Proc. Natl. Acad. Sci. USA
Vol. 90, pp. 2256–2260, March 1993
Biochemistry

Use of a zinc-finger consensus sequence framework and specificity rules to design specific DNA binding proteins

(database/modularity/prediction)

JOHN R. DESJARLAIS AND JEREMY M. BERG*

Thomas C. Jenkins Department of Biophysics, The Johns Hopkins University, Baltimore, MD 21218; and Department of Biophysics and Biophysical Chemistry, The Johns Hopkins University School of Medicine, Baltimore, MD 21205

Communicated by Thomas J. Kelly, Jr., December 8, 1992

ABSTRACT **We have designed three zinc-finger proteins with different DNA binding specificities. The design strategy combines a consensus zinc-finger framework sequence with previously characterized recognition regions such that the specificity of each protein is predictable. The first protein consists of three identical zinc fingers, each of which was expected to recognize the subsite GCG. This protein binds specifically to the sequence 5′-GCG-GCG-GCG-3′ with a dissociation constant of ≈11 μM. The second protein has three zinc fingers with different predicted preferred subsites. This protein binds to the predicted recognition site 5′-GGG-GCG-GCT-3′ with a dissociation constant of 2 nM. Furthermore, selection experiments indicate that this is the optimal binding site. A permuted version of the second protein was also constructed and shown to preferentially recognize the corresponding permuted site 5′-GGG-GCT-GCG-3′ over the nonpermuted site. These results indicate that earlier observations on the specificity of zinc fingers can be extended to generalized zinc-finger structures and realize the use of zinc fingers for the design of site-specific DNA binding proteins. This consensus-based design system provides a useful model system with which to study details of zinc-finger–DNA specificity.**

The prospect of designing functional proteins based on naturally occurring structural motifs has shown promise in recent years (1). The Cys_2His_2 zinc-finger motif represents a particularly attractive motif for design studies, as it is well-characterized structurally and has distinct metal binding and DNA binding properties (2–4). The large amount of sequence information available for the zinc-finger motif greatly facilitates and encourages a sequence-data-base-guided design of zinc-finger structure and DNA binding specificity. The DNA binding specificity of zinc-finger proteins has begun to be explored, revealing the possible existence of rules for specificity (4–10). Several of the change-of-specificity mutants were designed by following zinc-finger-sequence data-base distributions (6–8), demonstrating that the zinc-finger sequence data base is a useful guide for zinc-finger engineering. Structure determination (4) and mutagenesis studies (5–10) suggest that a region of only seven residues plays the dominant role in determining DNA binding specificity, whereas the remaining sequence is important for zinc binding, structural integrity, and nonspecific DNA contacts. In designing zinc-finger proteins, we hoped to create a model system for zinc fingers in which all of the zinc fingers were identical in sequence except for changes sufficient to confer different DNA binding specificities. If binding site specificities are transferable from one protein to another and if individual zinc-finger domains truly act as independent modules, then it should be possible to use this system to design proteins with predetermined binding-site preferences.

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

Our strategy for zinc-finger protein design is to combine a zinc-finger-framework sequence, which consists of a consensus sequence derived from 131 zinc-finger sequences, with specificity rules derived previously from native and mutant versions of Sp1 zinc fingers (6–8). Thus, each designed zinc finger is identical in sequence except for changes in one to four residues in its recognition region, which spans seven residues. The individual designed zinc fingers that have predicted subsites of three contiguous base pairs are then combined to yield proteins of three zinc fingers each, such that their predicted DNA recognition sites are the combinations of subsites determined by the order of zinc fingers in each protein.

