J|A|C|S
ARTICLES

Published on Web 10/25/2008

Crystal Structure of a Ten-Amino Acid Protein

Shinya Honda,* Toshihiko Akiba, Yusuke S. Kato,† Yoshito Sawada,
Masakazu Sekijima, Miyuki Ishimura, Ayako Ooishi, Hideki Watanabe,
Takayuki Odahara, and Kazuaki Harata

*National Institute of Advanced Industrial Science and Technology (AIST), Central 6,
Tsukuba 305-8566, Japan*

Received April 25, 2008; E-mail: s.honda@aist.go.jp

Abstract: What is the smallest protein? This is actually not such a simple question to answer, because there is no established consensus among scientists as to the definition of a protein. We describe here a designed molecule consisting of only 10 amino acids. Despite its small size, its essential characteristics, revealed by its crystal structure, solution structure, thermal stability, free energy surface, and folding pathway network, are consistent with the properties of natural proteins. The existence of this kind of molecule deepens our understanding of proteins and impels us to define an "ideal protein" without inquiring whether the molecule actually occurs in nature.

Introduction

Proteins being essential to all living organisms are linear polymers composed of amino acids. Amino acid polymers may also be referred to as polypeptides, although scientists do not customarily use these terms interchangeably.[1] "Protein" generally refers to naturally occurring molecules having a particular sequence and a defined 3-dimensional (3D) structure, whereas "polypeptide" can refer to any polymers of amino acids, regardless of length, sequence, and structure. "Peptide" is generally reserved for a short oligomer that often lacks a stable conformation. Nevertheless, distinctions between these terms are scientifically ambiguous,[2] and this occasionally leads to confusion. To resolve the ambiguity, a satisfactory definition of a protein is needed that stipulates its necessary properties in physical terms, without the context of evolution. Here we report a crystal structure of a miniature protein, termed CLN025, that is a synthetic molecule consisting of 10 naturally occurring amino acids. CLN025 maintains the same topology in an aqueous solution as in the crystal. It exhibited remarkable structural stability in solution, undergoing a reversible, cooperative transition upon thermal denaturation. Molecular dynamics (MD) simulations coupled with a dihedral angle (DA) space mapping analysis indicate that structures are distributed on a funnel-like energy surface. These observations are consistent with generally agreed-upon properties of proteins, except for the small size of CLN025. Consequently, we introduce the concept of an "ideal protein", which can be defined deductively in terms of physics, regardless of the absence or presence of molecular genealogy.

Results

Design and Stability of CLN025. A novel protein, CLN025, was designed based on the sequence of chignolin, another synthetic, 10-residue molecule that was designed previously,[3] which adopted a steady structure in a chilled solution. Its atomic coordinates were determined by NMR (PDB accession code 1uao). In designing chignolin, the central eight amino acids were a consensus sequence derived by statistical analysis. A Gly residue was attached to each terminus as a flanking spacer, but these terminal residues were not optimized. In the present study, we examined the thermal stabilities of chignolin variants in which Gly1 and Gly10 were replaced with various amino acids (Table S1, Supporting Information). CLN025, with sequence YYDPETGTWY, was one of the stable variants, showing the second largest enthalpy difference upon thermal unfolding (ΔH_m). Thermal unfolding of CLN025 was reversible (>98%) and cooperative (Figure 1). The circular dichroism (CD) melting data could be fitted on a theoretical curve corresponding to a two-state phase transition (Figure 1B). The melting temperature (T_m) was found to be 343 K, 28 degrees higher than chignolin. Using thermodynamic parameters obtained from fitting calculations, the molar fraction (f_F) of CLN025 in a folded state is estimated to be 99% at 273 K,[23] whereas f_F of chignolin is 83% at 273 K, suggesting that CLN025 maintains a unique conformation in a chilled solution longer than does chignolin.

