

Article

Multi-scale Ensemble Modeling of Modular Proteins with Intrinsically Disordered Linker Regions: Application to p53

Tsuyoshi Terakawa,¹ Junichi Higo,² and Shoji Takada^{1,*}¹Department of Biophysics, Graduate School of Science, Kyoto University, Kitashirakawa-Oiwakecho, Sakyo, Kyoto, 606-8502, Japan; and ²Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka, 565-0871, Japan

ABSTRACT In eukaryotic proteins, intrinsically disordered regions (IDRs) are ubiquitous and often exist in linker regions that flank the functional domains of modular proteins, regulating their functions. For detailed structural ensemble modeling of IDRs, we propose a multiscale method for IDRs that possess significant long-range order in modular proteins and apply it to the eukaryotic transcription factor p53 as an example. First, we performed all-atom (AA) molecular dynamics (MD) simulations of the explicitly solvated p53 linker region, without experimental restraint terms, finding fractional long-range contacts within the linker. Second, we fed this AA MD ensemble into a coarse-grained (CG) model, finding an optimal set of contact potentials. The optimized CG MD simulations reproduced the contact probability map from the AA MD simulations. Finally, we performed the CG MD simulation of the tetrameric p53 fragments including the core domains, the linker, and the tetramerization domain. Using the obtained ensemble, we theoretically calculated the small angle x-ray scattering (SAXS) profile of this fragment. The obtained SAXS profile agrees well with the experiment. We also found that the long-range contacts in the p53 linker region are required to reproduce the experimental SAXS profile. The developed framework in which we calculate the long-range contact probability map from the AA MD simulation and incorporate it to the CG model can be applied to broad range of IDRs.

INTRODUCTION

It has become clear that intrinsically disordered regions (IDRs) are ubiquitous in eukaryotic proteins: 30% to 50% of eukaryotic proteins have been predicted to have IDRs with at least 30 consecutive residues (1–3). IDRs often play crucial roles in molecular recognition and signaling, protein modification, molecular assembly, entropic chain activities, and so on (4). Furthermore, IDRs are related to various human diseases, such as cancer, cardiovascular diseases, amyloidosis, neurodegenerative diseases, and diabetes (5,6). Notably, in eukaryotic proteins, most IDRs exist either at the tails or at the linkers that flank folded domains in multidomain and modular proteins (see Fig. 1 A as an example) (7). Such flexible linkers can control the relative location of the flanking domains, which is important for the proteins to regulate their functions (8,9).

Biophysically, IDRs possess energy landscapes with many shallow and competing minimums at room temperature and thus encompass the broad spectrum of conformational ensembles (4,10–13). The dominant feature of these ensembles is the lack of persistent secondary and tertiary structures, with a flexible chain transiently sampling fractional local secondary structure as well as some long-range contacts (14–16). An accurate description of the conformational ensemble is crucial to fully understand its functions. It is, however, difficult to obtain detailed information of

conformational ensembles of IDRs solely by conventional experimental techniques. Therefore, combining experimental data with computational method has been used and shows some promise (14–16).

First, one can generate a large variety of conformations and then select the subset of conformations that have ensemble averages that agree with experimental data (17,18). Another approach uses a restrained MD simulation that is performed with experimental restraint terms for biasing the conformational sampling (19). These approaches help researchers obtain IDR conformational ensembles consistent with experiments (14–16). However, they commonly suffer from a so-called “degeneracy problem”—that is, given the low-resolution information from experiments, there can be many different ensembles that are consistent with the experiments. Sophisticated methods to determine weights for each conformation in the ensemble have been developed to mitigate the problem (20). An alternative way that does not suffer from the degeneracy problem is to perform MD simulations without the experimental restraint term. However, conformational sampling by AA MD itself is highly nontrivial for systems with transient long-range contacts. The purpose of this paper is to develop a multiscale method that can deal with IDRs with long-range residual contacts.

In addition, IDRs are often flanked by folded domains and thus the full-length proteins are large and beyond the reach of current AA MD simulations. Modular proteins, comprising two or more folded domains tethered by IDR linkers, are common in nature (21–23), among which is

Submitted February 24, 2014, and accepted for publication June 18, 2014.

*Correspondence: takada@biophys.kyoto-u.ac.jp

Editor: Martin Blackledge.

© 2014 by the Biophysical Society
0006-3495/14/08/0721/9 \$2.00



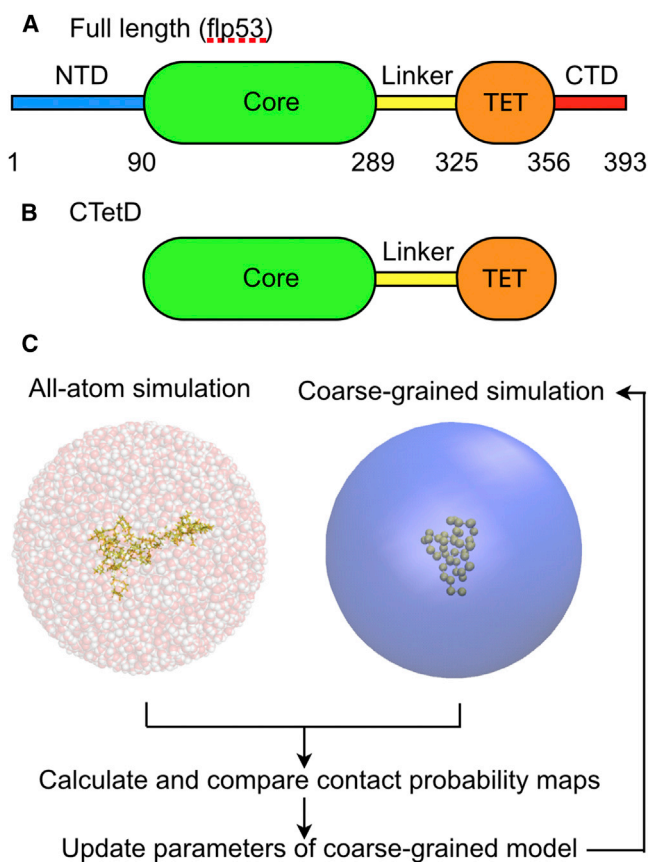


FIGURE 1 The domain maps of (A) the full length p53 and (B) the CTetD. The ellipses represent domains that have stable tertiary structure in solution, whereas the rectangles represent disorder regions. (C) The strategy to determine the parameters of the CG linker model based on the AA V-McMD simulation result. Snapshots show the AA V-McMD simulation of the linker region (left) and the CG MD simulation (right). To see this figure in color, go online.

the eukaryotic transcription factor p53, which we used in this study. The full-length p53 contains two folded domains (the core domain and the tetramerization domain (Fig. 1 A)) that are flanked with both N- and C-terminal disordered tails and the disordered linker region (24). Interestingly, of these five distinct regions, two regions bind to DNA: the core and the C-terminal domains. Our previous CG MD simulation study showed that tetrameric full-length p53 slides on DNA with its C-terminal IDR while its core domains repeat dissociation from and association to DNA (25). These are in accord with a recent single molecule experiment (26). This result points to the functionally important role of the p53 linker region that connects two DNA binding regions. Tidow et al. reported the SAXS profiles of the p53 fragment that lacks N- and C-terminal intrinsically disordered domains (CTetD; Fig. 1 B) and modeled a static structure that is consistent with the experimental data (27). As another study found, obtaining a unique static structure that could describe the SAXS profiles did not exclude the possibility that the CTetD is flexible in solution (28). In this construct, the

linker region controls a relative location between the core domain and the tetramerization domain. Therefore, the model of this linker region has the dominant effect on the overall shape of the CTetD and consequently its SAXS profile. We also know, a posteriori, that the linker region contains transient long-range order. This makes the CTetD construct an ideal system to verify quantitative modeling of IDRs in modular proteins.

