Supporting Information for

Multi-scale ensemble modeling of modular proteins with intrinsically disordered linker regions: Application to p53

Tsuyoshi Terakawa[1], Junichi Higo[2], and Shoji Takada[3]

[1]Department of Biophysics, Graduate School of Science, Kyoto University, Kitashirakawa-Oiwakecho, Sakyo, Kyoto, 6068502, Japan; [2]Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka, 5650871, Japan

Overview

In the following subsections, we provided 1) the set up of the all atom (AA) molecular dynamics (MD) simulation in the "All atom simulation of p53 linker region" subsection, 2) the detailed procedure of the virtual-system coupled multi-canonical MD (VMcMD) simulation in the "Details for the V-McMD method" and "Multicanonical force and energy distribution" subsection, 3) the potential energy function for the coarse-grained (CG) simulation in the " Coarse-grained simulation of p53 linker region" subsection, and 4) CG model parameter calibration procedure for inter-core domain interaction in the "Coarse-grained simulation of two core domains" subsection.

All atom simulation of p53 linker region

Here we describe the AA MD simulation method for the p53 linker. The system consists of the p53 linker segment with a few residue extensions in both ends (a 40-residue long, Residue ID: 288-327), which is solvated with water molecules. The amino-acid sequence is: Ace-NLRKKGEPHHELPPGSTKRALPNNTSSSPQPKKKPLDGET-Nme, where Ace and Nme are, respectively, the N-terminal acetyl and C-terminal N-methyl groups introduced to cap the segment termini.

We generated a random conformation of the linker segment for the initial conformation, and put it into a sphere (sphere 1; diameter = 82 Å), setting the center of mass at the center of the sphere 1. The water buffer had been equilibrated in advance at 1.0 g/cc and 300 K. Then, we randomly replaced 66 water molecules with 36 chlorine and 30 sodium ions to realize physiological ion concentration. The mismatch of the positive and negative ions neutralized the net charge of the linker. The system finally consisted of 30937 atoms (640 polypeptide atoms, 36 Cl^-, 30 Na^+, and 10077 water molecules). To avoid evaporation of the solvent from the sphere 1, a restoring force (harmonic potential) was applied to water-oxygen atoms or ions only when they were outside of the sphere 1. Another harmonic potential was applied to the linker heavy atoms when those atoms were outside of a smaller sphere (sphere2; diameter = 80 Å) concentric to the sphere 1. The sphere 2 was smaller than the sphere 1 because the linker should not be exposed to the sphere 1 surface. We fixed the linear and the angular momenta to zero by re-scaling velocities. The momentum and the angular momentum of the linker were fixed to zero during simulation. We did not use the periodic boundary condition in this study because the periodicity may cause artificially inter-chain entangling among the different periodic boxes. The solvent sphere (sphere 1) was set as large as possible, yet small enough so that the multi-canonical sampling can be done within a feasible simulation time.

We used PRESTO ver. 3 (Morikami et al., *Biopolymers Computers Chem*, 16:243, 1992) with which we implemented V-McMD (Higo et al., *J Chem Phys*, 138:184106, 2013). For time integration, we used the leap flog method (Hockney and Eastwood, Computer Simulation using particles, 1994). The MD time step was 1.0 fs. SHAKE (Ryckaert et al., *J Comput Phys*, 23:327, 1977) was used to constrain the covalent bonds between heavy atoms and hydrogen atoms. Long-range electrostatic interactions were calculated using a cell-multipole expansion (Ding et al., *J Chem Phys*, 97:4309, 1992). One of the advantages of the cell-multipole method is that we can apply it irrespective of the boundary condition. The temperature was controlled using a constant-temperature method (Schmidt et al., *J Phys Chem B*, 113:11959, 2009). The force field parameters for the polypeptides were from an AMBER-based hybrid force field (Kamiya et al., *Chem Phys Lett*, 401:312, 2005) defined as $V_{hybrid} = 0.25V_{94} + 0.75V_{96}$, where V_{94} and V_{96} respectively denote the AMBER parm94 (Cornell et al., *J Am Chem Soc*, 118:2309, 1995) and parm96 force fields (Kollman et al., Computer Simulation of Biomolecular Systems, 1997) Previous McMD simulations with Vhybrid revealed that a peptide with a helical propensity folds

