JCTC Journal of Chemical Theory and Computation

ARTICLE

pubs.acs.org/JCTC

The Fast-Folding Mechanism of Villin Headpiece Subdomain Studied by Multiscale Distributed Computing

Ryuhei Harada[†,‡,§] and Akio Kitao[*,†,‡,§]

[†]Department of Physics, Graduate School of Science, The University of Tokyo, Tokyo, 7-3-1, Hongo, Bunkyo-ku 113-0033, Japan
[‡]Institute of Molecular and Cellular Bioscience, The University of Tokyo, 1-1-1 Yayoi, Bunkyo-ku, Tokyo 113-0032, Japan
[§]Japan Science and Technology Agency, Core Research for Evolutional Science and Technology, 1-1-1 Yayoi, Bunkyo-ku,
Tokyo 113-0032, Japan

ABSTRACT: The fast-folding mechanism of a 35-residue mini-protein, villin headpiece subdomain (HP35), was investigated using folding free energy landscape analysis with the multiscale free energy landscape calculation method (MSFEL). A major and a minor folding pathway were deduced from the folding free energy landscape. In the major folding pathway, the formation of helices II and III was the rate-limiting step in the transition to an intermediate state, triggered by the folding of the PLWK motif. HP35 then folds into the native structure through the formation of the hydrophobic core located at the center of the three-helix bundle. Mutations in the motif and hydrophobic core that suppressed folding into the native state drastically changed the folding free energy landscape compared to the wild type protein. In the minor folding pathway, nucleation of the hydrophobic core preceded formation of the motif.

1. INTRODUCTION

The folding of proteins into their native structures plays an important role in many biological processes. Protein misfolding and aggregation into amyloid-like fibrils are thought to cause some diseases.[1−5] Recent dramatic advances in the power of computational methods have enabled researchers to trace a series of folding pathways from molecular dynamics (MD) simulations on the order of microsecond time scales. Significant progress has been made recently in *ab initio* folding simulations of some fast-folding mini-proteins, including chignolin (10 residues),[6,7] Trp-cage (20 residues),[8,9] and HP35 (35 residues).[10−13] Predicted structures of small proteins typically deviate from experimental structures determined by NMR or X-ray crystallography by 2−4 Å in the C_α atoms. Conventional MD (CMD) allows for direct observation of the time course of folding events. Information obtained in the form of MD trajectories provides dynamic pictures of conformational transitions. However, generating multiple trajectories sufficient for establishing kinetic views of protein folding is challenging.

The free energy landscape (FEL) introduces an important concept for investigating free energy changes on a given reaction coordinate space from a statistical point of view. However, the complex energy surface prevents accurate conformational sampling of the FEL due to trappings into local energy minima, resulting in bad conformational samplings. Therefore, samplings of rare events like conformational transitions jumping among minima are important to calculate the FEL accurately. To overcome this sampling problem, many computational methodologies have been developed, including the extended ensemble method like the multicanonical MD (McMD)[14,15] and replica exchange molecular dynamics (REMD)[16,17] methods. In the McMD method, a non-Boltzmann sampling enables random walks on the energy space without trapping in local energy minima. A target ensemble can be reconstructed by reweighting. In the REMD, a set of simulations is performed at different temperatures, and temperature exchanges are periodically attempted according to the Metropolis criterion,[18] which attains random walks in the temperature space and prompt escape from local energy minima. The combination of the REMD with the umbrella sampling like REUS (replica-exchange umbrella sampling)[19] and bias-exchange method[20] is also an effective approach to enhancing conformational sampling. In these methods, the umbrella potentials are exchanged in the REMD as well as temperatures. Metadynamics[21] is a powerful method that can be used both for calculating free energy and for accelerating rare events in systems. In this method, the normal evolution of the system is biased by a history-dependent potential constructed as a sum of Gaussians centered along the trajectory followed by a suitably chosen set of collective variables. The sum of Gaussians is used for reconstructing iteratively as an estimator of the free energy and forcing the system to escape from local minima. The Wang−Landau method[22] is an extended Monte Carlo method for calculating the density of state efficiently by performing independent random walks in different and restricted ranges of energy. The resultant density of states is modified continuously to produce locally flat histograms. This method permits us to directly access the free energy and entropy. The transition path sampling method[23] is a method for sampling a conformational transition called the "reactive path" that connects a given reactant and product conformation starting from an arbitrary initial transition path. The reactive transitional paths are sampled with the Metropolis criteria so as to hold a certain ensemble. This method has been successful in folding studies and conformational samplings of small proteins.[24−27] Transform and relax sampling (TRS) was successful in sampling protein domain motion and mini-protein folding.[28] Recently, we proposed a new approach to calculating the FEL, called the multiscale free energy calculation method (MSFEL).[29] In this method, multiple conformations are generated to cover a broad conformational space with a coarse-grained (CG) model. Distributed all-atom (AA)

Received: May 31, 2011
Published: December 08, 2011

MD simulations are then performed using umbrella sampling[30,31] to sample local energy landscapes in parallel. Finally, the FEL is calculated using the weighted histogram analysis method (WHAM).[32−34] The MSFEL method has been applied to the study of short peptides and mini-proteins, and the efficiency of the FEL calculation has been demonstrated.[29,35]

In this study, we investigated the fast-folding mechanism of the 35 amino acid residue villin headpiece subdomain (HP35) in explicit solvent using the MSFEL method. HP35 is an F-actin-binding domain located on the far C-terminus of the super villin.[36,37] HP35 can spontaneously fold into its native structure within microseconds without the assistance of disulfide bonds and metal ions. The native structure of HP35 has been determined using both NMR and X-ray crystallography in high resolution.[38,39] Since HP35 is small in size and folds quickly and cooperatively, it has also been studied extensively using kinetic experiments,[40,41] mutagenesis,[42,43] and computer simulations.[10−13] The folding FEL of HP35 by computer simulations has been first investigated in implicit solvent,[12] and intermediate and transitional conformations on the folding pathway have been reported.[13] The folding of HP35 has been also investigated in explicit solvent.[44−49] In experiments, mutational analyses of key residues have revealed the mechanism of fast folding.[50,51]

For further understanding of the folding mechanism of HP35, especially, the information of the folding FEL in explicit solvent, which is still difficult to measure by experimentation, is an important factor in the filling of gaps between experiments and computations. The accurate estimations of free energy difference from the folding FEL calculations are quite meaningful for comparing and supporting experimental data in addition to the elucidation of the folding pathway. Therefore, we focused on the folding FEL of HP35 in explicit solvent and more accurately calculated the folding FEL using the MSFEL.

