1(Z)
E2C-HS1(S)
(ELISA #57)
E2C-HS2(S)
(ELISA #10)
F2 Context
Multi-Finger Context
Expected Interactions
Consensus
FINGER 1
FINGER 2
FINGER 3
QSS-H-LVR
DCR-D-LRR
RSD-K-LVR
RKD-S-LVR
QSG-D-LRR
DCR-D-LAR
-112 3 456


HDII-HS2(S)
(ELISA #7)
B3-HS1(S)
(ELISA #56)
E1-HS2(S)
E2-HS2(S)
(ELISA #21)
F2 Context
Multi-Finger Context
Expected Interactions
Consensus
FINGER 1
FINGER 2
FINGER 3
DCR-D-LAR
TSG-E-LVR
RSD-K-LVR
DPG-N-LVR
QSS-H-LVR
RSD-T-LSN
RSD-V-LVR
QRA-N-LRA
vdw

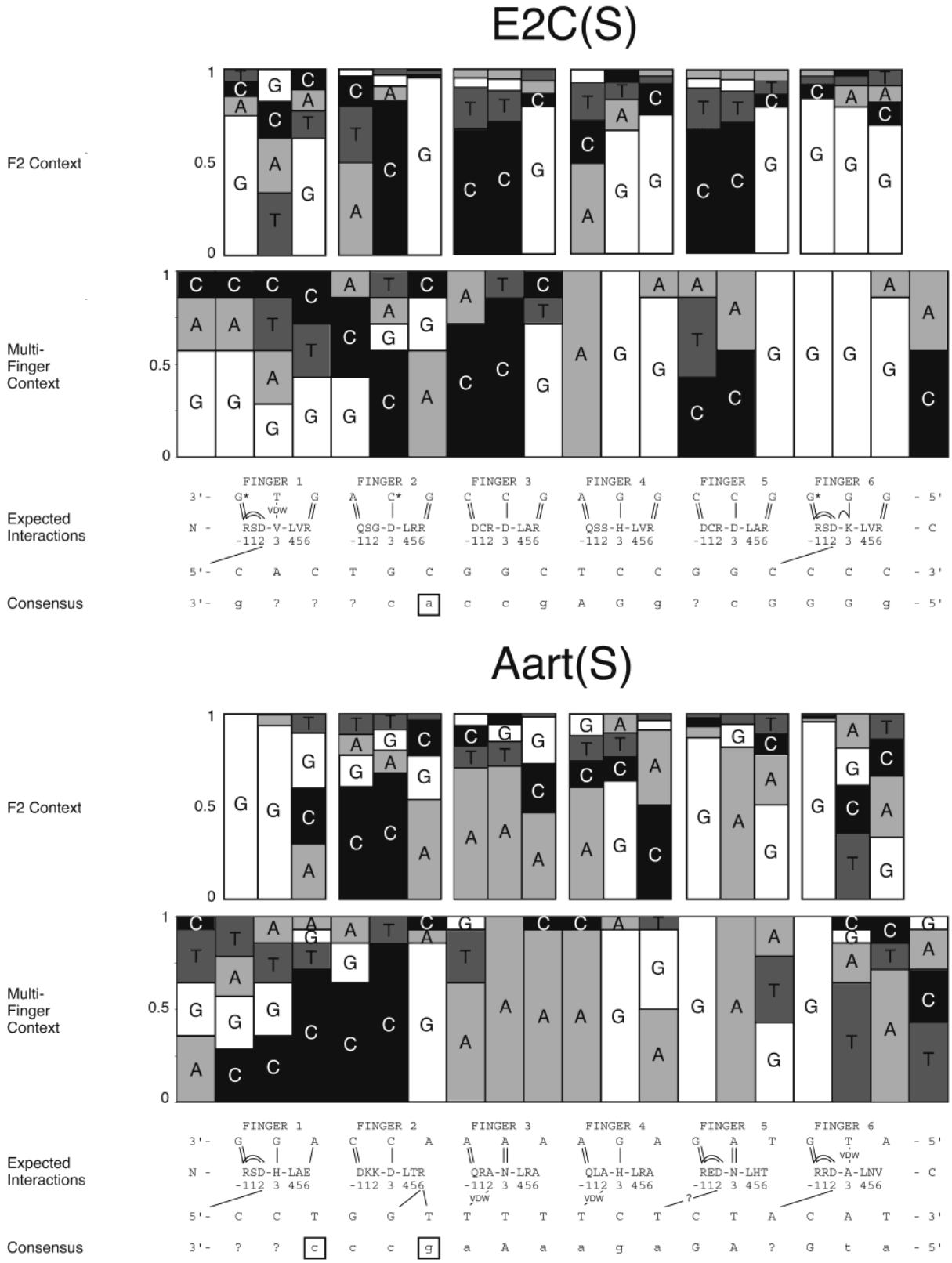


FIGURE 4: Results of the CAST assay. The name of the protein and a cross-reference (if available) to its position in the results of the multitarget ELISA specificity assay (Figure 2) are shown above each graph. Below the titles are bar graphs showing recalculated specificity data previously determined (*17*−*19*) when the domains were initially developed as finger 2 in a 3-finger protein (F2 context). The bars are shaded by nucleotide; their height represents the frequency with which each nucleotide was selected. Below the F2-context graphs are the CAST data of the domains assembled in multifinger proteins. Below this are the protein sequences, DNA target sequences, and expected interactions. Amino acids are numbered with respect to their position in the α-helix. The interactions are based on our previous computer models and analysis (*17*, *18*). Lines indicate expected hydrogen bonds. "VDW" indicates expected van der Waals interactions. "?" indicates an interaction that could potentially be destabilizing. The three asterisks next to nucleotides in the E2C(S) interactions indicate the positions that differ between the E2C and E3 binding sites (*4*). The consensus DNA-binding site is shown at the bottom. Capital letters indicate 100% conservation, lowercase letters indicate 50−99% conservation, and a question mark indicates less than 50% conservation. Boxes denote disagreement between the expected and observed nucleotides.

