

Domain-Swapped Dimeric Structure of a Stable and Functional *De Novo* Four-Helix Bundle Protein, WA20

Ryoichi Arai,^{*,†,‡,§} Naoya Kobayashi,[§] Akiho Kimura,^{§,⊥} Takaaki Sato,[†] Kyoko Matsuo,[§] Anna F. Wang,^{‡,▽} Jesse M. Platt,^{‡,#} Luke H. Bradley,^{‡,||} and Michael H. Hecht^{*,‡}

[†]International Young Researchers Empowerment Center, Shinshu University, Ueda, Nagano 386-8567, Japan

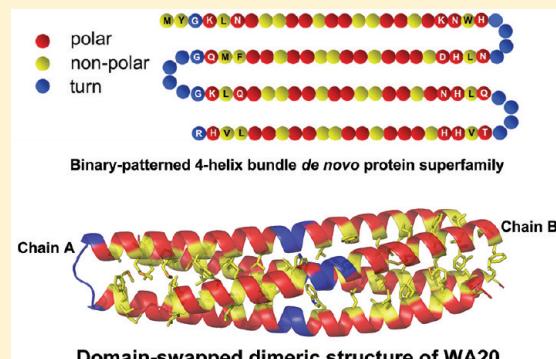
[‡]Department of Chemistry, Princeton University, Princeton, New Jersey 08544, United States

[§]Department of Applied Biology, Faculty of Textile Science & Technology, Shinshu University, Ueda, Nagano 386-8567, Japan

^{||}Departments of Anatomy & Neurobiology, Molecular & Cellular Biochemistry, Center of Structural Biology, University of Kentucky College of Medicine, Lexington, Kentucky 40536, United States

Supporting Information

ABSTRACT: To probe the potential for activity in unevolved amino acid sequence space, we created a third generation combinatorial library of *de novo* four-helix bundle proteins. The “artificial superfamily” of helical bundles was designed using binary patterning of polar and nonpolar residues, and expressed in *Escherichia coli* from a library of synthetic genes. WA20, picked from the library, is one of the most stable proteins in the superfamily, and has rudimentary activities such as esterase and lipase. Here we report the crystal structure of WA20, determined by the multiwavelength anomalous dispersion method. Unexpectedly, the WA20 crystal structure is not a monomeric four-helix bundle, but a dimeric four-helix bundle. Each monomer comprises two long α -helices that intertwist with the helices of the other monomer. The two monomers together form a 3D domain-swapped four-helix bundle dimer. In addition, there are two hydrophobic pockets, which may potentially provide substrate binding sites. Small-angle X-ray scattering shows that the molecular weight of WA20 is ~ 25 kDa and the shape is rod-like (the maximum length, $D_{\max} = \sim 8$ nm), indicating that WA20 forms a dimeric four-helix bundle in solution. These results demonstrate that our *de novo* protein library contains not only simple monomeric proteins, but also stable and functional multimeric proteins.



INTRODUCTION

The field of *de novo* protein design and engineering is motivated by two considerations: (i) Recapitulation of natural systems to ultimately test our understanding of biological systems including protein structure and function, and (ii) construction of novel, “tailor-made” proteins as an essential step toward future advances in biotechnology and synthetic biology. Progress toward the construction of novel proteins emanates mainly from two approaches: rational design^{1,2} and combinatorial methods.³ We have developed a semirational approach that incorporates elements of both rational design and combinatorial methods to produce focused libraries of novel proteins.^{4,5} In these libraries, the hydrophobic or hydrophilic nature of each amino acid side chain is rationally designed on the basis of the template structure of a globular protein, but the exact identities of the individual polar and nonpolar residues are varied combinatorially. Using this binary code strategy, we have constructed several libraries of *de novo* α -helical or β -sheet proteins.^{4–8} For example, the binary patterned design of amphipathic α -helical sequences places a hydrophobic amino acid every three or four residues in accordance with its secondary structure periodicity of 3.6

residues/turn, thereby generating the following pattern: O●O○●●○○●○○●●○, where O and ● represent polar and nonpolar residues, respectively.^{4,5} When four such helices are linked, the hydrophobic effect drives them to form a four-helix bundle with nonpolar residues forming a hydrophobic protein core and polar residues oriented to the aqueous solvent.

Using this approach, we have constructed three libraries of binary patterned four-helix bundles. The first library encoded 74-residue sequences and produced structures that were moderately stable but mostly dynamic.^{4,9} A second generation 102-amino acid residue library was designed to be well-folded, by increasing the number of hydrophobic core residues, similar to that observed in naturally occurring four-helix bundles.⁷ Of the five proteins characterized from this second generation library, four were shown to form stable well-ordered structures, and two of these structures, S-824 and S-836, were shown by