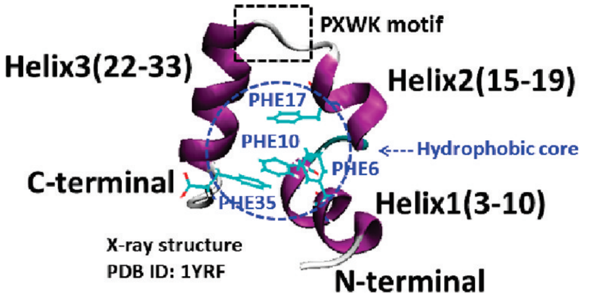

Figure 1. The native structure of HP35. The hydrophobic core residues, PLWK motif, and the definition of the two segments are shown. Figure created using VMD.[67]

2. MATERIAL AND METHODS

Implementation of the MSFEL. A MSFEL analysis consists of four stages. In the first stage, a CG MD simulation is performed to sample a broad conformational space. To address the folding of HP35, a replica exchange MD (REMD) simulation[17] with a CG model was used to further enhance the conformational sampling. The CG REMD simulations were performed by our original MD program developed for the MSFEL. We employed a C_α-based CG model[29] to widely sample the conformational space around a reference structure. The potential energy function was defined as the sum of bond, angle, torsion, and Lennard-Jones-type energy terms as follows:

$$\begin{aligned} V^{CG}(\vec{r}^{\,C\alpha}|\,\vec{r}^{\,C\alpha 0}) = & \sum_{|i-j|=1,\, i<j} k_{12}(r_{ij}^{C\alpha} - r_{ij}^{C\alpha 0})^2 \\ & + \sum_{|i-j|=2,\, i<j} k_{13}(r_{ij}^{C\alpha} - r_{ij}^{C\alpha 0})^2 \\ & + \sum_{|i-j|=3,\, i<j} k_{14}(r_{ij}^{C\alpha} - r_{ij}^{C\alpha 0})^2 \\ & + \sum_{\substack{|i-j|<rc,\\ |i-j|>3}} k_{LJ}\left[\left(\frac{r_{ij}^{C\alpha 0}}{r_{ij}^{C\alpha}}\right)^{12} - \left(\frac{r_{ij}^{C\alpha 0}}{r_{ij}^{C\alpha}}\right)^{6}\right] \end{aligned} \qquad (1)$$

where $r_{ij}^{C\alpha 0}$and $r_{ij}^{C\alpha}$ represent the distance between the ith and jth atoms of the reference and instantaneous structures, respectively. The forth term is associated with the atom pairs within the cutoff distance r_C = 10.0 Å in the native structure. To build the CG potential, the reference C_α coordinates were adopted from the X-ray structure of HP35 (Protein Data Bank ID: 1YRF)[38] shown in Figure 1 and were also used as the initial structures of the CG REMD. The global energy minimum of this potential function is designed to be the X-ray structure. Therefore, this CG model can be considered a Go-like model, a type of model widely used in protein folding studies.[52−54] The parameter $r_{ij}^{C\alpha 0}$ = 3.8 Å was determined first to reproduce the optimal distance between two adjacent C_α atoms. The value of k_{12} was determined to best reproduce the distribution with the relation between the variance of the Gaussian at 300 K, $\sigma = 5.651 \times 10^{-2}$ Å, $\beta = 5.919 \times 10^{-1}$ kcal/mol, and $k_{12} = 1/\beta\sigma^2$. The other parameters were defined by the ratios to k_{12} as $k_{13}/k_{12} = 1/5$ and $k_{14}/k_{12} = k_{LJ}/k_{12} = 1/100$, and the Newtonian equation of motion is integrated by a time step 15 ps. The CG REMD simulations were performed with 10 replicas at exponentially distributed temperatures of 200, 239, 286, 342, 404, 489, 585, 700, 853, and 1041 K. A 10^6-step production CG MD run after a 10^3-step equilibration was performed under a canonical ensemble with a Nosé−Hoover chain thermostat.[55] The two replicas with neighboring temperatures were exchanged every 10^3 steps, and a total of 10^4 snapshots were recorded every 10^2 steps. The average exchange rate between replicas was 0.31, which is sufficiently high to achieve efficient sampling.

In the second stage, multiple representative structures were chosen. These structures should roughly cover the entire conformational space sampled in the CG MD analysis and be distributed densely enough so that each AA MD trajectory in the third stage significantly overlaps with its neighboring trajectories. From structures obtained using the CG model, BBQ[56] was used to generate main-chain atoms from the C_α coordinates. Next, SCWRL[57] was employed to generate side-chain atoms. A total of 100 AA structures were constructed in this study. The C_α coordinates were picked up from 10 REMD trajectories with equal intervals.

In the third stage, independent AA MD simulations were conducted to investigate each local FEL more accurately around the distributed initial structures. The 100 AA structures were solvated in a rectangular box (63.9 Å × 55.2 Å × 48.1 Å) containing 3456 TIP3P water molecules.[58] Two chloride ions were also added to neutralize the system. The AA MD simulations were independently performed using the PMEMD module of the Amber 9.0 software[59] with the Amber parameter ff03 force field.[60] For the purpose of intensive local sampling, we employed the umbrella sampling method[30,31] and used harmonic positional restraints for the C_α atoms as the umbrella potentials.[29] Short energy minimizations and 300 ps relaxation MD runs that included density adjustments with an isothermal−isobaric ensemble at 300 K and 1 bar were then conducted using the Berendsen method. The systems were equilibrated with a canonical ensemble for 100 ps with harmonic restraints (1.0 × 10^{-4} kcal/mol/Å^2) imposed on the C$_\alpha$ atoms (except for the N- and C-terminal residues) for umbrella samplings. Production runs were performed for 1 ns × 100 trajectories, and each trajectory was recorded every 1.0 ps. In both the equilibration and production runs, the temperature was maintained at 300 K, and the SHAKE algorithm[61] was used to enable the use of a long time step of 2 fs. Electrostatic interactions were calculated using the particle-mesh Ewald method[62] with a real space cutoff distance of 9 Å.

In the final stage, the probability distributions obtained in the previous stage were reweighted and combined using the WHAM.[32−34]

3. RESULTS AND DISCUSSION

3.1. Overview of the MSFEL. In the MSFEL method, the CG MD simulation is performed in the first stage to efficiently sample the conformational space. Figure 2 shows the time series of the C$_\alpha$ root-mean square deviation (C$_\alpha$-RMSD) of two segments defined as described previously.[12,13] The A and B segments are composed of helices I−II (residues 3−21) and helices II−III (residues 15−33), respectively. Segments A and B overlap with helix II in order to consider local folding between helices I and III with respect to helix II. As shown in Figure 2a, both segments folded and unfolded frequently during the CG MD simulation, which is indicative of efficient conformational sampling. For comparison, we performed a conventional long (100 ns) AA MD simulation at 300 K in explicit solvent after a 1 ns equilibration starting from the native structure (Figure 2b). In the CMD simulation, HP35 was trapped around the native state, resulting in insufficient conformational sampling for calculating an accurate folding FEL. These results indicate that the CG MD simulation enables enhanced conformational searching with relatively low computational cost. We also examined the growth of the three helices, helix I (residues 3−10), helix II (15−19), and helix III (22−33) in the CG MD. We defined an order parameter Φ as the ratio of the helical residues in each helix. For each snapshot of the CG MD, main-chain atoms were generated by BBQ,[56] and the secondary structure assignment was done by STRIDE.[63] As shown in the time of evolution of Φ (Figures 2c−e), folding and unfolding of the helices were frequently observed.