Crystal and Solution Structures of CLN025. CLN025 crystals were obtained from an aqueous solution at 283 K using a conventional screening method (Figure S1, Supporting Information), and its structure was determined at 1.11 Å resolution (Table S2, Supporting Information). The crystal structure reveals that CLN025 forms a β-hairpin at the central region, allowing the ends to contact each other (Figure 2A). At least six intramolecular hydrogen bonds and one salt bridge are identified (Table S3, Supporting Information), which appear to be dominant factors to stabilize the structure. The N- and C-terminal segments are well packed, so that no internal cavities occur. Several water molecules hydrate polar atoms on the

[†] Present address: Astbury Centre, Institute of Molecular and Cellular Biology, University of Leeds, LS2 9JT Leeds, UK.

[1] Creighton, T. E. *Proteins: Structures and Molecular Properties*, 2nd ed.; W. H. Freeman and Company: New York, 1993; p 4.
[2] IUPAC-IUB Joint Commission on Biochemical Nomenclature, *Biochem. J.* **1984**, *219*, 345–373.
[3] Honda, S.; Yamasaki, K.; Sawada, Y.; Morii, H. *Structure* **2004**, *12*, 1507–1518.

10.1021/ja8030533 CCC: $40.75 © 2008 American Chemical Society

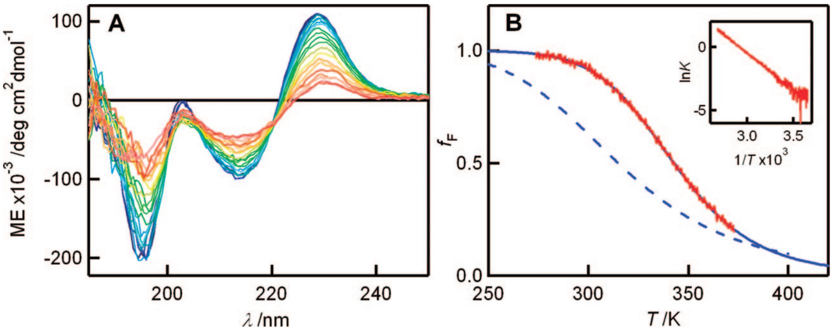

Figure 1. Thermal stability of CLN025. (A) Temperature dependence of CD spectra from 273 to 373 K (from blue to red). The protein was dissolved at approximately 1 mM in 20 mM potassium-phosphate buffer (pH 7.0). CD intensity is represented as molar ellipticity (ME). Isosbestic points are seen obviously. (B) Temperature dependence of f_F. Thermal unfolding data of CLN025 (red) were monitored at 229 nm at the heating rate of 1 K/min. A theoretical curve (blue) was introduced using a two-state transition model on the assumption of $\Delta C_p = 0$. A linear relation was found in the van't Hoff plot of the data (inset). A blue dashed line indicates f_F of chignolin that was previously reported.[3]

