

## Protein stitchery: Design of a protein for selective binding to a specific DNA sequence

CHANGMOON PARK\*†, JUDY L. CAMPBELL†‡, AND WILLIAM A. GODDARD III\*†§

\*Materials and Molecular Simulation Center, Beckman Institute (139-74), †Division of Chemistry and Chemical Engineering, and ‡Division of Biology, California Institute of Technology, Pasadena, CA 91125

Contributed by William A. Goddard III, June 3, 1992

**ABSTRACT** We present a general strategy for designing proteins to recognize DNA sequences and illustrate this with an example based on the “Y-shaped scissors grip” model for leucine-zipper gene-regulatory proteins. The designed protein is formed from two copies, in tandem, of the basic (DNA binding) region of v-Jun. These copies are coupled through a tripeptide to yield a “dimer” expected to recognize the sequence TCATCGATGA (the v-Jun–v-Jun homodimer recognizes ATGACTCAT). We synthesized the protein and oligonucleotides containing the proposed binding sites and used gel-retardation assays and DNase I footprinting to establish that the dimer binds specifically to the DNA sequence TCATCGATGA but does not bind to the wild-type DNA sequences, nor to oligonucleotides in which the recognition half-site is modified by single-base changes. These results also provide strong support for the Y-shaped scissors grip model for binding of leucine-zipper proteins.

We propose a general strategy for designing proteins to recognize specific DNA-binding sites: this strategy is to select segments of proteins, each of which recognizes particular DNA segments and to stitch these segments together via a short peptide with a cystine crosslink in a way compatible with each peptide being able to bind to its own DNA segment. This technique creates a protein that recognizes the composite site.

As a starting point we consider the gene-regulatory leucine-zipper proteins. They are characterized by two structural motifs (1–3): (i) the leucine zipper, which is responsible for dimerization, and (ii) the basic region for DNA binding (4–7). The basic regions of unbound leucine-zipper dimers are unfolded but fold into the  $\alpha$ -helix conformation upon binding to the specific site (8–10). The most plausible model for the conformation of leucine-zipper protein is the “Y-shaped scissors grip” model (1, 2), in which the basic region of each monomer interacts with DNA on either side of the dyad axis of the binding site. Thus, for yeast transcriptional activator GCN4 each arm recognizes the half-site AGTA (11, 12).

### DESIGN

Our design strategy assumes this Y-shaped scissors grip model (Fig. 1a). We design proteins by crosslinking (stitching together) various binding arms so as to be consistent with the orientation of the recognition helix in each half-site. Here we build upon the results of Kim and coworkers (5, 6), who showed that the leucine zipper of GCN4 can be replaced with linkers (Gly-Gly-Cys) at the C terminus of the DNA binding segment, which upon oxidation dimerize and bind to the same site (ATGACTCAT) as GCN4. As a model system to explore the design of additional DNA-binding proteins, we have chosen the v-Jun leucine-zipper dimer (Fig. 1a), which also

binds to the site ATGACTCAT as a homodimer with itself or as a heterodimer with Fos (4, 13–16), another member of this DNA-binding protein family. We will reverse the sequence relationship of the  $\alpha$ -helix to the target nucleotide of the binding arms by adding the Gly-Gly-Cys linker to the N terminus (rather than to the C terminus). As illustrated in Fig. 1b the designed protein then recognizes the DNA sequence TCATXATGA, where X represents 0–2 additional bases to accommodate the loop region of this dimer.

Several criteria were used in selecting v-Jun as the starting point: (i) To prevent nonspecific disulfide-bond formation, the protein must not contain cysteine in its basic region. (ii) Because we want to reverse the  $\alpha$ -helix relative to the target DNA sequence, the protein should have no residues (especially proline and probably glycine) that would interrupt  $\alpha$ -helices. (iii) Because we want to ensure that the protein can form the  $\alpha$ -helix when joined with the linker, the composition of amino acids in its basic region should strongly favor  $\alpha$ -helices [by the Chou–Fasman criterion (17)]. We considered 14 leucine-zipper proteins and found that v-Jun best satisfies the above criteria.