These results are consistent with other CAST studies. Ser^2 in finger 1 of the protein Sp1 failed to specify a neighboring nucleotide (*33*, *34*). Ser^2, which is present in 50% of all known zinc finger domains, has been shown to interact with all four nucleotides at the overlap position (*43*). A weak selection for G as the neighboring nucleotide was observed with His^2 in finger 1 of the sequentially selected protein NRE_{ZF}, but this preference was diminished when His^2 appeared in finger 1 of $p53_{ZF}$ (*39*). Thr^2 failed to specify a neighboring nucleotide in $TATA_{ZF}$ (*39*). Unlike Asp^2 in E1-HS2(S) of this study, Ala^2 did not dominate the neighboring Gln^{-1} recognition of 5′A in the code-derived protein Sint1 (36).

However, target site overlap is not only a consequence of the residue in position 2. Recent structural data suggest that the amino acid in position 1 can participate under some circumstances (*25*). In particular, Leu^1 in finger 3 of the sequentially selected $TATA_{ZF}$ was shown to interact with nucleotides on the opposite strand within the finger-2 triplet. Finger 2 contained an Ala^6, which did not contact any base in the structure (as expected) and therefore could not contribute to specificity. However, CAST analysis of this protein showed strong selection for a 5′ A in the finger-2 triplet, suggesting that the Leu^1 interactions from finger 3 were indeed specifying the base. It is intriguing to note that a similar situation exists in the case of finger 3 of Aart(S). Ala^6 of this domain is not expected to specify a 5′ nucleotide, and in fact none is specified when the domain appears as finger 3 of E2-HS2(S). However, 5′ A is strongly specified in the finger-3 triplet of Aart(S). Finger 4 of Aart(S) contains a Leu^1, which, by analogy to $TATA_{ZF}$, is likely to be responsible for the observed specificity. The caveat is that the two Leu^1-containing domains were created in different contexts. The entire recognition helix of finger 3 of $TATA_{ZF}$ was selected in a finger 3 context with A as the neighboring nucleotide, while finger 4 of Aart(S) was originally selected in a finger-2 context with G as the neighboring nucleotide. It is not clear how a Leu^1 selected in the latter context can so strongly specify A in the current context. Therefore, further studies will be required to determine if Leu or any other residue in position 1 is involved in a target site overlap interaction in the proteins described here.

