THE JOURNAL OF PHYSICAL CHEMISTRY B

Article

pubs.acs.org/JPCB

Comparative Molecular Dynamics Simulation Study of Crystal Environment Effect on Protein Structure

Tohru Terada*[,†,‡] and Akinori Kidera*[,†,§]

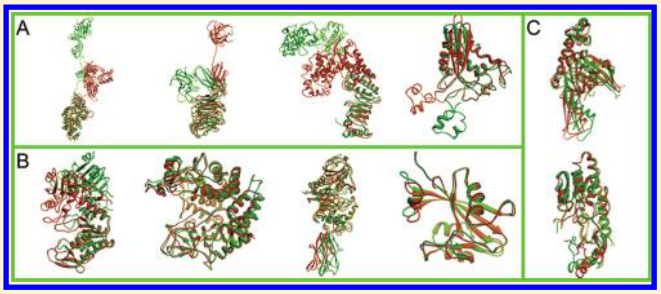
[†]Molecular Scale Team, Computational Science Research Program, RIKEN, 2-1 Hirosawa, Wako 351-0198, Japan

[‡]Agricultural Bioinformatics Research Unit, Graduate School of Agricultural and Life Sciences, The University of Tokyo, 1-1-1 Yayoi, Bunkyo-ku, Tokyo 113-8657, Japan

[§]Department of Supramolecular Biology, Graduate School of Nanobioscience, Yokohama City University, 1-7-29 Suehiro-cho, Yokohama 230-0045, Japan

Supporting Information

ABSTRACT: Crystal structures of proteins are under the influence from the crystal environment. In this study, we used molecular dynamics (MD) simulations to explore the possibility of eliminating the effect of the crystal packing and recovering the structure in solution. Ten representative proteins were chosen from the Protein Structural Change Database as the target systems, and 50 ns MD stimulations starting from two crystal structures having different domain arrangements were performed for each. The MD trajectories of the relaxation processes upon the release from the crystal environment revealed that the behaviors of the proteins were classified into three groups: "single domain linker", "harmonic motion", and "large barrier". We discuss the structural features common to the proteins in each group.



INTRODUCTION

The structure data in the Protein Data Bank (PDB)[1,2] provide information valuable for understanding the atomistic mechanism whereby a protein exerts its biological function. Nearly 90% of the protein structures in the PDB were determined by X-ray crystallography, but a protein structure in a crystalline state is liable to be affected by the crystal environment, that is, by nonphysiological contact with neighboring proteins in the crystal. Comparisons between protein structures determined in different crystal environments or with different methods (X-ray crystallography and nuclear magnetic resonance) have shown that the crystal environment significantly influences protein structure at both the backbone and the side-chain levels.[3−5] Nonphysiological contact has also been shown to affect the thermal fluctuations of protein atoms,[6] and the structural distortion it causes may severely limit our understanding of the mechanism of a protein's biological function. It is therefore important to know the extent to which a protein structure is affected by the crystal environment and, ultimately, to predict the solution structure from the crystal structure.

Amemiya et al. have recently developed the Protein Structural Change Database (PSCDB), in which is compiled a representative set of protein motions observed in pairs of crystal structures of identical proteins, one with a ligand molecule and the other without it.[7,8] The structural changes observed in 154 of the 528 PSCDB protein pairs exhibiting significant protein motion (72 exhibiting domain motion and 82 exhibiting local motion) are independent of ligand binding and are caused by change in the crystal environment.[7] These pairs of protein structures are a well-suited set for the analysis of the influence of the crystal environment on protein structures. Although the analyses of the crystal packing effects reported in ref 7 were successful to a certain extent, they used the elastic network model[9−11] and the linear response theory of protein structural change,[12,13] and these models are limited to the regime of the quasi-harmonic approximation.[14,15] Thus, they are not sufficient for fully analyzing the packing effects and for predicting the solution structures from the crystal structures. To go beyond the limitations of these analysis tools, in the present study we used molecular dynamics (MD) simulation.

To separate packing effects from crystal structures, we performed comparative MD simulations of 10 representative proteins whose structures had been determined in two different crystal environments. Comparative MD simulations have been used to extract characteristic structural and dynamical features from complex protein systems by searching for difference or similarity in multiple trajectories.[16−19] Here we used them to derive rules relating protein structural features to the crystal packing influence. Such rules are expected to be a guide to predicting solution structure from crystal structure.

From the PSCDB, we chose 10 representative pairs of proteins exhibiting domain motions caused by the change in

Special Issue: Harold A. Scheraga Festschrift

Received: December 29, 2011
Revised: March 2, 2012
Published: March 8, 2012

   dx.doi.org/10.1021/jp2125558 | J. Phys. Chem. B 2012, 116, 6810−6818

Table 1. Target Proteins for MD Simulations

ID[a]	protein name	PDB ID[b] structure 1	PDB ID[b] structure 2	rmsd[c] (Å)
CD.5	elongation factor 2	1n0v_D (2.85, 0.273)	1n0u_A (2.12, 0.254)	14.44
ID.31	cell division protein FtsA	1e4f_T (1.9, 0.249)	1e4g_T (2.6, 0.282)	1.92
ID.32	high-affinity zinc uptake system protein ZnuA	2ps3_A (2.47, 0.303)	2ps0_A (2.00, 0.250)	1.81
ID.38	sialidase	1eut_A (2.50, 0.214[d])	1w8n_A (2.1, 0.219)	1.66
ID.52	secreted effector protein	2qza_A (2.80, 0.291)	2qyu_A (2.10, 0.249)	4.72
ID.53	phosphoglycerate kinase 1	3c3b_A (1.80, 0.241)	2zgv_A (2.00, 0.258)	2.37
ID.56	pectate lyase	2v8i_A (1.50, 0.243)	2v8j_A (2.01, 0.247)	1.84
ID.59	chitinase A	3b8s_B (2.00, 0.205)	3b9d_A (1.72, 0.214)	1.71
ID.60	pseudecin	2ddb_A (1.90, 0.258)	2epf_D (2.30, 0.277)	1.53
ID.66	dATP pyrophosphohydrolase	2o1c_D (1.80, 0.289)	2o1c_A (1.80, 0.289)	1.27