In the second stage, 100 CG structures were selected from trajectories of CG REMD simulations. From each trajectory that contains 10^4 snapshots of the C$_\alpha$ coordinates, 10 snapshots are selected at every 10^3 intervals (10 snapshots × 10 replicas). Then, 100 AA structures were generated from the C$_\alpha$ coordinates using CG-AA mappings. The projections of the CG REMD trajectories onto the subspace spanned by C$_\alpha$-RMSDs for segments A and B of the native structure (Figure 3a) and those of the 100 selected CG structures (Figure 3b) shows that the representative structures roughly cover the conformational space sampled in the CG level. The overlaps of the AA MD trajectories with its neighboring trajectories (Figure 3c) also indicate that the representative structures were dense enough (also see Figures 5a−c later). Since accuracy in CG-AA mapping is important, mappings were examined with a benchmark coordinate set similar to that used in our previous work.[29] A total of 10 000 AA coordinates that included both native and unfolded structures were generated using AA MD. Next, C$_\alpha$ coordinates were picked, and AA coordinates were reconstructed using CG-AA mapping. The heavy-atom RMSD of the reconstructed structures from the original ones were also examined. For main-chain mapping, the distribution of the RMSD had a very sharp peak at around the average value 0.55 Å, with a standard deviation of 0.01 Å, which can be considered as sufficiently small compared to thermal fluctuation. This is because the arrangement of the main-chain, with the exception of the termini, is almost determined by the geometrical condition of the C$_\alpha$ coordinates.[56] However, for side-chain mapping, the RMSD distribution had a broad peak averaging 1.97 Å, with a standard deviation of 0.21 Å. The coordinates of the generated side chains are determined as one of the optimal arrangements.[57] Since the alternative side-chain arrangement is

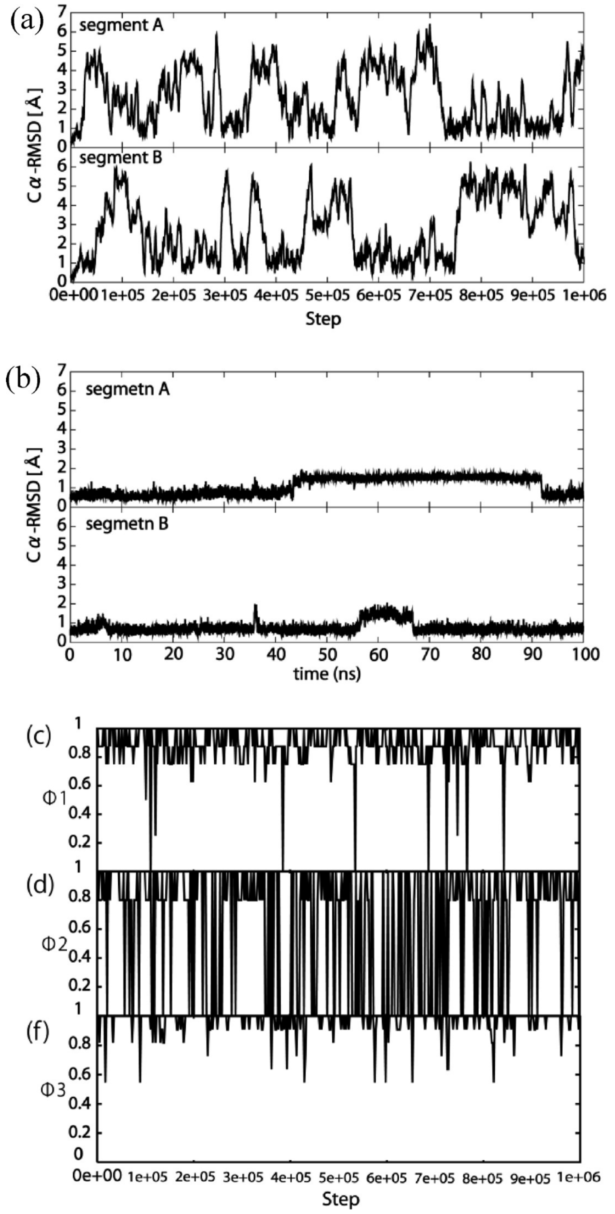

Figure 2. (a) Time series of the C$_\alpha$-RMSD of segment A (helices I and II) and segment B (helices II and III) from the X-ray structure during the CG MD simulation (replica 1). (b) Time series of the 100 ns all-atom CMD at 300 K in explicit solvent starting from the native structure. (c, d, e) Time series of the order parameters Φ for helices I, II, and III in the first replica of CG MD.

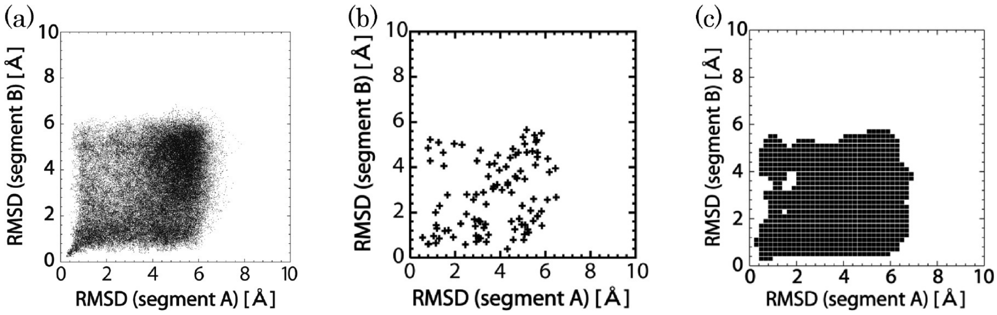

Figure 3. (a) Projections of the CG MD trajectories onto the subspace spanned by the C_α-RMSD of segment A and segment B and (b) those of the select 100 CG snapshots. (c) Overlapping regions among 100 distinct AA MD trajectories depicted by solid rectangles on the projected subspace. The overlap is counted if at least two distinct trajectories visit the rectangle. The size of each rectangle is 0.10 Å × 0.10 Å.

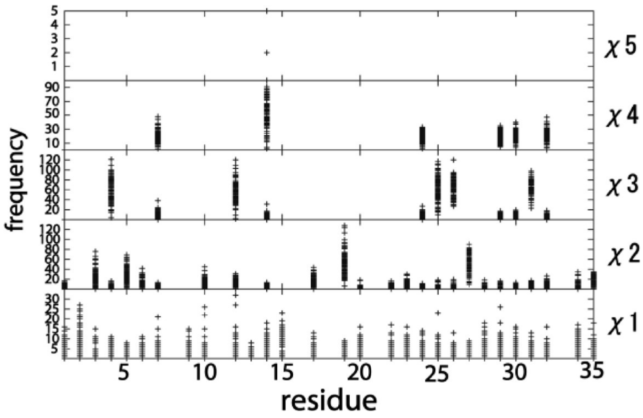

Figure 4. Transition frequencies of side-chain dihedral angles ($\chi_1 - \chi_5$) of each amino acid residue of HP35 in 100 1-ns AA MD simulations. Each symbol shows the transition frequency from one MD trajectory.

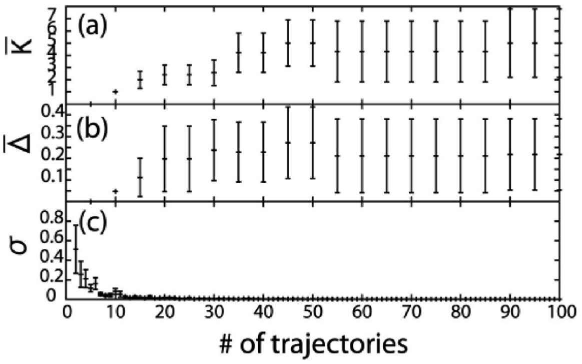

Figure 5. AA trajectory overlap and FEL convergence as a function of the number of trajectories considered, n. (a) The average number of overlapping trajectories per trajectory, $\bar{K}$. (b) The average fraction of overlapped snapshots between pairs of overlapping trajectories, . (c) The convergence of the probability distribution projected onto the corresponding subspace, σ.

possible, we examined whether rotamer transitions occur frequently during AA MD, as was shown in the case of Trpzip2.[29] Figure 4 shows rotamer transition frequencies during a 1 ns MD of HP35. Highly exposed side chains showed very high transition frequencies. Even in the well-packed side chains, the transition frequencies were sufficient to take the other rotamer states. This result suggests that the calculated FEL has minimal dependence on the choice of initial side-chain arrangements.

