Identification of Oligomerizing Peptides*

Anjali Dhiman, Michael E. Rodgers, and Robert Schleif‡

From the Biology Department, Johns Hopkins University, Baltimore, Maryland 21218

The AraC DNA binding domain is inactive in a monomeric form but can activate transcription from the arabinose operon promoters upon its dimerization. We used this property to identify plasmids encoding peptide additions to the AraC DNA binding domain that could dimerize the domain. We generated a high diversity library of plasmids by inserting 90-base oligonucleotides of random sequence ahead of DNA coding for the AraC DNA binding domain in an expression vector, transforming, and selecting colonies containing functional oligomeric peptide-AraC DNA binding domain chimeric proteins by their growth on minimal arabinose medium. Six of seven Ara" candidates were partially characterized, and one was purified. Equilibrium analytical centrifugation experiments showed that it dimerizes with a dissociation constant of \( \sim 2 \mu M \).

Coiled-coils about 40 residues long seem to be the smallest naturally occurring oligomerization domains yet found in proteins. Does the absence of smaller oligomerization elements mean that none are possible or that natural selective pressures for such elements have not existed? If very strong selective pressures were exerted, might new and unique structural motifs be found? And if they were, would the resulting information prove useful in understanding protein structure and for protein engineering?

The identification and genetic selection of oligomerizing elements is possible with systems where the monomeric form of a protein or domain of a protein is inactive but the dimeric or oligomeric form is active. Transcription regulators are particularly attractive for this application. Often, binding of their monomeric form to DNA is weak, and their binding at physiological concentrations of the protein requires dimeric protein and corresponding repeated DNA sites. These systems also have the virtue that they can be adjusted to select for strong or weak dimerizing abilities of the monomers. If the affinity of the monomeric form of the protein for its DNA site is relatively high, active protein need not be a dimer in solution; i.e. occasionally, monomers independently bind to the DNA and dimerize there, utilizing the additional interaction energy between the two protein monomers. The lifetime of the DNA of the dimer thus formed is significantly longer than the lifetimes of individual monomers on the DNA. On the other hand, if the DNA binding affinity of the monomers is relatively low, then the interaction energy between the monomers of the active protein must be higher, and the protein may well exist in solution at physiological concentrations as a dimer.

We have utilized the arabinose system of Escherichia coli for the selection of dimerizing peptides (Fig. 1). Monomers of the DNA binding domain of AraC are inactive in stimulating transcription of the genes required for the catabolism of L-arabinose, but attaching a 49-amino acid coiled-coil from C/EBP converts the AraC DNA binding domain to a fully active form (1). We sought to isolate 30 amino acid peptides that dimerize at physiological concentrations. Therefore, in the C/EBP coiled-coil-AraC DNA binding domain construct, we replaced the region coding for the coiled-coil with 90 bases of DNA of random sequence and selected for products capable of activating transcription of the arabinose genes. While this work was in progress, three reports of similar selections have appeared. Two of these utilized λ phage repressor (2, 3) and report the genetic characterization of candidate elements encoded by fragments of cloned DNA. A third (4) fused random sequence DNA coding for 15 amino acids to DNA coding for zinc finger domains, selected, and after further improvement in dimerization, examined dimerization with centrifugation experiments. In our work, candidate peptides with lengths from 6 to 32 residues were identified. We purified the chimeric peptide-AraC DNA binding domain product from a candidate with a 22-amino acid dimerizing peptide and found that it dimerized with a dissociation constant in the micromolar range.

EXPERIMENTAL PROCEDURES

General Methods—The oligonucleotides corresponding to peptides of random sequence were cloned into the Ncol and BamHI sites of a previously synthesized (1) derivative of pSE380 (Invitrogen, San Diego, CA), pGBO10, that contained residues 169–292 of the AraC DNA binding domain and transformed into SH321 (ara-leu1022 Δlac74 galK str thi1) (5). DNA inserted in the Ncol and BamHI region is transcribed under control of the lac regulatory system. Previously, a stop codon was inserted at the end of the AraC DNA binding domain region of pGBO10 (1), and DNA was prepared by using cesium chloride as described by Schleif and Wensink (6). Another pSE380-derived vector, pGBO07 (1), contained an in-frame coding region for the AraC DNA binding domain adjacent to the lac promoter of pSE380 and was used as the AraC DNA binding domain control construct.

