STRUCTURE PREDICTIONS OF MEMBRANE PROTEINS BY MOLECULAR SIMULATIONS

A Thesis
Presented to the Department of Functional Molecular Science
School of Physical Sciences
The Graduate University for Advanced Studies
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Science

by

Hironori Kokubo

September 2004

Acknowledgments

My most heartfelt thanks go to my thesis advisor, Professor Yuko Okamoto, for his patient guidance and constant encouragement. I am grateful to the members of the IMS theory groups for their generous support. I wish to thank the faculty and staff members of the Graduate University for Advanced Studies for their kindness.

Contents

Chapter 1

General Introduction

It is one of the most important problems in the structural genomics era to predict protein tertiary structures from the amino-acid sequence information. We can obtain various information about the function and stability by the knowledge of protein structures. Therefore many efforts are devoted to structural determination of proteins.

It is estimated that 20-30 % of all genes in most genomes encode membrane proteins [1, 2]. However, only a small number of detailed structures have been obtained for membrane proteins because of technical difficulties in experiments such as high quality crystal growth. About 25000 protein structures are currently registered on the Protein Data Bank (PDB) [3], but most of them are structures of soluble proteins, and the number of membrane protein structures are less than 100. The database analysis based on bioinformatics such as homology search are thus unreliable due to lack of enough samples. Therefore, it is desirable to develop a method for predicting membrane protein structures by computer simulations (for previous attempts, see, for instance, Refs. [4]–[9]).

Although the number of known membrane protein structures is small in PDB, we can still extract several features of their structures. Transmembrane regions of most membrane proteins in inner membrane are composed of helices, and those in outer membrane are composed of β-sheet. In other words, membrane proteins have only one type of secondary structures in these regions, and in this sense their structures are simpler than those of soluble proteins. Another feature is that membrane protein structures are known to be more tightly packed than soluble protein structures. These features should be taken into account and utilized when we consider membrane protein structure predictions.

The two-stage model was proposed for the structure formation of membrane proteins which are composed of several transmembrane helices in Ref.[10]. In the two-stage model, individual helices of a membrane protein are postulated to be stable separately as domains in a lipid bilayer and then side-to-side helix association is driven, resulting in a functional protein. In fact, some experimental evidence indicates that the formation of α-helices and the positioning of transmembrane helices are independent: Separated fragments of bacteriorhodopsin formed independently α-helical conformations in the membrane, and the native structure could be recovered by mixing the fragments [11, 12]. Therefore, it is reasonable to assume that processes of helix formation and positioning can be predicted separately.

Considering the difficulties in experiments and homology-based predictions, it is particularly desirable to develop effective prediction methods of membrane protein structures by molecular simulations. Molecular simulations allow us to understand the physical mechanism of the stability and functions of membrane proteins and help us to construct a unified view of their structures and functions.

Our prediction method consists of two parts. In the first part, amino-acid sequences of the transmembrane helix regions are obtained from database analyses [1][13]–[19]. In the second part, we perform a molecular simulation of these transmembrane helices with some constraints and identify the global-minimum-energy state as the predicted structure.

However, it is difficult to obtain a global-minimum state in potential energy surface by conventional molecular dynamics (MD) or Monte Carlo (MC) simulations. This is because there exist a huge number of local-minimum-energy states, and the simulations tend to get trapped in one of the local-minimum states. One popular way to overcome this multiple-minima problem is to perform a generalized-ensemble simulation (for reviews, see Refs. [20, 21]), which is based on non-Boltzmann probability weight factors so that a random walk in potential energy space may be realized. The random walk allows the simulation to go over any energy barrier and sample much wider configurational space than by conventional methods. One of well-known generalized-ensemble algorithms is the replica-exchange method (REM) [22]–[24] (the method is also referred to as parallel tempering [25]). We apply this method to the structure prediction of membrane proteins. We can obtain not only the global-minimum-energy state but also canonical-ensemble averages of physical quantities as functions of temperature from only one REM simulation run by using the multiple-histogram reweighting techniques [26, 27].

In this thesis we target helical membrane proteins and try to predict their structures using replica-exchange Monte Calro method, which is one of the generalized ensemble algorithms.

In Chapter2, we propose a method for predicting helical membrane protein structures by computer simulations. This method is used in Chapters 3, 4, and 5.

In Chapter 3, we apply the prediction method to the structure prediction of the dimeric transmembrane domain of glycophorin A (PDB code: 1AFO [28]) and analyze the predicted structures in detail.

In Chapter 4, the effectiveness of our classification and prediction method for transmembrane helix configurations of membrane proteins by replica-exchange simulations is tested with the glycophorin A transmembrane dimer. We classify low-energy configurations into clusters of similar structures by the principal component analysis. These clusters are identified as the global-minimum and local-minimum free energy states.

In Chapter 5, we examine by a molecular simulation whether or not the transmembrane helices of bacteriorhodopsin have the ability to self-assemble into the native configuration by themselves. Starting from random initial configurations of seven transmembrane helices and using essentially the same procedure as above, we examine whether similar structures to the experimental one (PDB code: 1C3W) are obtained by a replica-exchange Monte Carlo simulation.

Finally, Chapter 6 is devoted to conclusions.

Bibliography

[1] A. Krogh, B. Larsson, G.v. Heijne, E.L.L. Sonnhammer, J. Mol. Biol. 305 (2001) 567.

[2] S. Mitaku, Biophysics 42 (2002) 104 (in Japanese).

[3] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, Nucleic Acids Research 28 (2000) 235.

[4] W.R. Taylor, D.T. Jones, N.M. Green, Proteins 18 (1994) 281.

[5] M. Suwa, T. Hirokawa, S. Mitaku, Proteins 22 (1995) 363.

[6] P.D. Adams, D.M. Engelman, A.T. Brünger, Proteins 26 (1996) 257.

[7] R.V. Pappu, G.R. Marshall, J.W. Ponder, Nature Struct. Biol. 6 (1999) 50.

[8] T. Hirokawa, J. Uechi, H. Sasamoto, M Suwa, S. Mitaku, Protein Eng. 13 (2000) 771.

[9] N. Vaidehi, W.B. Floriano, R. Trabanino, S.E. Hall, P. Freddolino, E.J. Choi, G. Zamanakos, W.A. Goddard, III, Proc. Natl. Acad. Sci. U.S.A. 99 (2002) 12622.

[10] J.L. Popot, D.M. Engelman, Annu. Rev. Biochem. 69 (2000) 881.

[11] M.J. Liao, E. London, H.G. Khorana, J. Biol. Chem. 258 (1983) 9949.

[12] D.M. Engelman, B.D. Adair, J.F. Hunt, T.W. Kahn, J.L. Popot, Curr. Topics Membr. Transport 36 (1990) 71.

[13] P. Argos, J.K. Rao, P.A. Hargrave, Eur. J. Biochem. 128 (1982) 565.

[14] D. Eisenberg, R.M. Weiss, T.C. Terwilliger, W. Wilcox, Faraday Symp. Chem. Soc. 17 (1982) 109.

[15] K. Nakai, M. Kanehisa, Genomics 14 (1992) 897.

[16] D.T. Jones, W.R. Taylor, J.M. Thornton, Biochemistry 33 (1994) 3038-3049.

[17] B. Rost, R. Casadio, P. Fariselli, C. Sander, Protein Sci. 4 (1995) 521.

[18] T. Hirokawa, S. Boon-Chieng, S. Mitaku, Bioinformatics 14 (1998) 378.

[19] G.E. Tusnady, I. Simon, J. Mol. Biol. 283 (1998) 489.

[20] U.H.E. Hansmann, Y. Okamoto, in Annual Reviews of Computational Physics VI, edited by D. Stauffer (World Scientific, Singapore, 1999), p.129.

[21] A. Mitsutake, Y. Sugita, Y. Okamoto, Biopolymers (Pept. Sci.) 60 (2001) 96.

[22] K. Hukushima, K. Nemoto, J. Phys. Soc. Jpn. 65 (1996) 1604.

[23] K. Hukushima, H. Takayama, K. Nemoto, Int. J. Mod. Phys. C 7 (1996) 337.

[24] C.J. Geyer, Proceedings of the 23rd Symposium on the Interface, edited by E. Keramidas (Interface Foundation, Fairfax Station, 1991) p. 156.

[25] E. Marinari, G. Parisi, J.J. Ruiz-Lorenzo, in Spin Glasses and Random Fields, edited by A.P. Young (World Scientific, Singapore, 1998), p. 59.

[26] A.M. Ferrenberg, R.H.Swendsen, Phys. Rev. Lett. 63 (1989) 1195.

[27] S. Kumar, D. Bouzida, R.H. Swendesen, P.A. Kollman, J.M. Rosenberg, J. Comput. Chem. 13 (1992) 1011.

[28] K.R. MacKenzie, J.H. Prestegard, D.M. Engelman, Science 276 (1997) 131.

Chapter 2

Simulation Methods

H. Kokubo and Y. Okamoto, “Prediction of transmembrane helix configurations by replica-exchange simulations,” *Chemical Physics Letters* **383**, 397-402 (2004).

2.1 Simulation Protocols

In this section, we explain our prediction methods. Our method consists of two parts. In the first part, we obtain amino-acid sequences of the transmembrane helix regions of the target protein. Amino-acid sequences of the transmembrane helix regions can be predicted by analyzing mainly the hydrophobicity of the amino-acid sequences, without having any information about higher-order structures. There already exist many tools on WWW servers for this purpose such as TMHMM [1], MEMSAT [2], SOSUI [3], and HMMTOP [4]. Given the amino-acid sequence of a protein, they judge whether the protein is a membrane protein or not and (if yes) predict the regions in the amino-acid sequence that correspond to the transmembrane helices. However, the precision of these programs about amino-acid sequence information of transmembrane regions is about 85 % at present and needs improvement. We thus focus our attention on the effectiveness of the second part of our method, leaving this improvement to the developers of the WWW servers. Namely, we use the experimentally known amino-acid sequence of transmembrane regions (without relying on the WWW servers) and try to predict their native conformations.

In the second part, we assume transmembrane parts as helices and construct ideal canonical α-helices (3.6 residues per turn) of the sequences. We perform the replica-exchange simulations using these transmembrane helices with atomistic details and identify the global-minimum (free) energy structure as the predicted structure. In the above simulations the constraint energy terms explained in Sec. 2.2 are used, and the replica-exchange method is explained in Sec. 2.3, which is the method for obtaining the global-minimum free energy state without getting trapped in one of the local-minimum free energy states.

The following three approximations are introduced. Approximation (1) is that the backbone structures of the α-helices are treated as rigid body and only side-chain structures are made flexible. Each helix backbone thus has the freedom of rigid translation and rigid rotation only (in the future we will introduce some flexibility in the backbone structures). This is introduced following the two-stage model, in which each helix is stable as a domain and the native configurations are built mainly by the interactions between helices. We believe that the flexibility of side chains is also important because membrane proteins are very tightly packed and the packed structures are searched by varying side-chain structures. Approximation (2) is that only the transmembrane parts are used in the simulation and the rest of the amino acids of the membrane protein (such as loop regions) are neglected. The environment inside the membrane is very hydrophobic and that outside the membrane is very hydrophilic and both environments are much different, therefore, we assume that the regions inside the membrane form stable structures by themselves and Approximation (2) follows. Approximation (3) is that surrounding molecules such as lipids are not used in the simulations. This approximation is introduced in order to save computation time. If we perform the simulations using explicit lipids, large computational time will be necessary and it is difficult to search the wide conformational space. This approximation is justified by the two-stage model again, which implies that helix-helix interactions (and not helix-lipid interactions) are the main driving force of the structure formation. In the future we plan to treat the effects of surrounding lipids and water molecules more accurately. We remark that a generalized Born theory of lipids has been recently introduced [5].

In principle, we can also use molecular dynamics (MD) method, but we employ Monte Carlo (MC) algorithm here. We update configurations with rigid translations and rigid rotations of each α-helix and torsion rotations of side chains.

2.2 Constraint Energy for Membrane Proteins

We use a standard force field such as CHARMM [6, 7] for the potential energy of the system and add the following three simple harmonic constraints to the original force-field potential energy in order to make conformational sampling efficiency better and mimic the effect of membrane planes:

$$E_{\mathrm{constr}} = E_{\mathrm{constr1}} + E_{\mathrm{constr2}} + E_{\mathrm{constr3}}. \tag{2.1}$$

Here, E_{constr1} is the energy that constrains pairs of adjacent helices along the amino-acid chain not to be apart from each other too much (loop constraints) and is defined as follows:

$$E_{\mathrm{constr1}} = \sum_{i=1}^{N_{\mathrm{H}}-1} k_1 \; \theta\left(r_{i,i+1} - d_{i,i+1}\right) \left[r_{i,i+1} - d_{i,i+1}\right]^2 , \tag{2.2}$$

where N_{H} is the total number of transmembrane helices in the protein, $r_{i,i+1}$ is the distance between the C atom of the C-terminus of the i-th helix and the N atom of 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 constraints, respectively, and $\theta(x)$ is the step function:

$$\theta(x) = \begin{cases} 1 \ , & \text{for } x \geq 0 \ , \\ 0 \ , & \text{otherwise} \ . \end{cases} \tag{2.3}$$

This term has a non-zero value only when the distance $r_{i,i+1}$ becomes longer than $d_{i,i+1}$. Only the structures in which the distance between helices is short are searched because of this constraint term. Our purpose is to find the optimal packed configurations of helices, therefore it is reasonable to set this constraint energy.

The second term in Eq. (2.1) is the energy that constrains the helix N-terminus and C-terminus to be located near membrane boundary planes and is defined as follows:

$$\begin{aligned} E_{\mathrm{constr2}} = & \sum_{i=1}^{N_{\mathrm{H}}} \left\{ k_2 \, \theta\left(\left| z_i^{\mathrm{L}} - z_0^{\mathrm{L}} \right| - d_i^{\mathrm{L}} \right) \left[\left| z_i^{\mathrm{L}} - z_0^{\mathrm{L}} \right| - d_i^{\mathrm{L}} \right]^2 \right. \\ & + \left. k_2 \, \theta\left(\left| z_i^{\mathrm{U}} - z_0^{\mathrm{U}} \right| - d_i^{\mathrm{U}} \right) \left[\left| z_i^{\mathrm{U}} - z_0^{\mathrm{U}} \right| - d_i^{\mathrm{U}} \right]^2 \right\} , \end{aligned} \tag{2.4}$$

where k_2 is the force constant of the harmonic constraints, z_i^{L} and z_i^{U} are the z-coordinate values of the C_α (or C) atom of the N-terminus (or C-terminus) of the i-th helix near the fixed lower boundary value z_0^{L} and the upper boundary value z_0^{U} of the membrane, respectively, and d_i^{L} and d_i^{U} are the corresponding central value constants of the harmonic constraints. Here, the z-axis is defined to be the direction perpendicular to the membrane boundary planes. This term has a non-zero value when the C atom of each helix C-terminus or C_α atom of each helix N-terminus is apart more than d_i^{L} (or d_i^{U}). This constraint energy was introduced so that the helix ends are not too much apart from the membrane boundary planes.