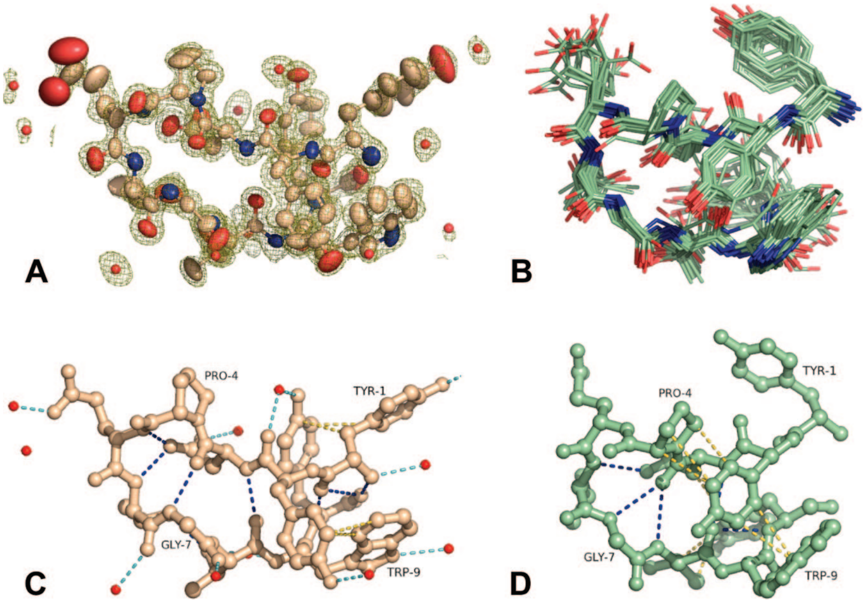

Figure 2. Structures of CLN025. (A) Crystal structure and electron density map. The structure is represented as a thermal ellipsoid model. Waters are shown as spheres. The electron density is drawn as a sigmaA-weighted $2F_o-F_c$ map contoured at 1σ. (B) Solution structures. Twenty structures having the lowest energy functions are superimposed. The bb-rmsd for 20 structures was 0.41 Å. (C) Primary interactions in the crystal structure. Blue and yellow dashes indicate polar interactions and hydrophobic contacts, respectively. (D) Primary interactions in the solution structure.

surface of CLN025, suggesting that the structure is also maintained by interactions with the hydration layer.

The solution structure of CLN025 was determined by NMR analysis (Table S4, Supporting Information). A sufficient number of distance constraints derived from homonuclear 2D measurements was obtained to give a satisfactory convergence in structure calculations (Figure 2*B*). The root-mean-square deviation of backbone coordinates (bb-rmsd) between the crystal and solution structures is 1.75 Å, showing that CLN025 retains a similar conformation in aqueous solution at 298 K. Analytical ultracentrifugation (AUC) measurements proved that the protein is monomeric in solution (Figure S2, Supporting Information), indicating that the crystal structure was not just a result of crystal packing forces, but that CLN025 forms a β-hairpin in solution without intermolecular interactions.

Further comparison of the two structures illustrates several similarities and differences (Figures 2C–D). Important interactions that determine this β-hairpin topology such as hydrogen bonds Asp3.N-Thr8.O, Asp3.O-Gly7.N, and Asp3.OD1-Thr6.N are conserved in both structures, whereas interactions between the terminal ends such as the hydrogen bond Tyr1.O-Tyr10.N and salt bridge Tyr1.N-Tyr10.OXT were not identified in the solution structure (Table S3, Supporting Information). The backbone dihedral angles ϕ and ψ of Asp3 and Gly7 differ in the solution and crystal structures, whereas the other residues are similar (Figure S3A, Supporting Information). Differences in side chain dihedral angles $\chi1$ and $\chi2$ of Tyr1, Tyr2, Thr6, Thr8, Trp9, and Tyr10 are relatively larger than those in Asp3, Pro4, and Glu5 (Figure S3B, Supporting Information). These comparisons suggests that crystallization would cause a slight structural change of CLN025, including a backbone distortion, reorientation of particular side chains, and anchoring of flexible regions (Figure S4, Supporting Information).

The topology of CLN025 is essentially the same as that of chignolin. The bb-rmsd between the solution structures of CLN025 and chignolin is 1.65 Å. Hydrogen bonds Asp3.N-Thr8.O, Asp3.O-Gly7.N, and Asp3.O-Thr8.N and an aromatic–aromatic interaction between the aromatic side chains of Tyr2 and Trp9 are found in both structures (Figure S5, Supporting Information). The latter interaction is supported by the characteristic positive CD peak at 229 nm (Figure 1A), which is probably caused by an edge-to-face exciton couplet[4–6] between Tyr2 and Trp9. An amide H–D exchange experiment revealed that amide protons of Gly7, Thr8, and Try10 in CLN025 are effectively shielded from solvent[23] (Table S5, Supporting Information), which are consistent with the hydrogen bonds identified in the solution structures (Table S3, Supporting Information). In contrast, the change in accessible surface area (ΔASA) after conversion from a folded to an extended structure differed significantly. The estimated total ΔASA, polar ΔASA, and nonpolar ΔASA of the CLN025 solution structure are 541, 165, and 376 Å^2, respectively, whereas those of chignolin are 471, 249, and 222 Å^2, obviously showing a greater nonpolar ΔASA was buried in CLN025 (Figure S6, Supporting Information). Therefore, the improved thermal stability of CLN025 may be attributed to additional interactions, especially hydrophobic interactions, of the terminal residues. This explains the large increment in ΔH_m of CLN025 (Table S1, Supporting Information) and the larger protection factors than chignolin (Table S5, Supporting Information). Judging from the NMR distance constraints, the interactions between Tyr1 and Asp3, Asp3 and Tyr10, and Thr8 and Tyr10 are most responsible for this effect (Figure S7, Supporting Information). No direct contact between Tyr1 and Tyr10 was seen in the solution structure.