into an α-helix, whereas a peptide with a β-hairpin propensity forms a β-hairpin (Kamiya et al., *Chem Phys Lett*, 401:312, 2005). Therefore, we used Vhybrid for the current study. We have successfully applied this force field to protein folding (Ikebe et al., *Chem Phys Lett,* 443:364, 2007; Ikebe et al. *Protein Sci,* 20:187, 2011; Ikebe et al., *J Comput Chem,* 32:1286, 2011) and an ensemble modeling of an IDP (Higo et al., *J Am Chem Soc,* 133:10448, 2011). Although there is no perfect atomistic force filed that can be applicable to any amino-acid sequence, our preceding works (Higo et al., *J Am Chem Soc,* 133:10448, 2011; Kamiya et al., *Chem Phys Lett*, 401:312, 2005) have suggested that the currently used force field does not have an apparent bias in secondary structure formation and is appropriate for IDR study. We used the TIP3P water model (Jorgensen et al., *J Chem Phys,* 79:926, 1983) for the water molecules.

The AA simulation procedure consists of two stages (For detail of the method, see the supporting information): the pre-V-McMD stage where 128 canonical MD runs were done at various temperatures, and the V-McMD stage, where 128 McMD runs were done. The first 128 pre-V-McMD simulations were performed with a high-temperature (719 K) for for each of the 128 runs starting from the random conformation generated above with different random seeds for the atomic velocity generation. Then, the second 128 pre-V-McMD simulations were performed at 671 K starting from the last snapshots of the first pre-V-McMD simulation. We repeated this procedure with decreasing temperatures to 296 K. After the pre-V- McMD simulations, the biased potential was computed for the first V-McMD simulation. Then, we started the first V-McMD simulations using the biased potential, where 128 runs were done independently starting from the first 128 pre-V-McMD simulations at 719 K. We repeated the V-McMD simulations for 16 times, where the iterations from the first to fifteenth V-McMD simulations were performed for the refinement of the biasing potential for the conformational sampling and the last iteration was the production run. The initial conformations of the 128 runs for the *i* th V-McMD simulation were the last snapshots of those for the *i*+1 th V-McMD simulation. The simulation length for the first to fifteenth V-McMD simulations ranged from 1.0×10^6 to 2.6×10^6 steps. Length of the production run was 1.2×10^7 steps for each of the 128 runs. Finally, we assigned a statistical weight at 300 K to each snapshot of the production run according to the re-weighting technique (Higo et al., *J Chem Phys,* 138:184106, 2013). We note that the V-McMD simulation is a generalized ensemble method, which is designed to obtain a wide conformational distribution by performing such short production simulation. We calculated

the root mean square deviation (RMSD) between the initial structure of the production V-McMD run and each snapshot in the trajectories, and plotted it for four representative cases in Fig. S4. From these plots, we see that, right after the beginning of simulations, the conformation rapidly changed drastically suggesting that it is unlikely that the initial conformation affected the sampling.

Details for the V-McMD method

Here we describe the methodology for the V-McMD method. More details are given in the paper (Higo et al., *J Chem Phys* **138**, 184106, 2013). In the pre-V-McMD stage, temperature T decreased as 629 K, 559 K, 503 K, 457 K, 419 K, 387 K, 359 K, 335 K, 315 K, and 296 K, where the inversed temperature T^{-1} was changed with the same interval: $\Delta T^{-1} = 0.2$. The pre-V-McMD stage covered an energy range of [–102300.0 kcal/mol, –69300.0 kcal/mol]. This energy range is called the entire energy range.

In the V-McMD stage, the entire energy range was divided into some energy zones (see Table S1), whose energy ranges are listed in Table S2. The number of energy partitioning decreased as proceeding with the V-McMD iterations in accordance with the original V-McMD method. The introduction of the zones is rationalized theoretically assuming that a virtual system interacts with the molecular system to be studied. Each of zones is assigned to a discrete state (i.e., the virtual state) of the virtual system.

In a V-McMD run, the molecular system confined in a virtual state (i.e., a zone) for a given period of simulation, and the molecular system moves to another virtual state at the end of the period. The virtual-state move is achieved with satisfying the detailed balance condition. An advantage of the V-McMD algorithm is: one can control arbitrarily the inter-virtual-state transition probability by setting the density of states for the virtual system.

A benefit of multicanonical sampling is that a canonical energy distribution at 300 K is derived from the sampling: $P(E,T)$ where E is the energy of a conformation and T = 300 K. The production run of the V-McMD sampling produced an ensemble of conformations, which have various energies. One can construct a canonical conformational distribution by assigning the statistical weight to the sampled conformations: i.e., the statistical weight of a conformation, whose energy is E', is $P(E',T)$. Those weighted conformations are used for analyses.

