Background: Evolutionary protein design provides a deeper understanding of how primordial proteins emerged. Results: An evolutionary model is proposed on the basis of structurally guided stepwise segment elongation. Conclusion: The structural guidance facilitates structural organization and gain-of-function of a generated 25-residue artificial protein. Significance: This study provides insights into how primordial protein evolution may have been promoted by structural guidance.

The understanding of how primordial proteins emerged has been a fundamental and longstanding issue in biology and biochemistry. For a better understanding of primordial protein evolution, we synthesized an artificial protein on the basis of an evolutionary hypothesis, segment-based elongation starting from an autonomously foldable short peptide. A 10-residue protein, chignolin, the smallest foldable polypeptide ever reported, was used as a structural support to facilitate higher structural organization and gain-of-function in the development of an artificial protein. Repetitive cycles of segment elongation and subsequent phage display selection successfully produced a 25-residue protein, termed AF.2A1, with nanomolar affinity against the Fc region of immunoglobulin G. AF.2A1 shows exquisite molecular recognition ability such that it can distinguish conformational differences of the same molecule. The structure determined by NMR measurements demonstrated that AF.2A1 forms a globular protein-like conformation with the chignolin-derived β-hairpin and a tryptophan-mediated hydrophobic core. Using sequence analysis and a mutation study, we discovered that the structural organization and gain-of-function emerged from the vicinity of the chignolin segment, revealing that the structural support served as the core in both structural and functional development. Here, we propose an evolutionary model for primordial proteins in which a foldable segment serves as the evolving core to facilitate structural and functional evolution. This study provides insights into primordial protein evolution and also presents a novel methodology for designing small sized proteins useful for industrial and pharmaceutical applications.

A fundamental question in biology and biochemistry is how primordial proteins emerged and evolved into modern proteins with their own functions and structures (1–4). Attempts to design artificial proteins provide the basis for a deeper understanding of the principles of primordial protein evolution (5–9). Designing a protein de novo, however, has remained challenging even for small proteins. This is due to the enormous diversity in sequence space of a protein, which makes it impractical to test all possible sequences by means of any conventional method. This holds true in the case of primordial protein evolution because such proteins emerged from prebiotic circumstances where modern biological systems did not exist (3). Thus, it is reasonable to suppose that primordial proteins must have efficiently evolved in a way that promoted their evolution. Examining reproducible mechanisms using an artificial protein would provide us better understanding of the emergence of primordial proteins.

The vast complexity of the sequence space implies that a long, functional polypeptide has never emerged by chance. It must have evolved from smaller molecules along particular pathways that can be described by a physical inevitability. Segment-based protein evolution, proposed in the context of exon shuffling (10, 11), hypothesizes that short peptide segments were assembled into single polypeptides and then evolved into modern proteins with their complex structures and functions. This hypothesis, although plausible, does not have enough empirical evidence to explain primordial protein evolution. In general, a short peptide segment is too flexible to fold into a specific structure, making folding and/or associating with other peptide segments energetically unfavorable.

Overcoming the inherent flexibility of peptides is indispensable for efficient segment-based evolution. The association between flexible peptide segments and subsequent folding requires the compensation of considerable entropic cost. To drive the structural organization, a small structural support such as metal coordination (7–9) may have been involved in primordial proteins. We hypothesized that a foldable short peptide segment would also be capable of serving as a structural support to promote segment-based protein evolution. In contrast to those studies using metal-coordinated motifs, however,
there have been few successful reports on segment-based de novo protein generation without relying on metal coordination. The lack of these trials is due to the scarcity of such short foldable peptide segments. Consequently, very little information is available concerning potential evolutionary mechanisms driven by foldable peptide segments.

To better understand the principles of primordial protein evolution, we propose a new evolutionary hypothesis, structurally guided stepwise segment elongation with a foldable short peptide segment. We previously reported on a 10-residue mini-evolution, we propose a new evolutionary hypothesis, structurally driven by foldable peptide segments.

**Structurally Guided Stepwise Segment Elongation**

**EXPERIMENTAL PROCEDURES**

**Materials**—The Fc region of human immunoglobulin G (IgG) used for phage display selection was purchased from Jackson ImmunoResearch. In addition to this commercially available Fc region, highly purified Fc of human monoclonal IgG (Chugai Pharmaceutical Co., Ltd.) was prepared by using a Fab preparation kit (Pierce) followed by anion exchange chromatography with DEAE-Sepharose (GE Healthcare) and size exclusion chromatography with a Superdex 200-pg column (GE Healthcare). Synthesized peptides were purchased from Bio-Synthesis, Inc. Synthesized Oligo DNAs were purchased from RIKAKEN (Japan).

**Library Preparation and Phage Display Selection**—The molecular evolution was carried out in two steps. The first generation library was constructed by elongating a randomized sequence of eight residues Xaa$_8$ at the C terminus of a chignolin-derived segment termed CLN. Here, the Xaa was encoded by a degenerate codon NNK (where N represents equal molar amounts of A, C, G, and T, and K is equal amounts of G and T). The gene fragments were synthesized by overlap extension polymerase chain reaction (PCR), digested by restriction enzymes EcoRI/HindIII, and then ligated into the C-terminal part of a T7 phage coat protein gene 10 (g10) of the T7Select 10–3b vector (Novagen). In vitro packaging and amplification of phages were carried out in accordance with the T7Select™ system manual (Novagen). (Novagen). The phages were precipitated with polyethylene glycol (M, 8000) and NaCl, collected by centrifugation (14,000 × g, 20 min), and resuspended in 1 ml of Tris-buffered saline containing 0.1% (w/v) Tween 20 (TBS-T). The phage library was biopanned against a biotinylated human Fc region (Jackson ImmunoResearch) immobilized on streptavidin MagneSphere® paramagnetic particles (Promega). After washing 10 times with TBS-T, bound phages were eluted with TBS buffer containing 1% (w/v) SDS. The eluted phages were amplified, followed by the next round of biopanning. After the final round of biopanning, 96 clones were randomly isolated and assayed by ELISA (see below). The clone identified as the highest affinity clone from ELISA and sequential analysis was used for the second generation library.