Arabinose isomerase levels were assayed as described (6). Cells were grown to an \( A_{600} \) of 0.5–0.8 in M10 minimal salts, 0.2% L-arabinose, 20 \( \mu M \) glucose, 10 \( \mu M \) thiamine, 20 \( \mu M \) CaCl\(_2\), and 10 \( \mu M \) MgCl\(_2\), and 1 ml was withdrawn. All manipulations at the DNA level were done by conventional molecular biology techniques. All candidates were sequenced using the SequiTherm EXCEL™ II DNA sequencing kit from Epicentre Technologies.

Construction, Isolation, and Characterization of Ara" Candidates—The sequence of the dimerization domain of the full-length leucine zipper-AraC DNA binding domain fusion construct is MAKQRNRVETGQKVELTSNDRLRLKKEQVLSREDPLLRLGPRQLPESSL (the underlined sequence is of the dimerization domain of the minimal leucine zipper-AraC DNA binding domain). The amino acid sequence of the AraC linker region is ESLHPPMDNRV.

For the minimal leucine zipper-AraC DNA binding domain construct,
an oligonucleotide, R1, with the following sequence was synthesized: CAGGAAAACAGCATGGTTCAGTGGCAATGACGGCTGGC-GAGCCGCTGGAACCGTGAACCGTCACTGACGTCGGTGATCTCCGG-GCGTGATCCGTCGGTCGTCGTCGTCGTCGTCGTCGTCGTCGTCG- (the NcoI and BamHI restriction sites are underlined). For generating the high diversity library of peptide-AraC DNA binding domain, a DNA oligonucleotide (R2) with the sequence CAGGAAAACAGCATGGTTCAGTGGCAATGACGGCTGGC-GAGCCGCTGGAACCGTGAACCGTCACTGACGTCGGTGATCTCCGG-GCGTGATCCGTCGGTCGTCGTCGTCGTCGTCGTCGTCGTCGTCG- (the NcoI and BamHI restriction sites are underlined), and the randomized region contains 30 repeats of NNK, where N represents A/T/G/C and K represents G/T was synthesized at the 200 nM level by Integrated DNA Technologies Co. An oligonucleotide (R3), complementary to the 3' end of R1 and R2, ATGGAGCGACTCGGATCC, was also used.

For synthesis of the double-stranded insert, equimolar amounts of either R1 or R2 were mixed with R3 in buffer containing 50 mM KCl, 20 mM Tris-Cl (pH 8.3), 1.5 mM MgCl2, 0.01% gelatin, and 0.2 mM each dNTP. 0.2 units of Tq polymerase was added, and annealing and extension were performed with the following cycling parameters: 95 °C for 5 min followed by a 1 °C per 45 s drop in temperature to 25 °C and a final extension at 75 °C for 5 min. The double-stranded nature of the oligonucleotide was verified by polyacrylamide gel electrophoresis.

The double-stranded oligonucleotide of random sequence was treated with 5 µg/ml protease K in 0.01 Tris-Cl (pH 7.8), 5 mM EDTA, and 0.5% SDS at 56 °C for 30 min. The sample was extracted with an equal volume of phenol followed by ethanol precipitation and digested with BamHI and NcoI endonucleases and electrophoresed on an 8% agarose gel. The double digested fragment was purified from the agarose gel. The doubly digested fragment was loaded into standard double sector cells with charcoal-filled epon centerpieces and quartz windows. On the sample side of the cell, 112 µl of sample was loaded with 12 µl of fuorobuco-43. On the solvent side, 125 µl of the buffer used for dialysis of the sample was loaded. The samples were centrifuged in a Beckman XL-I analytical ultracentrifuge at 20 °C at 34,000, 41,000, 48,000, and 54,000 revolutions/minute. At an improved variant of the equation was used when speed was 3 h apart were unchanged. Since absorbance readings at 280 nm gave very low values, data were collected at a wavelength of 230 nm. The partial specific volume of a monomer, v, of the protein was calculated as 0.7236 ml/g, and the density of the solvent, ρ, was estimated as 1.01728 g/ml using the program SEDNTERP (9). M1, the molecular weight of the monomer-dimer system, the accuracy of the extinction coefficient at 230 nm, ε230 can be estimated from this value in combination with an experimentally measured A230/A280 ratio. Due to low concentrations, we were unable to obtain a sufficiently accurate measure of this ratio. Thus, we assumed ε230 = 6 × ε280. An average value measured on several other proteins that were recently studied here in the analytical ultracentrifuge. Since Kc is linearly dependent on ε230 and the molecular weight value is insensitive to ε280.

All data sets were simultaneously analyzed using the global nonlinear least squares program, NONLIN (10). While the variance for the single species fit was adequate (variance of fit = 2.31 × 10−5), an improved fit was obtained when the centrifugation data were fitted to a monomer-dimer model. The distribution of residuals for the monomer dimer fit was measured on several other proteins that were recently studied here in the analytical ultracentrifuge. Since Kc is linearly dependent on ε230 and the molecular weight value is insensitive to ε280.