The third term in Eq. (2.1) is the energy that constrains all C_α atoms within the sphere (centered at the origin) of radius d_{C_α} and is defined as follows:

$$E_{\mathrm{constr3}} = \sum_{\mathrm{C}_\alpha} k_3 \, \theta\left(r_{\mathrm{C}_\alpha} - d_{\mathrm{C}_\alpha} \right) \left[r_{\mathrm{C}_\alpha} - d_{\mathrm{C}_\alpha} \right]^2 , \tag{2.5}$$

where r_{C_α} are the distance of C_α atoms from the origin, and k_3 and d_{C_α} are the force constant and the central value constant of the harmonic constraints, respectively. This term has a non-zero value only when C_α atoms go out of this sphere. The term is introduced so that the center of mass of the molecule stays near the origin. The radius of the sphere is set to a large value in order to guarantee that a wide configurational space is sampled.

These three constraint terms do not impose any constraints on the possible structures as membrane proteins if the constraint constants of d_i^L, d_i^U, and d_{C_α} are set large enough as we can understand from the fact that the step function is used.

2.3 Replica-Exchange Method

We now briefly review the replica-exchange method (REM) [8]–[10] (see Refs. [11, 12] for details). The system for REM consists of M non-interacting copies (or, replicas) of the original system in the canonical ensemble at M different temperatures T_m ($m = 1, \cdots, M$). Let $X = (\cdots, x_m^{[i]}, \cdots)$ stand for a state in this ensemble. Here, the superscript i and the subscript m in $x_m^{[i]}$ label the replica and the temperature, respectively. The state X is specified by M sets of $x_m^{[i]}$, which in turn is specified by the coordinates $q^{[i]}$ of all the atoms in replica i. A REM simulation is then realized by alternately performing the following two steps. Step 1: Each replica in canonical ensemble of the fixed temperature is simulated simultaneously and independently for a certain MC or MD steps. Step 2: A pair of replicas, say i and j, which are at neighboring temperatures T_m and T_n, respectively, are exchanged: $X = (\cdots, x_m^{[i]}, \cdots, x_n^{[j]}, \cdots) \rightarrow X' = (\cdots, x_m^{[j]}, \cdots, x_n^{[i]}, \cdots)$. The transition probability of this replica exchange is given by the Metropolis criterion:

$$w(X \rightarrow X') \equiv w(x_m^{[i]} | x_n^{[j]}) = \left\{ \begin{array}{ll} 1 \ , & \text{for } \Delta \leq 0, \\ \exp(-\Delta) \ , & \text{otherwise,} \end{array} \right. \tag{2.6}$$

where

$$\Delta = (\beta_m - \beta_n) \left(E(q^{[j]}) - E(q^{[i]}) \right) \ . \tag{2.7}$$

Here, $E(q^{[i]})$ and $E(q^{[j]})$ are the potential energy of the i-th replica and the j-th replica, respectively. In the present work, we employ Monte Carlo algorithm in Step 1. There are $2N_\mathrm{H} + N_\mathrm{D}$ kinds of MC moves, where N_D is the total number of dihedral angles in the side chains of N_H helices. The first term corresponds to the rigid translation and rigid rotation of the helices and the second to the dihedral-angle rotations in the side chains.

One MC step is defined to be an update of one fo these degrees of freedom, which is accepted or rejected according to the Metropolis criterion. One MC sweep is defined to consist of $2N_{\rm H} + N_{\rm D}$ updates that are randomly chosen from these MC moves with the Metropolis evaluation for each update (hence, it consists of $2N_{\rm H} + N_{\rm D}$ MC steps). We predict the native structure of membrane spanning regions as the global-minimum free energy state obtained by the REM simulations. The REM simulations can avoid getting trapped in one of the local-minimum (free) energy states because the temperature of each replica goes up and down by the temperature exchange. Therefore, this method is very useful for simulating systems which have multiple-minimum states such as proteins.

In principle, we can identify not only the global-minimum free energy state but also all the local-minimum free energy states from the replica-exchange simulations.

2.4 Calculation of Canonical Expectation Values

From only one REM simulation run, one can obtain not only the global-minimum structure but also canonical-ensemble averages of physical quantities as functions of temperature by using the multiple-histogram reweighting techniques [13, 14] (which is also referred to as Weighted Histogram Analysis Method, or WHAM [14]) (see also [15]) as follows. Suppose we have made M independent simulation runs at M different temperatures. Let $N_m(E)$ and n_m be the energy histogram and the total number of samples obtained at temperature T_m, respectively. The expectation value of a physical quantity A at any intermediate temperature T is given by

$$< A >_T = \frac{\sum\limits_E A(E) n(E) \exp(-\beta E)}{\sum\limits_E n(E) \exp(-\beta E)}, \tag{2.8}$$

where the density of states $n(E)$ is obtained by solving the following WHAM equations:

$$n(E) = \frac{\sum\limits_{m=1}^{M} g_m^{-1} N_m(E)}{\sum\limits_{m=1}^{M} g_m^{-1} n_m \exp(f_m - \beta_m E)} \ , \tag{2.9}$$

and

$$\exp(-f_m) = \sum_E n(E) \exp(-\beta_m E) \ . \tag{2.10}$$

Here, $g_m = 1+2\tau_m$, and τ_m is the integrated autocorrelation time at temperature T_m. For biomolecular systems the quantity g_m can safely be set to be a constant in the reweighting formulae [14], and so we set $g_m = 1$ throughout the analyses in the present work. Note that $n(E)$ and f_m are solved self-consistently by iteration. Moreover, ensemble averages of any physical quantity A (including those that cannot be expressed as functions of potential energy) can now be obtained from the "trajectory" of configurations of the production run. Namely, we first obtain f_m $(m = 1, \cdots, M)$ by solving Eqs. (2.9) and (2.10) self-consistently, and then we have [15]

$$< A >_T = \frac{\displaystyle\sum_{m=1}^{M} \sum_{x_m} A(x_m) \frac{g_m^{-1}}{\displaystyle\sum_{\ell=1}^{M} g_\ell^{-1} n_\ell \exp(f_\ell - \beta_\ell E(x_m))} \exp(-\beta E(x_m))}{\displaystyle\sum_{m=1}^{M} \sum_{x_m} \frac{g_m^{-1}}{\displaystyle\sum_{\ell=1}^{M} g_\ell^{-1} n_\ell \exp(f_\ell - \beta_\ell E(x_m))} \exp(-\beta E(x_m))} , \tag{2.11}$$

where x_m are the configurations at temperature T_m. Here, the trajectories x_m are taken for each temperature T_m separately.

In Chapter 4, we use more naive calculation method in order to calculate free energy differences. The canonical expectation value of a physical quantity A at temperature T_m $(m = 1, ..., M)$ can be calculated by averaging a physical quantity of the replica which has the temperature $T = T_m$ at MC step time t as follows:

$$< A >_{T_m} = \frac{1}{N_{sim}} \sum_{t=1}^{N_{sim}} A(x_m^i(t)), \tag{2.12}$$

where i is the replica number which has the temperature $T = T_m$ at MC step time t and N_{sim} is the total number of measurements.

2.5 Principal Component Analysis

Here we explain the procedure of the principal component analysis (PCA) [16]-[20] briefly. At first N conformations are chosen from a simulation. The structures are taken from the trajectory at even intervals, and we store N structures in total. Each structure is superimposed on the arbitrary reference structure. In this work we choose the structure of the NMR experiments as the reference structure. We then calculate the average structure of N structures and superimpose these N structures on this average structure. We define the following variance-covariance matrix:

$$C_{ij} =< (\vec{q}- < \vec{q} >)_i(\vec{q}- < \vec{q} >)_j >, \tag{2.13}$$

where $\vec{q} = (q_1, q_2, q_3, q_4, q_5, q_6, ..., q_{3n-2}, q_{3n-1}, q_{3n}) = (x_1, y_1, z_1, x_2, y_2, z_2, ..., x_n, y_n, z_n)$ and $< \vec{q} >= \Sigma_{k=1}^{N}\vec{q}(k)/N$, x_i, y_i, z_i are Cartesian coordinates of the i-th atom, and n is the total number of atoms. This symmetric matrix is diagonalized and the eigenvectors and eivenvalues are obtained. The first superposition is performed in order to remove large eigenvalues from the translations and rotations of the system because we want to analyze the differences of structures. Therefore, the six eigenvalues from the smallest one are very close to zero within the limit of arithmetic precision of a computer ($\sim 1.0 \times 10^{-12}$). We order the eigenvalues in the decreasing order of magnitude. The first and second principal component axes are thus defined as the eigenvectors corresponding to the largest and second-largest eigenvalues, respectively. The i-th principal component of each sampled structure is defined by the following inner product:

$$\mu_i = \vec{v}_i \cdot (\vec{q}- < \vec{q} >), \quad (i = 1, 2, ...), \tag{2.14}$$

where $\vec{v}_i$ is the i-th eigenvector.

Bibliography

[1] A. Krogh, B. Larsson, G.v. Heijne, E.L.L. Sonnhammer, J. Mol. Biol. 305 (2001) 567.

[2] D.T. Jones, W.R. Taylor, J.M. Thornton, Biochemistry 33 (1994) 3038.

[3] T. Hirokawa, S. Boon-Chieng, S. Mitaku, Bioinformatics 14 (1998) 378.

[4] G.E. Tusnady, I. Simon, J. Mol. Biol. 283 (1998) 489.

[5] W. Im, M. Feig, C.L. Brooks III, Biophys. J. 85 (2003) 2900.

[6] W.E. Reiher, III, Theoretical Studies of Hydrogen Bonding, Ph.D. Thesis, Department of Chemistry, Harvard University, Cambridge, MA,USA, 1985

[7] E. Neria, S. Fischer, M. Karplus, J. Chem. Phys. 105 (1996) 1902.

[8] K. Hukushima, K. Nemoto, J. Phys. Soc. Jpn. 65 (1996) 1604.

[9] K. Hukushima, H. Takayama, K. Nemoto, Int. J. Mod. Phys. C 7 (1996) 337.

[10] C.J. Geyer, Proceedings of the 23rd Symposium on the Interface, edited by E. Keramidas (Interface Foundation, Fairfax Station, 1991) p. 156.

[11] Y. Sugita, Y. Okamoto, Chem. Phys. Lett. 314 (1999) 141.

[12] A. Mitsutake, Y. Sugita, Y. Okamoto, Biopolymers (Pept. Sci.) 60 (2001) 96.

[13] A.M. Ferrenberg, R.H.Swendsen, Phys. Rev. Lett. 63 (1989) 1195.

[14] S. Kumar, D. Bouzida, R.H. Swendesen, P.A. Kollman, J.M. Rosenberg, J. Comput. Chem. 13 (1992) 1011.

[15] A. Mitsutake, Y. Sugita, Y. Okamoto, J. Chem. Phys. 118 (2003) 6664.

[16] M.M. Teeter, D.A. Case, J. Phys. Chem. 94 (1990) 8091.

[17] A. Kitao, F. Hirata, N. Go, Chem. Phys. 158 (1991) 447.

[18] A.E. Garcia, Phys. Rev. Lett. 68 (1992) 2696.

[19] R. Abagyan, P. Argos, J. Mol. Biol. 225 (1992) 519.

[20] A. Amadei, A.B.M. Linssen, H.J.C. Berendsen, Proteins 17 (1993) 412.

Chapter 3

Structure Prediction of Glycophorin A Transmembrane Dimer

H. Kokubo and Y. Okamoto, "Prediction of transmembrane helix configurations by replica-exchange simulations," *Chemical Physics Letters* **383**, 397-402 (2004).
H. Kokubo and Y. Okamoto, "Prediction of membrane protein structures by replica-exchange Monte Carlo simulations: case of two helices," *The Journal of Chemical Physics* **120**, 10837-10847 (2004).

3.1 Introduction

The dimeric transmembrane domain of glycophorin A is often used as a model system of helix-helix interaction of membrane proteins [1, 2]. In Ref. [1], Adams *et al.* performed many simulated annealing simualtions and tried to obtain the global-minimum energy structure. However, it is possible to get trapped in one of the local-minimum energy states if we use the simulated annealing method. Therefore we performed the replica-exchange simulations, which allow us to sample sufficiently without getting trapped in the initial or the local-minimum-energy states. Moreover, the simulated annealing method cannot be used to obtain the canonical-ensemble averages of physical quantities, but we can calculate them by the histogram reweighting techniques from the trajectories of the replica-exchange simulations. In fact we analyze several canonical physical quantities in subsection 3.3.2 and show which energy terms are important. Their prediction method needs the experimental mutagenesis data to determine the predicted structure, although out method can predict the native structure only by the simulations without the experimental data. In Ref. [2], Pappu *et al.* use the diffusion equation method proposed by Scheraga *et al.* and this method needs that the force field consisits of Gaussian-shaped functions. Therefore they developed their own Gaussian shaped force field, which were made by parameterizing to replace the pairwise Lennnard-Jones terms of the original OPLS force fields by a sum of two Gaussians, and the electrostatic terms were ignored. However we think that their procedure will be not be justified and there are some doubts whether their force field has the versatility or not. On the other hand, our method can use the original force field.

In this chapter, we test the second part of our prediction method, which has been described in Chapter 2, using this membrane protein. Namely, given the amino-acid sequences of the transmembrane helices of glycophorin A, we performed a replica-exchange Monte Carlo simulation to predict the helix configurations. Preliminary results have been already reported elsewhere [3]. Here, we give the details of the present approach and results.

In Sec. 3.2 the detail conditions of the REM simulations for the dimeric transmembrane domain of glycophorin A are explained. In Sec. 3.3 the results of the application to the structure prediction of the dimeric transmembrane domain of glycophorin A are given.