The second generation library was constructed by elongating Xaa$_n$ at the N terminus of the clone selected from the first generation library. Gene fragments encoding the polypeptide library were inserted into g10 of the T7Select 10-1b vector (Novagen). Biopanning was carried out as described above.

**Enzyme-linked Immunosorbent Assay (ELISA)**—A 96-well microtiter plate, MEDISORP 96 (Nunc), was coated with T7 phages displaying each polypeptide by incubating for 1 h at room temperature. After washing five times with TBS-T, binding to the Fc region was detected with Sunrise R (TECAN) by using a horseradish peroxidase-conjugated human Fc region (Jackson ImmunoResearch) and ABTS One-Component HRP Microwell Substrate (BioFX).

**Surface Plasmon Resonance (SPR)**—SPR assays were performed with a Biacore T100 (GE Healthcare). 1200 resonance units (RU) of human Fc region (Jackson ImmunoResearch) was immobilized on the Series S sensor chip CM5 (GE Healthcare) via primary amine groups using N-hydroxysuccinimide and 1-ethyl-3-[3-dimethylaminopropyl]carbodiimide hydrochloride. For a reference cell, another flow cell was blocked with 1 M ethanolamine. Binding assays were performed under HBS-T (10 mM HEPES (pH 7.4), 150 mM NaCl, 0.05% (w/v) Tween 20). The binding data were fitted to a 1:1 binding model to determine an equilibrium dissociation constant ($K_D$) using Biacore T100 evaluation software (GE Healthcare). The affinity data were fitted with the nonlinear van’t Hoff equation (Equation 1) (21) by using EGOR Pro (WaveMetrics, Inc.).

$$RT \ln K_D = \Delta H - T\Delta S + \Delta C_p (T - T_0) - \Delta C_p T \ln (T/T_0)$$

(Eq. 1)

where $R$ is the gas constant; $T$ is the temperature; $T_0$ is an arbitrary reference temperature (e.g. 298.15 K); $\Delta H$ and $\Delta S$ are the binding enthalpy and entropy at $T_0$; and $\Delta C_p$ is the change in heat capacity.

For the binding to a non-native conformer of IgG, acid-treated IgG was prepared by dialyzing human monoclonal IgG against 20 mM sodium acetate (pH 4.5) at a temperature of 323 K. Heat-treated IgG was prepared at a temperature of 343 K for 15 min in HBS-T buffer containing 2.7 mM KCl. Reduced Fc region was prepared by treating the purified Fc region with 50 mM 2-mercaptoethanol (Pierce) at a temperature of 310 K for 90 min. For SPR analysis, 3000 RU of AF.2A1 was immobilized on a sensor chip CM5. Obtained data were processed by using Biacore T100 evaluation software (GE Healthcare). Pulldown Assay—0.5 mg of AF.2A1 was immobilized on 100 μl of N-hydroxysuccinimide-activated Sepharose™ HP (GE Healthcare) in accordance with the manufacturer’s protocol.

$^2$The abbreviations used are: SPR, surface plasmon resonance; SEC, size exclusion chromatography; AUC, analytical ultracentrifugation.
Structurally Guided Stepwise Segment Elongation

Crude cell lysate was prepared by sonicating a 1-ml culture of *Escherichia coli* BLT 5403 strain in TBS-T. Acid-treated Fc region was prepared by dialyzing the Fc region against 10 mM Gly-HCl, 150 mM NaCl (pH 2.0) for 12 h. The crude cell lysate with or without the acid-treated Fc region was mixed with AF.2A1-immobilized Sepharose. After washing five times with TBS-T, the bound Fc region was eluted with 50 mM NaOH. The purities of samples were evaluated by SDS-PAGE.

Determination of Molecular Weight in Solution by Analytical Ultracentrifugation (AUC)—For AUC, sedimentation equilibrium data were obtained on a ProteomeLab XL-A (Beckman Coulter). AF.2A1 was dissolved in 10 mM sodium acetate-\(d_4\) buffer (pH 4.5). Plotted data are for 72 \(\mu\)M AF.2A1 at rotor speed of 40,000 rpm. The experimental molecular weight was calculated from the fitted slope. The monodispersity of AF.2A1 was also confirmed by size exclusion chromatography (SEC). SEC was performed with an AKTA purifier (GE Healthcare) using Superdex Peptide 10/300 GL (GE Healthcare). AF.2A1 was dissolved in 10 mM sodium acetate-\(d_4\) buffer (pH 4.5) and injected onto the column.

NMR Analysis—NMR spectra were obtained on a Bruker AMX 500 (500.13 MHz for \(^1\)H). Synthesized peptides were dissolved at 1.8 mM in 10 mM sodium acetate-\(d_4\) buffer (pH 4.5) containing 5% (v/v) \(D_2\)O and 0.01% sodium 2,2-dimethyl-2-silapentane-5-sulfonate. Homonuclear two-dimensional spectra (DQF-COSY, NOESY, and TOCSY) with water suppression) were performed at 283 K. The mixing time was 200 ms for NOESY, and 50 ms for TOCSY. \(^1\)H\(\alpha\) coupling values were obtained from cross-peaks in DQF-COSY spectrum. Resonance assignments were carried out with the standard sequential procedure and by using Sparky (22). An amide hydrogen-deuterium (H/D) exchange experiment was performed at 283 K in 10 mM sodium acetate-\(d_4\) buffer (pH 4.5) containing 0.01% sodium 2,2-dimethyl-2-silapentane-5-sulfonate, and the decay of the intensity was monitored in one-dimensional spectra from 4 to 60 min. Protection factors were determined by dividing the observed exchange rates by the intrinsic rates, which were calculated by the method of Bai et al. (23). The cross-peaks in the NOESY spectrum were divided into four classes according to their volumes, which corresponded to interproton distance restraints of 1.5–2.8, 1.5–3.5, 2.0–4.5, and 2.5–6.0 Å, respectively. Torsion angle constraints for \(\phi\) were determined from \(^3\)H\(\alpha\). They were classified into three categories as follows: 
- -65° ± 35°, -100° ± 70°, -120° ± 50° corresponding to \(^3\)H\(\alpha\) <6.5 and 6.5–8.0, and >8.0 Hz, respectively. Simulated annealing calculations were carried out using CNS (24). The distance restraint for a hydrogen bond was implemented in the final calculation. Finally, 20 structures having no distance violation exceeding 0.2 Å and no torsion angle violation beyond 3° were obtained out of 50 initial structures generated in random conformations.