As a whole, these results suggest that only target site overlap by Asp^2 presents an obstacle for modular construction. Asp^2 cannot be simply replaced in these domains. Aside from its undesired participation in target site overlap, Asp^2 forms buttressing contacts with Arg^{-1} that are thought to stabilize its orientation with respect to the DNA. Domains containing Arg^{-1} without Asp^2 display severely impaired specificity (*18*). However, it should be emphasized that Asp^2 appears in only 1/4 of all modular domains (those recognizing 5′-NNG-3′ sequences) and that complications are anticipated only when the neighboring nucleotide is A or C.

It should be noted that another recent study arrived at a contradictory conclusion, reporting biochemical evidence that Ser^2 is involved in target site overlap interactions (*44*). A potential explanation for this discrepancy may lie in the fact that the recognition helices examined here were displayed on the structurally regular Sp1C framework, while the other study investigated helices on finger 1 of the wild type Sp1 framework, which is known to interact with DNA differently than fingers 2 and 3. The structural differences underlying the two sets of observations would be insightful and deserve further study.

It is also interesting to note that in some instances a form of target site overlap appears to apply in the reverse direction. In particular, G was strongly specified 5′ to finger 3 of HDII-HS2(S) and finger 3 of B3-HS1(S). A similar interaction was described in the structure of the first three fingers of TFIIIA, in which a G 5′ to the finger-3 triplet is specified by an Arg in position 10 of the finger-3 helix (*27*). In our proteins, the residue at position 10 is always Thr, but at position 9 it is Arg. The C-terminal portion of the helix in finger 3 of TFIIIA is α-helical in nature, whereas this region in finger 3 of Zif268, Sp1, and our proteins is more likely to form a more compact 3_{10} helix (*45*). It is therefore possible that Arg^9 in our proteins could participate in a reverse target site overlap interaction to specify G. However, such a contact has not been reported in structural studies of Zif268 (*13*, *14*) or Sp1 (*46*), and none of the other proteins in the current study exhibit this behavior. Another explanation is that Arg^6, unsupported by an Asp^2-type buttressing interaction, could be free to interact with nucleotides 5′ to the binding site. However, Arg^6 also failed to specify a neighboring nucleotide in any other protein in the current study, and in E2C(S) there seems to be a weak preference for C. Two studies by other groups found that T was strongly selected as the 5′ neighbor to the finger-3 triplet of Zif268 (*32*, *39*). Finger 3 of this protein contains Arg^6 and Lys^9. In the protein NRE_{ZF}, there is a preference for G or A as the 5′ neighbor to the finger-3 triplet (finger 3 contains Ala^6 and Lys^9), and in $p53_{ZF}$ there is a weak preference for C (finger 3 contains Gln^6 and Lys^9) (*39*). CAST analysis of Sp1 has produced contradictory results on this issue (*34*, *35*). It is also possible that the nucleotides are conserved due to structural features of the DNA rather than a reverse target site overlap interaction from the protein. The basis for the apparent specificity remains unclear.

These studies further highlight the need for both structural and biochemical studies. Explanations for observed biochemical effects are weak without structural data, but structural studies alone are equally insufficient. For example, many structural studies have shown base contacts by Ser^2, but biochemical studies such as this one demonstrate that these contacts are not determinants of specificity. Claims that zinc finger domains specify a 4-bp, overlapping subsite have been largely exaggerated, due primarily to over-interpretation of too little or only one type of data.

Specificity as Modular Units. In general, the domains studied here maintained their original high specificity when placed in different positions in a new protein. The specificity data determined when the domain was created as finger 2 of a 3-finger protein (F2 context bar graphs in Figure 4) are excellent predictors of the specificity observed when that domain appears in a new polydactyl protein ("multifinger context" bar graphs). In several cases, the specificity in the new context was actually better, such as for the 5′-GTG-3′-recognition domains in finger 1 of E2C-HS2(S) and finger 2 of E1-HS2(S), the 5′-GGA-3′-recognition domain in finger 4 of E2C(S), and the 5′-ATG-3′-recognition domain in finger 6 of Aart(S). An interesting case where the specificity seems dependent on context is the 5′-GCC-3′-recognition domain. When this domain appears in finger 2 of E2C-HS1(S) it has perfect specificity, as it did in the original F2 context. In