MATERIALS AND METHODS

Construction of Genes and Protein Expression. Genes encoding the proteins were constructed as described in Fig. 2 and then subcloned into the expression vector pKK223-3 (Pharmacia) and/or into a vector containing a promoter for T7 RNA polymerase (11). Proteins were expressed in 71-18 *Escherichia coli* cells, when using the pKK223-3 vector, or in BL21 (DE3) pLysS cells, when using the T7 vector (12), and partially purified as described (6), starting with 250 ml of *E. coli* culture. Proteins expressed and purified from either system gave similar results. Further purification was achieved by KCl step elution of the protein from a heparin-Sepharose column (Pharmacia). When necessary, protein was concentrated with a Centricon-10 concentrator (Amicon). Protein was quantitated using an A_{275} value of 2800 $M^{-1}\cdot cm^{-1}$.

Quantitative DNase I Footprints. Footprinting experiments were performed essentially as described (13), except that data were quantitated and analyzed on a PhosphorImager (Molecular Dynamics, Sunnyvale, CA). Binding buffer was 31.5 mM Tris, pH 8.0/100 mM KCl/62.5 mM $ZnCl_2$/5 mM $MgCl_2$/1 mM $CaCl_2$/5 mM dithiothreitol/bovine serum albumin 50 μg/ml/poly(dI-dC)(1 μg/ml). The binding curves were fitted to the equation $f = a\{K_d/(K_d + [P])\} + b$ with the program KALEIDAGRAPH (Synergy Software, Reading, PA), where f is the fractional saturation, [P] is the protein concentration, and a and b are constants. We estimate the standard errors for the dissociation constants to be 50%.

Gel-Mobility-Shift Selection of Binding Sites from a Partially Biased Pool. We created a partially biased pool of randomized sequences by synthesizing oligonucleotides containing the sequence 5′-GCG-GCG-GCG-3′ (which differs from the predicted site of the QDR-RER-RHR protein by two bases) but with a 13% frequency of each other base at each of these nine positions (14). This oligonucleotide pool was PCR-amplified and labeled for selection by gel mobility shift, performed as described (6). Four rounds of gel shift selection were performed with PCR amplification after each round. The final

*To whom reprint requests should be addressed.

product was digested with appropriate restriction enzymes and subcloned, and individual clones were sequenced. The sequences were aligned by eye. Several observations suggest that the initial bias of sites did not significantly influence the final selected population. (*i*) No sequences identical to the basis sequence GCG-GCG-GCG were selected. (*ii*) Several of the sites are shifted in-frame from the basis sequence and one site was found on the opposite strand. (*iii*) Five of the 14 sequences are calculated to have concentrations in the starting pool equal to or less than that expected for a nonamer in a completely random population. As estimated visually from the fraction of DNA bound at each round, the apparent affinity of the protein for the pool was increased at least 100-fold from the starting round to the final round. Four sites were found that do not resemble the consensus sequence. Gel mobility shifts of the individual cloned sites (data not shown) showed that these sites do not show tight binding and were not used to generate the consensus sequence or in the histogram.

RESULTS AND DISCUSSION

A Consensus Sequence Framework for Zinc-Finger Protein Design. One hundred thirty-one zinc-finger sequences from an early zinc-finger sequence data base were aligned to yield a consensus zinc-finger sequence in which each residue in the consensus was the most frequent residue for that position (15). Most of the amino acid sequence of each zinc-finger domain described herein is derived from this consensus sequence. The only differences from the consensus are in the helical regions involved in the specific recognition of DNA. This approach was taken as the basis for design for several reasons. (*i*) The use of zinc fingers with identical frameworks allows for isolation of the properties necessary for specific DNA binding from other structural features. This approach should produce proteins with the expected DNA binding properties only if the activities of the recognition residues are transferable between domains without major contributions from other sequence features. (*ii*) The single consensus zinc-finger peptide has been shown to form a very stable zinc-finger structure (15). (*iii*) The sequence includes features involved in non-specific interactions with DNA and the frequently occurring Thr-Gly-Glu-Lys-Pro linker sequence. (*iv*) The extent of conservation of residues in the consensus sequence is fairly high in most of the sequence but clearly lower for residues involved in specific DNA recognition.