[a]Protein identification number in the PSCDB.[8] ID means independent domain motion, and CD means coupled domain motion. Elongation factor 2 is labeled CD because the domain motion of one of its components (components 2) is coupled. Component 1, with independent domain motion, was considered in this study. [b]PDB and chain identification codes. Resolution in Å and free *R* value are shown in parentheses. [c]Cα rmsd between structures 1 and 2. [d]Working set *R* value.

crystal environment. The MD simulations were performed for each protein in explicit water environment, starting from two different structures each for 50 ns (1 μs in total). The 20 MD trajectories thus obtained were classified according to the degree of deviation from the initial structures and the degree of overlap between the structural ensembles produced by the two simulations. From the results of these analyses, we infer key structural features relating to the crystal packing influence.

COMPUTATIONAL METHODS

MD Simulations. The PSDB contains 72 protein pairs exhibiting domain motions coupled with change in the crystal environment.[7,8] We selected as the simulation targets 10 of them meeting the following conditions: monomeric, water-soluble, globular, and showing pure domain motions or no local loop motion. They are listed in Table 1, where the two structures of each protein are designated "structure 1" and "structure 2".

The initial structures for the MD simulations were prepared as follows. The structure of missing residues were modeled using Modeler 9v5.[20] When the missing residues were at the chain terminus and more than three residues were missing, the structures were modeled up to three residues whose terminus was capped by the acetyl group or the *N*-methyl group. The protonation states of histidine residues were determined by a 100 ps constant-pH MD simulation[21] at pH 7.0, 300 K, and a salt concentration of 0.1 M. This simulation used the generalized Born model[22] (the OBC I model[23,24]) and used a force constant of 41.8 kJ mol^{-1} Å^{-2} restraining the non-hydrogen atoms to their initial positions. The initial structure thus prepared was then immersed in a water box using the leap module of AmberTools.[25] Sodium or chloride ions needed to neutralize the system were added, but no additional small molecules were included in the simulations. The dimensions of the boxes were determined so that the distance from protein atoms to the closest boundary was at least 16 Å. Each system was energy-minimized and equilibrated in a constant-*NPT* MD simulation at 300 K and 1.0 × 10^5 Pa. The harmonic position restraint was imposed on Cα atoms and gradually reduced from 41.8 to 0 kJ mol^{-1} Å^{-2} in the process of equilibration. Finally, a constant-*NPT* MD simulation was performed for 50 ns at 300 K and 1.0 × 10^5 Pa, while the coordinates were recorded every 10 ps. System temperature was controlled using Langevin dynamics,[26,27] and system pressure was controlled using the weak coupling method.[28] The Amber ff99SB force-field parameters[29] were used for the proteins, and the TIP3P model[30] was used for water molecules. The particle mesh Ewald method[31,32] was used to calculate the electrostatic interactions. The bond lengths involving hydrogen atoms were constrained with the SHAKE algorithm[33,34] to allow use of a time step of 2 fs. All the MD simulations were conducted using the Sander or the PMEMD module of Amber 10.[25] Throughout the rest of this paper, the simulations started from structure 1 are referred to as "simulation 1", and the simulation started from structure 2 are referred to as "simulation 2".

Calculation of the Average Structure. The definition of the average structure in the structural ensemble is a crucial step for comparing structures and identifying protein motions. In this study, protein motions were recognized as the relative motion of the moving domain against the fixed domain. Therefore, the fixed domain was defined simultaneously with the definition of the average structure. This was done using the following method. The largest structural domain determined by DynDom[35,36] was defined as the initial guess for the fixed domain. Structures were superimposed on the first snapshot of the trajectory to match the positions of the Cα atoms in the fixed domain. The average structure was calculated by averaging the positions of the atoms of the superimposed structures. Then structure was again superimposed similarly on the average structure to calculate a new average structure. This step was repeated until the average structure converged. For the average structure thus obtained, the average deviation d_k of the k-th Cα atom from that of the average structure was calculated as

$$d_k^2 = \frac{1}{T}\sum_{t=1}^{T} |\mathbf{x}_{kt} - \langle \mathbf{x}_k \rangle|^2 \quad (1)$$

where $\mathbf{x}_{kt}$ denotes the coordinates of the k-th Cα taken from the t-th snapshot in the ensemble (t = 1, ..., T) and $\langle \mathbf{x}_k \rangle$ denotes the average coordinates. The Cα atoms whose average deviations were less than 1.5 Å define the fixed domain. Using this fixed domain, we calculated the average structure in the same way as explained above. This cycle was repeated twice. The average structure and the fixed domain obtained in the last cycle were used in the following analyses.

Principal Component Analysis. To clarify the conformational distribution in the ensemble, we performed principal component analysis (PCA) on the ensemble obtained from

dx.doi.org/10.1021/jp2125558 | J. Phys. Chem. B 2012, 116, 6810–6818

Table 2. Root Mean Square Deviations from the Initial Structures and the Degrees of Overlap between Structural Ensembles