Special Issue: Harold A. Scheraga Festschrift

Received: December 23, 2011

Revised: March 2, 2012

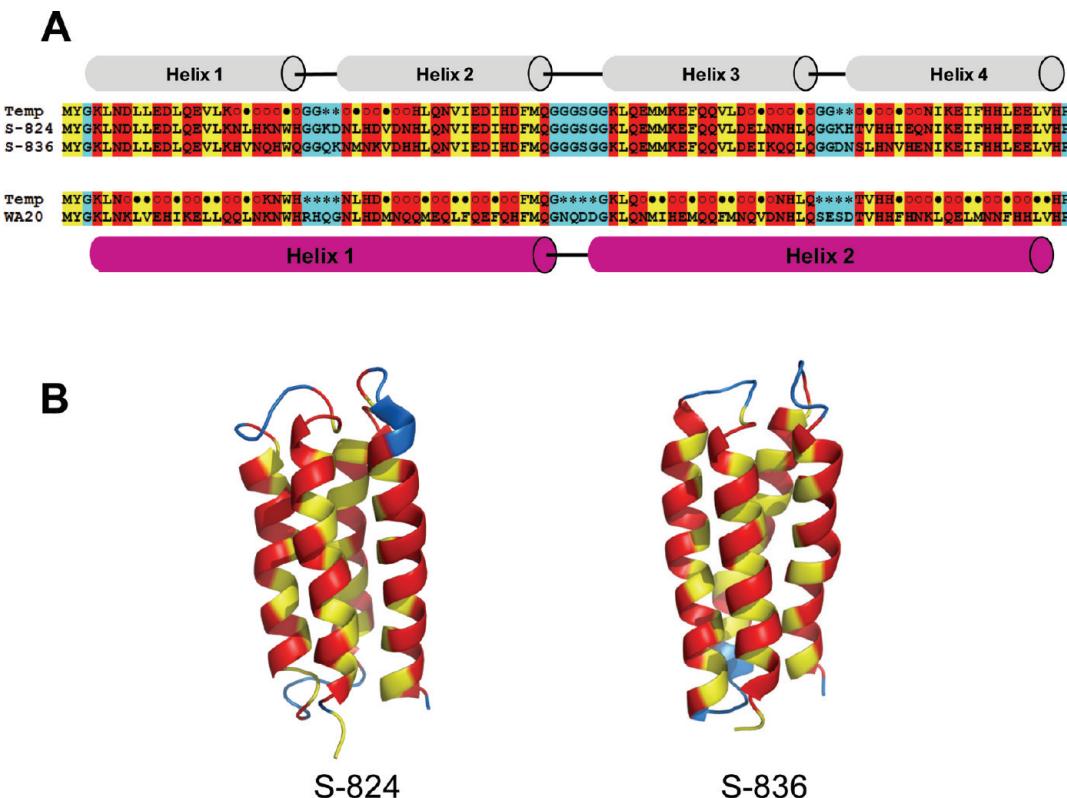


Figure 1. *De novo* four-helix bundle proteins. (A) Amino acid sequences of binary patterned four-helix bundle *de novo* proteins. Top: Design template for second generation library and the amino acid sequences of S-824 and S-836.⁷ The cylindrical shapes above the top represent designed regions of four helices. Bottom: Design template for third generation library and the amino acid sequences of WA20.¹² The magenta cylindrical shapes below the bottom represent two helices in the crystal structure of WA20 (chain A). The sequences follow the binary pattern design with red indicating polar residues and yellow indicating nonpolar residues. Turn residues are highlighted in blue. In the template (Temp) sequences, O, ●, and * represent polar, nonpolar, and turn residues, respectively (Figure S1, Supporting Information). (B) Ribbon representation of S-824 (PDB: 1P68)¹⁰ and S-836 (PDB: 2JUA)¹¹ structures determined by NMR. The color coding is the same as part A.

nuclear magnetic resonance (NMR) spectroscopy to form monomeric four-helix bundles with the hydrophobic residues sequestered in the core and the hydrophilic residues exposed on the exterior, as designed (Figure 1).^{10,11}

On the basis of this 102 amino acid template, a high-quality third generation combinatorial library comprising $\sim 10^6$ *de novo* four-helix bundle proteins was constructed to probe functional activity in unevolved amino acid sequence space.^{6,12} (Figure 1A and Figure S1 in the Supporting Information). Recently, we demonstrated that proteins from this third generation library of binary patterned four-helix bundles are rich in rudimentary enzymatic activities *in vitro*¹² and have the ability to restore cell growth in auxotrophic *E. coli* *in vivo*,¹³ thereby establishing the functional potential of an unevolved artificial superfamily of proteins. During our initial characterization of highly expressed sequences arbitrarily chosen from this collection of novel functional proteins, we identified WA20 as one of the most stable structures with a cooperative guanidine hydrochloride (GdnHCl) denaturation midpoint value (3.8 M) significantly higher than those determined for second generation proteins S-824 (3.2 M) and S-836 (3.0 M), respectively.^{7,14,15} (Bradley et al., unpublished). Moreover, WA20 has rudimentary activities as a peroxidase (with bound heme cofactor), esterase, and lipase.¹² To analyze the structural details of *de novo* protein WA20, we solved the crystal structure using multiwavelength anomalous dispersion (MAD).¹⁶ Here we present the novel structure and discuss its features.

EXPERIMENTAL METHODS

Protein Expression and Purification. WA20 was identified from a library of genes encoding the *de novo* proteins, as described previously (Figure 1A).^{6,12} The WA20 gene was cloned into an isopropyl β -D-1-thiogalactopyranoside (IPTG)-inducible protein expression vector, pET-3a (Novagen, MERCK, Darmstadt, Germany), with the T7 promoter and ampicillin resistance. WA20 protein was expressed in *E. coli* BL21 Star(DE3) (Invitrogen, Carlsbad, CA) using 2 L of LB medium at 30 °C. The selenomethionine (SeMet)-labeled WA20 protein was expressed in the methionine auxotroph *E. coli* B834(DE3) strain (Novagen, MERCK) using 2 L of LeMaster medium¹⁷ with 100 mg of L-SeMet at 30 °C. Expression was induced with 0.2 mM IPTG (at OD₆₀₀ = ~0.5), and cells were further cultured for 16 h at 30 °C. The protein was extracted from harvested cells using the freeze–thaw method¹⁸ in lysis buffer (50 mM sodium phosphate (pH 7.0), 300 mM NaCl, 10% glycerol). The protein was purified by immobilized metal ion affinity chromatography with TALON metal affinity resin (Clontech, Takara Bio, Mountain View, CA) according to the manufacturer's protocols (equilibration/wash buffer: 50 mM sodium phosphate buffer (pH 7.0) containing 300 mM NaCl; elution buffer: 50 mM sodium phosphate buffer (pH 7.0) containing 300 mM NaCl, 10% glycerol, and 250 mM imidazole). Even without a His-tag, the WA20 protein binds to a TALON metal affinity resin, presumably because of the relatively high percentage (12.7%) of histidine residues in its

sequence (Table S1, Supporting Information). WA20 was further purified by cation exchange chromatography (25 mM MES buffer (pH 6.0) containing 10% glycerol, with a linear gradient of NaCl from 0 to 1.5 M) with a Poros HS/M column (Perseptive Biosystems) and gel filtration chromatography (25 mM HEPES buffer (pH 7.0) containing 100 mM NaCl, 10% glycerol, and 1 mM dithiothreitol (DTT)) with a Superdex 75 10/300 GL column (GE healthcare, Little Chalfont, U.K.). The SeMet-labeled WA20 protein was used for crystallization, and the native WA20 protein was used for the other experiments.

Crystallization. The crystals of the SeMet-labeled WA20 protein were obtained in a drop composed of 0.5 μ L of the protein solution and 0.5 μ L of the reservoir solution (0.056 M sodium phosphate monobasic monohydrate, 1.344 M potassium phosphate dibasic, pH 8.2) by the sitting drop vapor diffusion method against 50 μ L of the reservoir solution at 4 °C. Rod-like crystals were obtained in several weeks (Figure S2, Supporting Information).

Data Collection, Structure Determination, and Refinement. The X-ray diffraction data were collected at the Photon Factory, BL-SA (KEK, Tsukuba, Japan). The data collection was carried out at 95 K with a mixture of equal parts of Paratone-N (Hampton Research, Aliso Viejo, CA) and paraffin oil as a cryoprotectant. All diffraction data were processed with the program HKL2000¹⁹ (Table 1).

The program SOLVE²⁰ was used to locate the selenium sites and to calculate the phases by the MAD method, and the program RESOLVE²¹ was used for the density modification and partial model building. The model was built and corrected with the program COOT²² and was refined with the program REFMAC5^{23,24} in the CCP4 suite.²⁵ All refinement statistics are presented in Table 1. The quality of the model was inspected by the programs PROCHECK²⁶ and MolProBity^{27–29} (Figure S3, Supporting Information). The atomic coordinates and the structure factors have been deposited in the Protein Data Bank, with the accession code 3VJF. The graphic figures were created using the program PyMOL (DeLano Scientific LLC).

