Observation of helix associations for insertion of a retinal molecule and distortions of helix structures in bacteriorhodopsin

Ryo Urano, and Yuko Okamoto

Citation: J. Chem. Phys. **143**, 235101 (2015); doi: 10.1063/1.4935964
View online: https://doi.org/10.1063/1.4935964
View Table of Contents: http://aip.scitation.org/toc/jcp/143/23
Published by the American Institute of Physics

Observation of helix associations for insertion of a retinal molecule and distortions of helix structures in bacteriorhodopsin

Ryo Urano[1,a)] and Yuko Okamoto[1,2,3,4,5,b)]
[1]*Department of Physics, Graduate School of Science, Nagoya University, Nagoya, Aichi 464-8602, Japan*
[2]*Center for Computational Science, Graduate School of Engineering, Nagoya University, Nagoya, Aichi 464-8603, Japan*
[3]*Structural Biology Research Center, Graduate School of Science, Nagoya University, Nagoya, Aichi 464-8602, Japan*
[4]*Information Technology Center, Nagoya University, Nagoya, Aichi 464-8601, Japan*
[5]*JST-CREST, Nagoya, Aichi 464-8602, Japan*

(Received 11 August 2015; accepted 28 October 2015; published online 15 December 2015)

We applied a newly proposed prediction method for membrane protein structures to bacteriorhodopsin that has distorted transmembrane helices in the native structure. This method uses an implicit membrane model, which restricts sampling space during folding in a membrane region, and includes helix bending. Replica-exchange simulations were performed with seven transmembrane helices only without a retinal molecule. Obtained structures were classified into clusters of similar structures, which correspond to local-minimum free energy states. The two lowest free energy states corresponded to a native-like structure with the correct empty space for retinal and a structure with this empty space filled with a helix. Previous experiments of bacteriorhodopsin suggested that association of transmembrane helices enables them to make a room for insertion of a retinal. Our results are consistent with these results. Moreover, distortions of helices in the native-like structures were successfully reproduced. In the distortions, whereas the locations of kinks for all helices were similar to those of Protein Data Bank's data, the amount of bends was more similar for helices away from the retinal than for those close to the retinal in the native structure. This suggests a hypothesis that the amino-acid sequence specifies the location of kinks in transmembrane helices and that the amount of distortions depends on the interactions with the surrounding molecules such as neighboring helices, lipids, and retinal. © *2015 AIP Publishing LLC.* [http://dx.doi.org/10.1063/1.4935964]

I. INTRODUCTION

Membrane proteins, in particular, α-helical integral membrane proteins, have important biological functions such as channel, transporter, and receptor. Although membrane protein folding has been studied for a long time,[1,2] it is difficult to determine their tertiary structures because of their complicated biomembrane environment. On the other hand, membrane proteins encoded in genomes are estimated to be about 30% of the total proteins by bioinformatics.[3] While more than a hundred thousand water soluble proteins are registered in the Protein Data Bank (PDB), only a few thousands membrane proteins are registered making up only about 2% in database.[4–6]

Computer simulations, on the other hand, have been developed in both algorithms and computational resources for decades. In the software, their efforts resulted in improvement of efficiency and precision. In particular, replica-exchange method (REM),[7,8] which is also referred to as parallel tempering, and its variants[9–11] greatly enhance efficiency of conformational sampling (for a review, see, e.g., Ref. 12).

However, it is still very difficult to perform simulations long enough to reach an experimental time scale for larger membrane proteins in general, and some implicit membrane model is in need. Because biomembranes have various characteristic properties such as chain order orientations of lipid molecules, pressure profile, free energy profile, and dielectric constant profile, there are many choices to make an implicit membrane model. Popular implicit membrane models such as HDGB model, DHDGB model, and IMM1 model are often focused on the free energy profile as a physical quantity for reducing system size and relaxation costs.[13–16]

Here, we used another implicit membrane model,[17–21] where four elementary harmonic restraints are added to the original Chemistry at Harvard Molecular Mechanics (CHARMM) potential energy function in order to mimic restrained membrane environment. Only the transmembrane helices are used in our simulations, and loop regions of the membrane proteins as well as lipid and water molecules were neglected. While membrane environment enhances the stability of helix structures, it also restricts sampling in conformational space in the membrane region. Thus, this model greatly reduces the search area in the conformational space during folding processes. This model is supported by many experimental data such as two-stage model (for a review, see Ref. 22). As for a simulation method, we employed the replica-exchange Monte Carlo (MC).

[a)]Current address: Chemistry Department, Boston University, 590 Commonwealth Avenue, Boston, Massachusetts 02215, USA.
[b)]Author to whom correspondence should be addressed. Electronic mail: okamoto@phys.nagoya-u.ac.jp. R. Urano and Y. Okamoto contributed equally to this work.

Although the previous method treated helix structures as rigid bodies and required the known helix structures from the native structure, this approach reproduced native-like structures of bacteriorhodopsin,[20,21] which has seven transmembrane helices with 249 amino acids and a retinal from Halobacterium salinarum (PDB ID: 1IW6,[23] 1PY6,[24] 1BRR[25]). This protein has a function of proton pump in bio-membrane with excitations of the retinal molecule by light. However, 60% of all transmembrane helix structures are distorted in PDB, which seems to be related to their functions. Thus, we also focus our interests in this work on the reproduction of their distortions in larger proteins such as bacteriorhodopsin which has seven distorted helices

To solve this problem, we extended the above method to include the flexible treatment of backbone structures and reproduced small membrane protein structures: glycophorin A of a transmembrane helix dimer and phospholamban of a distorted transmembrane helix.[26] We therefore applied this new method to introduce deformations of helix backbone structures of bacteriorhodopsin from initial ideal helix structures.

This article is organized as follows. In Section II, we present the method and simulation conditions. In Section III, we show results of the REM simulations of bacteriorhodopsin. Finally, Section IV is devoted to a discussion based on the results. After we confirm the results of our reproduction and prediction, we propose a hypothesis about the relation of minimum free energy states during the REM simulations and bends of transmembrane helices.

II. METHODS

We first obtain the amino-acid sequences of transmembrane helices of the target protein by bioinformatics servers such as SOSUI,[27] TMHMM,[28] MEMSAT,[29] and HMMTOP.[30] In the present work, however, the amino-acid sequences of transmembrane helices were taken from the experimental structure as in the previous work.[20,21] Different servers may predict different helix ends. In order to study end dependence, we also made shorter simulations with a little shorter helices and confirmed that we obtained similar results. We then perform REM simulations of these transmembrane helices. Our procedure for membrane protein structure predictions is summarized in Fig. 1.[26] The MC program is based on CHARMM macromolecular mechanics program,[31,32] and replica-exchange Monte Carlo method was implemented in it. Only the transmembrane helices were used in our simulations, and loop regions of the membrane proteins as well as lipid and water molecules were neglected. The membrane environment for this protein for the membrane thickness and the region of transmembrane region of the helices was taken from Orientation of Proteins in Membrane (OPM).[33] The amino-acid sequences of the transmembrane helices are EWIWLALGTALMGLGTLYFLVKG (9-31), KFYAITTLVPAIAFTMYLSMLL (41-62), IYWARYADWLFTTPLLLLDLALL (78-100), QGTILALVGADGIMIGTGLVGAL (105-127), RFVWWAISTAAMLYILYVLFFGF (134-156), TFKVLRNVTVVLWSAYPVVWLIGSE (170-194), and LNIETLLFMVLDVSAKVGFGLILL (201-224), which are the same as in Refs. 20 and 21, where two numbers in parentheses after each sequence correspond to the first and last amino-acid numbers of each helix. The N-terminus and the C-terminus of each helix were blocked with the acetyl group and the N-methyl group (first residue and last residue), respectively. The initial structure for each helix was an ideal helix structure and they were placed in the membrane region randomly.