ID	PDB ID[a]	rmsd[b] [Å]	PDB ID[a]	rmsd[b] [Å]	m[c]	overlap[d]
			Group A			
CD.5	1n0v_D	24.20 ± 0.98	1n0u_A	39.89 ± 1.53		
ID.38	1eut_A	27.15 ± 0.45	1w8n_A	7.41 ± 0.82		
ID.52	2qza_A	9.91 ± 0.55	2qyu_A	25.02 ± 0.51		
ID.60	2ddb_A	10.89 ± 0.62	2epf_D	3.21 ± 0.51		
			Group B			
ID.53	3c3b_A	1.93 ± 0.26	2zgv_A	4.85 ± 0.39	2	0.22, 0.11
ID.56	2v8i_A	1.78 ± 0.12	2v8j_A	1.52 ± 0.12	101	0.30, 0.35
ID.59	3b8s_B	4.20 ± 0.37	3b9d_A	3.06 ± 0.25	6	0.25, 0.38
ID.66	2o1c_D	2.15 ± 0.27	2o1c_A	2.20 ± 0.35	26	0.89, 0.86
			Group C			
ID.31	1e4f_T	2.11 ± 0.24	1e4g_T	2.14 ± 0.13	14	0.00, 0.00
ID.32	2ps3_A	1.81 ± 0.20	2ps0_A	2.66 ± 0.31	28	0.00, 0.00

[a]PDB and chain identification codes of the initial structures. [b]Average and standard deviation of rmsd's from the initial structure during the last 10 ns of the simulation. rmsd was calculated for all Cα atoms except those in the modeled regions. [c]The number of PCs used in the definition of overlap (eq 6). [d]The degree of overlap of the ensemble produced by simulation 1 (left) and simulation 2 (right).

each MD simulation and on the ensemble combining the trajectories of two MD simulations starting from the two structures of a protein. The PCA was based on the $3N \times 3N$ variance−covariance matrix $\mathbf{C}$ (= $\{C_{ij}\}$, $i, j = 1, ..., 3N$, N being the number of residues in a protein), whose element C_{ij} is given by

$$C_{ij} = \frac{1}{T}\sum_{t=1}^{T}(x_{it} - \langle x_i\rangle)(x_{jt} - \langle x_j\rangle) \quad (2)$$

The matrix was diagonalized with a unitary matrix $\mathbf{U}$ as

$$\mathbf{C} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^t \quad (3)$$

where $\mathbf{\Lambda}$ is a diagonal matrix containing $3N$ eigenvalues and $\mathbf{U}^t$ is the transpose of $\mathbf{U}$. The projection of the coordinates to the PCA space was given by

$$p_{lt} = \frac{1}{\sqrt{N}}\sum_{i=1}^{3N} u_{il}(x_{it} - \langle x_i\rangle) \quad (4)$$

where p_{lt} is the l-th component of the coordinates of t-th structure and u_{il} is the il-th element of matrix $\mathbf{U}$. The factor $1/\sqrt{N}$ was to derive the following relation between the squared sum of the projection and the rmsd

$$\sum_{l=1}^{3N} p_{lt}^2 = \frac{1}{N}\sum_{i=1}^{3N}(x_{it} - \langle x_i\rangle)^2 \quad (5)$$

This rmsd is slightly larger than the one calculated between the mutually superimposed structures.

To reduce noise and focus on domain motions, we measured the difference between two ensembles of a protein obtained in the two MD simulations starting from different structures by the rmsd defined by the m largest-amplitude principal components (PCs), where m was chosen as the smallest number satisfying that m PCs explained more than 90% of the total variance, as

$$(\mathrm{rmsd}_{st})^2 \approx \sum_{k=1}^{m}(p_{ks} - p_{kt})^2 \quad (6)$$

where rmsd_{st} is the value between the s-th structure of ensemble 1 and t-th structure of ensemble 2. The structure s is defined as having "overlap" with ensemble 2 if any structure t of ensemble 2 satisfies $\mathrm{rmsd}_{st} < 1$ Å. The degree of overlap between two ensembles was defined as a quantitative measure by the ratio of the number of overlapped structures to the total number of the structures in the ensemble. By definition, the degree of overlap can be defined either on ensemble 1 or on ensemble 2, and these two definitions are slightly different from each other (see Table 2).

Definition of the Causes of the Protein Structural Change. The method used in refs 7 and 8 to define the causes of the protein structural change in two crystal structures, one with and the other without a ligand molecule, is explained here briefly. The coupling between protein motions (structural difference between the two crystal structures) and ligand binding was identified by the location of the ligand binding site as in the definition of Qi and Hayward.[37] A protein motion was designated as "coupled" with ligand binding if the ligand molecule was located between the fixed and moving domains or if both the fixed and moving domains make contact with the ligand molecule in the ligand-bound form. When this criterion was not satisfied, the protein motion was designated "independent" of ligand binding. All the cases discussed in this study were those annotated in the PCSDB as independent of ligand binding.

The coupling with the crystal contact was determined for the protein pairs designated as independent in the following manner: First, two protein structures were respectively assigned open and closed on the basis of the distance between two domains. When the crystal environment in the open form was replaced by that in the closed form and any contact with neighboring molecules was found, the motion was classified as protein motion coupled with changes in the crystal environment.

RESULTS AND DISCUSSION

Classification of Proteins. One sees from results listed in Table 2 that four proteins showed large rmsd's (>10 Å) during the simulations, while the other six showed rmsd's less than 5 Å. These four proteins with large rmsd's were classified into group A. The remaining six were further examined to see whether the two ensembles generated by simulations that started from different crystal structures converged to a single distribution. This possibility was examined by calculating overlap between

