PROTEINS: Structure, Function, and Genetics 12:101–104 (1992)

RESEARCH ARTICLES

Redesigning the DNA-Binding Specificity of a Zinc Finger Protein: A Data Base-Guided Approach

John R. Desjarlais[1] and Jeremy M. Berg[1,2,3]

[1]Thomas C. Jenkins Department of Biophysics and [2]Department of Chemistry, Johns Hopkins University, Baltimore, Maryland 21218, and [3]Department of Biophysics and Biophysical Chemistry, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205

ABSTRACT **A peptide corresponding to the three zinc finger domains of the human transcription factor Sp1 has been expressed and found to bind a consensus Sp1 binding site with the sequence 5′-GGGGCGGGG-3′. Examination of the amino acid distributions within a large zinc finger sequence data base and chemical arguments suggested that a particular Arg to Gln sequence change might convert binding specificity to 5′-GGGGCAGGG-3′. Experimental tests of this hypothesis revealed that such a change could be induced only when two other sequence changes, deduced from examination of sequence correlations, were made as well. These results provide the most direct information to date about how zinc finger proteins might recognize adenine-containing binding sites and bear on the existence and nature of any code between zinc finger protein and binding site sequences.**

Key words: protein–DNA interactions, hydrogen bonding, Sp1, recognition code, amino acid correlations, protein–DNA interface

INTRODUCTION

Even prior to the determination of the structure of any sequence-specific DNA-binding protein, hypotheses were proposed concerning how particular amino acids could be used to directly contact nucleotide bases specifically.[1] Subsequent elucidation of the structures of several DNA-binding proteins and their complexes with oligonucleotides has revealed a complex set of interactions such that it is clear that no simple code can exist.[2] Nonetheless, comparison of the structures of cocrystals of the lambda and 434 repressor amino terminal domains have revealed the presence of similar modes of recognition.[3] In addition, for the lac repressor subclass of helix-turn-helix motif containing proteins, limited progress was made toward some rules relating protein and binding site sequences with use of extensive protein and binding site mutagenesis followed by screening.[4] The zinc finger proteins typified by *Xenopus* transcription factor IIIA represent a particularly exciting class of DNA binding proteins for code investigation.[5–7] These proteins contain tandem arrays of modules, each of which appears capable of interacting with three base pairs of DNA via the major groove. Thus, determination of the relationship between the amino acid sequence of a single module and its three base pair binding site could be rapidly extended to larger proteins. In addition, this appears to be a very large superfamily of eukaryotic gene regulatory proteins that includes members that play central roles in development and in human disease.[5]

We have chosen the human general transcription factor Sp1[8] as a system for investigating specific zinc finger protein–DNA interactions. This protein has three tandem zinc finger domains in its DNA binding region[9] and it binds to sequences that match the consensus 5′-(G,T)GGGCGG(G,A)(G,A)-3′. Methylation protection[10] and interference[11] experiments indicated that the protein interacts with the DNA in the major groove with many of the conserved guanines being directly contacted. The orientation of Sp1 on its binding site has been proposed based on comparison with Krox-20, another three zinc finger-containing protein that binds to a permuted version of the Sp1 binding site.[6] Thus, finger 1 appears to contact the 3′ triplet, finger 2 the central GCG, and finger 3 the 5′ triplet. Proposals of specific amino acids responsible for base recognition were made based on this comparison and were supported by mutagenesis studies. It was proposed that residues in positions 13, 15, 16, and 19 in the canonical zinc finger sequence of the form Pro_1 $(Tyr,Phe)_2$-X-Cys_4-X-X-Cys_7-X-X-X-Phe_{11}-X-X_{13}-X-

Received April 29, 1991; revision accepted June 27, 1991.

Address reprint requests to Jeremy M. Berg, Department of Biophysics and Biophysical Chemistry, Johns Hopkins University School of Medicine, 725 North Wolfe Street, Wood Basic Science Building, Baltimore, MD 21205.

```
                    M E K L R N G S G D P G K K K Q H I
C H I Q G C G K V Y G K T S H L R A H L R W H T G E R P F M
C T W S Y C G K R F T R S D E L Q R H K R T H T G E K K F A
C P E - - C P K R F M R S D H L S K H I K T H Q N K K G
```

Fig. 1. Amino acid sequence of the three zinc finger peptide Sp1-3F 2RSDE. The peptide consists of the zinc finger region of human Sp1 with a seven amino acid leader sequence. The zinc ligands are shown in bold. In the two variants Sp1-3F 2QSSE and Sp1-3F 2QSSD the overlined region is changed to QSSE and QSSD, respectively.