RESULTS

Development of Selection for Oligomerization—Appending the 49-amino acid leucine zipper coiled-coil from C/EBP to the DNA binding domain of AraC yields an active protein (1). In this work, we sought to identify peptide sequences of 30 amino acids or fewer that would similarly dimerize the DNA binding domain of AraC. To optimize the genetic engineering and get...
We found seven AraC medium. From an estimated 10,000 transformants, strains as quantitated from arabinose isomerase levels in SH321 cells. The binding domain activates pBAD cells, and AraC cannot grow into colonies on minimal arabinose plates in type AraC, and cells containing the domain instead of wild type AraC, we expect that translation reinitiates at a start codon close to the stop codon (12). We tested this idea on candidate 1 by deleting the region upstream of the nonsense codon and presumptive start codon (Fig. 2). The resulting construct was as active as the original in activating transcription from pRAD. We attribute the presence of peptides longer than 30 residues to additional nucleotides that were mistakenly incorporated during the chemical synthesis of the DNA oligomer. Possibly, such oligonucleotides were present at low levels in the synthesized DNA, but the selection method enriches for longer elements.

If the majority of the candidate oligomerizing peptides utilized a unique structure, we might expect to find an unusual distribution of amino acid content. Fig. 3 shows the amino acid composition expected from our “random” oligonucleotides and the amino acid composition found, excluding the initiating methionine. Ala, Phe, Met, and Asn are overrepresented, and Pro and Thr are underrepresented.

**Table I**

| Control proteins | Dimerization domain | Growth time | Activation ability |
|------------------|---------------------|-------------|--------------------|
| Wild type AraC   | 168                 | 1           | 1.00               |
| Leucine zipper-DNA binding domain | 50                   | 1           | 0.90               |
| Directly connected DNA binding domain | NA                  | 1           | 1.25               |
| Truncated zipper-DNA binding domain | 31                   | 1           | 0.90               |
| AraC DNA binding domain | NA                  | >5          | 0.05               |

* Number of amino acid residues including the initiating codon.  
* Time required for a colony to grow to approximately 3 mm in diameter on minimal medium containing 0.2% arabinose.  
* Relative in vivo transcriptional activation abilities of the control strains as quantitated from arabinose isomerase levels in SH321 cells.  
* NA, not applicable.

**Table II**

| Candidate | Peptide length | Growth time | Protein expression |
|-----------|----------------|-------------|--------------------|
| AraC      | 168            | 1           | Yes                |
| 1         | 22             | 1           | No                 |
| His6-1    | 28             | 1           | Yes                |
| 2         | 32             | >3          | No                 |
| His6-2    | 38             | N.T.        | No                 |
| 3         | 32             | 2.5         | Yes                |
| 4         | 17             | 1.5         | No                 |
| 5         | 6              | ~3          | No                 |
| 6         | 17             | ~3          | No                 |
| 7         | 32             | >3          | No                 |

* Length of inserted peptide including initiating methionine.  
* Time required for a colony to grow to approximately 3 mm in diameter on minimal medium containing 0.2% arabinose. N.T., not tested.  
* As assayed by visibility of a band on 14% SDS polyacrylamide gels of whole cells extracts from cells grown in the presence of 1 mM isopropyl-1-thio-β-D-galactopyranoside.

**Table III**

| Candidate | Sequence |
|-----------|----------|
| 1         | MFASNLVPFHMMHKMNQSC-SCQ- |
| 2         | MEKRGWVLASASHAHRGAEWEKVKYVSVMH- |
| 3         | METFWIFIRRTYRRRNMAFAILFASDMMKC- |
| 4         | MTKSSLVVLFGANSTS-TPS- |
| 5         | MRRFDI- |
| 6         | MALEWNNKQVGLVGVG- |
| 7         | MAMMAPRNSMLQVLSATNSANMFYSFQVV- |

**Fig. 2. Sequence of candidate 1.** The codons encoded by the sequence if the original or new translational start site is used is indicated by bars above or below the sequence, respectively.

**Fig. 3. Comparison of observed (dark shaded bar) and expected (lighter bar) frequencies of occurrence of amino acid residues in the peptide regions of the AraC candidates.** The translation initiating methionine was not included in the analysis.
the synthesis of the chimeric proteins is under the control of a strong promoter and the coding region is preceded by a strong ribosome binding site. To assess protease sensitivity and stability of the peptide-AraC DNA binding domain constructs, we examined the levels of the fusion protein expression of the candidates in crude cell extracts. Candidate 3 showed overexpression of three truncated protein fragments instead of the expected 18-kDa protein. This suggests possible cleavage by a cellular protease. Except for the first candidate after the addition of six histidines, none of the other candidates showed overexpression of a stable peptide-AraC DNA binding domain protein.