dx.doi.org/10.1021/jp2125558 | J. Phys. Chem. B 2012, 116, 6810−6818

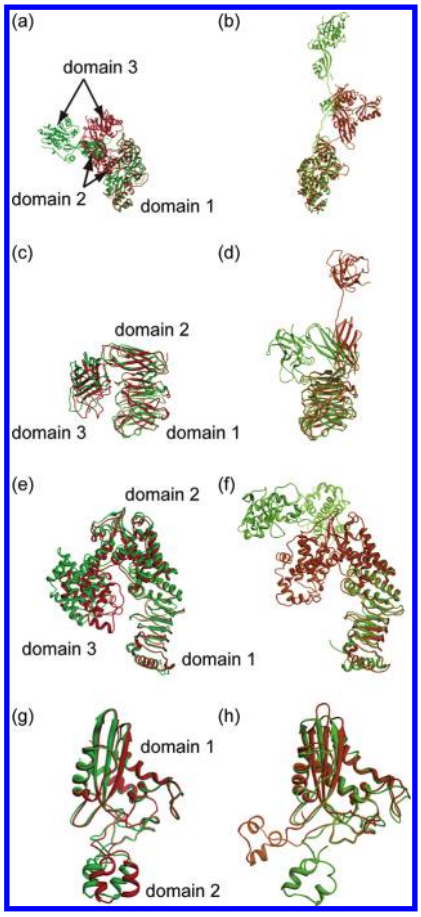


Figure 1. Crystal structures of (a) elongation factor 2 (CD.5), (c) sialidase (ID.38), (e) secreted effector protein (ID.52), (g) pseudecin (ID.60); structure 1 (red) and structure 2 (green). The final structures of the simulations starting from structure 1 (orange) and structure 2 (light green) of proteins (b) CD.5, (d) ID.52, (f) ID.38, and (h) ID.60. All structure images were generated with UCSF Chimera 1.5.3.[42]

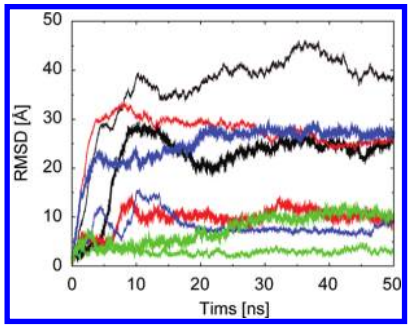


Figure 2. Time evolution of the Cα rmsd from the initial structure in simulation 1 (thin line) and simulation 2 (thick line) for elongation factor 2 (black), sialidase (blue), secreted effector protein (red), and pseudecin (green).

the pair of the structural ensembles (see Computational Methods). Four of the six proteins showed significant overlaps (>0.1), indicating that these proteins converged almost to a single basin within the simulation time, and were classified into group B. The other two proteins showed two different distributions depending on the initial structures and were classified into group C.

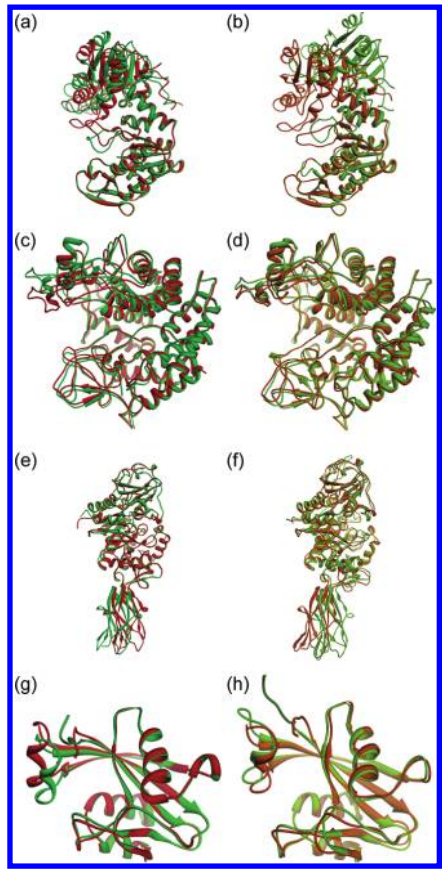


Figure 3. Crystal structures of structure 1 (red) and structure 2 (green) of (a) phosphoglycerate kinase 1 (ID.53), (c) pectate lyase (ID.56), (e) chitinase A (ID.59), and (g) dATP pyrophosphohydrolase (ID.66). The average structures of the last 10 ns of simulations starting from structure 1 (orange) and structure 2 (light green) for proteins (b) ID.53, (d) ID.56, (f) ID.59, and (h) ID.66.

Group A: Single Domain Linker. Proteins in group A showed large displacements from the initial structures during the 50 ns simulations, indicating that the domain configurations appearing in the crystal structures differ from those in solution. Since these proteins stably maintained each domain structure, the large motions are almost pure domain motions changing the mutual positions of the domains. Figure 1 shows the crystal structures and the average structures in the last 10 ns of the simulations. One sees that the crystal structures already show large differences (see rmsd values in Table 1), indicating that the domain arrangements of these proteins are susceptible to the change in the crystal environment. These differences were amplified when they were transferred to the solution environment.

The crystal structures of elongation factor 2 (CD.5) exhibit a motion of domain 2 (482−560) and domain 3 (561−841) against the N-terminal domain (domain 1: 2−481). In the simulations, the deviation was further enhanced. Particularly in simulation 2, domain 2 was further dissociated from domain 3. Sialidase (ID.38) shows the motion of domain 3 (506−647) against the N-terminal two domains (domain 1: 47−402; domain 2: 403−505) in the crystal. In solution, simulation 1 showed dissociation between domains 1 and 2. The secreted effector protein (ID.52) shows a motion of domain 3 (600−782) against the N-terminal two domains (domain 1: 163−386; domain 2: 387−599) in crystal. Simulation 1 showed the

dx.doi.org/10.1021/jp2125558 | J. Phys. Chem. B 2012, 116, 6810−6818

Table 3. Squared Correlation Coefficient between the Structural Difference between Crystal Structures and Large-Amplitude PCs

		CC^{2b}				CC^{2b}		
ID	PDB ID[a]	1st	2nd	sum	PDB ID[a]	1st	2nd	sum
				Group B				
ID.53	3c3b_A	0.42	0.15	0.57	2zgv_A	0.32	0.27	0.58
ID.56	2v8i_A	0.04	0.58	0.62	2v8j_A	0.29	0.40	0.69
ID.59	3b8s_B	0.00	0.83	0.83	3b9d_A	0.84	0.11	0.95
ID.66	2o1c_D	0.77	0.06	0.82	2o1c_A	0.69	0.03	0.72
				Group C				
ID.31	1e4f_T	0.20	0.03	0.22	1e4g_T	0.34	0.02	0.36
ID.32	2ps3_A	0.09	0.36	0.45	2ps0_A	0.06	0.00	0.06

[a]PDB and chain identification codes of the initial structure. [b]Square of the correlation coefficients of structural difference between crystal structures against first two principal axes calculated for the structural ensemble produced by the simulation starting from the structure of the corresponding PDB ID.