A Consensus-Based Protein with Three Identical Zinc Fingers. As a test of the design strategy, we first chose to design a protein containing three identical zinc fingers. The zinc-finger recognition region to be used in this first design was chosen such that it was likely to possess both reasonable affinity and predictable specificity. A particularly well-studied set of recognition residues is that present in the first and third zinc fingers of Zif268 (4, 5) and in the middle zinc finger of Sp1 (6–8, 16–18). In each case, residues Arg^{13}, Glu^{16}, and Arg^{19} are directly involved in DNA recognition, with Asp^{15} also being necessary for an important side-chain interaction with Arg^{13} (see Fig. 1 for numbering). This common set of residues recognizes the DNA subsite 5′-GCG-3′ in all cases studied. The sequence of the designed three-zinc-finger protein named RER-RER-RER for the recognition residues Arg^{13}, Glu^{16}, and Arg^{19} of each finger is shown in Fig. 1*A*.

A

H_2N-M E K L R N G S G D P G K K -

Q H A C P E C G K S F S R S D E L Q R H Q R T H T G E K -

P Y K C P E C G K S F S R S D E L Q R H Q R T H T G E K -

P Y K C P E C G K S F S R S D E L Q R H Q R T H Q N K K-COOH

13 16 19

Predicted binding site: 5'-GCG-GCG-GCG-3'

COOH- RER-RER-RER-NH_2

B

H_2N-M E K L R N G S G D P G K K -

Q H A C P E C G K S F S Q S S D L Q R H Q R T H T G E K -

P Y K C P E C G K S F S R S D E L Q R H Q R T H T G E K -

P Y K C P E C G K S F S R S D H L S R H Q R T H Q N K K-COOH

13 16 19

Predicted binding site: 5'-GGG-GCG-GCT-3'

COOH- RHR-RER-RDQ-NH_2

FIG. 1. Amino acid sequences of two designed zinc-finger proteins. Each protein contains three zinc fingers whose sequence is derived predominantly from a consensus of 131 zinc finger sequences (15). Residues important for specific DNA recognition or residues that vary from finger to finger are underlined. The sequences preceding the first cysteine zinc ligand and following the last histidine zinc ligand in each protein are taken from basic regions flanking the three zinc fingers of Sp1 (16). Shown below each protein sequence is its predicted binding site and the orientation of recognition residues relative to bases within the site. (*A*) Amino acid sequence of the protein RER-RER-RER, named for the residues in contact positions 13, 16, and 19 of each finger. (*B*) Amino acid sequence of the similarly named protein QDR-RER-RHR.

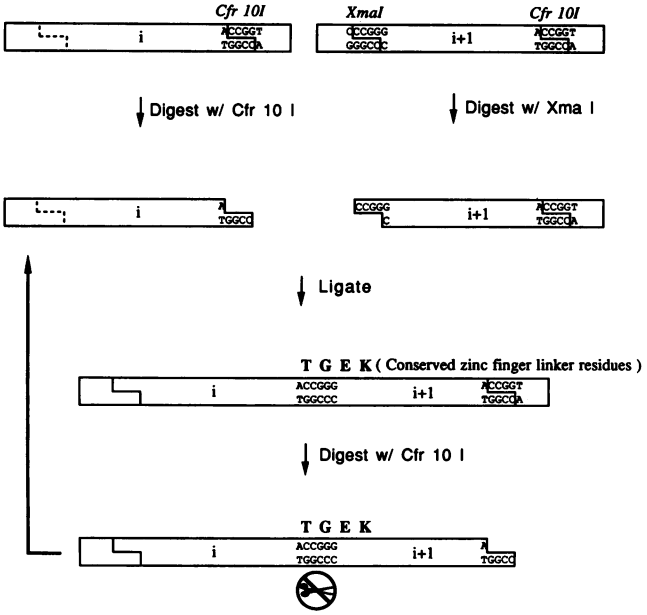


FIG. 2. General scheme for gene construction of tandem zinc-finger arrays. This stepwise subcloning approach makes use of *Xma* I and *Cfr*10I restriction sites which, when ligated together, yield a hybrid site that is no longer cleavable by either enzyme. The hybrid site ACCGGG encodes for the conserved Thr-Gly sequence in the zinc-finger linker region. Because the hybrid site is not cleavable, the ligated fragment is suitable for another round of zinc-finger addition, as indicated by the arrow. The genes for the proteins described in this paper were constructed using this scheme, except that the gene fragments encoding the first and last fingers contained additional restriction sites for subcloning into an expression vector.