3.2 Computational Details

Our method consists of two parts. In the first part, we obtain amino-acid sequences of the transmembrane helix regions from existing WWW servers such as those in Refs. [4, 5, 6, 7]. However, the precision of these programs in the WWW servers is about 85 % and needs improvement. We thus focus our attention on the effectiveness of the second part of our method, leaving this improvement to the developers of the WWW servers. Namely, we use the experimentally known amino-acid sequence of transmembrane regions (without relying on the WWW servers) and try to predict their conformations, following the prescription of the second part of our method described in the previous chapter. Here, we chose one of the simplest systems: the transmembrane dimer of glycophorin A (PDB code: 1AFO). The number of amino acids for each helix is 18 and the sequence is TLIIFGVMAGVIGTILLI.

We first constructed the ideal canonical α-helix (3.6 residues per turn) of this sequence. The N and C termini of this helix were blocked with acetyl and N-methyl groups, respectively. The force field that we used is the CHARMM param19 parameter set (polar hydrogen model) [8, 9]. No cutoff was introduced to the non-bonded energy terms, and the dielectric constant ϵ was set equal to 1.0. We have also studied the case of $\epsilon = 4.0$, because it is the value close to that for the lipid environment. The computer code based on the CHARMM macromolecular mechanics program [10] was used and the replica-exchange method was implemented in it. This helix structure was minimized subject to harmonic restraints on all the heavy atoms. The initial configuration for the REM simulation was that two α-helices of identical sequence and structure thus prepared were placed in parallel at a distance of 20 Å. These helices are quite apart from each other and the starting configuration is indeed very different from the native one. Note that the only information derived from the NMR experiments [11] is the amino-acid sequence of the individual helices.

The values of the constants for the constraints in Eq. (2.1)-(2.5) were set as follows: $N_{\mathrm{H}} = 2$, $k_1 = k_2 = 0.5$ kcal/(mol Å^2), $k_3 = 0.05$ kcal/(mol Å^2), $d_{i,i+1} = 20$ Å, $z_0^{\mathrm{L}} = -13.35$ Å, $z_0^{\mathrm{U}} = +13.35$ Å, $d_i^{\mathrm{L}} = d_i^{\mathrm{U}} = 1.0$ Å, and $d_{\mathrm{C}\alpha} = 50$ Å. The values for z_0^{L} and z_0^{U} were taken from the z-coordinates of the initial configuration (in Fig. 3.3(a) below; the z axis is placed vertically in the figure). In the present example of glycophorin A dimer, the first term in Eqs. (2.2) was imposed on both terminal ends (i.e., two kinds of $r_{i,i+1}$ were prepared: one is the distance between a pair of N atoms at the N-terminus of the two helices and the other is the distance between a pair of C atoms at the C-terminus of the two helices). As explained in Sec. 2.2, the constraint terms do not impose any constraints on the possible structures as membrane proteins if the constraint constants are set properly as we can understand from the fact that the step function is used.

We performed two REM MC simulations of 1,000,000 MC sweeps, starting from the parallel configuration of Fig. 3.3(a) below: one with the dielectric constant $\epsilon = 1.0$ and the other with $\epsilon = 4.0$. We used the following 13 temperatures: 200, 239, 286, 342, 404, 489, 585, 700, 853, 1041, 1270, 1548, and 1888 K, which are distributed almost exponentially. The highest temperature was chosen sufficiently high so that no trapping in local-minimum-energy states occurs. This temperature distribution was chosen so that all the acceptance ratios are almost uniform and sufficiently large ($>$ 10 %) for computational efficiency. Replica exchange was attempted once at each MC sweep.

3.3 Results and Discussion

3.3.1 Performances of the Replica-Exchange Simulations

We first examine whether the present REM simulations performed properly. The acceptance ratios of replica exchange are listed in Table 3.1 for both cases of dielectric constants. We see that the acceptance ratios of replica exchange between all pairs of neighboring temperatures are uniform and large enough ($>$ 10 %) for computational efficiency. The results in Table 3.1 imply that one should observe a free random walk in the replica space and temperature space.

In Figs. 3.1 and 3.2 we show the "time series" of the present REM simulations with the dielectric constant $\epsilon = 1.0$ and $\epsilon = 4.0$, respectively. In Figs. 3.1(a) and 3.2(a) the "time series" of replica exchange at the lowest temperature ($T = 200$ K) are shown. We see that every replica takes the lowest temperature many times, and we indeed observe a random walk in the replica space. The complementary picture to this is the temperature exchange for each replica. The results for one of the replicas (Replica 6) are shown in Figs. 3.1(b) and 3.2(b). We again observe random walks in the temperature space between the lowest and highest temperatures. Other replicas perform random walks similarly. In Figs. 3.1(c) and 3.2(c) the corresponding time series of the total potential energy are shown. We see that random walks in the potential energy space between low and high energy regions are also realized. Note that there is a strong correlation between the behaviors in Figs. 3.1(b) and 3.1(c) as there should. The same is true for Figs. 3.2(b) and 3.2(c). All these results confirm that the present REM simulations have been properly performed.

We now study how widely the configurational space was sampled during the present simulations. We first examine the case for $\epsilon = 1.0$. We plot the time series of the root-mean-square (RMS) deviation of the backbone atoms from the NMR structure [11] in Fig. 3.1(d). When the temperature becomes high, the RMS deviation takes a large value (the largest value in Fig. 3.1(d) is 14.1 Å, and the maximum value among all the replicas is 15.7 Å), and when the temperature becomes low, the RMS deviation takes a small value (the smallest value in Fig. 3.1(d) is 0.55 Å, and the minimum value among all the replicas is also 0.55 Å). By comparing Fig. 3.1(c) and Fig. 3.1(d), we see that there is a strong correlation between the total potential energy and the RMS deviation values. In particular, it is remarkable that when the energy is the lowest (around -1490 kcal/mol), most of the RMS values are as small as about 0.5 Å. This implies that the global-minimum-energy state is indeed very close to the native structure.

We now examine the case with $\epsilon = 4.0$. We plot the time series of the RMS deviation of the backbone atoms from the NMR structure [11] in Fig. 3.2(d). When the temperature becomes high, the RMS deviation takes a large value (the largest value in Fig. 3.2(d) is 14.2 Å, and the maximum value among all the replicas is 14.8 Å). The RMS deviation sometimes takes a small value (the smallest value in Fig. 3.2(d) is 0.58 Å, and the minimum value among all the replicas is 0.56 Å), and when this occurs, the temperature is low and the potential energy takes a small value. When the temperature becomes low and the potential energy is low, however, the RMS deviation is not always small and takes several values (around 0.7 Å, 2.5 Å, 3.5 Å, 4.4 Å, or 5.7 Å) contrary to the case with $\epsilon = 1.0$. This implies that there are several stable structures at low temperatures. We will discuss this matter more in detail below.

In Fig. 3.3 typical snapshots from the REM simulations of Figs. 3.1 and 3.2 are shown. Fig. 3.3(a) is the initial configuration of our simulations, in which the two helices are placed in parallel. Figs. 3.3(b1)-3.3(b4) and Figs. 3.3(c1)-3.3(c4) are typical snapshots with the dielectric constant $\epsilon = 1.0$ and $\epsilon = 4.0$, respectively. These figures confirm that our simulations indeed sampled wide configurational space. We see that the REM simulations perform random walks not only in energy space but also in conformational space and that they do not get trapped in one of a huge number of local-minimum-energy states.

3.3.2 Canonical Probability Distributions and Averages of the Potential Energy Terms

In Fig. 3.4(a) and Fig. 3.5(a) the canonical probability distributions of the total potential energy obtained at the chosen 13 temperatures from the REM simulation with the dielectric constant $\epsilon = 1.0$ and $\epsilon = 4.0$ are shown, respectively. We see that there are enough overlaps between all neighboring pairs of distributions, indicating that there will be sufficient numbers of replica exchange between pairs of replicas. In Fig. 3.4(b) (the case for $\epsilon = 1.0$) and Fig. 3.5(b) (the case for $\epsilon = 4.0$), the average of the total potential energy $E_{\rm tot}$ and averages of its component terms, namely, the electrostatic energy $E_{\rm c}$, van der Waals energy $E_{\rm v}$, torsion energy $E_{\rm t}$ (these three terms are from the CHARMM force field), and constraint energy $E_{\rm cons}$ as a function of temperature T are shown. The multiple-histogram reweighting techniques in Eq. (2.11) were used for these calculations. We see that as the temperature becomes low, $E_{\rm tot}$ becomes low mainly because $E_{\rm v}$ and $E_{\rm t}$ become low. The changes of $E_{\rm c}$ and $E_{\rm cons}$ as the temperature is varied are small and the contribution of $E_{\rm c}$ and $E_{\rm cons}$ can be said to be smaller on the average than those of $E_{\rm v}$ and $E_{\rm t}$. In the case for $\epsilon = 4.0$, this temperature variation of $E_{\rm c}$ is particularly small and less than 1.0 kcal/mol.

3.3.3 Comparison of the Predicted Structures and the Experimental Structure

In Fig. 3.6 the configuration obtained by the NMR experiments [11] and the global-minimum-energy configurations with the dielectric constant $\epsilon = 1.0$ and $\epsilon = 4.0$ obtained by the REM simulations are compared. The predicted structure with $\epsilon = 1.0$ is in remarkable agreement with that from the NMR experiments. At first sight, it is rather surprising that the result with $\epsilon = 1.0$ is much closer to the experimental result than that with $\epsilon = 4.0$, because the dielectric constant for a lipid system is closer to 4.0 than to 1.0. However, on second thoughts we understand that the present results are reasonable because the pairs of helices in transmembrane proteins are tightly packed and almost no lipid molecules can exist between helices. This implies that helix-helix interactions are the main driving force in the final stage of the structure formation of membrane proteins.

For further understanding we compare some properties of the three structures from Fig. 3.6 in Table 3.2. We see that the structure with $\epsilon = 4.0$ is stabilized by $E_{\rm v}$ and $E_{\rm t}$ compared with the structure with $\epsilon = 1.0$. The difference of $E_{\rm cons}$ between the structures of $\epsilon = 1.0$ and $\epsilon = 4.0$ is smaller than those of $E_{\rm v}$ and $E_{\rm t}$ (0.62 kcal/mol versus 5.4 and 2.0 kcal/mol). Thus the existence of the constraint terms is not the major cause for the difference between $\epsilon = 1.0$ and $\epsilon = 4.0$. The predicted structure with $\epsilon = 4.0$

had smaller solvent accessible surface area and was more packed than the native one. This means that although only four hydrophilic amino acids are included in 36 amino acids used in our simulations, the electrostatic energy term contributes to the stability of this membrane protein and forces the native structure to be a little less packed than the case with weakened electrostatic interactions. In other words, the stability of the native structure is determined by the balance of E_{c}, E_{v}, and E_{t}, and the contribution of electrostatic energy is also important. The interhelical crossing angle in this table is defined as the angle between the principal axes of moment of inertia for each helix. From the interhelical crossing angle, the structure with $\epsilon = 4.0$ is near parallel. This can also be understood from Fig. 3.6(c). One of the helices in the structure with $\epsilon = 4.0$ appears to be slightly off from the membrane boundary. We see that the structure with $\epsilon = 1.0$ is indeed close to the native structure in every property.

In Table 3.3 the interhelical distances of the three structures in Fig. 3.6 are compared with those of the solid-state NMR experiments [12]. The structure obtained from the solid-state NMR experiments is considered to be closer to the native structure than those obtained from the solution NMR experiments because it is crystallized in lipid bilayer as in the native state. On the other hand, the solution NMR structures were determined in detergent micelles. In the case for $\epsilon = 1.0$, three distances out of six are between the solid-state NMR values and the solution NMR values and the distance between Gly79 C and Val80 C is closer to the solid-state NMR experiments. This suggests that the predicted structure in Fig. 3.6(b) is closer to the native structure than to the solution NMR structure in Fig. 3.6(a). The distances of the structure with $\epsilon = 4.0$ are totally different from those of the solid-state and solution NMR structures as we can expect from Fig. 3.6(c).

3.3.4 Average Physical Quantities as Functions of Temperature

In Fig. 3.7 we show the average values of the RMS deviation, the radius of gyration, the interhelical crossing angle, and the solvent accessible surface area as functions of temperature T for the REM simulations with $\epsilon = 1.0$. The multiple-histogram reweighting techniques of Eq. (2.11) were used again. In Fig. 3.7(a) we see that the average RMSD decreases monotonically as the temperature is lowered. This means that when the temperature is high the structures that are very distant from the native one are often sampled (RMSD is as large as 8.0 Å) and when the temperature is low the structures that are close to the native one are mainly sampled (RMSD is as small as 1.0 Å). Fig. 3.7(b) implies that the packed conformations are searched when the temperature is low and disjointed structures are searched when the temperature is high. When the temperature is 200 K, the average value of the radius of gyration is close to the native one (10.0 Å; see Table 3.2). In Fig. 3.7(c) the average interhelical crossing angle is about 50 degrees at low temperatures and about 40 degrees at high temperatures. The crossing angle at the low temperature (about 50 degrees) is a little larger (about 5 degree) than that of the native structure (45.8°; see Table 3.2). From Fig. 3.7(d) we see that the average solvent accessible surface area is large when the temperature is high and it is small when the temperature is low. This is reasonable because the packed conformations with small surface area are searched at low temperatures and disjointed conformations are searched at high temperatures. When the temperature is 200 K, the value of the surface area is indeed close to the native one (3152Å^2; see Table 3.2).

Similarly in Fig. 3.8, we show the corresponding quantities of Fig. 3.7, which were calculated from the results for the case of $\epsilon = 4.0$. In Fig. 3.8(a) we see that the average RMS deviation is always more than 3.0 Å and the simulation often samples different structures from the native one. The RMS deviation at 600 K is smaller than that at 200 K, suggesting that structures more similar to the native one are sampled. The similar nature is observed in Figs. 3.8(b), 3.8(c), and 3.8(d). Namely, the quantities around 600 K are closer to those of the native structure than at other temperatures. This point is discussed further below around Fig. 3.10. The values at high temperatures in Fig. 3.8 are similar to the corresponding ones in Fig. 3.7, because the simulations search various structures without getting trapped in one or a few of local-minimum-energy structures. At low temperatures the simulations sample one or a few of them. In Fig. 3.8(a) the average RMS deviation at 200 K is close to that of the structure in Fig. 3.6(c) (see Table 3.2). The averages of the radius of gyration (Fig. 3.8(b)), the interhelical crossing angle (Fig. 3.8(c)), and solvent accessible surface area (Fig. 3.8(d)) also take values close to those of the structure in Fig. 3.6(c). This implies that at the lowest temperature of 200 K, only the structure in Fig. 3.6(c) is mostly sampled.

