Ab initio simulation of a 57-residue protein in explicit solvent reproduces the native conformation in the lowest free-energy cluster

Jinzen Ikebe,¹ Daron M. Standley,² Haruki Nakamura,³ and Junichi Higo^{3*}

¹Graduate School of Frontier Biosciences, Osaka University, Open Laboratories for Advanced Bioscience and Biotechnology, Suita, Osaka 565-0874, Japan

²Systems Immunology Laboratory, Immunology Frontier Research Center (IFReC), Osaka University, Suita, Osaka 565-0871, Japan

³Laboratory of Protein Informatics, Institute for Protein Research, Osaka University, Suita, Osaka 565-0871, Japan

Received 6 September 2010; Revised 12 October 2010; Accepted 24 October 2010

DOI: 10.1002/pro.553

Published online 16 November 2010 proteinscience.org

Abstract: An enhanced conformational sampling method, multicanonical molecular dynamics (McMD), was applied to the ab initio folding of the 57-residue first repeat of human glutamyl-prolyl-tRNA synthetase (EPRS-R1) in explicit solvent. The simulation started from a fully extended structure of EPRS-R1 and did not utilize prior structural knowledge. A canonical ensemble, which is a conformational ensemble thermodynamically probable at an arbitrary temperature, was constructed by reweighting the sampled structures. Conformational clusters were obtained from the canonical ensemble at 300 K, and the largest cluster (i.e., the lowest free-energy cluster), which contained 34% of the structures in the ensemble, was characterized by the highest similarity to the NMR structure relative to all alternative clusters. This lowest free-energy cluster included native-like structures composed of two anti-parallel α -helices. The canonical ensemble at 300 K also showed that a short Gly-containing segment, which adopts an α -helix in the native structure, has a tendency to be structurally disordered. Atomic-level analyses demonstrated clearly that inter-residue hydrophobic interactions drive the helix formation of the Gly-containing segment, and that increasing the hydrophobic contacts accompanies exclusion of water molecules from the vicinity of this segment. This study has shown, for the first time, that the free-energy landscape of a structurally well-ordered protein of about 60 residues is obtainable with an all atom model in explicit water without prior structural knowledge.

Keywords: folding simulation; free-energy landscape; multicanonical molecular dynamics simulation; enhanced conformational sampling; EPRS-R1; explicit solvent; all-atom model

Introduction

Since Anfinsen¹ proposed that the native structure of a protein is determined by only its amino acid sequence under physiological conditions, the reproduction of native protein folds by first-principles simulations has remained one of the most difficult problems in computational biology. In spite of significant improvements in

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: Japan Society for the Promotion of Science (JSPS); Grant number: Grant-in-aid for Scientific Research (B) (20370061); Grant sponsor: Ministry of Education, Culture, Sports, Science and Technology (MEXT) Japan; Grant number: Scientific Research on Priority Areas "Structures of Biological Macromolecular Assemblies" (513-20051013); Grant sponsor: MEXT; Grant sponsor: New Energy and Industrial Technology Development Organization (NEDO) of Japan; J.H. was supported by Grant-in-Aid for Scientific Research on Innovative Areas (21113006) from MEXT.

*Correspondence to: Junichi Higo, Open Laboratories for Advanced Bioscience and Biotechnology, 6-2-3, Furuedai, Suita, Osaka 565-0874, Japan. E-mail: higo@protein.osaka-u.ac.jp

Figure 1. a: Native structure of EPRS-R1 (NMR model 1; PDB ID: 1FYJ), where the H1 region (residues 2–21) is shown in gray, and the H2 region (residues 26–47) in black. Segment of residues 2–47 is called a “core-region,” which involve H1 and H2. b: The initial elongated structure for the simulation. c: An example of conformation with water sphere (sphere 1) at 700 K. d: Some conformations sampled at 700 K.

computing hardware, we still cannot trace the folding process by conventional (canonical) molecular dynamics (MD) simulations, because the practical simulation time is considerably shorter than that required to fold a typical protein.² This difficulty results from the fact that, at room temperature, a polypeptide chain undergoing simulated folding is drawn into local energy attractors and remains trapped in these minima for a significant amount of time before escaping. To reproduce the native structure by computer simulations, the sampling problem must first be overcome.

Generalized-ensemble methods,^{3–5} such as multicanonical MD (McMD)^{3,5} or replica exchange MD (REMD),^{6,7} have been developed to resolve the sampling problem. To date, McMD simulations^{8–10} implemented in Cartesian coordinates¹¹ have been applied to polypeptides of up to 40 amino acid residues, using all-atom protein models in explicit solvent, starting from a fully unfolded structure. However, these polypeptides are still shorter than typical proteins, and it remains a challenge to reproduce native structures for proteins up to 50 amino acid residues in explicit water by such methods.