In this study we extended previous multiscale approaches (29–32) and proposed a multiscale ensemble modeling method that can be used on IDRs with long-range residual contacts. For the p53 linker, we first performed atomistic structure modeling by taking the recently developed enhanced sampling techniques, a virtual-system coupled multicanonical MD (V-McMD) simulation (33). This AA MD based ensemble was utilized to obtain an optimal set of CG interaction parameters (Fig. 1 C). Using the obtained CG model of the p53 linker, we performed CG MD simulations for the p53 CTetD, and theoretically calculated the SAXS profile of the obtained CG conformational ensemble. We find that the profile agrees well with that of the experiment. Finally, we investigated the effect of the long-range order in the linker on biological functions, focusing on the contact probabilities between two core domains. The results suggest that the linker conformations modulate the inter-core domain contacts to a certain degree. In this work, we successfully modeled the p53 linker region that has long-range contacts and obtained structural ensemble of p53 CTetD that cannot be obtained easily by the previously established modeling methods. The framework in which we calculate the long-range contact probability map from the AA MD simulation and incorporate it to the CG model can be applied to broad range of IDRs.

METHODS

Multiscale method for intrinsically disordered region

We next outline the multiscale method for IDR modeling, in which we use AA MD simulations to tune parameters in a CG model. We suppose the case that a relatively large modular protein contains IDRs and that the IDRs possess long-range residual order. Here, the long-range residual order means a structural order stabilized by nonlocal interactions. For this IDR region, first, we obtained the conformational ensemble by performing AA MD simulations. Then, using this AA MD-based ensemble, we tuned the CG model.

For the IDR that has long-range order, we write the potential energy function V of our CG model as follows:

$$V = V_0 + \sum_{i>j+3} \epsilon_{ij} \left[5 \left(\frac{r_{ij}^0}{r_{ij}} \right)^{12} - 6 \left(\frac{r_{ij}^0}{r_{ij}} \right)^{10} \right], \quad (1)$$

where V_0 represents a CG potential previously developed for IDRs that are supposed to approximate local residual order (see the following sections for the explicit formula); r_{ij} is the distance between the i -th and the j -th CG

particles; and ϵ_{ij} and r_{ij}^0 are parameters to be determined via the multiscale method.

First, from the AA ensemble, we calculated the modes of distances between CG particle pairs and set them to the r_{ij}^0 parameters. We also calculated the contact probability map from the AA ensemble (r_{ij}^{AA}). We considered a residue pair was contacted, if the distance between its CG particles (the C_α atoms of these residues for AA ensemble) was less than 8.5 Å.

To determine the ϵ_{ij} parameters, we first performed CG MD simulations of this region with the ϵ_{ij} parameters set to 0.0. Then, we calculated the contact probability map from the CG ensemble (p_{ij}^{CG}). By comparing the contact probability map p_{ij}^{CG} with p_{ij}^{AA} , we updated the ϵ_{ij} parameters that were initially set to 0.0, according to the following equation:

$$\epsilon_{ij}^{new} = \epsilon_{ij}^{prev} - \ln \frac{p_{ij}^{CG}}{p_{ij}^{AA}} \quad (2)$$

Then, we performed the CG MD simulation again, using the new ϵ_{ij}^{new} parameters. We repeated 1), the CG MD simulation, 2), the contact probability calculation, and 3), the parameter update until the contact probability map from the CG ensemble converged to that from the AA ensemble as shown below. The strategy is outlined in Fig. 1 C. Previously, a similar procedure was utilized to incorporate interaction information from experiments into their CG model (34).

All-atom simulation of p53 linker region

In this section we describe the AA MD simulation method for the p53 linker (for details of the method, see the [Supporting Material](#)). The system consists of the p53 linker segment with a few residue extensions at both ends (40 residues long, residue ID: 288 to 327), which is solvated with water molecules. The amino-acid sequence is Ace- NLRKKGEPHHELPPGSKRALPNNTSSSPQPKKKPLDGET-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 solvent sphere (diam. = 80 Å), setting the center of mass at the center of the sphere. The system consisted of 30,937 atoms (640 polypeptide atoms, 36 Cl⁻, 30 Na⁺, and 10,077 water molecules). To sample the conformation of the linker region with reduced influence of the boundary condition, we fixed the linear and the angular momenta of the linker segment to zero by rescaling the velocities. We did not use the periodic boundary condition in this study because the periodicity may cause artificially interchain entangling among the different periodic boxes. The solvent sphere was set as large as possible, yet small enough so that the multiconformational sampling can be done within a feasible simulation time.

The force field parameters for the polypeptides were from an AMBER-based hybrid force field (35) defined as $V_{hybrid} = 0.25V_{94} + 0.75V_{96}$, where V_{94} and V_{96} denote the AMBER parm94 (36) and parm96 force fields, respectively (37). Previous McMD simulations with V_{hybrid} 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 (35). Therefore, we used V_{hybrid} for the present study. We have successfully applied this force field to protein folding (38–40) and an ensemble modeling of an intrinsically disordered protein (IDP) (41). Although there is no perfect atomistic force field that can be applicable to any amino-acid sequence, our preceding works (35,41) have suggested that the force field we used in the present study does not have an apparent bias in secondary structure formation and is appropriate for IDRs.

The AA simulation procedure consists of two stages: 128 pre-V-McMD simulations were performed with a high temperature. These 128 simulations were all started from different random conformations. Then, for 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 (see the [Supporting Material](#) for the accuracy of the biased potential estimation). Each of these simulations was started from the final conformation of each of the pre-V-McMD simulations. The 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 reweighting technique (33).

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 (“pure-CG” model in (42)). This model does not take into account long-range contacts. The potential energy function of that model is as follows:

$$V_0 = V_{w/o \text{ contact}} = V_{bond} + V_{angle} + V_{ele} + V_{ex}, \quad (3)$$

where V_{bond} , V_{ele} , and V_{ex} are the bond-stretching term, the electrostatics term, and the excluded volume effect term, respectively. (For complete description of the potential energy function, refer to an earlier study (42)). 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.

The molecular system of the CG MD simulation was the same as that of the AA MD simulation except for the absence of the cap of the segment termini. We used the one-bead-per-one-amino-acid CG model and put a CG particle on each C_α position of the 40-residues-long linker segment. We generated a random conformation of the linker segment for the initial conformation and put it into a sphere. The diameter of the sphere was 80 Å, the same as that of sphere 2 in the AA MD simulation. Because the diameter of the sphere was same between the AA and the CG MD simulations, the confinement was expected to affect similarly the AA and the CG conformational samplings. Therefore, we thought that the confinement effect on the parameter calibration procedure was negligible. Production runs for the CG simulations were performed by Langevin dynamics for 10^8 MD steps using CafeMol 2.0 (43).

Coarse-grained simulation of two core domains

Experimentally, it has been shown that two p53 core domains form a loose dimer with the dissociation constant of 2 mM at 100 mM monovalent ion (44). 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 (27). To model this intercore-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. S1 A). The initial coordinate of the core domain was taken from the x-ray crystal structure (45) (PDB ID: 2XWR). We put two core domains into a sphere with the diameter of 50 Å. We adopted recently developed Go-like AICG2 model (46) for the intramolecular potential energy function that stabilizes the native structure (45) (PDB ID: 2XWR). The intercore-domain potential energy function is defined as follows:

$$V_{inter-core} = V_{ele} + V_{ex} + \sum_{i>j+3}^{native \text{ contact}} \epsilon_{ij} \left[5 \left(\frac{r_{ij}^0}{r_{ij}} \right)^{12} - 6 \left(\frac{r_{ij}^0}{r_{ij}} \right)^{10} \right], \quad (4)$$

where V_{ele} and V_{ex} were electrostatics term and excluded volume effect term, respectively (for complete description of these terms, refer to the [Supporting Material](#)). The i and j run over the CG particle pairs that contacted in the experimentally indicated interface in the x-ray crystal structure (47) (PDB ID: 3KMD). The r_{ij}^0 is the distance between two CG particles i and j in the native structure. The ϵ_{ij} 's are the AICG2 model parameters (46). 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 interprotein-interaction. Accordingly, to reproduce the dissociation constant, we scaled the intermolecular native contact interaction by an additional factor $\epsilon = 0.65$ (the determination process of ϵ is described in detail in the [Supporting Material](#)). The ion strength was set to the same value as that of the experiment (100 mM) (44).

In previous NMR studies (27,48), the other interdomain interactions (i.e., core-linker, core-Tet, and linker-Tet interactions) were not identified. Therefore, we imposed only repulsive and electrostatic interactions to the other interdomain interactions.