Small-Angle X-ray Scattering (SAXS). We performed SAXS measurements on WA20 (5.4 mg/mL) to examine the static structure in solution (25 mM HEPES buffer (pH 7.0) containing 100 mM NaCl, 10% glycerol, and 1 mM DTT). We used a SAXSess camera (Anton Paar, Graz, Austria) attached to a sealed tube anode X-ray generator (GE Inspection Technologies, Germany). The generator was operated at 40 kV and 50 mA. A Göbel mirror and a block collimator provide a focused monochromatic X-ray beam of Cu K α radiation ($\lambda = 0.1542$ nm) with a well-defined shape. A thermostatted sample holder unit (TCS 120, Anton Paar) was used to control the sample temperature. The two-dimensional scattering patterns recorded by an imaging-plate (IP) detector (Cyclone, Perkin-Elmer) were integrated into one-dimensional scattered intensities, $I(q)$, as a function of the magnitude of the scattering vector $q = (4\pi/\lambda) \sin(\theta/2)$ using SAXSQuant software (Anton Paar), where θ is the total scattering angle. For all experiments, we monitored the attenuated primary beam at $q = 0$ using a semitransparent beam stop. All the measured intensities were semiautomatically calibrated for transmission by normalizing a zero- q primary intensity to unity. The background scattering contributions from capillary and solvent were corrected. The absolute intensity calibration was made by using water intensity as a secondary standard.³⁰

Table 1. X-ray Data Collection and Refinement Statistics

	peak	edge	remote
Data Collection ^a			
space group	$P_{2_1}2_12$		
unit-cell parameters (Å)	$a = 65.95$		
	$b = 102.86$		
	$c = 31.34$		
	$\alpha = \beta =$		
	$\gamma = 90.00^\circ$		
wavelength (Å)	0.97881	0.97908	0.90000
resolution (Å)	50.0–2.20 (2.28–2.20)	50.0–2.20 (2.28–2.20)	50.0–2.20 (2.28–2.20)
unique reflections	11102	11112	11028
average redundancy	6.1 (5.5)	6.1 (5.3)	6.1 (5.3)
completeness (%)	97.4 (85.3)	97.1 (83.6)	96.2 (79.1)
$I/\sigma(I)$	11.6 (5.4)	15.3 (4.8)	12.5 (3.1)
R_{sym}^b (%)	9.8 (30.6)	7.9 (32.0)	8.4 (36.8)
MAD Analysis			
resolution (Å)	50.0–2.20		
no. of Se sites ^c	13		
FOM _{MAD} ^d	0.63		
FOM _{RESOLVE} ^e	0.76		
Refinement			
resolution (Å)	50.0–2.20		
no. of reflections	10479		
no. of protein atoms	1635		
no. of water molecules	55		
no. of other atom	1		
R_{work} (%)	23.3		
R_{free}^f (%)	25.5		
rmsd bond length (Å)	0.009		
rmsd bond angle (deg)	1.17		
average B-factor ^g (Å ²)	51.2		
Ramachandran Plot ^h			
favored regions (%)	99.5		
allowed regions (%)	0.5		
disallowed regions (%)	0.0		

^aAll numbers in parentheses represent last outer shell statistics. ^b $R_{\text{sym}} = \sum |I_i - I_{\text{avg}}| / \sum I_i$, where I_i is the observed intensity and I_{avg} is the average intensity. ^cNumber of selenium sites located using SOLVE.

^dFigure of merit after SOLVE phasing. ^eFigure of merit after RESOLVE density modification. ^f R_{free} is calculated for 5% of randomly selected reflections excluded from refinement. ^gAverage B-factor is average of sum of TLS (Translation, Libration, and Screw-rotation) and residual B-factors. ^hThe Ramachandran plot is shown in Figure S3 in the Supporting Information.

Assuming the structure factor $S(q) = 1$ for dilute samples, $I(q)$ is given by Fourier transformation of the so-called pair-distance distribution function $p(r)$, i.e., the spatial autocorrelation function of the electron density fluctuations, $\Delta p(r)$, as

$$I(q) = 4\pi \int_0^\infty p(r) \frac{\sin qr}{qr} dr$$

where r is the distance between two scattering centers chosen inside the molecule. We used the indirect Fourier transformation (IFT) technique^{31–33} to calculate $p(r)$.

Refolding and Gel Filtration Chromatography. The WA20 protein was denatured by 6 M guanidine hydrochloride (GdnHCl) for 2 h at 4 °C in 25 mM HEPES buffer (pH 7.0) containing 100 mM NaCl, 10% glycerol, and 1 mM DTT. For refolding, the denatured WA20 proteins at different concentrations (2.1, 0.21, and 21 µg/mL) were dialyzed three times for ~4 h (×3) against 200× volume of 25 mM HEPES buffer (pH 7.0) containing 100 mM NaCl, 10% glycerol, and 1 mM DTT. The refolded WA20 proteins were analyzed by gel filtration chromatography (25 mM HEPES buffer (pH 7.0) containing 100 mM NaCl, 10% glycerol, and 1 mM DTT) on a Superdex 75 10/300 GL column (GE healthcare). Absorbance was monitored at 280 nm. The calibration curve for molecular weight (M_w) estimation was plotted with a Gel Filtration Calibration kit LMW (GE Healthcare) (Figure S4, Supporting Information).

Differential Scanning Fluorimetry (DSF). A real-time PCR device, MiniOpticon (Bio-Rad, Hercules, CA), was used to monitor protein unfolding by the increase in the fluorescence of the fluorophore SYPRO Orange (Invitrogen) with affinity for hydrophobic parts of the protein, which are exposed as the protein unfolds.^{34,35} WA20 protein (1 mg/mL) samples (20 µL) with SYPRO Orange (5× concentration) in 25 mM HEPES buffer (pH 7.0) containing 100–1000 mM NaCl, 10% glycerol, and 1 mM DTT were analyzed in 48-well PCR microplates (Bio-Rad). The relative fluorescence intensity was plotted as a function of temperature; this generates a sigmoidal curve that can be described by a two-state transition.³⁵ The inflection point of the transition curve (T_m) was calculated using the curve fitting function in KaleidaGraph (Synergy Software, Reading, PA) with the following equation:

$$y = LL + \frac{UL - LL}{1 + \exp\left(\frac{T_m - x}{a}\right)}$$

where LL and UL are the values of minimum and maximum intensities, respectively, and a denotes the slope of the curve within T_m .³⁵

RESULTS AND DISCUSSION

Overall Structure. The crystal structure of the *de novo* four-helix bundle protein WA20 was solved by the MAD method, and refined to 2.2 Å. The crystallographic data are summarized in Table 1. The WA20 crystal contains two protein molecules per asymmetric unit. The final model includes 189 amino acid residues of two WA20 monomers, 55 water molecules, and one metal ion in the asymmetric unit. The metal ion is probably potassium, because of the metal–ligand geometry.³⁶ The N-terminal and C-terminal residues and some loop residues (chain A, 1–2, 102; chain B, 1–4, 49–55, 102) are invisible due to disorder.