Here, we carried out an McMD simulation of a 57-residue protein, the first repeat from human glutamyl-prolyl-tRNA synthetase (EPRS-R1), in explicit water. The native structure (NMR structure; PDB ID: 1FYJ) has a helix-hairpin fold, where H1 and H2 regions (residues 2–21 and 26–47, respectively) adopt α -helices and make contact in an anti-parallel

orientation [Fig. 1(a)]. The simulation was started from a fully extended structure. The most thermodynamically stable cluster in the conformational ensemble at 300 K contained the native-like structures. Further analysis showed that the native fold is stabilized by inter-helix hydrophobic contacts that exclude water molecules from the vicinity.

Results and Discussion

We present a table in Supporting Information, Terms and quantities used for analyses, which is a quick reference table of terms and quantities used in this section.

Variety of sampled structures

Our McMD simulation is free from native structure-guided interactions and started from a fully extended conformation of EPRS-R1 [Fig. 1(b)]. The production McMD simulation covered a temperature range of 280 to 700 K, and example snapshots are shown in Figure 1(c,d). In Figure 1(c), water molecules are shown to illustrate the size of the solvent boundary. EPRS-R1 adopted various conformations during the simulation. This structural variety also demonstrates that the solvent boundary was of sufficient size to explore many different conformations. More significantly, the production simulation reproduced native-fold structures, as described below.

Sampled structures similar to native

We generated a conformational ensemble by reweighting the entire set of sampled structures at

Figure 2. a: 3D conformational space at 300 K constructed by MDS.^{12,13} A sphere represents a cluster. Meaning of the size and tone of spheres is given in the main text. Arrows v_1 , v_2 and v_3 are eigenvectors with the largest, second largest, and the third largest eigenvalues constructing the 3D conformational space (See Supporting Information, Multidimensional scaling). Numbers 1–3 indicate the largest three clusters. b: D_{NTV} value for each cluster.

300 K. We refer to this ensemble, consisting of 2,501 structures, as the “300 K dataset.” Next, a pseudo distance [Eq. (5)], which quantifies a structural dissimilarity of the core-region (residues 2–47) between two structures, was computed among the structures in the 300 K dataset. Then, we performed cluster analysis based on the pseudo distances using the average linkage method and obtained 20 clusters. Note that it is a coincidence that the number of obtained clusters and the number of NMR models are both 20. Hereafter, the cluster ID decreases with the cluster size.

The largest, second largest, and third largest clusters had proportions of 34%, 17%, and 13% of the 300 K dataset, respectively. Thermodynamically, these are the lowest, second lowest, and third lowest free-energy clusters. To visualize the distribution of the 20 clusters, we created a 3D conformational space by multidimensional scaling, MDS,^{12,13} as described in detail in Supporting Information, Multidimensional scaling. Figure 2(a) demonstrates the cluster distribution. The larger the sphere (cluster), the greater the number of structures in the cluster; the darker the tone of the sphere, the smaller the average pseudo distance (D_{NTV}) between the structures in the cluster and the 20 NMR models. Figure 2(b) represents D_{NTV} for each cluster. We concluded that the largest three clusters have higher similarity with the native structure than the other clusters. Below we analyze these clusters.

First, we analyzed reproduction of NOE signals in each cluster. In an NMR experiment, an NOE signal assigned to a pair of hydrogen atoms i and j is converted to an atomic distance (NOE distances, R_{NOE}^{exp}), and NMR models are reconstructed from the set of NOE distances. EPRS-R1 yielded 173 NOE distances from the core-region.¹⁴ The NOE distance is a statistical quantity defined as $R_{NOE}^{exp}(i,j) = \langle 1/r(i,j)^6 \rangle^{-1/6}$, where $r(i,j)$ is the distance between the two hydrogen

atoms of an instantaneous structure, and $\langle \dots \rangle$ denotes an ensemble average. We have 20 conformational clusters obtained above. The NOE distance for each cluster is computed as: $R_{NOE}^{calc}(i,j) = \langle 1/r(i,j)^6 \rangle^{-1/6}$, where $\langle \dots \rangle$ is the average over the structures in a cluster. We calculate the reproduction ratio for the 173 NOE distances with a tolerance ΔR_{NOE} : If $R_{NOE}^{calc} \leq R_{NOE}^{exp} + \Delta R_{NOE}$ is satisfied for an NOE distance, the distance is reproduced. Figure 3(a) depicts the reproduction ratio of the NOE distances in each cluster with two tolerances: $\Delta R_{NOE} = 0$ Å (green bars) and $\Delta R_{NOE} = 1$ (red bars). The later tolerance is suitable for recognizing a structure whose overall fold is similar to the native.¹⁵ The lowest free-energy cluster was characterized by the largest ratio, and the unsatisfied NOE pairs assembled in the turn region connecting H1 and H2 and in the N-terminus of H1 [Fig. 3(b)]. Thus, the lowest free-energy cluster reproduced the helix-hairpin fold overall, as shown below. We also noticed that the second and third lowest free-energy clusters had the second and third highest reproduction ratios, respectively.