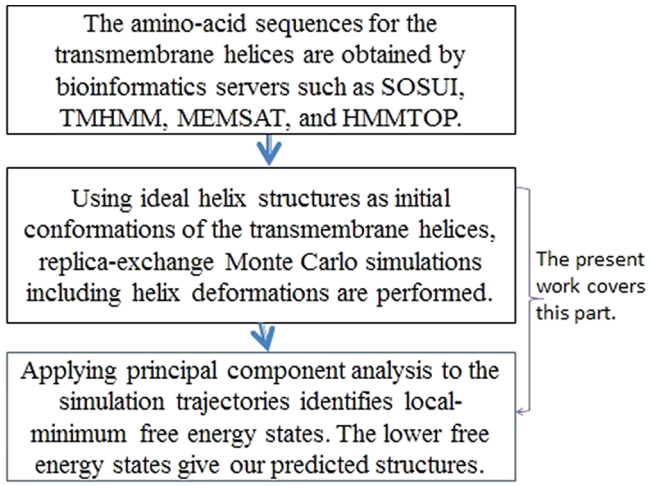

FIG. 1. Our prediction procedure for membrane protein structures.

We added the following four elementary harmonic restraints as a simple implicit membrane model to the original CHARMM potential energy function of bond length, bond angle, torsion angle, van der Waals, electrostatic interaction, etc., in order to mimic the restrained membrane environment. The restraint energy function is given by

$$E_{\mathrm{restr}} = E_{\mathrm{restr1}} + E_{\mathrm{restr2}} + E_{\mathrm{restr3}} + E_{\mathrm{restr4}}, \tag{1}$$

where each term is defined as follows:

$$E_{\mathrm{restr1}} = \sum_{i=1}^{N_{\mathrm{H}}-1} k_1\, \theta\left(r_{i,i+1} - d_{i,i+1}\right)\left[r_{i,i+1} - d_{i,i+1}\right]^2, \tag{2}$$

$$E_{\mathrm{restr2}} = \sum_{i=1}^{N_{\mathrm{H}}} \left\{ k_2\, \theta\left(\left|z_i^{\mathrm{L}} - z_{0,i}^{\mathrm{L}}\right| - d^{\mathrm{L}}\right)\left[\left|z_i^{\mathrm{L}} - z_{0,i}^{\mathrm{L}}\right| - d^{\mathrm{L}}\right]^2 + k_2\, \theta\left(\left|z_i^{\mathrm{U}} - z_{0,i}^{\mathrm{U}}\right| - d^{\mathrm{U}}\right)\left[\left|z_i^{\mathrm{U}} - z_{0,i}^{\mathrm{U}}\right| - d^{\mathrm{U}}\right]^2 \right\}, \tag{3}$$

$$E_{\mathrm{restr3}} = \sum_{\mathrm{C}_\alpha} k_3\, \theta\left(r_{\mathrm{C}_\alpha} - d_{\mathrm{C}_\alpha}\right)\left[r_{\mathrm{C}_\alpha} - d_{\mathrm{C}_\alpha}\right]^2, \tag{4}$$

$$E_{\mathrm{restr4}} = \sum_{j=1}^{N_{BD}} k_4\theta(|\,\phi_j - \phi_0\,| - \alpha_j^{\phi})[|\,\phi_j - \phi_0\,| - \alpha_j^{\phi}]^2 + \sum_{j=1}^{N_{BD}} k_5\theta(|\,\psi_j - \psi_0\,| - \alpha_j^{\psi})[|\,\psi_j - \psi_0\,| - \alpha_j^{\psi}]^2. \tag{5}$$

E_{restr1} is the energy that restrains pairs of adjacent helices along the amino-acid chain not to be apart from each other too much (loop restraints), where $r_{i,i+1}$ is the distance between the C atom of the C-terminus of the ith helix and the C^{α} atom of

TABLE I. Acceptance ratios of replica exchange corresponding to pairs of neighboring temperatures from the REM simulations.

Pairs of T	Acceptance ratio	Pairs of T	Acceptance ratio
400 ⟷ 415	0.467	1369 ⟷ 1460	0.232
415 ⟷ 435	0.382	1460 ⟷ 1558	0.222
435 ⟷ 455	0.405	1558 ⟷ 1662	0.200
455 ⟷ 485	0.253	1662 ⟷ 1774	0.201
485 ⟷ 518	0.247	1774 ⟷ 1892	0.207
518 ⟷ 552	0.269	1892 ⟷ 2019	0.210
552 ⟷ 589	0.245	2019 ⟷ 2154	0.227
589 ⟷ 629	0.244	2154 ⟷ 2298	0.230
629 ⟷ 671	0.260	2298 ⟷ 2452	0.266
671 ⟷ 716	0.259	2452 ⟷ 2616	0.257
716 ⟷ 764	0.247	2616 ⟷ 2791	0.267
764 ⟷ 815	0.248	2791 ⟷ 2978	0.281
815 ⟷ 870	0.237	2978 ⟷ 3177	0.265
870 ⟷ 928	0.229	3177 ⟷ 3390	0.246
928 ⟷ 990	0.231	3390 ⟷ 3616	0.225
990 ⟷ 1056	0.251	3616 ⟷ 3808	0.265
1056 ⟷ 1127	0.240	3808 ⟷ 4050	0.184
1127 ⟷ 1202	0.232	4050 ⟷ 4250	0.162
1202 ⟷ 1283	0.223	4250 ⟷ 4500	0.073
1283 ⟷ 1369	0.210		

the N-terminus of the $(i + 1)$th helix, and k_1 and $d_{i,i+1}$ are the force constant and the central value constant of the harmonic restraints, respectively. Each $d_{i,i+1}$ is proportional to the loop length connected between helices. $\theta(x)$ is the step function, which is 1 when x is larger than or equal to 0, otherwise zero. N_{H} is the total number of transmembrane helices in the protein.

E_{restr2} is the energy that restrains helix N-terminus and C-terminus to be located near membrane boundary planes. Here, the z-axis is defined to be the direction perpendicular to the membrane boundary planes. k_2 is the force constant of the harmonic restraints. $z^{\mathrm{L}}_{0,i}$ and $z^{\mathrm{U}}_{0,i}$ are the z-coordinate values of the C^{α} atom of the N-terminus or C-terminus of the ith helix near the fixed lower membrane boundary and the upper membrane boundary, respectively. $z^{\mathrm{L}}_{0,i}$ and $z^{\mathrm{U}}_{0,i}$ are the fixed lower boundary z-coordinate value and the upper boundary z-coordinate value of the membrane planes, respectively, and here they depend on each helix atoms due to the known data from OPM[33] although constant membrane plane region is also possible like a previous research condition. d^{L} and d^{U} are the corresponding central value constants of the harmonic restraints. This term has a non-zero value only when the C^{α} atoms of the N-terminus or C-terminus of the ith helix are apart more than d^{L}_i (or d^{U}_i). This restraint energy was introduced so that the helix ends are not too much apart from the membrane boundary planes.

E_{restr3} is the energy that restrains all C^{α} atoms within the sphere (centered at the origin) of radius $d_{C^{\alpha}}$. $r_{C^{\alpha}}$ is the distance of C^{α} atoms from the origin, and k_3 and $d_{C^{\alpha}}$ are the force constant and the central value constant of the harmonic restraints, respectively.

E_{restr4} is the energy that restraints the dihedral angles of the main chain so that helix structures may not deviate too much from ideal helix structures, preventing them from forming random-coil structures. N_{BD} is the total number of (ϕ, ψ) angles in the helix backbones. Here, all the backbone dihedral angles ϕ_j and ψ_j $(j = 1, \ldots, N_{BD})$ are restrained. ϕ_0 and ψ_0 are the reference value of the harmonic restraint to keep the helix structures without forming random coil structure, and α^{ϕ}_j and α^{ψ}_j are the ranges of the harmonic restraints.

We set $k_1 = 5.0$, $d_{i,i+1} = (46, 53, 34, 19, 95, 30)$, where $i = 1, 2, \ldots, 6$, $k_2 = 5.0$, $z^{\mathrm{L}}_{0,i} = (-14, -16, -20, -15, -19, -24, -18)$, where $i = 1, 2, \ldots, 7$, $z^{\mathrm{U}}_{0,i} = (12, 14, 15, 15, 14, 11, 12)$, where $i = 1, 2, \ldots, 7$, $d^{\mathrm{U}} = d^{\mathrm{L}} = 2.0$, $k_3 = 0.5$, $d_{C^{\alpha}} = 80$, $k_4 = 30.0$, $k_5 = 30.0$, $\phi_0 = -62$, $\psi_0 = -40$, $\alpha^{\phi}_j = 16$, and $\alpha^{\psi}_j = 13$.