After 100 independent AA MD simulations in the third stage, the FEL was calculated in the fourth stage using the WHAM. To calculate the FEL properly, AA MD trajectories should significantly overlap with a sufficient number of neighboring trajectories. In addition, the convergence of the calculated probability density should be examined. We calculated the number of overlaps between trajectories per trajectory, K. A pair of trajectories is regarded as overlapped if the C_α-RMSD between the average structures is smaller than 1.0 Å. For each pair of overlapping trajectories, an all-to-all comparison is made among the snapshots, and the fraction of overlap Δ is estimated using the same C_α-RMSD criterion. One trajectory overlapped with $\bar{K} = 5.0 \pm 2.8$ trajectories, and $= 21.8 \pm 16.4\%$ of the snapshots overlapped in each pair of overlapping trajectories on average. We confirmed that all of the trajectories were not isolated and that all of the snapshots were connected in conformational space. To examine the convergence of $\bar{K}$ and values, 10 distinct series of the trajectories were prepared in random order. The value for the number of trajectories, $n = 30$, for example, indicates the quantity calculated with 30 trajectories and averaged over 10 distinct sets. The $\bar{K}$ and values were calculated as a function of n (Figures 5a,b). The $\bar{K}$ value rapidly increased from 0 to 5, but the rate of increase became slower in the range $n \geq 40$ (Figure 5a). The values almost converged when 30 or more trajectories were considered (Figure 5b). We also examined the convergence of probability distributions projected onto a two-dimensional subspace spanned by the reaction coordinates. As shown in Figure 5c, σ almost converged at around $n = 30$, corresponding to the convergence of trajectory overlap. From these results, we concluded that the calculated FEL is well-converged with 100 trajectories.

To examine the convergence of the FEL versus the tertiary packing, we focused on the hydrophobic core formed as aromatic stacking of PHE6, PHE10, PHE17, and PHE35 shown in Figure 1. The fraction of the native contacts among the four phenylalanine residues (NC) was chosen as the reaction coordinate to describe the tertiary packing. To consider whether the 1 ns AA MD simulation is sufficient, one-dimensional FELs projected onto the NC calculated from the first and second halves of the all-atom 1 ns trajectories and their difference are shown in Figure 6. Since the difference is significantly smaller than k_BT, we judged that the 1 ns AA MD simulations were sufficient in length.

3.2. Major and Minor HP35 Folding Pathways. Figure 7a shows the folding FEL of HP35 obtained using the MSFEL method (1 ns × 100 runs) projected onto the two-dimensional

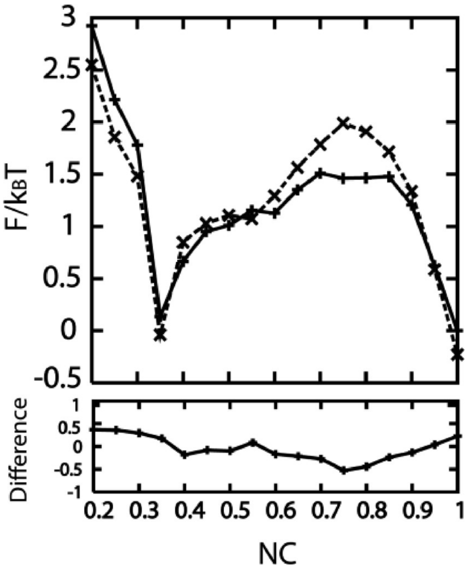

Figure 6. The one-dimensional free energy landscapes defined as the function of the fraction of the native contacts among hydrophobic core residues, PHE6, PHE10, PHE17, and PHE35. The solid and broken lines correspond to the free energy landscapes calculated from the first and second halves of the 1 ns trajectories, respectively. Hereafter, free energy value, *F*, is scaled by k_BT. Their difference is also shown below.

subspace spanned by the C_α-RMSD of two segments (R_A and R_B) from the crystal structure (PDB ID: 1YRF). In the folding FEL, four distinct states exist: denatured (D), major intermediate 1 (I1), minor intermediate 2 (I2), and native folded (N) (Figure 7a). To select the representative structures of these states, we first divided the subspace into four regions, D ($R_A > 2.0$ Å, $R_B > 2.0$ Å), I1 ($R_A > 2.0$ Å, $0 < R_B < 2.0$ Å), I2 ($0 < R_A < 2.0$ Å, $R_B > 2.0$ Å), and N ($0 < R_A < 2.0$ Å, $0 < R_B < 2.0$ Å), and calculated the weighted probability densities for grid points. The snapshot closest to the weighted average structure of the highest density grid in each region was selected as the representative structure.

To extract putative dynamic folding pathways, we first focused on the folding FEL (Figure 7a). As the first folding processes, the denatured protein folds into one of near intermediate states, I1 or I2. These two intermediate states can be distinguished from the denatured states by partial folding of segment A or B. Therefore, one-dimensional FEL projected onto the C_α-RMSD of each segment can describe the first folding processes, D→I1 (occurred in 1.0 Å < R_A) or D→I2 (occurred in 1.0 Å < R_B). The one-dimensional FEL on each first folding process was calculated as a double-well shape (Figure 9a,b). As the second folding processes, the intermediate structures fold into the native ones through partial folding of the remaining segment, I1→N (occurred in 1.0 Å > R_B) or I2→N (occurred in 1.0 Å > R_A). Thereby, the second folding process can be also described by the one-dimensional FEL projected onto the C_α-RMSD of each segment. The folding FEL on the second folding process also shows a double-well shape (Figure 9c,d). These pictures are the putative dynamical folding pathways extracted from the folding FEL shown in Figure 10a,b. As shown in Figure 9a,d, each state during D→I2→N was separated by relatively high free energy barriers compared to D→I1→N (Figure 9b,c). Therefore, D→I1→N is a more favorable route in the folding and defined as the major folding pathway, whereas D→I2→N as the minor folding pathway.

In the major folding pathway (D→I1→N), segment B forms first (D→I1), and then the remaining region in segment A docks with segment B through I1 to reach N (I1→N). This can be interpreted as a two-step folding process. This major folding pathway agrees well with previous results calculated using REMD simulation (20 replicas × 400 ns) with implicit solvent[13] and multicanonical replica-exchange (MUCAREM) molecular dynamics (MD) simulations (total 2.28 μs) with explicit solvent.[46]

In this folding process, the formation of segment B precedes the hydrophobic residue contacts (Phe6, Phe10, and Phe17) that form the hydrophobic core necessary for stabilizing the native structure shown in Figure 1. In the minor folding pathway (D→I2→N) on the other hand, formation of segment A precedes formation of segment B. This minor folding pathway was not described in previous reports.[12,13] Segment A in I2 is considered to be unstable because there is no hydrophobic side-chain stacking with segment B. As the energy barrier from I2 to N is relatively high (~6 k_BT), the minor folding process from I2 to N would be expected to occur infrequently.