The root mean square deviation ($RMSD_{core}$) of the backbone (N, C α , and C atoms) for the core-region was calculated between each cluster and the first NMR model (PDB ID: 1FYJ). The Q value was also calculated for each cluster. The Q value is the reproduction rate of the inter-residual contacts formed in the native structure (See Supporting Information, Q value). The $RMSD_{core}$ and Q value for each cluster are shown in Figure 4(a) and 4(b), respectively. As the figure shows, the largest three clusters had better $RMSD_{core}$ and Q values than did the other clusters.

Several sampled structures at 300 K are shown in Figure 5. The lowest free-energy cluster involves native-like structures exhibiting the helix-hairpin fold (see the upper panel for the lowest free-energy cluster). This cluster also included structures (see the lower panel) where the H1 region is disordered but

Figure 3. a: Reproduction ratio of NOE distances in each cluster. Green and red bars are ratios with tolerances $\Delta R_{\text{NOE}} = 0 \text{ \AA}$ and $1 \text{ \AA}$, respectively. b: Locations of satisfied/unsatisfied NOE pairs in the lowest free-energy cluster. Depicted structure is NMR model 1 (PDB ID: 1FYJ). Four color-cylinders indicate the locations of the pairs with four tolerances (blue: $\Delta R_{\text{NOE}} = 0 \text{ \AA}$, green: $\Delta R_{\text{NOE}} = 1 \text{ \AA}$, orange: $\Delta R_{\text{NOE}} = 2 \text{ \AA}$, and red: $\Delta R_{\text{NOE}} > 2 \text{ \AA}$).

the H2 helix was well-formed. Interestingly, the H1 helix in the second and third lowest free-energy clusters were all disordered. Thus, it is likely that the amino-acid sequence of H1 contains a factor to prevent helix formation. This point is discussed below. Minor clusters provided largely disordered structures, some of which involved a short β sheet (Fig. 5).

Instability of the H1 helix

We next analyzed the structural instability of H1, observed above. We computed the helix frequency at each residue site in a conformational ensemble by the DSSP program.¹⁶ This program checks the hydrogen bond patterns from the atomic coordinates of a given protein structure and assigns a secondary structure state to each residue. We refer to a residue, to which α -helix is assigned, as a “helical residue.” The structure with the most helical residues, which was part of the lowest free-energy cluster, contained two helices, in

the H1 and H2 regions (see the dot-dash-line in Fig. 6). Recall that H1 and H2 regions are residues 2–21 and 26–47, respectively. In the 300 K dataset, however, EPRS-R1 had a significantly less stable helix frequency for residues 9–13 than the other regions. This unstable region, referred as “Region(9–13),” is a part of the H1 helix in the native structure. The instability appears to result from a helix-breaker,¹⁷ Gly, at residue 13. Helix breakers (Gly and Pro) exist also in non-helical regions (residues 25 and 48–57) as indicated in Figure 6. Gly13 is the only helix breaker in the middle of the H1 helix in the native structure. We wondered how Region(9–13) could form an α -helix in the native structure in spite of the presence of a helix breaker.

We next computed the temperature dependence of the helix frequency at various temperatures. As expected, the helix frequency decreased monotonically with increasing temperature, and the structural instability of Region(9–13) was observed at all of the temperatures examined, as also shown in Figure 6.

Figure 4. a: RMSD_{core}. b: Q value of each cluster at 300 K.

Figure 5. Structures taken from the “lowest,” “second lowest,” and “third lowest” free-energy clusters. Two structures are displayed for each cluster. Structures from minor clusters are denoted as “minor.” The H1 and H2 regions are shown in gray and black, respectively.

Hydrophobic contacts required for proper folding

We hypothesized that the helix in Region(9–13) could be stabilized by hydrophobic contacts with the rest of the protein because the H1 and H2 helices interact through hydrophobic interactions in the NMR models.¹⁴ The hydrophobic contacts should presumably prevent hydrogen-bonds between Region(9–13) and water by exclusion of the water molecules, and it should stabilize inter-residual hydrogen bonds necessary for helix formation of Region(9–13).

