Consolidating critical binding determinants by noncyclic rearrangement of protein secondary structure

Ramon K. Tabtiang*, Brent O. Cezairliyan, Robert A. Grant, Jesse C. Cochrane†, and Robert T. Sauer‡

Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139

Contributed by Robert T. Sauer, December 21, 2004

We designed a single-chain variant of the Arc repressor homodimer in which the β strands that contact operator DNA are connected by a hairpin turn and the α helices that form the tetrahelical scaffold of the dimer are attached by a short linker. The designed protein represents a noncyclic permutation of secondary structural elements in another single-chain Arc molecule (Arc-L1-Arc), in which the two subunits are fused by a single linker. The permuted protein binds operator DNA with nanomolar affinity, refolds on the sub-millisecond time scale, and is as stable as Arc-L1-Arc. The crystal structure of the permuted protein reveals an essentially wild-type fold, demonstrating that crucial folding information is not encoded in the wild-type order of secondary structure. Noncyclic rearrangement of secondary structure may allow grouping of critical active-site residues in other proteins and could be a useful tool for protein design and minimization.

circular permutation | protein folding | protein structure

Imagine a protein in which the α -helices and β -strands are disconnected but otherwise arranged properly in three dimensions, so that the hydrophobic core and most native interactions remain intact. There are probably many ways to connect the secondary-structure elements that are compatible with stable folding. For example, cyclic permutation of secondary structure (Fig. 1A) is tolerated in numerous proteins and occurs naturally in some protein families (1–10). After cyclic permutation, a wild-type block of N-terminal sequence is shifted to the C terminus of the permuted protein. Three-dimensional domain swapping, in which a monomeric protein fold is reproduced with structural parts donated by different subunits in a multimer (Fig. 1B), provides another instance of rearranged structural elements (11, 12). These examples suggest that the essential feature of a protein fold is the complementary packing of secondary structural elements and not the precise manner in which these elements are connected. If this model is correct, then many permutations of protein structural elements should be allowed. However, to our knowledge, no successful examples of noncyclic permutations of protein structural elements have been reported.

Arc repressor from bacteriophage P22 is a small transcription factor that folds as a symmetric homodimer. Two Arc dimers bind cooperatively to adjacent subsites in *arc* operator DNA (13–16). Each Arc subunit consists of a β -strand and two α -helices (α A and α B); the packing of these elements from both subunits of the dimer forms a single hydrophobic core. The β -strands pair to form an antiparallel β -sheet, which contacts the major groove of operator DNA. A flexible N-terminal arm and residues near the beginning of helix B make other DNA contacts. Joining two Arc subunits with a 15-residue linker results in a single-chain molecule (Arc-L1-Arc) that is more stable than wild-type Arc and binds operator DNA at lower protein concentrations (17, 18). The wild-type order of secondary structure is preserved in Arc-L1-Arc, with each β -strand followed by α A and then α B. Here, we show that the secondary structure of single-chain Arc can be rearranged

in a noncyclic fashion with retention of structure and operator-binding activity and without significant changes in stability or folding kinetics. Noncyclic permutation should be a useful addition to the toolkit of protein design and engineering.

Methods

Molecular Biology. The gene for the permuted pArc molecule was constructed in a plasmid system by using a combination of PCR amplification of portions of the *arc-st11* gene from pSA700 and synthesis of oligonucleotide cassettes for the linker and β -strand regions. The structure of the final construct was verified by DNA sequencing (Massachusetts Institute of Technology Biopolymer Facility).

Protein Purification. *Escherichia coli* XL1Blue cells (Stratagene) containing the overexpression plasmid were grown at 37°C to an OD₆₀₀ of 0.5 in 1 liter of LB broth plus 100 μ g/ml ampicillin. Isopropyl β -D-thiogalactoside was added (0.1 g/liter), and cells were harvested after an additional 2 h of growth. Cells were lysed in buffer A [6 M guanidine hydrochloride (Gdn·HCl)/50 mM sodium phosphate/10 mM Tris/5 mM imidazole, pH 8.0]. The lysate was cleared by centrifugation and loaded onto a 1-ml Ni²⁺-NTA column (Qiagen, Valencia, CA) equilibrated in buffer A. The column was washed with buffer A and eluted with 10 ml of 6 M Gdn·HCl/0.2 M acetic acid. The eluate was dialyzed overnight against 4 liters of buffer B (25 mM sodium phosphate/5% glycerol, pH 6.0) at 4°C and loaded on a SP-Sephadex column equilibrated in buffer B. The column was washed with buffer B, and pArc protein was eluted with 10 mM Tris, pH 8.0/2 M NaCl/0.2 mM EDTA. At this point, pArc protein was >95% pure. For storage, pArc was dialyzed extensively against water and lyophilized. The concentration of pArc was determined by absorbance using an extinction coefficient at 280 nm of 15,312 M⁻¹·cm⁻¹.