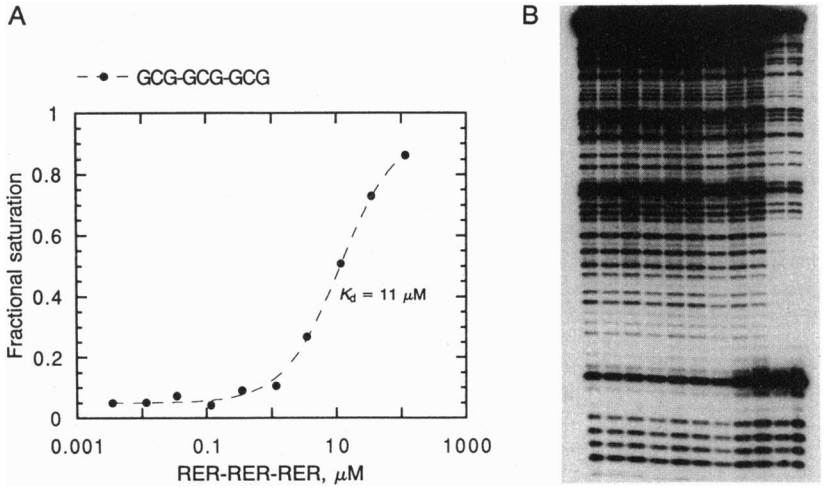


FIG. 3. Specific DNA binding activity of the protein RER-RER-RER to the predicted site 5′-GCG-GCG-GCG-3′, as determined by quantitative DNase I footprinting (19). The equilibrium binding isotherm for binding of RER-RER-RER to the predicted DNA site 5′-GCG-GCG-GCG-3′ within a 79-bp probe demonstrates concentration-dependent specific binding to the protected region. The data shown in *B* fit to an isotherm with a dissociation constant of 11 μM (*A*).

To produce such proteins, a strategy that enables one to build up a gene encoding an array of zinc-finger domains in a stepwise manner was devised and is shown in Fig. 2. A gene for the RER-RER-RER protein was synthesized in this manner and the protein was expressed in *E. coli*. The purified protein binds specifically to the predicted site 5′-GCG-GCG-GCG-3′ as demonstrated by DNase I footprinting experiments shown in Fig. 3. No binding to the Sp1 binding site 5′-GGG-GCG-GGG-3′ was detected under equivalent conditions, even though this site differs from the predicted site in only two of nine positions. We have also compared binding of the protein to its predicted site versus a partially randomized pool of G+C-rich sequences in which the concentration of the predicted site is 1% of the total. Gel mobility shifts (not shown) reveal that the affinity for the predicted site is at least 50-fold higher than the apparent affinity for the pool, indicating a significant level of specificity. DNase I footprinting titrations with RER-RER-RER on its predicted site yielded an equilibrium binding isotherm that could be fit with a dissociation constant of 11 μM, as shown in Fig. 3. Similar experiments with a three-zinc-finger peptide derived directly from Sp1 (6), assayed with the Sp1 binding site revealed a dissociation constant of 0.3 μM for that complex (data not shown). Thus, the designed protein appears to bind with reasonable specificity and affinity to its predicted binding site.