Construction of coarse-grained conformational ensemble of CTetD

To validate the parameters of the CG linker model, we performed the CG MD simulation of the tetrameric CTetD (Fig. 1 B), obtained the CG conformational ensemble, and theoretically calculated the SAXS profile from this conformational ensemble using the CRY SOL program (49) after reconstructing AA model using the PULCHRA program (50). We used 2XWR (45) and 1AIE (51) as the template structure for the core (residues 91 to 289) and the tetramerization (residues 326 to 356) domains, respectively, and modeled the linker region (residues 290 to 325) as a random coil. We used Eq. 1 as the potential energy function (with and without intrachain contact interactions) for the linker region, the AICG2 model as that for the core and the tetramerization domains, and Eq. 4 as that for the intrachain-domain interaction. The ion strength was set to the same value as the experiment (150 mM) (27). Each production run was performed by Langevin dynamics for 10^9 MD steps.

RESULTS AND DISCUSSIONS

All-atom simulation of p53 linker region

First, we performed the AA V-McMD simulation of the p53 linker region with a few residue extensions in both ends (residue ID: 288 to 327). We obtained the well-converged conformational ensemble (see the end of this section for the convergence test). To characterize structures in the ensemble, we grouped the structures into clusters. The clustering was done by the gromos algorithm (52) implemented in GROMACS 4.5.5 (53) using the root mean square deviation (RMSD) between each pair of structures as a distance metric. In the clustering analysis, we chose 200 conformations that had large Boltzmann weights at 300 K out of 30,000 stored conformations. Using the RMSD cutoff 3.0 Å, we obtained 31 clusters. In Fig. 2, we show the representative structures of the top-six largest clusters, together with the ranking by the cluster size. We find that the conformational ensemble is very diverse.

Interestingly, each conformation has its specific secondary structure and tertiary contact pattern. For example, the top-cluster structure contains a helical region (light green in Fig. 2), although the same region in the other five clusters does not contain a helical region. The secondary structure prediction also indicated that this region have high helical probability (Fig. S2 in [Supporting Material](#)). A recent study revealed that the iASPP protein, whose function is to modulate p53-dependent apoptosis, is bound to the p53 linker region although its binding site in the linker region has

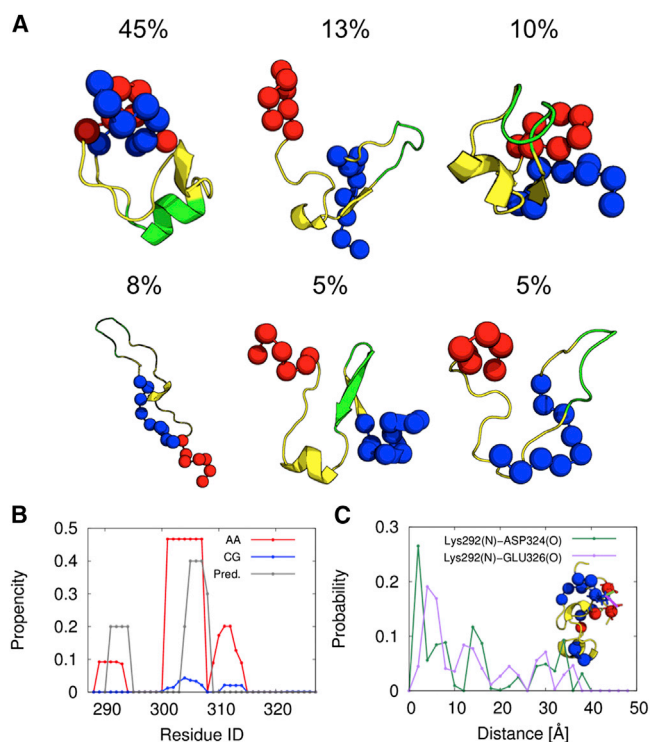


FIGURE 2 The representative structures of the top-six largest clusters obtained using the AA V-McMD simulation. The blue spheres represent the C_{α} atoms of the N-terminal residues (residue ID: 288 to 297). The red spheres represent the C_{α} atoms of the C-terminal residues (residue ID: 321 to 327). The helical region of the largest cluster (residue ID: 301 to 307) is colored green. To see this figure in color, go online.

not been elucidated yet (54). It was also reported that a significant number of molecular recognition events often involved loosely structured regions within IDRs (55). Taken together, we speculate that either or both of these regions (residue ID: 289 to 293 and 301 to 307) with relatively high helical propensities recognize the iASPP by the “coupled folding and binding” manner.

Moreover, the N-terminal region (blue) and C-terminal region (red) in the simulated peptide form the long-range contacts in the most and the third-most populated clusters, but not in the other clusters (Fig. 2). We speculate that some of these long-range contacts are caused by the electrostatic interaction (Fig. S3). Overall, it is suggested that the conformational ensemble of the p53 linker region cannot be described by the simple random coil model and is composed of the structures with the transient secondary structures and long-range contacts.

We assessed the convergence of the conformational ensemble obtained by V-McMD simulations. To do so, we randomly divided the 128 simulation trajectories into two groups, extracted the structures from each of them to make two conformational ensembles (“AA_1” and “AA_2”), calculated contact probability maps and the distributions of the radius of gyration, and compared them (left panel in Fig. 3 A). In drawing the contact

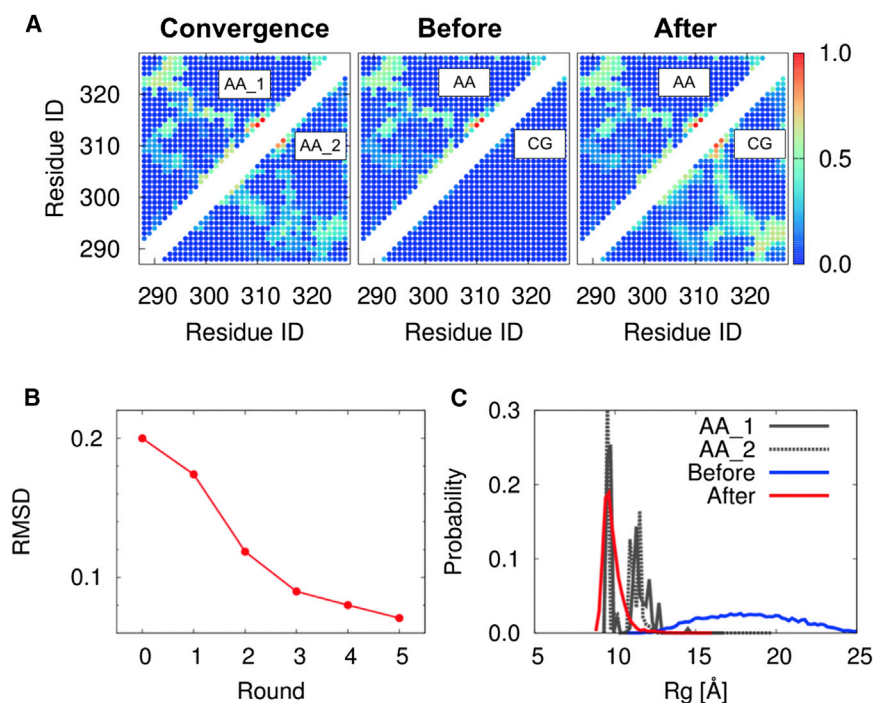


FIGURE 3 The determination of parameters of the CG linker model. (A) The contact probability maps from AA V-McMD simulation (left) using two halves of ensemble (AA_1 and AA_2; see text for details). The contact probability maps from the AA V-McMD simulation (center; above diagonal), from the coarse-grained (CG) simulation with all of the ϵ 's in Eq. 1 set to 0.0 (center; below diagonal), and from the CG simulation with the updated parameters (right; below diagonal). (B) The RMSD between the contact probability maps from the AA V-McMD simulation and those from the CG MD simulation of each round of the parameter update procedure. (C) The probability distributions of the radius of gyration from the AA V-McMD simulation (“AA_1” and “AA_2”), from the CG simulation with all of the ϵ 's in Eq. 1 set to 0.0 (blue), and from the CG simulation with the updated parameters (red). To see this figure in color, go online.

maps, we identify two residues forming contact when the distance between two C_α atoms is less than 8.5 Å. From Fig. 3 A, we see that these two contact maps are similar, suggesting that the AA ensemble converges relatively well in this perspective. This information is directly incorporated into the CG model below. Therefore, the convergence in this perspective is critically important. From this map, we also see that, in addition to the transient long-range contacts between N- and C-terminal regions, there are several fractional but noticeable contacts in the AA conformational ensemble. We listed the prominent contacts in Table S3. In Fig. 3 C, distributions of radius of gyration from AA_1 and AA_2 are completely overlapped, suggesting that the AA ensemble converges relatively well in this perspective, too.