TABLE II. Acceptance ratios (ARs) of replica exchange corresponding to pairs of neighboring temperatures in four REM simulations.

Pairs of T	AR: simulation 1/2/3/4	Pairs of T	AR: simulation 1/2/3/4
400 ⟷ 415	0.413/0.430/0.471/0.523	1369 ⟷ 1460	0.232/0.285/0.212/0.218
415 ⟷ 435	0.362/0.366/0.370/0.415	1460 ⟷ 1558	0.221/0.249/0.198/0.225
435 ⟷ 455	0.400/0.400/0.389/0.424	1558 ⟷ 1662	0.250/0.145/0.168/0.222
455 ⟷ 485	0.243/0.241/0.249/0.271	1662 ⟷ 1774	0.213/0.160/0.200/0.217
485 ⟷ 518	0.259/0.249/0.215/0.262	1774 ⟷ 1892	0.223/0.229/0.133/0.242
518 ⟷ 552	0.267/0.289/0.246/0.277	1892 ⟷ 2019	0.227/0.172/0.208/0.220
552 ⟷ 589	0.259/0.269/0.220/0.242	2019 ⟷ 2154	0.245/0.215/0.170/0.266
589 ⟷ 629	0.258/0.264/0.226/0.236	2154 ⟷ 2298	0.243/0.259/0.189/0.237
629 ⟷ 671	0.268/0.285/0.239/0.258	2298 ⟷ 2452	0.282/0.237/0.263/0.274
671 ⟷ 716	0.270/0.270/0.248/0.252	2452 ⟷ 2616	0.264/0.252/0.197/0.303
716 ⟷ 764	0.261/0.239/0.227/0.259	2616 ⟷ 2791	0.272/0.255/0.196/0.327
764 ⟷ 815	0.264/0.257/0.235/0.240	2791 ⟷ 2978	0.303/0.277/0.218/0.317
815 ⟷ 870	0.270/0.275/0.193/0.226	2978 ⟷ 3177	0.278/0.278/0.214/0.290
870 ⟷ 928	0.244/0.261/0.204/0.221	3177 ⟷ 3390	0.330/0.245/0.152/0.262
928 ⟷ 990	0.269/0.242/0.218/0.208	3390 ⟷ 3616	0.298/0.223/0.134/0.248
990 ⟷ 1056	0.275/0.272/0.246/0.227	3616 ⟷ 3808	0.310/0.219/0.202/0.310
1056 ⟷ 1127	0.277/0.229/0.228/0.230	3808 ⟷ 4050	0.204/0.200/0.133/0.203
1127 ⟷ 1202	0.268/0.261/0.206/0.211	4050 ⟷ 4250	0.174/0.198/0.129/0.159
1202 ⟷ 1283	0.262/0.283/0.153/0.216	4250 ⟷ 4500	0.073/0.065/0.081/0.071
1283 ⟷ 1369	0.229/0.247/0.175/0.203		

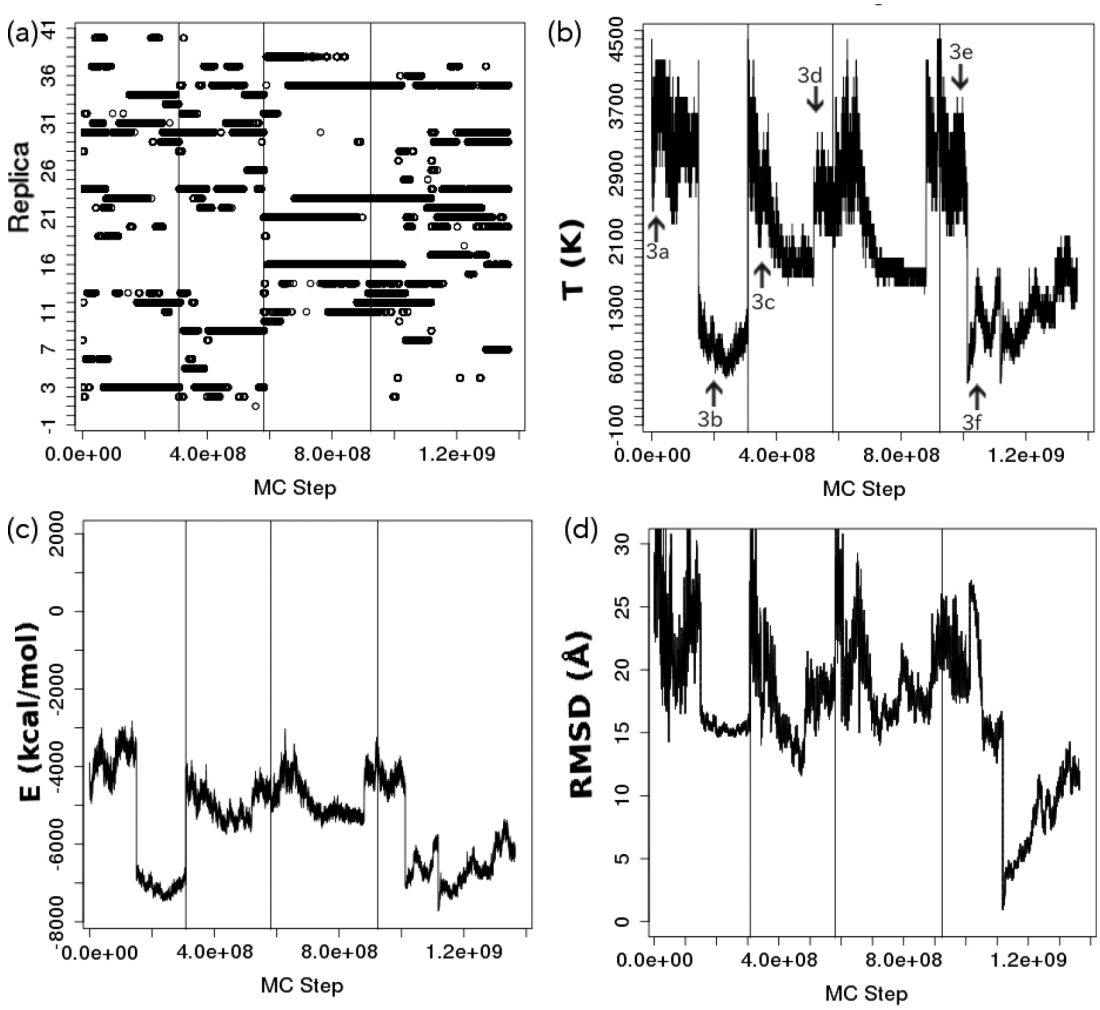

FIG. 2. Time series of various quantities for the REM simulations. (a) Replica index at temperature 400 K, (b) temperature change, (c) total potential energy change, and (d) RMSD (in Å) of all the C^{α} atoms from the PDB structure. (b)-(d) are the results of one of the replicas (Replica 27). The vertical lines show the boundary between four independent REM simulations. The arrows in (b) correspond to structures in Fig. 3.