Fluorescence Analysis—Tryptophan intrinsic fluorescence spectra were measured at 283 K with a Jasco FP-6500 spectrofluorometer (Jasco). Each sample was dissolved at 10 \(\mu\)M in 10 mM sodium acetate-\(d_4\) buffer (pH 4.5). Emission spectra were recorded between 300 and 400 nm with an excitation wavelength of 295 nm.

The 2nd gen. (N-terminal lib.)
- phage - (Xaa)_4 - CLN - selected seq.

The 1st gen. (C-terminal lib.)
- phage - CLN - (Xaa)_6 - Gly

**FIGURE 1.** Scheme for structurally guided stepwise segment elongation. The center inset shows the three-dimensional structure of chignolin (Protein Data Bank code 1UAO). Eight-residue randomized regions Xaa\(_4\) were elongated from the chignolin-derived segment (termed CLN) in two steps. In the first generation library (1st gen. C-terminal lib.), Xaa\(_6\) was elongated from the C terminus of CLN and displayed on T7 phage. After affinity selection to a target, a second generation library (2nd gen. N-terminal lib.) was generated in which Xaa\(_8\) was elongated from the N terminus of an amino acid sequence selected from the first generation library.

Competitive Selection from a Spiked Library—A control library that consisted of polypeptides with an unstructured segment was constructed by replacing chignolin segment of the second generation library (Table 1) with a glycine-rich sequence (Gly\(_3\)-Ser-Gly\(_4\)). Here, the control library and conventional chignolin-based library are termed “Gly library” and “CLN library,” respectively. The Gly library and CLN library were mixed with a ratio of Gly library/CLN library = 10:1 to prepare a spiked library. Biopanning selection from the spiked library against the Fc region was performed in accordance with the same procedure described above. 48 clones were randomly isolated from each group (the spiked library and polyclonal population after each round of biopanning selection). Amino acid sequences of the isolated clones were determined three separate times. The ratio of clones from CLN library and the ratio of unique amino acid sequences were calculated.

RESULTS

Functional Selection of Affinity Clones from Segment-elongated Protein Libraries—The designed protein consists of three segments as follows: a foldable chignolin-derived segment (hereafter referred to as CLN) and two elongated segments fused with the N and C termini of CLN (Fig. 1). CLN is an eight-residue segment (Tyr-Asp-Pro-Xaa-Thr-Gly-Thr-Trp) obtained by truncating the N- and C-terminal glycine residues of chignolin and introducing a randomized amino acid at the middle position where the original glutamic acid residue has been found not to contribute to the formation of the chignolin structure (12). We performed the segment elongation in two steps. The first library was generated by elongating a segment containing an eight-residue randomized region and a glycine.
Structurally Guided Stepwise Segment Elongation

Detailed functional analysis revealed the exquisite molecular recognition ability of AF.2A1. Interestingly and unexpectedly, AF.2A1 bound to the Fc region that we purchased from a manufacturer but not to any monoclonal IgGs, including a commercially available therapeutic antibody, nor to the purified Fc region that we prepared from monoclonal IgG (Fig. 2, a–c). We then tested whether or not AF.2A1 recognized a panel of IgG and Fc regions treated under various conditions as follows: slightly acidic conditions (pH 4.0), heat-treated conditions, and reducing conditions (Fig. 2, d–f). AF.2A1 bound to all of those treated IgG and Fc regions. These results imply that AF.2A1 did not recognize natively formed IgG nor the Fc region, but rather non-native conformers caused upon acidification or heat treatment (25–27).

We analyzed the binding ability of a 25-residue protein, AF.2A1, by using SPR. Chemically synthesized AF.2A1 showed high affinity to the IgG Fc region with an equilibrium dissociation constant \(K_D\) of \(7.9 \pm 2.8 \times 10^{-5}\) (M), showing that the second segment elongation enhanced the affinity by more than 3 orders of magnitude.

We carried out biopanning selection against the human IgG Fc region used as a model target. The first generation library was used for the second generation library. The second generation library was prepared by elongating Xaa8 at the N terminus of the selected clone. After ELISA screening, the most converged clone (17/21 clones) termed H6 was isolated. We then elongated its N-terminal segment to evolve the H6 into a higher affinity binder. Sequential comparisons of the selected clones revealed notable convergence of amino acid sequences. The first generation library of \(5 \times 10^6\) clones was displayed on T7 phage. The fourth round of selection enriched clones with distinctive sequences (Table 1). The second generation library was prepared by biopanning selection against a target molecule. A selected clone termed 2A1 (18/30 clones) (Table 1).