Determination of parameters of coarse-grained linker model

For the p53 linker region, we seek a set of CG model ϵ_{ij} parameters in Eq. 1 that can reproduce the residue-residue contact probabilities computed by the AA simulations as close as possible.

First, we performed the CG MD simulation with the ϵ_{ij} parameters in Eq. 1 set to zero. Then, we calculated the contact probability map from the obtained CG conformational ensemble (the bottom-right triangle region in the center panel of Fig. 3 A) using the same cutoff distance as above for the definition of contacts. The map shows almost no fractional contact in this CG conformational ensemble. For comparison, the top-left triangle region in the same

panel shows the contact map by AA simulations. We see clear difference between the two halves.

Then, we updated the ϵ_{ij} parameters according to the Eq. 2. Using the updated parameters, we repeated the CG MD simulation and calculated the contact probability maps. As a result, the difference between the AA map and the CG map became smaller (the RMSD plotted in Fig. 3 B). We repeated this cycle five times. Fig. 3 B shows that, as we repeated the cycle, the RMSD monotonically decreased and finally converged to a small value. In the bottom-right triangle region of the right panel of Fig. 3 A, we show the contact probability map calculated from the final CG conformational ensemble. By comparing the elements above (the contact map by AA simulations) and below the diagonal, we see that the CG map and the AA maps are essentially the same. Thus, the parameter set in the fifth round reproduced the AA contact probability map well and was used for the subsequent CG MD simulations of the CTetD (Fig. 1 B).

To investigate the characteristics of the CG conformational ensemble in the final round, we calculated and plotted probability distributions of radius of gyration (R_g ; Fig. 3 C). The plot shows that the average R_g of the ensemble (red) is smaller than that in the initial round with all of the ϵ_{ij} parameters set to zero (blue). In Fig. 3 C, we also plotted the probability distribution of R_g of the AA conformational ensemble (black or gray, see the previous section). Interestingly, there are three peaks in this probability distribution, which indicates that several distinct states with different R_g 's coexist in the AA conformational ensemble. By comparing these probability distributions, we can see that

the most intense peak with the smaller average R_g of the AA probability distribution is reproduced by the CG conformational ensemble in the final round, but not by the initial round. This suggests that the AA conformational ensemble is more compact than expected from the random coil model and the fractional long-range contacts are required to reproduce this conformational ensemble. We can also see that the other two peaks are reproduced by neither of the CG conformational ensembles. This suggests that the Lenard-Jones-type contact term in Eq. 1 is not sophisticated enough to reproduce the transition between several distinct states in the AA conformational ensemble. The development of more sophisticated CG model for contact interaction is desirable in future. For example, a simple way to improve the model is to impose the contact potential (Eq. 1) only to the interaction between beads representing a hydrophobic or bulky and polar amino acid (i.e., Leu, Ala, Asn, and Gln). However, we consider that the simple model can be used for the p53 linker region, because these less populated conformations do not seem to affect the conformational ensemble of the CTetD.

Validation of parameters of the coarse-grained model

To validate the parameters of the CG linker model obtained above, we performed 10^9 -step CG MD simulation of the tetrameric CTetD (Fig. 1 B), obtained the CG conformational ensemble, theoretically calculated the SAXS profile from this ensemble using CRY SOL (49) after reconstructing the AA model using PULCHRA (50), and compared it with that of the previous experiment (27). We note that, except for the linker region, the other parts of the CTetD have stable tertiary structures in solution (Fig. 1 B). Therefore, the overall shape of the CTetD is mostly decided by the flexible linker region, which makes this system suitable for validation of the parameters of the CG linker model.

First, we compared the normalized SAXS intensity profile calculated from the CG MD simulation with that of the experiment (Fig. 4 A). We see that the SAXS profile from the CG MD simulation with the contact interactions in the linker region (all of the ϵ 's in Eq. 1 calibrated based on the AA MD simulation; red in Fig. 4 A) reproduces that of the experiment well (gray in Fig. 4 A) ($\chi = 0.38$, where χ is the sum of square difference of each data point). On the other hand, the SAXS profile from the CG MD simulation without the contact interactions (all of the ϵ 's in Eq. 1 were set to 0; blue in Fig. 4 A) exhibits clear deviation from the experimental data ($\chi = 2.60$).

Second, we compared the Kratky plots ($s^2 I(s)$ versus s , where s is a scattering vector and $I(s)$ is a scattering intensity) (Fig. 4 B). The experimental Kratky plot (gray) shows a single pronounced peak that is indicative of a spatial decorrelation between the different globular domains (28). This peak is nearly perfectly reproduced by the CG MD simula-

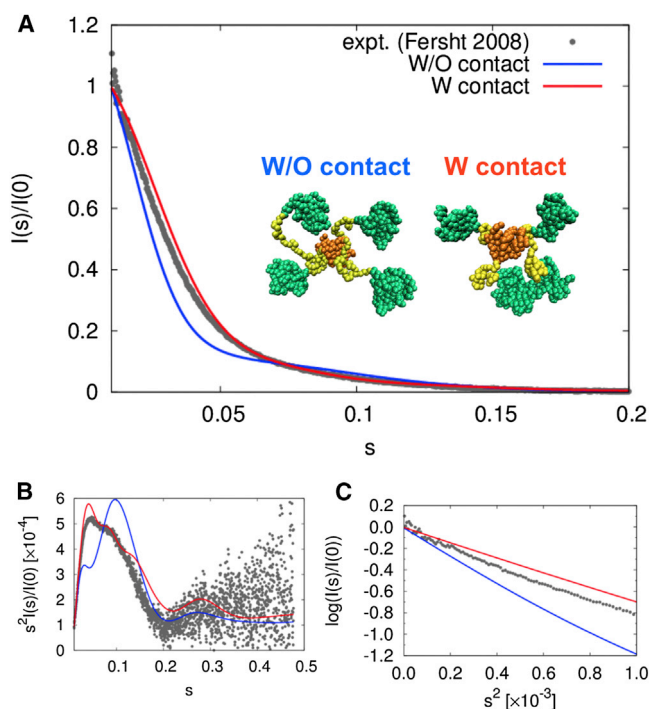


FIGURE 4 Comparison of the experimentally observed and theoretically calculated SAXS profiles. (A) The SAXS intensity profile from the experiment (gray points), from the CG MD simulation with the updated parameters (red solid line), and from the CG MD simulation with all of the ϵ 's in Eq. 1 set to 0.0 (blue solid line). Inset shows representative structures from the CG MD simulation with all the ϵ 's in Eq. 1 set to 0.0 (left) and from the CG MD simulation with the updated parameters (right). (B) Kratky plot. The color assignment is same as that of A. (C) Guinier plot. The color assignment is the same as that of A. To see this figure in color, go online.

tion with the contact interaction in the linker region (red in Fig. 4 B), whereas the peak position is clearly different when all of the ϵ 's in Eq. 1 are set to zero. This result indicates that the relative position of the different core domains is well decorrelated in the CG MD simulation with the optimal ϵ 's.

Third, we compared the Guinier plots ($\log(I(s))$ versus s^2) (Fig. 4 C). From the slope of the linear region in the small-angle limit, we can estimate the radius of gyration of the molecule. From this figure, we can see that the slope of the CG MD simulation with the contact interactions in the linker region is essentially the same as that of the experiment, whereas that of the CG MD simulation without the contact interaction is quite different. This result indicates that the experimental radius of gyration (52.2 Å) is closer to the CG MD simulation with the contact interactions in the linker region (45.8 Å). The latter is significantly smaller than that of the CGMD simulation without long-range order in the linker (64.8 Å). This result also indicates that the structures from the CG MD simulation with the contact interaction (right panel of the inset of Fig. 4 A) tend to be more compact than those from the CG MD simulation without them (left panel of the inset of Fig. 4 A). In the CG MD simulation with

the contact interaction, the average radius of gyration is smaller than that of the experiment. We think that this can partly be attributable to the bias in the conformational ensemble obtained using the AA MD simulation to the relatively compact structures. This is a previously reported problem in almost all of the current generation force fields (56). Thus, it is desired that next generation force fields are designed to mitigate this problem.

