AIP | The Journal of Chemical Physics

Physical origins of the high structural stability of CLN025 with only ten residues

Satoshi Yasuda, Tomohiko Hayashi, and Masahiro Kinoshita

Citation: The Journal of Chemical Physics **141**, 105103 (2014); doi: 10.1063/1.4894753
View online: http://dx.doi.org/10.1063/1.4894753
View Table of Contents: http://scitation.aip.org/content/aip/journal/jcp/141/10?ver=pdfcov
Published by the AIP Publishing

THE JOURNAL OF CHEMICAL PHYSICS **141**, 105103 (2014)

Physical origins of the high structural stability of CLN025 with only ten residues

Satoshi Yasuda, Tomohiko Hayashi, and Masahiro Kinoshita[a)]
Institute of Advanced Energy, Kyoto University, Uji, Kyoto 611-0011, Japan

(Received 29 June 2014; accepted 22 August 2014; published online 10 September 2014)

CLN025, a peptide with only 10 residues, folds into a specific β-hairpin structure (this is referred to as "native structure"). Here we investigate the stabilization mechanism for CLN025 using our free-energy function F. F comprises two components, the hydration entropy and the component related to the energetic dehydration effect. The former component is calculated using the hybrid of the angle-dependent integral equation theory (ADIET) and our recently developed morphometric approach. The ADIET is a statistical-mechanical theory applied to a molecular model for water. The latter component is calculated in a simple but judicious manner accounting for physically the most important factors: the break of polypeptide-water hydrogen bonds and formation of polypeptide intramolecular hydrogen bonds upon structural change to a more compact one. We consider the native structure, compact nonnative structures newly generated, and a set of random coils mimicking the unfolded state. F and its components are calculated for all the structures considered. The loss of the polypeptide conformational entropy upon structural transition from the unfolded state to a compact structure is also estimated using a simple but physically reasonable manner. We find that the key factor is the water-entropy gain upon folding originating primarily from an increase in the total volume available to the translational displacement of water molecules in the system, which is followed by the reduction of water crowding. The amino-acid sequence of CLN025 enables it not only to closely pack the backbone and side chains including those with large aromatic groups but also to assure the intramolecular hydrogen bonding upon burial of a donor and an acceptor when the backbone forms the native structure. The assurance leads to essentially no enthalpy increase upon folding. The close packing brings a water-entropy gain which is large enough to surpass the conformational-entropy loss. By contrast, it is not possible for the design template of CLN025, GPM12, to realize the same type of structure formation. There are significantly many compact structures which are equally stable in terms of F, and due to the conformational-entropy effect, the unfolded state is favorably stabilized.
 [http://dx.doi.org/10.1063/1.4894753]

I. INTRODUCTION

A protein, which is a polypeptide comprising 20 kinds of amino-acid residues, folds into its unique three-dimensional structure called the native structure (NS) in aqueous solution under the physiological condition.[1] The elucidation of the folding mechanism is one of the most challenging problems in chemical physics, biophysics, and biochemistry. The number of residues usually exceeds 50. In general, a short polypeptide does not form a specific folded structure and takes an ensemble of unfolded structures which resemble random coils. For example, GPM12 (GYDDATKTFG), whose central residues correspond to the central 8 residues of G-peptide (a dissected fragment comprising residues 41–56 of the B1 domain of protein G) is known to be in the unfolded state.[2] Chignolin with only 10 residues[2] (GYDPETGTWG) is a notable exception (it was designed using GPM12 as the template). A nuclear magnetic resonance (NMR) experiment has shown that chignolin forms a specific β-hairpin structure though the denaturation temperature T_m is only 315 K and the content of the denatured state coexisting at 273 K is significantly high (∼20%).[2] CLN025 (YYDPETGTWY) possesses a variant of the amino-acid sequence of chignolin designed to improve the thermal stability.[3] According to X-ray crystallography and NMR experiments, CLN025 forms the β-hairpin structure in common with chignolin, but its T_m is 28 degrees higher than that of chignolin (i.e., 343 K).[3] Revealing the physical origins of the unexpectedly high thermal stability of CLN025 expands the understanding of the protein folding mechanism and provides a physical basis for predicting the foldability of a polypeptide.

It has been suggested on the basis of molecular dynamics (MD) simulation[3–6] and experimental[7] results that the formation of intramolecular hydrogen bonds, salt bridge between the charged termini, aromatic-aromatic (Ar–Ar) attractive interaction, and hydrophobic interaction between nonpolar groups are responsible for the unusually high stability of CLN025. (The Ar–Ar attractive interaction consists of van der Waals and electrostatic attractive interactions.[4]) The first two factors are based on the contact of polypeptide groups or atoms with unlike charges followed by the gain of electrostatic attractive interactions. We note that such a contact in aqueous solution does not lead to sufficiently high

[a)]Author to whom correspondence should be addressed. Electronic mail: kinoshit@iae.kyoto-u.ac.jp.

stabilization. Before the contact, the groups or atoms with negative charges interact with water hydrogens with positive charges and those with positive charges interact with water oxygens with negative charges, maintaining group-water or atom-water electrostatic attractive interactions. The contact is unavoidably accompanied by the loss of these interactions. The gain of intramolecular electrostatic attractive interactions and the loss of electrostatic attractive interactions *with water molecules* are cancelled out,[8] or the loss is often larger.[9] Likewise, a gain of intramolecular van der Waals attractive interactions and the loss of van der Waals attractive interactions *with water molecules* accompanied are somewhat compensating.[8,9] Taken together, even when the MD simulation is performed in explicit water, a calculation of the free-energy landscape alone,[3] an analysis on intramolecular interaction energies,[4,5] or the visualization of the folding process[6] for CLN025 is not complete: A free energy fully accounting for the effect of *dehydration* must be decomposed into various constituents, and sufficiently many different structures including the native structure are to be compared in terms of the constituents. These conclusions were drawn in our earlier works[8,9] using statistical-mechanical theories for hydration of biomolecules combined with molecular models for water. The hydrophobic interaction mentioned above, which represents the contact of nonpolar groups, is based on the conventional view:[10] The water adjacent to a nonpolar group is entropically unstable due to water structuring, and the contact of nonpolar groups leads to the reduction of such unfavorable water followed by a water-entropy gain. In this view, however, only the water molecules near nonpolar groups are considered. As explained in the next paragraph, water molecules *in the system* are to be taken into account in discussing the water entropy.

We have pointed out that protein folding is driven by the water-entropy gain originating primarily from the excluded-volume (EV) effect.[11–17] Upon protein folding, the EV (i.e., the volume of the space which the centers of water molecules cannot enter) decreases to a large extent, which is followed by a corresponding increase in the total volume available to the translational displacement of water molecules in the system and a reduction in the water crowding. The folding thus leads to a large gain of the water entropy. The formation of α-helix or β-sheet leads to a significant decrease in the EV (see Figs. 1(a) and 1(b)). At the same time, the formation compensates the loss of hydrogen bonds with water molecules by ensuring intramolecular hydrogen bonds. Hence, these secondary structures are very advantageous units. Close, efficient packing of side chains, which reduces the EV to a considerable extent (see Fig. 1(c)), is also crucial. Protein folding is characterized by the formation of as much α-helix and β-sheet as possible and close packing of the backbone and side chains.[15] Using our methods wherein the water-entropy effect is treated as the key factor, we have been successful in elucidating the mechanisms of protein folding[11–17] and cold,[18–20] pressure,[21,22] and heat denaturating of a protein.[23,24] Terazima *et al.*[12,25] developed a novel experimental technique which allows us to directly measure the enthalpic change upon protein folding at a prescribed temperature. They showed that apoplastocyanin (apoPC) folding at

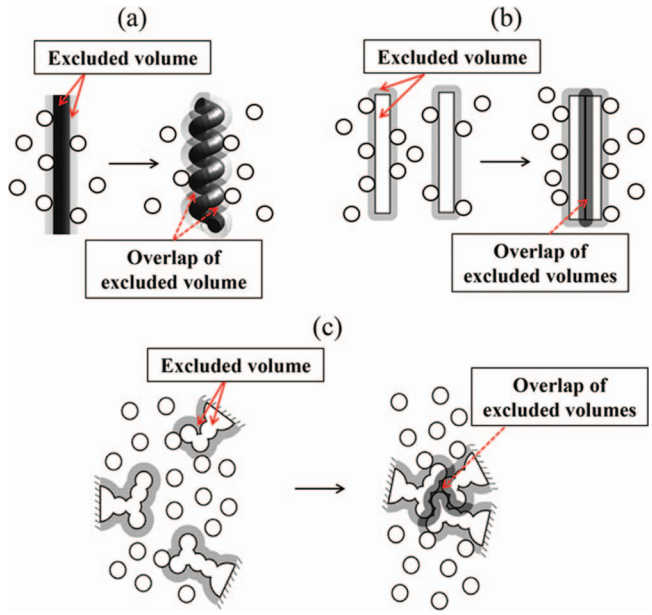

FIG. 1. (a) Formation of α-helix by a portion of the backbone. (b) Lateral contact of (formation of β-sheet by) portions of the backbone. (c) Close packing of side chains. The total excluded volume decreases by the overlapped volume marked in dark gray, leading to a corresponding increase in the total volume available to the translational displacement of water molecules in the system.