X_{15}-X_{16}-Leu_{17}-X-X_{19}-His_{20}-X-X-X-His_{24} play the most direct roles in determining the binding site sequence. These conclusions have been independently developed based on a crystal structure determination of the three zinc finger domains of Zif268 bound to an oligonucleotide containing its binding site.[7]

MATERIALS AND METHODS

Construction and Expression of Sp1-3F 2RSDE and Variants

The region of DNA encoding the three zinc fingers of Sp1 was subcloned from pSp1-516C[9] into the *Eco*RI and *Hin*dIII sites of the expression vector pKK223-3 (Pharmacia), with use of polymerase chain reaction (PCR) methods[12] to incorporate the restriction sites, a T7 Shine–Dalgarno sequence, and a short leader sequence (for purposes of expression) into a DNA fragment encoding the zinc fingers. This expression construct, pKKSp1-3F, was used to overexpress the peptide Sp1-3F 2RSDE, shown in Figure 1. PCR mutagenesis[13] of pKKSp1-3F was used to create variants of this sequence. Correct sequences were confirmed by double stranded sequencing[14] of the constructs with Sequenase (United States Biochemical).

E. coli 71-18 cells harboring the expression constructs were grown to an OD_{600} of 0.7 in 30 ml of LB media supplemented with 100 μM $ZnCl_2$ and 100 μg/ml Ampicillin. Isopropylthiogalactoside (1 mM) was added and after 6 hr of further growth the cells were collected by centrifugation of 1.5-ml aliquots. The supernatant was discarded and the pellets stored at −80°C. To prepare partially pure peptide a single pellet was resuspended in 42.5 μl of 25 mM Tris-Cl pH 8.0, 100 μM $ZnCl_2$, 5 mM dithiothreitol, and heated in a boiling water bath for 4 min. KCl was added to 150 mM and the extract was centrifuged for 10 min at 4°C. This procedure yields a peptide approximately 70% pure as judged by SDS-PAGE.[15]

Construction of Binding Site Probes

The four DNA probes, each containing one of the possible binding sites 5′-GGGGCNGGG-3′, were created by cloning synthetic complementary deoxyoligonucleotides with this sequence and flanking sequence into the pEMBL vector[16] polylinker and confirmed by double-stranded sequencing of individual clones. For each probe, a 79 base pair *Eco*RI–*Hin*dIII restriction fragment containing the binding site was end labeled with [α-^{32}P]dATP and gel purified.

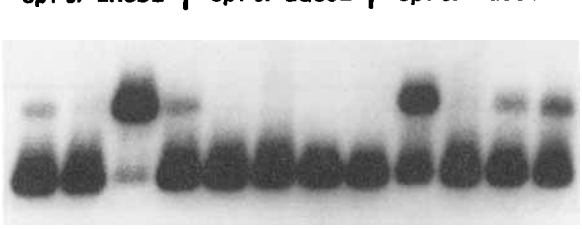

Fig. 2. Gel retardation assay. Binding of Sp1-3F 2RSDE and the two variants to DNA probes containing one of the binding sites 5′-GGGGCNGGG-3′, where N is A, C, G, or T as shown below each lane. The upper bands represent protein-bound DNA and the lower bands represent free DNA.

Gel Retardation Assay

One to four microliters (~0.5 μg) of partially purified peptide was added to a labeled DNA probe (1–10 ng) containing one of the four binding site sequences so that the final reaction conditions in a 10 μl volume were as follows: 35 mM Tris-Cl pH 8.0, 60 mM KCl, 90 μM $ZnCl_2$, 3 mM DTT, 300 μg/ml BSA, 20 μg/ml poly(dI-dC), and 10% glycerol. After incubating at room temperature for 15 min, the samples were electrophoresed on a 1.8% Seaplaque (FMC) agarose gel. The gel was dried and autoradiographed.