Crystallography. Crystals of pArc or selenomethionine-substituted pArc were grown at 20°C by hanging-drop vapor diffusion. Lyophilized protein was dissolved in 10 mM Hepes·KCl (pH 7.5) at a concentration of 10 mg/ml, and 1 μ l of protein solution was mixed with 1 μ l of well solution. Well solutions containing 17.5–30% polyethylene glycol (PEG) 3350 and 100–250 mM KI yielded visible crystals within 2 days. The native crystal used to collect the 1.9-Å data set was grown under similar conditions, but with NaF substituting for KI. Crystals were frozen in liquid nitrogen without addition of

Abbreviation: Gdn·HCl, guanidine hydrochloride.

Data deposition: The atomic coordinates have been deposited at the Protein Data Bank, www.pdb.org (PDB ID code 1U9P).

*Present address: Fish and Richardson PC, 225 Franklin Street, Boston, MA 02110.

†Present address: Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520.

‡To whom correspondence should be addressed. E-mail: bobsauer@mit.edu.

© 2005 by The National Academy of Sciences of the USA

Fig. 1. Permuted protein secondary structure. (A) Cartoon of cyclic permutation. (B) Cartoon of domain swapping. (C) The secondary structure in pArc is a noncyclic permutation of the structure in single-chain Arc-L1-Arc. The order of structural elements in Arc-L1-Arc is shown in the upper portion of the panel with the first Arc repeat colored blue and the second repeat colored green. The sequence of pArc is shown in the bottom portion of the panel, with new linker sequences colored red and sequence elements from Arc-L1-Arc colored blue and green as in the upper cartoon. The C-terminal sequence of pArc (black) includes a H₆-affinity tag and charged and polar residues to prevent intracellular degradation (23).

cryoprotectant. The native data set was collected on a Rigaku rotating anode source equipped with Yale/MSU mirrors and a R-Axis IV detector. Selenomethionine (SeMet) crystal data were collected at Advanced Photon Source beamline 8BM (NE-CAT, Argonne, IL). Anomalous diffraction data were collected at the peak wavelength of the x-ray absorption spectrum. Data were integrated and scaled by using HKL2000 or DENZO/SCALEPACK (19). CNS (20) was used to locate two selenium sites by Patterson correlation methods, calculate phases from these sites, and refine phases by density modification. The resulting phases produced a 3-Å electron-density map in which the overall features of the Arc molecule were evident and a difference-Fourier map that clearly indicated the position of a third selenium site. The model was built by using O (21) and refined with CNS against the 2.0-Å SeMet data set. The final cycles of refinement by CNS were performed by using the 1.9-Å native data after rigid body refinement of the model into the nearly isomorphous cell of the native crystal.

Protein Chemistry and DNA Binding. For equilibrium denaturation, 5 μM pArc in S buffer (50 mM Tris-HCl, pH 7.5/250 mM KCl/0.2 mM EDTA) was mixed with increasing amounts of 5 μM pArc in S buffer plus 6 M Gdn-HCl, the mixture was allowed to equilibrate at 25°C, and circular-dichroism ellipticity at 234 nm was averaged for 60 s by using an Aviv 60DS instrument. Refolding kinetics at 25°C were monitored by changes in tryptophan fluorescence (excitation, 280 nm; emission, >300 nm) by using an Applied Photosystems stopped-flow instrument after mixing of 50 μM pArc in S buffer plus 6 M Gdn-HCl with a 5-fold excess of S-buffer alone or S-buffer containing lower concentrations of Gdn-HCl. Analytical ultracentrifugation (25,000 rpm at 20°C) was performed by using a Beckman Optima XL-A instrument. The pArc protein (64 μM initial concentration) in S buffer was centrifuged until equilibrium was established, and 10 scans at 280 nm were averaged and analyzed by standard methods (22). Binding of pArc to a ³²P-labeled *arc* operator duplex was assayed by gel mobility-shift experiments as described (14).