Surprisingly, the WA20 crystal structure is not a monomeric four-helix bundle like the *de novo* proteins S-824¹⁰ and S-836¹¹ (Figure 1B) but a dimeric four-helix bundle (Figure 2). Each monomer comprises two long α -helices, which span residues 4–50 (α 1), 54–100 (α 2) in chain A (Figure 1A) and residues 8–47 (α 1), 58–100 (α 2) in chain B. The helices intertwine with the helices of the other monomer, and the two monomers together form a 3D domain-swapped^{37,38} dimer. The four α -helices wrap around into a left-handed coiled coil.³⁹ The overall shape of WA20 is cylindrical with a length of ~8 nm and a diameter of ~3 nm. Helices α 1 (chain A) and α 2 (chain B) are roughly parallel. Helices α 1 (B) and α 2 (A) are also roughly

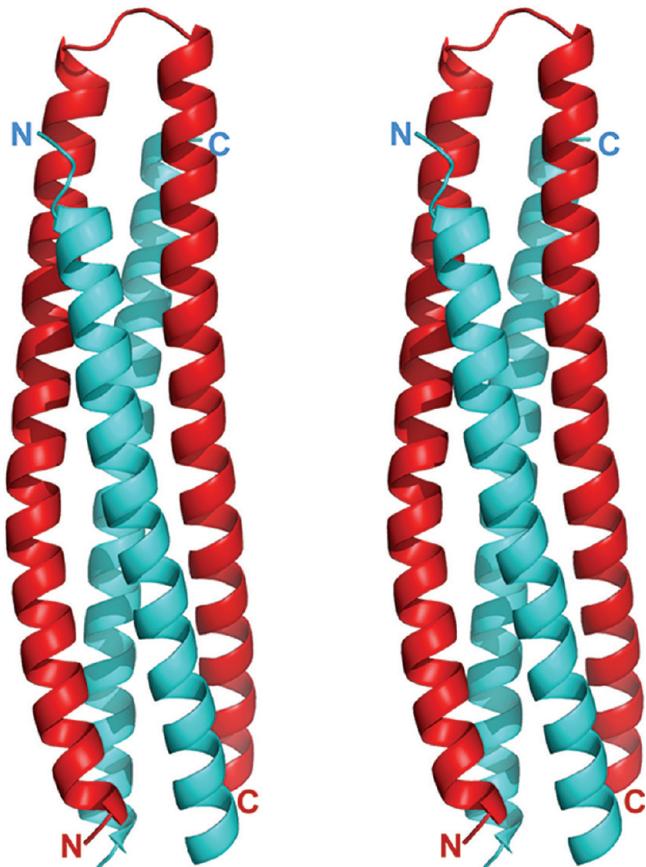


Figure 2. Ribbon representation of the crystal structure of WA20 (stereoview). Chains A and B are shown in red and cyan, respectively.

parallel. In contrast, the angles between helices α 1 (A) and α 1 (B) and helices α 2 (A) and α 2 (B) are about 20°, similar to the angle found in the “knobs-into-holes” packing of many natural α -helical proteins.⁴⁰ Four-helix bundles, in which some angles are ~20° and others are more parallel (or antiparallel), also occur in nature (e.g., cytochrome b_{562})⁴¹ and in *de novo* proteins (e.g., S-824 and S-836).^{10,11}

In more detailed views, there are differences of architecture between the four-helix bundle *de novo* proteins, monomeric S-824 and dimeric WA20. In the monomeric S-824, helices 1 (residues 5–20) and 2 (residues 28–48), and helices 3 (residues 56–72) and 4 (residues 80–99) are roughly antiparallel, and the angle between helices 1 and 4 and between helices 2 and 3 is ~20°. In contrast, in the view of the upper half part of WA20, their corresponding regions 1 (residues 5–20 in chain B) and 3 (residues 56–72 in chain A), and regions 2 (residues 28–48 in chain A) and 4 (residues 80–99 in chain B) are roughly parallel, and the angle between regions 1 and 2 and between regions 3 and 4 is ~20°.

The Binary Patterned Structure and the Dimer Interface. Figure 3A shows that the hydrophobic residues of WA20 form core regions in the four-helix bundle dimeric structure. The head-on views in Figure 3B and C indicate that the side chains in the crystal structure are clearly partitioned with nonpolar residues (yellow) in the interior and polar residues (red) on the surface, as specified by the binary code design strategy. In addition, Figure 3B shows that four helices are located on the diamond shape at the end part of the WA20 structure. In the diagonal distance of the diamond shape, the helices at the loop region are proximal (ca. 1.2–1.3 nm) and

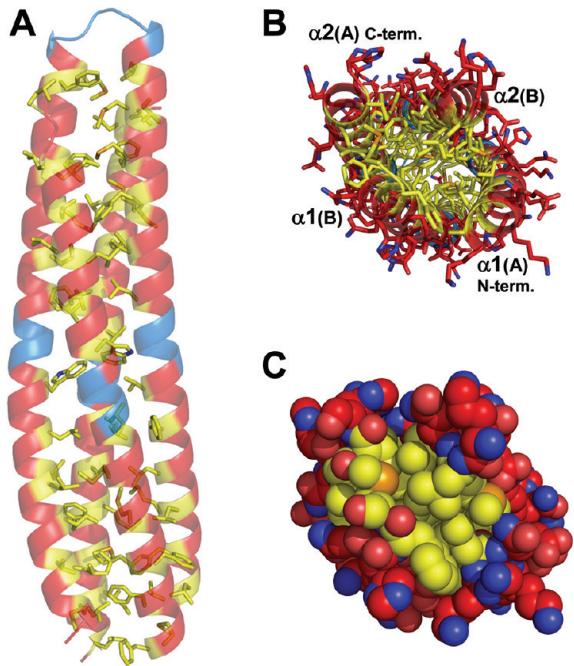


Figure 3. The binary patterned structure of WA20. (A) The hydrophobic core of WA20 in the four-helix bundle dimeric structure. Nonpolar side chains are shown as stick models. The color coding is the same as Figure 1A. Polar (red), nonpolar (yellow), and turn (cyan) residues as the design template. (B) Head-on view (from the side with disordered loop region in chain B) with the polar (red) and nonpolar (yellow) side chains shown as stick models. (C) Same as part B but in space-filling representation.

the helices at the terminal regions are distant (ca. 1.8–1.9 nm). The dimer interface is predominantly hydrophobic with several hydrophobic clusters (Figure 4A and Figure S5, Supporting Information). The major residues, involved in these hydrophobic interactions, include Val9, Ile12, Leu16, Leu19, Trp23, Leu30, Met33, Met37, Leu40, Phe41, Phe44, Met48, Met64, Phe67, Val71, Leu75, Phe85, Leu89, Leu92, Phe96, and Leu99 (Figure 4A). We suggest that one reason for the stable structure of WA20 is because there are roughly twice as many hydrophobic interactions in a dimeric structure of WA20 relative to a monomeric structure like S-824.