both cases a target site overlap interaction aids, perhaps, in the specification of a 5′ G. When the domain appears in finger 3 of E2C-HS2(S), the specificity changed to 5′-CCC-3′. There is no target site overlap to aid the specification of 5′ G. However, structurally it is not clear why this would be necessary. There is also no expected target site overlap when the same domain appears in finger 3 of E2C(S); yet the specificity for 5′ G has been restored. Finally, the domain which had perfect specificity as finger 2 of the 3-finger E2C-HS1(S) has rather poor specificity as finger 5 of the 6-finger E2C(S). The structural basis for these observation is unclear. Possible explanations include context-dependent reorientation of the α-helix or increased sensitivity to differences in local DNA structure.

Another recent study involving analysis of zinc finger domains derived from rational design and selection, similar in many cases to those described here, also reported exceptionally specific recognition based on CAST analysis (*40*). The similarity of the domains used suggests that CAST analysis may generally produce a "cleaner" specificity profile than noniterative techniques such as the multitarget ELISA assay used in our earlier work (*17*−*19*). This caveat should be considered when interpreting the results from all such studies. More importantly, the other study demonstrated a clear positional dependence for many of the domains, a result in contrast to the findings reported here. However, the positional effects seemed to be restricted exclusively to finger 1 of their 3-finger constructs, which again may be a consequence of using a wild-type Sp1 framework. As noted above, finger 1 of Sp1 is known to interact with DNA differently than fingers 2 and 3. The resolution of this issue has important implications for the application of modular assembly and deserves further investigation.

5′-ANN-3′-recognition domains also maintained their original specificity well, but their performance was somewhat obscured by the fact that recognition of 5′ A is much less robust than for 5′ G. None of the various interactions that emerged from our previous study (*17*), small hydrophobics, Glu^6, Gln^6, or Arg^6, were able to stringently specify 5′ A in the current study. Consequently, specificity of this nucleotide can often be dominated by target site overlap interactions. In the absence of such interactions, results were confusing. Arg^6, which had been strongly selected to recognize 5′-ACN-3′ type sequences, reverted in finger 2 of Aart(S) to its more traditional role of specifying 5′ G. This came as somewhat of a surprise, since others had shown that the bases of a 5′-ACN-3′ triplet were correctly specified when Arg^6 appeared in finger 2 of the sequentially selected $p53_{ZF}$ (*39*). Gln^6, which had poor 5′ specificity originally, unexplainably specified 5′ C in finger 1 of Aart(S), while Ala^6, which also had poor specificity originally, was nonspecific in finger 3 of E1-HS2(S). However, more interesting than the failures are examples in fingers 3, 4, and 6 of Aart(S) where 5′ A was correctly specified. In all three cases, the position 6 residue was a small hydrophobic amino acid, which by computer modeling and structural analysis should be too far away from the DNA to influence specificity (*13*, *17*). Correct specification of 5′ A in the finger-3 triplet may be due to a target site overlap interaction as mentioned earlier. In the case of finger 4, 5′ A was partially specified despite a target site overlap interaction from finger 5 that was expected to specify either G or T. 5′ A was strongly specified in the finger-6 triplet in the absence of any potential target site overlap. It is therefore not at all clear what structural features are responsible for the observed specificity. Structural analysis is indicated.

Framework Effects and Higher-Order Proteins. The specificity of protein E2C-HS1 changed very little as the backbone was changed from F2, to Zif, to Sp1C. A much more dramatic change occurred when E2C-HS1(S) and E2C-HS2(S) were linked together as E2C(S). In particular, it is not clear why fingers 1 and 2, which displayed perfect specificity in E2C-HS2(S), displayed diminished specificity in E2C(S). E2C-HS2(S) and fingers 1−3 of E2C(S) are the same, thus ruling out influences from neighboring domains or differences in local DNA structure. One explanation is that the increased number of contacts in the 6-finger protein elevates the binding energy to a point where individual residue:base mismatches are insufficient to prevent binding. Alternatively, the fact that so many contacts are made to one strand of the DNA may "pull" the protein toward that strand and mis-orient some fingers.