Taken together, these results show that we can obtain the CG conformational ensemble of the CTetD that fairly well reproduces the experimental SAXS profile using the CG MD simulation with the contact interaction in the linker region. Thus, the parameters of the CG linker model obtained based on the AA MD simulation are valid.

Intercore-domain interaction in CTetD

In the “Determination of parameters for intercore-domain interaction” section, we tuned the CG model ϵ_{ij} parameters in Eq. 4 for the intercore-domain interaction. The parameters were determined so that the experimental dissociation constant of the intercore-domain interaction was reproduced. In the tetrameric CTetD, four core domains are tethered by tetramerization domains and thus tethering by the linker modulates the inter-core domain associations. To reveal the effect of the tethering on the inter-core-domain association, we plot the probability distributions of intercore-domain Q-score of each pair of core domains in Fig. 5. The Q-score represents the ratio of the transiently formed contacts to the natively formed contacts.

We note that the tetramerization domain of p53 takes dimer-of-dimers. The primary dimer makes tight contacts including interchain β sheet formation. Interactions between two primary dimers are via helix-helix contacts and are

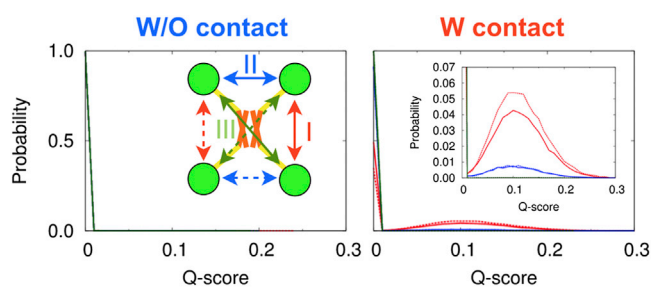


FIGURE 5 Probability distributions of the inter-core-domain Q-score from the CG MD simulation with all of the ϵ 's in Eq. 1 set to 0.0 (left) and from the CG MD simulation with the updated parameters (right and right inset). The Q-score represents the ratio of the transiently formed contacts to the natively formed contacts. Natively formed contacts are defined using the x-ray crystal structure in which four core domains bind to its specific DNA (PDB ID: 3KMD) as a template structure. We calculated the probability distribution of each pair of the core domains. Because of the symmetry of the molecule, these pairs can be divided into three classes. Therefore, we mapped these three classes on to the cartoon of tetramerized CTetD (left inset). The color assignment is the same as that of Fig. 1 A and B. To see this figure in color, go online.

weaker. Because of the nature of the dimer-of-dimer form of tetramerization, inter-core-domain interactions have three types of pairings: 1), the pairing of core domains, of which chains form the primary dimer in the tetramerization domain; 2), the pairing of core domains, of which chains form contacts via secondary dimer interface in the tetramerization domain; and 3), the pairing of core domains, of which chains are most distant and thus form the least contacts in the tetramerization domain (arrows in the inset of the left panel of Fig. 5). In Fig. 5, we used different colors for different types of pairings: red for the type 1, blue for the type 2, and green for the type 3. Because of the symmetry of the molecule, there are two pairs in each of the three types. We distinguished the two pairs using the solid and the dashed curves.

The right panel of Fig. 5 shows that, when we included long-range interaction in the linker region, we see a weak yet significant probability of contacts between core domains (a low and broad peak around the intercore domain Q-score of 0.1). On the other hand, we do not see significant probability of intercore domain contact when the linker does not contain long-range interactions (left panel of Fig. 5: the correlation coefficient between red (blue) curves in left and right panels of Fig. 5 is around 0.15). Note that the inter-core-domain contact strengths are identical between the two simulations and that the origin of the difference is purely in the treatment of the linker region. This result indicates that the long-range contact interaction in the linker region increases the local effective concentration of the core domains and thereby enhances the inter-core-domain association. However, even with the contact interaction in the linker region, the contact probability is rather low. Therefore, our data show that the p53 predominantly exists as an open form, i.e., the core domains are not in contact, in the absence of its response element (RE) on DNA, whereas there is a low probability to take a topologically closed form, i.e., the core domains are in contact. This may facilitate p53 to wrap the DNA when p53 finds the response element. When p53 binds to its RE, the core domains form interchain contacts, taking the closed form. Bound to the nonspecific DNA, p53 would primarily take an open form although intercore domain contact probability may be slightly higher than that in the absence of DNA. This conformation could be somewhat different from the RE wrapping conformation previously observed using cryo-electron microscope (27). On the nonspecific DNA, the inter-core domain contact probability would be quite low. Thus, we can reasonably argue that the inclusion of the inter-core domain does not affect the main conclusion of the previous work.

From this figure, we also see the peak of type 1 interaction (red) is the most pronounced. The distribution of the intercore-domain Q-score of one core-domain-pair is similar to that of the other core-domain-pair (compare the solid and dashed line in the right panel in Fig. 5), suggesting that the initial-structure-dependency is almost diminished by

repeated dissociation and association of the core domains. This result indicates that each core cannot freely diffuse because of the tethering and that the type 1 interaction is preferred. Because the interaction energy parameters for each pair of the core domains are set identically, this preference arises from the topology of the tetramerization domain and from the restraint of the linker region.

CONCLUSION

Although the SAXS profile is not sensitive enough to test the detail of the model, the SAXS can monitor the shape of the molecular envelope. Therefore, comparison of the SAXS profile provided validation of the compactness of the compact linker structural ensemble obtained in our CG MD simulations.

At the moment, limited experimental information is available for structural and dynamic properties of the p53 isolated linker domain. In this study, we found that the long-range contacts in the linker region alter the structure of the p53 as a whole, affecting the function of this protein. Thus, more structural study of this region would be beneficial. To address biological functions of p53 more directly, we need to characterize conformations of p53 binding to the RE on DNA. This is beyond the scope of the present work and should be addressed in future studies.

The structural analysis of p53 CTetD indicates that the long-range contact in the linker region increases the local effective concentration of the core domains and thereby enhances the inter-core-domain association, though the contact probability is rather low. Therefore, our data show that the p53 predominantly exists as an open form, whereas it takes a closed form in a low probability. We speculate that this low-probability closed form in DNA-free state may facilitate the closed form on the DNA and to wrap its recognition element.

Modular proteins comprising two or more folded domains tethered by intrinsically disordered linker regions are ubiquitous in nature (21–23). Our results strongly suggest that the multiscale modeling strategy employed in this work can be used in the conformational ensemble modeling of modular proteins that usually have fractional long-range contacts in its disordered regions. Although the method itself is general, the CG potential function obtained in this work is specific to the target molecule, and is not transferable to other systems. For each of the target molecules, we started with the AA MD simulation to obtain conformational ensemble of a disordered region because different amino-acid sequences have different conformational ensembles. Therefore, the applicable range of the proposed method is limited by the capability of obtaining an equilibrium AA conformational ensemble of a disordered region, i.e., the longer the IDR, the more difficult the conformational sampling. Although various promising methods including the one we used in this work (33) have been devel-

oped, further improvement is definitely desired to overcome the limits of the present method.

SUPPORTING MATERIAL

Six figures and three tables are available at [http://www.biophysj.org/biophysj/supplemental/S0006-3495\(14\)00666-3](http://www.biophysj.org/biophysj/supplemental/S0006-3495(14)00666-3).

We thank Prof. Alan Fersht for kindly providing with experimental SAXS profile of the CTetD.