Multicanonical force and energy distribution

In the V-McMD simulation (Higo et al., *J Chem Phys* **138**, 184106, 2013), the force $\boldsymbol{f}_{mc}(r_i)$ acting on the atom i is given by

$$\boldsymbol{f}_{mc}(r_i) = RT\frac{d\ln[n(E)]}{dE}\boldsymbol{f}(\boldsymbol{r}_i) \tag{S3}$$

where R is gas constant, T is temperature, and $\boldsymbol{f}(\boldsymbol{r}_i)$ is force acting on atom i based on the potential energy E. $n(E)$ is the density of states of the system, which we do not know *a priori*. In the actual process of McMD, instead of the $n(E)$, we obtain $d\ln[n(E)]/dE$ by the iterative simulations described above. We plotted $d\ln[n(E)]/dE$ obtained in the current work against E in Fig. S5.

If we accurately estimate $d\ln[n(E)]/dE$ and perform AA simulation using $\boldsymbol{f}_{mc}(r_i)$ in eq. S3 instead of $\boldsymbol{f}(\boldsymbol{r}_i)$, ideally, we obtain a flat distribution of the potential energy (i.e., $P(E) \approx const$). Therefore, the flat distribution of the potential energy indicates the accurate estimation of $d\ln[n(E)]/dE$. We plotted the distribution of the potential energy in Fig. S6. From this figure, we can see fairly good flatness, showing the accurate estimation of $d\ln[n(E)]/dE$ and efficient conformational sampling. From this $d\ln[n(E)]/dE$, we can obtain canonical energy distributions $P_c(E,T)$, at an arbitral temperature T as $P_c(E,T) \propto n(E)exp[-E/RT]$. We plotted the canonical energy distributions at 300 K and at 700 K in Fig. S6.

Coarse-grained simulation of p53 linker region

As a starting point of development of a new CG model, we began with a concise CG model that we developed previously (Terakawa et al., *Biophys J,* 101:1450, 2011). This model does not take into account long-range contacts. The potential energy function of that model is

$$V_0 = V_{without\ contact} = V_{bond} + V_{angle} + V_{ele} + V_{ex} \tag{S3}$$

where V_{bond}, V_{ele}, and V_{ex} are the bond stretching term, the electrostatics term, and the excluded volume effect term, respectively. V_{bond} is the potential energy for bond stretching and is defined as

$$V_{bond} = k_b\left(r_{ij} - b\right)^2 \tag{S4}$$

where parameters were set as k_b = 110.4 (kcal/mol·Å^2) and b = 3.8 (Å). r_{ij} is the length of

the virtual bond. V_{angle} is the potential energy for two kinds of angles and is defined as

$$V_{angle} = -k_B T \ln \frac{P(\theta)}{\sin\theta} - k_B T \ln P(\eta) \quad \text{(S5)}$$

where θ (η) were virtual bond (dihedral) angles that were defined by the coordinates of three (four) consecutive CG particles. $P(\theta)$ ($P(\eta)$) was the probability distribution of θ (η) in loop regions of PDB structures. V_{ele} is the potential energy for electrostatics and is defined as

$$V_{ele} = \sum_{i<j} \frac{q_i q_j}{4\pi\epsilon_0\epsilon_k r_{ij}} \exp\left(-\frac{r_{ij}}{\kappa_D}\right) \quad \text{(S6)}$$

where q_i is charge (charge is set as +1 for Lys, Arg, and His and -1 for Asp and Glu), ϵ_0 is the dielectric constant of vacuum, ϵ_k = 78.0 is the dielectric constant, and κ_D is the Debye length defined as

$$\kappa_D = \left(\frac{\epsilon_0\epsilon_k k_B T}{2N_A e^2 I}\right)^{\frac{1}{2}} \quad \text{(S7)}$$

where k_B is the Boltzmann constant, T = 300.0 (K) is temperature, N_A is Avogadro number, e is the elementary electric charge, and I is the ionic strength. The conformation of the IDRs changes dynamically. Accordingly, it is supposed that the protonation state of histidines continually changes. Ideally, it is desired to calculate pKa of the histidine in each MD time step calculation to decide the protonation state of histidines. However, the pKa calculation method is not established for CG protein model. Thus, in the current work, we performed the AA and CG MD simulation based on the assumption that histidine is always protonated. V_{ex} is the excluded volume potential and is defined as