3.3.5 Canonical Probability Distributions of RMSD

In Figs. 3.7 and 3.8 we saw average quantities as functions of temperature. Here, we study how many dominant structures contribute to the averages by examining the probability distribution of RMSD. In Fig. 3.9 the results with $\epsilon = 1.0$ at four temperatures are shown. From Fig. 3.9(a) we see that the structures close to the native one (RMSD=0.7 Å) are mainly sampled at T=200 K. There also exists a small contribution around RMSD=2.5 Å. This cannot be understood from the average properties. Actually, we do observe this second structure in Fig 3.1. Namely, we see that the structures around RMSD=2.5 Å are certainly sampled at low temperatures (compare Figs. 3.1(b) and 3.1(d)). Similar behavior, although less conspicuous, is observed at T=342 K (Fig. 3.9(b)). As the temperature becomes higher, the distributions become broader. The two peaks become almost equally important at T=585 K (Fig. 3.9(c)). This suggests that structures are sampled around only two local-minimum-energy states below 585 K. At the highest temperature of 1888 K (Fig. 3.9(d)) we no longer see any peaks in the histogram and various structures are sampled, and the simulation does not get trapped in local-minimum-energy structures at this temperature.

The corresponding probability distribution for the case of $\epsilon = 4.0$ are shown in Fig. 3.10. From Fig. 3.10(a) the situation is more complicated than the case of $\epsilon = 1.0$. We see that the structures around RMSD=4.5 Å are mainly sampled at 200 K. However in this case as many as four other contributions also exist (RMSD $\sim$ 0.7 Å, 2.5 Å, 3.5 Å, and 5.7 Å). As the temperature becomes high (Fig. 3.10(b) or Fig. 3.10(c)) from 200 K (Fig. 3.10(a)), the peak around RMSD=4.5 Å becomes small, and the four peaks (RMSD $\sim$ 0.7 Å, 2.5 Å, 4.5 Å, and 5.7 Å) become almost equally important. This is the reason why the average values around 600 K are closer to those of the native structure than those around 200 K in Fig. 3.8. These results show that the simulation with $\epsilon = 4.0$ also samples the native structure as one of the local-minimum-energy states (RMSD=0.7 Å), but it is not the dominant contributions at the lowest temperature.

We list the average properties of the structures that correspond to the peaks of the histograms in Figs. 3.9(a) and 3.10(a) in Table 3.4. The global-minimum-energy structure for $\epsilon = 1.0$ in Fig. 3.6(b) belongs to peak1 for $\epsilon = 1.0$ in Table 3.4, and that for $\epsilon = 4.0$ in Fig. 3.6(c) belongs to peak4 for $\epsilon = 4.0$ in Table 3.4. We see that not only the RMS deviation but also radius of gyration, interhelical crossing angle, and solvent accessible surface area of peak1 in both cases of $\epsilon = 1.0$ and $\epsilon = 4.0$ are similar to one another. Hence, we conclude that the structures of peak1 for $\epsilon = 1.0$ are essentially identical with those of peak1 for $\epsilon = 4.0$ (they correspond to the native structure, see Table 3.2). Likewise, the structures of peak2 for $\epsilon = 1.0$ are the same as those of peak2 for $\epsilon = 4.0$. Therefore, we can say that the low-energy structures that were sampled in the simulation with $\epsilon = 1.0$ are subsets of structures of those that were sampled with $\epsilon = 4.0$. However, only the case with $\epsilon = 1.0$ gives the native structures as the global-minimum-energy state.

Bibliography

[1] P.D. Adams, D.M. Engelman, A.T. Brünger, Proteins 26 (1996) 257.

[2] R.V. Pappu, G.R. Marshall, J.W. Ponder, Nature Struct. Biol. 6 (1999) 50.

[3] H. Kokubo, Y. Okamoto, Chem. Phys. Lett. 383 (2004) 397.

[4] A. Krogh, B. Larsson, G.v. Heijne, E.L.L. Sonnhammer, J. Mol. Biol. 305 (2001) 567.

[5] D.T. Jones, W.R. Taylor, J.M. Thornton, Biochemistry 33 (1994) 3038-3049.

[6] T. Hirokawa, S. Boon-Chieng, S. Mitaku, Bioinformatics 14 (1998) 378.

[7] G.E. Tusnady, I. Simon, J. Mol. Biol. 283 (1998) 489.

[8] W.E. Reiher, III, Theoretical Studies of Hydrogen Bonding, Ph.D. Thesis, Department of Chemistry, Harvard University, Cambridge, MA,USA, 1985

[9] E. Neria, S. Fischer, M. Karplus, J. Chem. Phys. 105 (1996) 1902.

[10] B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan, M. Karplus, J. Comput. Chem. 4 (1983) 187.

[11] K.R. MacKenzie, J.H. Prestegard, D.M. Engelman, Science 276 (1997) 131.

[12] S.O. Smith, D. Song, S. Shekar, M. Groesbeek, M. Zilioz, S. Aimoto, Biochemistry 40 (2001) 6553.

Table 3.1: Acceptance ratios of replica exchange corresponding to pairs of neighboring temperatures with the dielectric constant $\epsilon = 1.0$ and $\epsilon = 4.0$

Pairs of temperatures	Acceptance ratio ($\epsilon = 1.0$)	Acceptance ratio ($\epsilon = 4.0$)
200 ⟷ 239 K	0.41	0.37
239 ⟷ 286 K	0.40	0.37
286 ⟷ 342 K	0.39	0.37
342 ⟷ 404 K	0.40	0.41
404 ⟷ 489 K	0.32	0.34
489 ⟷ 585 K	0.34	0.36
585 ⟷ 700 K	0.33	0.36
700 ⟷ 853 K	0.28	0.30
853 ⟷ 1041 K	0.29	0.31
1041 ⟷ 1270 K	0.36	0.36
1270 ⟷ 1548 K	0.42	0.42
1548 ⟷ 1888 K	0.46	0.47

Table 3.2: Various properties of the native structure and the global-minimum-energy structure by the REM simulation with the dielectric constant $\epsilon = 1.0$ and $\epsilon = 4.0$. The following abbreviations were used: the total potential energy $E_{\rm tot}$, van der Waals energy $E_{\rm v}$, electrostatic energy $E_{\rm c}$, dihedral energy $E_{\rm t}$, constraint energy $E_{\rm cons}$, RMS deviation RMSD, radius of gyration RGYR, interhelical crossing angle IHCA, and solvent accessible surface area SA. The energy is in kcal/mol, distance is in Å, angle is in degrees, and area is in Å^2.

	solution NMR	global-minimum	global-minimum
	structure	structure ($\epsilon = 1.0$)	structure ($\epsilon = 4.0$)
$E_{\rm tot}$	–	-1497.8	-509.6
$E_{\rm v}$	–	-219.6	-225.0
$E_{\rm c}$	–	-1322.1	-327.0
$E_{\rm t}$	–	10.5	8.5
$E_{\rm cons}$	–	0.12	0.74
RMSD	–	0.64	4.48
RGYR	10.00	10.12	10.83
IHCA	45.8	42.7	14.0
SA	3152.3	3133.5	3087.2

Table 3.3: The interhelical distances (in Å) of the solid-state NMR [12], the solution NMR [11], and the global-minimum-energy structure obtained from the REM simulation with the dielectric constant $\epsilon = 1.0$ and $\epsilon = 4.0$.

	solid-state NMR	solution NMR	global-minimum	global-minimum
	structure	structure	structure	structure
			($\epsilon = 1.0$)	($\epsilon = 4.0$)
Gly79 C Gly79 CA	4.1	4.7	4.5	8.9
Gly79 CA Ile76 C	4.8	4.8	5.7	12.1
Gly83 C Gly83 CA	4.3	5.1	4.7	9.8
Gly83 CA Val80 C	4.2	4.3	4.3	12.4
Gly79 C Val80 C	4.0	2.9	4.4	12.4
Gly83 C Val84 C	4.0	3.7	3.8	13.1

Table 3.4: Various properties averaged over structures correspond to each peak in histograms in Fig.3.9(a) and Fig.3.10(a). The abbreviations are the same as in Table 3.2. The energy is in kcal/mol, distance is in Å, angle is in degrees, and area is in Å^2.

$\epsilon = 1.0$		
	peak1	peak2
E_{tot}	−1489.1	−1488.2
E_{v}	−214.2	−215.6
E_{c}	−1322.0	−1320.8
E_{t}	13.2	13.0
E_{cons}	0.66	2.02
RMSD	0.73	2.52
RGYR	10.13	10.47
IHCA	50.7	52.1
SA	3137.9	3214.9

$\epsilon = 4.0$					
	peak1	peak2	peak3	peak4	peak5
E_{tot}	−498.2	−499.0	−498.6	−500.8	−499.1
E_{v}	−216.2	−217.8	−217.6	−219.9	−219.8
E_{c}	−328.0	−328.0	−327.4	−326.9	−326.0
E_{t}	12.1	11.7	12.6	11.9	12.8
E_{cons}	0.68	1.90	0.53	0.94	0.78
RMSD	0.70	2.47	3.49	4.41	5.74
RGYR	10.13	10.44	10.28	10.84	10.60
IHCA	50.3	51.2	38.2	16.1	21.0
SA	3130.6	3202.0	3077.7	3097.8	3049.9

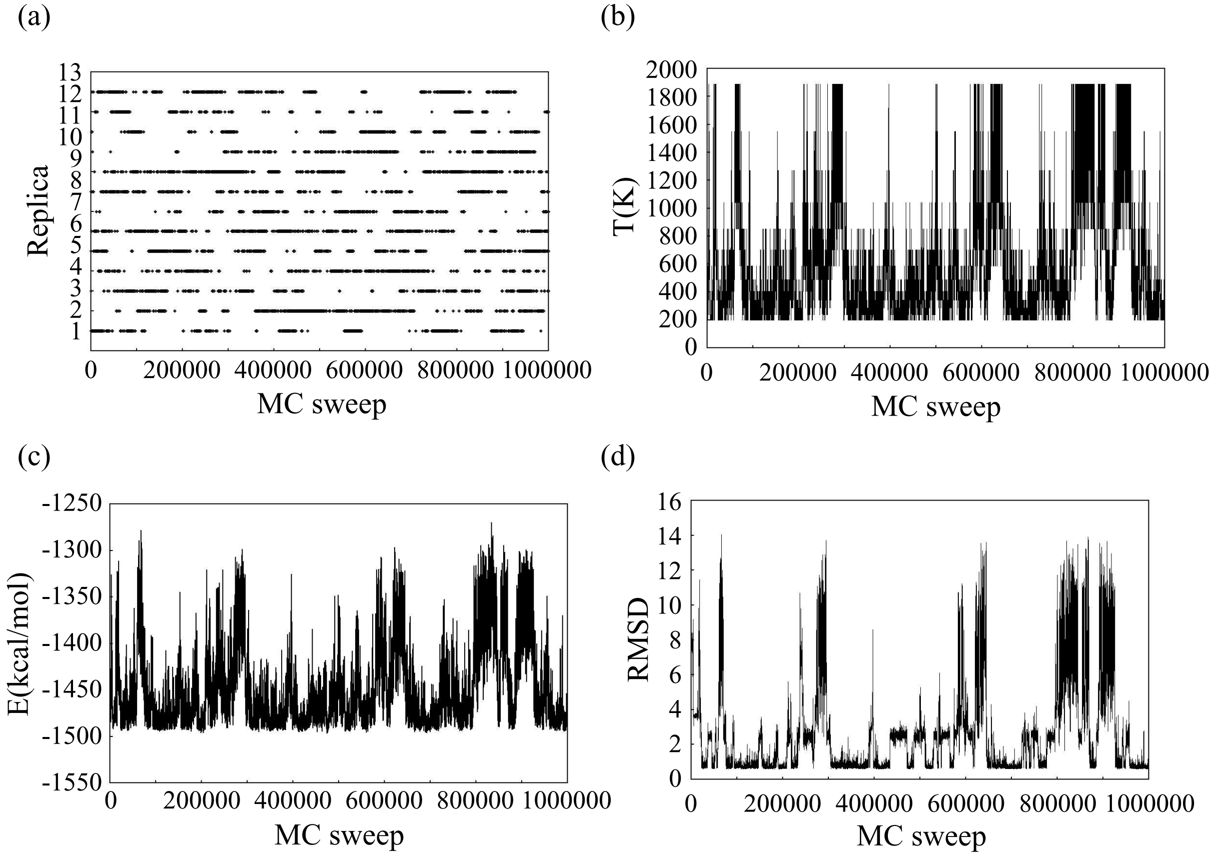

Figure 3.1: Time series of replica exchange at $T = 200$ K (a), temperature exchange for one of the replicas (Replica 6) (b), the total potential energy for Replica 6 (c), and the RMS deviation (in Å) of backbone atoms from the NMR structure for Replica 6 (d) with the dielectric constant $\epsilon = 1.0$.