We now explain the replica-exchange method[7,8] briefly. This method prepares M non-interacting replicas at M different temperatures. Let the label i (=1, . . ., M) correspond to the replica index and label m (=1, . . ., M) to the temperature index. We consider exchanging a pair of temperatures T_m and T_n, corresponding to replicas i and j. The transition probability $\omega(X \to X')$ of Metropolis criterion is given by

$$\omega(X \to X') \equiv \omega(x_m^{[i]} \mid x_n^{[j]}) = \min\left(1, \frac{W_{\mathrm{REM}}(X')}{W_{\mathrm{REM}}(X)}\right) = \min(1, \exp(-\Delta)), \quad (6)$$

where $\Delta = (\beta_m - \beta_n)(E(q^{[j]}) - E(q^{[i]}))$. While conventional canonical MC simulation is performed for each replica, temperature exchange between pairs of replicas corresponding to temperatures is attempted at a fixed interval based on the Metropolis criterion. Because each replica reaches various temperatures followed by replica exchange, the REM method performs a random walk in temperature space during the simulation. We used 40 replicas and the following temperatures: 400, 415, 435, 455, 485, 518, 552, 589, 629,

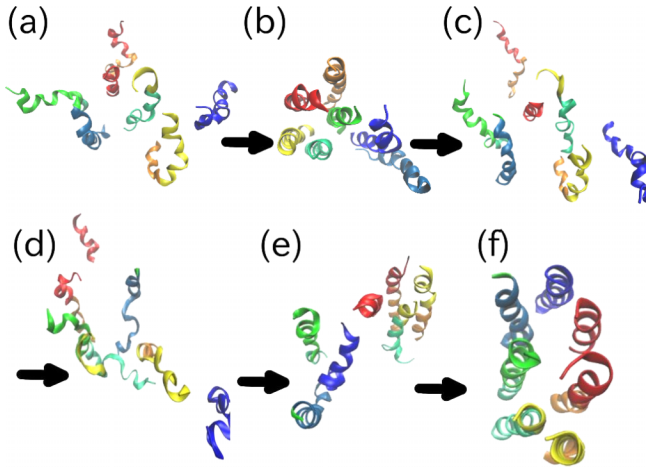

FIG. 3. Typical snapshots of one of the replicas (Replica 27) from the REM simulations. The corresponding temperatures are high (a), low (b), intermediate (c), high (d), intermediate (e), and low (f). They correspond to arrows in Fig. 2(b). The color of the helices from the N-terminus is as follows: helix A (blue), helix B (light blue), helix C (green), helix D (deep green), helix E (yellow), helix F (orange), and helix G (red) rendered by VMD.

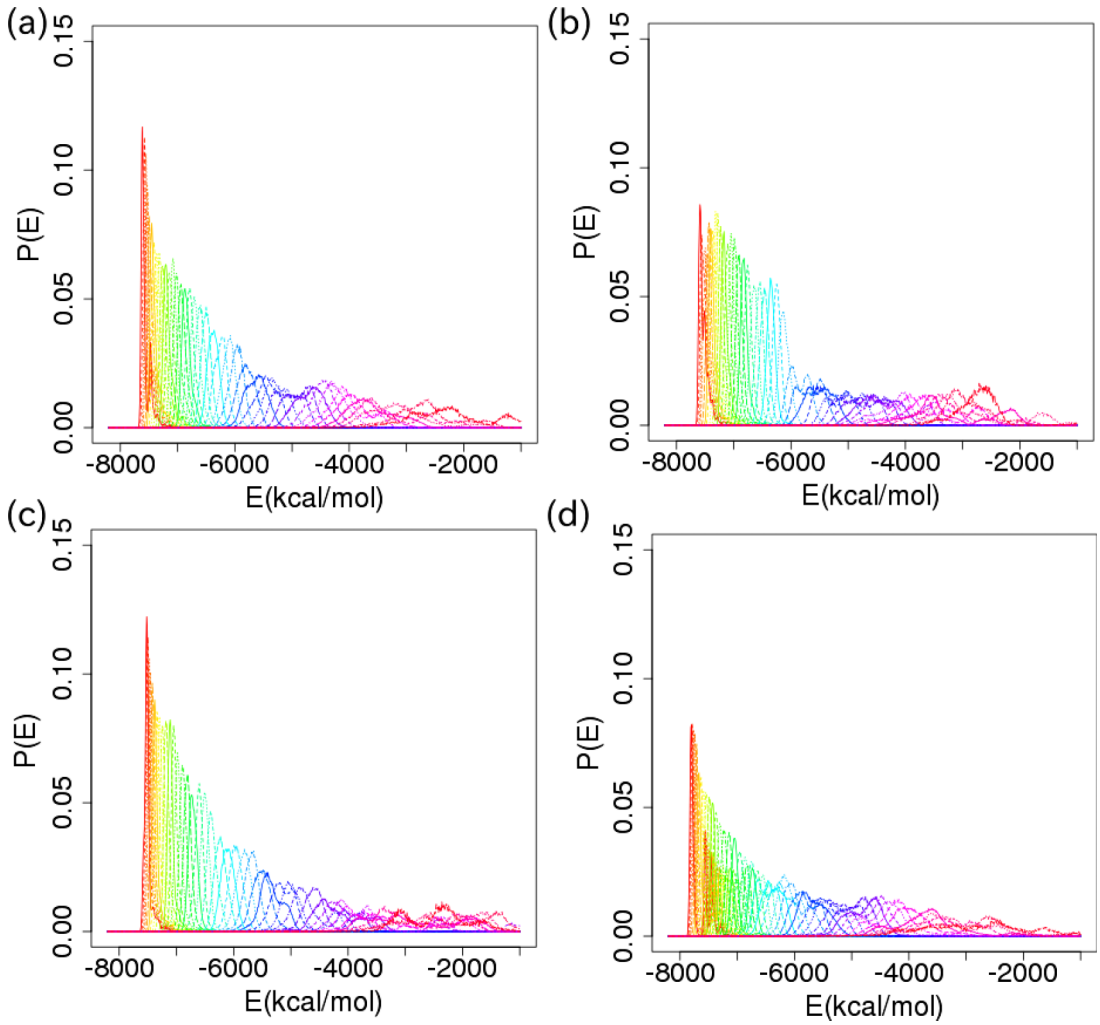

FIG. 4. Canonical probability distributions of total potential energy at each temperature from each REM simulation ((a) simulation 1, (b) simulation 2, (c) simulation 3, and (d) simulation 4). The distributions correspond to the following temperatures (from left to right): 400, 415, 435, 455, 485, 518, 552, 589, 629, 671, 716, 764, 815, 870, 928, 990, 1056, 1127, 1202, 1283, 1369, 1460, 1558, 1662, 1774, 1892, 2019, 2154, 2298, 2452, 2616, 2791, 2978, 3177, 3390, 3616, 3808, 4050, 4250, and 4500 K.

671, 716, 764, 815, 870, 928, 990, 1056, 1127, 1202, 1283, 1369, 1460, 1558, 1662, 1774, 1892, 2019, 2154, 2298, 2452, 2616, 2791, 2978, 3177, 3390, 3616, 3808, 4050, 4250, and 4500 K. We remark that because short simulations suggested that the choice of the minimum temperature of 300 K did not change overall helix orientations compared to 400 K, we set the minimum temperature to 400 K instead of 300 K to reduce the number of replicas. We used rather high temperature values compared to experimental conditions. This is because our implicit membrane model guarantees the helix stability and enhances conformational sampling. Replica exchange was attempted at every 50 MC steps. We performed four independent simulations in total of 1 363 925 000 MC steps (the number of MC steps in each simulation was 308 000 000, 250 000 000, 343 200 000, and 462 775 000 MC steps).

We used the CHARMM19 parameter set (polar hydrogen model) for the potential energy of the system.[34,35] No cutoff was introduced to the non-bonded terms. Each helix structure was first minimized subjected to harmonic restraint on all the

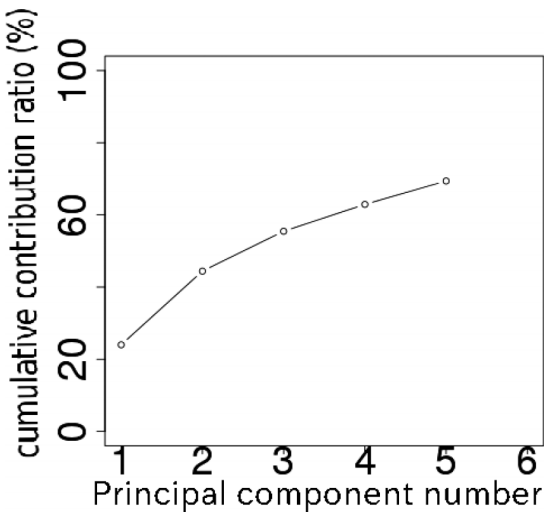

FIG. 5. Cumulative contribution ratio of the first five eigenvalues in the principal component analysis from sampled structures of the REM simulation at 400 K.