$$V_{ex} = \sum_{i<j-3}^{non-native} \epsilon_{ex} \left(\frac{C}{r_{ij}}\right)^{12} \quad \text{(S8)}$$

where ϵ_{ex} = 0.2 (kcal/mol) and C = 4.0 (Å) are constant parameters. This model reproduced the SAXS profile of the p53 N-terminal IDR whose conformational ensemble did not have extensive long-range contacts. However, the direct application to the system with fractional long-range contacts fails to reproduce the SAXS profile, as is shown below.

We used CafeMol 2.0 (Kenzaki et al., *J Chem Theory Comput,* 7:1979, 2011) for all the CG MD simulations in this work. Production runs for the CG simulations were performed by

Langevin dynamics for 10^8 MD steps with friction coefficient of 0.02 and with temperature of 300 K.

Coarse-grained simulation of two core domains

Experimentally, it has been revealed that two p53 core domains form a loose dimer with the dissociation constant of 2 mM at 100 mM monovalent ion (Rippin et al., *J Mol Biol*, 319:351, 2002). Using NMR spectroscopy, Tidow et al. revealed that transient interaction between core domains in solution involved the same interface as that observed in the crystal structure of the core domain–DNA complex (Tidow et al., *Proc Natl Acad Sci USA,* 104:12324, 2007). To model this inter-core-domain interaction so that the dissociation constant was essentially the same as that measured in the previous experiment, we performed the CG MD simulation of the system containing the two core domains (Fig. S1A). The initial coordinate of the core domain was taken from the X-ray crystal structure (Natan et al., *J Mol Biol,* 409:358, 2011) (PDB ID: 2XWR). We put two core domains into a sphere with the diameter of 50 Å. We used the one-bead-per-one-amino-acid CG model. We adopted recently developed state of the art Go-like AICG2 model (Li et al., *Proc Natl Acad Sci USA*, 109:17789, 2012) for the intra-molecular potential energy function that stabilizes the native structure (Natan et al., *J Mol Biol*, 409:358, 2011) (PDB ID: 2XWR). The inter-core-domain potential energy function was defined as

$$V_{inter\_core} = V_{ele} + V_{ex} + \sum_{i>j+3}^{native\ contact} \epsilon\epsilon_{ij}\left[5\left(\frac{r_{ij}^0}{r_{ij}}\right)^{12} - 6\left(\frac{r_{ij}^0}{r_{ij}}\right)^{10}\right] \quad \text{(S9)}$$

where V_{ele} and V_{ex} were electrostatics term and excluded volume effect term, respectively (see above for complete description of these terms. i and j run over the CG particle pairs that contacted in the experimentally indicated interface in the X-ray crystal structure in which the four core domains specifically bound to the recognition element (Chen et al., *Structure*, 18:246, 2010) (PDB ID: 3KMD). We considered that two CG particles contacted if one of the heavy atoms represented by one CG particle was within 6.5 Å from that represented by the other particle. The r_{ij}^0 was the distance between two CG particles i and j in the native structure. The ϵ_{ij}s are the AICG2 model parameters (Li et al., *Proc Natl Acad Sci USA*, 109:17789, 2012). These parameters were tuned so that the fluctuation of isolated proteins was reproduced. Thus, there is no guarantee that these parameters reproduce the strength of inter-protein-interaction.

The ϵ_{ij}s were the AICG2 model parameters (Li et al., *Proc Natl Acad Sci USA*, 109:17789, 2012) calibrated for intra-molecular interaction. Accordingly, to reproduce the dissociation constant, we scaled the inter-molecular native contact interaction by an additional factor ϵ. The ion strength was set to the same value as that of the experiment (100 mM) (Rippin et al., *J Mol Biol,* 319:351, 2002). The ϵ in the equation above was varied from 0.1 to 1.0 with a step of 0.1 and from 0.6 to 0.7 with a step of 0.01.

Each production run was performed by Langevin dynamics for 10^9 MD steps with friction (damping) coefficient of 0.02 and with temperature of 300 K. For time integration, we used a simple algorithm developed by Honneycutt and Thirumalai (Honeycutt and Thirumalai, *Biopolymers,* 32:695, 1992) to solve an approximated Langevin equation. More sophisticated algorithm was proposed by Paterlini and Ferguson (Paterlini and Ferguson, Chemical Physics, 236:243, 1998) to solve the generalized Langevin equation. The integration time step was 0.1. The friction force was uniformly and independently applied to all the CG beads.