Structural Dynamics of CLN025. To gain deeper insight into dynamic behavior of CLN025, one-microsecond MD simulations initiated from the extended conformation were carried out in explicit waters. The bb-rmsd between the MD and NMR structures suggest that folding/unfolding events occurred several times *in silico* (Figure 3A). These MD structures were then mapped on a DA space to visualize their distribution as shown in Figure 3B (see also Figure S8, Supporting Information). The space is defined by axes that were independently determined using principal component analysis (PCA) of protein segments having representative backbone conformations.[7] As shown in Figure 3B, the MD structures converge on a certain area that overlaps with both the crystal and solution structures, indicating that the MD simulation accurately reproduces the protein's environment. The simulation time was apparently long enough to reach an equilibrium state, so a free energy surface of CLN025 was produced by calculating the potential of mean force (PMF).[8] The resulting surface was funnel-shaped, consisting of a deep well and surrounding shallows (Figure 3C). The most populated MD structure, that is, the structure locating at the bottom of the well, agrees with the experimental structures (the bbRMSD to crystal and solution structures are 0.86 and 1.40 Å, respectively), including the orientations of side chains (Figure 3C). A similar MD analysis of a control peptide with sequence DTYGYWEPYT, shuffled-025, yielded completely different results (Figures 3E−G). Hence, the converged distribution and the funnel-like energy surface are consistent with a single structure for CLN025 that was demonstrated by crystallization. In contrast, the scattered distribution and flat energy surface of shuffled-025 suggests that the control peptide would fluctuate randomly in water.

Fifty thousand MD structures for CLN025 were classified into 519 subclusters using a supervised nearest-neighbor algorithm, and then their relationship was illustrated using an undirected network diagram (Figure 3D). In this folding pathway network, the node sizes indicate subcluster populations, and the link thicknesses denote the degree of traffic. Traffic along each link was almost symmetrical (Figure S9, Supporting Information), proving that the simulation reached the equilibrium state, where microstates interconvert at the almost same frequency. Although the network appears complicated, most of the nodes belong to rare subclusters. If the rare subclusters are dismissed, the CLN025 network appears rather simple as compared with that of shuffled-025 (Figure S10, Supporting Information). This implies that the structural dynamics of CLN025 in the equilibrium state can be described effectively as transitions between a relatively small numbers of dominant subclusters. The networks of CLN025 and shuffled-025 involve several power-law relations (Figures 3H−I and S11, Supporting Information), as reported in other protein simulations,[9] indicating that the global topology of both folding pathways can be considered as a scale-free and small world network, and that the nodes size and link thickness for dominant subclusters follow a Zipf-like distribution only in case of the CLN025 network (see Supporting Information for detailed analysis).

Discussion

In 1840, F. L. Hünfeld found that tabular protein crystals grew in a drop of blood.[10] J. B. Sumner crystallized urease isolated from jack beans in aqueous−organic solvents in 1926.[11] A significant outcome of these discoveries is the concept that a protein is a macromolecule possessing a distinct 3D structure. The concept cemented by the pioneering X-ray crystallography of myoglobin and hemoglobin, accomplished by J. C. Kendrew and M. F. Perutz in 1958−60.[12,13] Since then, a great number of proteins crystal structures have been reported.