**DNA Binding of the Chimeras**—Although none of the Ara<sup>+</sup> candidates dramatically overexpressed stable protein, it was possible to determine their DNA binding activity in cell extracts, because this assay requires only very small amounts of protein. This assay also allows estimation of the molecular weight of the DNA-binding protein and, hence, detection of several of the more likely artifacts that could masquerade as dimerizing peptides. Using whole cell extracts, we examined DNA binding for all candidates, except candidate 7, to the I<sub>1</sub>I<sub>2</sub> DNA template (Fig. 4). No DNA binding activity was observed for the AraC DNA binding domain itself. The truncated leucine zipper-AraC DNA binding domain protein induces transcriptional activation to 90% of wild type AraC levels in *vivo* but does not bind to the DNA in *vitro*. Candidate 2 also showed no binding to DNA, whereas the five other Ara<sup>+</sup> candidates bound stably to DNA in *vitro*.

In theory, a short peptide could oligomerize the AraC DNA binding domains by fortuitous association with an oligomeric protein. Such a possibility appears likely, since the interaction of two short peptides to form a stable oligomer seems difficult, whereas peptide-domain interactions are not infrequent (13). This association could create a high molecular weight protein complex that would cause an anomalously large shift of the DNA template in DNA migration retardation assays. Comparison of the DNA retardation rates of the candidates with the DNA retardation rates of other control proteins indicated, however, that none are associated with auxiliary proteins (Fig. 4).

**Purification and Sedimentation**—We chose candidates 1 and 2 for further analysis and inserted hexahistidine tags at their N termini for Ni<sup>2+</sup>-His<sub>6</sub> affinity purification. The first candidate then showed overexpression of a full-length protein. Apparently, the hexahistidine tag increased its overall stability, and we purified this candidate for further analysis. The protein behaved poorly, and only 20% of the total overexpressed protein was soluble. We were, however, able to obtain 0.2 mg of >95% pure peptide-AraC DNA binding domain protein from 1 liter of cell culture. The purified protein retained the same DNA binding activity as the unpurified and untagged protein.

Up to this point, no direct evidence of dimerization or oligomerization has been presented. As another alternative to direct dimerization, it is possible that peptide-monomers could bind to DNA independently but much more tightly than the AraC DNA binding domain alone. A simple test of this possibility is to decrease progressively the concentration of protein in a binding assay. At some concentration, DNA with a single bound monomer would then be observed. Experiments with purified candidate 1 (Fig. 5) showed no evidence for such DNA-monomer intermediates. Similar experiments with candidate 5 in crude extracts also showed no evidence for monomer binding (data not shown). We therefore conclude that at least these two proteins dimerize.

We performed sedimentation equilibrium experiments to examine the strength and nature of oligomerization of the purified protein. The best fit of the sedimentation data to a single ideal nonassociating model yields a predicted molecular mass of 29 kDa, much different from the 17.5-kDa molecular mass derived from the protein sequence. Fitting the data from three protein concentrations and four different centrifugation speeds to a monomer-dimer equilibrium (Fig. 6) not only produced a good fit; it also yielded a predicted monomer molecular mass of 17.9 ± 1.1 kDa, very close to the molecular mass predicted from the protein sequence. The dimerization equilibrium constant was 1.8 with a 67% probability of lying in the interval 0.8–3.4 μM. We did not try fitting to other models, because the monomer-dimer model gave an excellent fit to the data.
DISCUSSION

We replaced the dimerization domain of AraC with peptides of random sequence, and from the resulting library of randomized peptide-AraC DNA binding domain fusions we identified seven peptides of 32 residues or shorter that can confer activation from \( p_{BAD} \). One peptide could not be studied, but the remaining six appear to oligomerize.