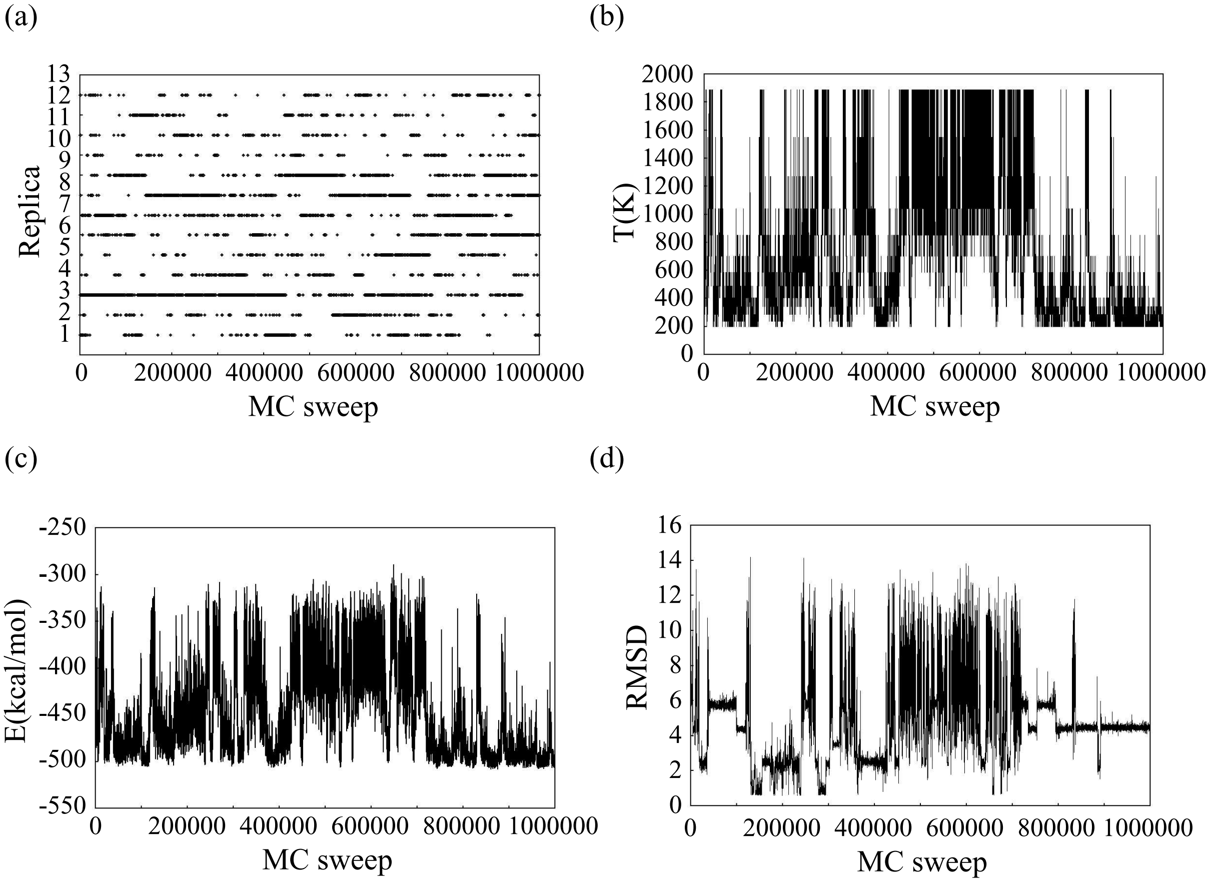

Figure 3.2: Time series of replica exchange at $T = 200$ K (a), temperature exchange for one of the replicas (Replica 6) (b), the total potential energy for Replica 6 (c), and the RMS deviation (in Å) of backbone atoms from the NMR structure for Replica 6 (d) with the dielectric constant $\epsilon = 4.0$.

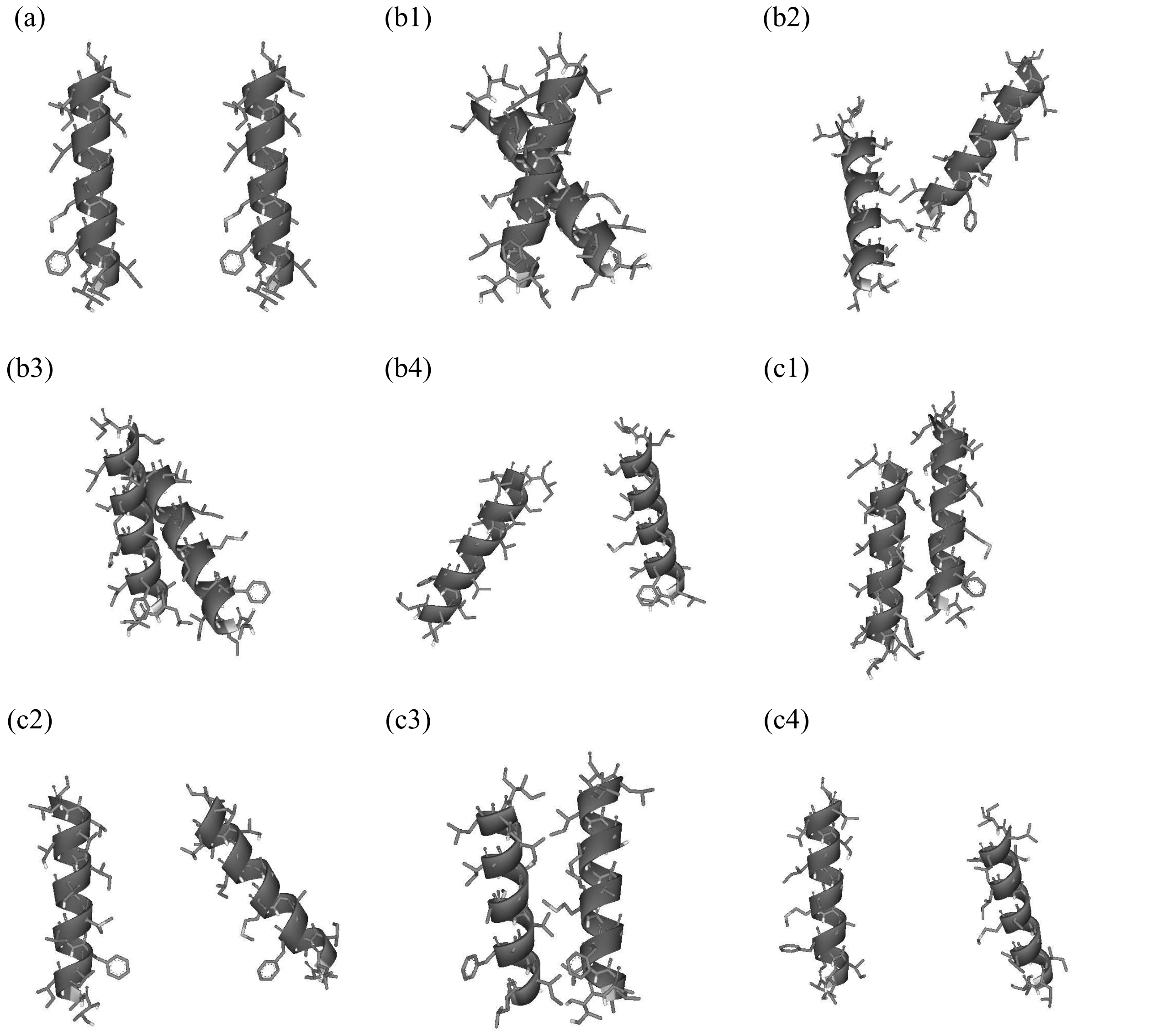

Figure 3.3: Typical snapshots from the REM simulation. The initial configuration (a), the configurations with the dielectric constant $\epsilon = 1.0$ (b1)-(b4) and with the dielectric constant $\epsilon = 4.0$ (c1)-(c4).

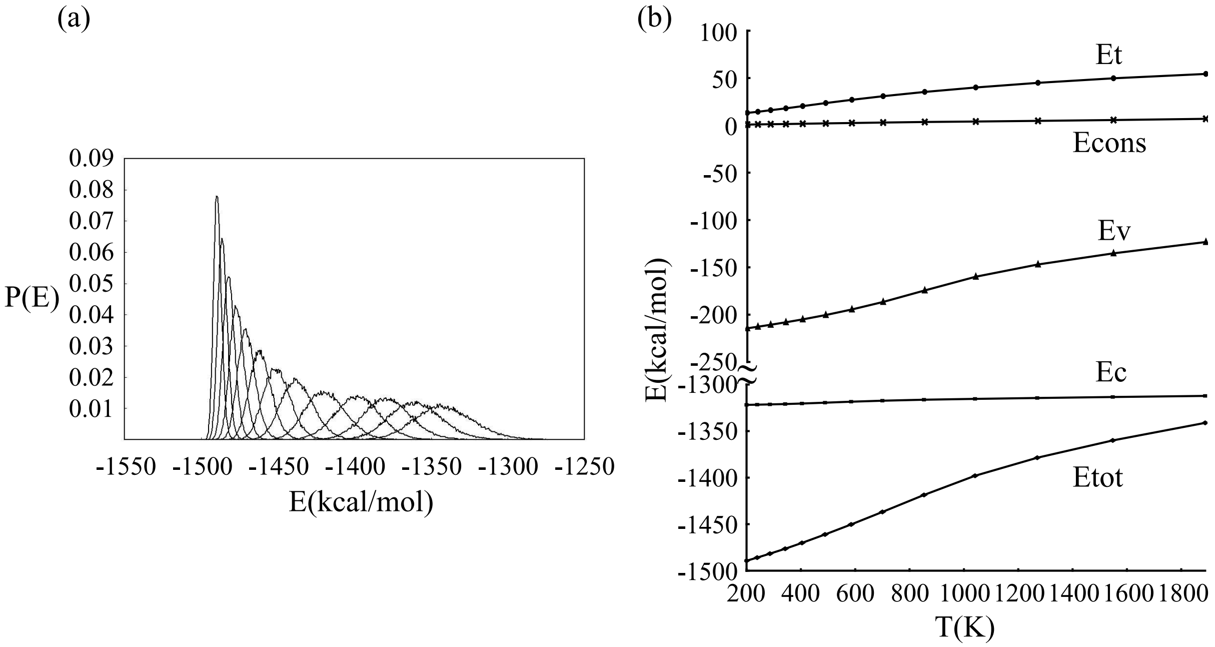

Figure 3.4: (a) The canonical probability distributions of the total potential energy obtained from the replica-exchange MC simulation at the thirteen temperatures with the dielectric constant $\epsilon = 1.0$. The distributions correspond to the following temperatures (from left to right): 200, 239, 286, 342, 404, 489, 585, 700, 853, 1041, 1270, 1548, and 1888 K. (b) The averages of the total potential energy $E_{\rm tot}$ and its component terms: electrostatic energy $E_{\rm c}$, van der Waals energy $E_{\rm v}$, dihedral energy $E_{\rm t}$, and constraint energy $E_{\rm cons}$ as functions of temperature T with the dielectric constant $\epsilon = 1.0$. The values were calculated by the multiple-histogram reweighting techniques.

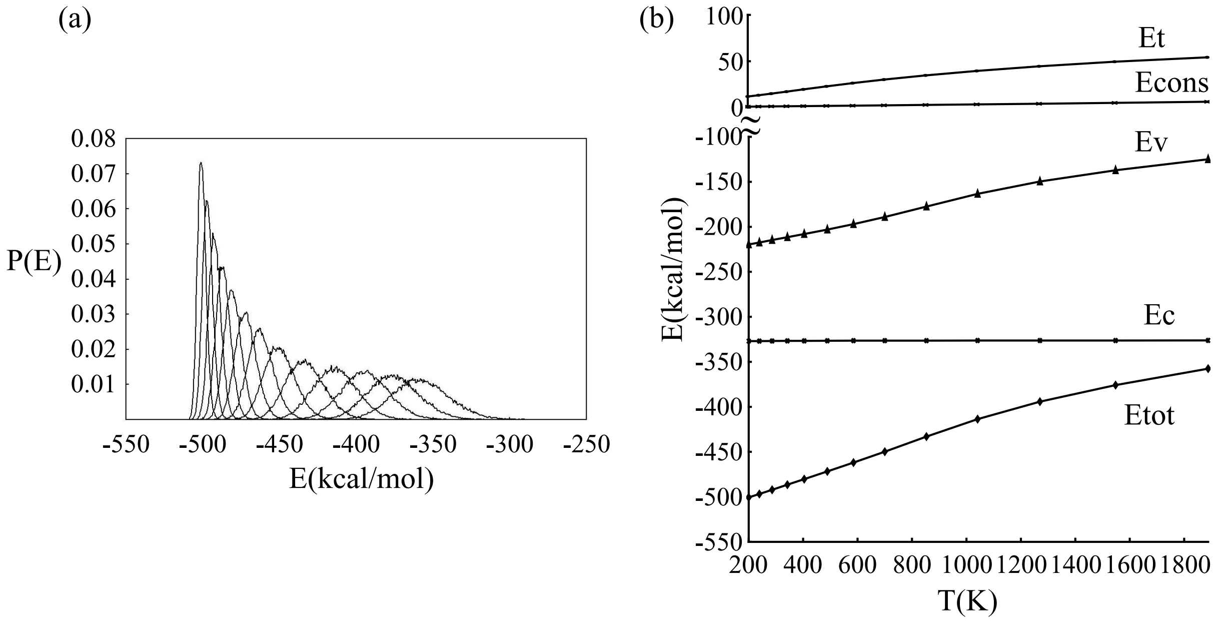

Figure 3.5: (a) The canonical probability distributions of the total potential energy obtained from the replica-exchange MC simulation at the thirteen temperatures with the dielectric constant $\epsilon = 4.0$. The distributions correspond to the following temperatures (from left to right): 200, 239, 286, 342, 404, 489, 585, 700, 853, 1041, 1270, 1548, and 1888 K. (b) The averages of the total potential energy $E_{\rm tot}$ and its component terms: electrostatic energy $E_{\rm c}$, van der Waals energy $E_{\rm v}$, and dihedral energy $E_{\rm t}$ as functions of temperature T with the dielectric constant $\epsilon = 4.0$. The values were calculated by the multiple-histogram reweighting techniques.