Consensus-Based Proteins with Three Different Zinc Fingers. The second designed protein represents a more dramatic effort to utilize zinc fingers as building blocks, as it is made up of three zinc fingers with different predicted preferred subsites. This protein is similar to the first in that its sequence is predominantly consensus derived and, as such, provides a more stringent test for the context independence of the specificity determining residues. We have previously shown that the DNA binding specificity of an Sp1-based three-zinc-finger peptide can be altered systematically by making a small set of changes, chosen on the basis of pairwise data-base distributions, in the recognition region of the second zinc finger of Sp1 (6–8). By changing some of the residues involved in DNA recognition from Arg^{13}, Glu^{16}, Arg^{19} to Arg^{13}, His^{16}, Arg^{19} or Gln^{13}, Asp^{16}, Arg^{19} (see Fig. 1 for numbering), we could change the specificity of these zinc fingers from GCG to GGG or GCT, respectively. Note, Asp^{15} was changed to Ser^{15} when changing Arg^{13} to Gln^{13}, as described (6). The orientation of recognition residues in each finger is antiparallel to the 5′ → 3′ orientation of bases, such that residue 19 recognizes the first base of a triplet and residue 13 recognizes the third base of a triplet. We designed and produced a second consensus-based protein that incorporates these recognition regions. This protein, termed QDR-RER-RHR, has the sequence shown in Fig. 1*B*. Since interfinger effects may be an additional factor in DNA binding, the order of zinc fingers in this protein was chosen such that all of the junctions resulted in adjacent Arg^{19}—Arg^{13} interfinger residues, as this situation is more frequently observed in our data base than are Arg^{19}—Gln^{13} junctions. We predicted that this protein would bind optimally to the DNA sequence 5′-GGG-GCG-GCT-3′, based on specificities observed in other protein contexts and on the antiparallel orientation of the recognition residues with respect to the G-rich strand of

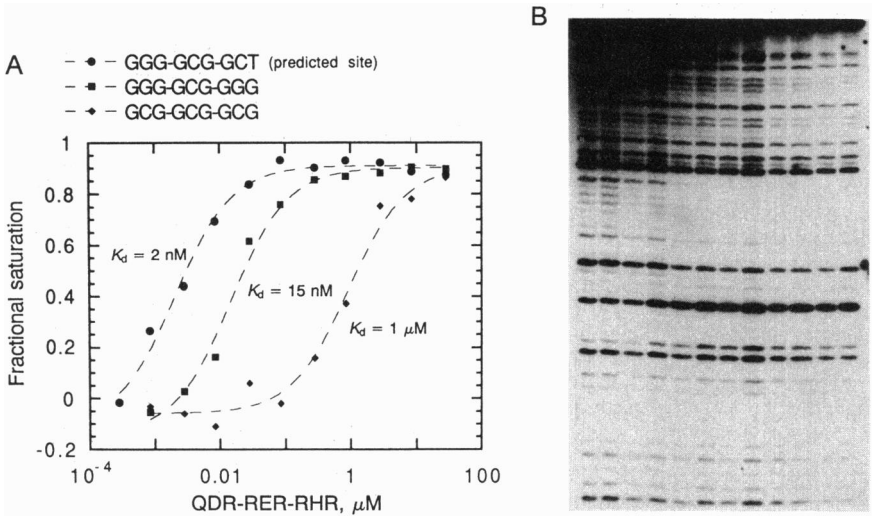


FIG. 4. (*A*) Equilibrium binding isotherms for the interaction of the designed protein QDR-RER-RHR with three related binding sites. The calculated dissociation constants for the fitted curves are shown for each titration. (*B*) The actual footprinting data for the predicted site, each lane corresponding to a point on the curve with the first lane as a zero-protein reference. The isotherms show that the protein binds with highest affinity to its predicted site.