We took as our standard protein the 31 residues at the N terminus of v-Jun joined with the linker (Gly-Gly-Cys) (Fig. 2a). The subsequent protein (v-Jun–NN) is designed to bind specifically to the site TCATXATGA, where X might contain 0–2 base pairs (bp). As indicated in Fig. 2b, we considered four cases for X: (i)  $X = \phi$  (no base pairs), denoted as NNS– $\phi$ ; (ii)  $X = C$  (which is equivalent to  $X = G$ ), denoted as NNS–C/G; (iii)  $X = CG$ , denoted as NNS–CG; and (iv)  $X = GC$ , denoted as NNS–GC. We excluded using adenine or thymine on the assumption that the methyl group of thymine (which sits in the major groove) might interfere with binding of the protein.

### TESTS OF THE DESIGN

We carried out gel-retardation assays using four DNA sequences: (a) the sequence TCATCGATGA (case *iii* above), NNS–CG; (b) the binding sequence for v-Jun, ATGACTCAT; (c) the complementary double-base-pair substitution ( $C^2 \rightarrow A^2$  and  $G^3 \rightarrow T^3$ ) of a: TAATCGATTA; and (d) the complementary double-base-pair substitution ( $A^3 \rightarrow C^3$  and  $T^8 \rightarrow G^8$  of a: TCCTCGAGGA).

The results (Fig. 3) indicate that v-Jun–NN binds to the DNA sequence a as a homodimer with a  $K_d$  of  $<1$  nM at 4°C. On the other hand, v-Jun–NN does not bind significantly to the wild-type site b or to the mutant sites c and d.

To establish the specific binding site for v-Jun–NN, we used deoxyribonuclease (DNase) I footprinting. These results (Fig. 4) show that v-Jun–NN protects the exact binding site predicted for the designed protein. Thus, we conclude that each arm of DNA-bound v-Jun–NN retains the same structure as in native v-Jun. The DNase I footprinting results (Fig. 4) also indicate that NNS–CG has the strongest binding affinity for

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.

§To whom reprint requests should be addressed.

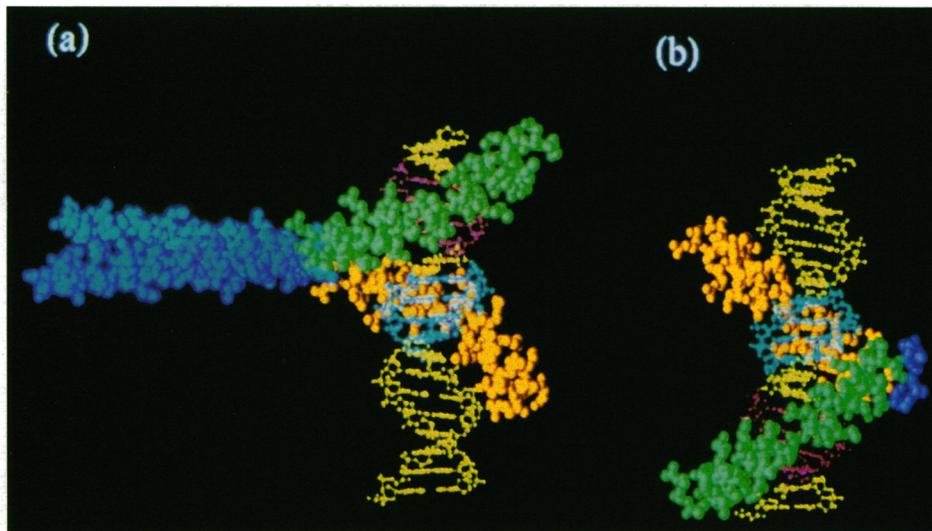


FIG. 1. (a) The Y-shaped scissors grip model for the v-Jun-v-Jun homodimer bound to the ATGACTCAT site. (b) The designed protein v-Jun-NN. After removing the leucine-zipper region (blue and light-blue) of each monomer, the upper arm (green) and its DNA-binding site (pink, ATGA) were shifted just below the lower arm (orange). In b the shifted upper arm and DNA fragment retain their original green and pink colors, respectively. Different linkers (Gly-Gly-Cys, purple) were added at the N termini of both arms, and a disulfide bond was made. v-Jun-N was dimerized by oxidized dithiothreitol and purified by HPLC. Protein concentration was determined by the method of Bradford with the Bio-Rad protein assay kit. Thus, the designed protein is expected to bind to TCATXATGA (where cyan joins to pink in b) and does not bind to ATGACTCAT (where pink joins to cyan in a), where X (filling bases) fit the loop introduced between the peptide monomers to make the dimer.