In Fig. S1B, we illustrate a time trajectory of the Q-score of inter-molecular contacts in the case where the ϵ is set to 0.65. The Q-score represents the ratio of the transiently formed contacts to the contacts formed in the reference crystal structure (Chen et al., *Structure*, 18:246, 2010). With this interaction strength, the core domains repeatedly associate (Q-score is around 1.0) and dissociate (Q-score is around 0.0) each other. The probability distribution of the interaction energy, shown in Fig. S1C, is composed of a broad peak around -13.0 kcal/mol and a sharp peak around 0.0 kcal/mol, which correspond to the bound and unbound states, respectively. This bimodal distribution allows us to set a threshold (-3.0 kcal/mol) between these two states and to calculate the fraction of the bound state (f_b) (Ganguly et al., *Proteins*, 79:1251, 2011; Okazaki et al., *J Am Chem Soc*, 134:8918, 2012). Using this f_b, we can estimate the K_d by the equation,

$$K_d = \frac{2C(1-f_b)^2}{f_b} \tag{S10}$$

where C is the concentration of the core domains (6.3 mM based on the radius of the sphere). We plot the calculated K_ds against the ϵs in Fig. S1D. From this figure, we can see that, when the ϵ is set to 0.65, the order of magnitude of the dissociation constant agrees with the experimentally measured dissociation constant (red horizontal line in Fig. S1D). Therefore, we utilized the potential energy function V_{inter_core} (eq. S9) with the ϵ of 0.65 for inter-core-domain interaction in all the simulations described below.

Table S1. Virtual state setting and simulation length.

Iteration No.	Number of virtual states	Simulation length ($\times 10^6$ steps)[a]
#1-5	7	1.0
#6	7	1.2
#7-8	7	1.4
#9	5	2.0
#10-11	4	2.0
#12	4	2.4

#13-15	4	2.6
#16[b]	4	12.0

[a]Simulation length (number of MD steps) for each of 128 runs.

[b]The sixteenth simulation is the production run.

Table S2. Energy zone for virtual states.

Iteration No.	Energy zone[a]
#1-8	[0.0, 0.25], [0.125, 0.375], [0.25, 0.5], [0.375, 0.625], [0.5, 0.75], [0.625, 0.875], [0.75, 1.0]
#9	[0.0, 0.15], [0.075, 0.27], [0.15, 0.39], [0.27, 0.58], [0.39, 1.0]
#10-16	[0.0, 0.125], [0.0625, 0.25], [0.125, 0.5], [0.25,1.0]

[a]Energy zone $[E_i^{low}, E_i^{up}]$ for the i-th virtual state is given in a normalized form as $[\lambda_i^{low}, \lambda_i^{up}]$, where $E_i^{low} = \lambda_i^{low}\Delta E + E_{low}$ and $E_i^{up} = \lambda_i^{up}\Delta E + E_{low}$. The quantity ΔE is the width for the entire energy range: $\Delta E = E_{up} - E_{low}$, where E_{up} and E_{low} are the upper and lower value for the entire energy range: $[E_{low}, E_{up}] = [-102300\ \text{kcal/mol}, 69300\ \text{kcal/mol}]$.

Table S3. Prominent contact in all atom simulation of linker region

Rank	Residue 1	Residue 2	Prob.	Rank	Residue 1	Residue 2	Prob.
1	ASN23	SER27	0.978047	26	PRO13	GLN30	0.565987
2	ASN24	SER28	0.953985	27	GLY6	TYR40	0.565547
3	PRO22	SER27	0.715654	28	GLY6	GLU39	0.565313
4	LEU21	SER26	0.645219	29	LYS4	ASP37	0.561774
5	ASN23	SER28	0.630158	30	LYS4	LEU36	0.561774