298 K is accompanied by a large enthalpic increase. The break of protein-water hydrogen bonds upon folding cannot always be compensated by the formation of intramolecular hydrogen bonds, giving rise to this enthalpy increase. Thus, the large loss of the protein conformational entropy plus the enthalpic loss caused by protein folding is surpassed by a great water-entropy gain. The water-entropy gain comprises the translational and rotational components. However, the translational component takes ~95% (the rotational one takes only ~5%) of the total gain.[12]

On the basis of the above picture of protein folding, we have recently developed a free-energy function (FEF) F.[26,27] F is dependent on the protein structure and expressed as $F = \Lambda - TS_{\mathrm{VH}}$ where Λ and S_{VH}, respectively, are the energetic and entropic components. S_{VH} is the hydration entropy calculated using our statistical-thermodynamic theory with a molecular model for water. $-S_{\mathrm{VH}}$ represents magnitude of the water-entropy loss upon protein insertion. In the calculation of the energetic component Λ, a fully extended structure is treated as the reference one: Λ represents the energy change for the protein and water upon transition from the fully extended structure to a prescribed, more compact one. Λ is calculated in a simple manner which still accounts for physically the most important factors: the break of protein-water hydrogen bonds (i.e., the principal component of the energetic dehydration) and formation of protein intramolecular hydrogen bonds. In our earlier works,[26,27] it was demonstrated that F exhibits exceptionally high performance in discriminating the native fold from a number of misfolded decoys.

In the present study, we analyze the structural stability of CLN025 and GPM12 by applying the FEF and its energetic and entropic components to small polypeptides. The entropic

component is further decomposed into the contributions from the backbone and side-chain packing efficiencies. The analyses on these constituents of the FEF allow us to reveal the true physical origins of the unexpectedly high thermal stability of CLN025. The NS, compact nonnative structures, and a set of random coils mimicking the unfolded state are considered for CLN025. Since GPM12 does not take a specific folded structure, only compact nonnative structures and a set of random coils are considered for it. The compact nonnative structures are generated via the two routes, MD simulation and coarse-grained normal mode analysis (CGNMA).[28] Using these two methods, we can generate a great variety of compact structures. The compact nonnative structures include those which are quite similar to the NS. The FEF and its constituents are calculated for all the structures considered. When only the compact structures are compared, there is no need to account for the effect of the polypeptide conformational entropy. When the compact structures are compared with the unfolded state, however, the conformational-entropy effect must also be incorporated in the FEF: The incorporation is made in a simple but physically reasonable manner.

Our major findings are as follows. The key factor is the water-entropy gain upon folding originating primarily from an increase in the total volume available to the translational displacement of water molecules in the system, which is followed by the reduction of water crowding. The NS of CLN025 is characterized by an amino-acid sequence which enables it not only to assure the intramolecular hydrogen bonding upon burial of a donor and an acceptor (i.e., to make Λ zero) but also to accomplish close packing of the backbone and side chains when the backbone forms the β-hairpin structure. A fortuitously large gain in the water entropy is realized by the following two features: There are four aromatic side chains with large sizes (TYR1, TYR2, TRP9, and TYR10) in the sequence; and the backbone and side chains including these are efficiently packed. Close packing in which aromatic side chains with large sizes participate is crucially important with respect to the water-entropy gain because it provides considerably large reduction in the EV. (Although the α-helix structure can assure the formation of intramolecular hydrogen bonds, close packing of the backbone and side chains is not achievable.) The NS is considerably more stable than the unfolded state though the latter is much more favorable from the standpoint of the conformational entropy. CLN025 folding, upon which essentially no enthalpy change occurs, is driven by the water-entropy gain. The behavior of GPM12 is substantially different. Structures with $\Lambda = 0$ can be formed, but they always lack sufficiently close packing of the backbone and side chains. On the other hand, structures with close packing of the backbone and side chains can be formed, but they suffer positive values of Λ that are large enough to vitiate the close packing. Since GPM12 has only two aromatic side chains with large sizes (TYR2 and PHE9), the water-entropy gain brought by close packing of the backbone and side chains is not as large as that in the case of CLN025. The conformational-entropy effect is predominant and the unfolded state becomes the most stable. These results are in good accord with the experimental observations described above.

II. FREE-ENERGY FUNCTION

A. Definition

Our free-energy function F is expressed for a prescribed structure of a protein as

$$F = E_{\mathrm{I}} + \mu, \tag{1}$$

where E_{I} is the protein intramolecular energy and μ is the hydration free energy (i.e., excess chemical potential) that is the most important thermodynamic quantity of protein hydration. The hydration free energy is given by

$$\mu = E_{\mathrm{VH}} - TS_{\mathrm{VH}}, \tag{2}$$

where E_{VH} represents the hydration energy (i.e., the protein-water interaction energy generated plus the energy change due to the structural reorganization of water upon protein insertion) and S_{VH} represents the hydration entropy. Defining Λ by

$$\Lambda = E_{\mathrm{I}} + E_{\mathrm{VH}}, \tag{3}$$

we obtain

$$F = \Lambda - TS_{\mathrm{VH}}. \tag{4}$$

F is scaled by $k_{\mathrm{B}}T_0$ (k_{B} is the Boltzmann constant and $T_0 = 298$ K) and T is set at T_0 in the present study. $-S_{\mathrm{VH}}$ is always positive and Λ is either positive or zero. Λ, S_{VH}, and F are largely dependent on the protein structure. The procedures of calculating S_{VH} and Λ are briefly described below (the FEF is explained in our earlier publications[26,27] as well).

It should be noted that μ is independent of the protein insertion condition, isobaric or isochoric, but E_{VH} and S_{VH} are not.[29] We consider isochoric condition for the following reasons: (i) It is free from the effects of compression or expansion of the bulk water and more suited to physical interpretation of a change in a thermodynamic quantity of hydration; (ii) the structural transition of a protein occurs with the system pressure and volume almost unchanged[12,25] (the EV of a more compact structure is smaller but the partial molar volume is almost independent of the compactness); and (iii) it is much more convenient in a theoretical treatment. It follows from (ii) that the energy change upon structural transition is equal to the enthalpy change.

B. Entropic component

A feature of our FEF is that we do not regard water as a dielectric continuum. A water molecule is modeled as a hard sphere with diameter $d_{\mathrm{S}} = 0.28$ nm in which a point dipole and a point quadrupole of tetrahedral symmetry are embedded.[30,31] We employ the angle-dependent integral equation theory (ADIET),[29–34] a statistical-mechanical theory for molecular liquids. In the ADIET the effect of the molecular polarizability is taken into account using the self-consistent mean field (SCMF) theory.[30,31] At the SCMF level the many-body induced interactions are reduced to pairwise additive potentials involving an effective dipole moment. The effective dipole moment thus determined at 298 K and 1 atm is about 1.42 times larger than the *bare* gas-phase dipole moment. We calculated the hydration free energy of

a hard-sphere solute with diameter 0.28 nm using our water model and the angle-dependent integral equation theory with the hypernetted-chain closure:[34] The value obtained is 3.56 kcal/mol at 300 K that is in excellent agreement with the values from Monte Carlo simulations for more popular water models: 3.56 kcal/mol at 300 K[35] for TIP4P and 3.65 kcal/mol at 298 K[36] for SPC/E.

S_{VH} is fairly insensitive to the solute-water interaction potential as proved in our earlier work.[37] For example, the three quantities, μ, S_{VH}, and E_{VH}, are calculated for a spherical solute with diameter 0.28 nm at 298 K using the ADIET. For the hard-sphere solute with zero charge, the calculated values are $\mu = 5.95k_BT_0$, $S_{VH} = -9.22k_B$, and $E_{VH} = -3.27k_BT_0$. When the point charge $-0.5e$ (e is the elementary electric charge) is embedded at its center, the calculated values are $\mu = -32.32k_BT_0$, $S_{VH} = -10.11k_B$, and $E_{VH} = -42.43k_BT_0$. Therefore, a protein can be modeled as a set of fused hard spheres just for calculating its S_{VH}. We note that E_{VH}, which is influenced by the protein-water interaction potential, is separately treated in Λ.

The calculation of S_{VH} is performed by combining the ADIET with the morphometric approach (MA).[38–40] This combination allows us to finish the calculation quite rapidly notwithstanding the employment of a molecular model for water and the application to a complexly shaped solute like a protein. The idea of the MA is to express S_{VH} by the linear combination of only four geometric measures of a solute molecule,

$$S_{VH}/k_B = C_1V_{ex} + C_2A + C_3X + C_4Y. \tag{5}$$

Here, V_{ex} is the EV, A is the water-accessible surface area, and X and Y are the integrated mean and Gaussian curvatures of the accessible surface, respectively. The water-accessible surface is the surface that is accessible to the centers of water molecules. The volume that is enclosed by this surface is the EV. C_1 is completely independent of the solute-water interaction potential. Though S_{VH} is influenced by all the four terms, C_1V_{ex} is the principal term at normal temperature and pressure. This is the reason for the fair insensitivity of S_{VH} to the solute-water interaction potential. The contribution from the water molecules near the solute molecule is represented by $C_2A + C_3X + C_4Y$. In the MA, the solute shape enters S_{VH} only via the four geometric measures. The four coefficients (C_1–C_4), which are dependent only on the thermodynamic state of water, can be determined in simple geometries. They are calculated from the values of S_{VH} for hard-sphere solutes with various diameters immersed in our model water.

The procedure of calculating S_{VH} of a protein with a prescribed structure comprises the following four steps.

(1) S_{VH} of a hard-sphere solute with diameter d_U is calculated using the ADIET. The values of S_{VH} are prepared for sufficiently many different values of d_U ($0.6 \le d_U/d_S \le 10$).

(2) The four coefficients are determined by the least square fitting applied to the following equation for hard-sphere solutes (i.e., Eq. (5) applied to hard-sphere solutes):

$$S_{VH}/k_B = C_1(4\pi R^3/3) + C_2(4\pi R^2) + C_3(4\pi R) + C_4(4\pi),\ R = (d_U + d_S)/2. \tag{6}$$

(3) The four geometric measures of a protein (V_{ex}, A, X, and Y) with a prescribed structure are calculated by means of an extension[39] of Connolly's algorithm.[41,42] The x-y-z coordinates of the protein atoms used as part of the input data to account for the polyatomic structure at the atomic level. The diameter of each atom is set at the sigma-value of the Lennard-Jones (LJ) potential parameters which are taken from the f99SB force field[43] employed in the Amber12 program package.[44]

(4) S_{VH} of a protein with a prescribed structure is obtained from Eq. (5) in which the four coefficients determined in step (2) are used. Smaller $-S_{VH}$ implies a closer, more efficient packing of the backbone and side chains.