RESULTS AND DISCUSSION

Expression and Binding of a Three Zinc Finger Peptide

In order to study determinants of the specificity of interaction of zinc fingers with DNA and to explore possibilities for designing changes of specificity, we first constructed a system for the overexpression of a peptide representing the three zinc fingers of human Sp1. The sequence of this peptide, Sp1-3F 2RSDE, is shown in Figure 1. This peptide binds with high affinity to the sequence 5′-GGGGCGGGG-3′ as demonstrated by a gel retardation assay (see Fig. 2, lane 3) and DNase I footprinting (data not shown). Thus, Sp1-3F 2RSDE, representing only the three zinc fingers of human Sp1, appears to bind DNA with the same specificity as the full length protein, paralleling observations with other zinc finger proteins.[7,17] Larger fragments of Sp1 had been previously shown to bind DNA specifically.[9,18]

Redesign of DNA Binding Specificity

In attempting to redesign the DNA binding specificity of the three zinc finger peptide Sp1-3F 2RSDE,

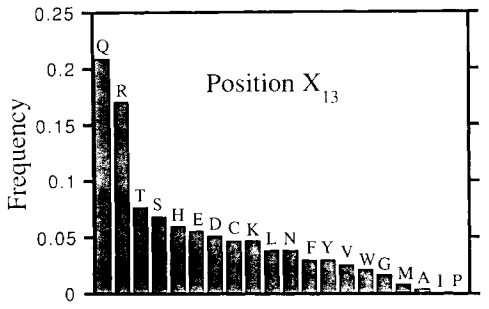

Fig. 3. The amino acid distribution for position 13 from a data base of over 200 zinc finger sequences. The distribution is dominated by potential hydrogen-bonding residues including especially Gln and Arg.

we have looked to large sequence database sequences of zinc finger proteins for information concerning particular choices of amino acid changes or combinations of changes that might be effective. The data base contained the aligned sequences of over 200 zinc finger domains.[5] The amino acid distribution for position 13 is shown in Figure 3. This distribution strongly suggests that residues in this position are important for determining DNA-binding specificity in that the distribution is dominated by potential hydrogen bond donor- and acceptor-containing residues but with considerable variation. The second most frequently occurring residue in this position is Arg which is present in domains 2 and 3 of Sp1. It has been proposed[6] and demonstrated[7] that this Arg contacts a guanine at the 3′ end of the three base pair unit contacted by each zing finger via two hydrogen bonds. The most frequently occurring residue at this position is Gln. Since Gln can make an analogous doubly bound contact with adenine rather than guanine,[1-3] we hypothesized that changing the Arg to Gln might change the binding specificity from guanine to adenine. Further examination of the sequence data base revealed amino acid correlations which suggested that accompanying changes in positions 15 and 16 might be necessary to achieve the desired change of specificity. The amino acid distributions at positions 15 and 16 are strongly correlated with the identity of the residue at position 13. This is illustrated in Figure 4. When Arg is present in position 13, residue 15 tends to be Asp and residue 16 is most commonly Glu or His. In contrast, when position 13 is Gln, residue 15 tends to be Ser and residue 16 is most commonly Asp. There are no examples of $Gln_{13}Asp_{15}$ sequences in the data base. To test whether these sequence correlations are, indeed, important for DNA binding, two sequence variants of the second domain of the peptide Sp1-3F 2RSDE were prepared: a double variant with Arg_{13} to Gln and Asp_{15} to Ser (Sp1-3F 2QSSE) and a triple variant with Glu_{16} changed to Asp (Sp1-3F 2QSSD). Binding studies revealed that the double variant was incapable of high affinity binding to any of the target sequences. However, the triple variant was found to bind 5′-GGGGCAGGG-3′ specifically as shown in Figure 2. The overall affinity of this peptide for its cognate binding site appears to be somewhat lower than that observed for the wild-type. Additional studies of other variants (Sp1-3F 2QSDE, Sp1-3F 2QSDD) revealed that all three sequence changes were required to produce the observed change in DNA-binding specificity (data not shown). These results indicate that the Arg to Gln conversion does induce the predicted change in DNA-binding specificity, but in a context-dependent manner.