The folding FEL of HP35 indicates that the formation of I1 is the rate-limiting step in the process. Previous experimental work supports this hypothesis, suggesting that a well-conserved, solvent-exposed PLWK motif (residues 21−24) in segment B is critical for fast folding.[50] This site is considered to function as a structural gatekeeper in the HP35 folding process. The rigid Pro21 situated in the linker region between helices II and III plays a crucial role in restricting the movement of the two helices. The formation of this site initiates the folding of segment B.

To characterize the formation of the PLWK motif and segment B, segment B2 (residues 3−24) consisting of helix II and segment B3 (residues 21−33) consisting of helix III were examined for overlap of the PLWK motif (21−24) in the two segments. Characterization of overlap in this region enabled us to determine whether formation of the PLWK motif is correlated with the formation of segment B2 or segment B3. Figure 8 shows the FELs using the RMSD of segments B2 and B3 from the native structure as the reaction coordinates. In the major folding pathway, formation of segment B precedes formation of segment A (D→I1). Therefore, in the first stage of folding, the FEL can be calculated without considering the formation of segment A ($R_A > 2.0$ Å). Furthermore, the first stage of the folding FEL is divided into two types depending upon whether the motif forms or not. Figure 8a,b shows the FEL in the first stage of folding. The two folding FELs obviously show that segment B begins to form with the formation of the motif (Figure 8b) and that segment B does not form without formation of the motif (Figure 8a). These results support the experimentally derived hypothesis that the motif is the structural gatekeeper of HP35.[50] Figure 8b also shows that the formation of helix II is followed by the formation of helix III. Therefore, the following folding pathway is suggested: formation of the PLWK motif is triggered first, and then segment B is formed from C-terminal helix III to helix II.

We hypothesize that in the second folding stage (I1→N) formation of the hydrophobic core acts as a driving force for folding into the native state. To examine this possibility, fractions of native contacts between hydrophobic residues (Phe6, Phe10, and Phe17) and the C_α-RMSD of segment A were employed as reaction coordinates. The folding FEL of the second stage was calculated under the condition that segment B has already

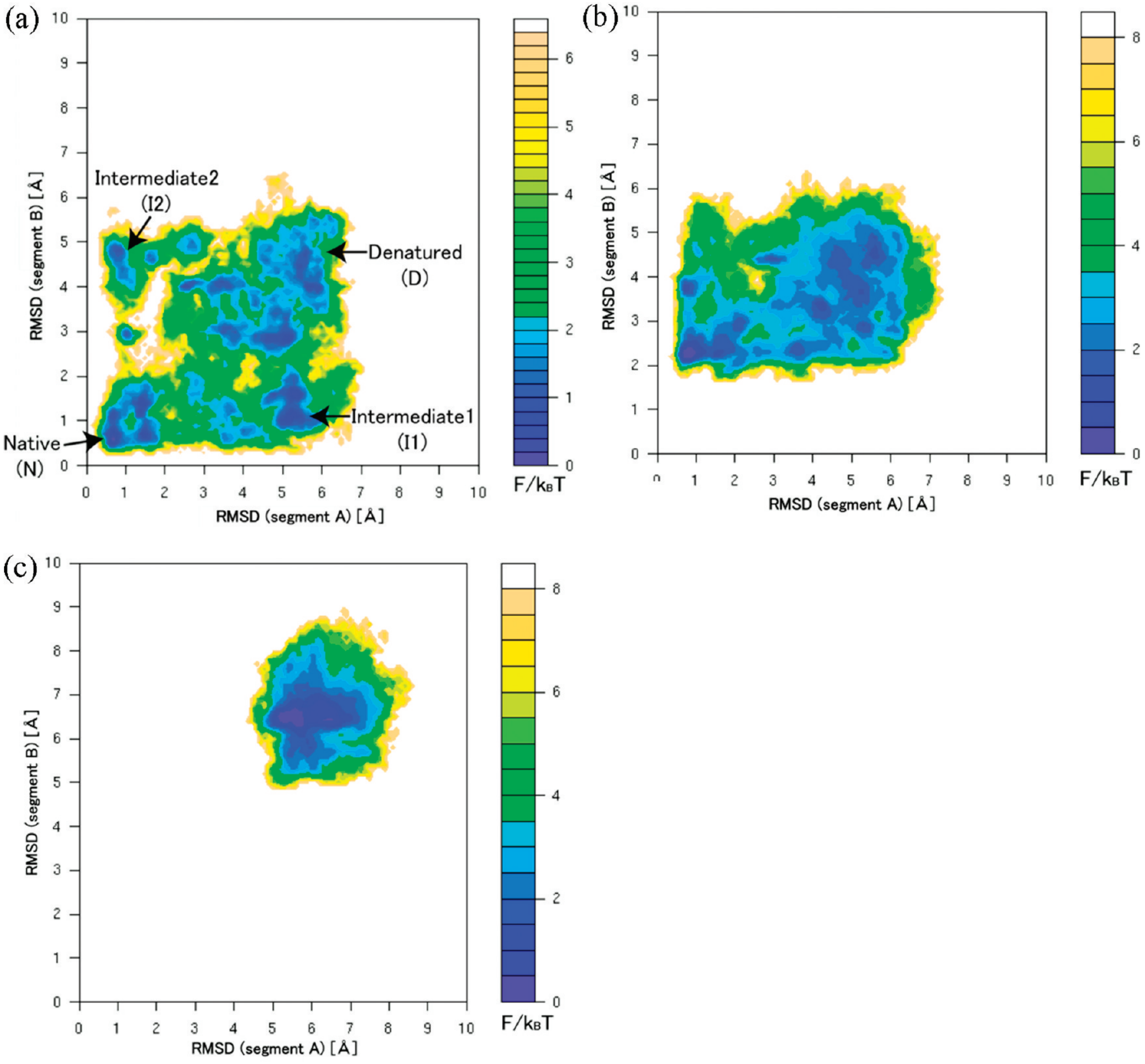

Figure 7. (a) The folding FEL of HP35 projected onto the subspace spanned by the C_α-RMSD of segment A and segment B calculated using the MSFEL method. (b) Folding FEL of the P21A and (c) F6A/F10A/F17A mutants.

formed (R_B < 2.0 Å). Figure 8c shows the folding FEL in the second stage of folding and indicates that the formation of segment A is dependent upon formation of the hydrophobic native contacts. The calculated free energy difference during the folding process (~4 k_BT) was slightly smaller than the experimentally derived value (~5 k_BT).[40] This is a better agreement than previous work with implicit solvent.[13]

In the minor folding pathway, the order of segment formation is reversed; the formation of segment A (D→I2) precedes the formation of segment B (D→I1). The trigger for the minor pathway is contact of the hydrophobic residues to form the hydrophobic core, not the formation of the motif as in the major pathway. Figure 8d shows the FEL of the first folding stage in the minor folding pathway when segment B was not formed (R_B > 2.0 Å). After the formation of segment A (R_A < 2.0 Å), segment B is formed through the formation of the motif in the same order (helix III→helix II) as shown in Figure 8e,f. Graphical summary representations of the major and minor folding pathways are shown in Figure 10a,b. This minor folding pathway is consistent with the recent result of triplet−triplet-energy transfer (TTET) experiment.[64] The high-free-energy intermediate found to be accessible from the N state is expected to correspond to I2 in the MSFEL. The experimental activation barrier between D and I2 was reported to be ~7 k_BT,[64] which is comparable to our calculation ~6 k_BT. The slight underestimation may be due to the choice of the force field. It has been reported that the AMBER ff03 force field had higher helical stability than in experiments. Folding enthalpies were also less than half of the experimental value.[65] Furthermore, it has been pointed that the ff03 favors a helical unfolded state and a diffusion-collision-type folding mechanism.