A

```
ATAT GGGGCGGCT TACCG
 ATA GGGGTGGAG TTACAG
 ATA GGGGCGGCG TTACAG
 ATA GGGGTGGCA TTACAG
ATAT GGGGCGGCT TACTG
 GTT CAGGCGGCT GCAA
 ATA GGGGAGGCT TTACAG
 ATA GGGGTGGCC TTACAG
 ATA GGGGCGGCA TTACAG
 ATA GAGGCGGCG TTACAG
ATAG AGGGTGGCT TACAG
 AAA GGGGAGGCT ATGAT
 ATA GCGGAGGCG TCTGAA
ATAC GGGGCGGCT TACAG

Consensus: GGGGCGGCT
```

B

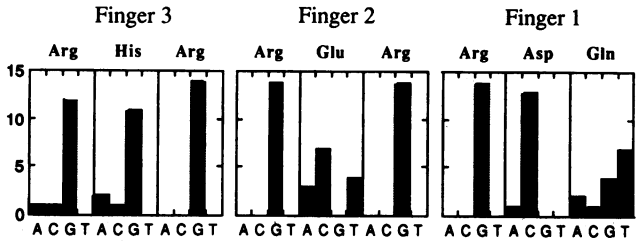
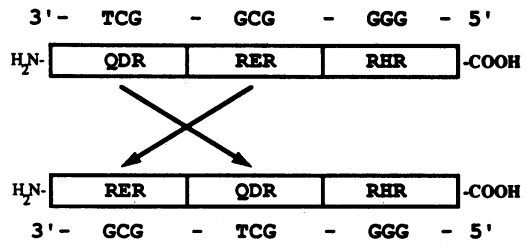


FIG. 5. Results from a binding-site-refinement experiment for the QDR-RER-RHR protein. (*A*) The consensus of the aligned sequences, which agrees perfectly with the predicted site, is indicated below. Several sequences identical to the predicted site are underlined. Some flanking sequence is shown for each clone to reflect the shift in-frame of several sites. (*B*) A histogram of base frequencies for each of the nine positions of the aligned binding sites. The amino acid presumed to recognize each base is shown above the four-base histogram for each position. The distributions of specificities seen here are consistent with previously reported measurements of the specificities of the same recognition residues in other contexts (4–9, 16–18).

the binding site. Footprinting experiments with QDR-RER-RHR on this site and on the sites 5′-GGG-GCG-GGG-3′ and 5′-GCG-GCG-GCG-3′ reveal binding isotherms with dissociation constants of 2 nM, 15 nM, and 900 nM, respectively, as shown in Fig. 4. These results indicate that the QDR-RER-RHR protein binds to the predicted site with high affinity and that this high-affinity binding allows detectable, albeit weaker, binding to other sites. The relative affinities for the three sites with the protein QDR-RER-RHR are reasonably consistent with those predictable from earlier work with Sp1 variants.

To determine the optimal site for the protein QDR-RER-RHR, we performed a binding-site-refinement experiment by gel-mobility-shift selection of high-affinity binding sites, starting with a partially randomized pool of DNA probes (14). The advantages of starting with the biased population are an ability to use the previously determined affinities for sites to determine a useful starting protein concentration for getting a desired degree of selection and a reduction in the number of rounds needed for near complete enrichment of sites. The results from the selection are shown in Fig. 5*A*. The aligned sequences reveal a consensus sequence 5′-GGG-GCG-GCT-3′, which agrees perfectly with the predicted site. The selected sequences are consistent with previous results on the specificity of individual recognition residues as seen in the histogram of base frequencies for each position, shown in Fig. 5*B*: all arginines show a striking specificity for guanine (4, 6–9, 18); the glutamic acid in the middle finger, recognizing the central base of the middle subsite, shows relatively