### TABLE 1

| 1\textasciitilde generation | -YD PTGTVXXXXX XXXXG Convergence |
|-----------------------------|----------------------------------|
| H6                          | -R- RSSI AYGG-                   | 17/21 |
| H3                          | -G- RSYL RFGG-                   | 1/21  |
| G11                         | -V- TSSI AGWM-                   | 1/21  |
| D5                          | -Y- RSSI WVL-                    | 1/21  |
| E7                          | -G- RSWL SNV-                    | 1/21  |
| 2\textasciitilde generation | XXXXYYYYYD PTGTVRSSSL AYGGG Convergence |
| 2A1                         | GVRQWSSG--                      | 18/30 |
| 2A4                         | AGSSRARG-                       | 5/30  |
| 2F12                        | ASVRSWSS-                       | 1/30  |
| 2F2                         | TRGRSAR-                        | 1/30  |
| 2F4                         | HWNGREG-                        | 1/30  |
| 2F6                         | ERWITWSG-                       | 1/30  |
| 2F7                         | GSVRRWLG-                       | 1/30  |
| 2F9                         | GAVRERSFG-                      | 1/30  |
| 2H1                         | SWRRNGSS-                       | 1/30  |

Functional Analysis of an Evolved Protein Having the Converged Sequence—We analyzed the binding ability of a 25-residue protein, AF.2A1, possessing the converged sequence of 2A1, by using SPR. Chemically synthesized AF.2A1 showed high affinity to the IgG Fc region with an equilibrium dissociation constant \(K_D\) of \(2.2 \pm 0.2 \times 10^{-8}\) (M). The affinity of AF.H6, a 17-residue peptide corresponding to the clone H6, a selected clone from the first generation library, was determined to be \(K_D = 7.9 \pm 2.8 \times 10^{-5}\) (M), showing that the second residue Xaa8-Gly at the C terminus of CLN, then followed by biopanning selection against a target molecule. A selected clone from the first generation library was used for the second generation library. The second generation library was prepared by elongating Xaa8 at the N terminus of the selected clone.

We carried out biopanning selection against the human IgG Fc region used as a model target. The first generation library was used for the second generation library. The second generation library was prepared by biopanning selection against a target molecule. A selected clone termed 2A1 (18/30 clones) (Table 1). We then elongated its N-terminal segment to evolve the H6 into a higher affinity binder. The fifth round of biopanning selection from the second generation library of \(5 \times 10^6\) clones successfully enriched a converged clone termed 2A1 (18/30 clones) (Table 1).

Sequential comparisons of the selected clones revealed notable convergence of amino acid sequences. The first generation library was used for the second generation library. The second generation library was prepared by biopanning selection against a target molecule. A selected clone termed 2A1 (18/30 clones) (Table 1).

The three-dimensional structure of AF.2A1 revealed that the chignolin-derived \(\beta\)-hairpin was retained after fusion with the two elongated segments. The \(\beta\)-hairpin of AF.2A1 was stabilized by the edge-to-face \(\pi-\pi\) interaction between Tyr-9 and Trp-16, which was derived from chignolin (Fig. 5b). Being positioned via the \(\beta\)-hairpin structure of CLN, the two elongated segments interacted with each other. This interaction made the overall structure of AF.2A1 globular and not extended.

Each elongated segment adopted a structurally different conformation (Fig. 5a). The N-terminal segment formed a relatively extended structure and associated mostly with the C-terminal segment. In contrast, the C-terminal segment formed a compact structure and associated with both CLN and the N-terminal segment. This difference was quantitatively illustrated by a contact map (Fig. 6). The inter-residue contacts between the C-terminal segment and CLN were twice as large as those between the N-terminal segment and CLN.

AF.2A1 formed a hydrophobic core within its structure as seen in globular proteins. Trp-16 was surrounded by the side...
chains of Ile-20, Tyr-9, and Gln-5. Most of the Trp-16 side chain was buried (Fig. 7a), serving as the core residue. To survey the tryptophan environment, we measured the tryptophan intrinsic fluorescence spectra of AF.2A1, as well as fragmented AF.2A1, which was an equimolar mixture of the N- and C-terminal elongated segments and CLN (Fig. 7b). The fluorescent peak of AF.2A1 was 7.5-nm blue-shifted compared with those of fragmented AF.2A1 and chignolin, thus strongly indicating the burial of Trp-16 into the interior of the protein. The results of the fluorescence experiment that implied the formation of the hydrophobic core were highly consistent with the structural characteristics given by NMR.

Mutation Studies—The structural analysis demonstrated that CLN guides the overall structure of AF.2A1. To evaluate the effect of the structural support of CLN on the function, we assayed the binding of a protein variant termed AF.2A1Gly in which the CLN sequence was substituted into a flexible glycine-rich sequence of Gly3-Ser-Gly4. The substitution resulted in a significant decrease of the affinity by more than 3 orders of magnitude; the $K_D$ of AF.2A1Gly was determined to be $2.3 \pm$
FIGURE 3. AUC (a) and SEC (b) of AF.2A1. The experimental molecular weight calculated from the fitted slope in a is 2158, which is close to the theoretical molecular weight of AF.2A1, 2753. mAU, milli-absorbance units.

FIGURE 4. $^1$H NOESY spectra of the fingerprint region (a) and aromatic region (b) of AF.2A1 at 283 K. The labels in a are sequential assignments. Cross-peaks in b show long range NOEs between Tyr9-H and Trp16-He1 or Trp16-NH, and between Tyr9-He and Trp16-He1 or Trp16-NH.
1.9 \times 10^{-5} (\text{M}). Cleavage of AF.2A1 into three segments, i.e. the fragment mixture of CLN and the two elongated segments, resulted in almost complete loss of the binding function, and a meaningful affinity constant could not be quantified. This indicated that the CLN segment itself was not involved in the binding directly. Thus, the affinity loss observed in AF.2A1Gly was attributed to the lack of the CLN structure.