Results

Design. As shown in Fig. 1C, the order of structural elements in single-chain Arc-L1-Arc is arm-β-αA-αB-linker-arm-β-αA-αB (17). We reengineered the sequence to encode a

permuted single-chain protein (pArc) with structural elements rearranged in the order arm-β-linker-β-αA-αB-linker-αA-αB (Fig. 1C). In pArc, unlike Arc-L1-Arc, the β strands that contact operator DNA are now positioned together in the N-terminal part of the single-chain molecule, whereas the C-terminal portion comprises the tetra helical scaffold. Another consequence of the pArc design is deletion of the second arm sequence in Arc-L1-Arc. Simple model building suggested that a three-residue GGG sequence could form a hairpin turn between the β-strands in pArc and an eight-residue sequence GTGGSGGG would suffice to connect residues C-terminal to the first αB helix with the N terminus of the second αA helix. To facilitate expression and purification, a H₆KNQHE sequence was appended to the C terminus of the pArc construct (23).

Expression and Structure. The pArc protein was expressed in *E. coli* and was soluble and well behaved during purification. Purified pArc sedimented as a monomer during equilibrium analytical ultracentrifugation (Fig. 2A). The pArc protein showed no propensity to form higher oligomers, indicating that swapping of structural elements from different pArc molecules must not occur to any significant extent.

We crystallized pArc and solved the structure (Fig. 3A) by single-wavelength anomalous dispersion (SAD) phasing of a selenomethionine derivative. The final structure was refined to 1.9-Å resolution with *R* and *R*_{free} values of 0.249 and 0.297 (Table 1). Except for the designed permutations and linkers, the structure of pArc was very similar to the wild-type Arc dimer, with a 1.04-Å rms deviation for common main-chain atoms. Electron density was not observed for the middle of the GTGGSGGG linker, and only the terminal residues of this linker could be traced. The GGG linker connecting the β-strands of pArc was visible, but had high thermal factors. Relative to wild-type Arc, the structures of the β sheet and α helices in pArc were generally unperturbed, except for small distortions in the position of the second copy of helix A (Fig. 3B and C). The N terminus of this helix moves several angstroms to avoid steric clashes with the engineered linker between the pArc β strands and/or because it is not constrained by the wild-type linkage to a β-strand. Overall, however, the crystal structure demonstrates that the noncyclic permutation of secondary structure in pArc is tolerated with

Fig. 2. Folding and stability of pArc. (A) Analytical equilibrium ultracentrifugation shows that pArc behaves as a single solution species with a calculated molecular mass (12.2 kDa) close to that expected for the monomer (13.7 kDa). (B) pArc and Arc-L1-Arc have similar equilibrium stabilities to Gdn-HCl denaturation. The fits are for two-state transitions between native and denatured protein. In the absence of denaturant, ΔG_u was 5.9 kcal/mol for pArc and 6.0 kcal/mol for Arc-L1-Arc; the m -values were 2.4 and 2.5 kcal/mol, respectively. (C) Refolding of pArc after a jump from 7 to 2.4 M Gdn-HCl. The residuals of the fit to a single exponential function with a refolding rate constant of 14.2 s^{-1} are shown. (D) Semilog plot of pArc refolding rate constants as a function of denaturant concentration. The fit is to the equation $y = 9,816 \times 10^{-0.99 \times x}$.

minimal changes in tertiary structure. Although the general fold of pArc closely resembles that of wild-type Arc and related repressors, the topology of secondary structural elements in pArc is unique.

Stability and Folding Kinetics. Rearranging the order of secondary structural elements in pArc affected neither global stability nor the ability of the permuted protein to fold extremely rapidly to its native conformation. Incubation of pArc or Arc-L1-Arc with increasing concentrations of Gdn-HCl resulted in similar equilibrium denaturation curves (Fig. 2B), with fitted ΔG and m values within experimental error. After a jump from 7.1 to 2.4 M Gdn-HCl, pArc refolding was complete in $\approx 0.2 \text{ s}$ (Fig. 2C). The refolding rate constant calculated by extrapolation of experimental values to zero denaturant was $\approx 9,800 \text{ s}^{-1}$ (Fig. 2D), corresponding to a refolding half-life of 70 μs . Arc-L1-Arc refolds at essentially the same rate (18). We conclude that the precise order of secondary structure elements in these single-chain proteins is not a critical factor in determining the energy landscape for folding.

DNA-Binding Activity. The structural changes in pArc did not preclude its binding with high affinity to an *arc* operator DNA fragment, as assayed by gel-mobility shift assays (Fig. 4). As observed for wild-type Arc (14, 15), operator binding was highly cooperative, and intermediates corresponding to occupancy of a single operator subsite were not observed. The concentration of

pArc required for half-maximal binding was $\approx 5 \text{ nM}$. This value is higher than that observed with wild-type Arc (0.25 nM; ref. 14) or with the single chain Arc-L1-Arc molecule (1.7 pM; ref. 17). We presume that this weaker operator binding by pArc results from the minor structural distortions and because it contains only a single N-terminal arm, whereas the wild-type dimer or Arc-L1-Arc have two arms. Both arms in wild-type Arc make numerous operator contacts and contribute significantly to binding affinity (16, 24).