We emphasize that the four geometric measures, V_{ex}, A, X, and Y, are largely dependent on the protein structure. When the solvent is simple fluid (i.e., the solvent particles interact through radial-symmetric potential like the LJ one), the solvation entropy of a protein with a prescribed structure can be calculated via two routes. One of them is the three-dimensional integral equation theory[11,45] which is capable of accounting for the structure in the atomic details. The other is the hybrid of the radial symmetric integral equation theory and the MA. We calculated the solvation entropies for protein G (the number of residues is 56) with a number of different structures via the two routes.[39] The error in the hybrid is only within ±2%. Thus, the MA is highly accurate in accounting for the polyatomic structure.

The high reliability of the ADIET-MA hybrid in calculating S_{VH} has been demonstrated for such subjects as the quantitative reproduction of the experimentally measured changes in thermodynamic quantities upon apoPC folding,[12] elucidation of the molecular mechanisms of cold[18–20] and pressure[21] denaturating of a protein, proposal of a reliable measure of the thermal stability of a protein,[23,24] and characterization of experimentally determined native-structure (NS) models of a protein.[46] S_{VH} comprises the translational and rotational components. However, the rotational component, which possesses no EV term, is much smaller than the translational one.[12–14,17]

C. Energetic component

Λ defined by Eq. (3) is calculated by choosing a fully extended structure as the reference one. The fully extended structure possesses the maximum number of hydrogen bonds with water molecules but no intramolecular hydrogen bonds ($\Lambda = 0$). When the protein structure changes from the fully extended structure to a more compact one, some donors and acceptors (there are four different donor-acceptor combinations: (N, O), (O, N), (O, O), and (N, N)) are buried in the interior after the break of hydrogen bonds with water molecules (e.g., NH···W and O···W where W denotes a water molecule). There is no problem if intramolecular

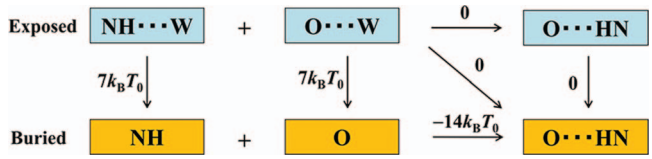

FIG. 2. Thermodynamic cycle for calculating Λ. "W" and "···" represent a water molecule and a hydrogen bond, respectively. $T_0 = 298$ K. This figure is illustrated for the case where N is the donor and O is the acceptor.

hydrogen bonds (e.g., NH···O) are formed. However, such bonds are not always formed, often giving rise to positive Λ.

Our procedure of calculating Λ can be summarized in the thermodynamic cycle illustrated in Fig. 2. When a donor and an acceptor are buried in the interior after the break of hydrogen bonds with water molecules, if they form an intramolecular hydrogen bond, we impose no penalty. On the other hand, when a donor or an acceptor is buried with no intramolecular hydrogen bond formed, we impose the penalty of $7k_BT_0$. The value $7k_BT_0$ is based on the energy-decrease of $-14k_BT_0$ arising from hydrogen-bond formation between two formamide molecules in a nonpolar liquid.[47] To determine if each of the donors and acceptors is buried or not, its water-accessible surface area is calculated by means of Connolly's algorithm.[41,42] The donor or acceptor is considered buried if the area is smaller than 0.001 Å^2. To determine if an intramolecular hydrogen bond is formed or not, we use the criteria proposed by McDonald and Thornton.[48] We examine all the donors and acceptors for backbone-backbone, backbone-side chain, and side chain-side chain intramolecular hydrogen bonds and calculate Λ.

It is assumed that upon structural change from the fully extended structure to a more compact one, the gain of intramolecular van der Waals attractive interactions and the loss of van der Waals attractive interactions with water molecules accompanied are cancelled out. Further, the energetic component is not considered for nonpolar groups. These are justifiable because the break of hydrogen bonds with water molecules, when they are not compensated by the intramolecular hydrogen bonding, should be the most serious and form the dominating factor of the energetic component. The torsion energy is not considered, either. The structures to be treated share the property that the torsion energy is reasonably low (i.e., only the structures with sufficiently low torsion energies are chosen), and the difference between two structures in the torsion energy makes no essential contribution to the difference in the energetic component.

D. Performance of discriminating native fold from misfolded decoys

We have tested our FEF F for a total of 133 proteins in 8 decoy sets and demonstrated that it discriminates the native fold from the misfolded decoys with almost 100% accuracy.[26,27] The discrimination was not successful only for five proteins. For them, however, the NS models were determined using NMR under acidic conditions or portions of the terminus sides were removed with the result of very high percentages of the secondary structures lost (i.e., they were not real). F is far superior to any of the previously reported functions. The approximations employed in calculating Λ, S_{VH}, and F can thus be justified by these successful results. It is worthwhile to emphasize that the calculation of F is finished in less than 1 s per structure on a standard workstation even for a very large protein.

III. STRUCTURE MODELS CONSIDERED

A. Native structure models for CLN025

For the NS of CLN025, we consider one model and 20 models (Models 1−20) taken from X-ray crystallography and NMR experimental results, respectively.[3] Hereafter, they are referred to as "X-ray model" and "NMR models," respectively. The coordinates of hydrogen atoms cannot be obtained by the X-ray diffraction. We give hydrogen atoms to the X-ray model using the Amber12 program package[44] with the f99SB force field.[43] There is no NS model for GPM12.

B. Compact nonnative structures obtained via route 1

Compact nonnative structures (NNSs) are generated for CLN025 and GPM12 via two routes. In the first route, we use MD simulations in the NPT ensemble at 298 K and 1 atm. The Amber12 program package[44] with the f99SB force field[43] and the Generalized-Born (GB) model[49,50] is employed. The SHAKE algorithm[51] is used to fix the length of the covalent bonds involving hydrogen atoms. Ten independent runs are carried out using 10 different random coils (RCs) as the initial structures, respectively. How to generate the RCs are described in "Unfolded State." In each simulation performed for 10 ns, the structure at every 10 ps is stored on a file. Thus, a total of 10 000 structures are obtained for each of CLN025 and GPM12.

C. Compact nonnative structures obtained via route 2

In order to assure that sufficiently many compact NNSs are considered in the present study, we generate a set of structures near a fixed structure using the coarse-grained normal mode analysis (CGNMA).[28] Starting from the fixed structure, the CGNMA constructs a number of modified structures. For CLN025, the fixed structure is taken to be the NS model giving the FEF the lowest value (Model 14). For GPM12, the structure with the lowest FEF is chosen as the fixed structure from among the NNSs obtained via route 1. A total of 2000 structures are obtained for each of CLN025 and GPM12. The structures generated by the CGNMA do not largely differ from the fixed structure: The root mean square deviation (RMSD) for C_α atoms from the fixed structure is averaged at 1.27 Å for CLN025 and 1.39 Å for GPM12. A more detailed description of the CGNMA was given in the literature.[28,52]

In the CGNMA, the structures are described only with the positions of Cα atoms. For each of the structures, we reconstruct the all-atomic structure with the PULCHRA software package.[53] When a fixed structure is not stable enough in terms of the FEF, more stable structures can be generated

by the CGNMA (see the Appendix). It is verified that the fixed structure we employ is quite stable because the FEF of any of the structures generated by the CGNMA is higher than that of the fixed structure.

Via routes 1 and 2, we can generate a large variety of compact NNSs: For CLN025, the radius of gyration R_g of these structures is averaged at 6.51 Å, which is comparable to R_g of the X-ray model, 6.33 Å; and the minimum value of R_g is 5.61 Å and even smaller than R_g of the NS model (Model 14 in the NMR models), 5.82 Å. It follows that the NNSs obtained are fairly compact.

D. Complete α-helix structure

The complete α-helix structure is generated for CLN025 and GPM12 using the TINKER program package[54] with the dihedral angles (φ, ψ, ω) set at (60°, 45°, 180°). This structure, which is referred to as "all α" hereafter, is contrastive to the NS of CLN025: It is interesting to compare the characteristics of "all α" and the β-hairpin structure.

E. Unfolded state

We generate a set of RCs by assigning a random number to each of the dihedral angles, φ and ψ, for the main chain using the TINKER program package[54] (ω is set at 180°). The random numbers are limited to the ranges, $-180° \le \varphi \le -30°$, $-180° \le \psi \le -150°$, and $-90° \le \psi \le 180°$, except for glycine and proline. For glycine, the ranges allowed are $-180° \le \varphi \le -30°$, $30° \le \varphi \le 180°$, and $-180° \le \psi \le 180°$. The dihedral angles for proline are set at $-65°$ and 180°, respectively. From among these RCs, only those with $\Lambda = 0$ are chosen. This is because a rather extended structure like RCs undergoing positive Λ should be significantly unstable. A total of ~650 different RCs thus obtained are assumed to form the unfolded state for each of CNL025 and GPM12. A physical quantity averaged over the RCs is regarded as the physical quantity of the unfolded state. For CLN025, the average value of R_g for the RCs is 8.74 Å which is much larger than that for the NNSs, 6.51 Å.