Sequential comparisons of selected clones implied that converged amino acid residues played an important role in the function. We investigated residues important for the binding by alanine scanning mutagenesis. Of the converged residues (Fig. 8a), the side chains of Val-3 and Ile-20 were buried in the protein, and those of five residues (Trp-6, Ser-7, Arg-17, Ser-18, and Ser-19) were exposed to the solvent. Alanine scanning mutagenesis showed that four of the five solvent-exposed resi-


**Evolutionary Significance of Introducing a Foldable Peptide Segment**—AF.2A1 showed characteristic features of a globular protein-like structure and function, rather than those of a flexible polypeptide, although its amino acid sequence is far different from that of any native protein motif. Functional, structural, and evolutionary analysis on the basis of SPR, NMR, and competitive selection strongly suggests that the foldable chignolin, even if unrelated to the function itself, can contribute to promoting evolution. CLN played a major role in the structural organization of AF.2A1. Formation of the overall structure was mainly assisted by the chignolin-derived β-hairpin, thereby allowing the two elongated segments to be positioned closely to each other. AF.2A1 formed a hydrophobic core in which Trp-16 was almost entirely shielded from the solvent, as confirmed by tryptophan intrinsic fluorescence. The core needs to be supported by surrounding residues (Fig. 7a). For instance, alanine substitution of isoleucine at position 20 significantly reduced the protein’s binding affinity (Table 3). This corresponds to the convergence of aliphatic residues at position 20. The hydrophobic core containing Trp-16 was stabilized at least 40-fold by CLN (Fig. 5b). This stabilization probably made Trp-16 suitable for serving as a hydrophobic core of AF.2A1. Trp-16 (red), Tyr-9 (in CLN and the C-terminal segment (left panel)). The side chain of Trp-16 was almost entirely shielded from the solvent, as confirmed by tryptophan intrinsic fluorescence. The peak positions for AF.2A1, a fragment mixture, and CLN are 347.6 ± 0.04, 355.1 ± 0.19, and 357.6 ± 0.36 nm, respectively. AU, arbitrary units.

**DISCUSSION**

**Evolutionary Significance of Introducing a Foldable Peptide Segment**—AF.2A1 showed characteristic features of a globular protein-like structure and function, rather than those of a flexible polypeptide, although its amino acid sequence is far different from that of any native protein motif. Functional, structural, and evolutionary analysis on the basis of SPR, NMR, and competitive selection strongly suggests that the foldable chignolin, even if unrelated to the function itself, can contribute to promoting evolution. CLN played a major role in the structural organization of AF.2A1. Formation of the overall structure was mainly assisted by the chignolin-derived β-hairpin, thereby allowing the two elongated segments to be positioned closely to each other. AF.2A1 formed a hydrophobic core in which Trp-16 was almost entirely shielded from the solvent, as confirmed by tryptophan intrinsic fluorescence. The core needs to be supported by surrounding residues (Fig. 7a). For instance, alanine substitution of isoleucine at position 20 significantly reduced the protein’s binding affinity (Table 3). This corresponds to the convergence of aliphatic residues at position 20. The hydrophobic core containing Trp-16 was stabilized at least 40-fold by CLN (Fig. 5b). This stabilization probably made Trp-16 suitable for serving as

**TABLE 3**

| Proteins                | $K_D$ (M) |
|-------------------------|-----------|
| AF.H6                   | 7.9 ± 2.8 × 10^{-5} |
| AF.2A1                  | 2.2 ± 0.2 × 10^{-6} |
| AF.2A1Gly               | 2.3 ± 1.9 × 10^{-5} |
| Fragment mixture         | >10^{-5}   |
| AF.2A1(Q5R)             | 8.7 ± 0.2 × 10^{-9} |
| AF.2A1(W6A)             | 1.8 ± 0.3 × 10^{-7} |
| AF.2A1(S7A)             | 1.1 ± 0.2 × 10^{-7} |
| AF.2A1(R17A)            | 3.5 ± 0.1 × 10^{-7} |
| AF.2A1(S18A)            | 2.3 ± 0.1 × 10^{-8} |
| AF.2A1(S19A)            | 2.2 ± 0.3 × 10^{-7} |
| AF.2A1(I20A)            | 9.2 ± 0.7 × 10^{-10} |

* Equimolar mixture of CLN and the N- and C-terminal segments.
* A meaningful value was not determined because of too low an affinity.

**FIGURE 7.** a, tryptophan-mediated hydrophobic core of AF.2A1. Trp-16 (red) is surrounded by Gln-5 in the N-terminal region, Tyr-9 in CLN, and Ile-20 in the C-terminal segment. The side chain of Trp-16 was buried as a result of folding (right panel). b, tryptophan intrinsic fluorescence spectra for AF.2A1 (red), a fragment mixture of the N-terminal segment, C-terminal segment, and CLN (blue), and CLN (black). The peak positions for AF.2A1, a fragment mixture, and CLN are 347.6 ± 0.04, 355.1 ± 0.19, and 357.6 ± 0.36 nm, respectively. AU, arbitrary units.

**FIGURE 8.** a, sequence logo for selected amino acid sequences. b, mapping of the deduced hot spot region (red) in the interaction. CLN and the elongated segments are colored in light blue and wheat, respectively.

clones were important in the binding (Table 3). Sequentially or structurally, these residues were located in the vicinity of CLN and clustered at a particular area of the molecular surface as generally seen in protein-protein interaction hot spots (Fig. 8b) (28). The alanine substitution at position 20, where hydrophobic residues (Ile or Leu) converged and formed the hydrophobic core with Trp-16, caused a 40-fold reduction in the affinity. This indicates the importance of the stable structure of AF.2A1 on its function. In contrast, the point mutant Q5R enhanced the affinity by a factor of 2.5 (Table 3). The sequential comparison showed an arginine residue converged at position 5, and the
sequences of the clones were determined three separate times. Amino acid sequences, respectively. 48 clones were isolated from each group (the spiked library and each round of biopanned population). Amino acid sequences of the clones were determined three separate times. Error bars are standard deviations.

FIGURE 9. Competitive selection from a spiked library. The bar and line graph indicate the ratio of clones from CLN library and the ratio of unique amino acid sequences, respectively. 48 clones were isolated from each group (the spiked library and each round of biopanned population). Amino acid sequences of the clones were determined three separate times. Error bars are standard deviations.