We first examined the number of residue–residue hydrophobic contacts (N_{hc1}) between Region(9–13) and the other protein regions (residues 1–7 and 15–57) at 300 K. The detailed definition for the hydrophobic contacts and N_{hc1} is given in subsection “Hydrophobic contacts” of the Materials and methods section. We classified the structures in the 300 K dataset into three structure groups: Group_{4–5} consisting of structures with 4–5 helical residues in Region (9–13), Group_{1–3} consisting of those with 1–3 residues, and Group₀ consisting of those with no helical residues. The definition of a helical residue was given in the preceding subsection. Figure 7(a) plots the probability of detecting N_{hc1} in each structural group. This figure presents a correlation between the hydrophobic-contact formation and helix formation in Region(9–13): The greater the number of hydrophobic contacts, the more the helical content. Next, we investigated the correlation between N_{hc1} and the number of water molecules (N_{wat}) in the vicinity of Region(9–13). Water molecules around Region(9–13) were defined as those within 5.5 Å of a protein heavy atom in Region(9–13). This definition was based on the sum of the van der Waals radius of a heavy atom (≈ 2.0 Å), the radius of a water molecule (≈ 1.5 Å), and a constant value for tolerance (2.0 Å). Figure 7(b) clearly illustrates that increasing N_{hc1} excludes water molecules from the environment of Region(9–13).

Thus, the hydrophobic interactions have an essential role for helix formation in Region(9–13). In particular, the hydrophobic contacts between the H1 and H2 regions are expected to be important for the proper folding of EPRS-R1 because Region(9–13) interacts with H2 in the native topology. Then, we computed the number of residue–residue hydrophobic contacts between Region(9–13) and H2 (N_{hc2}), where the residue–residue contacts were counted between the hydrophobic residues in Region(9–13) and those only in H2. We calculated N_{hc2} for the following two structure groups: the structures in the lowest free-energy cluster and those in the 300 K dataset. Clearly, in the lowest free-energy cluster, the helical residues in Region(9–13) appeared to be the most probable at a large N_{hc2} value ($= 6$) [Fig. 7(c)], indicating that the hydrophobic contacts between Region(9–13) and H2 are important for the proper folding. In the 300 K dataset, the N_{hc2}

Figure 6. Helix frequency at each amino acid residue computed from the 300 K dataset (thick solid line), 500 K (thin solid line), 700 K (thin dotted line), and a structure with the most helical residues (dot-dash-line). Black and white circles represent locations of helix breakers, Gly and Pro, respectively.

Figure 7. a: Probability distribution of residue-residue hydrophobic constant (N_{hc1}) in three structural groups, Group₄₋₅, Group₁₋₃, and Group₀, of the 300 K dataset. See main text for the definition of the groups. Y-axis presents probability of N_{hc1} in Group₄₋₅ (thick solid line), Group₁₋₃ (thick dot line), and Group₀ (thick dot-dash-line). Distribution from 20 NMR models is shown by thin solid line. b: Relation between N_{hc1} and N_{wat} . c: Probability distribution of N_{hc2} computed from the lowest free-energy cluster. d: Probability Distribution of N_{hc2} from the 300 K dataset.

distribution for the 4–5 helical residues was broader than that of the lowest free-energy cluster [Fig. 7(d)]. Note that in the 300 K dataset, structures with 4–5 helical residues involve hydrophobic contacts between Region(9–13) and other regions. As shown in Figure 7(a), the helix in Region(9–13) is stabilized by these general hydrophobic interactions, while the proper folding requires the particular hydrophobic contacts between Region(9–13) and H2.

The above analysis shows the importance of proper hydrophobic contacts to form the native-fold structure. However, this does not necessarily mean that hydrophobic interactions are sufficient to drive native-like folding. Another possibility is that electrostatic interactions drive the attraction between the two helices. We analyzed this possibility and concluded that the electrostatic interactions do not considerably contribute to the attraction between H1 and H2 (Supporting Information Correlation between charged residues). This conclusion is consistent with the NMR models where charged amino acids are exposed to the solvent. Nevertheless, although there is no significant ionic charge distribution that would support the folding of EPRS-R1,

the approach of the two anti-parallel helices may be influenced by the anti-parallel helical dipoles (Supporting Information Helix-helix dipole interaction).

We have used an explicit water model for the solvent in our McMD simulations^{8,9,18–21} to gain detailed understanding of the protein folding process. Implicit solvent models such as the generalized Born (GB) method^{22–26} are widely used in protein simulations. Recently, a REMD simulation of a 57-residue protein, the same chain length as our target protein, was performed using GB, starting from the native structure to compute the protein heat capacity.²⁷ Implicit solvent is a convenient simulation tool to save processing time for large systems. However, such simulations cannot explicitly detect the water-protein hydrogen bonds or water-protein contacts. Furthermore, some studies have indicated that GB models may induce serious errors in structural and/or dynamical properties of polypeptides.^{28–31} For example, we performed McMD simulations of a 25-residue polypeptide using both explicit and implicit GB solvent models and found that the GB model has a strong propensity to enhance helix formation.¹⁰ Here, we show that the use of an explicit

water model is critical for providing key insights into the folding process of 57-residue protein.