F. Slight modification of structures

When unrealistic overlaps of the constituent atoms occur in a structure, they are removed by the minimization of the energy function using the Amber12 program package[44] with the f99SB force field[43] and the GB model.[49,50]

IV. EFFECT OF POLYPEPTIDE CONFORMATIONAL ENTROPY

In later sections, we compare the compact structure having the lowest FEF (FEF is F defined by Eq. (4)) with the unfolded state to examine the foldability of CLN025 and GPM12. In such a comparison, the conformational-entropy effect is to be taken into account. The conformational entropy of the compact structure is considered to be essentially zero. That of the unfolded state can be estimated as follows. For the backbone, per residue there are two dihedral angles that can rotate and each angle has three stable values. Therefore, the number of possible combinations is $3^2 = 9$ and the contribution to the conformational entropy is $k_B \ln(9)$. Based on the computer simulation study by Doig and Sternberg,[55] we regard the contribution from the side chain to the conformational entropy as $1.7k_B$ per residue. Hence, for a polypeptide with N_r residues, $-TS_C$ ($T = T_0 = 298$ K) is added to the FEF for the unfolded state, where

$$S_C/k_B = -N_r\{\ln(9) + 1.7\}. \tag{7}$$

We remark that Eq. (7) was not employed in our earlier works[26,27] for discriminating the native fold from the misfolded decoys because only compact structures were considered.

V. CHARACTERIZATION OF NATIVE STRUCTURE MODELS FOR CLN025

A total of 21 NS models are available for CLN025: the X-ray model and 20 NMR models (Models 1−20). The EV, $-S_{VH}$, Λ, and F of the 21 NS models are compared in Table I. They vary largely from model to model (the maximum difference reaches $\sim 50k_BT_0$). In general, even when the NS models drawn in figures are not significantly different from one another in sight, their thermodynamic quantities of hydration as well as intramolecular electrostatic and LJ energies can be substantially different. Therefore, the selection of the best model is a very important step.[9,46] It has already been demonstrated that our FEF and its energetic and entropic

TABLE I. Excluded volume (EV), $-S_{VH}$, Λ, and F of X-ray and NMR models for CLN025. ($F-F_{NS}$) (F_{NS} represents value of F for Model 14) is also given.

Model	EV (Å^3)	$-S_{VH}/k_B$	$\Lambda/(k_BT_0)$	$F/(k_BT_0)$	$(F-F_{NS})/(k_BT_0)$
1	2795.17	449.93	7	456.93	10.65
2	2773.07	447.25	21	468.25	21.97
3	2800.45	451.15	14	465.15	18.87
4	2783.89	448.62	7	455.62	9.34
5	2808.96	452.38	7	459.38	13.10
6	2770.31	447.31	7	454.31	8.03
7	2798.21	450.22	7	457.22	10.94
8	2799.86	450.49	7	457.49	11.21
9	2787.55	448.81	7	455.81	9.53
10	2769.84	447.24	14	461.24	14.96
11	2782.76	447.38	0	447.38	1.10
12	2800.49	450.80	7	457.80	11.52
13	2836.11	453.95	14	467.95	21.67
14	2770.17	446.28	0	446.28	0.00
15	2807.40	451.35	7	458.35	12.07
16	2848.12	456.92	7	463.92	17.64
17	2800.74	451.48	7	458.48	12.20
18	2845.02	457.07	35	492.07	45.79
19	2912.28	467.42	28	495.42	49.15
20	2881.19	464.03	21	485.03	38.75
X-ray	2887.41	458.84	7	465.84	19.56

components are very useful in characterizing the NS models and determining the best model.[46]

F of Model 14 is the lowest: It is the most stable in terms of both $-S_{\mathrm{VH}}$ and Λ. The backbone and side chains of Model 14 are more efficiently packed than any other model. At the same time, an intramolecular hydrogen bond is always formed upon burial of a donor and an accepter in Model 14. Model 19 has the highest F. This structure suffers large Λ $(= 28k_{\mathrm{B}}T_0)$. In Model 19, four donors and accepters, ASP3.N, GLU5.N, THR8.N, and THR8.O, are buried without forming intramolecular hydrogen bonds, giving rise to large Λ. In Model 14, by contrast, these donors and acceptors are either exposed to water or forming the intramolecular hydrogen bonding with the result of $\Lambda = 0$. Model 19 is unstable in terms of $-S_{\mathrm{VH}}$ as well. We find the following: In Model 19, the side chain of TYR10 is not closely packed with the backbone and side chains of some other residues. As for the X-ray model, it is not very stable in terms of Λ because TYR8.N is buried without any partner for the intramolecular hydrogen bonding. The X-ray model is not stable in terms of $-S_{\mathrm{VH}}$, either: The side chains of TYR1 and TYR2 stick out, leading to lack of close packing.

From the results described above, we choose Model 14 as the NS of CLN025. It should be noted that the nonnative compact structures (NNSs) include those which are quite similar to the NS. An NS model is not always more stable than the NNSs in terms of the FEF (see Sec. VI A).

VI. RESULTS AND DISCUSSION

A. Structural stability of CLN025

We compare the NS (Model 14 in the NMR models), the other NS models (the X-ray Model, Models 1−13, and Models 15−20), NNSs obtained via routes 1 and 2, and "all α" with regard to the structural stability for CLN025. Figure 3 shows the plot of $(F - F_{\mathrm{NS}})$ (the subscript "NS" denotes the value of the NS) against the RMSD form the NS. It is clear that F of the NS is the lowest. Furthermore, there is a strong trend that F becomes progressively lower as the RMSD from the NS decreases. A MD simulation study[56] has shown the following: CLN025 exhibits a funnel-like free energy landscape; and it possesses two minima which can be reproduced not in implicit water but in explicit one. It is interesting to note that the remnant of the two minima is observed in Fig. 3.

In what follows, we analyze what factor is responsible for the highest stability of the NS. To this end, F is decomposed into entropic and energetic terms. We define X and Y as

$$X = \Lambda - \Lambda_{\mathrm{NS}} \tag{8}$$

and

$$Y = -S_{\mathrm{VH}} - (-S_{\mathrm{VH,NS}}), \tag{9}$$

respectively. X and Y, respectively, denote Λ and $-S_{\mathrm{VH}}$ of a structure *relative to* those of the NS $(F - F_{\mathrm{NS}} = X + Y)$. Closer, more efficient packing of the backbone and side chains leads to a smaller EV generated, less serious water crowding, and smaller $-S_{\mathrm{VH}}$. As the number of donors and acceptors buried without the intramolecular hydrogen bonding increases, Λ becomes larger. $Y < 0$, for example, implies that the backbone and side-chain packing of the structure is more efficient than that of the NS.

In Figure 4, X is plotted against Y for the NS, the other NS models, NNSs obtained via routes 1 and 2, and "all α." Since $\Lambda_{\mathrm{NS}} = 0$, there is no structure with $X < 0$. There are significantly many structures in the NNSs obtained via route 1 with $Y < 0$, but they all possess large, positive X. There is no structure giving $X + Y < 0$: The NS is optimized in terms of $X + Y$. This can also be appreciated in Table II comparing the EV, $-S_{\mathrm{VH}}$, Λ, and F of "all α," unfolded state, NS, and some

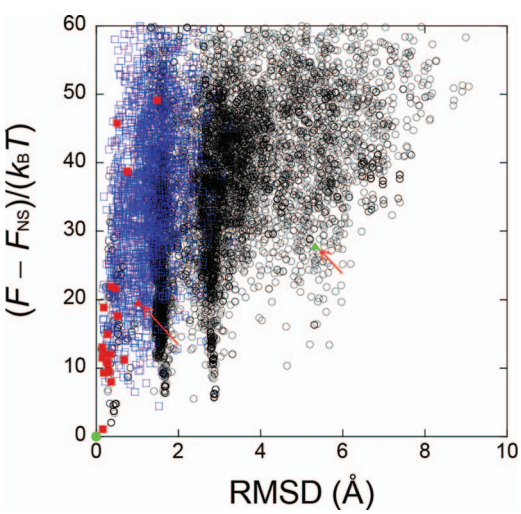

FIG. 3. $(F - F_{\mathrm{NS}})$ (the subscript "NS" denotes the value of the NS) plotted against RMSD form the NS for CLN025. Black circles: data points for the NNSs obtained via route 1. Green circle: data point for the NS (Model 14 in the NMR models). Red closed squares: data points for the NMR models other than Model 14. Red triangle: data point for the X-ray model (this is indicated by the long arrow). Blue squares: data points for the NNSs obtained via route 2. Green triangle: data point for "all α" (this is indicated by the short arrow).

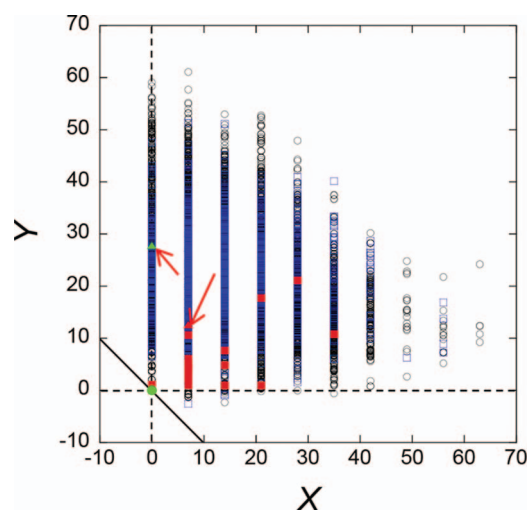

FIG. 4. Y plotted against X for CLN025. Black circles: data points for the NNSs obtained via route 1. Green circle: data point for the NS (Model 14 in the NMR models). Green closed squares: data points for the NMR models other than Model 14. Red triangle: data point for the X-ray model (this is indicated by the long arrow). Blue squares: data points for the NNSs obtained via route 2. Green triangle: data point for "all α" (this is indicated by the short arrow). The vertical broken line, horizontal broken line, and diagonal solid line represent $X = 0$, $Y = 0$, and $X + Y = 0$, respectively.