This context dependence can be rationalized in two ways. First, side chain–side chain interactions have been shown to be important in protein–DNA interactions. Such interactions have been observed in, for example, the lambda repressor and phage 434 repressor cocrystal structures.[3] Furthermore, in the Zif268 cocrystal structure, such interactions are seen involving the $Arg_{13}Asp_{15}$pairs.[7] Thus,the correlation between Gln_{13} and Ser_{15} may be due to a hydrogen-bonding interaction between the hydroxyl group of the Ser with the carbonyl group of the Gln side chain. The orientation of the carboxamide of the Gln residue is fixed by virtue of the proposed bidentate interaction with adenine. Alternatively, the Ser residue could hydrogen bond to the thymine O4 that is base paired to the adenine residue. The Asp over Glu preference in position 16 obviously cannot be explained in terms of differences in functional groups. However, Gln and Ser are two and one nonhydrogen atoms shorter than are Arg and Asp, respectively, so that Glu may be too long to fit into the protein–DNA interface within the context of the Gln-Ser interaction.

CONCLUSION

We have successfully designed and constructed a zinc finger array variant that shows altered DNA-binding specificity. Our design was based on chemical arguments supplemented by use of a sequence data base to elucidate secondary sequence changes that were required to produce an appropriate context. The success of this approach suggests that a code that relates zinc finger protein sequences to the sequences of their binding sites may exist and that this code may be rationalizable or even predictable in terms of specific bonding interactions. However, this code must involve sets of amino acid residues that, together, form effective surfaces for specifically interacting with DNA.

ACKNOWLEDGMENTS

We wish to thank Drs. Laura Zawadzke and Charles Vinson for helpful suggestions for the expression and purification of peptides, Nikola Pavletich and Professor Carl Pabo for communication of

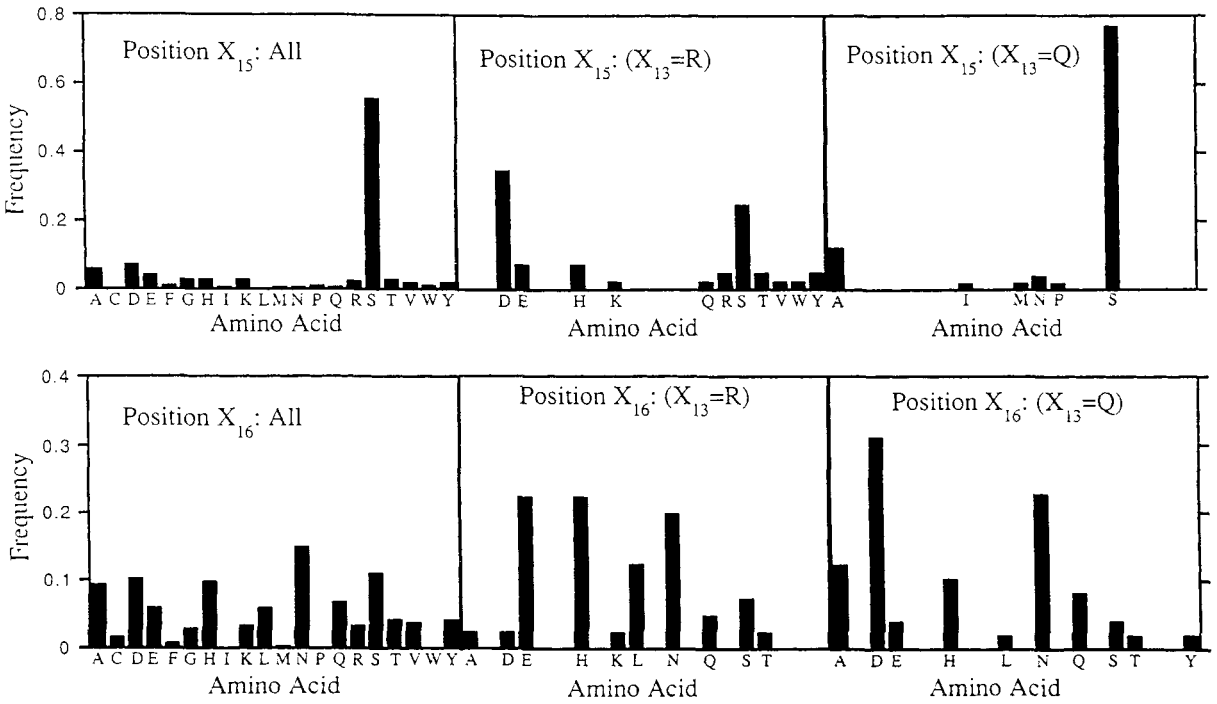

Fig. 4. The amino acid distributions for positions 15 and 16. The upper three panels show the distributions at position 15. The distributions shown are for all sequences, for those sequences that have Arg at position 13, and those sequences that have Gln at position 13. Overall, Ser dominates the distribution but when Arg is in position 13, Asp becomes the most frequently occurring residue. The bottom three panels show the corresponding distributions for position 16. A more diverse group of residues occurs at this position. However, when Arg is present at position 13, Glu and His are most frequently found whereas when Gln is present at position 13, Asp and Asn are more prevalent.