A third explanation is that the DNA-contacting residues of the longer protein fail to align properly with the DNA bases. This phenomenon is supported by a growing consensus in the field and is attributed to the use of consensus TGEKP linkers between the domains. One consequence of the awkward alignment is that the protein exhibits lower affinity because binding energy is consumed contorting the DNA or simply lost due to missing DNA contacts. We originally discussed this concern when we reported the first studies of 6-finger proteins (*21*). Several subsequent studies have found that using longer linkers in various arrangements can produce proteins of higher affinity (*47*−*49*). Another logical consequence of framework-imposed misalignment could be the observed loss in specificity in the E2C(S) protein. However, since this work constitutes the first CAST analysis of a designed 6-finger protein, more research will be required to establish the relationship between framework constraints and specificity.

An interesting question raised by these results is whether the 6-finger proteins in this study can bind to more or less sequences than a 3-finger protein. A site for a 3-finger protein such as E2C-HS2(S), with near perfect specificity for its 9-bp site, should occur every 2.6×10^5 bp in a genome of random nucleotides ($[4 \times \{1 =$ the frequency of consensus nucleotide$\}]^9$), or around 13 000 times in the human genome (3.5×10^9 bp). In theory, an 18-bp site should occur once every 6.9×10^{10} bp ($[4 \times \{1\}]^{18}$), meaning that it would be unique in the human genome. However, the degenerate specificity of E2C(S) would lower this number to around one every 5.3×10^7 bp ($4^{18} \times \{0.57 \times 0.29 \times 0.43 \times 0.43 \times 0.57 \times 0.57 \times 0.71 \times 0.86 \times 0.71 \times 1 \times 1 \times 0.86 \times 0.43 \times 0.57 \times 1 \times 1 \times 1 \times 0.86\}$) or roughly 66 times in human. A consensus site for Aart(S) would occur around once per 1.2×10^8 bp ($4^{18} \times \{0.29 \times 0.36 \times 0.71 \times 0.64 \times 0.86 \times 0.86 \times 0.64 \times 1 \times 0.93 \times 0.93 \times 0.93 \times 0.50 \times 1 \times 1 \times 0.43 \times 1 \times 0.64 \times 0.70\}$) or 29 times in human. Therefore, the data support that these 6-finger proteins are still significantly more specific than an ideal 3-finger protein.

It should also be emphasized that the number of available binding sites in the genome will be somewhat lower than the theoretical total because many of the sites will be inaccessible due to structure chromatin. Furthermore, since

less than 1% of the human genome is coding region (*12*), most binding sites will occur in regions that will not affect the regulation of any gene. Previous studies have shown that only proteins that bind their target with an affinity of 10 nM or better are productive regulators. Therefore, even if a protein binds a site in a regulatory region that is related but nonconsensus, it may not have sufficient affinity to elicit a biological response.

In another study, we showed that E2C(S) can functionally discriminate in vivo at the level of endogenous gene regulation between its 18-bp cognate site in *erb*B-2 and another site, E3 in *erb*B-3, containing only three bp mismatches (*4*). In vitro, these three mismatches resulted in a 15-fold loss in affinity. The position of the mismatches are marked with asterisks on the expected interactions line of the E2C(S) CAST data (Figure 4). The discrimination can be rationalized in light of the CAST results; all mismatches correspond to nucleotides that are more than 50% conserved, one is 100% conserved. However, the CAST data also suggest that mismatches in other positions would affect specificity differently.

Zinc finger domains are the largest single class of identifiable folded domains in the human genome (4,500 examples identified), comprise the most common type of DNA-binding motif found in eukaryots, and represent the best characterized and simplest DNA-binding fold. Although there is considerable heterogeneity in the way naturally occurring zinc finger domains interact with DNA, many domains have been shown to interact in a manner similar to those used in this study. Therefore, the detailed analysis of these modified proteins should also contribute to our understanding of how this most important class of natural proteins recognizes DNA.