TABLE II. Excluded volume (EV), $-S_{VH}$, Λ, and F of "all α," unfolded state (random coils: RCs), NS (the native structure: Model 14 in the NMR models), and some representative structures chosen from the compact, nonnative structures (NNS1−NNS10) for CLN025. F is given by Eq. (4). For the unfolded state, however, $-TS_C$ ($T = T_0 = 298$ K) is added to F. ($F - F_{NS}$) (F_{NS} represents value of F for the NS) is also given.

Structure	EV (Å^3)	$-S_{VH}/k_B$	$\Lambda/(k_BT_0)$	$-S_C/k_B$	$F/(k_BT_0)$	$(F-F_{NS})/(k_BT_0)$
NNS1	2763.90	443.98	14	0	457.98	11.70
NNS2	2756.20	446.05	21	0	467.05	20.77
NNS3	2762.85	445.68	35	0	480.68	34.40
NNS4	2769.97	445.03	7	0	452.03	5.75
NNS5	2765.43	446.26	28	0	474.26	27.98
NNS6	2828.10	451.78	0	0	451.78	5.50
NNS7	2823.66	452.64	0	0	452.64	6.36
NNS8	2825.88	452.90	0	0	452.90	6.62
NNS9	2822.28	452.97	0	0	452.97	6.69
NNS10	2853.03	457.45	0	0	457.45	11.18
All α	2995.56	474.07	0	0	474.07	27.79
RCs	3205.34	492.55	0	− 38.97	453.58	7.30
NS	2770.17	446.28	0	0	446.28	0

representative structures chosen from the NNSs. The representative structures are those which have especially low values of $-S_{VH}$ (NNS1–NNS5) and those with $\Lambda = 0$ (NNS6–NNS10). In the table, $-TS_C$ ($T = T_0$) is added to F just for the unfolded state. Though the structures with $Y < 0$ can be constructed (e.g., NNS1–NNS5), such structures suffer large values of X. There can be many structures with $X = 0$ (e.g., NNS6–NNS10), but they unavoidably possess large values of Y. For "all α," which can form the maximum number of intramolecular hydrogen bonds between donors and accepters in the backbone, $X = 0$. However, its Y is positive and quite large: The backbone and side chains can be packed only much less efficiently than in the NS.

B. Importance of backbone and side-chain packing in structural stability of CLN025

It is physically insightful to separate the effect of side chains from that of the backbone for the entropic component of the FEF. To perform this separation, we replace all residues in each structure by Gly using the CHARMM (Ref. 57) and MMTSB (Ref. 58) programs. The replacement is carried out after the slight modification of the structure described above. The structure thus made has essentially no side chains (hereafter, these are referred to as "structures without side chains"). $-S_{VH}$ represents the water-entropy loss upon insertion of a protein with a prescribed structure. The information on the effect of side chains is contained in "$-S_{VH}$ of a structure with side chains" − "$-S_{VH}$ of the corresponding structure without side chains (i.e., with the backbone alone)": The latter is denoted by $-S_b$: $-S_{VH} = -S_b + (-S_{sc})$ where $-S_b$ and $-S_{sc}$ represent the contributions from the backbone and side chains to $-S_{VH}$, respectively. We then define Y_b and Y_{sc} as

$$Y_b = -S_b/k_B - (-S_{b,NS}/k_B) \tag{10a}$$

and

$$Y_{sc} = -S_{sc}/k_B - (-S_{sc,NS}/k_B), \tag{10b}$$

respectively.

Smaller $-S_b$ implies closer, more efficient packing of the backbone. Smaller $-S_{sc}$ implies more efficient packing of the backbone as well as side chains. For a large protein $-S_{sc}$ is governed by the side-chain packing,[15,16] but for a small polypeptide like CLN025 and GPM12 the packing of side chains with the backbone also plays essential roles (this is discussed in a later section). In the case of $Y_b < 0$, the backbone of the structure is more efficiently packed than that of the NS (Model 14). $Y_{sc} < 0$ is indicative that the backbone and side chains of the structure are more efficiently packed than those of the NS.

For the NS, the other NS models, NNSs obtained via routes 1 and 2, and "all α," we plot Y_b and Y_{sc} against the RMSD form the NS in Fig. 5. "All α" is more stable than the NS in terms of $-S_b/k_B$, which implies that it can achieve very close packing of the backbone (see Fig. 1(a)). There are many NNSs with $Y_b < 0$. However, the number of the NNSs with $Y_{sc} < 0$ is considerably smaller. "All α" is not stable in terms of $-S_{sc}/k_B$, and it is lacking close packing of the backbone and side chains. Choosing only the structures with $\Lambda = 0$, we replot Y_b and Y_{sc} against the RMSD form the NS in Fig. 6. While there are still significantly many structures with $Y_b < 0$, there is no structure with $Y_{sc} < 0$. The NS becomes quite stable in terms of $-S_{VH}$ through the most efficient, closest packing of the backbone and side chains.

C. Comparison between conformational-entropy loss and water-entropy gain upon folding for CLN025

The conformational-entropy loss upon folding from the unfolded state to the NS (Model 14) is calculated to be $-38.97k_B$ (Eq. (7) is employed). The water-entropy gain is $46.27k_B$ (46.27 = 492.55 − 446.28) (see Table II). Since the intramolecular hydrogen bonding is always formed upon burial of a donor and an acceptor ($\Lambda = 0$ in the NS), the system energy remains unchanged (we can consider that the system enthalpy also remains unchanged because the folding occurs with the system pressure and volume almost

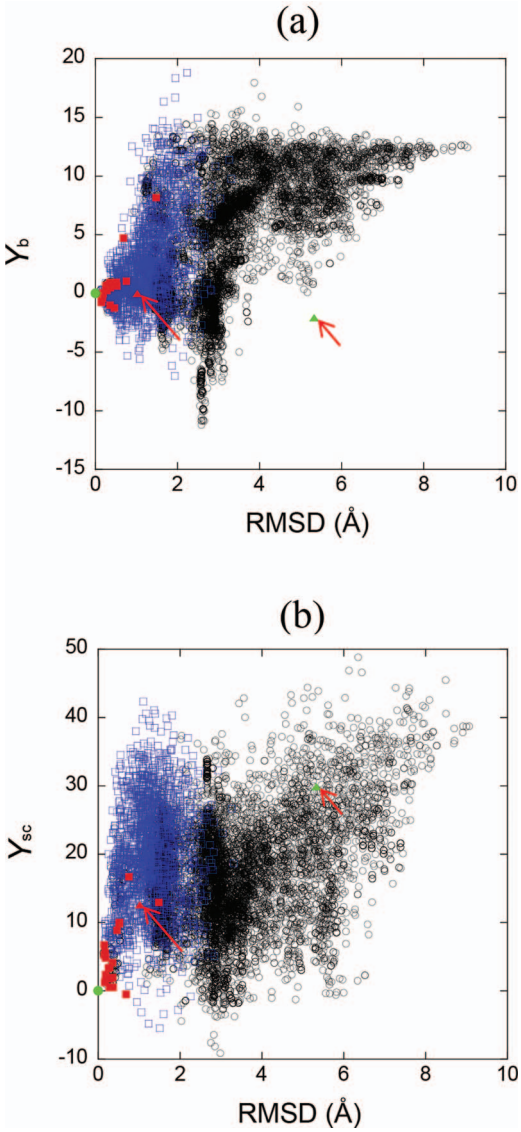

FIG. 5. (a) Y_b or (b) Y_{sc} plotted against RMSD from the NS (Model 14 in the NMR models) for CLN025. Black circles: data points for the NNSs obtained via route 1. Green circle: data point for the NS. Red closed squares: data points for the NMR models other than Model 14. Red triangle: data point for the X-ray model (this is indicated by the long arrow). Blue squares: data points for the NNSs obtained via route 2. Green triangle: data point for "all α" (this is indicated by the short arrow).

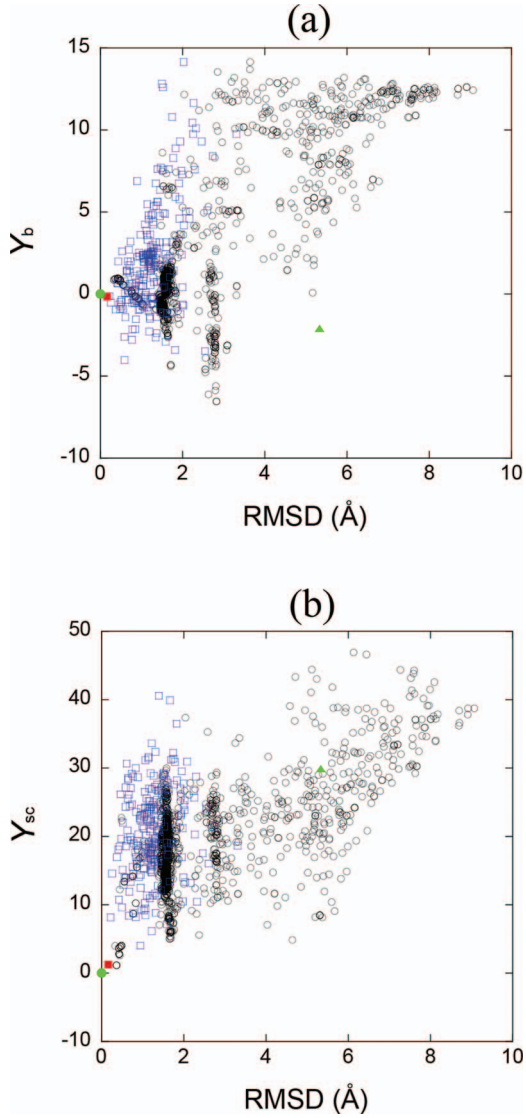

FIG. 6. (a) Y_b or (b) Y_{sc} plotted against RMSD from the NS (Model 14 in the NMR models) for CLN025. Only the structures with $\Lambda = 0$ are chosen. Green circle: data point for the NS. Red closed squares: data points for the NMR models other than Model 14. Blue squares: data points for the NNSs obtained via route 2. Green triangle: data point for "all α."

unchanged). Therefore, the system free energy changes by ~ -18 kJ/mol.