6	SER16	ALA20	0.610774	31	LYS4	PRO35	0.561774
7	THR17	LEU21	0.610753	32	LYS5	GLU39	0.561083
8	LYS34	GLY38	0.607818	33	PRO13	THR17	0.560192
9	LEU12	SER16	0.601879	34	LEU2	LEU36	0.553724
10	GLY15	ARG19	0.601842	35	LYS5	ASP37	0.550361
11	LEU12	THR17	0.600647	36	ASN24	PRO29	0.542103
12	LYS5	GLY38	0.596418	37	GLU7	PRO35	0.538107
13	PRO14	LYS18	0.592074	38	HIS10	GLN30	0.537345
14	GLY6	GLY38	0.583776	39	GLU7	TYR40	0.534104
15	GLY6	SER28	0.583504	40	LEU2	PRO35	0.533431
16	THR17	PRO22	0.581628	41	SER27	PRO31	0.526967
17	LYS18	PRO22	0.568143	42	THR17	SER27	0.518567
18	HIS10	PRO31	0.566331	43	HIS10	LYS32	0.518497
19	LYS18	ASN23	0.566127	44	PRO8	ASP37	0.503605
20	PRO13	PRO29	0.566097	45	GLY6	ASP37	0.503605
21	GLY6	PRO35	0.566007	46	GLU7	ASP37	0.503589
22	GLY6	LYS34	0.566007	47	LYS5	SER27	0.490885
23	LYS5	PRO35	0.566007	48	PRO8	GLY38	0.488531
24	ARG3	PRO35	0.566007	49	ARG3	LYS34	0.479736
25	ARG3	LEU36	0.566004	50	GLY6	LEU36	0.479611

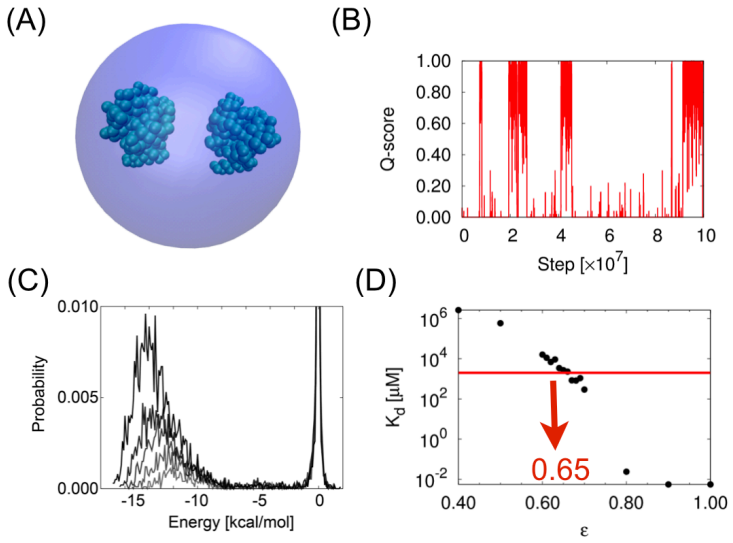

Fig. S1 The determination of the parameters of the inter-core-domain interaction. (A) The initial structure of the coarse-grained simulation for the determination of the ε in eq. 4. (B) The time trajectory of the inter-core-domain Q-score (Time is not physical time, but reduced time, i.e. time step of MD simulation). Q-score represents the ratio of the transiently formed contacts to the natively formed contacts. Natively formed contacts are defined using the X-ray crystal structure in which four core domains bind to its specific DNA (60) (PDB ID: 3KMD). (C) Probability distribution of the inter-core-domain interaction energy. (D) The inter-core-domain dissociation constant versus ε. The red line represents the experimentally measured value (2 mM)

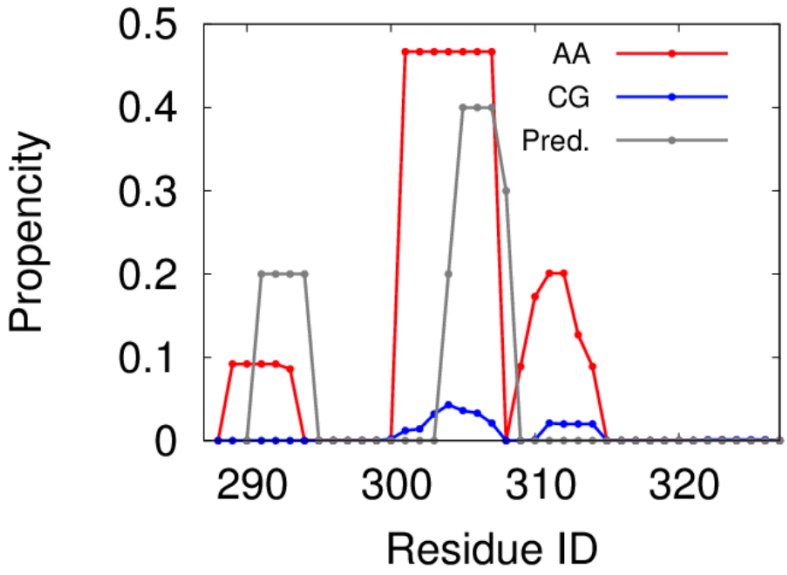

Fig. S2 Ahelical propensity (the population of structures with a helical structure) for each residue calculated from the obtained atomistic conformational ensemble (red) and estimated from only the amino acid sequence using the AGADIR (grey).