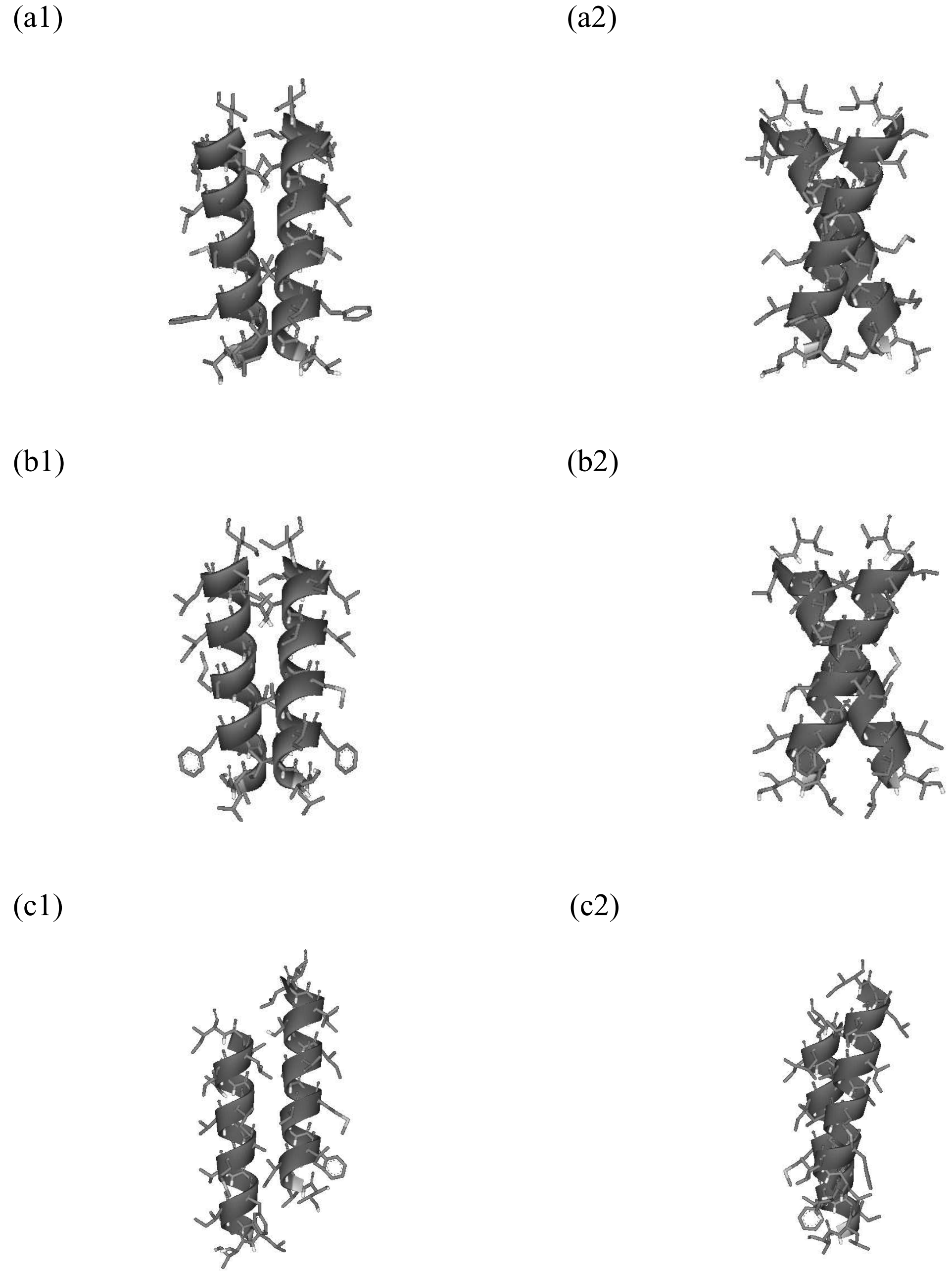

Figure 3.6: The NMR configuration (PDB code 1AFO, MODEL 16) (a1)(a2), the global-minimum-energy configuration that was obtained by the REM simulation with the dielectric constant $\epsilon = 1.0$ (b1)(b2), and the global-minimum-energy configurations that was obtained by the REM simulation with the dielectric constant $\epsilon = 4.0$ (c1)(c2). The pair (a1) and (a2) correspond to the same structure viewed from different angles. Likewise, the pair (b1) and (b2) and the pair (c1) and (c2) correspond to the same structures viewed from different angles. The RMS deviation from the native configuration (a) is 0.64 Å (b) and 4.48 Å (c) with respect to all backbone atoms, and it is 1.31 Å (b) and 5.55 Å (c) with respect to all atoms.

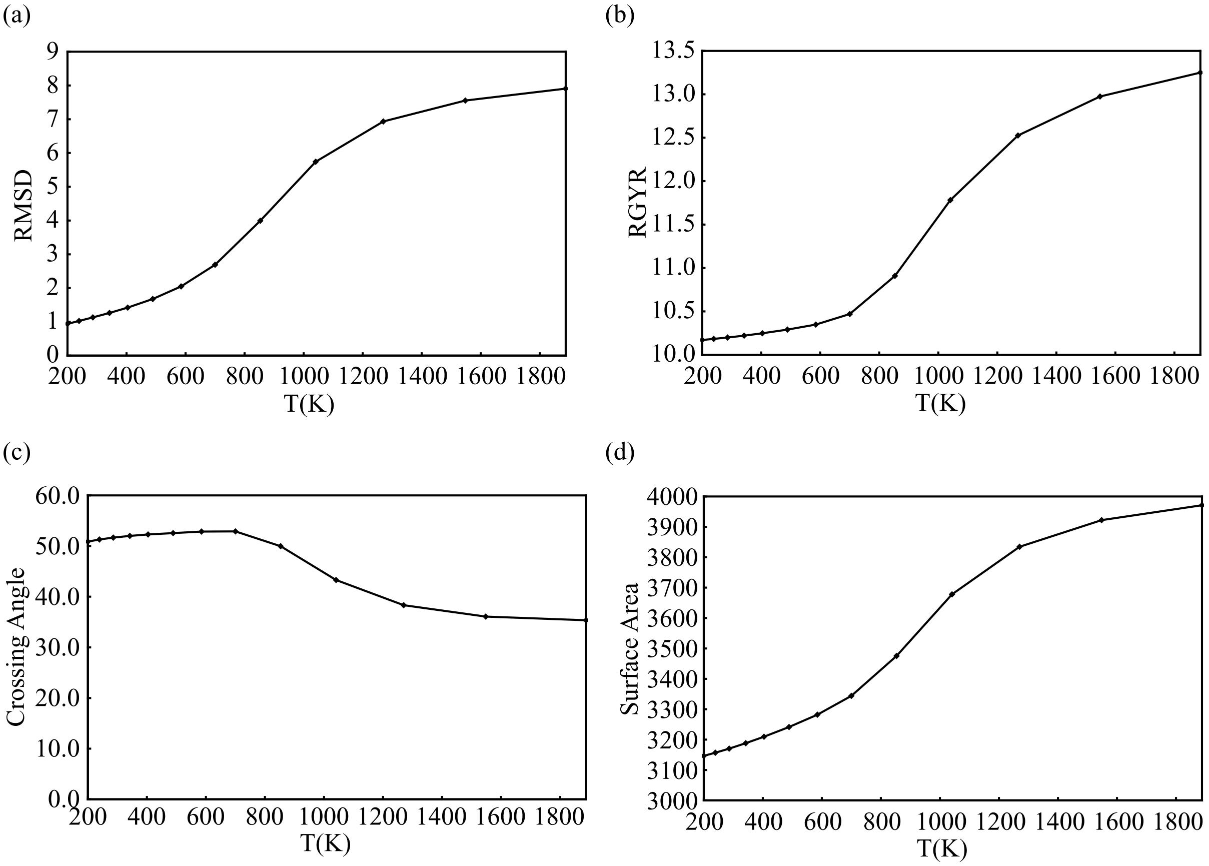

Figure 3.7: The averages of the RMS deviation (a),the radius of gyration (RGYR) (b), the interhelical crossing angle (c), and the solvent accessible surface area (d) as functions of temperature T with the dielectric constant $\epsilon = 1.0$. The values were calculated by the multiple-histogram reweighting techniques. The distance is in Å, angle is in degrees, and area is in Å^2.

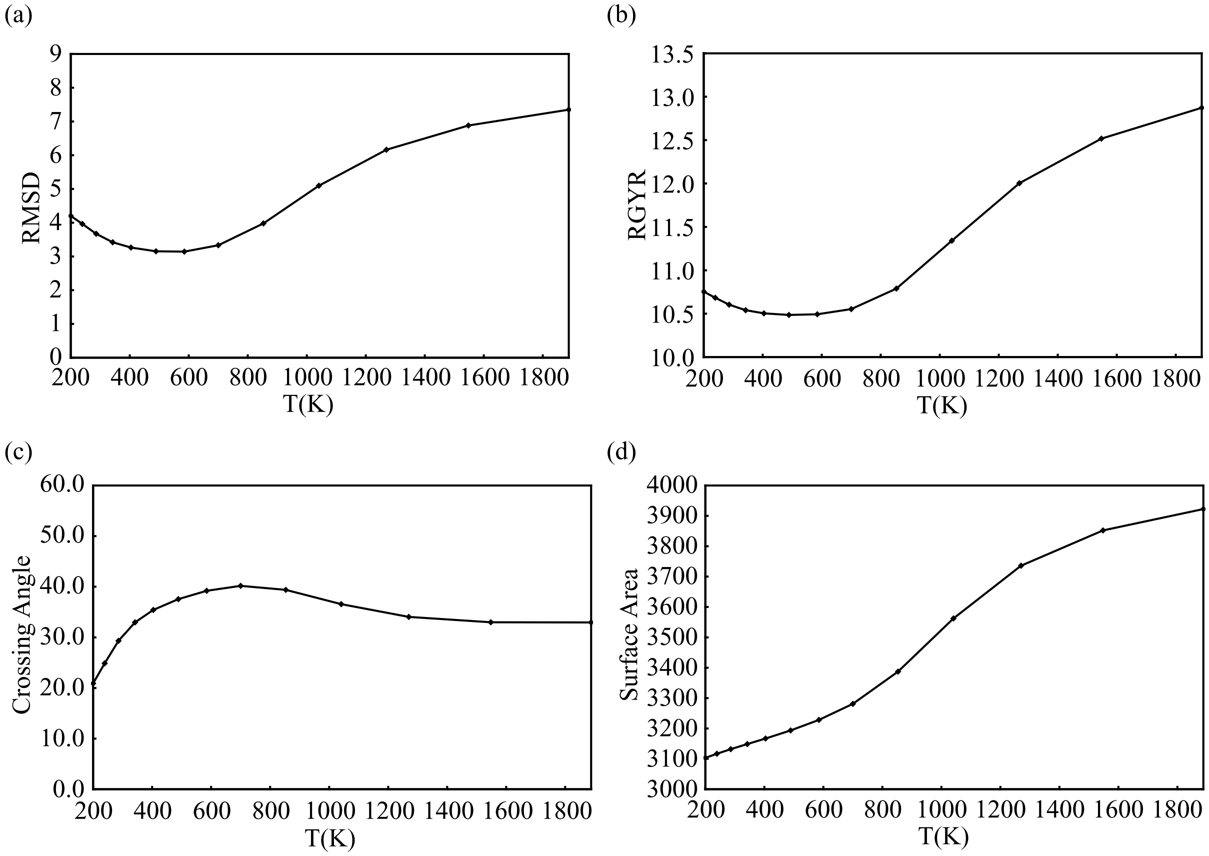

Figure 3.8: The averages of the RMS deviation (a),the radius of gyration (RGYR) (b), the interhelical crossing angle (c), and the solvent accessible surface area (d) as functions of temperature T with the dielectric constant $\epsilon = 4.0$. The values were calculated by the multiple-histogram reweighting techniques. The distance is in Å, angle is in degrees, and area is in Å^2.

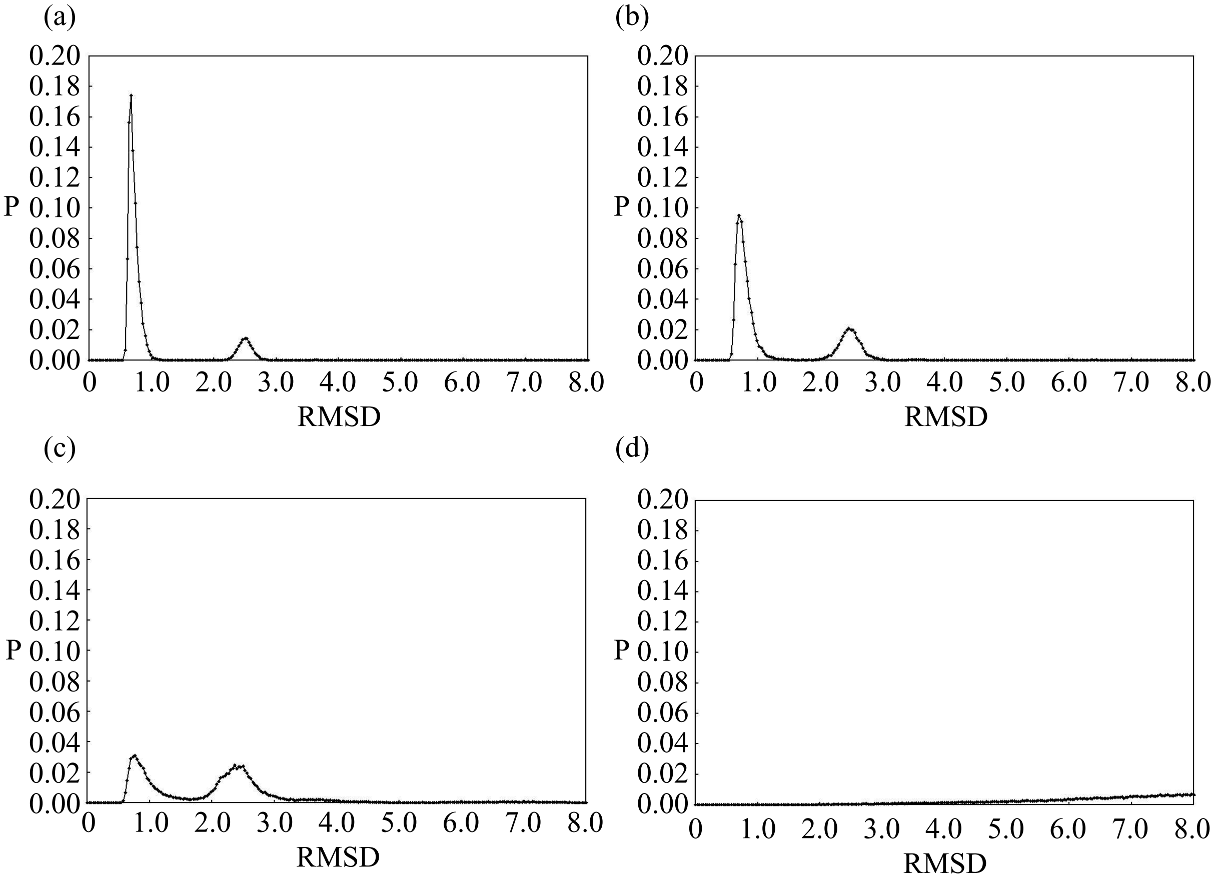

Figure 3.9: The probability distributions of the RMSD obtained from the replica-exchange MC simulation with the dielectric constant $\epsilon = 1.0$ at the chosen four temperatures. The distributions correspond to the following temperatures: 200 K (a), 342 K (b), 585 (c), and 1888 K (d).