v-Jun-NN. The decrease in binding for shorter X may result from the strain required at the loop region of the dimer to place the binding segments along the binding region. Particularly interesting is the difference in specificity observed between NNS-GC and NNS-CG (Fig. 4, compare lanes 9 and 12 for top strands and lanes 21 and 24 for bottom strands). These results indicate that X plays more than the role of spacer.

**DISCUSSION**

These results support the idea that the N-terminal region of v-Jun contributes to the binding to DNA through specific interaction with the DNA (because in v-Jun-NN this region is forced to contact the DNA). This result supports the angulated bend conformation (1). Our results help differen-

tiate the respective roles of the basic region and of the leucine-zipper region in recognition and binding. The basic region of v-Jun is sufficient for specific binding. Although the leucine-zipper region is not directly involved in DNA binding, our results indicate that its position relative to the basic region plays an important role in determining which target sequence of DNA the protein recognizes.

Summarizing, we have designed a protein (stitched together from segments derived from the natural protein) to recognize a specific DNA-binding site, and we have established specific binding of the designed protein to this site. Note that use of the Gly-Gly-Cys linker is not essential in the design. We could just as well replace the cysteine and make a continuous ≈70-amino acid protein that should recognize a predictable site (14). In addition, this strategy is not limited to two arms. We could have stitched together three, four, or more arms with appropriate linkers to design proteins that would recognize DNA sequences with 15, 20, or 25 bp. Such systems with EDTA-Fe (18) or other nucleases would presumably cut very specific sites, allowing the genome to be cut into much longer segments. The design is not limited to v-Jun. Any protein or other molecule that recognizes a specific

<b>a Protein</b>	
vJun-br:	S QERIKAERKR MRNRIAASKS RKRKLERIAR
vJun-N:	CGG S QERIKAERKR MRNRIAASKS RKRKLERIAR
<b>b DNA</b>	
NNS-φ:	..... TCATATGA .....
NNS-C/G:	..... TCAT(C/G) ATGA .....
NNS-CG:	..... TCATCGATGA .....
NNS-GC:	..... TCATGCATGA .....
α:	..... TCATCGATGA .....
β:	..... ATGACTCAT .....
γ:	..... TAAATCGATTA .....
δ:	..... TCCTCGAGGA .....

FIG. 2. Sequences of protein and oligonucleotides used in gel retardation and footprinting. v-Jun-br contains the basic region of v-Jun, and Gly-Gly-Cys is added to make v-Jun-N. Single-letter amino acid code is used. The protein corresponding to the residues from 218 to 346 of v-Jun was chemically synthesized at the Biopolymer Synthesis Center at the California Institute of Technology. The automated stepwise solid-phase syntheses were done on an Applied Biosystems model 430A peptide synthesizer with an optimized synthetic protocol of the N-t-butoxycarbonyl (t-Boc) chemistry. The peptide was purified by reverse-phase HPLC on a Vydac C<sub>18</sub> column. A linear gradient of 0–50% aqueous/acetonitrile/0.1% trifluoroacetic acid was run over 120 min. Mass spectroscopy data are as follows: calculated, 4039.3; experimental, 4041.8.