A



B

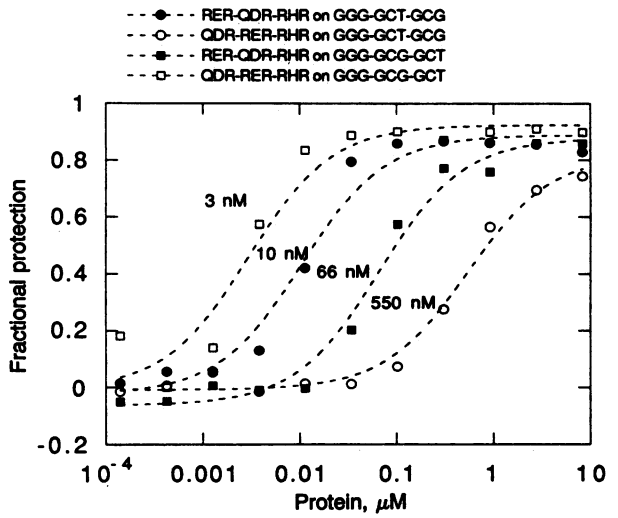


FIG. 6. Permutation of the first two zinc fingers of the QDR-RER-RHR protein and demonstration of modular specificity. (*A*) We performed a finger-permutation experiment in which the positions of the first two zinc fingers of the protein QDR-RER-RHR were exchanged to yield the protein RER-QDR-RHR. As indicated by the DNA sites shown along with each protein, a corresponding permutation of subsites recognized by each finger predicts a change of specificity from 5′-GGG-GCG-GCT-3′ to 5′-GGG-GCT-GCG-3′. (*B*) DNase I footprinting experiments performed with the two proteins and the two binding sites. Open symbols, titrations done with the protein QDR-RER-RHR; solid symbols, titrations done with the permuted protein RER-QDR-RHR; squares, the DNA site GGG-GCG-GCT; circles, the site GGG-GCT-GCG. The calculated dissociation constants are shown next to each fitted curve. The 3 nM dissociation constant shown for the QDR-RER-RHR protein on its preferred site is consistent, within experimental error, with the 2 nM constant reported for the same interaction in Fig. 4. These results show that each protein has a significantly higher affinity for its predicted site than for the permuted version. However, the level of discrimination between the two DNA sites is clearly not conserved when the fingers are swapped.

low specificity with some preference for cytosine, and strong selection against guanine (8, 17, 18); the aspartic acid, recognizing the central base of the last subsite, is selective for cytosine (8).

The two designed proteins have different absolute affinities for their respective binding sites. The RER-RER-RER protein contains six arginines as the only potential hydrogen bonding contacts, as the glutamic acids do not directly contact the DNA, as seen in the Zif268 cocrystal structure (4). In contrast, the QDR-RER-RHR protein contains five arginines, a histidine, an aspartic acid, and a glutamine, all of which are presumed to recognize bases via direct hydrogen bonding interactions. These differences may account for the 5000-fold difference in relative affinities of these proteins with their respective binding sites although it is possible that this difference is also due to more subtle structural effects.

Permutation of Two Zinc Fingers as a Test of Modular Specificity. We prepared a third protein, RER-QDR-RHR, that contains the same zinc fingers as QDR-RER-RHR, but with the order of the first two fingers exchanged. The predicted binding site for this protein is 5′-GGG-GCT-GCG-3′, based on a corresponding permutation of the second and third subsites of the binding site for QDR-RER-RHR. As a demonstration of the modular specificities of the QDR-RER-RHR and RER-QDR-RHR proteins, we performed DNase I footprinting analysis of both proteins with each predicted binding site. As shown in Fig. 6, each protein binds well to its predicted site and with lesser affinity to the alternate site, implying that these proteins, although having predicted sites that differ by only two bases, are able to discriminate between them. However, while the absolute affinities of both proteins for their predicted sites differ by only 3-fold, the discrimination between sites is markedly different for the two proteins. The QDR-RER-RHR protein discriminates between its preferred site and the permuted site by a factor of ≈180, whereas the RER-QDR-RHR protein discriminates between the two sites by a factor of only 7, with the order of affinities reversed as expected. This result demonstrates that, although zinc fingers can be qualitatively modular, in that specificities can be swapped along with individual fingers, zinc fingers are not always quantitatively modular. This lack of perfect modularity is at least partially caused by end effects, where the specificity due to a given contact residue may vary depending on whether it occurs in a central finger within an array or at one of the ends. This effect is seen here at the level of recognition residues only, as the sequence of the remainder of each zinc finger in this system is identical.