The dimer interface is further stabilized by interchain salt bridges and/or hydrogen bonds between the atoms of Gln27(A)-Thr81(B), Glu63(A)-Lys13(B), Asp72(A,B)-His86-(B,A), His74(A)-Asn20(B), and Ser79(A)-His83(B) (Figure 4B and Figure S5, Supporting Information). To estimate the key residues of dimerization, we carefully compared the structure and amino acid sequence of dimeric WA20 with those of monomeric S-824. Significant differences occur in the loop regions of residues 25–28 and 77–80. The regions (GGKD and GGKH) in S-824 are glycine-rich loops. In contrast, the regions (RHQG and SESD) in WA20 are located in the middle of α -helices, which are stabilized by intrachain salt bridges or hydrogen bonds between His26 and Glu78 (Figure 4C). These observations suggest that the His26 and Glu78 residues in the designed loop regions are potential key residues leading to the formation of the domain-swapped dimer, since the interactions at loops affect the free energy difference between the monomer and the 3D domain-swapped multimer.³⁸

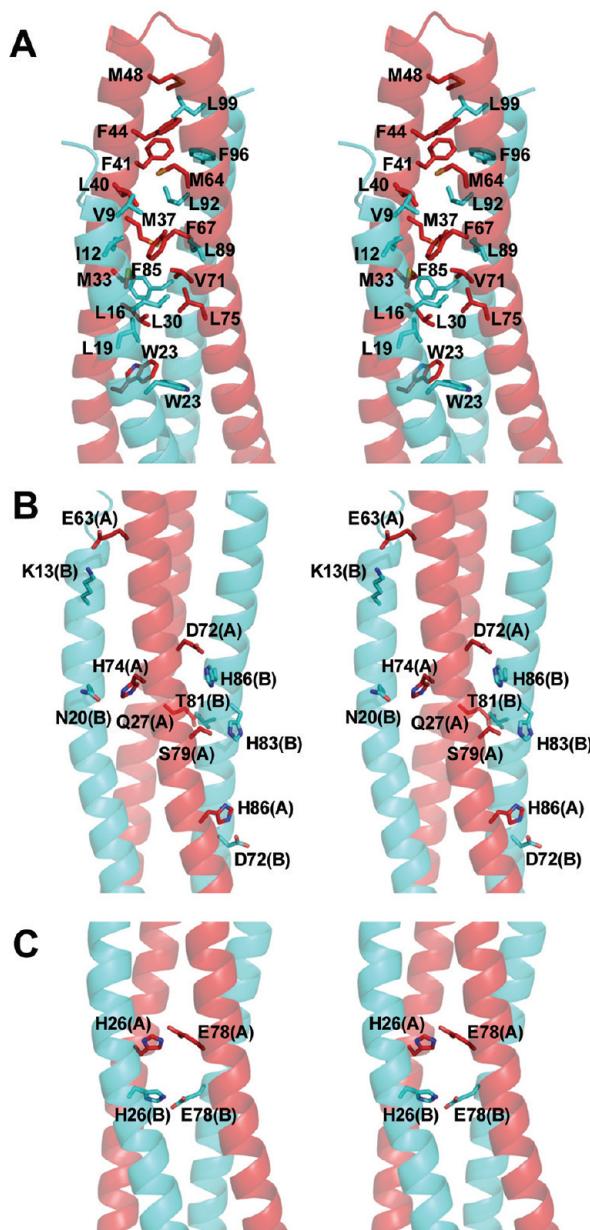


Figure 4. Close-up view of the dimer interface of WA20 (stereoview). Chains A and B are shown in red and cyan, respectively. (A) The major clusters of hydrophobic residues in the dimer interface are shown as stick models. (B) The residues of interchain salt bridges and hydrogen bonds, as determined by DIMPLOT in LIGPLOT,⁴⁹ are shown as stick models. (C) Potential key residues of intrachain salt bridges or hydrogen bonds for dimer formation are shown as stick models.

Small-Angle X-ray Scattering (SAXS). To examine the solution structure of WA20, we performed SAXS. Figure S5A and B show that X-ray scattering intensities of WA20 and chicken egg lysozyme as a molecular weight reference (lysozyme $M_w = 14.3$ kDa). Assuming that WA20 and lysozyme have identical scattering length densities, no aggregation in solution, and the structure factor $S(q) = 1$ for dilute samples, the forward scattering intensity $I(q \rightarrow 0)$ is proportional to the molecular weight (M_w) at the same concentration (5.4 mg/mL). The $I(q \rightarrow 0)$ of WA20 and lysozyme are 0.0812 and 0.0459 cm^{-1} , respectively. The molecular weight of WA20 was estimated to be 25.3 kDa.