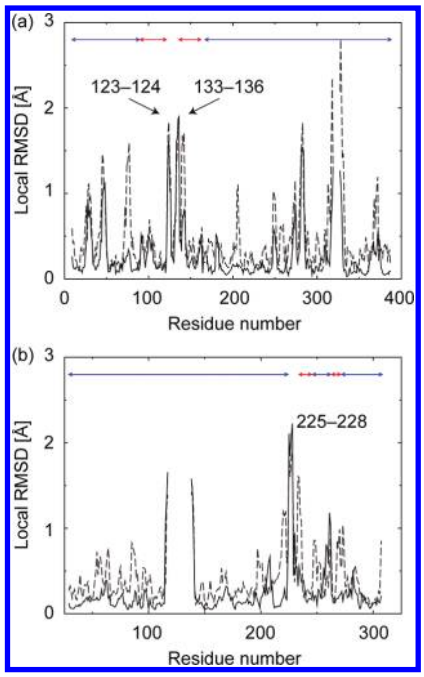


Figure 4. Plots of local rmsd's between the crystal structures (solid lines) and between average structures of last-10-ns trajectories of MD simulations (dashed lines) for (a) cell division protein FtsA (ID.31) and (b) high-affinity zinc uptake system protein ZnuA (ID.32). The local rmsd at residue i was calculated between Cα atoms of residues $i-2$, $i-1$, i, $i+1$, and $i+2$ of the two structures after superimposing the Cα atoms of one structure on those of the other structure. Ranges of fixed and moving domains (five or more residues long) assigned by DynDom for the crystal structures are indicated by arrows (blue for fixed domains and red for moving domains). Only residues that show local rmsd's larger than 1.5 Å in the moving domains (or in the regions between the fixed and moving domains) are indicated. Residues 321–326 and 118–137 are missing in FtsA and ZnuA.

structural ensemble similar to the conformational space defined by the two crystal structures, but simulation 2 largely separated the three domains from each other. Pseudecin (ID.60) shows the smallest rmsd value in the crystal structures of group A. The smallness of the domain motion (domain 2: 171–208) was caused by the ionic and hydrogen bond interactions between the two domains (between Tyr113/Lys114 and the C-terminal carboxyl group). In simulation 2, the interaction between Lys114 and the C-terminus was maintained, though the interaction involving Tyr113 was broken, so that the motion was suppressed to give a small rmsd. On the other hand, simulation 1 broke both interactions to cause a much larger motion of domain 2.

Figure 2 shows that the time course of the relaxation process upon the transfer from the crystal to the solution environment. One sees that the trajectories attained a plateau within 10 ns. This may imply that the domain–domain interactions observed in the crystal structures are not strong enough to maintain the domain contacts. The weak interdomain interactions were already evident in most of the crystal structures of group A because those domain configurations were easily changed by subtle changes in the crystal environment. The exception is Pseudecin (ID.60), which shows a longer time course of the structural change because of the ionic bonding side-chain interactions. At about 20 ns in simulation 1, however, those interactions were broken and caused a large deformation of the domain arrangement.

The MD simulations indicated that most domains of the proteins in group A tended to be isolated from the other domains and were likely to be fully solvated. The crystal environment appears to simply make the proteins take a compact form, thus making domain–domain contacts that are not stable in solution. As seen from the simulation structures (Figure 1), the domains of all four of these proteins are connected by single domain linkers. Each domain of such proteins is composed of a continuous region of the primary sequence. In contrast to the proteins in group A, those in groups B and C have multiple domain linkers. Consequently, when a protein shows a domain motion caused by the difference in the crystal environment and has a single domain linker between the two domains, it is predicted that these domains are fully flexible and work independently of each other in solution.

Group B: Harmonic Motion. The structures of the proteins in group B showed small deviations from the crystal structures (Table 2). From the beginning, the two crystal structures have small differences (Table 1). As indicated by the overlap degrees in Table 2, the 50 ns MD simulations succeeded in merging the two structural distributions produced by the two simulations started from the two crystal structures. Actually, the time required to attain the overlaps was very short, less than 4 ns in these four proteins. For illustrative purposes, structural comparisons are given in Figure 3, and detailed maps of the distributions projected onto the first and second principal components (PCs) are presented in Figure S1 (Supporting Information), where PCs were calculated with