Materials and Methods

Target protein

EPRS-R1 is located between two catalytic domains of EPRS, and the solution structure has been determined by NMR spectroscopy (PDB ID: 1FYJ).¹⁴ We chose this segment because of its appropriate length (57 residues) for this study and because the experimental conditions (temperature: 303 K; solvent: water; pH 5.0) could be reproduced in our simulation. The amino acids Asp, Glu, Arg, and Lys in the protein were treated as charged in our simulation to reproduce the experimental environment at pH 5.0. The protein consists of 913 atoms and adopts a helix-hairpin fold with two anti-parallel α -helices (H1 and H2) [Fig. 1(a)]. The two helices interact with each other via hydrophobic residues.¹⁴ The terminal regions (residues 1 and 48–57) have no secondary structure; thus, we refer to the 46-residue region (residues 2–47) forming the secondary structures as the “core-region.” We used the core-region for structural similarity analyses.

Computational procedure

We prepared the initial conformation for the simulation as follows. First, we generated a fully extended conformation of the protein (Figure 1b). Then, we performed a high-temperature (1000 K) canonical MD simulation for 10 ns starting from the extended conformation to embed the protein into the interior of a sphere (called sphere 1; diameter = 64 Å), which was filled by solvent molecules (4234 water molecules and 2 chloride ions). The number of water molecules was determined to set the density of water to 1 g/cc in advance, and the ions were introduced to compensate the net charge of protein (+2.0), respectively. Finally, the total number of atoms in the system was 13617. In the high-temperature run, a harmonic potential was applied to the protein heavy atoms only when those atoms were outside a smaller sphere (diameter = 56 Å), concentric to sphere 1. Similarly, to avoid evaporation of water molecules from sphere 1, another harmonic potential was applied to water-oxygen atoms only when they were outside sphere 1. These two restoring potentials were also used during the subsequent McMD simulation, explained later.

To position the protein in the middle of sphere 1, we moved the protein mass center to the center of sphere 1 in the high-temperature simulation. However, no positional shift was done for the protein in the subsequent McMD simulation. Instead, the linear and angular momenta of the protein were constrained to zero.

We used force field parameters taken from an AMBER-based hybrid force field³² and a TIP3P water model³³ for the protein and the water

molecules, respectively. The AMBER-based hybrid force field (E_{hybrid}) is a mixture of AMBER parm94³⁴ (E_{94}) and parm96³⁵ (E_{96}) force fields: $E_{\text{hybrid}}(w) = (1 - w)E_{94} + wE_{96}$, where w is a weighting factor. We previously assessed the physicochemical validity of E_{hybrid} by comparing the free-energy landscape of short peptides from an McMD calculation with one from a quantum chemical calculation³² and showed that E_{hybrid} for $0.45 \leq w \leq 0.95$ is better than either E_{94} or E_{96} . Additionally, McMD simulations of a 25-residue polypeptide using different w values⁸ indicated that E_{hybrid} (0.70) optimally reproduced the α and β secondary-structure contents experimentally measured, so we used E_{hybrid} (0.70) for this work.

We used the computer program PRESTO ver. 3³⁶ for McMD with a time step of 1 fs, using the SHAKE algorithm³⁷ to constrain the geometry of atom groups X-H_{*i*} (X is a heavy atom and *i* is a number of hydrogen atoms covalently bonding to X), with the cell-multipole expansion method³⁸ to compute long-range electrostatic interactions, and a constant-temperature method³⁹ to control the temperature.

McMD simulation

The McMD method¹¹ can efficiently sample a wide conformational space without being trapped in energy local minima. We will briefly explain the method below. First, we describe a canonical energy distribution conventionally as:

$$P_c(E, T) = n(E) \exp[-E/RT] / Z_c(T), \quad (1)$$

where E is the potential energy, $n(E)$ the density of states, T the temperature of the system, R is the gas constant, and $Z_c(T) = \int n(E) \exp[-E/RT] dE$ is the partition function at T .

A multicanonical sampling algorithm introduces a modified potential energy to increase the conformational sampling efficiency:

$$E_{\text{mc}}(E) = RT_0 \ln[n(E)] = E + RT_0 \ln P_c(E, T_0), \quad (2)$$

where T_0 is the simulation temperature. McMD is a canonical MD simulation with using E_{mc} to derive forces acting on atoms: $\text{force} = -\nabla E_{\text{mc}}$. Suppose that the canonical energy distribution $P_c(E, T_0)$ is accurately estimated in an energy range, then McMD produces a flat energy distribution $P_{\text{mc}}(E, T_0)$ in the energy range as:

$$\begin{aligned} P_{\text{mc}}(E, T_0) &= \frac{n(E) \exp[-E_{\text{mc}}/RT_0]}{Z_{\text{mc}}(T_0)} \\ &= \frac{n(E) \exp[-E/RT_0]}{P_c(E, T_0) Z_{\text{mc}}(T_0)} = \text{const}, \end{aligned} \quad (3)$$

where $Z_{\text{mc}}(T_0) = \int n(E) \exp[-E_{\text{mc}}/RT_0] dE$.