Adoption of a well-defined 3D structure is now considered crucial to accurate and efficient functioning of proteins. Therefore, the ability to be crystallized is a distinct property of proteins that may distinguish them from other nonvital polypeptides. Furthermore, the protein structures are not rigid. It is generally accepted that flexibilities such as hinge-bending motions, ligand-induced conformational changes, and global unfolding are important for proteins to carry out molecular recognition, catalysis, signal transduction, and transportation across membranes.

The present study demonstrates the defined 3D structure of CLN025. This structure also shows the dynamic behavior, including the reorientation of side chains and reversible folding. Thus, by the criteria above, CLN025 can be regarded as a protein. Probably, however, some scientists may argue that CLN025 is a peptide, rather than a protein, because of its small size. To our knowledge, CLN025 is the smallest linear polypeptide that has been shown to adopt a unique structure in aqueous solutions, that undergoes a cooperative structural transition, and that can be crystallized. A somewhat arbitrary borderline has customarily been used to discriminate between peptide and protein, whereby a molecule consisting of 50 or fewer amino acids is considered a peptide.[2] However, this borderline seems

[4] Grishina, I. B.; Woody, R. W. *Faraday Discuss.* **1994**, *99*, 245–262.

[5] Cochran, A. G.; Skelton, N. J.; Starovasnik, M. A. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 5578–5583.

[6] Guvench, O.; Brooks, C. L., III. *J. Am. Chem. Soc.* **2005**, *127*, 4668–4674.

[7] Sawada, Y.; Honda, S. J. *Biophys. J.* **2006**, *91*, 1213–1223. Sawada, Y.; Honda, S. *J. Comput. Aided Mol. Des.*, in press (DOI http://dx.doi.org/10.1007/s10822-008-9248-x).

[8] Zhou, R.; Berne, B. J.; Germain, R. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 14931–14936.

[9] Rao, F.; Caflisch, A. *J. Mol. Biol.* **2004**, *342*, 299–306.

[10] Hunefeld, F. L. *Die Chemismus in der thierischen Organization*; Brockhaus: Leipzig, 1840; pp 158−163.

[11] Sumner, J. B. *J. Biol. Chem.* **1926**, *69*, 4355–4441.

[12] Kendrew, J. C.; Bodo, G.; Dintzis, H. M.; Parrish, R. G.; Wyckoff, H.; Phillips, D. C. *Nature* **1958**, *181*, 662–666.

[13] Perutz, M. F.; Rossman, M. G.; Cullis, A. F.; Muirhead, H.; Will, G.; North, A. C. T. *Nature* **1960**, *185*, 416–422.