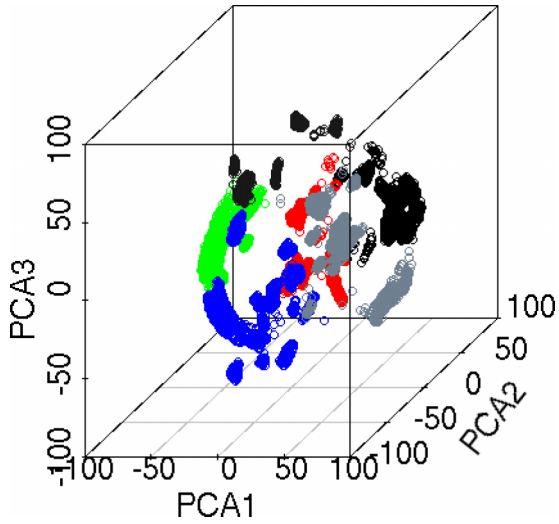

FIG. 6. Projection of sampled structures at temperature 400 K on the first, second, and third principal axes from the REM simulations. Structures are classified into clusters of similar structures by k-means method and analyzed in detail. Clusters are highlighted by different colors: red, blue, green, gray, and black. PCA1, PCA2, and PCA3 represent the principal component axes 1, 2, and 3, respectively.

heavy atoms. In order to prepare random initial conformations, we first performed regular constant temperature MC simulations of all the replicas for 3 000 000 MC steps at 4500 K. We then performed equilibrium MC simulation for 3 000 000 MC steps at the above 40 temperatures, and the last conformation for each replica was the initial structure for the REM simulations. We repeated this process four times for four independent REM simulations with different random number seeds. In those simulations, the dielectric constant was set to $\epsilon = 1.0$ as in the previous works[17–21] because the works confirmed that the results with $\epsilon = 1.0$ value were in better agreement with the experimental structures than $\epsilon = 4.0$. The interpretation was that in the native structure there are few lipid molecules between helices. In MC move, we updated conformations with a rigid translation and rotation of each α-helix, a rotation of torsion angles of backbones by directional manipulation and concerted rotation,[36–38] and torsion rotations of side-chains. There are $2N_{\mathrm{H}} + N_{\mathrm{SD}} + N_{\mathrm{BD}} + N_{\mathrm{CR}}$ kinds of MC moves, where N_{SD} is the total number of dihedral angles in the side-chains of N_{H} helices and N_{CR} is the total number of the combination of seven successive backbone torsion angles by the concerted rotation in the helix backbone. One MC step in this article is defined to be an update of one of these degrees of freedom, which is accepted or rejected according to the Metropolis criterion.

We analyzed the simulation data by the principal component analysis.[39–44] At first, $N = 42\,238$ conformational data were chosen at each temperature from the REM simulations. The structures were chosen from the trajectories at a fixed interval of 25 000 steps. The structures were superimposed on an arbitrary reference structure, for example, the native structures of PDB code:1PY6. Images were rendered by Visual Molecular Dynamics (VMD).[45] The variance-covariance matrix is defined by

$$C_{ij} = \langle (q_i - \langle q_i \rangle)(q_j - \langle q_j \rangle) \rangle, \tag{7}$$

where $\vec{q} = (q_1, q_2, q_3, \ldots, q_{3n-1}, q_{3n}) = (x_1, y_1, z_1, \ldots, x_n, y_n, z_n)$ and $\langle q_i \rangle = \sum_{i=1}^{N} q_i/N$, x_i, y_i, z_i are the Cartesian coordinates of the ith atom, n is the total number of atoms, and N is the total number of samples. The first superposition is performed to remove large eigenvalues from the translations and rotations of the system because we want to analyze the internal differences of structures. The eigenvalues are ordered in the decreasing order of magnitude. The ith component of each sampled structure $\vec{q}$ is defined by the inner product

$$\mu_i = \vec{v}_i \cdot (\vec{q} - \langle \vec{q} \rangle), \quad (i = 1, 2, \ldots, n), \tag{8}$$

where $\vec{v}_i$ is the (normalized) ith eigenvector. The calculations were performed by the R program package,[46] and the clustering in the 3-dimensional PCA space (see Fig. 6) was performed by the k-means clustering method.[47] The optimal number of clusters, k, in the k-mean method was determined as follows. We first chose a large number of $k = 20$ as an initial value and performed the k-mean clustering for several times with randomly chosen initial set of k centers. For each clustering, the representative structure for each cluster was chosen to be the one with the largest density of points in the cluster. If the representative structures of different clusters had the same helix topology (or, relative positions of neighboring helices), then we reduced the value of k to, say, 15, and repeated the above clustering. We continued this procedure until all the obtained clusters essentially have different helix topology. Our optimal value of k was 6. After k was set to 6, we performed the clustering with different randomly chosen 6 initial centers for twenty times. We finally selected five clusters that have the most frequently chosen regions of centers. For the analysis of helix kink, we used the algorithms[48,49] by HELANAL-Plus.[50]

TABLE III. Various average quantities for each cluster and at the temperature of 400 K. The following abbreviations are used: Str: number of structures, E_{tot}: average total potential energy, E_{elec}: average electrostatic energy, E_{vdw}: average Lennard-Jones energy, E_{dih}: average dihedral energy, E_{geo}: average restraint energy (all in kcal/mol), RMSD: average root-mean-square deviation of all C^{α} atoms (in Å).

	Str	E_{tot}	E_{elec}	E_{vdw}	E_{dih}	E_{geo}	RMSD
Cluster 1	9 123	−7587 ± 355	−7006 ± 51	−1197 ± 83	202 ± 14	106 ± 215	6.2 ± 5.6
Cluster 2	13 146	−7422 ± 182	−6976 ± 45	−1166 ± 44	213 ± 11	147 ± 123	12.2 ± 2.4
Cluster 3	7 457	−7287 ± 210	−6943 ± 31	−1143 ± 71	211 ± 10	229 ± 142	14.9 ± 2.5
Cluster 4	5 121	−7421 ± 109	−7010 ± 36	−1145 ± 43	213 ± 10	160 ± 71	16.3 ± 1.1
Cluster 5	4 418	−7300 ± 309	−6954 ± 41	−1096 ± 67	207 ± 11	184 ± 258	16.7 ± 1.0
400 K	42 240	−7412 ± 263	−6979 ± 48	−1158 ± 68	210 ± 12	166 ± 170	12.5 ± 4.8

III. RESULTS

We first examine how our REM simulations performed. Table I lists the acceptance ratio (AR) of replica exchange between all pairs of neighboring temperatures in whole trajectories. We find that the acceptance ratio is high enough (>0.1) in all temperature pairs. Similarly, Table II lists the acceptance ratio of replica exchange between all pairs of neighboring temperatures in each simulation. Figure 2(a) shows the time series of the replica index at the lowest temperature of 400 K. We see that many replicas visited the minimum temperature many times during the REM simulations, and we observe a random walk in the replica space. The complementary picture is the temperature exchange for each replica in the temperature space. We show the results for one of the replicas (Replica 27) in Figs. 2(b)–2(d). Replica 27 visited various temperatures during the REM simulations as can be seen in Fig. 2(b). We observe random walks in the temperature space between the lowest and highest temperatures. Other replicas also performed random walks similarly. Sudden changes of temperature (around 2100 K) and potential energy (around −6000 kcal/mol) occurred near 2.0×10^8 and 1.0×10^9 MC steps. Helices are apart from each other above these values and in contact below. A large decrease occurs when more than two helices become in contact. Figure 2(c) shows the corresponding time series of the total potential energy. These results imply that these REM simulations were properly performed. We next examine how widely the conformational space was sampled during the REM simulations. We plot the time series of the root-mean-square deviation (RMSD) of all the C^α atoms from the experimental structure (PDB code: 1PY6) for Replica 27 in Fig. 2(d). When the temperature becomes high, the RMSD takes large values, and when the temperature becomes low, the RMSD falls into smaller values. By comparing Figs. 2(b) and 2(d), we see that there is a strong correlation between the temperature and RMSD values. The fact that RMSD at high temperature was large implies that our simulations did not get trapped in local-minimum potential-energy states so that the distances among helices are large. Typical snapshots are shown in Fig. 3 (see Fig. 2 also). Fig. 4 shows the probability distributions of total potential energy at each temperature in four REM simulations. The distributions correspond to the following temperatures (from left to right): 400, 415, 435, 455, 485, 518, 552, 589, 629, 671, 716, 764, 815, 870, 928, 990, 1056, 1127, 1202, 1283, 1369, 1460, 1558, 1662, 1774, 1892, 2019, 2154, 2298, 2452, 2616, 2791, 2978, 3177, 3390, 3616, 3808, 4050, 4250, and 4500 K. They are reasonably similar to each other, covering the same ranges of potential energy. They indeed exhibit wide conformational changes. In Fig. 5, we show the percentage of the cumulative contribution ratio of the first five eigenvalues at 400 K. We see that more than 55% of the total fluctuations at the lowest temperature is expressed by the first three principal components.