The calculations in this work were, in part, performed by using the super-computer of ACCMS, Kyoto University. T. T. thanks to the support of a JSPS fellowship. This work was supported by Grant-in-Aid for JSPS Fellows, by Grant-in-Aid for Scientific Research on Innovative Area, by Strategic Program for Innovative Research, by the Global COE Program “Formation of a Strategic Base for Biodiversity and Evolutionary Research: from Genome to Ecosystem” of the Ministry of Education Culture, Sports, Science, and Technology (MEXT) and by the Excellent Graduate School Program “Biodiversity and Evolution from Genome to Ecosystem.” J. H. appreciates a Grant-in-Aid for Scientific Research on Innovative Areas (21113006) provided by the MEXT. J. H. was supported by grants from the New Energy and Industrial Technology Development Organization (NEDO) Japan.

REFERENCES

1. Ward, J. J., J. S. Sodhi, ..., D. T. Jones. 2004. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.* 337:635–645.
2. Xue, B., A. K. Dunker, and V. N. Uversky. 2012. Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. *J. Biomol. Struct. Dyn.* 30:137–149.
3. Peng, Z., M. J. Mizianty, and L. Kurgan. 2014. Genome-scale prediction of proteins with long intrinsically disordered regions. *Proteins.* 82:145–158.
4. Dunker, A. K., C. J. Brown, ..., Z. Obradović. 2002. Intrinsic disorder and protein function. *Biochemistry.* 41:6573–6582.
5. Uversky, V. N., C. J. Oldfield, and A. K. Dunker. 2008. Intrinsically disordered proteins in human diseases: introducing the D² concept. *Annu Rev Biophys.* 37:215–246.
6. Metallo, S. J. 2010. Intrinsically disordered proteins are potential drug targets. *Curr. Opin. Chem. Biol.* 14:481–488.
7. Good, M. C., J. G. Zalatan, and W. A. Lim. 2011. Scaffold proteins: hubs for controlling the flow of cellular information. *Science.* 332:680–686.
8. Chong, P. A., H. Lin, ..., J. D. Forman-Kay. 2010. Coupling of tandem Smad ubiquitination regulatory factor (Smurf) WW domains modulates target specificity. *Proc. Natl. Acad. Sci. USA.* 107:18404–18409.
9. Smaghe, B. J., P.-S. Huang, ..., T. A. Springer. 2010. Modulation of integrin activation by an entropic spring in the beta-knee. *J. Biol. Chem.* 285:32954–32966.
10. Uversky, V. N., J. R. Gillespie, and A. L. Fink. 2000. Why are ‘natively unfolded’ proteins unstructured under physiologic conditions? *Proteins.* 41:415–427.
11. Dyson, H. J., and P. E. Wright. 2005. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* 6:197–208.
12. Huang, A., and C. M. Stultz. 2009. Finding order within disorder: elucidating the structure of proteins associated with neurodegenerative disease. *Future Medicinal Chem.* 1:467–482.
13. Tompa, P. 2011. Unstructural biology coming of age. *Curr. Opin. Struct. Biol.* 21:419–425.
14. Eliezer, D. 2009. Biophysical characterization of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* 19:23–30.

15. Fisher, C. K., and C. M. Stultz. 2011. Constructing ensembles for intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* 21:426–431.
16. Jensen, M. R., R. W. H. Ruigrok, and M. Blackledge. 2013. Describing intrinsically disordered proteins at atomic resolution by NMR. *Curr. Opin. Struct. Biol.* 23:426–435.
17. Marsh, J. A., C. Neale, ..., J. D. Forman-Kay. 2007. Improved structural characterizations of the drkN SH3 domain unfolded state suggest a compact ensemble with native-like and non-native structure. *J. Mol. Biol.* 367:1494–1510.
18. Nodet, G., L. Salmon, ..., M. Blackledge. 2009. Quantitative description of backbone conformational sampling of unfolded proteins at amino acid resolution from NMR residual dipolar couplings. *J. Am. Chem. Soc.* 131:17908–17918.
19. Dedmon, M. M., K. Lindorff-Larsen, ..., C. M. Dobson. 2005. Mapping long-range interactions in alpha-synuclein using spin-label NMR and ensemble molecular dynamics simulations. *J. Am. Chem. Soc.* 127:476–477.
20. Fisher, C. K., A. Huang, and C. M. Stultz. 2010. Modeling intrinsically disordered proteins with Bayesian statistics. *J. Am. Chem. Soc.* 132:14919–14927.
21. Lim, W. A. 2002. The modular logic of signaling proteins: building allosteric switches from simple binding domains. *Curr. Opin. Struct. Biol.* 12:61–68.
22. Levitt, M. 2009. Nature of the protein universe. *Proc. Natl. Acad. Sci. USA.* 106:11079–11084.
23. Ekman, D., A. K. Björklund, ..., A. Elofsson. 2005. Multi-domain proteins in the three kingdoms of life: orphan domains and other unsigned regions. *J. Mol. Biol.* 348:231–243.
24. Joerger, A. C., and A. R. Fersht. 2008. Structural biology of the tumor suppressor p53. *Annu. Rev. Biochem.* 77:557–582.
25. Terakawa, T., H. Kenzaki, and S. Takada. 2012. p53 searches on DNA by rotation-uncoupled sliding at C-terminal tails and restricted hopping of core domains. *J. Am. Chem. Soc.* 134:14555–14562.
26. Tafvizi, A., F. Huang, ..., A. M. van Oijen. 2011. A single-molecule characterization of p53 search on DNA. *Proc. Natl. Acad. Sci. USA.* 108:563–568.
27. Tidow, H., R. Melero, ..., A. R. Fersht. 2007. Quaternary structures of tumor suppressor p53 and a specific p53-DNA complex. *Proc. Natl. Acad. Sci. USA.* 104:12324–12329.
28. Bernadó, P. 2010. Effect of interdomain dynamics on the structure determination of modular proteins by small-angle scattering. *Eur. Biophys. J.* 39:769–780.
29. Kamerlin, S. C. L., S. Vicatos, ..., A. Warshel. 2011. Coarse-grained (multiscale) simulations in studies of biophysical and chemical systems. *Annu. Rev. Phys. Chem.* 62:41–64.
30. Hyeon, C., and D. Thirumalai. 2011. Capturing the essence of folding and functions of biomolecules using coarse-grained models. *Nat. Commun.* 2:487.
31. Takada, S. 2012. Coarse-grained molecular simulations of large biomolecules. *Curr. Opin. Struct. Biol.* 22:130–137.
32. Vuzman, D., and Y. Levy. 2012. Intrinsically disordered regions as affinity tuners in protein-DNA interactions. *Mol. Biosyst.* 8:47–57.
33. Higo, J., K. Umezawa, and H. Nakamura. 2013. A virtual-system coupled multiconformational molecular dynamics simulation: principles and applications to free-energy landscape of protein-protein interaction with an all-atom model in explicit solvent. *J. Chem. Phys.* 138:184106.
34. Matysiak, S., and C. Clementi. 2004. Optimal combination of theory and experiment for the characterization of the protein folding landscape of S6: How far can a minimalist model go? *J. Mol. Biol.* 343:235–248.
35. Kamiya, N., Y. S. Watanabe, ..., J. Higo. 2005. AMBER-based hybrid force field for conformational sampling of polypeptides. *Chem. Phys. Lett.* 401:312–317.
36. Cornell, W. D., P. Cieplak, ..., P. A. Kollman. 1996. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* 117:5179–5197.
37. Kollman, P., R. Dixon, ..., A. Pohorille. 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 Biomolecular Systems, Vol. 3.* van Gunsteren, W. F., P. K. Weiner, and A. J. Wilkinson, editors. Springer, New York, pp. 83–96.
38. Ikebe, J., N. Kamiya, ..., J. Higo. 2007. Conformational sampling of a 40-residue protein consisting of alpha and beta secondary-structure elements in explicit solvent. *Chem. Phys. Lett.* 443:364–368.
39. Ikebe, J., D. M. Standley, ..., J. Higo. 2011. Ab initio simulation of a 57-residue protein in explicit solvent reproduces the native conformation in the lowest free-energy cluster. *Protein Sci.* 20:187–196.
40. Ikebe, J., K. Umezawa, ..., J. Higo. 2011. Theory for trivial trajectory parallelization of multiconformational molecular dynamics and application to a polypeptide in water. *J. Comput. Chem.* 32:1286–1297.
41. Higo, J., Y. Nishimura, and H. Nakamura. 2011. A free-energy landscape for coupled folding and binding of an intrinsically disordered protein in explicit solvent from detailed all-atom computations. *J. Am. Chem. Soc.* 133:10448–10458.
42. Terakawa, T., and S. Takada. 2011. Multiscale ensemble modeling of intrinsically disordered proteins: p53 N-terminal domain. *Biophys. J.* 101:1450–1458.
43. Kenzaki, H., N. Koga, ..., S. Takada. 2011. CafeMol: a coarse-grained biomolecular simulator for simulating proteins at work. *J. Chem. Theory Comput.* 7:1979–1989.
44. Rippin, T. M., S. M. Freund, ..., A. R. Fersht. 2002. Recognition of DNA by p53 core domain and location of intermolecular contacts of cooperative binding. *J. Mol. Biol.* 319:351–358.
45. Natan, E., C. Baloglu, ..., A. C. Joerger. 2011. Interaction of the p53 DNA-binding domain with its n-terminal extension modulates the stability of the p53 tetramer. *J. Mol. Biol.* 409:358–368.
46. Li, W., T. Terakawa, ..., S. Takada. 2012. Energy landscape and multi-route folding of topologically complex proteins adenylate kinase and 2ouf-knot. *Proc. Natl. Acad. Sci. USA.* 109:17789–17794.
47. Chen, Y., R. Dey, and L. Chen. 2010. Crystal structure of the p53 core domain bound to a full consensus site as a self-assembled tetramer. *Structure.* 18:246–256.
48. Bista, M., S. M. Freund, and A. R. Fersht. 2012. Domain-domain interactions in full-length p53 and a specific DNA complex probed by methyl NMR spectroscopy. *Proc. Natl. Acad. Sci. USA.* 109:15752–15756.
49. Svergun, D., C. Barberato, and M. H. J. Koch. 1995. CRYSOLE—a program to evaluate x-ray solution scattering of biological macromolecules from atomic coordinates. *J. Appl. Cryst.* 28:768–773.
50. Rotkiewicz, P., and J. Skolnick. 2008. Fast procedure for reconstruction of full-atom protein models from reduced representations. *J. Comput. Chem.* 29:1460–1465.
51. Mittl, P. R., P. Chène, and M. G. Grütter. 1998. Crystallization and structure solution of p53 (residues 326–356) by molecular replacement using an NMR model as template. *Acta Crystallogr. D Biol. Crystallogr.* 54:86–89.
52. Daura, X., K. Gademann, ..., A. E. Mark. 1999. Peptide folding: when simulation meets experiment. *Angew. Chem.* 38:236–240.
53. Pronk, S., S. Páll, ..., E. Lindahl. 2013. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics.* 29:845–854.
54. Ahn, J., I.-J. L. Byeon, ..., A. M. Gronenborn. 2009. Insight into the structural basis of pro- and antiapoptotic p53 modulation by ASPP proteins. *J. Biol. Chem.* 284:13812–13822.
55. Kotta-Loizou, I., G. N. Tsaousis, and S. J. Hamodrakas. 2013. Analysis of molecular recognition features (MoRFs) in membrane proteins. *Biochim. Biophys. Acta.* 1834:798–807.
56. Piana, S., J. L. Klepeis, and D. E. Shaw. 2014. Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations. *Curr. Opin. Struct. Biol.* 24:98–105.