In conclusion, vast arrays of 3-finger proteins can be rapidly and reliably assembled from predetermined domains originally constructed in a F2 context. The 3-finger proteins constructed using this methodology generally recapitulate the specificity observed for each constituent domain. The robust domain specificity observed within 3-finger proteins weakens somewhat when two 3-finger proteins are directly linked. Even with some losses in domain specificity, the genomic targeting potential of 6-finger proteins is greatly improved over 3-finger proteins. The relationship between the longer proteins and specificity deserves further investigation. Since the loss of specificity clearly does not correlate with the original F2-context specificity of individual domains nor with the specificity of constituent 3-finger proteins, a higher order phenomenon must be responsible. Until better insight is obtained, our ability to predict in detail the specificity and affinity of 6-domain zinc finger proteins is limited. There would be cause for optimism if this framework explanation were proven true, for that would imply that specificity could be improved through further protein engineering. The alternative would be to accept that affinity and specificity are often opposing forces, and that one comes at the expense of the other.

REFERENCES

1. Beerli, R. R., and Barbas, C. F., III (2002) *Nat. Biotech. 20*, 135−141.
2. Segal, D., and Barbas, C. F., III (2001) *Curr. Opin. Biotech. 12*, 632−637.
3. Segal, D. J., and Barbas, C. F., III (2000) *Curr. Opin. Chem. Biol. 4*, 34−39.
4. Beerli, R. R., Dreier, B., and Barbas, C. F., III (2000) *Proc. Natl. Acad. Sci. U.S.A. 97*, 1495−1500.
5. Liu, P. Q., Rebar, E. J., Zhang, L., Liu, Q., Jamieson, A. C., Liang, Y., Qi, H., Li, P. X., Chen, B., and Mendel, M. C., et al. (2001) *J. Biol. Chem. 276*, 11323−11334.
6. Zhang, L., Spratt, S. K., Liu, Q., Johnstone, B., Qi, H., Raschke, E. E., Jamieson, A. C., Rebar, E. J., Wolffe, A. P., and Case, C. C. (2000) *J. Biol. Chem. 275*, 33850−33860.
7. Guan, X., Stege, J., Kim, M., Dahmani, Z., Fan, N., Heifetz, P., Barbas, C. F., III, and Briggs, S. P. (2002) *Proc. Natl. Acad. Sci. U.S.A. 99*, 13296−13301.
8. Ordiz, M. I., Barbas, C. F., III, and Beachy, R. N. (2002) *Proc. Natl. Acad. Sci. U.S.A. 99*, 13290−13295.
9. Xu, L., Zerby, D., Huang, Y., Ji, H., Nyanguile, O. F., de los Angeles, J. E., and Kadan, M. J. (2001) *Mol. Ther. 3*, 262−273.
10. Bibikova, M., Golic, M., Golic, K. G., and Carroll, D. (2002) *Genetics 161*, 1169−75.
11. Holmes-Son, M. L., Appa, R. S., and Chow, S. A. (2001) *Adv. Genet. 43*, 33−69.
12. Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., et al. (2001) *Science 291*, 1304−1351.
13. Pabo, C. O., and Nekludova, L. (2000) *J. Mol. Biol. 301*, 597−624.
14. Elrod-Erickson, M., Rould, M. A., Nekludova, L., and Pabo, C. O. (1996) *Structure 4*, 1171−1180.
15. Pavletich, N. P., and Pabo, C. O. (1991) *Science 252*, 809−817.
16. Narayan, V. A., Kriwacki, R. W., and Caradonna, J. P. (1997) *J. Biol. Chem. 272*, 7801−7809.