The flatness of $P_{\text{mc}}(E, T_0)$ ensures that the conformation overcomes energy barriers that exist in the conformational space. Since $P_c(E, T_0)$ is unknown a priori, iterative simulations are required to accurately determine $P_c(E, T_0)$. We set the last snapshot of the i th iteration to the initial structure for the $(i + 1)$ -th iteration. After the final iteration (41-th iteration in the current work), we perform the production McMD simulation to generate a conformational ensemble, which is used for subsequent analyses.

The canonical distribution $P_c(E, T)$ at any temperature T , which is not the simulation temperature T_0 , is computed from $n(E)$ by a reweighting scheme⁴⁰ as follows:

$$P_c(E, T) = \frac{n(E) \exp[-E/RT]}{Z_c(T)} = \frac{P_{\text{mc}}(E, T_0) \exp[E_{\text{mc}}(T)/RT_0] \exp[-E/RT]}{Z_c(T)/Z_{\text{mc}}(T_0)}. \quad (4)$$

Since $Z_c(T)/Z_{\text{mc}}(T_0)$ is regarded as a normalization factor for $P_c(E, T)$, we do not compute the absolute values for Z_{mc} and Z_c . A canonical conformational ensemble at T is generated by conformations taken from the entire ensemble with the statistical weight of $P_c(E, T)$.

The McMD simulation trajectory is not equivalent to a time series of conformational motions. Unlike the case of a canonical MD simulation, a protein does not trace realistic conformational changes in an McMD simulation due to use of the modified potential energy E_{mc} . The meaningful quantity from the McMD simulation is the conformational ensemble, which is reweighted at each temperature.

NER score to estimate conformational similarity

To assess the conformational similarity among the structures sampled from the McMD simulation, we used the Number of Equivalent Residues (NER) score as implemented in the program ASH, which is a structure alignment package made available to the public by the Protein Data Bank, Japan (PDBj) (<http://www.pdbj.org/ash/>).^{41,42} This score has been used to discriminate between different structural folds.⁴³ The NER score between two structures i and j is defined as: $\text{NER}_{ij} = \sum_k \exp[-(d_k/d_{\text{cut}})^2]$, where k is the ordinal number of a residue in the core-region (residues 2–47), d_k is the distance between C α atoms of the k th residue in two structures, when the core-region is superimposed, and d_{cut} is set empirically to 6 Å. We converted NER_{ij} into a pseudo distance to construct a conformational space (Supporting Information Multidimensional scaling):

$$d(i, j) = N_{\text{res}} - \text{NER}_{ij}, \quad (5)$$

where N_{res} is the number of residues compared (46 residues involved in the core-region). Both NER_{ij}

and $d(i, j)$ are nondimensional quantities.⁴³ The maximum pseudo distance is 46, which means that the two conformations are largely different from each other, and the minimum is zero, which means that the core-regions of the two structures are exactly the same. Hou *et al.* used a similar pseudo distance in their work.⁴⁴

Structure clustering

Average linkage clustering was carried out for the sampled structures as follows: First, each structure in a conformational ensemble was treated as a cluster. We denote the number of clusters as N_{clust} . Next, the nearest-cluster pair in the $N_{\text{clust}}(N_{\text{clust}} - 1)/2$ intercluster pairs was merged as a new cluster (a merger step). The new intercluster distance is an average of the pseudo distances [Eq. (5)] between the structures belonging to the two clusters. Here, we denote the intercluster distance for the nearest clusters as r_{nearest} . Repeating the merger, N_{clust} decreases by one. We must terminate the clustering at a merger step, because N_{clust} is not set in advance. If the clustering is terminated when r_{nearest} becomes larger than a threshold D_{thre} , the resultant N_{clust} is expressed as a function of D_{thre} . In the 300 K dataset, N_{clust} had a notable inflection point at $D_{\text{thre}} = 32$ (Supporting Information Dependence of N_{clust} on D_{thre}). Thus, we set $D_{\text{thre}} = 32$ for the clustering.