Structurally Guided Stepwise Segment Elongation

Note that the clone 2A1 was selected not by its structural stability but by binding affinity. This suggests that functional evolution could promote structural organization. Similarly, in earlier work, functional selection from random libraries yielded folded proteins stabilized by a zinc finger motif (29, 30), although it is also reported that a functional selection gave rise to a high affinity binder whose structural organization was restricted to a part of overall structure (31). Structural organization through functional selection probably depends on selection pressure where structural restraints enhance function. Under such circumstances, structure was organized by functional selection, and function was enhanced by structural organization. A similar evolutionary process was illustrated by an in silico study (32) where fixation of inter-residue contacts around an active site in early evolution gradually developed into overall structure formation. Considering the enormous diversities of sequential and conformational spaces of a protein, it is quite reasonable to deduce such a stepwise cyclic relationship between the structural organization and functional enhancement.

Chignolin-derived structural support was involved in the gain-of-function. The amino acid convergence after functional selection mainly occurred at positions sequentially or structurally in the vicinity of CLN (e.g. Val-3, Trp-6, Ser-7, Gly-8, Arg-17, and Ser-18). Most of the converged residues with solvent-exposed side chains were located on the same side of the surface (Fig. 8b) and critical to the binding function as revealed by the mutation study. These residues composed a hot spot region as seen in native protein-protein interactions. Positions in the vicinity of CLN were the most structurally stabilized in the elongated segments and were probably more suitable for the generation of an active site by minimizing unfavorable entropic costs. Tryptophan and arginine are frequently occurring hot spot residues in protein-protein interactions as statistically deduced (28). The features of the AF.2A1 functional region strongly correspond to the typical properties of protein-protein interactions. Therefore, it is plausible that introducing a structural support into a polypeptide could promote gain-of-function in protein evolution as well as structural organization.

The above observation that CLN affected both the structure and function led us to propose an evolutionary model in which a foldable segment serves as the core for both structural organization and gain-of-function, and thereby promotes primordial protein evolution. Indeed, the result of the competitive selection strongly suggested that CLN promoted functional evolution (Fig. 9). The contribution of CLN to molecular function was also confirmed by a mutation study that showed that substitution of CLN to a glycine-rich linker caused significant affinity reduction by 3 orders of magnitude (Table 3). In light of our functional, structural, and evolutionary observation, we propose the term “evolving core” for the foldable segment in primordial protein evolution. Several protein folding studies have revealed that a partially folded segment in a protein, described as the “folding core” (33–37), promotes the folding process probably by restricting the conformational space accessible to a polypeptide. Analogously, in our model, evolving core promotes evolution by giving a partially stabilized structure to a polypeptide. It has been reported that the recombination of 30–40-residue polypeptide segments produces a novel folded protein in which each segment retains its own structural features and acts as a template for the fold (38). Our results suggest that similar events occur in the evolution of much smaller and more primitive proteins.

Structural stabilization is key to promoting evolution as has been supported by simulation studies (39) and experimental observations (40). De novo generated proteins that have been reported also utilize structural supports for stabilization (29), leading to a general view that metal coordination plays a major role in primordial protein evolution (9). This view is persuasive because a relatively short peptide could also form a metal coordination motif. Meanwhile, some short peptide segments can also fold into their own structures (12, 13, 41, 42). As demonstrated here, introducing a foldable segment with the size of a local structure of a modern protein can give rise to a structured
protein as a result of functional selection. The stabilization was attributed to secondary structural formation of a β-hairpin and hydrophobic core formation through the tryptophan residue. It should be noted that tryptophan-mediated hydrophobic cores are found in other artificial or naturally occurring small sized proteins such as the Trp cage (43) or Pin WW domain (44); thus, it is reasonable to assume that such hydrophobic cores formed in early protein evolution. We conclude that a foldable peptide segment possesses great potential to promote primordial protein evolution.

**Molecular Design of Small Sized Proteins for Practical Applications**—We should emphasize the advantage of structurally guided segment elongation from a practical viewpoint. To date, the protein engineering technique has developed and utilized small structural motifs as a scaffold such as zinc finger motif (7), tryptophan zipper (31, 45), Trp cage motif (46), and disulfide-cyclized β-hairpin (47) to generate a wide variety of functional mini-proteins of tens of residues. By comparison with those scaffold-based approaches, our strategy is based on an evolutionary concept of “structurally guided stepwise segment elongation” that has a characteristic that an artificial protein evolves from a short flexible peptide to a more structured functional protein in a stepwise manner with the aid of a structural support. Similar evolutionary mechanisms have been reported in enzyme evolution (48) and antibody affinity maturation (49) where a flexible molecule with cross-reactivity or weak activity evolves toward a more specific and higher activity protein along with entropic reduction. Those studies have proposed that structural flexibility in the early stage could increase structural and thereby functional diversity. In light of these preceding studies, we can expect a similar mechanism in structurally guided stepwise segment elongations.

The use of the 10-residue chignolin has the potential to give rise to small-sized proteins useful for industrial and pharmaceutical applications. AF.2A1 showed high affinity and specificity without depending on any additional stabilization such as chemical cross-linking (50), protein splicing (51), or metal coordination (7, 9). Because of not relying on a disulfide bond, AF.2A1 was capable of functioning under reducing conditions. This characteristic would expand the range of use, including a specific binder that functions in cytosol. This small sized protein exhibited resistance against misfolding and/or aggregation even after being boiled at 373 K (Fig. 10). Such a small sized protein can be a promising alternative to high molecular weight proteinaceous binders. Molecular design concepts based on theories of primordial protein evolution will meet the growing demand for low molecular weight binders with superior thermal and storage stabilities and low manufacturing costs.

*Acknowledgment*—We thank Dr. H. Takahashi for technical support in AUC measurements.