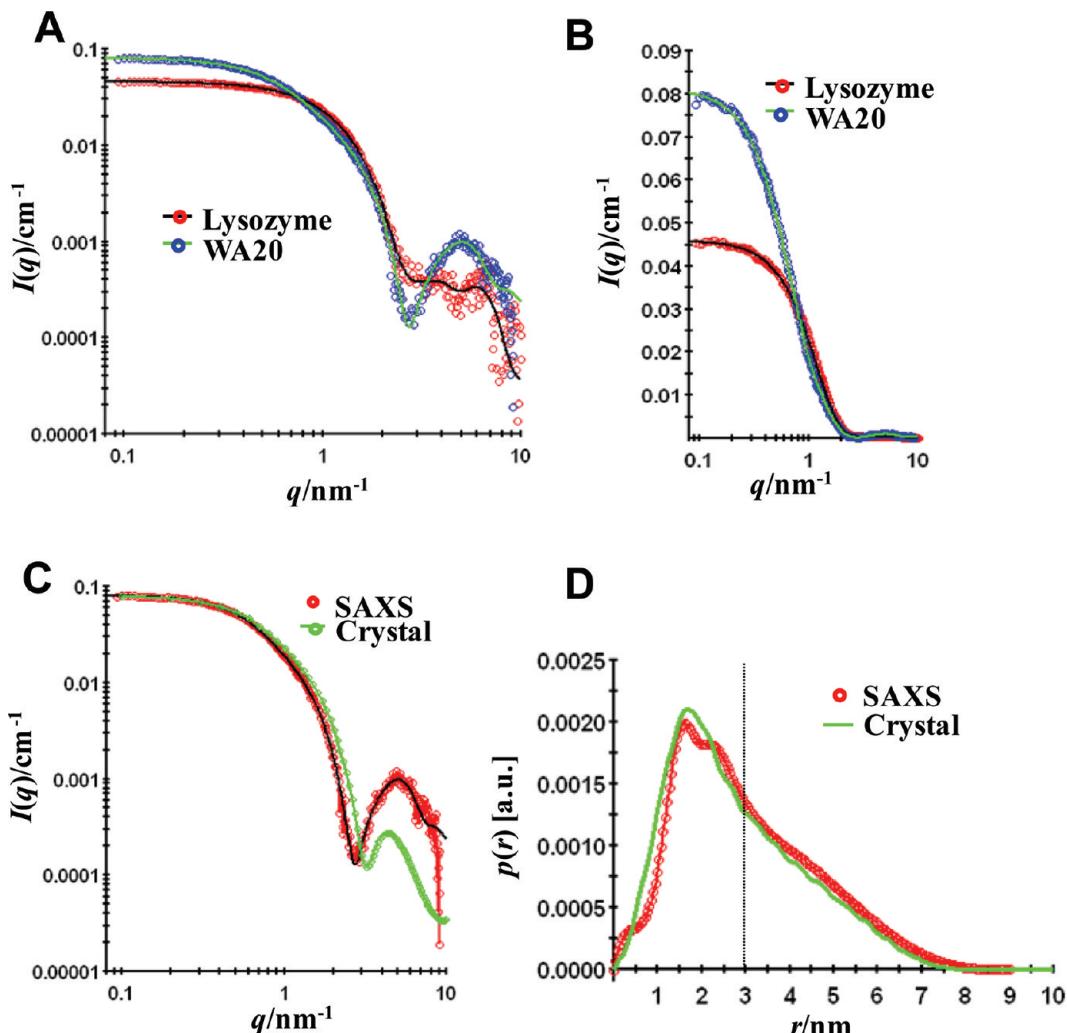


Figure 5. Small-angle X-ray scattering (SAXS) analyses. SAXS intensities, $I(q)$, in (A) log scale and (B) linear scale, of WA20 (5.4 mg/mL) and chicken egg lysozyme (5.4 mg/mL) in HEPES buffer solution at 25 °C on absolute scale. (C) SAXS intensities, $I(q)$, of WA20 solution and that simulated from the crystal structure of WA20. (D) The corresponding pair distance distribution function, $p(r)$, of WA20 obtained as output of indirect Fourier transformation (IFT) analysis and that simulated from the crystal structure of WA20.

Since the molecular weight of WA20 monomer is 12.5 kDa, these data show that WA20 forms a dimer in solution. To extract intuitive real-space information via a virtually model-free routine, we obtained the pair-distance distribution function, $p(r)$, of WA20 using an indirect Fourier transformation (IFT) technique (Figure 5D). The $p(r)$ indicates that the maximum diameter, D_{\max} , is ~ 8 nm, which is consistent with the crystal structure. The observed pronounced peak of $p(r)$ in the low- r regime and extended linear tail in the high- r regime are significant features of rod-like structure.⁴² The inflection point located on the higher- r side of the maximum in $p(r)$, highlighted by broken lines in Figure 5D, gives a measure of the cross section diameter, $D_{c,\max}$. The $D_{c,\max}$ value of WA20 is roughly ~ 3 nm, which is also consistent with the crystal structure. Furthermore, $I(q)$ and $p(r)$ of WA20 resemble those simulated from the crystal structure of WA20 (Figure 5C and D). These SAXS results show that WA20 forms the dimeric four-helix bundle structure in solution.

Concentration Independence of Dimerization in Refolding. To examine the concentration dependence of dimerization, we tried refolding of WA20 at different concentrations, and analyzed the resulting protein by gel

filtration chromatography. WA20 was denatured by 6 M GdnHCl and refolded by dialysis at different concentrations of protein (2.1 mg/mL, 0.21 mg/mL, and 21 µg/mL). There was no precipitation during refolding. The circular dichroism (CD) spectra (Figure S6, Supporting Information) show that WA20 was denatured completely by 6 M GdnHCl, and the α -helical content of WA20 was recovered by refolding. Table 2 shows the elution volume and estimated molecular weight by gel filtration chromatography with the calibration curve (Figure S4, Supporting Information). The dimer peak of WA20 was clearly detected by gel filtration chromatography in each tested concentration following refolding, and no monomer peak was detected (Figure S7, Supporting Information). These data

Table 2. Elution Volume and Molecular Weight Estimated by Gel Filtration Chromatography

WA20 sample	V_e (mL)	K_{av}	M_w (kDa)	state
native	11.91	0.254	25.6	dimer
refolded at 2.1 mg/mL	11.83	0.249	26.4	dimer
refolded at 0.21 mg/mL	11.85	0.250	26.2	dimer
refolded at 21 µg/mL	11.87	0.251	26.0	dimer

indicate that, in the range of concentrations tested, the refolding of WA20 into its dimeric structure is independent of concentration, thereby suggesting that the dimer form of WA20 is much more stable than the monomer.

Thermal Denaturation. To examine protein stability, we tested thermal denaturation by differential scanning fluorimetry (DSF)^{34,35} (Figure 6). The temperature at which a protein

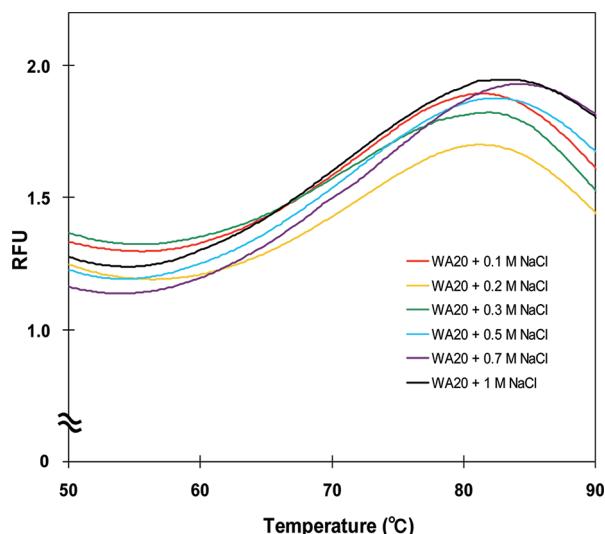


Figure 6. Thermal denaturation curves of WA20 by differential scanning fluorimetry (DSF) at various salt concentrations (0.1–1 M NaCl) (RFU, Relative Fluorescence Units).

unfolds is measured by an increase in the fluorescence of the SYPRO orange dye with affinity for hydrophobic parts of the protein, which are exposed as the protein unfolds.^{34,35} The melting temperature (T_m) of WA20 is about 70 °C at various salt concentrations (0.1–1 M NaCl) (Table 3), indicating that

Table 3. Melting Temperature (T_m) of WA20 at Various Salt Concentrations, Analyzed by Differential Scanning Fluorimetry (DSF)

NaCl conc. (M)	T_m (°C)
0.1	70.1
0.2	70.3
0.3	69.9
0.5	69.7
0.7	70.8
1	69.7

the high stability of the dimeric structure is independent of salt. In addition, to test whether the temperature-induced unfolding of WA20 is reversible, we monitored the fluorescence by DSF again after the first thermal denaturation and cooling. The fluorescence curve of the second thermal denaturation is significantly different from that of the first denaturation (Figure S8, Supporting Information), implying that the temperature-induced unfolding of WA20 is not reversible in the test condition.