Hydrophobic contacts

To estimate intramolecular residue–residue hydrophobic contacts in protein, we first defined hydrophobic side-chain heavy atoms as: all side-chain heavy atoms of eight hydrophobic amino acids (Ala, Val, Leu, Ile, Phe, Pro, Met, and Trp); C β of Asn and Asp; C β and C γ of Gln and Glu; C β , C γ , C δ atoms of Arg and Lys; C γ 2 of Thr; and C β , C γ , C δ 1, C δ 2 C ϵ 1, and C ϵ 2 of Tyr. Although the side-chain tips of the latter eight amino acids are charged or polar, the non-polar heavy atoms at the root of their side-chains can participate to hydrophobic contacts. When the distance between two hydrophobic heavy atoms was smaller than 6.5 Å, we judged that the two residues, to which the heavy atoms belong, formed a residue–residue hydrophobic contact. By counting these contacts for a given tertiary structure, we obtained the number intramolecular residue–residue hydrophobic contacts (N_{hc1}).

Conclusions

We performed an McMD simulation of a 57-residue protein, EPRS-R1, in explicit solvent starting from a fully extended conformation. Native-like structures were found in the lowest free-energy cluster at 300 K, using a force field that was able to produce both α - and β -secondary structures for typical short peptides. The H1 helix was unstable due to the presence

of Gly13. Subsequent analysis showed that hydrophobic contacts are essential to stabilize the helix in Region(9–13) in the native-fold structure. Growth of the hydrophobic contacts excludes water molecules from the environment of Region(9–13), resulting in the emergence of a hydrophobic environment. Furthermore, we confirmed that electrostatic interactions due to ionic residues did not correlate with the approach of the two helices. This study provides an *ab initio* folding simulation that reproduced the native-like fold within the free energy minimum cluster. Moreover, it revealed key factors in the folding of a small protein in explicit solvent using a physics-based force field without any reference to structural templates or to the native structure.

References

1. Anfinsen CB (1973) Principles that govern the folding of protein chains. *Science* 181: 223–230.
2. Ding F, Dokholyan NV (2005) Simple but predictive protein models. *Trends Biotechnol* 23: 450–455.
3. Hansmann UHE, Okamoto Y, Eisenmenger F (1996) Molecular dynamics, Langevin and hybrid Monte Carlo simulations in a multicanonical ensemble. *Chem Phys Lett* 259: 321–330.
4. Kidera A (1995) Enhanced conformational sampling in Monte Carlo simulations of proteins: application to a constrained peptide. *Proc Natl Acad Sci USA* 92: 9886–9889.
5. Mitsutake A, Sugita Y, Okamoto Y (2001) Generalized-ensemble algorithms for molecular simulations of biopolymers. *Biopolymers* 60: 96–123.
6. Sugita Y, Okamoto Y (1999) Replica-exchange molecular dynamics method for protein folding. *Chem Phys Lett* 314: 141–151.
7. Wei G, Shea JE (2006) Effects of solvent on the structure of the Alzheimer amyloid-beta(25–35) peptide. *Biophys J* 91: 1638–1647.
8. Ikebe J, Kamiya N, Ito J, Shindo H, Higo J (2007) Simulation study on the disordered state of an Alzheimer's beta amyloid peptide Aβ(12–36) in water consisting of random-structural, beta-structural, and helical clusters. *Protein Sci* 16: 1596–1608.
9. Ikebe J, Kamiya N, Shindo H, Nakamura H, Higo J (2007) Conformational sampling of a 40-residue protein consisting of α and β secondary-structure elements in explicit solvent. *Chem Phys Lett* 443: 364–368.
10. Mitomo D, Watanabe YS, Kamiya N, Higo J (2006) Explicit and GB/SA solvents: each with two different force fields in multicanonical conformational sampling of a 25-residue polypeptide. *Chem Phys Lett* 427: 399–403.
11. Nakajima N, Nakamura H, Kidera A (1997) Multicanonical ensemble generated by molecular dynamics simulation for enhanced conformational sampling of peptides. *J Phys Chem B* 101: 817–824.
12. Gale Y, Householder A (1938) Discussion of a set of points in terms of their mutual distances. *Psychometrika* 3: 19–22.
13. Torgerson W (1952) Multidimensional scaling. I. Theory and method. *Psychometrika* 17: 401–419.
14. Jeong EJ, Hwang GS, Kim KH, Kim MJ, Kim S, Kim KS (2000) Structural analysis of multifunctional peptide motifs in human bifunctional tRNA synthetase: identification of RNA-binding residues and functional implications for tandem repeats. *Biochemistry* 39: 15775–15782.
15. Baumketner A, Shea J (2007) The structure of the Alzheimer amyloid[®] 10–35 peptide probed through replica-exchange molecular dynamics simulations in explicit solvent. *J Mol Biol* 366: 275–285.
16. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22: 2577–2637.
17. Levitt M (1978) Conformational preferences of amino acids in globular proteins. *Biochemistry* 17: 4277–4285.
18. Higo J, Ito N, Kuroda M, Ono S, Nakajima N, Nakamura H (2001) Energy landscape of a peptide consisting of alpha-helix, 3(10)-helix, beta-turn, beta-hairpin, and other disordered conformations. *Protein Sci* 10: 1160–1171.
19. Ikeda K, Galzitskaya O, Nakamura H, Higo J (2003) beta-Hairpins, alpha-helices, and the intermediates among the secondary structures in the energy landscape of a peptide from a distal beta-hairpin of SH3 domain. *J Comput Chem* 24: 310–318.
20. Ikeda K, Higo J (2003) Free-energy landscape of a chameleon sequence in explicit water and its inherent alpha/beta bifacial property. *Protein Sci* 12: 2542–2548.
21. Kamiya N, Higo J, Nakamura H (2002) Conformational transition states of a beta-hairpin peptide between the ordered and disordered conformations in explicit water. *Protein Sci* 11: 2297–2307.
22. Bashford D, Case DA (2000) Generalized Born models of macromolecular solvation effects. *Annu Rev Phys Chem* 51: 129–152.
23. Feig M, Im W, Brooks III CL (2004) Implicit solvation based on generalized Born theory in different dielectric environments. *J Chem Phys* 120: 903–911.
24. Scarsi M, Apostolakis J, Caflisch A (1997) Continuum electrostatic energies of macromolecules in aqueous solutions. *J Phys Chem A* 101: 8098–8106.
25. Schaefer M, Karplus M (1996) A comprehensive analytical treatment of continuum electrostatics. *J Phys Chem* 100: 1578–1599.
26. Still WC, Tempczyk A, Hawley RC, Hendrickson T (1990) Semianalytical treatment of solvation for molecular mechanics and dynamics. *J Am Chem Soc* 112: 6127–6129.
27. Yeh IC, Lee MS, Olson MA (2008) Calculation of protein heat capacity from replica-exchange molecular dynamics simulations with different implicit solvent models. *J Phys Chem* 112: 15064–15073.
28. Cheung MS, Garcia AE, Onuchic JN (2002) Protein folding mediated by solvation: water expulsion and formation of the hydrophobic core occur after the structural collapse. *Proc Natl Acad Sci USA* 99: 685–690.
29. Masunov A, Lazaridis T (2003) Potentials of mean force between ionizable amino acid side chains in water. *J Am Chem Soc* 125: 1722–1730.
30. Nymeyer H, Garcia AE (2003) Simulation of the folding equilibrium of alpha-helical peptides: a comparison of the generalized Born approximation with explicit solvent. *Proc Natl Acad Sci USA* 100: 13934–13939.
31. Zhou R (2003) Free energy landscape of protein folding in water: explicit vs. implicit solvent. *Proteins* 53: 148–161.
32. Kamiya N, Watanabe YS, Ono S, Higo J (2005) AMBER-based hybrid force field for conformational sampling of polypeptides. *Chem Phys Lett* 401: 312–317.
33. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. *J Chem Phys* 79: 926–935.

34. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA (1995) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J Am Chem Soc* 117: 5179–5197.
35. Kollman PA, Dixon RW, Cornell WD, Chipot C, Pohorille A (1997) The development/application of a ‘minimalist’ organic/biochemical molecular mechanic force field using a combination of ab initio calculations and experimental data. In *computer simulations of biological systems*. Dordrecht: Springer.
36. Morikami K, Nakai T, Kidera A, Saito M, Nakamura H (1992) PRESTO (PRotein Engineering SimulaTOR): a vectorized molecular mechanics program for biopolymers. *Comput Chem* 16: 243–248.
37. Ryckaert J, Ciccotti G, Berendsen H (1977) Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J Comput Phys* 23: 327–341.
38. Ding H-Q, Karasawa N, Goddard WA, III (1992) Atomic level simulations on a million particles: the cell multipole method for Coulomb and London nonbond interactions. *J Chem Phys* 97: 4309–4315.
39. Evans DJ, Morriss GP (1983) The isothermal/isobaric molecular dynamics ensemble. *Phys Lett A* 98: 433–436.
40. Shirai H, Nakajima N, Higo J, Kidera A, Nakamura H (1998) Conformational sampling of CDR-H3 in antibodies by multicanonical molecular dynamics simulation. *J Mol Biol* 278: 481–496.
41. ASH. Available at: <http://www.pdbj.org/ash/>.
42. Standley DM, Toh H, Nakamura H (2007) ASH structure alignment package: sensitivity and selectivity in domain classification. *BMC Bioinform* 8: 116.
43. Standley DM, Toh H, Nakamura H (2004) Detecting local structural similarity in proteins by maximizing number of equivalent residues. *Proteins* 57: 381–391.
44. Hou J, Sims GE, Zhang C, Kim SH (2003) A global representation of the protein fold space. *Proc Natl Acad Sci USA* 100: 2386–2390.