We investigated the free energy landscape obtained by the principal component analysis. We classified the sampled structures at the minimum temperature of 400 K into clusters of similar structures by the k-means clustering method.[47] The cumulative contribution in the whole fluctuation was 55% at 400 K by the third principal component as shown in Fig. 5. Although we can express the system more accurately as we use more principal axes in k-means clustering, we here classify and analyze the sampled structures at the lowest

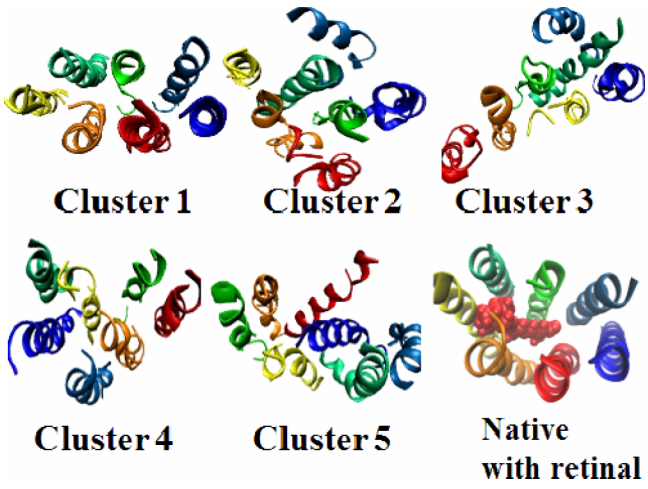

FIG. 7. Typical structures in each cluster selected in the highest density region. The RMSD from the native conformation with respect to all C^α atoms is 3.6 Å, 8.8 Å, 15.8 Å, 15.9 Å, and 16.6 Å for cluster 1, cluster 2, cluster 3, cluster 4, and cluster 5, respectively. Helices are colored from the N-terminus to the C-terminus: blue (helix A), light blue (helix B), green (helix C), deep green (helix D), yellow (helix E), orange (helix F), and red (helix G).

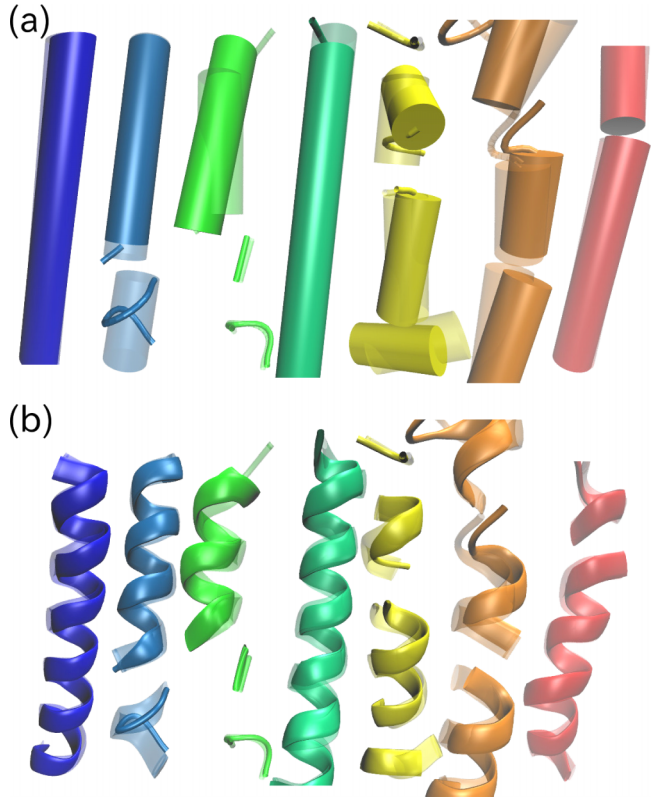

FIG. 8. Comparison of distortions of each helix structure in cluster 1 with the native helix structure by two different representation methods, (a) cartoon and (b) ribbon. The transparent structures are the native structures. The RMSD from the native conformation with respect to backbone atoms is 0.7 Å, 0.9 Å, 1.6 Å, 2.7 Å, 1.1 Å, 2.2 Å, and 0.4 Å for helix A, helix B, helix C, helix D, helix E, helix F, and helix G, respectively. The coloring of helices is the same as in Fig. 7.

TABLE IV. Properties of kink, twist, and bending of each helix in the representative structure of cluster 1 and PDB analyzed HELANAL-Plus. Helices A, B, C, D, E, F, and G correspond to 9E-31G, 41K-62L, 78I-100L, 105Q-127L, 134R-156F, 170T-194E, and 201L-224L, respectively. The following abbreviations are used: Twist: Average unit twist of the helix (deg.), n: Average number of residues per turn of the helix, h: Average unit height of the helix (Å), Aver vtor: Average virtual torsion angle defined by four C^{α} atoms (deg.), Aver BA: Average bending angle between successive local helix axes (deg.), Max BA: Maximum bending angle between successive local helix axes (deg) (the residue number and name are given in parentheses), Ge.: Overall geometry of the helix, namely, linear (L), curved (C), kinked (K), or unassigned (-).

	Helix	Twist	n	h	Aver vtor	Aver BA	Max BA	Ge.
1PY6	A	98.2 ± 2.08	3.67 ± 0.08	1.49 ± 0.09	48.7 ± 3.6	4.3 ± 1.8	9.4 (5W)	C
1PY6	B	104.5 ± 15.29	3.45 ± 0.40	1.60 ± 0.32	57.8 ± 22.7	14.7 ± 11.9	37.1 (34V)	K
1PY6	C	121.4 ± 42.79	2.96 ± 0.74	1.75 ± 0.58	95.0 ± 79.0	42.6 ± 26.4	91.2 (61P)	K
1PY6	D	98.7 ± 3.06	3.65 ± 0.11	1.48 ± 0.14	48.8 ± 5.9	7.9 ± 5.0	16.4 (85T)	L
1PY6	E	113.8 ± 19.38	3.16 ± 0.48	1.69 ± 0.26	68.2 ± 25.3	11.8 ± 6.5	25.9 (100S)	L
1PY6	F	116.2 ± 41.25	3.10 ± 0.65	1.76 ± 0.34	76.2 ± 56.6	16.5 ± 15.7	64.0 (135I)	K
1PY6	G	101.2 ± 15.76	3.56 ± 0.47	1.50 ± 0.29	52.6 ± 21.1	11.3 ± 5.2	23.9 (155K)	L
Cluster 1	A	99.3 ± 2.86	3.63 ± 0.10	1.51 ± 0.11	49.9 ± 4.9	7.6 ± 4.4	14.6 (8L)	C
Cluster 1	B	104.3 ± 16.08	3.45 ± 0.44	1.57 ± 0.37	57.0 ± 23.7	15.0 ± 13.3	47.2 (30A)	K
Cluster 1	C	116.3 ± 39.30	3.09 ± 0.65	1.69 ± 0.44	85.7 ± 74.2	38.0 ± 30.4	97.1 (61P)	K
Cluster 1	D	100.6 ± 5.37	3.58 ± 0.19	1.51 ± 0.22	51.3 ± 10.0	12.9 ± 9.2	31.6 (77G)	K
Cluster 1	E	115.2 ± 19.84	3.13 ± 0.48	1.74 ± 0.29	71.5 ± 27.6	15.0 ± 8.4	25.6 (100S)	K
Cluster 1	F	108.4 ± 23.24	3.32 ± 0.77	1.68 ± 0.45	65.7 ± 31.5	23.0 ± 10.0	35.9 (134L)	K
Cluster 1	G	101.4 ± 14.48	3.55 ± 0.46	1.51 ± 0.30	52.7 ± 18.0	15.9 ± 8.1	32.8 (155K)	K