Discussion

The Arc fold was maintained after noncyclic rearrangement of its secondary structural elements, which required the addition of two nonnatural linkers. Despite these changes, the single-chain pArc molecule had the same thermodynamic stability as the nonpermuted Arc-L1-Arc protein, refolded in the sub-millisecond time regime, and bound operator DNA with nanomolar affinity.

Noncyclic permutation of protein structure may be a useful addition to the tools of protein engineering and design, which include directed mutagenesis, protein fusions, computational identification of sequences that adopt a specified target fold, and experimental isolation of desired variants by randomization and library selection technologies (25–29). Residues from distant regions of a sequence often form binding surfaces and active sites in proteins. This is the case for Arc, where the β -sheet that mediates DNA binding is formed by β -strands that are >60 residues apart in the normal single-chain protein. This distance presents a problem for mutagenesis/selection experiments, because highly efficient randomization methods, like cassette mutagenesis, are difficult

Fig. 3. Crystal structure of pArc. (A) Ribbon representation of pArc structure (colored blue $\rightarrow$ red from N $\rightarrow$ C terminus). In wild-type Arc, the antiparallel β -sheet contacts the major groove of operator DNA (16). (B and C) The pArc structure is compared with the wild-type Arc homodimer (A-subunit in gray; B-subunit in brown; Protein Data Bank ID code 1PAR). Residues 19–62 of pArc, which have wild-type sequence and topology, were aligned with the A-subunit of wild-type Arc. This maximizes structural differences between wild type and the remaining portions of pArc, which are most dissimilar in terms of secondary-structure topology. Dotted lines represent sequence linkage not visible in electron-density maps.

when the target codons are far apart. However, in the pArc molecule, the entire DNA-binding β sheet is encoded in a short stretch of contiguous residues. As a result, complete cassette randomization of all of the DNA contact residues would be straightforward. Noncyclic permutation may be an especially useful

tool in protein minimization (30, 31), where design and selection methods are combined with the goal of maintaining functionality with the fewest number of residues possible.

Proteins are generally remarkably robust to amino acid substitutions, cyclic permutations, insertions, and even small deletions (1–4, 6–9, 32–35). The studies reported here add noncyclic permutation to the list of seemingly abusive sequence manipulations that proteins can tolerate. However, not all cyclic permutations are tolerated, and it seems likely that noncyclic permutations will be accepted at somewhat lower frequencies because the sequence changes and linker additions represent more dramatic alterations. Tsuji *et al.* (36) did not find noncyclic permutants of barnase that folded like the wild-type protein, but they made no attempt to link permuted structural elements in a fashion compatible with wild-

Table 1. Crystallographic data, phasing, and refinement statistics

Crystal forms		
SeMet crystal	Space group P2 ₁ <i>a</i> = 30.0 Å, <i>b</i> = 62.6 Å, <i>c</i> = 29.1 Å, β = 104.4° One monomer per asymmetric unit	
Native crystal	Space group P2 ₁ <i>a</i> = 29.2 Å, <i>b</i> = 62.9 Å, <i>c</i> = 30.4 Å, β = 102.7° One monomer per asymmetric unit	
Data collection		
Data sets	Se peak	Native
Wavelength, Å	0.979	1.541
Resolution, Å	50–2.0	20–1.9
<i>R</i> _{sym} , %	4.9 (18.0)	6.7 (19.5)
Completeness, %	96.4 (80.9)	95.0 (75.8)
No. of observations	61,531	34,489
No. of unique hkl	13,545	8,069
Average <i>I</i> /σ	16.2	11.5
Phasing (data truncated to 2.5 Å in cns)		
Anomalous differences, %		3.6
Number of Se sites used for phasing		3
Figure of merit		0.359 (0.263)
Refinement		
<i>R</i> , %	24.9	
<i>R</i> _{free} , %	29.7	
rms deviation		
Bonds, Å	0.006	
Angles, degrees	1.1	

SeMet, selenomethionine.

pArc concentration (M)

Fig. 4. Binding of pArc to a 21-bp *arc* operator DNA fragment ($\approx 10 \text{ pM}$) assayed by gel electrophoretic mobility shifts. The fraction of operator DNA bound varied with the square of the pArc concentration with half saturation at $\approx 5 \text{ nM}$ protein.