results prior to publication, and Professors James Kadonaga and Robert Tjian for providing us with the plasmid pSp1-516C. This research was supported by grants from the National Institutes of Health (GM-38230,GM-07231) and the National Science Foundation (DMB-8850069). J.M.B. is a fellow of the Alfred P. Sloan Foundation.

REFERENCES

1. Seeman, N.C., Rosenberg, J.M., Rich, A. Sequence-specific recognition of double helical nucleic acids by proteins. Proc. Natl. Acad. Sci. U.S.A. 73:804–808, 1976.
2. Steitz, T.A. Structural studies of protein-nucleic acid interaction: The sources of sequence-specific binding. Q. Rev. Biophys. 23:205–280, 1990.
3. Pabo, C.O., Aggarwal, A.K., Jordan, S.R., Beamer, L.J., Obeysekare, U.R., Harrison, S.C. Conserved residues make similar contacts in two repressor-operator complexes. Science 247:1210–1213, 1990.
4. Lehming, N., Sartorius, J., Kisters-Woike, B., von Wilcken-Bergmann, B., Müller-Hill, B. Mutant lac repressors with new specificities hint at rules for protein-DNA recognition. EMBO J. 9:615–621, 1990.
5. Berg, J.M. Zinc finger domains: hypothesis and current knowledge. Annu. Rev. Biophys. Biophys. Chem. 19:405–421, 1990.
6. Nardelli, J., Gibson, T.J., Vesque, C., Charnay, P. Base sequence discrimination by zinc-finger DNA-binding domains. Nature (London) 349:175–178, 1991.
7. Pavletich, N.P., Pabo, C.O. Zinc finger-DNA recognition: crystal structure of a Zif268–DNA complex at 2.1 Å. Science 252:809–817, 1991.
8. Kadonaga, J.T., Jones, K.A., Tjian, R. Promoter-specific activation of RNA polymerase II transcription by Sp1. Trends Biochem. Sci. 11:20–23, 1986.
9. Kadonaga, J.T., Carner, K.C., Masiarz, F.R., Tjian, R. Isolation of cDNA encoding transcription factor Sp1 and functional analysis of the DNA binding domain. Cell 51:1079–1090, 1987.
10. Gidoni, D., Dynan, W.S., Tjian, R. Multiple specific contacts between a mammalian transcription factor and its cognate promoter. Nature (London) 312:409–413, 1984.
11. Westin, G., Schaffner, W. A zinc-responsive factor interacts with a metal-regulated enhancer element (MRE) of the mouse metallothionein-I gene. EMBO J. 7:3763–3770, 1988.
12. Saiki, R.K., Gelfand, D.H., Stoffel, S., Scharf, S.J., Higuchi, R., Horn, G.T., Mullis, K.B., Erlich, H.A. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science 239:487–491, 1988.
13. Higuchi, R., Krummel, B., Saiki, R.K. A general method of *in vitro* preparation and specific mutagenesis of DNA fragments: Study of protein and DNA interactions. Nucl. Acids Res. 16:7351–7366, 1988.
14. Lee, S., Rasheed, S. A simple procedure for maximum yield of high quality plasmid DNA. BioTechniques 9:676–679, 1990.
15. Laemmli, U.K. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature (London) 227:680–685, 1970.
16. Dente, L., Cesareni, G., Cortese, R. pEMBL: A new family of single stranded plasmids. Nucl. Acids Res. 19: 1645–1654, 1983.
17. Nagai, K., Nakaseko, Y., Nasmyth, K., Rhodes, D. Zinc-finger motifs expressed in *E. coli* and folded *in vitro* direct specific binding to DNA. Nature (London) 332:284–286, 1988.
18. Kuwahara, J., Coleman, J.E. Role of the zinc(II) ions in the structure of the three-finger DNA binding domain of the Sp1 transcription factor. Biochemistry 29:8627–8631, 1990.