In principle, the selection technique requiring growth on minimal arabinose could yield five different classes of peptides: 1) peptides that themselves self-associate to form homodimers or other higher order structures; 2) peptides that bind to the DNA binding domains, but for steric reasons the arms cannot bind cis, and thus cross-binding between two chimeric molecules generates a dimer; 3) peptides that bind other cellular structures that provide an oligomeric framework; 4) peptides that bind nonspecifically to DNA; and 5) peptides that stabilize the normally rather unstable DNA binding domain of AraC. Structurally, the existence of class two candidates appears to be incompatible with binding to direct repeat DNA half-sites. None of the candidates seems to be in class three, since the migration retardation assays showed no evidence of anomalously high retardation, and two of the candidates, candidates 1 and 5, were shown not to be in class four or five, since the possibility of monomeric binding was eliminated by titration experiments. Except for one candidate, the peptide-AraC chimeras found in this work all bound stably in vitro to DNA containing the AraC binding site consisting of the I₁ half-site, to which AraC binds tightly, and the I₂ half-site, to which AraC binds weakly. Comparisons of the DNA migration retardation rates with other control proteins confirmed that no higher order oligomeric structures other than the dimeric species were formed.

We purified one of our dimerizing domain-AraC domain proteins. Although it was not sparingly soluble, it was possible to obtain equilibrium sedimentation data that showed it to dimerize rather tightly, with a \( K_d \) of 1.8 \( \mu M \). In the selection experiments reported by Wang and Pabo (4), 15-mer dimerizing peptides were isolated. The fusion proteins exist as monomers rather tightly, with a \( K_d \) of 1.8 \( \mu M \), and thus cross-binding between two chimeric molecules generates a dimer.

The natural interdomain linker of AraC was retained in our constructs. It is thus possible that these 8 amino acids interact with the 22 amino acids of the added peptide and contribute to the dimerization of the chimeric protein. Direct experiments with peptide should resolve this issue. The occurrences of Ala, Phe, Met, Asn, Pro, and Thr in the collection of oligomerizing peptides we found were greater than random. The probabilities that these particular amino acids would individually have deviated as much or more than what we found were 3.5, 1.02, 0.37, 5.6, 0.27, and 3%, respectively. These numbers are derived by noting that since there were 151 total amino acids, excluding the initiating methionines, in the peptides, and 31 possible codons, the probability that the set contains \( n \) or more alaines, for which there are two possible codons in the 31, is as follows.

\[
\sum_{n=31}^{151} \binom{151}{n} \left( \frac{29}{31} \right)^n \left( \frac{2}{31} \right)^{151-n}
\]

Since the abnormal distribution extends across most of the peptides and because a number of amino acids are involved, the distribution probably reflects some general principles relevant to the association of short amino acid sequences. The actual sequences do not obviously reveal their secondary or tertiary structures. None looks like a leucine zipper coiled-coil, although the program PredictProtein (14–16) does predict that 16 contiguous residues out of 22 in the peptide of candidate 1 form an \( \alpha \)-helix. Clearly, additional studies on both the selection method and the peptides that are found should yield much interesting information.

Acknowledgment—We thank Ula Gryczynski for assistance.

REFERENCES

1. Bustos, S. A., and Schleif, R. F. (1993) Proc. Natl. Acad. Sci. U. S. A. 90, 5638–5642
2. Zhang, Z., Murphy, A., Hu, J. C., and Kodadek, T. (1999) Curr. Biol. 9, 417–420
3. Japelli, R., and Brenner, S. (1999) Biochem. Biophys. Res. Commun. 266, 243–247
4. Wang, B. S., and Pabo, C. O. (1999) Proc. Natl. Acad. Sci. U. S. A. 96, 9568–9573
5. Hahn, S., Dunn, T., and Schleif, R. F. (1984) J. Mol. Biol. 186, 60–72
6. Schleif, R. F., and Wensink, P. (1981) Practical Methods in Molecular Biology, Springer-Verlag, New York
7. Hendrickson, W., and Schleif, R. F. (1984) J. Mol. Biol. 178, 611–628
8. Bradford, M. (1976) Anal. Biochem. 72, 248–254
9. Laue, T. M., Shah, B. D., Ridgeway, T. M., and Pelletier, S. L. (1992) in Analytical Ultracentrifugation in Biochemistry and Polymer Science (Harding, S., and Rowe, A., eds) Royal Society of Chemistry
10. Johnson, M. L., Correia, J. J., Yphantis, D. A., and Halvorson, H. R. (1981) Biophys. J. 36, 575–588
11. Pace, C. N., Vajdos, F., Fee, L., Grimsley, G., and Gray, T. (1995) Protein Sci. 4, 2411–2423
12. Fikes, J. G., Weber, K., and Miller, J. H. (1974) Proc. Natl. Acad. Sci. U. S. A. 71, 667–670
13. Schleif, R. F. (1999) Proteins 34, 1–3
14. Rost, B., and Sander, C. (1993) J. Mol. Biol. 232, 584–599
15. Rost, B., and Sander, C. (1994) Proteins 20, 216–226
16. Rost, B. (1996) Methods Enzymol. 266, 525–539