*REFERENCES*

1. Seidel, H. M., Pompliano, D. L., and Knowles, J. R. (1992) Exons as microgenes? *Science* 257, 1489–1490
2. Söding, J., and Lupas, A. N. (2003) More than the sum of their parts: on the evolution of proteins from peptides. *BioEssays* 25, 837–846
3. van der Gulik, P., Massar, S., Gilis, D., Buhman, H., and Rooman, M. (2009) The first peptides: the evolutionary transition between prebiotic amino acids and early proteins. *J. Theor. Biol.* 261, 531–539
4. Carvunis, A. R., Rolland, T., Wapinski, I., Calderwood, M. A., Yildirim, M. A., Simonis, N., Charlooteaux, B., Hidalgo, C. A., Barbette, J., Santha-nam, B., Brar, G. A., Weissman, J. S., Regev, A., Thierry-Mieg, N., Cusick, M. E., and Vidal, M. (2012) Proto-genes and de novo gene birth. *Nature* 487, 370–374
5. Riechmann, L., and Winter, G. (2000) Novel folded protein domains generated by combinatorial shuffling of polypeptide segments. *Proc. Natl. Acad. Sci. U.S.A.* 97, 10068–10073
6. Keefe, A. D., and Szostak, J. W. (2001) Functional proteins from a random-sequence library. *Nature* 410, 715–718
7. Seelig, B., and Szostak, J. W. (2007) Selection and evolution of enzymes from a partially randomized non-catalytic scaffold. *Nature* 448, 828–831
8. Riechmann, L., and Winter, G. (2006) Early protein evolution: building domains from ligand-binding polypeptide segments. *J. Mol. Biol.* 363, 460–468
9. Salgado, E. N., Ambroggio, X. I., Brodin, J. D., Lewis, R. A., Kuhlman, B., and Tezcan, F. A. (2010) Metal templated design of protein interfaces. *Proc. Natl. Acad. Sci. U.S.A.* 107, 1827–1832
10. Blake, C. C. (1978) Do genes-in-pieces imply proteins-in-pieces? *Nature* 273, 267
11. Gilbert, W. (1978) Why genes in pieces? *Nature* 271, 501
12. Honda, S., Yamasaki, K., Sawada, Y., and Morii, H. (2004) 10 residue folded peptide designed by segment statistics. *Structure* 12, 1507–1518
13. Honda, S., Akiba, T., Kato, Y. S., Sawada, Y., Sekijima, M., Ishimura, M., Ooishi, A., Watanabe, H., Ohadara, T., and Harata, K. (2008) Crystal structure of a ten-amino acid protein. *J. Am. Chem. Soc.* 130, 15327–15331
14. Satoh, D., Shimizu, K., Nakamura, S., and Terada, T. (2006) Folding free-energy landscape of a 10-residue mini-protein, chignolin. *FEBS Lett.* 580, 3422–3426
15. Xu, W., Lai, T., Yang, Y., and Mu, Y. (2008) Reversible folding simulation by hybrid Hamiltonian replica exchange. *J. Chem. Phys.* 128, 175105
16. Terada, T., Satoh, D., Mikawa, T., Ito, Y., and Shimizu, K. (2008) Understanding the roles of amino acid residues in tertiary structure formation of chignolin by using molecular dynamics simulation. *Proteins* 73, 621–631
17. Roy, S., Goedecker, S., Field, M. J., and Penev, E. (2009) A minima hopping
Structurally Guided Stepwise Segment Elongation

study of all-atom protein folding and structure prediction. J. Phys. Chem. B 113, 7315–7321.

18. Hatfield, M. P., Murphy, R. F., and Lovas, S. (2010) Molecular dynamics analysis of the conformations of a β-hairpin miniprotein. J. Phys. Chem. B 114, 3028–3037.

19. Kurniawan, N. A., Enemark, S., and Rajagopalan, R. (2012) Crowding alters the folding kinetics of a β-hairpin by modulating the stability of intermediates. J. Am. Chem. Soc. 134, 10200–10208.

20. Davis, C. M., Xiao, S., Raleigh, D. P., and Dyer, R. B. (2012) Raising the speed limit for β-hairpin formation. J. Am. Chem. Soc. 134, 14476–14482.

21. Tabata, S., Kuroki, K., Wang, J., Kajikawa, M., Shiratori, I., Kohda, D., Arase, H., and Maenaka, K. (2008) Biophysical characterization of O-glycosylated CD99 recognition by paired Ig-like type 2 receptors. J. Biol. Chem. 283, 8893–8901.

22. Goddard, T. D., and Kneller, D. G. (2006) SPARKY, Version 3, University of California, San Francisco.

23. Bai, Y., Milne, J. S., Mayne, L., and Englander, S. W. (1993) Primary structure effects on peptide group hydrogen exchange. Proteins 17, 75–86.

24. Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T., and Warren, G. L. (1998) Crystallography & NMR system: A new software suite for macromolecular structure determination. Acta Crystallogr. D Biol. Crystallogr. 54, 905–921.

25. Thies, M. J., Kammermeier, R., Richter, K., and Buchner, J. (2001) The alternatively folded state of the antibody C(H)3 domain. J. Mol. Biol. 309, 1077–1085.

26. Buchner, J., Renner, M., Lilie, H., Hinz, H. J., Jaenicke, R., Kiefhaber, T., and Rudolph, R. (1991) Alternatively folded states of an immunoglobulin. Biochemistry 30, 6922–6929.

27. Kannert, D., Borrows, A. C., Jonsson, B. H., and Enander, K. (2011) Thermal induction of an alternatively folded state in human IgG Fc. Biochemistry 50, 981–988.

28. Bogan, A. A., and Thorn, K. S. (1998) Anatomy of hot spots in protein interfaces. J. Mol. Biol. 280, 1–9.

29. Lo Surdo, P., Walsh, M. A., and Sollazzo, M. (2004) A novel ADP- and zinc-binding fold from function-directed in vitro evolution. Nat. Struct. Mol. Biol. 11, 382–383.