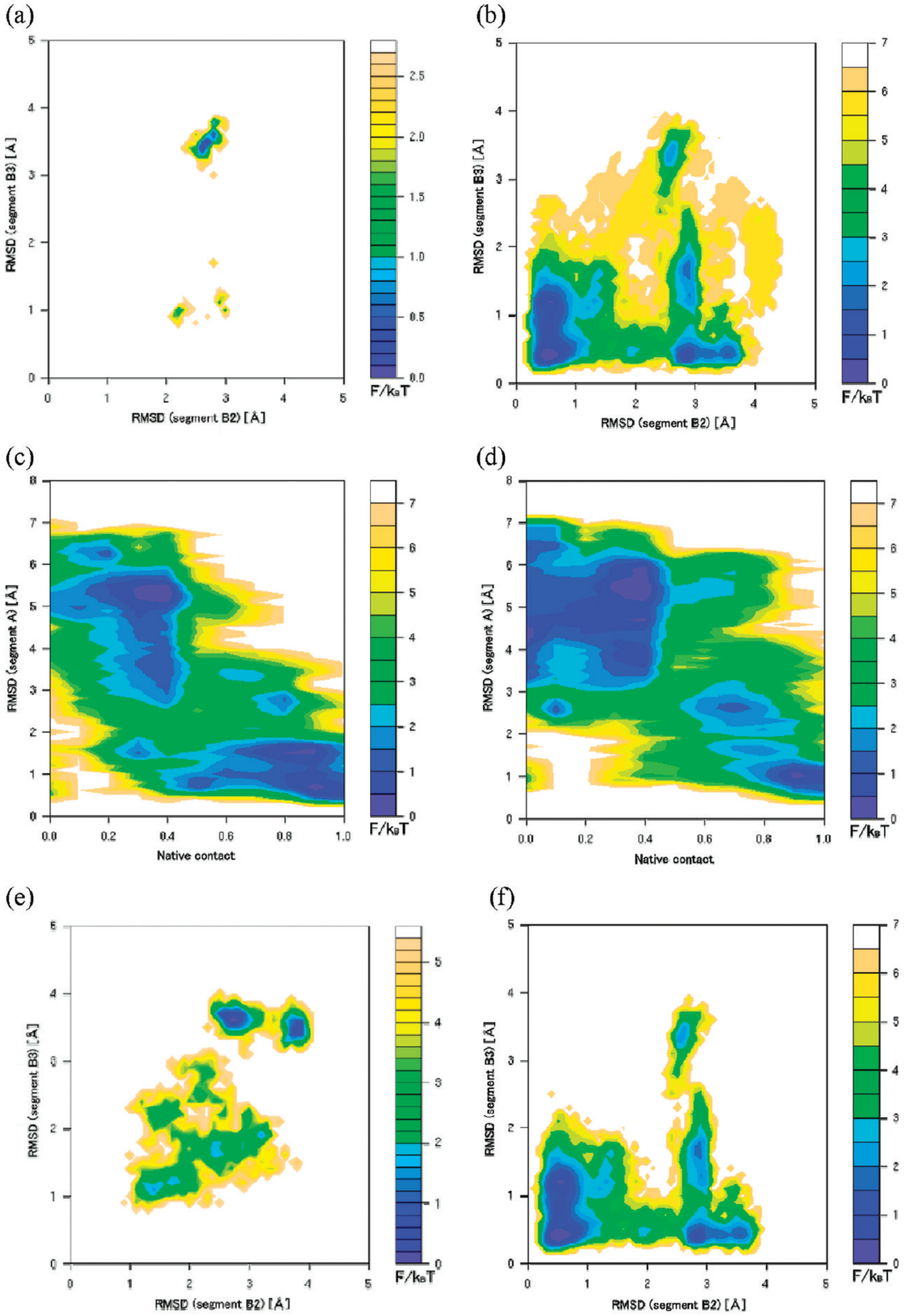

Figure 8. (a, b) The folding FEL of HP35 in the first stage of the major pathway (D→I1, $R_A > 2.0$ Å) projected onto the two-dimensional subspace spanned by the C_α-RMSD of segments B2 and B3, (a) without considering the formation of the PLWK motif and (b) considering formation of the PLWK motif. (c) The folding FEL of HP35 in the second stage of the major pathway (I1→N, $R_A < 2.0$ Å) projected onto the two-dimensional subspace spanned by fractions of the native contact between the hydrophobic residues (Phe6, Phe10, and Phe17) and the C_α-RMSD of segment A. (d) The folding FEL of HP35 in the first stage of the minor pathway (D→I2, $R_B > 2.0$ Å) projected onto the same reaction coordinates as for c. The folding FEL of the second stage in the minor pathway (D→I2, $R_A < 2.0$ Å) (e) when the motif was not formed and (f) when the motif was formed.

Individual helices were rather stable in isolation and dock together to form the folded state. Actually, this interpretation agrees with the folding pathways shown as Figure 10a,b; each segment docks together into the native conformation when each helix is almost folded. Therefore, the stability of helical structures with the ff03 force field may lead the underestimation of the D to I2 free energy difference.

The folding pathways observed in this work are also compared to the result of all-atom unbiased 100 μs MD simulations[66] in which the same force field (AMBER ff03) was employed.

The folding free energy barrier height along the 1D major pathway at 300 K derived from Figure 9b,c corresponds to ~1.2 kcal/mol (~2.0 k_BT), which is 6.0 times higher than the value at 390 K (~0.2 kcal/mol).[66] This difference will be originated from the temperature difference because the frequent observation of the folding−unfolding transitions at 390 K should be caused by the reduction of the free energy barrier. The order of helix formation on the major folding pathway derived from Figure 8 showed good agreement with that of the all-atom MD simulation in the following points: (1) helix 1 is relatively unstable in the unfolded sate; (2) helices 2 and 3 form during the early stage of the folding process; and (3) helix 1 is nearly always the last to form.

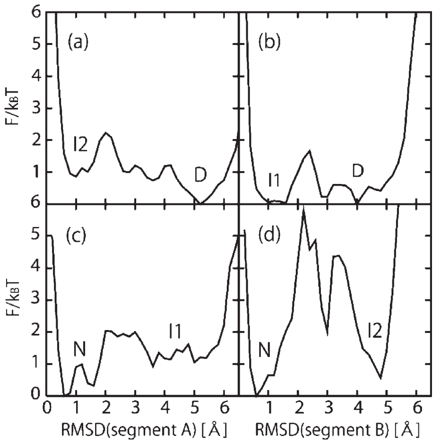

Figure 9. (a,b) The one-dimensional free energy landscapes projected onto the C_α-RMSD of segment A and B on the first folding processes of the minor and major folding pathways and (c,d) those of the second folding processes of the major and minor folding pathways.