temperature by the first three principal components. In Fig. 6, the projection of sampled structures from the REM simulations at 400 K on the first, second, and third principal component axes is shown. We obtained five distinct clusters of similar structures. If we perform constant temperature simulations at the lowest temperature, the simulations will get trapped in any of the clusters in Fig. 6, depending on the initial conformations of the simulations. However, each replica during the REM simulations did not get trapped in one of the local-minimum free energy states, by going through high temperature regions. Every replica could overcome energy barriers at higher temperatures during the simulations. This is the advantage of the replica-exchange method. Table III lists average quantities of five clusters of similar structures and at the temperature of 400 K. The rows of cluster 1, cluster 2, cluster 3, cluster 4, and cluster 5 represent various average values for the structures that belong to each cluster. The number of structures in each cluster (the total number was 42238 structures) was 9123, 13 146, 7457, 5121, and 4418 for cluster 1, cluster 2, cluster 3, cluster 4, and cluster 5, respectively. Thus, the global-minimum free energy state is cluster 2, and the second-lowest minimum state is cluster 1.

Figure 7 shows the representative structure in each cluster from the highest density region. The RMSD value of each representative structure with respect to the C^{α} atoms was 3.6 Å, 8.8 Å, 15.8 Å, 15.9 Å, and 16.6 Å for cluster 1, cluster 2, cluster 3, cluster 4, and cluster 5, respectively. From these RMSD values, we see that the native-like structure is the second-lowest free energy state (cluster 1) and that the global-minimum free energy state (cluster 2) is the second closest to the native structure. In the structure of cluster 2, the space where the retinal molecule occupies in the native structure is filled with a helix, and this increases the contact between helices and seems to stabilize this structure more than the native-like structure of cluster 1 with the empty space for the retinal molecule. Moreover, this result that a helix occupies the retinal space is consistent with previous works[20,21] which did not include the flexibility of helix structures. However, the previous works were not able to obtain the native-like structure such as cluster 1. Hence, the extension of including the freedom of helix structure distortion has improved the accuracy of prediction for membrane protein structure determination by simulation. Our results suggest that in the simulations without a retinal molecule the structures can interchange between the structures of cluster 1 and cluster 2. After an insertion of a retinal, it then stabilizes the native-like structure. It is important that the association of helices enabled them to make a room for an insertion of a retinal molecule. This is consistent with the experimental results of bacteriorhodopsin, which observed the spontaneous insertion of a retinal molecule by a helix association.[51]

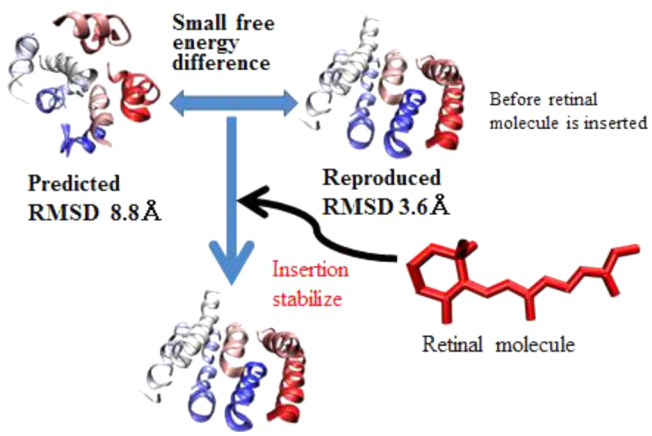

FIG. 9. Hypothesis about the relation between the global-minimum free energy state and the second-minimum. The effect of an insertion of a retinal molecule causes the stabilization of the native-like structure.

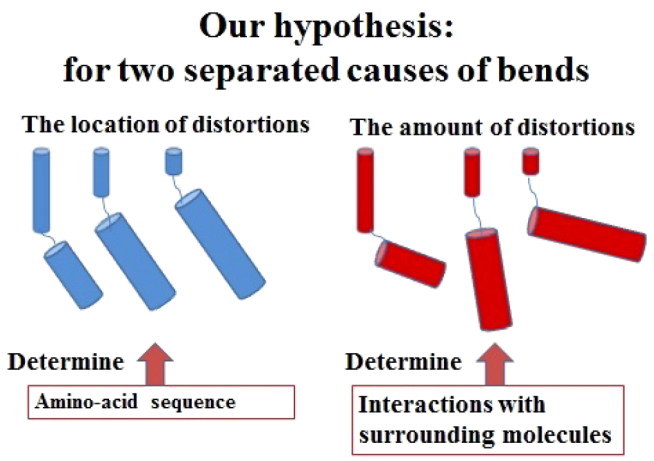

FIG. 10. Hypothesis for distortions of transmembrane helix structures: the location of bending is determined by the amino-acid sequences whereas the amount of bending is determined by the interactions with their surroundings such as neighboring helices, retinal, and lipid molecules.

We now examine the distortions of each helix of native-like structure in cluster 1. They are compared with the native structures in Fig. 8. Note that the positions of the kinks of helices are all correctly reproduced. The RMSD values with respect to backbone atoms are 0.7 Å, 0.9 Å, 1.6 Å, 2.7 Å, 1.1 Å, 2.2 Å, and 0.4 Å for helix A, helix B, helix C, helix D, helix E, helix F, and helix G, respectively. These helix structures reproduced the bending of helices with smaller RMSD values for the helices apart from the retinal molecule in the native state (helices A, B, and G), whereas the helices with larger RMSD values of about 2.0 Å are for the helices in close contact with the retinal in the native structure (helices C, D, E, and F). They suggest that the larger RMSD values came from the neglect of the retinal molecule in our simulations.

In order to study bendings of helices more quantitatively, we analyzed the data by the software HELANAL-Plus. The results for representative structures of PDB and cluster 1 are summarized in Table IV. They are in good agreement between PDB and cluster 1 structures. Most properties of helices are similar between them, and because of the similarity in maximum bending and average bending with standard errors, this suggests that the tendency of bending was highly reproduced, although overall helix geometry is assigned differently (L vs. K) in helix D, helix E, and helix G. However, there is disagreement in data about the maximum bending location in helix D (85T for PDB and 77G for cluster 1). This seems to result from the lack of interactions with the retinal molecule because the retinal binding positions are 79D, 82M, and 86G near the C-terminus side.

These results imply that the position of the kinks of helices is determined by their amino-acid sequences and that the amount of bends of helices is determined by the interactions with surrounding molecules such as other helices, retinal, and lipid molecules.

IV. CONCLUSIONS AND DISCUSSION

In this work, we applied our new prediction method for membrane protein structures to bacteriorhodopsin without a retinal molecule. This method includes the flexibility of helix backbone structures for reproduction of distorted helix structures frequently observed in experimental structures. Our results reproduced the native-like structure determined by X-ray diffraction with a retinal molecule as the second global-minimum free energy state.