17. Dreier, B., Beerli, R. R., Segal, D. J., Flippin, J. D., and Barbas III, C. F. (2001) *J. Biol. Chem. 276*, 29466−29478.
18. Dreier, B., Segal, D. J., and Barbas, C. F., III (2000) *J. Mol. Biol. 303*, 489−502.
19. Segal, D. J., Dreier, B., Beerli, R. R., and Barbas, C. F., III (1999) *Proc. Natl. Acad. Sci. U S A 96*, 2758−2763.
20. Beerli, R. R., Segal, D. J., Dreier, B., and Barbas, C. F., III (1998) *Proc. Natl. Acad. Sci. U S A 95*, 14628−14633.
21. Liu, Q., Segal, D. J., Ghiara, J. B., and Barbas, C. F., III (1997) *Proc. Natl. Acad. Sci. U.S.A. 94*, 5525−5530.
22. Desjarlais, J. R., and Berg, J. M. (1993) *Proc. Natl. Acad. Sci. U.S.A. 90*, 2256−2260.
23. Isalan, M., Choo, Y., and Klug, A. (1997) *Proc. Natl. Acad. Sci. U.S.A. 94*, 5617−5621.
24. Jamieson, A. C., Wang, H., and Kim, S.-H. (1996) *Proc. Natl. Acad. Sci. U.S.A. 93*, 12834−12839.
25. Wolfe, S. A., Grant, R. A., Elrod-Erickson, M., and Pabo, C. O. (2001) *Struture 9*, 717−723.
26. Pabo, C. O., Peisach, E., and Grant, R. A. (2001) *Annu. Rev. Biochem. 70*, 313−340.
27. Wuttke, D. S., Foster, M. P., Case, D. A., Gottesfeld, J. M., and Wright, P. E. (1997) *J. Mol. Biol. 273*, 183−206.
28. Jamieson, A. C., Kim, S.-H., and Wells, J. A. (1994) *Biochemistry 33*, 5689−5695.
29. Isalan, M., Klug, A., and Choo, Y. (2001) *Nat. Biotechnol. 19*, 656−660.
30. Greisman, H. A., and Pabo, C. O. (1997) *Science 275*, 657−661.
31. Wright, W. E., Binder, M., and Funk, W. (1991) *Mol. Cell. Biol. 11*, 4104−4110.
32. Swirnoff, A. H., and Milbrandt, J. (1995) *Mol. Cell. Biol. 15*, 2275−2287.
33. Thiesen, H. J., and Bach, C. (1990) *Nucleic Acids Res. 18*, 3203−3209.
34. Shi, Y., and Berg, J. M. (1995) *Chem. Biol. 2*, 83−89.
35. Nagaoka, N., Shiraishi, Y., and Sugiura, Y. (2001) *Nucleic Acids Res. 29*, 4920−4929.
36. Corbi, N., Libri, V., Fanciulli, M., and Passananti, C. (1998) *Biochem. Biophys. Res. Commun. 253*, 686−692.
37. Corbi, N., Perez, M., Maione, R., and Passananti, C. (1997) *FEBS Lett. 417*, 71−74.
38. Desjarlais, J. R., and Berg, J. M. (1992) *Proteins: Struct., Funct., Genet. 12*, 101−104.
39. Wolfe, S. A., Greisman, H. A., Ramm, E. I., and Pabo, C. O. (1999) *J. Mol. Biol. 285*, 1917−1934.
40. Liu, Q., Xia, Z., Zhong, X., and Case, C. C. (2002) *J. Biol. Chem. 277*, 3850−3856.
41. Corbi, N., Libri, V., Fanciulli, M., Tinsley, J. M., Davies, K. E., and Passananti, C. (2000) *Gene Ther. 7*, 1076−1083.

42. Choo, Y. (1998) *Nucleic Acids Res. 26*, 554−557.
43. Kim, C. A., and Berg, J. M. (1995) *J. Mol. Biol. 252*, 1−5.
44. Nagaoka, M., Shiraishi, Y., Uno, Y., Nomura, W., and Sugiura, Y. (2002) *Biochemistry 41*, 8819−8825.
45. Laity, J. H., Dyson, H. J., and Wright, P. E. (2000) *J. Mol. Biol. 295*, 719−727.
46. Kim, C. A., and Berg, J. M. (1996) *Nat. Struct. Biol. 3*, 940−945.
47. Kim, J. S., and Pabo, C. O. (1998) *Proc. Natl. Acad. Sci. U.S.A. 95*, 2812−2817.
48. Moore, M., Klug, A., and Choo, Y. (2001) *Proc. Natl. Acad. Sci. U.S.A. 98*, 1437−1441.
49. Nagaoka, M., Nomura, W., Shiraishi, Y., and Sugiura, Y. (2001) *Biochem. Biophys. Res. Commun. 282*, 1001−1007.

BI026806O