3.3. Mutation in the PLWK Motif. The importance of Pro21 in the PLWK motif has been shown by the point mutation experiments.[50,51] The point mutation P21A had a dramatic effect on the folding of HP35. It has also been suggested that Pro21 may responsible for critical interactions for folding into the native structure. To examine the effect of this mutation in the PLWK motif from a point of view of FEL, we constructed a mutant (P21A) and calculated its folding FEL (Figure 7b). The formation of segment B was clearly suppressed in this mutant, indicating that the rigid Pro21 in the linker region between helices II and III plays an important role in stabilizing segment B. This mutation only suppressed the formation of segment B; therefore, Pro21 is essential for controlling the major folding pathway through its influence on the formation of the PLWK motif. This result was in good agreement with previous NMR and CD experimental and computational results of the point mutation.[50]

3.4. Mutation in the Hydrophobic Core. To address the effect of mutations in the hydrophobic core shown in Figure 1, we constructed a triple mutant, F6A/F10A/F17A. Figure 7c

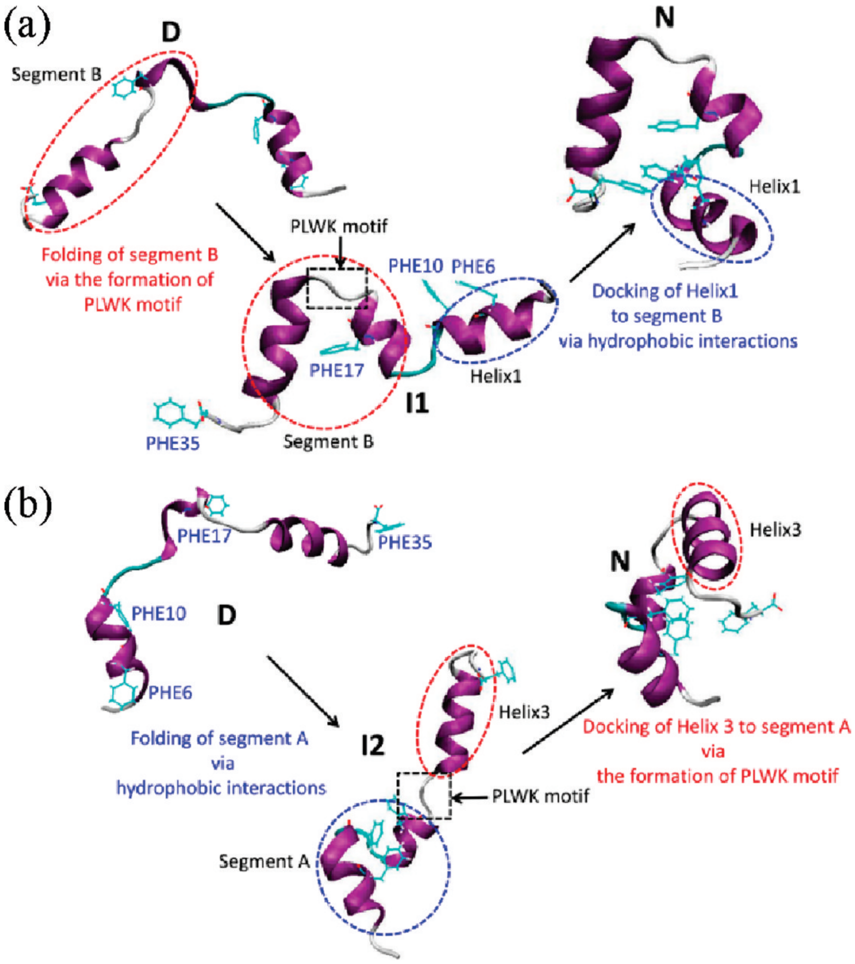
(a)
D
Segment B
Folding of segment B via the formation of PLWK motif
PLWK motif
PHE10 PHE6
PHE17
Helix1
I1
PHE35
Segment B
N
Helix1
Docking of Helix1 to segment B via hydrophobic interactions
(b)
PHE17
PHE35
PHE10
D
PHE6
Folding of segment A via hydrophobic interactions
I2
Helix3
PLWK motif
Segment A
N
Helix3
Docking of Helix 3 to segment A via the formation of PLWK motif

Figure 10. Schematic representations of (a) the major and (b) minor folding pathways. Figure created using VMD.[67]

shows the folding FEL of the mutant. The mutations are expected to suppress the formation of segment A as the formation of the hydrophobic core is the driving force behind the formation of segment A in the minor folding pathway. The folding FEL shown in Figure 7c indicated that there was suppression of the formation of both segments A and B. These observations agree well with the previously reported results of mutagenesis experiments,[42] suggesting that these three phenylalanine residues play crucial roles in stabilizing the native structure and in the folding of HP35.

4. CONCLUSION

In the present work, we studied the folding process of the fast-folding mini-protein HP35 by investigating the folding FEL in explicit solvent using the MSFEL method. A previously unreported minor folding pathway in the order of D→I2→N was identified in this work in addition to the major folding pathway, D→I1→N, described previously.[12,13] In the minor pathway, the driving force behind folding is the formation of the hydrophobic core (residues F6, F10, and F17) located at the center of the native structure, while the formation of the PLWK motif (residues P21−K24) is considered to be the trigger in the major pathway. Mutations in the PLWK motif (P21A) and hydrophobic core showed that these residues play important roles in the folding of HP35. The P21A mutation partially suppressed folding, especially the formation of helices II and III, while mutations in the hydrophobic core completely prevented the overall folding of HP35.

■ AUTHOR INFORMATION

Corresponding Author

*Phone/Fax: +81-3-5841-2297. E-mail: kitao@iam.u-tokyo.ac.jp.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENT

This work was supported by the Next Generation Super Computing Project, Nanoscience Program, the Strategic Programs for Innovative Research (SPIRE), and Computational Material Science Initiative (CMSI), Grants-in-Aid for Science Research (B), and Grants-in-Aid for Science Research in Priority Areas from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan to A.K. The computations were performed in part using supercomputers at the Research Center for Computational Science, Okazaki Research Facilities, National Institute of Natural Science.

■ REFERENCES

(1) Bevivino, A. E.; Loll, P. J. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98*, 11955–11960.

(2) Cooper, J. K.; Schilling, G.; Peters, M. F.; Herring, W. J.; Sharp, A. H.; Kaminsky, Z.; Masone, J.; Khan, F. A.; Delanoy, M.; Borchelt, D. R.; Dawson, V. L.; Dawson, T. M.; Ross, C. A. *Hum. Mol. Genet.* **1998**, *7*, 783–790.

(3) Georgalis, Y.; Starikov, E. B.; Hollenbach, B.; Lurz, R.; Scherzinger, E.; Saenger, W.; Lehrach, H.; Wanker, E. E. *Proc. Natl. Acad. Sci. U. S. A.* **1998**, *95*, 6118–6121.

(4) Scherzinger, E.; Lurz, R.; Turmaine, M.; Mangiarini, L.; Hollenbach, B.; Hasenbank, R.; Bates, G. P.; Davies, S. W.; Lehrach, H.; Wanker, E. E. *Cell* **1997**, *90*, 549–558.

(5) Singer, S. J.; Dewji, N. N. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 1546–1550.

(6) Satoh, D.; Shimizu, K.; Nakamura, S.; Terada, T. *FEBS Lett.* **2006**, *580*, 3422–3426.

(7) Suenaga, A.; Narumi, T.; Futatsugi, N.; Yanai, R.; Ohno, Y.; Okimoto, N.; Taiji, M. *Chem.—Asian J.* **2007**, *2*, 591–598.

(8) Juraszek, J.; Bolhuis, P. G. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 15859–15864.

(9) Zhou, R. H. *Proteins: Struc., Funct., Genet.* **2003**, *53*, 148–161.

(10) Duan, Y.; Kollman, P. A. *Science* **1998**, *282*, 740–744.

(11) Freddolino, P. L.; Schulten, K. *Biophys. J.* **2009**, *97*, 2338–2347.