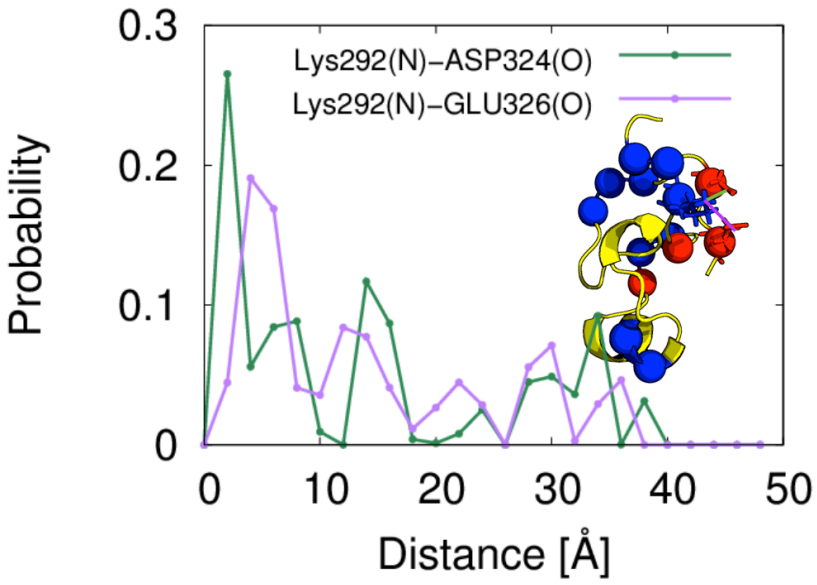

Fig. S3 The probability distributions of the distance between the nitrogen atom of LYS 292 and the oxygen atom of ASP 324 (Green) and between the nitrogen atom of LYS 292 and the oxygen atom of GLU 326 (purple). (Inset) The representative structures of the largest cluster. The blue (red) sphere represents Cα atoms of the positively (negatively) charged residues. The green (purple) line is depicted between the nitrogen atom of LYS 292 and the oxygen atom of ASP 324 (GLU 326).

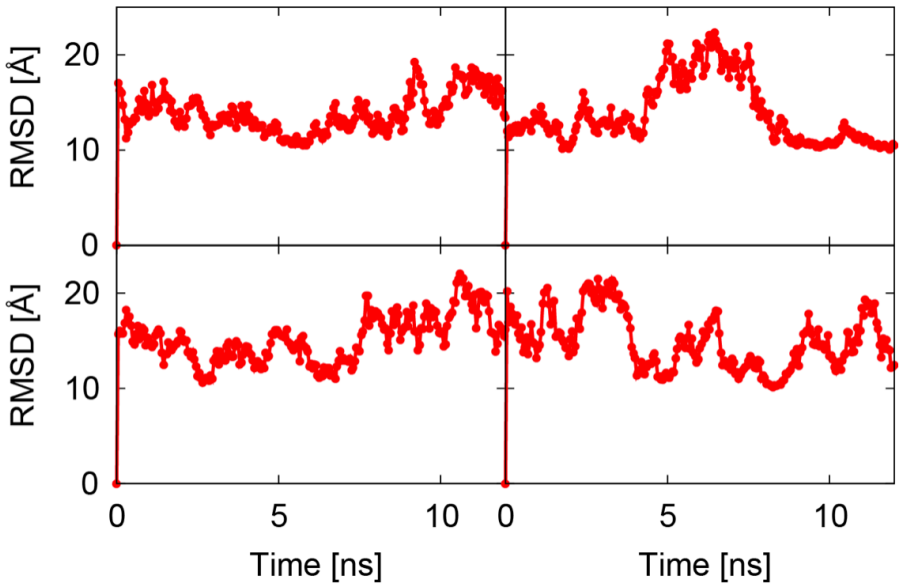

Fig. S4 Root mean square deviation of each snapshot in the representative AA V-McMD simulation. The reference structure is the initial structure of each production simulation.