CLN025 includes proline in its amino-acid sequence. Because dihedral angle of proline is more restricted than the other amino acids, the conformational entropy calculated by Eq. (7) is overestimated for CLN025. When dihedral angles of proline, φ and ψ, are fixed to -65° and 180°, respectively, the conformational entropy loss reduces to $-36.78k_B$ and the change in the system free energy becomes ~ -24 kJ/mol.

D. Structural stability of GPM12

Figure 7 shows the plot of $(F - F_{NNS6})$ (the subscript "NNS6" denotes the value for NNS6, the most stable structure in terms of F) against the RMSD form the NNS6 for GPM12. We note that NNS6 possesses the β-hairpin structure. There are significantly many structures sharing almost the same value of F in the RMSD range 0–2 Å, which is in contrast to Fig. 3 drawn for CLN025.

Table III gives the EV, $-S_{VH}$, Λ, F, and $(F - F_{NNS6})$ of "all α," unfolded state, and some representative structures chosen from the NNSs. The representative structures are those which have especially low values of $-S_{VH}$ (NNS1–NNS5) and those with $\Lambda = 0$ (NNS6–NNS10). In the table, $-TS_C$ ($T = T_0$) is added to F just for the unfolded state. Relatively close packing of the backbone and side chains can be achieved in NNS1–NNS5. However, these structures undergo large Λ (i.e., some donors and accepters are unavoidably buried without forming intramolecular hydrogen bonds). The intramolecular hydrogen bonding is always formed upon burial of a donor and an acceptor in NNS6–NNS10, but sufficiently close

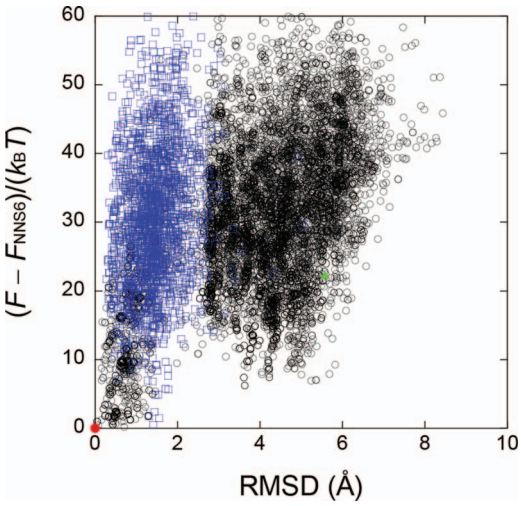

FIG. 7. $(F - F_{NNS6})$ (the subscript "NNS6" denotes the value of NNS6) plotted against RMSD form NNS6 for GPM12. Black circles: data points for the NNSs obtained via route 1. Red circle: data point for NNS6 (the most stable structure in terms of F). Blue squares: data points for the NNSs obtained via route 2. Green triangle: data point for "all α."

packing of the backbone and side chains cannot simultaneously be attained. Unlike CLN025, GPM12 is incapable of optimizing its structure in terms of both of the energetic and entropic components.

E. Comparison between conformational-entropy loss and water-entropy gain upon folding for GPM12

If the folding from the unfolded state to NNS6 took place, the conformational-entropy loss of $-38.97k_B$ would be caused. On the other hand, the water-entropy gain would be $36.92k_B$ ($36.92 = 412.48 - 375.56$) that is much smaller than in the case of CLN025 (see Table III). Since $\Lambda = 0$ in NNS6, the system free energy would change by $\sim$+5 kJ/mol. Therefore, GPM12 is not likely to fold, and the unfolded state is stabilized.

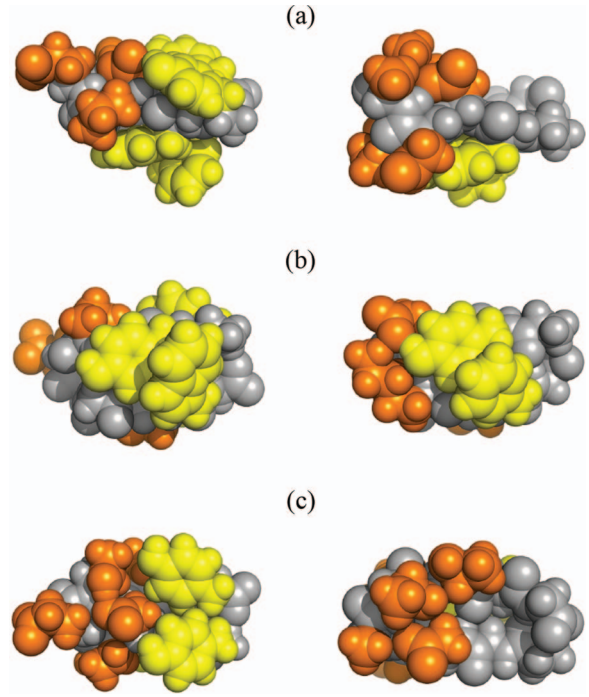

FIG. 8. Space-filling models of the NS (Model 14) for CLN025 (left) and NNS6 for GPM12 (right). (a) Side, (b) bottom, and (c) top view. The backbone, aromatic side chains, and the other side chains are colored in gray, yellow, and orange, respectively. This figure was drawn by PyMol 1.3.[59]

F. Origin of large water-entropy gain upon CLN025 folding

In the β-hairpin structure, the side chains in the upper and lower sides of the sheet can closely be packed together with the backbone if their geometric characteristics are amenable to such close packing. Figure 8 compares the β-hairpin structure of the NS of CLN025 and that of NNS6 of GPM12 (this figure was drawn by PyMol 1.3 (Ref. 59)). Close packing

TABLE III. Excluded volume (EV), $-S_{VH}$, Λ, and F of "All α," unfolded state (random coils: RCs), and some representative structures chosen from the compact, nonnative structures (NNS1−NNS10) for GPM12. NNS6 is the most stable structure in terms of F. F is given by Eq. (4). For the unfolded state, however, $-TS_C$ ($T = T_0 = 298$ K) is added to F. $(F-F_{NNS6})$ (F_{NNS6} represents value of F for NNS6) is also given.

Structure	EV (Å^3)	$-S_{VH}/k_B$	$\Lambda/(k_B T_0)$	$-S_C/k_B$	$F/(k_B T_0)$	$(F-F_{NNS6})/(k_B T_0)$
NNS1	2341.05	373.53	7	0	380.53	4.97
NNS2	2338.34	373.95	7	0	380.95	5.39
NNS3	2355.81	374.45	14	0	388.45	12.89
NNS4	2347.57	376.98	28	0	404.98	29.42
NNS5	2335.99	375.26	28	0	403.26	27.7
NNS6	2362.92	375.56	0	0	375.56	0
NNS7	2373.19	376.56	0	0	376.56	1
NNS8	2415.80	381.77	0	0	381.77	6.21
NNS9	2403.75	382.48	0	0	382.48	6.92
NNS10	2443.05	383.39	0	0	383.39	7.83
All α	2542.12	397.89	0	0	397.89	22.33
RCs	2699.35	412.48	0	− 38.97	373.50	− 2.06

in which aromatic side chains with large sizes participate plays essential roles in increasing the water entropy. CLN025 has four aromatic side chains (TYR1, TYR2, TRP9, and TYR10). The large water-entropy gain upon CLN025 folding is brought primarily by a large decrease in the EV arising from very efficient, close packing of the backbone and side chains including those with aromatic groups. Geometrical features (sizes, overall shapes, and details of the polyatomic structure) rather than chemical properties of side chains are crucially important. On the other hand, GPM12 has only two aromatic side chains (TYR2 and PHE9). Though the packing of the backbone and side chains is considerably close, the water-entropy gain due to the packing is not as large as that in the case of CLN025. This is because, as observed in Fig. 8(c), the side chains in the upper side of the sheet are not adequately packed with the backbone due to the absence of aromatic side chains, which is in contrast with the case of CLN025. The water-entropy gain upon GPM12 folding would be smaller than that upon CLN025 folding by $\sim 10k_B$.

G. On a peptide which is foldable into α-helix structure

As mentioned in the last section, in the β-hairpin structure it is possible to closely pack the backbone and side chains in the upper and lower sides of the sheet. In the α-helix structure, on the other hand, the side chains are not sufficiently close to one another and the close packing on the same level is inherently unachievable (see (a) and (b) in Fig. 9). It is unlikely that a small α-helix structure is stabilized by itself though it merits $\Lambda = 0$. A peptide which folds into the α-helix structure is the trp-cage[60] (PDB code: 2JOF) comprising 20 residues. This peptide possesses a long tail-region as well as the α-helix structure as illustrated in Fig. 9(c). The side chains of the residues constructing the α-helix structure are closely packed with those forming the long tail-region. The trp-cage should be stabilized by a sufficiently large gain of the water entropy arising from such close packing (TRP6 is at the center of the packing).