(12) Lei, H. X.; Duan, Y. *J. Mol. Biol.* **2007**, *370*, 196–206.

(13) Lei, H. X.; Wu, C.; Liu, H. G.; Duan, Y. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 4925–4930.

(14) Nakajima, N.; Nakamura, H.; Kidera, A. *J. Phys. Chem. B* **1997**, *101*, 817–824.

(15) Hansmann, U. H. E.; Okamoto, Y.; Eisenmenger, F. *Chem. Phys. Lett.* **1996**, *259*, 321–330.

(16) Hukushima, K.; Nemoto, K. *J. Phys. Soc. Jpn.* **1996**, *65*, 1604–1608.

(17) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141–151.

(18) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *J. Chem. Phys.* **1953**, *21*, 1087–1092.

(19) Sugita, Y.; Kitao, A.; Okamoto, Y. *J. Chem. Phys.* **2000**, *113*, 6042–6051.

(20) Bolhuis, P. G.; Juraszek, J. *Biophys. J.* **2010**, *98*, 646–656.

(21) Laio, A.; Gervasio, F. L. *Rep. Prog. Phys.* **2008**, 71.

(22) Wang, F. G.; Landau, D. P. *Phys. Rev. Lett.* **2001**, *86*, 2050–2053.

(23) Dellago, C.; Grunwald, M. *J. Chem. Phys.* **2007**, 127.

(24) Garcia, A. E.; Onuchic, J. N. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100*, 13898–13903.

(25) Gnanakaran, S.; Nymeyer, H.; Portman, J.; Sanbonmatsu, K. Y.; Garcia, A. E. *Curr. Opin. Struct. Biol.* **2003**, *13*, 168–174.

(26) Jang, S.; Kim, E.; Pak, Y. *J. Chem. Phys.* **2008**, *128*, 105102.

(27) Zhang, J.; Qin, M.; Wang, W. *Proteins: Struct., Funct., Bioinf.* **2006**, *62*, 672–685.

(28) Kitao, A. *J. Chem. Phys.* **2011**, *135*, 045101.

(29) Harada, R.; Kitao, A. *Chem. Phys. Lett.* **2011**, *503*, 145–152.

(30) Torrie, G. M.; Valleau, J. P. *Chem. Phys. Lett.* **1974**, *28*, 578–581.

(31) Torrie, G. M.; Valleau, J. P. *J. Comput. Phys.* **1977**, *23*, 187–199.

(32) Ferrenberg, A. M.; Swendsen, R. H. *Phys. Rev. Lett.* **1989**, *63*, 1195–1198.

(33) Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. *J. Comput. Chem.* **1992**, *13*, 1011–1021.

(34) Souaille, M.; Roux, B. *Comput. Phys. Commun.* **2001**, *135*, 40–57.

(35) Harada, R.; Kitao, A. *J. Phys. Chem. B* **2011**, *115*, 8806–8812.

(36) Tang, Y. F.; Grey, M. J.; McKnight, J.; Palmer, A. G.; Raleigh, D. P. *J. Mol. Biol.* **2006**, *355*, 1066–1077.

(37) Vardar, D.; Chishti, A. H.; Frank, B. S.; Luna, E. J.; Noegel, A. A.; Oh, S. W.; Schleicher, M.; McKnight, C. J. *Cell Motil. Cytoskeleton* **2002**, *52*, 9–21.

(38) Chiu, T. K.; Kubelka, J.; Herbst-Irmer, R.; Eaton, W. A.; Hofrichter, J.; Davies, D. R. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 7517–7522.

(39) McKnight, C. J.; Matsudaira, P. T.; Kim, P. S. *Nat. Struct. Biol.* **1997**, *4*, 180–184.

(40) Kubelka, J.; Eaton, W. A.; Hofrichter, J. *J. Mol. Biol.* **2003**, *329*, 625–630.

(41) Kubelka, J.; Chiu, T. K.; Davies, D. R.; Eaton, W. A.; Hofrichter, J. *J. Mol. Biol.* **2006**, *359*, 546–553.

(42) Frank, B. S.; Vardar, D.; Buckley, D. A.; McKnight, C. J. *Protein Sci.* **2002**, *11*, 680–687.

(43) Lei, H.; Deng, X.; Wang, Z.; Duan, Y. *J. Chem. Phys.* **2008**, *129*, 155104.

(44) Piana, S.; Laio, A.; Marinelli, F.; Van Troys, M.; Bourry, D.; Ampe, C.; Martins, J. C. *J. Mol. Biol.* **2008**, *375*, 460–470.

(45) Raleigh, D. P.; Wickstrom, L.; Okur, A.; Song, K.; Hornak, V.; Simmerling, C. L. *J. Mol. Biol.* **2006**, *360*, 1094–1107.

(46) Yoda, T.; Sugita, Y.; Okamoto, Y. *Biophys. J.* **2010**, *99*, 1637–1644.

(47) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y. B.; Wriggers, W. *Science* **2010**, *330*, 341–346.

(48) Schulten, K.; Freddolino, P. L. *Biophys. J.* **2009**, *97*, 2338–2347.

(49) Kollman, P. A.; Duan, Y. *Science* **1998**, *282*, 740–744.

(50) Vermeulen, W.; Van Troys, M.; Bourry, D.; Dewitte, D.; Rossenu, S.; Goethals, M.; Borremans, F. A. M.; Vandekerckhove, J.; Martins, J. C.; Ampe, C. *J. Mol. Biol.* **2006**, *359*, 1277–1292.

(51) Raleigh, D. P.; Xiao, S. F. *J. Mol. Biol.* **2010**, *401*, 274–285.

(52) Brooks, C. L., 3rd *Curr. Opin. Struct. Biol.* **1998**, *8*, 222–226.

(53) Go, N. *Annu. Rev. Biophys. Bioeng.* **1983**, *12*, 183–210.

(54) Mirny, L.; Shakhnovich, E. *Annu. Rev. Biophys. Biomol. Struct.* **2001**, *30*, 361–396.

(55) Martyna, G. J.; Klein, M. L.; Tuckerman, M. *J. Chem. Phys.* **1992**, *97*, 2635–2643.

(56) Gront, D.; Kmiecik, S.; Kolinski, A. *J. Comput. Chem.* **2007**, *28*, 1593–1597.

(57) Canutescu, A. A.; Shelenkov, A. A.; Dunbrack, R. L. *Protein Sci.* **2003**, *12*, 2001–2014.

(58) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.

(59) Case, D. A.; Darden, T. A.; Cheatham, T. E., III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Pearlman, D. A.; Crowley, M.; Walker, R. C.; Zhang, W.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Wong, K. F.; Paesani, F.; Wu, X.; Brozell, S.; Tsui, V.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Mathews, D. H.; Schafmeister, C.; Ross, W. S.; Kollman, P. A. *AMBER 9*; University of California, San Francisco, 2006.

(60) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G. M.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J. M.; Kollman, P. *J. Comput. Chem.* **2003**, *24*, 1999–2012.

(61) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327–341.

(62) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577–8593.

(63) Frishman, D.; Argos, P. *Proteins: Struct., Funct., Genet.* **1995**, *23*, 566–579.

(64) Kiefhaber, T.; Reiner, A.; Henklein, P. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 4955–4960.

(65) Shaw, D. E.; Piana, S.; Lindorff-Larsen, K. *Biophys. J.* **2011**, *100*, L47–L49.

(66) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. *Biophys. J.* **2011**, *100*, L47–L49.

(67) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33–38.