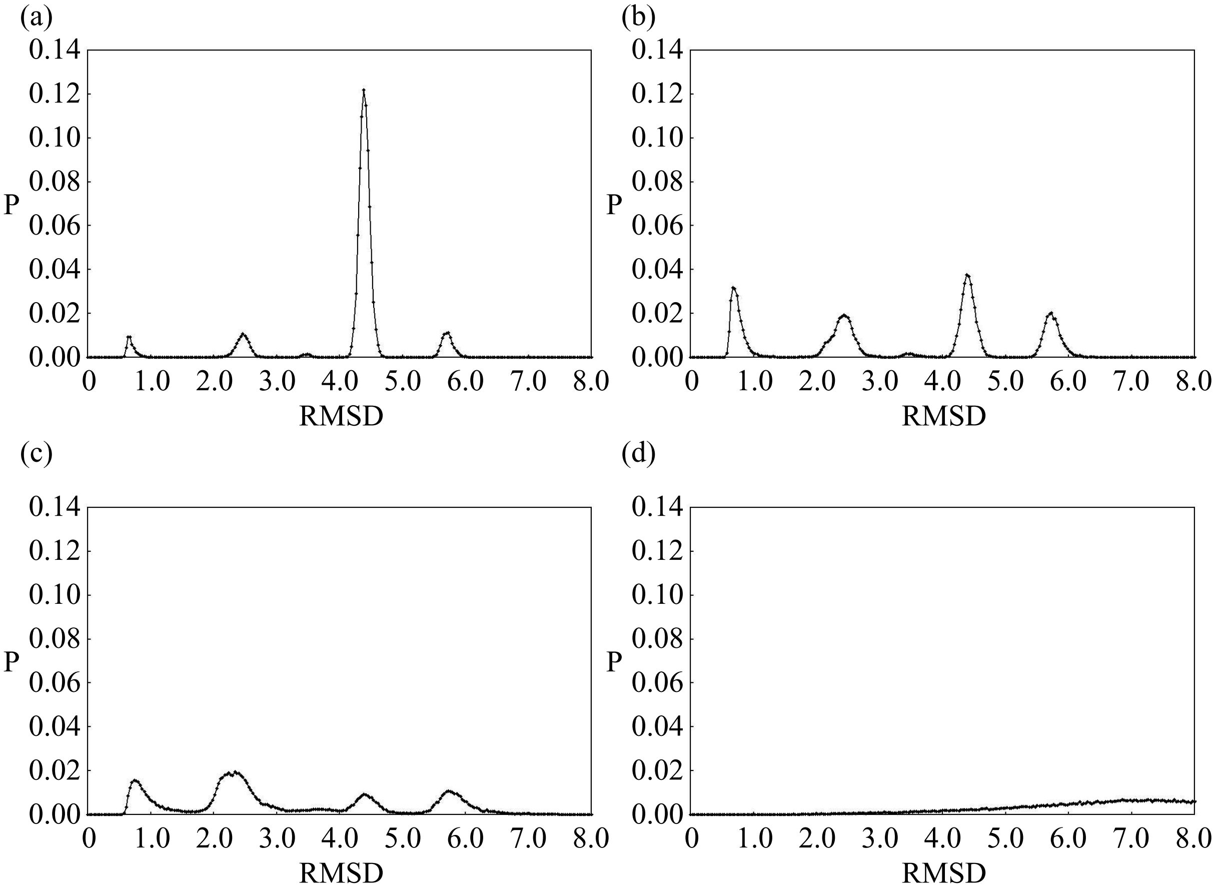

Figure 3.10: The probability distributions of the RMSD obtained from the replica-exchange MC simulation with the dielectric constant $\epsilon = 4.0$ at the chosen four temperatures. The distributions correspond to the following temperatures: 200 K (a), 342 K (b), 585 (c), and 1888 K (d).

Chapter 4

Classification of Low-Energy Configurations of Glycophorin A Transmembrane Dimer

H. Kokubo and Y. Okamoto, “Classification and prediction of low-energy membrane protein helix configurations by replica-exchange Monte Carlo method,” *Journal of the Physical Society of Japan* **73**, 2571-2585 (2004).

4.1 Introduction

The principal component analysis (PCA) [1]-[5] is often used to investigate protein dynamics (for a review, see Ref. [6]). PCA can be used to represent a significant protein motion by a low-dimensional subspace. In Refs. [6]-[8] the molecular dynamics trajectory of the protein BPTI was analyzed using PCA to investigate the harmonic and unharmonic aspects of the dynamics of BPTI. Free energy landscape of protein folding was recently studied by PCA, which was applied to generalized-ensemble simulation results [9, 10, 11]. In this chapter we employ PCA for the structure prediction of membrane proteins by molecular simulations and classify the structures sampled by the simulations into distinctly-defined clusters.

In the previous chapter we analyzed the global-minimum-energy structure obtained from the replica-exchange simulations [12], but the analysis of the only one structure is not enough to understand the statistical properties of the native structure because the native state is fluctuating. Non-native structures also contribute to a better understanding of the characteristics of the native structure. Therefore, in this chapter we effectively classify the low-energy structures obtained from the REM simulations into clusters by PCA and identify clusters as the global-minimum and local-minimum free energy states. We characterize each cluster by analyzing conformational properties averaged over structures in the cluster.

This chapter is organized as follows. Sec. 4.2 briefly summarizes the computational details. In Sec. 4.3 the results of the application to the structure classification and prediction of the dimeric transmembrane domain of glycophorin A are given.

4.2 Computational Details

As explained in detail in the previous chapter (see Sec. 3.2), we performed two REM MC simulations of 60,000,000 MC steps, starting from the parallel configuration at a distance 20 Å with each other: one with the dielectric constant $\epsilon = 1.0$ and the other with $\epsilon = 4.0$. We used the following 13 temperatures: 200, 239, 286, 342, 404, 489, 585, 700, 853, 1041, 1270, 1548, and 1888 K, which are distributed almost exponentially. The highest temperature was chosen sufficiently high so that no trapping in local-minimum-energy states occurs. This temperature distribution was chosen so that all the acceptance ratios are almost uniform and sufficiently large (> 10 %) for computational efficiency. One MC sweep corresponds to 58 MC steps that consist of the rigid translation and rotation of two helices (four degrees of freedom) and the dihedral-angle rotations ($27 \times 2 = 54$ degrees of freedom). Hence, 60,000,000 MC steps corresponds to about 1,000,000 MC sweeps. Replica exchange was attempted once every 50 MC steps. The structures were stored every 1,000 MC steps for later analyses (the total number of structures is thus $60,000 \times 13$). For the principal component analysis we used 6,000 such structures at each chosen temperature (one tenth of the stored data for each temperature).

4.3 Results and Discussion

4.3.1 MC Time Series of the Replica-Exchange Simulations

We now present the results of further analysis of the REM simulations in Chapter 3. In Fig. 4.1 we show time series of various quantities that we obtained with the dielectric constant $\epsilon = 1.0$ and $\epsilon = 4.0$. Fig. 4.1(a1) and Fig. 4.1(b1) are time series of temperature exchange of one of the replicas with $\epsilon = 1.0$ and $\epsilon = 4.0$, respectively. We see that the temperature goes back and forth between the lowest value and the highest one many times and that enough sampling has been achieved. The acceptance ratios of replica exchange between all pairs of neighboring temperatures were almost uniform and large enough (staying in the range between 28 % and 47 %). Fig. 4.1(a2) and Fig. 4.1(b2) are time series of the total potential energy corresponding to Fig. 4.1(a1) and Fig. 4.1(b1), respectively. When the temperature is high, the total potential energy is high, and when the former is low, the latter is also low. There is a strong correlation between Figs. 4.1(a1) and 4.1(a2), as there should. Likewise, there is a strong correlation between Figs. 4.1(b1) and 4.1(b2). These figures show that the replica-exchange simulations have been properly performed. We next examine how widely the conformational space was sampled. Figs. 4.1(a3) and 4.1(b3) are time series of the root-mean-square deviation (RMSD) from the structure of the solution NMR experiments [13]. We see that the RMSD goes back and forth between small values and large ones many times and that wide conformational space has indeed been searched in the simulations. Comparing Fig. 4.1(a3) with Figs. 4.1(a1) and 4.1(a2), we see that the structures close to the experimental one (RMSD $\sim$ 0.7 Å) were often sampled when both the temperature and the potential energy were low. However, the RMSD values around 2.5 Å were also sampled at low temperatures. This suggests that there exists a local minimum free energy state around 2.5 Å in the case of $\epsilon = 1.0$. Similarly, comparing Fig. 4.1(b3) with Figs. 4.1(b1) and 4.1(b2), we see that the RMSD takes on several values (RMSD $\sim$ 0.7, 2.5, 3.5, 4.4, and 5.7 Å) when both the temperature and the potential energy are low. This suggests that there exist several local-minimum free energy states at low temperatures in the case of $\epsilon = 4.0$. We discuss this point more in detail below.

4.3.2 Classifications and Structure Predictions by the Principal Component Analysis

In Fig. 4.2 we show the percentage (the left ordinate) and amplitude (the right ordinate) of the first ten principal components at the chosen four temperatures in the case of $\epsilon = 1.0$. The principal components were calculated from Eq. (2.14). Fig. 4.3 is the corresponding figure in the case of $\epsilon = 4.0$. We see from the percentage values in Figs. 4.2 and 4.3 that more principal component axes are needed to represent the fluctuation of the system as the temperature becomes higher, as expected. We observe from the amplitude values in Figs. 4.2 and 4.3 that the amplitude becomes larger as the temperature becomes higher. This is reasonable because as the temperature becomes higher, the fluctuations of the system become larger and the simulations sample wider conformational space. In the case of $\epsilon = 4.0$, the amplitude at 200 K is larger than that in the case of $\epsilon = 1.0$. This implies that local-minimum free energy states at the lowest temperature in the case of $\epsilon = 4.0$ are distributed in wider conformational space than in the case of $\epsilon = 1.0$. This point is further elaborated around Fig. 4.5 below. In Fig. 4.2(a) we see that more than 80 % of the total amplitude is expressed by the first two principal components in the case of $\epsilon = 1.0$. Similarly, we see from Fig. 4.3(a) that more than 90 % amplitude is expressed by the first two principal components in the case of $\epsilon = 4.0$. Therefore, we can classify and analyze properly the sampled structures at the lowest temperature by only the first two principal components. The fact that most of the amplitude of fluctuations in this protein system is represented only by a small number of principal components supports that protein folding dynamics can be expressed as the diffusion over a low-dimensional free energy surface as is elucidated in the energy landscape theory [14]. At high temperatures it can be said from Figs. 4.2(d) and 4.3(d) that more principal component axes are needed to analyze the sampled structures properly. The sampled structures are sometimes analyzed by other reaction coordinates (for example, native contact, RMSD, and radius of gyration) and the free energy surface is drawn by projecting on these axes. However, we cannot tell how many reaction coordinates we need to be able to identify important local-minimum free energy states in the free energy landscape. The principal component analysis naturally provides us with the information as to how many reaction coordinates (principal components) we need for such investigations.

In Fig. 4.4 the structures obtained from the replica-exchange simulation with $\epsilon = 1.0$ are projected on the first and second principal component axes at chosen four temperatures. There are three distinct clusters at the lowest temperature in Fig. 4.4(a). As the temperature becomes higher, these clusters become less distinct and first and second principal components become larger (note that the scales of both axes are expanded). This implies that as the temperature becomes higher, wider conformational space is sampled without getting trapped in local-minimum free energy states. If we perform constant temperature simulations at the lowest temperature, the simulations will get trapped in one of the clusters in Fig. 4.4(a), depending on the initial configurations of the simulations. However, each replica of the replica-exchange simulations will not get trapped in one of the local-minimum free energy states, because the temperature of each replica goes up and down by temperature exchange. The three clusters in Fig. 4.4(a) lie in the ranges $(-60 \sim -30, -25 \sim -10)$, $(-20 \sim 20, -10 \sim 10)$, and $(30 \sim 60, -15 \sim 0)$, which we refer to as Cluster 1, Cluster 2, and Cluster 3, respectively. Note that the basis for classifying the structures in this way exist in that more than 80 % of the total amplitude is represented by the first two principal components as was shown in Fig. 4.2(a).

In Fig. 4.5 the structures obtained from the replica-exchange simulation with $\epsilon = 4.0$ are projected on the first and second principal component axes at the chosen four temperatures. The similar features as in the case of $\epsilon = 1.0$ are observed. In Fig. 4.5(a) the sampled structures exist in wider region along the second principal axis than in Fig. 4.4(a). Therefore, the amplitude (right-hand ordinate) in Fig. 4.3(a) is larger than that in Fig. 4.2(a). We classified the structures at the lowest temperature in Fig. 4.5(a) in nine clusters. Clusters 1, 2, 3, 4, 5, 6, 7, 8, and 9 correspond to the structures that lie in the ranges $(-50 \sim -35, -10 \sim 10)$, $(-10 \sim 5, 50 \sim 60)$, $(18 \sim 28, 55 \sim 70)$, $(40 \sim 65, 55 \sim 70)$,$(70 \sim 85, -20 \sim -5)$, $(45 \sim 65, -55 \sim -45)$, $(-35 \sim -15, -55 \sim -40)$, $(0 \sim 8, 5 \sim 20)$, and $(27 \sim 35, 22 \sim 32)$, respectively. Note again that the basis for classifying the structures in this way exists in that more than 90 % of the total amplitude is represented by the first two principal components as was shown in Fig. 4.3(a).

In Table 4.1 various average properties of the three clusters classified in the case of

$\epsilon = 1.0$ are compared. The average values at 200 K are those of all the structures sampled at 200 K and the average values of each cluster are those of the structures included in each cluster classified at 200 K. Similarly the standard deviations at 200 K are those of all the structures sampled at 200 K and the standard deviations of each of the three clusters are those of the structures included in each cluster. We see that the standard deviations of the RMSD of the three clusters are much smaller than that of all the structures at 200 K. The standard deviations of the radius of gyration and the solvent accessible surface area are also smaller than that of all the structures sampled at 200 K. This implies that the structures sampled at 200 K are sharply divided into three clusters. It is found that Cluster 2 has a very close structure to the native one and that Clusters 1 and 3 have essentially the same structures, judging from the comparison of the RMSD, the radius of gyration, the interhelical crossing angle, and the solvent accessible surface area with those of the NMR structure. In fact, we confirm this in the next figure. From Table 4.1 we see that the structures of Cluster 2 are more stabilized by the electrostatic energy than those of Clusters 1 and 3 (the electrostatic energy of Cluster 2 is less by 1.3 and 1.5 kcal/mol than that of Clusters 1 and 3) and that the structures of Clusters 1 and 3 are more stabilized by the van der Waals energy than those of Cluster 2 (the van der Waals energy of Clusers 1 and 3 is less by 1.2 kcal/mol than that of Cluster 2). This implies that although the amino acids used in our simulations consist of only four hydrophilic amino acids (threonine) out of 36 amino acids in total, the electrostatic energy also contributes to the stability of the native structure. The simulations sample one of three clusters at low temperatures which correspond to global- and local-minimum free energy states. If we classify the structures by other indices, it is not guaranteed to be able to divide the structures into clusters properly in this way.