Putative Primitive Active Site. We have shown previously that WA20 binds heme, and that this protein/heme complex has rudimentary peroxidase activity.¹² As shown previously¹² (and illustrated in Figure S9 in the Supporting Information), the heme complex of WA20 produces an absorbance spectrum

with a typical Soret peak at 410 nm. In natural heme proteins, histidine and methionine are generally used as axial ligand residues for heme.⁴³ WA20 is relatively rich in these residues, with 26 histidine and 16 methionine residues in the dimer (twice the 102-residue primary sequence, Table S1, Supporting Information). From this structure, several pairs of putative candidates for heme-ligand residues between chains A and B (e.g., His11-Met33, His24-His74, His31-His84, Met48-His101, His62-His97, His62-His101) are estimated by reference to the distance and geometry between axial ligand residues in cytochrome b_{562}^{41} and cytochrome c^{44} . Further studies are necessary to confirm the binding site.

We have also shown previously that WA20 in the absence of heme has low levels of esterase and lipase activity.¹² Although substantially less active (~10000-fold) than natural enzymes, this novel protein produces rate enhancements (k_{cat}/k_{uncat}) that are ~400-fold and ~500-fold above background for esterase and lipase activities, respectively.¹² To find putative substrate binding sites for hydrolase activities (esterase and lipase), we searched for pocket sites in the WA20 dimeric structure. Two relatively large pockets (volumes: 205 and 174 Å³), comprised of Leu30, Met33, Asn34, Met37, His74, and Leu75 (in chain A/B) and Leu16, Leu19, Phe85, Leu89, and Leu92 (in chain B/A) are detected using the programs Pocket-Finder⁴⁵ (<http://www.modelling.leeds.ac.uk/pocketfinder/>) and Cavor⁴⁶ (<http://www.caver.cz/>) (Figure 7). Similar sized pockets were also found in the F64A mutant of monomeric protein S-824.⁴⁷ We hypothesize that these hydrophobic pockets may

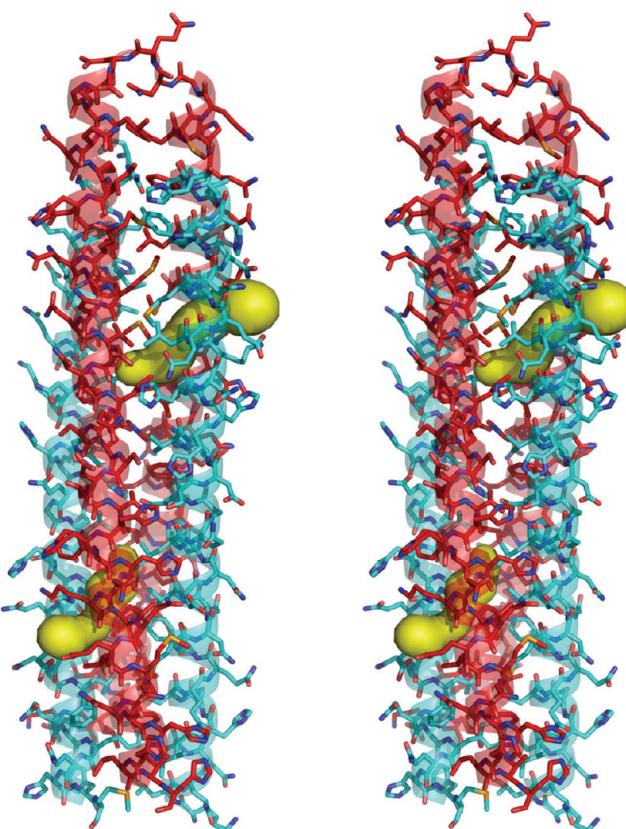


Figure 7. The pocket sites on the surface of WA20 (stereoview). There are two major pockets (yellow shapes) on the WA20 structure, depicted using the programs Cavor⁴⁶ and PyMOL. Chains A and B are shown in red and cyan, respectively.

serve as substrate binding sites. In natural hydrolases, a carboxyl peptidase family, Eqolisin, hydrolyzes peptide bonds using Glu and Gln as catalytic residues.⁴⁸ Although putative catalytic residues are still unclear, we speculate that candidates for primitive catalytic residues may possibly be Glu38, Glu91, Asn34, Gln35, and Asn95 around the putative substrate binding pockets, roughly similar to the Eqolisin family. The *de novo* proteins including WA20 showed similar K_m values to that of natural enzymes, but k_{cat} values of the *de novo* proteins were ~10000-fold lower than those of natural enzymes.¹² This is consistent with our speculation that the binding pockets exist on the WA20 structure but the catalytic residues are unevolved and not optimized. We expect that the four-helix bundle dimeric structure of WA20 with the pockets may serve as a simple framework for the evolution of *de novo* enzymes.

CONCLUSIONS

In this study, we show that the *de novo* protein WA20 forms a stable domain-swapped four-helix dimer. These results demonstrate that our *de novo* library of proteins contains not only simple monomeric proteins but also self-assembling, stable, and functional multimeric proteins. This structure also suggests potential binding sites for heme cofactor binding and coordination, as well as sites for hydrolase-substrate binding. These data suggest that the binary patterning strategy can be used to design libraries of more complicated multimeric structures, which may pave the way for the discovery of new functions with applications in synthetic biology and biotechnology.

ASSOCIATED CONTENT

Supporting Information

Supplementary information, additional results, Table S1, Figures S1–S9, and related descriptions. This material is available free of charge via the Internet at <http://pubs.acs.org>.

AUTHOR INFORMATION

Corresponding Author

*E-mail: rarai@shinshu-u.ac.jp (R.A.); hecht@princeton.edu (M.H.H.). Phone: +81-268-21-5881 (R.A.); 609-258-2901 (M.H.H.). Fax: +81-268-21-5881 (R.A.); 609-258-6746 (M.H.H.).

Present Addresses

[†]KAIGEN Co., Ltd, Osaka, 541-0045, Japan.

[#]Dept. of Pathology and Laboratory Medicine, University of Pennsylvania School of Medicine, Philadelphia, PA, 19104, United States.

[▽]Dept. of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138, United States.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank Dr. Phil Jeffrey at Princeton University for helpful advice on crystallography. We also thank staff of the Structural Biology Beamlines at the Photon Factory, KEK, for data collection. This work was supported by JSPS Postdoctoral Fellowships for Research Abroad, Grants-in-Aid for Scientific Research, and Program for Dissemination of Tenure-Track System to R.A., funded by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan. This work

was also supported by NSF grants MCB-0817651 and MCB-1050510 to M.H.H.