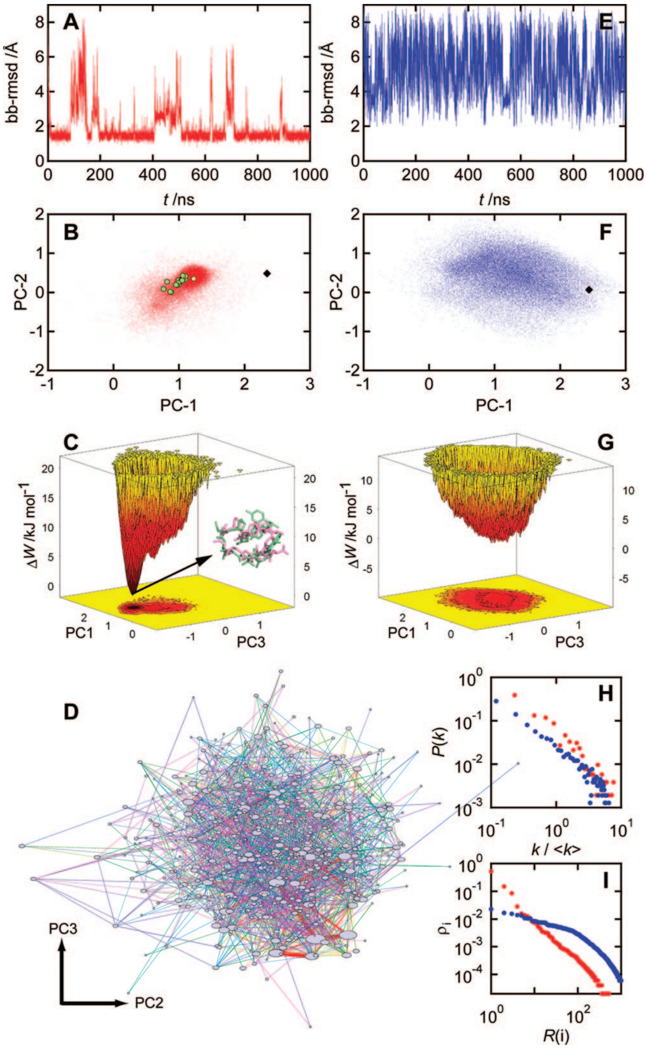

Figure 3. Structural dynamics of CLN025. (A) Time course of the bb-rmsd of MD structures ($T = 373$ K) from NMR structure. (B) Distribution of the MD structures (red dots) mapped on the DA space (PC1-PC2 axes). Crystal structure (yellow circle), solution structures (green circles), and extended structure as an initial condition of the MD calculation (black diamond) are also mapped. (C) Free energy surface energy projected on the DA space (PC1−PC3 axes). Comparison of the most populated MD structure with the solution structure is also shown. (D) Folding pathway network mapped on the DA space (PC2−PC3 axes). Nodes and links correspond to the subclusters and trajectories, respectively. Thickness of each link corresponds to the sum of forward and reverse traffic. (E−G) Structural dynamics of the control peptide, shuffled-025, having a shuffled sequence DTYGYWEPYT. (H) Degree distribution of the folding networks. Probability $P(k)$ of a node having k degree follows: $P(k) \approx k^{-\gamma}$. The order of power-law γ is 1.63 and 1.44 for CLN025 and shuffled-025, respectively. (I) Zipf distribution of node size of the folding networks. Plot of normalized population ρ of the i-th subcluster vs its rank R follows: $\rho_i \approx R(i)^{-\gamma}$. The order of power-law γ is 1.63 for CLN025.

to have no logical basis (Figure S12, Supporting Information). Rather, it resulted from practical experiences indicating that most small peptides isolated from nature or resulting from protein digestion did not show characteristics like those seen for CLN025. Some scientists may not regard CLN025 as a protein because it is not a natural product. We would argue that this is not an appropriate criterion to standardize a term in molecular science, because the properties of a molecule are specified by its chemical structure, not by its origin.

The sequence space of a polypeptide is too vast to investigate systematically, as this would require a time greater than the age of the universe.[14] Thus, nature has not sampled all possible protein sequences. Protein evolution is also considered to involve some irrationality attributed to accidental events that occur periodically in history. Therefore, a taxonomic and inductive definition based on proteins currently in existence may not be logically consistent. We therefore conclude that a physical and deductive definition of proteins is indispensable for clarifying the ambiguous usage described above. Here we propose the definition of an “ideal protein”: a polypeptide that possesses a single and smooth funnel-like energy surface. In other words, an ideal protein is defined as a polypeptide that perfectly satisfies the consistency principle or the minimum frustration principle. Today, the funnel-like energy surface,[15] consistency principle,[16] and minimum frustration principle[17] are widely accepted as terms characterizing essential protein properties. In the present study, we use these terms not to describe properties, but as a definition of a protein.

The suggested criterion of an ideal protein refers only to the shape of its energy surface, and is associated with neither the length nor the origin of the molecule. Furthermore, an ideal protein can be theoretically discussed, regardless of whether it occurs in nature or will be synthesized in the future. The term “ideal protein” was coined by analogy with the chemical terms “ideal gas” and “ideal solution”. Like an ideal gas and an ideal solution, an ideal protein that perfectly fits its definition may not exist in the real world. However, defining an ideal molecule and analyzing the deviations of real molecules from an ideal one will lead to deeper understanding of protein entity.