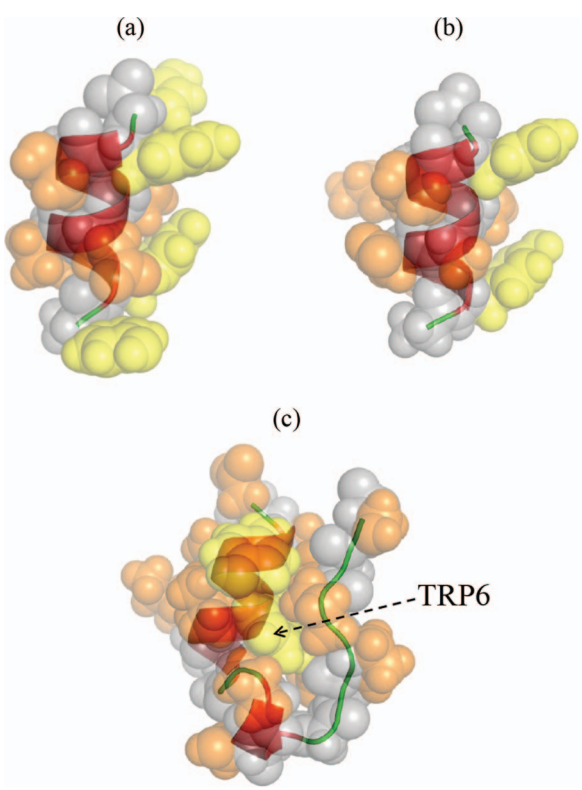

FIG. 9. Space-filling and ribbon representation of "all α" for CLN025 (a), "all α" for GPM12 (b), and the trp-cage (PDB code: 2JOF) (c). The backbone, aromatic side chains, and the other side chains are colored in gray, yellow, and orange, respectively. This figure was drawn by PyMol 1.3.[59]

VII. CONCLUSION

We have analyzed the structural stability of CLN025 and GPM12 using our free-energy function (FEF);[26,27] $F = \Lambda - TS_{VH}$ ($T = T_0 = 298$ K) and its energetic and entropic components, Λ and S_{VH}, respectively. S_{VH} is the hydration entropy calculated using a hybrid of the angle-dependent integral equation theory[29–34] and the morphometric approach.[38–40] A water molecule is modeled as a hard sphere in which a point dipole and a point quadrupole of tetrahedral symmetry are embedded.[30,31] As the backbone and side chains (especially the latter) are more efficiently packed, the positive quantity $-S_{VH}$ becomes smaller. This is ascribed primarily to a smaller EV, a larger total volume available to the translational displacement of water molecules, a larger number of accessible translational configurations of water, less water crowding, and higher water entropy.[11–17] Λ is calculated in a simple manner which still accounts for physically the most important factors: the break of protein-water hydrogen bonds and formation of protein intramolecular hydrogen bonds. When compared with the fully extended structure, which has the maximum number of hydrogen bonds with water molecules and no intramolecular hydrogen bonds, in a more compact structure some donors and acceptors are buried in the interior after the break of hydrogen bonds with water molecules. As the number of donors and acceptors buried without the intramolecular hydrogen bonding increases, Λ becomes larger (Λ is not a negative quantity). When only the compact structures are compared, there is no need to account for the effect of the polypeptide conformational entropy. When compact structures are compared with random coils, we add the conformational-entropy component estimated in a simple but physically reasonable manner to F.

As pointed out in our earlier work,[11] there is a general trend that the water entropy gain originating from close packing of the backbone and side chains (the overall, close side-chain packing is especially important for a large protein) increases more than in proportion to the number of residues. For a large protein, the water-entropy gain can be powerful enough to surpass the conformational-entropy loss plus the energy increase due to an insufficient number of intramolecular hydrogen bonds formed to compensate for the break of hydrogen bonds with water molecules upon folding. This is not the case for a short polypeptide. Close packing of the backbone and side chains is absolutely required, but the resulting

water-entropy gain is usually not very large. Hence, it is required that the break of hydrogen bonds with water molecules be completely compensated with the intramolecular hydrogen bonding (i.e., Λ be zero). It is rather difficult to achieve both of a sufficiently large water-entropy gain and $\Lambda = 0$ by a short amino-acid sequence. CLN025 is a good exception.

CLN025 is characterized by the amino-acid sequence which enables it not only to closely pack the backbone and side chains including those with aromatic groups having large sizes but also to always form an intramolecular hydrogen bond upon burial of a donor and an acceptor when the backbone forms the β-hairpin structure. There is essentially no enthalpy increase upon folding, and the water-entropy gain is large enough to surpass the conformational-entropy loss. The α-helix structure is distinguished from the β-hairpin structure in the sense that with the former the intramolecular hydrogen bonding is assured but sufficiently close packing of the backbone and side chains cannot be accomplished. By contrast, it is not possible for GPM12 to achieve both of a sufficiently large water-entropy gain and $\Lambda = 0$, even with the β-hairpin structure. There are significantly many compact structures which are equally stable in terms of F, and due to the conformational-entropy effect, the unfolded state is favorably stabilized. The differences between CLN025 and GPM12 are illustrated in Fig. 10. Here, it is worthwhile to add an important remark. According to the usual concept of the Ar–Ar attractive interaction, the aromatic groups in the β-hairpin structure are stabilized by stacking in pairs. One might think that a short polypeptide with only aromatic side chains (the number is even) is capable of achieving close packing. However, this is not always true. The packing can be closer when the aromatic side chains are packed together with the other side chains with different geometries and portions of the backbone.

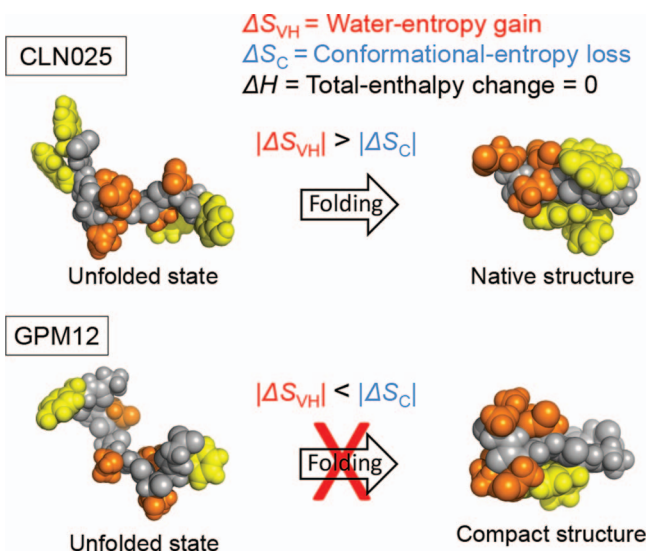

FIG. 10. Differences between CLN025 and GPM12. CLN025, for which the water-entropy gain predominates over the conformational-entropy loss, is able to fold. For GPM12, the water-entropy gain which would occur upon folding cannot be larger than the conformational-entropy loss accompanied: The unfolded state is favored. This figure was drawn by PyMol 1.3.[59]

Specifying the factor highly stabilizing the folded structure of CLN025 despite its very small size (the number of residues is only 10) has been made possible by decomposing our FEF into physically insightful constituents. Such decomposition is not possible in molecular dynamics simulations. If the combination of the effect of polypeptide conformational entropy and our FEF is applicable to a polypeptide irrespective of its length, the prediction of the foldability of any polypeptide given will become feasible. Work in this direction is in progress.

ACKNOWLEDGMENTS

The computer program for the morphometric approach was developed with R. Roth and Y. Harano.[39] The computer program for the CGNMA was written by S. Du.[52] This work was supported by JSPS (Japan Society for the Promotion of Science) Grant-in-Aid for Scientific Research (B) (No. 25291035: M. Kinoshita) and by Grant-in-Aid for JSPS fellows (S. Yasuda).

APPENDIX: USEFULNESS OF COARSE-GRAINED NORMAL MODE ANALYSIS (CGNMA)

In the present study, the structures near a fixed structure are generated using the CGNMA.[28] The fixed structure is the NS (Model 14) for CLN025 or NNS6 for GPM12. We find that none of the structures generated possesses the FEF which is lower than that of the fixed structure.

We perform two more analyses for CLN025, analyses A and B, by choosing structures A and B as the fixed structures, respectively. The two structures are taken from the compact nonnative structures. Table IV compares the values of the EV, $-S_{\mathrm{VH}}$, Λ, and F of the NS (Model 14 in the NMR models), structure A, and structure B. $-S_{\mathrm{VH}}$ of structure A is fairly close to that of the NS. However, structure A suffers a considerably larger value of Λ. Structure B and the NS share the same value of Λ (= 0), but $-S_{\mathrm{VH}}$ of the former is substantially larger. For the structures generated by the CGNMA in analysis A, the FEF is plotted against the RMSD form structure A in Fig. 11(a). There are a lot of structures whose FEF is lower than that of structure A. It is observed in a similar plot for analysis B shown in Fig. 11(b) that there are also a lot of structures whose FEF is lower than that of structure B. These examinations are suggestive that the CGNMA is capable of generating more stable structures than the fixed structure if they actually exist.

TABLE IV. Excluded volume (EV), $-S_{\mathrm{VH}}$, Λ, and F of the native structure (NS) of CLN025, structure A, and structure B.