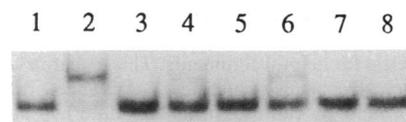


FIG. 3. Gel-retardation assay for binding of v-Jun-NN to various DNA segments. The dimer selectively binds to the predicted site NNS-CG. Even-numbered lanes, no protein; odd-numbered lanes, v-Jun-NN. Lanes: 1 and 2, probe DNA α (NNS-CG); 3 and 4, probe DNA β (wild-type site); 5 and 6, probe DNA γ; 7 and 8, probe DNA δ (Fig. 2b). The binding solution contains bovine serum albumin at 50 μg/ml, 10% (vol/vol) glycerol, 20 mM Tris-HCl (pH 7.5), 4 mM KCl, 2 mM MgCl<sub>2</sub>, and 1.56 nM v-Jun-NN in 10-μl reaction volume. After 5000 cpm of each 5'-<sup>32</sup>P-labeled probe DNA (25- and 26-mer) was added, the solutions were stored at 4°C for 1 hr and loaded directly on an 8% nondenaturing polyacrylamide gel in Tris/EDTA buffer at 4°C. As determined by titration of the gel shift, v-Jun-NN binds to the predicted sequence NNS-CG with a K<sub>d</sub> of ≈0.3 nM at 4°C.

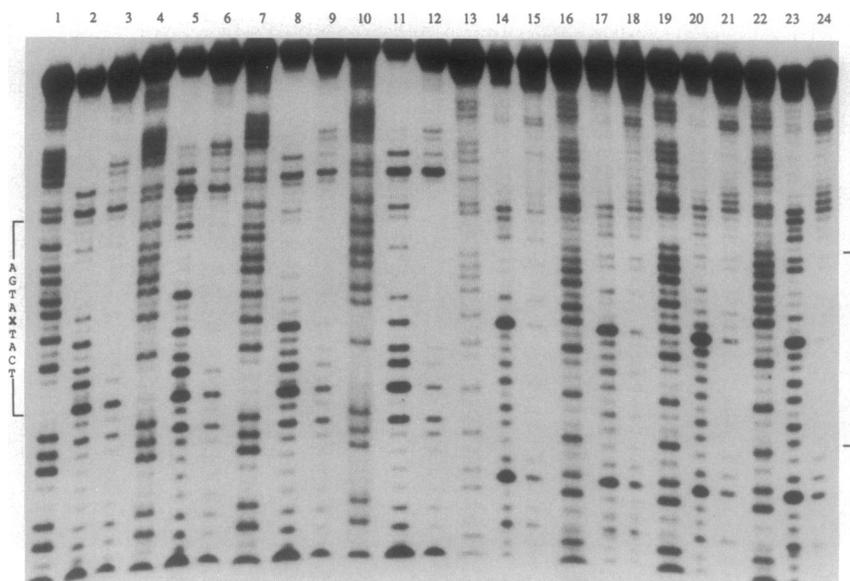


FIG. 4. DNase I footprinting assay of v-Jun-NN with DNA containing the predicted binding sites NNSs. This result shows that the protein protects the target-binding site with the best protection for NNS-CG. Lanes: 1–12, labeled at the 5' end of top strand; 13–24, labeled at the 5' end of bottom strand; 1–3 and 13–15, NNS- $\phi$ ; 4–6 and 16–18, NNS-C/G; 7–9 and 9–21, NNS-GC; 10–12 and 22–24, NNS-CG. The first lane for each probe DNA (lanes 1, 4, 7, 10, 13, 16, 19, and 22) contains G+A marker; the middle lane for each probe DNA (lanes 2, 5, 8, 11, 14, 17, 20, and 23) contains no protein; and the last lane for each probe DNA (lanes 3, 6, 9, 12, 15, 18, 21, and 24) contains v-Jun-NN. The footprinting assay solution (in 50  $\mu$ l) contains bovine serum albumin at 50  $\mu$ g/ml, 5% (vol/vol) glycerol, 20 mM Tris-HCl (pH 7.5), 4 mM KCl, 2 mM MgCl<sub>2</sub>, 1 mM CaCl<sub>2</sub>, 20,000 cpm of each 5'-<sup>32</sup>P-labeled probe DNA (60- to 62-mer), and 50 nM v-Jun-NN. This solution was stored at 4°C for 1 hr. After 5  $\mu$ l of DNase I diluted in 1 $\times$  footprinting assay buffer was added, the solutions were stored 1 min more at 4°C. The DNase I digestion was stopped by adding 100  $\mu$ l of DNase I stop solution containing 15 mM EDTA (pH 8.0), 100 mM NaCl, sonicated salmon sperm DNA at 25  $\mu$ g/ml, and yeast tRNA at 25  $\mu$ g/ml. This mixture was phenol/chloroform-extracted, ethanol-precipitated, and washed with 70% (vol/vol) ethanol. The pellet was resuspended in 5  $\mu$ l of formamide loading buffer, denatured at 90°C for 4 min, and analyzed on 10% polyacrylamide sequencing gel (50% urea).