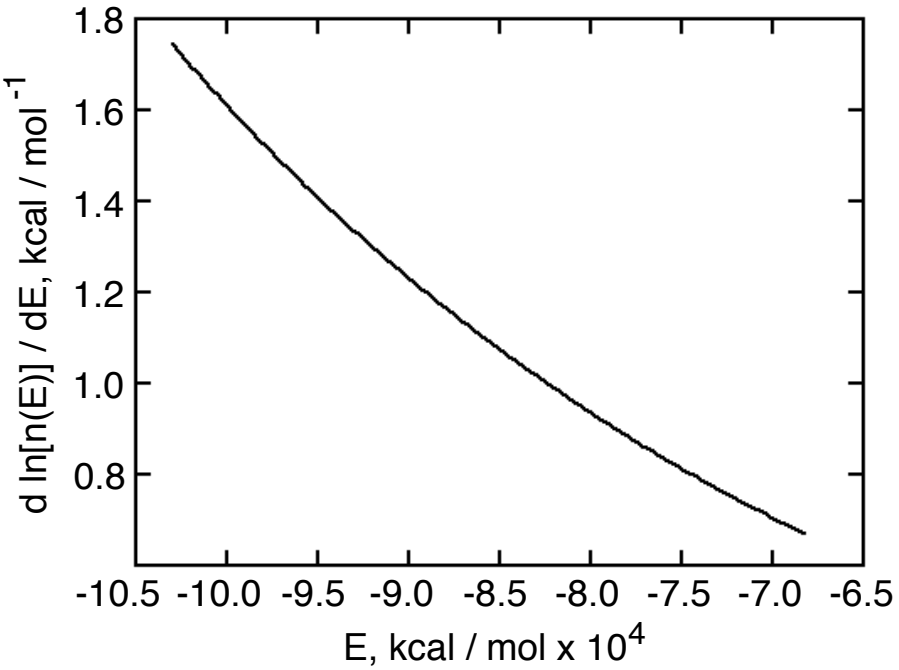

Fig. S5 The horizontal axis represents potential energy and the vertical axis represents the $d\ln[n](E)./dE$ in eq. S3

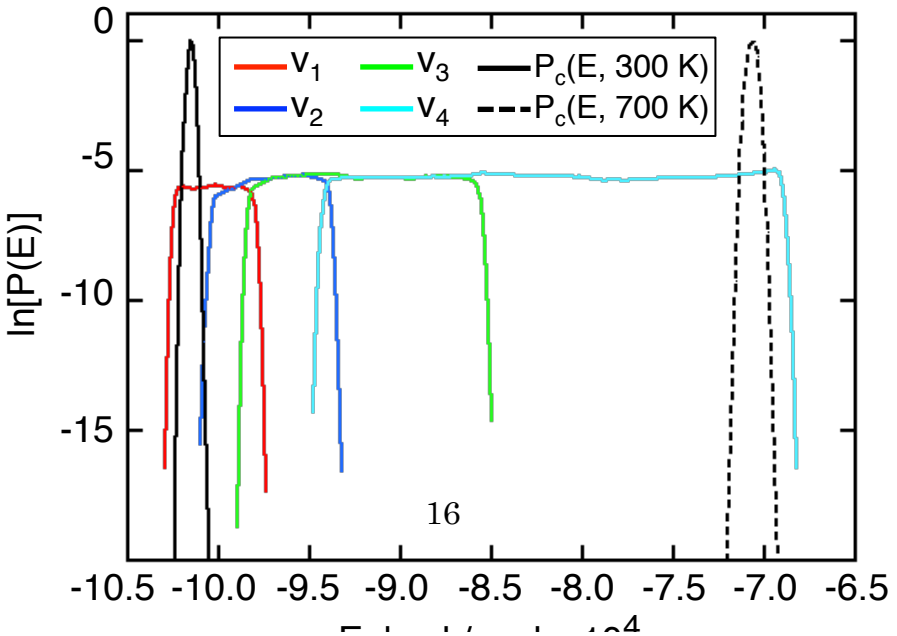

Fig. S6 Energy distributions in log scale. Each colored line represents the energy distribution in each virtual state (v_1, v_2, v_3, and v_4). Black solid and dashed line represent the canonical energy distributions $P_c(E,T)$ at 300 K and at 700 K, respectively.