dx.doi.org/10.1021/jp2125558 | J. Phys. Chem. B 2012, 116, 6810–6818

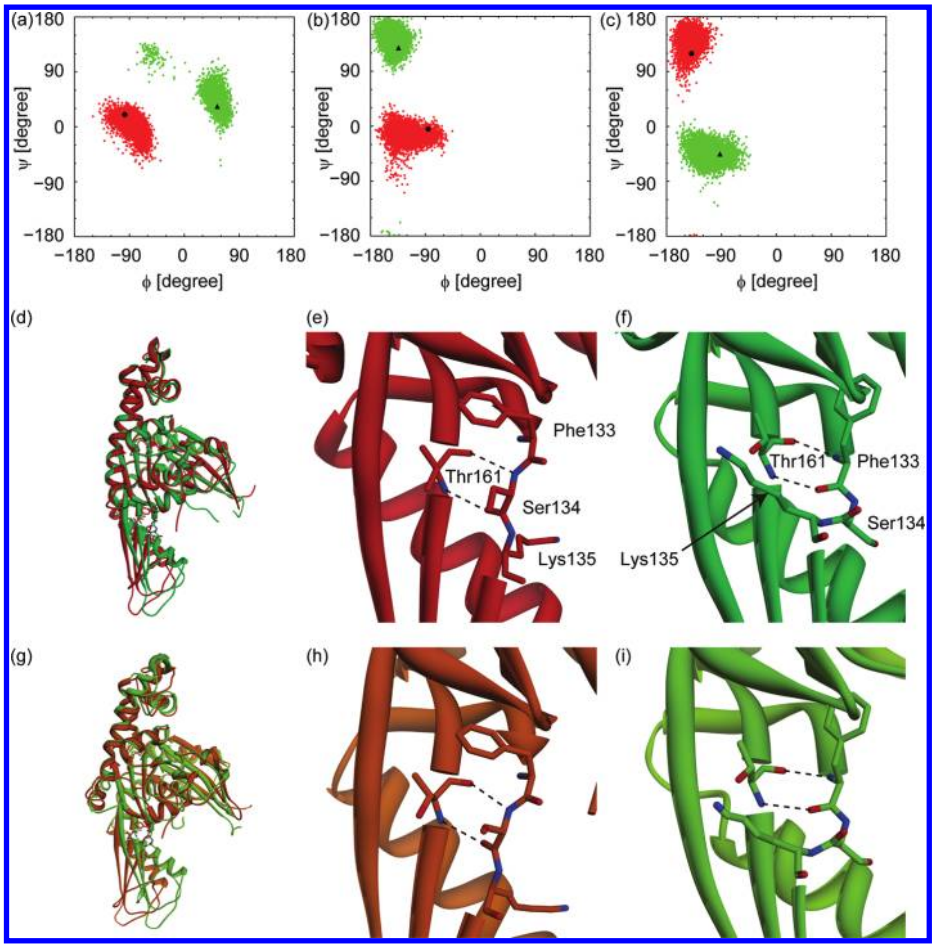


Figure 5. Distributions of $\phi-\psi$ dihedral angles of (a) Asn125, (b) Phe133, and (c) Lys135 of cell division protein FtsA (ID.31) in simulations 1 (red points) and 2 (green points). The $\phi-\psi$ dihedral angles of these residues of the crystal structures are indicated for structure 1 by black circles and for structure 2 by black triangles. Ribbon representations of the FtsA crystal structures of structures 1 (red) and 2 (green) (d) and their close-up images (e and f). Ribbon representations of the average structures of the last-10-ns trajectory of simulations 1 (orange) and 2 (light green) for FtsA (g) and their close-up images (h and i). The non-hydrogen atoms of Phe133, Ser134, Lys135, and Thr161 are shown in stick model. Carbon atoms are colored the same color as the ribbon and nitrogen, and oxygen atoms are colored red and blue. The main-chain hydrogen bonds are indicated by broken lines.

the combined trajectories of the two simulations. These overlaps were not only due to the smallness of the differences in the initial structures but also occurred because these differences in the crystal structures were in the space of the large-amplitude PCs. Table 3 shows the squared correlation coefficient between the structural difference between the two crystal structures and the two largest-amplitude PCs defined by each of the two MD simulations. It clearly indicates that the differences in the crystal structures are in the space of largely fluctuating domain motions observed in the simulations or within a single basin of the free energy surface.

From these analysis results, one can infer the following scenario for the formation of two different crystal structures: In the crystallization process, proteins are fluctuating largely in the space of the large-amplitude PCs observed in the simulations or with large domain motions. Nucleation occurs with a domain arrangement randomly chosen from the ensemble. Even in these proteins of group B, however, the surface of the basin for the domain motions is not perfectly smooth but has a certain ruggedness. Figure S1 (Supporting Information) shows that the ensembles generated by the two simulations are not perfectly merged into a single distribution but exhibit two distinct basins even though the barrier between these two basins is surmountable in a short-time simulation. This implies that the two structures in crystal may last long enough to stay in the different structures during the nucleation process, reflecting some ruggedness of the potential surface. If this time scale was short, the two crystal structures did not appear. The proteins in group C discussed below are those having much larger ruggedness in the free energy surface.

Therefore, the common feature to the proteins in group B is easiness of conformational transitions between the structures in the crystals. Although the correlation between the difference between the crystal structures and the large-amplitude PCs is only a necessary condition for this feature, it will be a good indicator.

Group C: Large Barrier. Although the cell division protein FtsA (ID.31) and the high-affinity zinc uptake system protein ZnuA (ID.32) showed small rmsd's from their initial structures during the MD simulations, their structural ensembles did not overlap after 50 ns (Table 2 and Figure S2, Supporting Information). To understand the causes that produced the barrier between the two structural ensembles, we focused on the local deformation of the backbone structures. Structural comparisons of five-residue structures, sliding a five-residue window along the primary sequence, are shown in Figure 4 for

dx.doi.org/10.1021/jp2125558 | J. Phys. Chem. B 2012, 116, 6810−6818

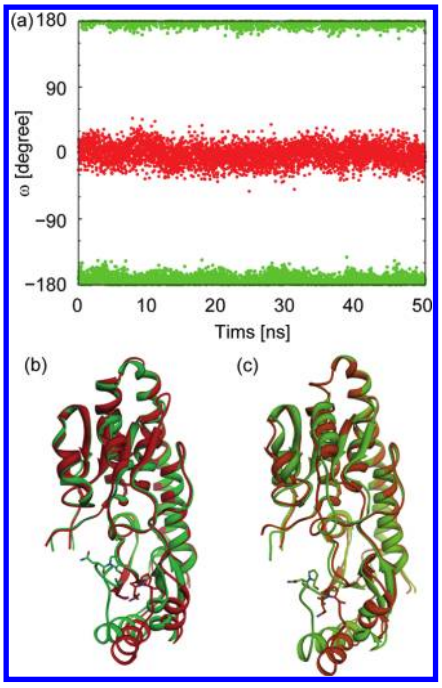


Figure 6. (a) Time evolutions of ω dihedral angles of the peptide bond between Asn228 and Pro229 of the high-affinity zinc uptake system protein ZnuA (ID.32) during simulations 1 (red points) and 2 (green points). (b) Ribbon representations of the ZnuA crystal structures of structures 1 (red) and 2 (green). (c) Ribbon representations of the average structures of the last-10-ns trajectory of simulations 1 (orange) and 2 (light green) for ZnuA. Structures of residues 118−137, which are missing in the crystal structures, are not shown. The non-hydrogen atoms of Asn228 and Prop229 are shown in stick model. Carbon atoms are colored the same color as the ribbon and nitrogen and oxygen atoms are colored red and blue.

the two crystal structures and for the average structures of the last-10-ns trajectories of the two simulations. The rmsd values remained large at residues 123−124 and 133−136 of FtsA and residues 225−228 of ZnuA. The large difference remaining at residues 282−283 of FtsA, certainly caused by the crystal packing effect, is not discussed here because it is irrelevant to the domain motion.