Structure	EV (Å^3)	$-S_{\mathrm{VH}}/k_{\mathrm{B}}$	$\Lambda/(k_{\mathrm{B}}T_0)$	$F/(k_{\mathrm{B}}T_0)$
NS	2770.17	446.28	0	446.28
Structure A	2772.34	448.18	35	483.18
Structure B	3125.98	483.16	0	483.16

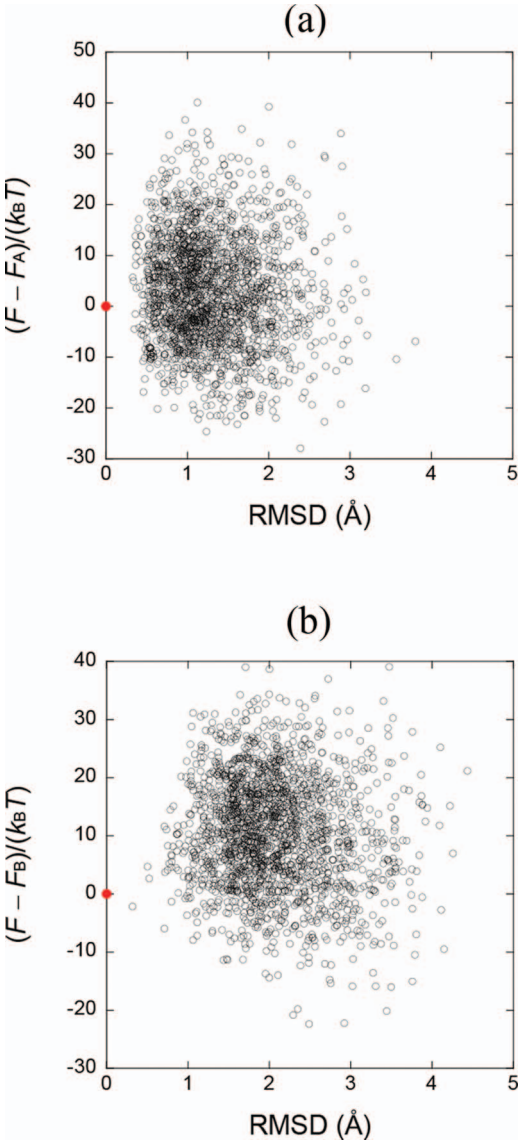

FIG. 11. (a) F plotted against RMSD form Structure A in analysis A. (b) F plotted against RMSD form Structure B in analysis B. Analyses A and B are performed for CLN025.

Taken together, it is reasonable to consider that the NS (Model 14) for CLN025 or NNS6 for GPM12 is truly the most stable structure.

[1] C. M. Dobson, Nature **426**, 884 (2003).
[2] S. Honda, K. Yamasaki, Y. Sawada, and H. Morii, Structure **12**, 1507 (2004).
[3] S. Honda, T. Akiba, Y. S. Kato, Y. Sawada, M. Sekijima, M. Ishimura, A. Ooishi, H. Watanabe, T. Odahara, and K. Harata, J. Am. Chem. Soc. **130**, 15327 (2008).
[4] M. P. D. Hatfield, R. F. Murphy, and S. Lovas, J. Phys. Chem. B **114**, 3028 (2010).
[5] M. P. D. Hatfield, R. F. Murphy, and S. Lovas, J. Phys. Chem. B **115**, 4971 (2011).
[6] G.-J. Zhao and C.-L. Cheng, Amino Acids **43**, 557 (2012).
[7] C. M. Davis, S. Xiao, D. P. Raleigh, and R. B. Dyer, J. Am. Chem. Soc. **134**, 14476 (2012).
[8] T. Imai, Y. Harano, M. Kinoshita, A. Kovalenko, and F. Hirata, J. Chem. Phys. **126**, 225102 (2007).
[9] T. Hayashi, H. Oshima, T. Mashima, T. Nagata, M. Katahira, and M. Kinoshita, Nucleic Acids Res. **42**, 6861 (2014).
[10] K. A. Dill, Biochemistry **29**, 7133 (1990).
[11] Y. Harano and M. Kinoshita, Biophys. J. **89**, 2701 (2005).
[12] T. Yoshidome, M. Kinoshita, S. Hirota, N. Baden, and M. Terazima, J. Chem. Phys. **128**, 225104 (2008).
[13] M. Kinoshita, Front. Biosci. **14**, 3419 (2009).
[14] M. Kinoshita, Int. J. Mol. Sci. **10**, 1064 (2009).
[15] S. Yasuda, T. Yoshidome, H. Oshima, R. Kodama, Y. Harano, and M. Kinoshita, J. Chem. Phys. **132**, 065105 (2010).
[16] S. Yasuda, H. Oshima, and M. Kinoshita, J. Chem. Phys. **137**, 135103 (2012).
[17] M. Kinoshita, Biophys. Rev. **5**, 283 (2013).
[18] T. Yoshidome and M. Kinoshita, Phys. Rev. E **79**, 030905(R) (2009).
[19] H. Oshima, T. Yoshidome, K. Amano, and M. Kinoshita, J. Chem. Phys. **131**, 205102 (2009).
[20] T. Yoshidome and M. Kinoshita, Phys. Chem. Chem. Phys. **14**, 14554 (2012).
[21] Y. Harano, T. Yoshidome, and M. Kinoshita, J. Chem. Phys. **129**, 145103 (2008).
[22] T. Yoshidome, Y. Harano, and M. Kinoshita, Phys. Rev. E **79**, 011912 (2009).
[23] K. Amano, T. Yoshidome, Y. Harano, K. Oda, and M. Kinoshita, Chem. Phys. Lett. **474**, 190 (2009).
[24] K. Oda, R. Kodama, T. Yoshidome, M. Yamanaka, Y. Sambongi, and M. Kinoshita, J. Chem. Phys. **134**, 025101 (2011).
[25] N. Baden, S. Hirota, T. Takabe, N. Funasaki, and M. Terazima, J. Chem. Phys. **127**, 175103 (2007).
[26] T. Yoshidome, K. Oda, Y. Harano, R. Roth, Y. Sugita, M. Ikeguchi, and M. Kinoshita, Proteins **77**, 950 (2009).
[27] S. Yasuda, T. Yoshidome, Y. Harano, R. Roth, H. Oshima, K. Oda, Y. Sugita, M. Ikeguchi, and M. Kinoshita, Proteins **79**, 2161 (2011).
[28] M. M. Tirion, Phys. Rev. Lett. **77**, 1905 (1996).
[29] N. M. Cann and G. N. Patey, J. Chem. Phys. **106**, 8165 (1997).
[30] P. G. Kusalik and G. N. Patey, J. Chem. Phys. **88**, 7715 (1988).
[31] P. G. Kusalik and G. N. Patey, Mol. Phys. **65**, 1105 (1988).
[32] M. Kinoshita and M. Harada, Mol. Phys. **81**, 1473 (1994).
[33] M. Kinoshita and D. R. Bérard, J. Comput. Phys. **124**, 230 (1996).
[34] M. Kinoshita, J. Chem. Phys. **128**, 024507 (2008).
[35] M. Ikeguchi, S. Shimizu, S. Nakamura, and K. Shimizu, J. Phys. Chem. B **102**, 5891 (1998).
[36] N. Matubayasi and M. Nakahara, J. Chem. Phys. **112**, 8089 (2000).
[37] T. Imai, Y. Harano, M. Kinoshita, A. Kovalenko, and F. Hirata, J. Chem. Phys. **125**, 024911 (2006).
[38] P. M. König, R. Roth, and K. R. Mecke, Phys. Rev. Lett. **93**, 160601 (2004).
[39] R. Roth, Y. Harano, and M. Kinoshita, Phys. Rev. Lett. **97**, 078101 (2006).
[40] R. Kodama, R. Roth, Y. Harano, and M. Kinoshita, J. Chem. Phys. **135**, 045103 (2011).
[41] M. L. Connolly, J. Appl. Crystallogr. **16**, 548 (1983).
[42] M. L. Connolly, J. Am. Chem. Soc. **107**, 1118 (1985).
[43] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling, Proteins **65**, 712 (2006).
[44] D. A. Case, T. A. Darden, T. E. Cheatham III, C. L. Simmerling, J. Wang, R. E. Duke, R. Luo, R. C. Walker, W. Zhang, K. M. Merz, B. Roberts, S. Hayik, A. Roitberg, G. Seabra, J. Swails, A. W. Goetz, I. Kolossváry, K. F. Wong, F. Paesani, J. Vanicek, R. M. Wolf, J. Liu, X. Wu, S. R. Brozell, T. Steinbrecher, H. Gohlke, Q. Cai, X. Ye, J. Wang, M.-J. Hsieh, G. Cui, D. R. Roe, D. H. Mathews, M. G. Seetin, R. Salomon-Ferrer, C. Sagui, V. Babin, T. Luchko, S. Gusarov, A. Kovalenko, P. A. Kollman, AMBER 12, University of California, San Francisco, 2012.
[45] M. Kinoshita, J. Chem. Phys. **116**, 3493 (2002).
[46] H. Mishima, S. Yasuda, T. Yoshidome, H. Oshima, Y. Harano, M. Ikeguchi, and M. Kinoshita, J. Phys. Chem. B **116**, 7776 (2012).
[47] S. F. Sneddon, D. J. Tobias, and C. L. Brooks III, J. Mol. Biol. **209**, 817 (1989).
[48] I. K. McDonald and J. M. Thornton, J. Mol. Biol. **238**, 777 (1994).
[49] G. D. Hawkins, C. J. Cramer, and D. G. Truhlar, Chem. Phys. Lett. **246**, 122 (1995).
[50] G. D. Hawkins, C. J. Cramer, and D. G. Truhlar, J. Phys. Chem. **100**, 19824 (1996).
[51] J. P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen, J. Comput. Phys. **23**, 327 (1977).
[52] S. Du, Y. Harano, M. Kinoshita, and M. Sakurai, Biophysics **8**, 127 (2012).
[53] P. Rotkiewicz and J. Skolnick, J. Comput. Chem. **29**, 1460 (2008).

[54]J. W. Ponder and F. M. Richards, J. Comput. Chem. **8**, 1016 (1987).
[55]A. J. Doig and M. J. E. Sternberg, Protein Sci. **4**, 2247 (1995).
[56]A. Rodriguez, P. Mokoema, F. Corcho, K. Bisetty, and J. J. Perez, J. Phys. Chem. B **115**, 1440 (2011).
[57]B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, J. Comput. Chem. **4**, 187 (1983).
[58]M. Feig, J. Karanicolas, and C. L. Brooks III, J. Mol. Graphics Modell. **22**, 377 (2004).
[59]The PyMol Molecular Graphics System, Version 1.3, Schrödinger, LLC.
[60]B. Barua, J. C. Lin, V. D. Williams, P. Kummler, J. W. Neidigh, and N. H. Andersen, Protein Eng. Des. Sel. **21**, 171 (2008).