Conclusions. We have demonstrated that it is possible to design functional zinc-finger proteins based on multiple consensus-based zinc-finger motifs with appropriate changes endowing specificity to the individual zinc fingers. The design approach of combining a zinc-finger consensus sequence framework with different specificity-determining regions has been successful, although different levels of affinity and specificity for the different proteins are apparent. The specificity-determining residues were previously characterized in wild-type and mutated versions of zinc-finger sequences from other proteins. Importantly, they appear to retain the original specificities independent of framework sequence and domain order. Our results strongly favor both the dominant importance of the varied residues in determining individual subsite preferences and the qualitatively modular behavior of this class of zinc-finger domains. We have demonstrated that zinc fingers may not be quantitatively modular, suggesting that in general, design of zinc-finger proteins with specificities and affinities predictable in detail may not always be straightforward. As more information accumulates concerning the binding specificities of individual zinc-finger domains, this system should prove to be a powerful vehicle for the production of desired site-specific DNA binding proteins. The consensus approach to design may also prove to be generally useful for studying model systems of a variety of structural motifs, especially as sequence information for structural families continues to accumulate.

We thank the National Institutes of Health and the Lucille B. Markey Charitable Trust for support of this work.

1. DeGrado, W. F., Raleigh, D. P. & Handel, T. (1991) *Curr. Opinion Struct. Biol.* **1**, 984–993.
2. Rhodes, D. & Klug, A. (1986) *Cell* **46**, 123–132.
3. Berg, J. M. (1990) *Annu. Rev. Biophys. Biophys. Chem.* **19**, 405–421.
4. Pavletich, N. P. & Pabo, C. O. (1991) *Science* **252**, 809–817.
5. Nardelli, J., Gibson, T. J., Vesque, C. & Charnay, P. (1991) *Nature (London)* **349**, 175–178.
6. Desjarlais, J. R. & Berg, J. M. (1992) *Proteins* **12**, 101–104.
7. Desjarlais, J. R. & Berg, J. M. (1992) *Proteins* **13**, 272.
8. Desjarlais, J. R. & Berg, J. M. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 7345–7349.
9. Klevit, R. E. (1991) *Science* **253**, 1367.
10. Thukral, S. K., Morrison, M. L. & Young, E. T. (1992) *Mol. Cell. Biol.* **12**, 2784–2792.
11. Alexander, P., Fahnestock, S., Lee, T., Orban, J. & Bryan, P. (1992) *Biochemistry* **31**, 3597–3603.
12. Studier, F. W., Rosenberg, A. H., Dunn, J. J. & Dubendorff, J. W. (1990) *Methods Enzymol.* **185**, 61–89.
13. Brenowitz, M. & Senear, D. F. (1989) in *Current Protocols in Molecular Biology*, eds. Ausubel, F., Kingston, R., Moore, D., Seidman, J., Smith, J. & Struhl, K. (Wiley, New York), Suppl. 7, p. 12.4.1.
14. Ellington, A. D. & Szostak, J. W. (1990) *Nature (London)* **346**, 818–822.
15. Krizek, B. A., Amann, B. T., Kilfoil, V. J., Merkle, D. L. & Berg, J. M. (1991) *J. Am. Chem. Soc.* **113**, 4518–4523.
16. Kadonaga, J. T., Jones, K. A. & Tjian, R. (1986) *Trends Biochem. Sci.* **11**, 20–23.
17. Letovsky, J. & Dynan, W. S. (1989) *Nucleic Acids Res.* **17**, 2639–2653.
18. Thiesen, H. J. & Bach, C. (1990) *Nucleic Acids Res.* **18**, 3203–3209.
19. Brenowitz, M., Senear, D. F., Shea, M. A. & Ackers, G. K. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 8462–8466.