The crystal structures of FtsA, that Asn125 adopted a right-handed helical conformation in structure 1 and a left-handed helical conformation in structure 2 (Figure 5a), and Phe133 and Lys135, respectively, had helical and extended conformations in structure 1 but had different conformations in structure 2 (Figures 5b and 5c). These main-chain conformations were maintained during the simulation. It was found in a close analysis of the local region that these two conformations were stabilized by different hydrogen bond pairings in a β-sheet. The residue pairs for the hydrogen bond pairing are Ser134:Thr161 in structure 1 and Phe133:Thr161 in structure 2 (Figures 5d−5i). One residue shift in the β-sheet inverted the orientation of the side chains of residues 132−135 of the two structures. These differences caused the different conformations at residues 123−125 and 133−136.

The crystal structures of ZnuA appear to undergo the *cis*−*trans* isomerization of the peptide bonds between structure 1 and structure 2, *cis* to *trans* in Asn228 and *trans* to *cis* in Gln236. Asn228 precedes Pro229 along the sequence, and thus this is a prototype of *cis* peptide in X-Pro. The sequence Gln236-Arg237 is a nonproline *cis* peptide. Although the conformations of Asn228-Pro229 were maintained in the simulations (Figure 6a), the *cis* conformation of Gln236-Arg237 in structure 2 was converted into the *trans* conformation during the equilibration process. Thus, the local structural difference appears only in residues 225−228. Comparisons of the overall structures are shown in Figures 6b and 6c.

Substitution of the hydrogen bond pairs of a β-sheet, which occurred in FtsA, has to accompany unfolding of the β-sheet. The *cis*−*trans* isomerization of the peptide bond, found in ZnuA, may also require unfolding.[38,39] Therefore, the simulation of the native state cannot be expected to surmount these barriers and merge two conformational ensembles into a single distribution even if the simulation time were extended further. These two structures should be considered results of conformational polymorphism and may also be stable even in solution. As seen above these structural variations are not related to the domain motions and thus are not in the space of the large-amplitude PCs. This was indicated in the low correlation between the structural difference of the two crystal structures and the first two PCs obtained in each of the two MD simulations (Table 3). Therefore, the structural features in group C are the local distortions indicating a high potential barrier and are low correlation with the large amplitude PCs.

CONCLUDING REMARKS

Molecular dynamics simulations of the 10 representative pairs of protein structures, exhibiting significant domain motions caused by the change in crystal environment, revealed that the structures could be classified into three groups—A, "single domain linker"; B, "harmonic motion"; and C, "large barrier"—with regard to their response to the transfer from the crystal environment to the solution state. The results of the comparative simulations on the 10 proteins provided the information needed to decide which group each of these proteins belongs to. Group A is characterized by large displacement and a single domain linker, group B by a small displacement strongly correlated with the principal component, and group C by the large potential barrier indicated by some local distortions.

Here, we consider a possibility to use the elastic network model for the calculation of the PCs.[9−11] Since the elastic network model gives similar results to those of MD simulation when the amplitude of domain motion is small enough, the small motions in groups B and C will allow us to use the elastic network model for MD simulations. Table S1 (Supporting Information) shows the squared correlation coefficients between the structural changes in the crystal structures and the first two PCs derived by the elastic network model. It is shown that the elastic network model gives results consistent with those obtained in the MD simulations (Table 3) except for one case (FtsA (ID.31)). Therefore, prior to performing MD simulations, we can have a good estimation of the classification of a query protein, group B or group C, using the elastic network model. Once a protein is classified into group B, it is feasible to run MD simulations to predict its solution structure from the crystal structures.

Finally, let us consider the process of crystallization based on the two models of protein dynamics: the induced-fit model[40] and the population shift model.[41] Suppose a protein having two domains. The induced-fit scenario is as follows: this protein takes a certain stable domain configuration in solution. When it binds to the crystal surface, strong interactions from the crystal surface deform the domain arrangement to form a crystal

dx.doi.org/10.1021/jp2125558 | J. Phys. Chem. B 2012, 116, 6810−6818

structure. The way of deformation depends on the shape and the physical properties of the crystal surface. On the other hand, the population shift scenario presumes fluctuating domain configurations in solution. From a large variety of structures in solution is selected a configuration that fits the crystal surface. Largely populated species in solution may be selected for the crystal structure. The induced-fit model may require a large packing force to constrain the protein molecule, whereas the population shift model simply captures the configuration most dominant in solution. Group B is considered to be more closely related to the population shift scenario than Group A. Group C is neither the induced-fit nor the population shift, but it is obtained from two distinct structures generated by the different folding processes in each solution condition. However, we cannot completely rule out the possibility that the difference in Group C is simply due to a misassignment of the electron density of low resolution data.