Moreover, the global-minimum free energy state has the structure that has a helix filling in the empty space where a retinal molecule is in the native structure. This filling is reasonable because the increase of contact among helices stabilizes the structure. This is also consistent with previous results.[21] Our results confirmed that a space for a retinal was created by the association of transmembrane helices in a membrane. This suggests that after the insertion of a retinal at the same space, the interaction among helices and the retinal stabilizes the native-like structure as shown in Fig. 9. The thermal equilibrium between the empty space and the filled space for retinal insertion is also consistent with the experiments of bacteriorhodopsin observing a spontaneous insertion using the association of helices to make the room for co-molecule and three stage model about the co-molecule.[51,52] This consistency supports the accuracy of our results. As a result, we obtained native-like structure with an empty space at the center and the structure filled with a helix at the position as the global-minimum free energy state. The difference of at the same space shows the association of helices can make an empty space for an insertion of a molecule. This space confirmed the model suggested by previous experiments and three stage model.[52]

The distortion of each helix structure was examined in the native-like structure of cluster 1. In the native-like conformation, the distortion of helices with similar interactions to the native state apart from the position of a retinal molecule was similar to experimental one. On the other hand, the distortions of helices with different interactions due to lack of a retinal had about two times larger RMSD values although the bends started at similar positions of native ones. Both these results and previous results of phospholamban suggest a hypothesis (summarized in Fig. 10) about bends that the causes of distortions are divided into two factors: the location of bends is determined by the amino-acid sequence information and the amount of bends is determined by the interactions with surroundings. In particular, some previous works observed the sensitivity of helix kinks to their environment, for example, membrane surface curvature.[53] To confirm our hypothesis about the effects of a retinal molecule, we are now performing simulations of the same system including a retinal molecule. The improvement for more accuracy is also important, and for that, combination with the other implicit membrane models such as GB models for membranes will be useful to improve the accuracy of free energy surfaces during simulations.

ACKNOWLEDGMENTS

Some of the computations were performed on the supercomputers at the Institute for Molecular Science, at the Supercomputer Center, Institute for Solid State Physics, University of Tokyo, and Center for Computational Sciences,

University of Tsukuba. This work was supported, in part, by the Grants-in-Aid for Scientific Research (A) (No. 25247071), for Scientific Research on Innovative Areas ("Dynamical Ordering and Integrated Functions"), for the Computational Materials Science Initiative, for High Performance Computing Infrastructure, and by the Program for Leading Graduate Schools "Integrative Graduate Education and Research in Green Natural Sciences" from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan.

[1]S. Fiedler, J. Broecker, and S. Keller, Cell. Mol. Life Sci. **67**, 1779 (2010).
[2]G. von Heijne, Annu. Rev. Biochem. **80**, 157 (2011).
[3]R. Sawada and S. Mitaku, J. Biochem. **151**, 189 (2012).
[4]G. Tusnády, Z. Dosztányi, and I. Simon, Bioinformatics **20**, 2964 (2004).
[5]G. Tusnády, Z. Dosztányi, and I. Simon, Nucleic Acids Res. **33**, 275 (2005).
[6]P. Raman, V. Cherezov, and M. Caffrey, Cell. Mol. Life Sci. **63**, 36 (2006).
[7]K. Hukushima and K. Nemoto, J. Phys. Soc. Jpn. **65**, 1604 (1996).
[8]Y. Sugita and Y. Okamoto, Chem. Phys. Lett. **314**, 141 (1999).
[9]Y. Sugita, A. Kitao, and Y. Okamoto, J. Chem. Phys. **113**, 6042 (2000).
[10]P. Liu, B. Kim, R. A. Friesner, and B. J. Berne, Proc. Natl. Acad. Sci. U. S. A. **102**, 13749 (2005).
[11]A. Mitsutake and Y. Okamoto, J. Chem. Phys. **130**, 214105 (2009).
[12]A. Mitsutake, Y. Sugita, and Y. Okamoto, Biopolymers **60**, 96 (2001).
[13]W. Im, M. Feig, and C. L. Brooks III, Biophys. J. **85**, 2900 (2003).
[14]S. Tanizaki and M. Feig, J. Chem. Phys. **122**, 124706 (2005).
[15]T. Lazaridis, Proteins: Struct., Funct., Bioinf. **58**, 518 (2005).
[16]A. Panahi and M. Feig, J. Chem. Theory Comput. **9**, 1709 (2013).
[17]H. Kokubo and Y. Okamoto, Chem. Phys. Lett. **383**, 397 (2004).
[18]H. Kokubo and Y. Okamoto, J. Chem. Phys. **120**, 10837 (2004).
[19]H. Kokubo and Y. Okamoto, J. Phys. Soc. Jpn. **73**, 2571 (2004).
[20]H. Kokubo and Y. Okamoto, Chem. Phys. Lett. **392**, 168 (2004).
[21]H. Kokubo and Y. Okamoto, Biophys. J. **96**, 765 (2009).
[22]J. L. Popot and D. M. Engelman, Annu. Rev. Biochem. **69**, 881 (2000).
[23]Y. Matsui, K. Sakai, M. Murakami, Y. Shiro, S.-i. Adachi, H. Okumura, and T. Kouyama, J. Mol. Biol. **324**, 469 (2002).
[24]S. Faham, D. Yang, E. Bare, S. Yohannan, J. P. Whitelegge, and J. U. Bowie, J. Mol. Biol. **335**, 297 (2004).
[25]L.-O. Essen, R. Siegert, W. D. Lehmann, and D. Oesterhelt, Proc. Natl. Acad. Sci. U. S. A. **95**, 11673 (1998).
[26]R. Urano, H. Kokubo, and Y. Okamoto, J. Phys. Soc. Jpn. **84**, 084802 (2015).
[27]T. Hirokawa, S. Boon-Chieng, and S. Mitaku, Bioinformatics **14**, 378 (1998).
[28]A. Krogh, B. Larsson, G. von Heijne, and E. Sonnhammer, J. Mol. Biol. **305**, 567 (2001).
[29]D. T. Jones, W. R. Taylor, and J. M. Thornton, Biochemistry **33**, 3038 (1994).
[30]G. Tusnady and I. Simon, Bioinformatics **17**, 849 (2001).
[31]B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, J. Comput. Chem. **4**, 187 (1983).
[32]J. Hu, A. Ma, and A. R. Dinner, J. Comput. Chem. **27**, 203 (2006).
[33]M. Lomize, A. Lomize, I. Pogozheva, and H. Mosberg, Bioinformatics **22**, 623 (2006).
[34]W. Reiher, "Theoretical studies of hydrogen bonding," Ph.D. thesis, Harvard University, 1985.
[35]E. Neria, S. Fischer, and M. Karplus, J. Chem. Phys. **105**, 1902 (1996).
[36]A. R. Dinner, J. Comput. Chem. **21**, 1132 (2000).
[37]N. Gō and H. A. Scheraga, Macromolecules **3**, 178 (1970).
[38]L. Dodd, T. Boone, and D. Theodorou, Mol. Phys. **78**, 961 (1993).
[39]M. Teeter and D. Case, J. Phys. Chem. **94**, 8091 (1990).
[40]A. Kitao, F. Hirata, and N. Gō, Chem. Phys. **158**, 447 (1991).
[41]A. Garcia, Phys. Rev. Lett. **68**, 2696 (1992).
[42]R. Abagyan and P. Argos, J. Mol. Biol. **225**, 519 (1992).
[43]A. Amadei, A. Linssen, and H. Berendsen, Proteins **17**, 412 (1993).
[44]A. Kitao and N. Gō, Curr. Opin. Struct. Biol. **9**, 164 (1999).
[45]W. Humphrey, A. Dalke, and K. Schulten, J. Mol. Graphics **14**, 33 (1996).
[46]R. Ihaka and R. Gentleman, J. Comput. Graph. Stat. **5**, 299 (1996).
[47]J. MacQueen, in *Fifth Berkeley Symposium on Mathematical Statistics and Probability*, edited by L. M. Le Cam and J. Neyman (University of California Press, Berkeley, CA, 1967), pp. 281–297.
[48]H. Sugeta and T. Miyazawa, Biopolymers **5**, 673 (1967).
[49]C. M. Shakarji *et al.*, J. Res. Natl. Inst. Stand. Technol. **103**, 633 (1998).
[50]P. Kumar and M. Bansal, J. Biomol. Struct. Dyn. **30**, 773 (2012).
[51]J. L. Popot, S. E. Gerchman, and D. M. Engelman, J. Mol. Biol. **198**, 655 (1987).
[52]D. M. Engelman, Y. Chen, C.-N. Chin, A. R. Curran, A. M. Dixon, A. D. Dupuy, A. S. Lee, U. Lehnert, E. E. Matthews, Y. Reshetnyak, A. Senes, and J.-L. Popot, FEBS Lett. **555**, 122 (2003).
[53]L. Dominguez, S. C. Meredith, J. E. Straub, and D. Thirumalai, J. Am. Chem. Soc. **136**, 854 (2014).