30. Chao, F. A., Morelli, A., Haugner, J. C., 3rd, Churchfield, L., Hagmann, L. N., Shi, L., Masterson, L. R., Sarangi, R., Veglia, G., and Seelig, B. (2013) Structure and dynamics of a primordial catalytic fold generated by in vitro evolution. Nat. Chem. Biol. 9, 81–83.

31. Kim, S., Kim, D., Jung, H. H., Lee, I. H., Kim, J. I., Suh, J. Y., and Joon, S. (2012) Bio-inspired design and potential biomedical applications of a novel class of high-affinity peptides. Angew. Chem. Int. Ed. Engl. 51, 1890–1894.

32. Nagao, C., Terada, T. P., Yomo, T., and Sasai, M. (2005) Correlation between evolutionary structural development and protein folding. Proc. Natl. Acad. Sci. U.S.A. 102, 18950–18955.

33. Woodward, C. (1993) Is the slow exchange core the protein folding core? Trends Biochem. Sci. 18, 359–360.

34. Orban, J., Alexander, P., and Bryan, P. (1994) Hydrogen-deuterium exchange in the free and immunoglobulin G-bound protein G B-domain. Biochemistry 33, 5702–5710.

35. Raschke, T. M., Kho, J., and Marqusee, S. (1999) Confirmation of the hierarchical folding of RNase H: a protein engineering study. Nat. Struct. Biol. 6, 825–831.

36. Kuwata, K., Shastry, R., Cheng, H., Hoshino, M., Batt, C. A., Goto, Y., and Roder, H. (2001) Structural and kinetic characterization of early folding events in β-lactoglobulin. Nat. Struct. Biol. 8, 151–155.

37. Curnow, P., Di Bartolo, N. D., Moreton, K. M., Ajoie, O. O., Saggese, N. P., and Booth, P. J. (2011) Stable folding core in the folding transition state of an α-helical integral membrane protein. Proc. Natl. Acad. Sci. U.S.A. 108, 14133–14138.

38. de Bono, S., Riechmann, L., Girard, E., Williams, R. L., and Winter, G. (2005) A segment of cold shock protein directs the folding of a combinatorial protein. Proc. Natl. Acad. Sci. U.S.A. 102, 1396–1401.

39. Bloom, J. D., Labthavikul, S. T., Otey, C. R., and Arnold, F. H. (2006) Protein stability promotes evolvability. Proc. Natl. Acad. Sci. U.S.A. 103, 869–874.

40. Tokuriki, N., and Tawfik, D. S. (2009) Stability effects of mutations and protein evolvability. Curr. Opin. Struct. Biol. 19, 596–604.

41. Honda, S., Kobayashi, N., and Munekata, E. (2000) Thermodynamics of a β-hairpin structure: evidence for cooperative formation of folding nucleus. J. Mol. Biol. 295, 269–278.

42. Blanco, F. J., Rivas, G., and Serrano, L. (1994) A short linear peptide that folds into a native stable β-hairpin in aqueous solution. Nat. Struct. Biol. 1, 584–590.

43. Neidigh, J. W., Fesinmeyer, R. M., and Andersen, N. H. (2002) Designing a 20-residue protein. Nat. Struct. Biol. 9, 425–430.

44. Jäger, M., Nguyen, H., Crane, J. C., Kelly, I. W., and Gruebele, M. (2001) The folding mechanism of a β-sheet: the WW domain. J. Mol. Biol. 311, 373–393.

45. Cochran, A. G., Skelton, N. J., and Starovasnik, M. A. (2001) Triptophan zippers: stable, monomeric β-hairpins. Proc. Natl. Acad. Sci. U.S.A. 98, 5578–5583.

46. Herman, R. E., Badders, D., Fuller, M., Makienko, E. G., Houston, M. E., Jr., Quay, S. C., and Johnson, P. H. (2007) The Trp cage motif as a scaffold for the display of a randomized peptide library on bacteriophage T7. J. Biol. Chem. 282, 9813–9824.

47. Cochran, A. G., Tong, R. T., Starovasnik, M. A., Park, E. J., McDowell, R. S., Theaker, J. E., and Skelton, N. J. (2001) A minimal peptide scaffold for β-turn display: optimizing a strand position in disulfide-cycZed β-hairpins. J. Am. Chem. Soc. 123, 625–632.

48. Tomatis, P. E., Fabiane, S. M., Simona, F., Carloni, P., Sutton, B. J., and Vila, A. J. (2008) Adaptive protein evolution grants organismal fitness by improving catalysis and flexibility. Proc. Natl. Acad. Sci. U.S.A. 105, 20605–20610.

49. James, L. C., Roversi, P., and Tawfik, D. S. (2003) Antibody multispecificity mediated by conformational diversity. Science 299, 1362–1367.

50. Dias, R. L., Fasan, R., Moehle, K., Renard, A., Obrecht, D., and Robinson, J. A. (2006) Protein ligand design: from phage display to synthetic protein epitope mimetics in human antibody Fc-binding peptides. J. Am. Chem. Soc. 128, 2726–2732.

51. van Lieshout, J. F., Perez Gutierrez, O. N., Vroom, W., Planas, A., de Vos, W. M., van der Oost, J., and Koutsopoulos, S. (2012) Thermal stabilization of an endoglucanase by cyclization. Appl. Biochem. Biotechnol. 167, 2039–2053.

52. Laskowski, R. A., Rullmann, J. A., MacArthur, M. W., Kaptein, R., and Thornton, J. M. (1996) AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. J. Biomol. NMR 8, 477–486.

53. Vehlow, C., Stehr, H., Winkelmann, M., Duarte, J. M., Petzold, L., Dinne, J., and Lappe, M. (2011) CMView: interactive contact map visualization and analysis. Bioinformatics 27, 1573–1574.