REFERENCES

- Dahiyat, B. I.; Mayo, S. L. *Science* **1997**, *278*, 82–87.
- Kuhlman, B.; Dantas, G.; Ireton, G. C.; Varani, G.; Stoddard, B. L.; Baker, D. *Science* **2003**, *302*, 1364–1368.
- Keefe, A. D.; Szostak, J. W. *Nature* **2001**, *410*, 715–718.
- Kamtekar, S.; Schiffer, J. M.; Xiong, H.; Babik, J. M.; Hecht, M. H. *Science* **1993**, *262*, 1680–1685.
- Hecht, M. H.; Das, A.; Go, A.; Bradley, L. H.; Wei, Y. *Protein Sci.* **2004**, *13*, 1711–1723.
- Bradley, L. H.; Kleiner, R. E.; Wang, A. F.; Hecht, M. H.; Wood, D. W. *Protein Eng., Des. Sel.* **2005**, *18*, 201–207.
- Wei, Y.; Liu, T.; Sazinsky, S. L.; Moffet, D. A.; Pelczer, I.; Hecht, M. H. *Protein Sci.* **2003**, *12*, 92–102.
- West, M. W.; Wang, W.; Patterson, J.; Mancias, J. D.; Beasley, J. R.; Hecht, M. H. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 11211–11216.
- Roy, S.; Hecht, M. H. *Biochemistry* **2000**, *39*, 4603–4607.
- Wei, Y.; Kim, S.; Fela, D.; Baum, J.; Hecht, M. H. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 13270–13273.
- Go, A.; Kim, S.; Baum, J.; Hecht, M. H. *Protein Sci.* **2008**, *17*, 821–832.
- Patel, S. C.; Bradley, L. H.; Jinadasa, S. P.; Hecht, M. H. *Protein Sci.* **2009**, *18*, 1388–1400.
- Fisher, M. A.; McKinley, K. L.; Bradley, L. H.; Viola, S. R.; Hecht, M. H. *PLoS One* **2011**, *6*, e15364.
- Wang, A. F. Senior Thesis, Princeton University, NJ, 2006.
- Platt, J. M. Senior Thesis, Princeton University, NJ, 2007.
- Hendrickson, W. A. *Science* **1991**, *254*, 51–58.
- LeMaster, D. M.; Richards, F. M. *Biochemistry* **1985**, *24*, 7263–7268.
- Johnson, B. H.; Hecht, M. H. *Bio/Technology* **1994**, *12*, 1357–1360.
- Otwinowski, Z.; Minor, W. *Methods Enzymol.* **1997**, *276*, 307–326.
- Terwilliger, T. C.; Berendzen, J. *Acta Crystallogr., Sect. D* **1999**, *55*, 849–861.
- Terwilliger, T. C. *Acta Crystallogr., Sect. D* **2002**, *58*, 1937–1940.
- Emsley, P.; Lohkamp, B.; Scott, W. G.; Cowtan, K. *Acta Crystallogr., Sect. D* **2010**, *66*, 486–501.
- Murshudov, G. N.; Vagin, A. A.; Dodson, E. J. *Acta Crystallogr., Sect. D* **1997**, *53*, 240–255.
- Murshudov, G. N.; Skubak, P.; Lebedev, A. A.; Pannu, N. S.; Steiner, R. A.; Nicholls, R. A.; Winn, M. D.; Long, F.; Vagin, A. A. *Acta Crystallogr., Sect. D* **2011**, *67*, 355–367.
- Collaborative Computational Project, *Acta Crystallogr., Sect. D* **1994**, *50*, 760–763.
- Laskowski, R. A.; Macarthur, M. W.; Moss, D. S.; Thornton, J. M. *J. Appl. Crystallogr.* **1993**, *26*, 283–291.
- Chen, V. B.; Arendall, W. B., 3rd; Headd, J. J.; Keedy, D. A.; Immormino, R. M.; Kapral, G. J.; Murray, L. W.; Richardson, J. S.; Richardson, D. C. *Acta Crystallogr., Sect. D* **2010**, *66*, 12–21.
- Davis, I. W.; Leaver-Fay, A.; Chen, V. B.; Block, J. N.; Kapral, G. J.; Wang, X.; Murray, L. W.; Arendall, W. B., 3rd; Snoeyink, J.; Richardson, J. S.; Richardson, D. C. *Nucleic Acids Res.* **2007**, *35*, W375–383.
- Lovell, S. C.; Davis, I. W.; Arendall, W. B., 3rd; de Bakker, P. I.; Word, J. M.; Prisant, M. G.; Richardson, J. S.; Richardson, D. C. *Proteins* **2003**, *50*, 437–450.
- Orthaber, D.; Bergmann, A.; Glatter, O. *J. Appl. Crystallogr.* **2000**, *33*, 218–225.
- Glatter, O. *Acta Phys. Austriaca* **1980**, *52*, 243–256.
- Glatter, O. *J. Appl. Crystallogr.* **1980**, *13*, 577–584.
- Glatter, O.; Kratky, O. *Small-Angle X-Ray Scattering*; Academic Press: New York, 1982.
- Vedadi, M.; Niesen, F. H.; Allali-Hassani, A.; Fedorov, O. Y.; Finerty, P. J., Jr.; Wasney, G. A.; Yeung, R.; Arrowsmith, C.; Ball, L. J.; Berglund, H.; Hui, R.; Marsden, B. D.; Nordlund, P.; Sundstrom, M.,

- Weigelt, J.; Edwards, A. M. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 15835–15840.
- (35) Niesen, F. H.; Berglund, H.; Vedadi, M. *Nat. Protoc.* **2007**, *2*, 2212–2221.
- (36) Harding, M. M. *Acta Crystallogr., Sect. D* **2002**, *58*, 872–874.
- (37) Bennett, M. J.; Choe, S.; Eisenberg, D. *Protein Sci.* **1994**, *3*, 1444–1463.
- (38) Liu, Y.; Eisenberg, D. *Protein Sci.* **2002**, *11*, 1285–1299.
- (39) Mason, J. M.; Arndt, K. M. *ChemBioChem* **2004**, *5*, 170–176.
- (40) Crick, F. H. C. *Acta Crystallogr.* **1953**, *6*, 689–697.
- (41) Lederer, F.; Glatigny, A.; Bethge, P. H.; Bellamy, H. D.; Matthew, F. S. *J. Mol. Biol.* **1981**, *148*, 427–448.
- (42) Sato, T.; Shimozawa, T.; Fukasawa, T.; Ohtaki, M.; Aramaki, K.; Wakabayashi, K.; Ishiwata, S. *Biophysics* **2010**, *6*, 1–11.
- (43) Reedy, C. J.; Gibney, B. R. *Chem. Rev.* **2004**, *104*, 617–649.
- (44) Finzel, B. C.; Weber, P. C.; Hardman, K. D.; Salemme, F. R. *J. Mol. Biol.* **1985**, *186*, 627–643.
- (45) Hendlich, M.; Rippmann, F.; Barnickel, G. *J. Mol. Graphics Modell.* **1997**, *15*, 359–363.
- (46) Beneš, P.; Chovancová, E.; Kozlíková, B.; Pavelka, A.; Strnad, O.; Brezovský, J.; Šustr, V.; Klvaňa, M.; Szabó, T.; Gora, A.; Zamborský, M.; Biedermannová, L.; Medek, P.; Damborský, J.; Sochor, J. *CAVER 2.1 Software*; Masaryk University: Brno, Czech Republic, 2010.
- (47) Das, A.; Wei, Y.; Pelczer, I.; Hecht, M. H. *Protein Sci.* **2011**, *20*, 702–711.
- (48) Fujinaga, M.; Cherney, M. M.; Oyama, H.; Oda, K.; James, M. N. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 3364–3369.
- (49) Wallace, A. C.; Laskowski, R. A.; Thornton, J. M. *Protein Eng.* **1995**, *8*, 127–134.