Our definition of an ideal protein posits a defined 3D structure that exhibits some flexibility. The defined 3D structure corresponds to the bottom of the funnel and the flexibility corresponds to transitions among structures near the bottom. Thus, the results of the present study indicate that CLN025 is close to an ideal protein. In CLN025, the bottom of the funnel accorded with a hub node that has the most links in its folding pathway network. What does it mean that such a small molecule shows characteristics of an ideal protein? Considering sequence diversity, we should not undervalue the existence of such a small molecule, because the ratio of the number of near-ideal proteins like CLN025 (at least one) against the number of all possible sequences for a 10-residue polypeptide ($20^{10} \approx 10^{13}$) is much larger than the ratio of the number of known proteins (10^{10} at the maximum) against the number of all possible sequences for a 100-residue polypeptide ($20^{100} \approx 10^{130}$). In fact, CLN025 is not solely the exception. Recent progress in molecular design has produced many short linear peptides that adopt stable structures (see refs 18–22 and references therein). As examples,

(14) Morton, J. S. *Impact* **1980**, *90*, 1.
(15) Onuchie, J. N.; Wolynes, P. G.; Luthey-Schulten, Z.; Socci, N. D. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 3626–3630.
(16) Go, N. *Annu. Rev. Biophys. Bioeng.* **1983**, *12*, 183–210.
(17) Bryngelson, J. D.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 7524–7528.
(18) Gellman, S. H. *Curr. Opin. Chem. Biol.* **1998**, *2*, 717–725.
(19) Ramirez-Alvarado, M.; Kortemme, T.; Blanco, F. J.; Serrano, L. *Bioorg. Med. Chem.* **1999**, *7*, 93–103.
(20) Searle, M. S. *J. Chem. Soc., Perkin Trans. 2* **2001**, 1011–1020.
(21) Andersen, N. H.; Olsen, K. A.; Fesinmeyer, R. M.; Tan, X.; Hudson, F. M.; Eidenschink, L. A.; Farazi, S. R. *J. Am. Chem. Soc.* **2006**, *128*, 6101–6110.
(22) Hughes, R. M.; Waters, M. L. *Curr. Opin. Struct. Biol.* **2006**, *16*, 514–524.
(23) According to an empirical method reported by Lin et al., the protection factors of Asp3, Gly7, Thr8, and Tyr10 determined in the amide H–D exchange experiments suggest that the molar fraction of CLN025 in a folded state is 96−98% at 278 K, which quantitatively agrees with the value estimated from the CD melting data. Lin, J. C.; Barua, B.; Andersen, N. H. *J. Am. Chem. Soc.* **2004**, *126*, 13679–13684.

stabilities of 10–16-residue β-hairpins are summarized in Figure S13 (Supporting Information). Although it has not been reported whether these peptides can be crystallized or not, these successes suggest that the number of short peptides that satisfy the definition of an ideal protein may be much more than our current knowledge.

The number of naturally occurring proteins is very limited compared to the vast sequence space of a polypeptide. Nevertheless, the present world was created from this limited area. Therefore, it would be worthwhile to investigate how large the area of ideal proteins is, to better understand the present world, especially when evaluating "chance and necessity" in the chemical evolution of ancient proteins. It should be emphasized that whereas the sequence space of a 100-residue polypeptide is too vast to investigate completely, the sequence space of a 10-residue polypeptide may be explored by means of high-throughput experiments and high-performance calculations.

Acknowledgment. We thank Beckman Coulter K. K. (Tokyo) for help with AUC measurements, NMR Centre, Astbury Centre, University of Leeds, and Dr. Arnout P. Kalverda for the dedicated support for the H–D exchange experiments. This work was funded in part by a grant from the New Energy and Industrial Technology Development Organization (NEDO) of Japan.

Supporting Information Available: Materials and Methods (Protein synthesis and characterization, Structure determination, MD simulation and DA space mapping analysis, Visualization), Results and Discussion (Topological analysis of folding pathway networks), Figures S1−S13, Tables S1−S5, and CIF files for the crystal and solution structures of CLN025. This material is available free of charge via the Internet at http://pubs.acs.org.

JA8030533