■ ASSOCIATED CONTENT

Supporting Information

Correlation coefficients between the structural changes in the crystal structures and the two largest amplitude PCs derived by the elastic network model. Distributions of structures of proteins in groups B and C as projected onto the first and second PCs calculated with the combined trajectories of the two simulations. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: tterada@iu.a.u-tokyo.ac.jp (T.T.); kidera@tsurumi.yokohama-cu.ac.jp (A.K.).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was partly supported by Grants-in-Aid for Scientific Research (no. 23247027) from the Ministry of Education, Culture, Sports, Science, and Technology of Japan to A.K. and by Research and Development of the Next-Generation Integrated Simulation of Living Matter, a part of the Development and Use of the Next-Generation Supercomputer Project of the Ministry of Education, Culture, Sports, Science, and Technology of Japan. The computations were performed on the RIKEN Integrated Cluster of Clusters (RICC). We thank Dr. Takayuki Amemiya at Nagoya University for helpful discussion.

■ REFERENCES

(1) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(2) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F. Jr.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. *J. Mol. Biol.* **1977**, *112*, 535−542.
(3) Andrec, M.; Snyder, D. A.; Zhou, Z.; Young, J.; Montelione, G. T.; Levy, R. M. *Proteins* **2007**, *69*, 449−465.
(4) Eyal, E.; Gerzon, S.; Potapov, V.; Edelman, M.; Sobolev, V. *J. Mol. Biol.* **2005**, *351*, 431−442.
(5) Jacobson, M. P.; Friesner, R. A.; Xiang, Z.; Honig, B. *J. Mol. Biol.* **2002**, *320*, 597−608.
(6) Hinsen, K. *Bioinformatics* **2008**, *24*, 521−528.
(7) Amemiya, T.; Koike, R.; Fuchigami, S.; Ikeguchi, M.; Kidera, A. *J. Mol. Biol.* **2011**, *408*, 568−584.
(8) Amemiya, T.; Koike, R.; Kidera, A.; Ota, M. *Nucleic Acids Res.*, in press.
(9) Atilgan, A. R.; Durell, S. R.; Jernigan, R. L.; Demirel, M. C.; Keskin, O.; Bahar, I. *Biophys. J.* **2001**, *80*, 505−515.
(10) Hinsen, K. *Proteins* **1998**, *33*, 417−429.
(11) Tirion, M. M. *Phys. Rev. Lett.* **1996**, *77*, 1905−1908.
(12) Ikeguchi, M.; Ueno, J.; Sato, M.; Kidera, A. *Phys. Rev. Lett.* **2005**, *94*, 078102.
(13) Omori, S.; Fuchigami, S.; Ikeguchi, M.; Kidera, A. *J. Comput. Chem.* **2009**, *30*, 2602−2608.
(14) Brooks, C. L.; Karplus, M.; Pettitt, B. M. *Proteins: A Theoretical Perspective of Dynamics, Structure, and Thermodynamics*; J. Wiley: New York, 1988.
(15) Fuchigami, S.; Fujisaki, H.; Matsunaga, Y.; Kidera, A. Protein Functional Motions: Basic Concepts and Computational Methodologies. In *Advancing Theory for Kinetics and Dynamics of Complex, Many-Dimensional Systems: Clusters and Proteins*; John Wiley & Sons, Inc.: New York, 2011; pp 35−82.
(16) Hashido, M.; Ikeguchi, M.; Kidera, A. *FEBS Lett.* **2005**, *579*, 5549−5552.
(17) Lama, D.; Sankararamakrishnan, R. *Proteins* **2008**, *73*, 492−514.
(18) Ozboyaci, M.; Gursoy, A.; Erman, B.; Keskin, O. *PLoS ONE* **2011**, *6*, e14765.
(19) Rui, H.; Lee, J.; Im, W. *Biophys. J.* **2009**, *97*, 787−795.
(20) Šali, A.; Blundell, T. L. *J. Mol. Biol.* **1993**, *234*, 779−815.
(21) Mongan, J.; Case, D. A.; McCammon, J. A. *J. Comput. Chem.* **2004**, *25*, 2038−2048.
(22) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127−6129.
(23) Onufriev, A.; Bashford, D.; Case, D. A. *Proteins* **2004**, *55*, 383−394.
(24) Tsui, V.; Case, D. A. *Biopolymers (Nucleic Acid Sci.)* **2001**, *56*, 275−291.
(25) Case, D. A.; Darden, T. A.; Cheatham, T. E., III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Crowley, M.; Walker, R. C.; Zhang, W.; et al. *AMBER 10*; University of California: San Francisco, 2008.
(26) Pastor, R. W.; Brooks, B. R.; Szabo, A. *Mol. Phys.* **1988**, *65*, 1409−1419.
(27) Loncharich, R. J.; Brooks, B. R.; Pastor, R. W. *Biopolymers* **1992**, *32*, 523−535.
(28) Berendsen, H. J. C.; Postma, J. P. M; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684−3690.
(29) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins* **2006**, *65*, 712−725.
(30) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926−935.
(31) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089−10092.
(32) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577−8593.
(33) Miyamoto, S.; Kollman, P. A. *J. Comput. Chem.* **1992**, *13*, 952−962.
(34) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327−341.
(35) Hayward, S.; Berendsen, H. J. C. *Proteins* **1998**, *30*, 144−154.
(36) Hayward, S.; Lee, R. A. *J. Mol. Graphics Modell.* **2002**, *21*, 181−183.
(37) Qi, G.; Hayward, S. *BMC Struct. Biol.* **2009**, *9*, 13.
(38) Brandts, J. F.; Halvorson, H. R.; Brennan, M. *Biochemistry* **1975**, *14*, 4953−4963.
(39) Wedemeyer, W. J.; Welker, E.; Scheraga, H. A. *Biochemistry* **2002**, *41*, 14637−14644.
(40) Koshland, D. E. *Proc. Natl. Acad. Sci. U.S.A.* **1958**, *44*, 98−104.
(41) Kumar, S.; Ma, B.; Tsai, C.-J.; Sinha, N.; Nussinov, R. *Protein Sci.* **2000**, *9*, 10−19.

(42) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. *J. Comput. Chem.* **2004**, *25*, 1605−1612.