Supporting Information for

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

Tsuyoshi Terakawa¹, Junichi Higo², and Shoji Takada³

¹Department of Biophysics, Graduate School of Science, Kyoto University, Kitashirakawa-Oiwakecho, Sakyo, Kyoto, 6068502, Japan; ²Institute 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 V_{hybrid} 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 field 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 $[-102300.0 \text{ kcal/mol}, -69300.0 \text{ kcal/mol}]$. 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 $T = 300 \text{ K}$. 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 $\mathbf{f}_{mc}(r_i)$ acting on the atom i is given by

$$\mathbf{f}_{mc}(r_i) = RT \frac{d\ln[n(E)]}{dE} \mathbf{f}(r_i) \quad (\text{S3})$$

where R is gas constant, T is temperature, and $\mathbf{f}(r_i)$ 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 $\mathbf{f}_{mc}(r_i)$ in eq. S3 instead of $\mathbf{f}(r_i)$, ideally, we obtain a flat distribution of the potential energy (i.e., $P(E) \approx \text{const}$). 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

$$V_0 = V_{\text{without contact}} = V_{\text{bond}} + V_{\text{angle}} + V_{\text{ele}} + V_{\text{ex}} \quad (\text{S3})$$

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

$$V_{\text{bond}} = k_b (r_{ij} - b)^2 \quad (\text{S4})$$

where parameters were set as $k_b = 110.4$ (kcal/mol $\cdot\text{\AA}^2$) and $b = 3.8$ (\AA). r_{ij} is the length of

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

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

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

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

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, $\epsilon_k = 78.0$ is the dielectric constant, and κ_D is the Debye length defined as

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

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

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

where $\epsilon_{ex} = 0.2$ (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

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

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 Honeycutt 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,

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

where C is the concentration of the core domains (6.3 mM based on the radius of the sphere). We plot the calculated K_d s against the ϵ s 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

^aSimulation length (number of MD steps) for each of 128 runs.

^bThe 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]

^aEnergy 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

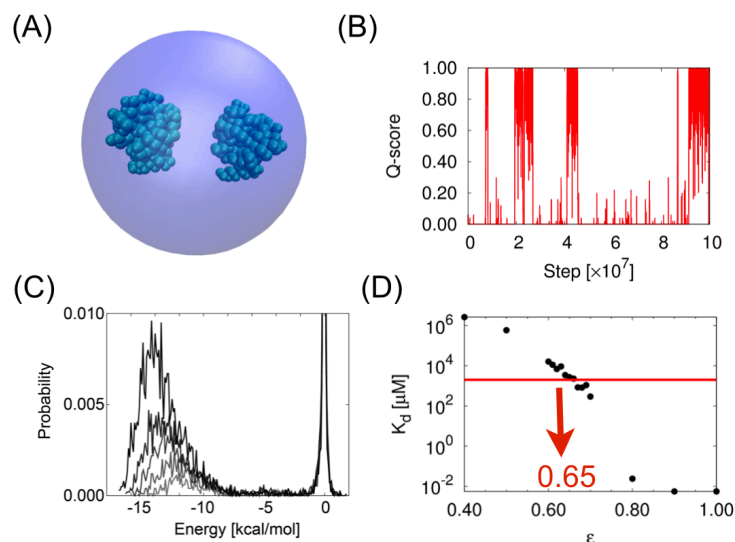


Fig. S1 The determination of the parameters of the inter-core-domain interaction. (A) The initial structure of the coarse-grained simulation for the determination of the ϵ in eq. 4. (B) The time trajectory of the inter-core-domain Q-score (Time is not physical time, but reduced time, i.e. time step of 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)

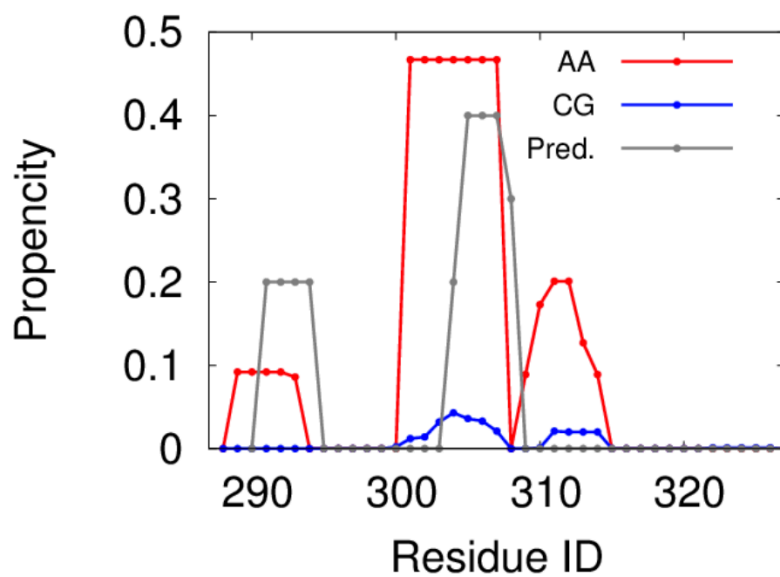


Fig. S2 Ahelical propensity (the population of structures with a helical structure) for each residue calculated from the obtained atomistic conformational ensemble (red) and estimated from only the amino acid sequence using the AGADIR (grey).