DNA sequence by binding along the major groove could be a candidate. Many such cases are now known so that we already have a collection of available partial-binding sites that could be combined to form composite target-binding sites for designing binding proteins. Of course, the segments of these proteins should be designed so that the intramolecular interactions are not so strong as to compete with binding to the DNA.

Our results support the idea that each monomer arm of the dimer binds along the major groove to the half of the binding site of the dimer (11, 12); this strongly supports the Y-shaped scissors grip model for leucine-zipper proteins (1). Our strategy can be used to investigate the interaction between DNA and protein and the structure of DNA-protein complex.

We thank Prof. Mel Simon (California Institute of Technology) for helpful suggestions and discussion and for use of his laboratory resources (funded by National Science Foundation Grant-DMB-90-18536). We also thank Prof. David Eisenberg (University of California, Los Angeles) for helpful comments. This research was supported by a grant from Department of Energy-Advanced Industrial Concepts Division and by funding from the Materials and Molecular Simulation Center of the Beckman Institute. This is contribution number 8547 from the Division of Chemistry and Chemical Engineering.

1. Vinson, C. R., Sigler, P. B. & McKnight, S. L. (1989) *Science* **246**, 911–916.

2. Pu, W. T. & Struhl, K. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 6901–6905.
3. Landschultz, W. H., Johnson, P. F. & McKnight, S. L. (1988) *Science* **240**, 1759–1764.
4. Neuberger, M., Schuermann, M., Hunter, J. B. & Muller, R. (1989) *Nature (London)* **338**, 589–590.
5. Tlalian, R. V., McKnight, C. J. & Kim, P. S. (1990) *Science* **249**, 769–771.
6. O'Shea, E. K., Rutkowski, R. & Kim, P. S. (1989) *Science* **243**, 538–542.
7. Agre, P., Johnson, P. F. & McKnight, S. L. (1989) *Science* **246**, 922–926.
8. Weiss, M. A. (1990) *Biochemistry* **29**, 8020–8024.
9. Weiss, M. A., Ellenberger, T., Wobbe, C. R., Lee, J. P., Harrison, S. C. & Struhl, K. (1990) *Nature (London)* **347**, 575–578.
10. Saudek, V., Pasley, H. S., Gibson, T., Gausepohl, H., Frank, R. & Pastore, A. (1991) *Biochemistry* **30**, 1310–1317.
11. Oakley, M. G. & Dervan, P. B. (1990) *Science* **248**, 847–850.
12. O'Neil, K. T., Hoess, R. H. & DeGrado, W. F. (1990) *Science* **249**, 774–778.
13. Struhl, K. (1987) *Cell* **50**, 841–846.
14. Bos, T. J., Ramscher, F. J., III, Curran, T. & Vogt, P. K. (1989) *Oncogene* **4**, 123–126.
15. Abate, C., Luk, D., Gentz, R., Ramscher, F. J., III, & Curran, T. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 1032–1036.
16. Turner, R. & Tjian, R. (1989) *Science* **243**, 1689–1694.
17. Chou, P. Y. & Fasman, G. D. (1973) *J. Mol. Biol.* **74**, 263–281.
18. Mack, D. P., Iverson, B. L. & Dervan, P. B. (1988) *J. Am. Chem. Soc.* **110**, 7572–7574.
19. Nakabeppu, Y. & Nathans, D. (1989) *EMBO J.* **8**, 3833–3841.