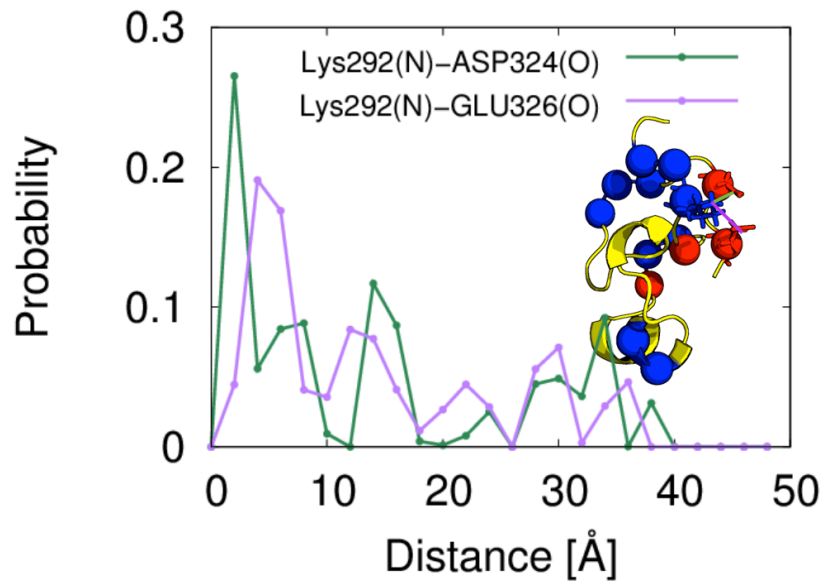


Fig. S3 The probability distributions of the distance between the nitrogen atom of LYS 292 and the oxygen atom of ASP 324 (Green) and between the nitrogen atom of LYS 292 and the oxygen atom of GLU 326 (purple). (Inset) The representative structures of the largest cluster. The blue (red) sphere represents Ca 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).

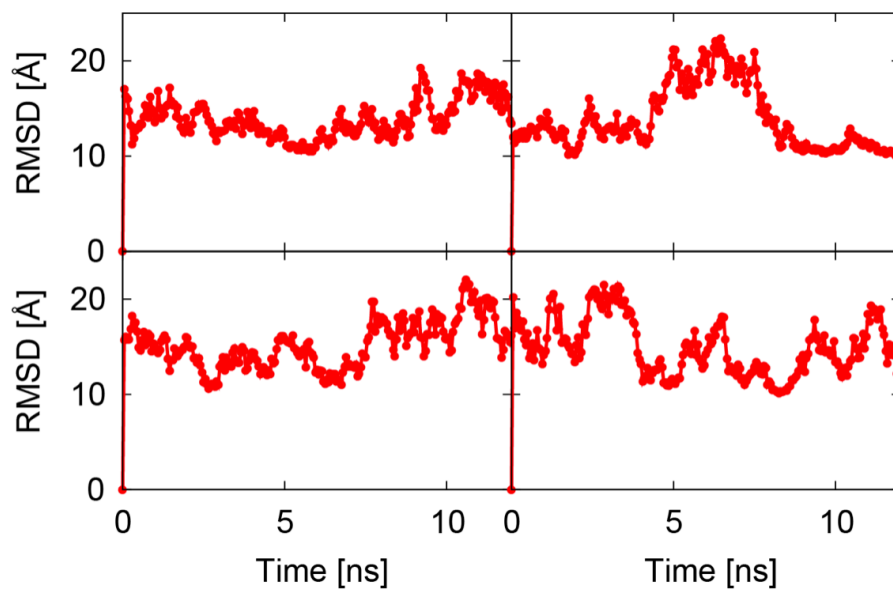


Fig. S4 Root mean square deviation of each snapshot in the representative AA V-McMD simulation. The reference structure is the initial structure of each production simulation.

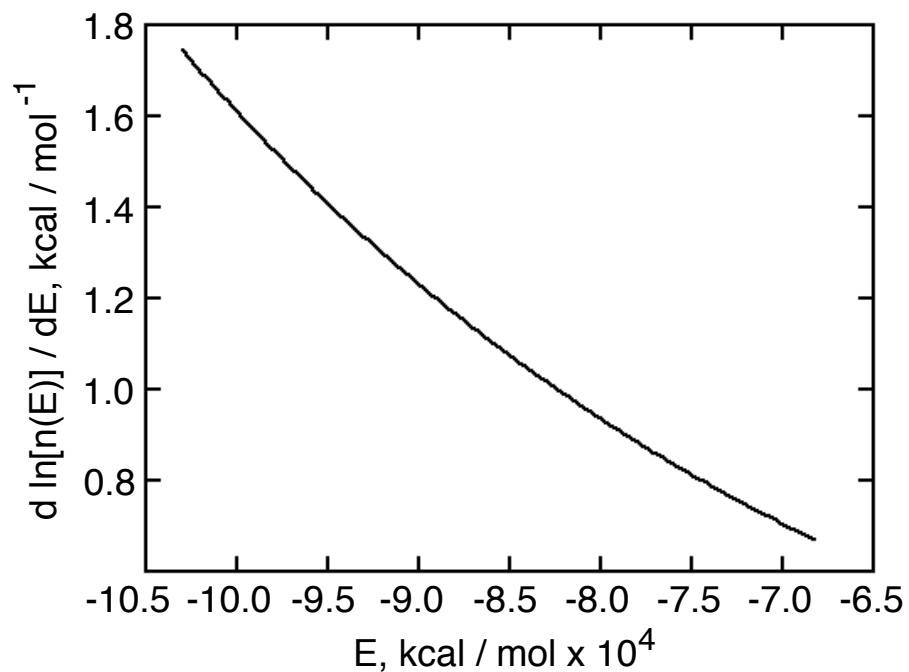


Fig. S5 The horizontal axis represents potential energy and the vertical axis represents the $d \ln[n](E) / dE$ in eq. S3

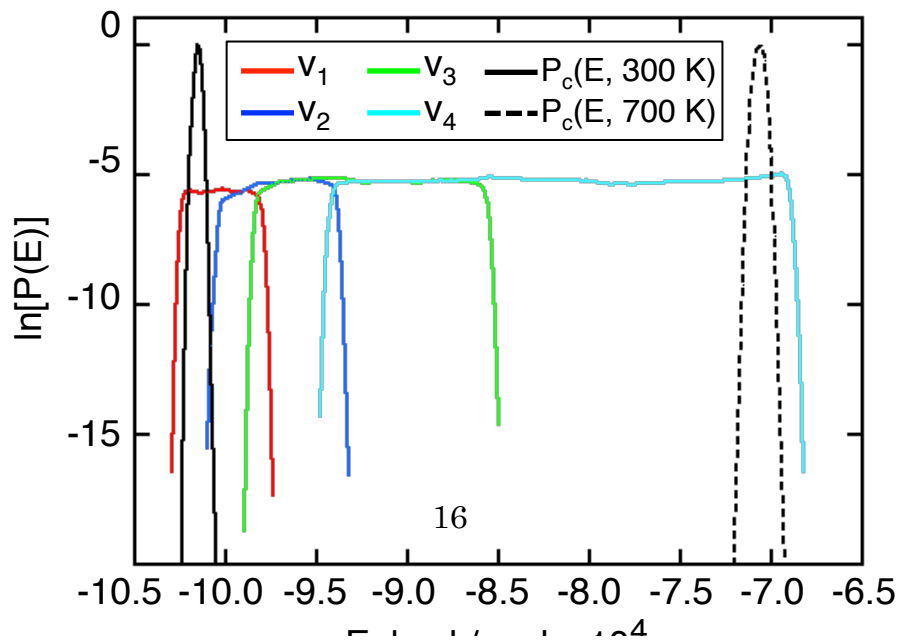


Fig. S6 Energy distributions in log scale. Each colored line represents the energy distribution in each virtual state (v_1 , v_2 , v_3 , and v_4). Black solid and dashed line represent the canonical energy distributions $P_c(E, T)$ at 300 K and at 700 K, respectively.