The surface area and the radius of gyration of Cluster 2 are less than those of Clusters 1 and 3, and the interhelical crossing angle of Cluster 2 is similar to that of Clusters 1 and 3. The standard deviation of interhelical crossing angle of each cluster is almost the same as that of all the structures at 200 K, but this is because the average values are close to each other. If the average values are much different from each other among the clusters, the standard deviations for clusters will become much less than that of all the structures at 200 K. In fact, we will see that this is true in Table 4.2 below. The standard deviations of the total energy and its component terms of each cluster are nearly equal to those of all the structures.

Because this membrane protein is a dimer and the two helices consist of the same amino-acid sequences, one configuration can have two different numbering of atoms. Therefore, the principal component analysis treats the same configurations as different clusters if it does not have C_2-symmetry in structure. In Table 4.1 Cluster 2, which is close to the native structure, is C_2-symmetric and therefore becomes only one cluster, while Clusters 1 and 3 have no such symmetry and actually have the same structure.

In Fig. 4.6 the typical structure of each cluster in Table 4.1 (in the case of $\epsilon = 1.0$) and the solution NMR structure (PDB code: 1AFO [13]) are shown. The standard deviation of RMSD of each cluster is very small and the structures included in each cluster have essentially the same backbone structures. We confirm that the structures of Clusters 1 and 3 are almost the same and the structures of Cluster 2 are indeed very close to the experimental one as expected in Table 4.1. The structures of Clusters 1 and 3 are a little off from the membrane boundary, and their constraint energy in Table 4.1 is higher than that of Cluster 2.

We next examine the case of $\epsilon = 4.0$. Table 4.2 lists various properties of the nine clusters classified by the principal component analysis. Features similar to the case of $\epsilon = 1.0$ are found. We see that the standard deviation of the RMSD of each cluster is much smaller than that of total structures sampled at 200 K. The standard deviations of the radius of gyration, interhelical crossing angle, and solvent accessible surface area of each cluster are also much smaller than those of total structures sampled at 200 K. This implies that the structures sampled at 200 K are effectively divided into clusters similarly as in the case of $\epsilon = 1.0$. By comparing the RMSD, radius of gyration, crossing angle, and surface area, it can be said that Cluster 3 is close to the native structure and Clusters 1 and 5, Clusters 2 and 4, Clusters 6 and 7, and Clusters 8 and 9 are the same structures, respectively. This point is also confirmed in Fig. 4.7. From Tables 4.1 and 4.2 we see that Clusters 1 and 3 in the case of $\epsilon = 1.0$ are the same as Clusters 2 and 4 in the case of $\epsilon = 4.0$ and that Cluster 2 in the case of $\epsilon = 1.0$ and Cluster 3 in the case of $\epsilon = 4.0$ correspond to the native structure. Therefore, the structures sampled at low temperatures in the case of $\epsilon = 1.0$ are subsets of those in the case of $\epsilon = 4.0$. We see that Cluster 3 (in the case of $\epsilon = 4.0$), which is close to the native structure, is more stabilized by the electrostatic energy than other clusters. However, the number of samples for Cluster 3 is smaller than that for the corresponding cluster (Cluster 2) in the case of $\epsilon = 1.0$, because the stability by the electrostatic energy is more weakened in the case of $\epsilon = 4.0$ than in the case of $\epsilon = 1.0$. Because membrane proteins are often tightly packed it is important to estimate the contribution of the electrostatic energy, van der Waals energy, and torsion energy to the stability of the native structure accurately. Therefore, we think that the simulations with atomistic details are essential to predict the native structure and understand its stability from the physical standpoint. The most sampled clusters are Clusters 1 and 5. Their van der Waals energy is the lowest among all the clusters. Cluster 3 has the smallest radius of gyration and it is the closest to the native one. Clusters 1, 5, 6, 7, 8, and 9 have smaller surface area than Cluster 3. Note that the cluster which has a smaller surface area does not always have smaller van der Waals energy. Clusters 2 and 4 have larger surface area than Cluster 3 but smaller van der Waals energy.

In Fig. 4.7 the typical cluster structures in Table 4.2 (in the case of $\epsilon = 4.0$) are shown. It is confirmed that Cluster 3 (Fig. 4.7(c)) is very close to the native structure (Fig. 4.6(a)) and Clusters 1 and 5, Clusters 2 and 4, Clusters 6 and 7, and Clusters 8 and 9 are the same structures, respectively, as expected from Table 4.2. We also confirm that Clusters 2 and 4 in the case of $\epsilon = 4.0$ (Figs. 4.7(b) and 4.7(d)) are the same clusters with Clusters 1 and 3 in the case of $\epsilon = 1.0$ (Figs. 4.6(b) and 4.6(d)). Hence, the structures sampled at 200 K in the case of $\epsilon = 1.0$ are indeed subsets of those in the case of $\epsilon = 4.0$.

4.3.3 Probability Distributions of Sampled Structures

We now examine the probability distributions of several properties in the case of $\epsilon = 1.0$ in order to confirm the effectiveness of our classifications. They are shown in Fig. 4.8. We see that the RMSD distribution in Fig. 4.8(a) becomes broad at the highest temperature. and the distributions have two peaks at low temperatures. We can understand that the simple average values will make no sense when the distribution has multiple peaks with narrow widths in this way. One of the peaks lies around 0.7 Å and is due to Cluster 2 in Table 4.1 (see also Fig. 4.6(c)), which is close to the native structure. Another peak is found around 2.5 Å and due to Clusters 1 and 3 in Table 4.1 (see also Figs. 4.6(b) and 4.6(d)). This can be understood from the fact that the standard deviations of RMSD in Table 4.1 is small. The distribution of the radius of gyration in Fig. 4.8(b) has similar features. The peak around 10.13 Å is due to Cluster 2 and that around 10.46 Å is due to Clusters 1 and 3 in Table 4.1. In Fig. 4.8(c) the probability distribution of interhelical crossing angle is shown. When the temperature is high, the simulations sample wide angles. It appears that there is only one peak at the low temperatures although in fact two types of different structures are sampled. Hence, the interhelical crossing angle is not a good measure for distinguishing structures in the present system. In Fig. 4.8(d) the distribution of surface area is shown. At the highest temperature various structures are sampled but the surface area has the maximum fixed value for the backbone structures when the two helices are apart. The peak around 4,150 Å^2 at T=1888 K is due to this maximum value. At low temperatures Clusters 1, 2, and 3 are sampled, but it is difficult to distinguish the peak around 3,138 Å^2 for Cluster 2 from that around 3,215 Å^2 for Clusters 1 and 3 (see Table 4.1). In summary, we have to choose physical quantities carefully in order to classify structures. As we have seen in Figs. 4.4 and 4.5, the principal component analysis yields a natural means of structure classifications. How many principal components are necessary for the classification can be determined by considering the proportion of eigenvalues to the total amplitude.

In Fig. 4.9 the corresponding probability distributions in the case of $\epsilon = 4.0$ are shown. Similar behaviors to the case of $\epsilon = 1.0$ are seen. In Fig. 4.9(a) the peaks of RMSD around 0.7 Å, 2.5 Å, 4.4 Å, 5.7 Å are due to Cluster 3, Clusters 2 and 4, Clusters 1 and 5, Clusters 6 and 7 in Table 4.2, respectively. The numbers of samples for Clusters 8 and 9 are relatively small and their peaks are not clearly seen in the Figure. Clusters 1 and 5 (RMSD $\sim$ 4.4 Å) are sampled most at 200 K but sampled less as the temperature becomes higher. Therefore, the "entropy" of Clusters 1 and 5 seem to be less than that of other clusters. In Fig. 4.9(b) the peaks of the radius of gyration around 10.13 Å, 10.45 Å, 10.60 Å, and 10.84 Å are due to Cluster 3, Clusters 2 and 4, Clusters 1 and 5, Clusters 6 and 7 in Table 4.2, respectively. The peak which is due to Clusters 8 and 9 are likewise not clearly seen. In Fig. 4.9(c) the probability distribution of the interhelical crossing angle is shown. It seems that there are only two peaks, but in fact they consist of nine clusters (five different types of structures). It is impossible to distinguish these clusters from this figure. In Fig. 4.9(d) the probability distribution of the surface area is shown. It is also impossible to distinguish clusters from this figure because the standard deviation of the surface area of each cluster is large compared with the difference of the surface area between clusters in Table 4.2. Therefore, it is not useful to draw the free energy landscape with these quantities as reaction coordinates in order to identify local-minimum free energy states.

4.3.4 Comparison of Interhelical Distances

We now compare the distance between helices with the experimental data. In Table 4.3 the interhelical distances of the solid-state NMR [15] and solution NMR [13] structures are compared with average values of Cluster 2 obtained from the REM simulation with $\epsilon = 1.0$ and Cluster 3 obtained from the REM simulation with $\epsilon = 4.0$. These clusters correspond to the native-like conformations (see Tables 4.1 and 4.2 and Figs. 4.6(c) and 4.7(c)). The solid-state NMR structure was obtained from the structure crystallized in lipid bilayer as in the native state. On the other hand, the solution NMR experiments were carried out in detergent micelles. We confirm that Cluster 2 of $\epsilon = 1.0$ and Cluster 3 of $\epsilon = 4.0$ have the interhelical distances close to those of the solid-state and solution NMR experiments, and this means that both clusters are close to the native structure as expected. Other clusters have totally different values from these experimental data (data not shown in the Table). The simulation results are more in accord with those of the solid-state NMR experiments that the solution NMR in the sense that all the distances agree in the former, while one distance (between Gly 79 and Val 80) is in disagreement in the latter.

4.3.5 Free Energy Analyses

In Fig. 4.10 the free energy surface with respect to the first and second principal component axes at 200 K is shown in the case of $\epsilon = 1.0$. We use the following equation to calculate the free energy as a function of the first and second principal components x_1 and x_2:

$$F(x_1, x_2) = -k_{\rm B}T \ln P(x_1, x_2), \tag{4.1}$$

where $k_{\rm B}$ is the Boltzmann constant, T is absolute temperature, and $P(x_1, x_2)$ is the probability to find the structure with the first and second principal component values x_1 and x_2. We see that Cluster 2 is the lowest free energy state. This figure shows that the cluster which is very close to the native structure has indeed the global-minimum free energy structure. The free energy of Clusters 1 and 3 is larger by about 1.1 kcal/mol than that of Cluster 2. We remark that the free energy in the Figure is reliable only in the vicinity of the bottoms of valleys; more accurate results in principle can be obtained in wider regions including transition states by the reweighting techniques.

In Fig. 4.11 we show the free energy surface in the case of $\epsilon = 4.0$. In this case Cluster 3 which is close to the native structure is not the global-minimum free energy state but a local-minimum free energy state. Clusters 1 and 5 correspond to the global-minimum free energy state instead. The free energy of Clusters 2, 3, and 4 is large about 1.0 kcal/mol than that of Clusters 1 and 5. The free energy of Clusters 6 and 7 is larger by about 1.2 kcal/mol than that of Clusters 1 and 5. The free energy of Clusters 8 and 9 is larger by about 1.7 kcal/mol than that of Clusters 1 and 5.

The free energy differences depicted in Figs. 4.10 and 4.11 are not estimated accurately because the clusters which have the same structures are treated separately in our principal component analysis. We have already seen that Clusters 1 and 3 in the case of $\epsilon = 1.0$, and Clusters 1 and 5, Clusters 2 and 4, Clusters 6 and 7, and Clusters 8 and 9 in the case of $\epsilon = 4.0$ have the same structures, respectively. Therefore, in Table 4.4 we treat the clusters which have the same structures as one group and calculate the free energy as follows:

$$F = -k_{\rm B}T \ln P, \tag{4.2}$$

where P is now the probability to find the structure in each of the groups of the same structures (for each we added all the histogram entries and divided them by the total number of samples). The differences of the free energy, internal energy, and entropy have the following relation:

$$\Delta F = \Delta U - T\Delta S, \tag{4.3}$$

where the internal energy U is defined to be the average total potential energy. The free energy and internal energy of the group which has the global-minimum free energy are set to zero as baseline in both cases of the dielectric constants in this Table. The difference of the internal energy, ΔU, between groups was calculated by averaging the total potential energy of the structures which belong to each group. The entropy difference, ΔS, was calculated from ΔF and ΔU by using the relation in Eq. (4.3). We see that the group which has the global-minimum free energy also has the global-minimum total potential energy and the least entropy values. Because the entropy of the local-minimum free energy clusters is larger than that of the global-minimum free energy cluster, the differences of the average total potential energy between clusters are larger than those of the free energy.

Bibliography

[1] M.M. Teeter, D.A. Case, J. Phys. Chem. 94 (1990) 8091.

[2] A. Kitao, F. Hirata, N. Go, Chem. Phys. 158 (1991) 447.

[3] A.E. Garcia, Phys. Rev. Lett. 68 (1992) 2696.

[4] R. Abagyan, P. Argos, J. Mol. Biol. 225 (1992) 519.

[5] A. Amadei, A.B.M. Linssen, H.J.C. Berendsen, Proteins 17 (1993) 412.

[6] A. Kitao, N. Go, Curr. Opin. Struc. Biol. 9 (1999) 164.

[7] A. Kitao, S. Hayward, N. Go, Proteins 33 (1998) 496.

[8] S. Hayward, A. Kitao, N. Go, Protein Sci. 6 (1994) 936.

[9] J. Higo, N. Ito, M. Kuroda , S. Ono, N. Nakajima and H. Nakamura, Protein Sci. 10 (2001) 1160.

[10] A. Garcia and K. Y. Sanbonmatsu, Proteins 42 (2001) 345.

[11] R. Zhou, B. J. Berne and R. Germain, Proc. Natl. Acad. Sci. USA 98 (2001) 14931.

[12] H. Kokubo, Y. Okamoto, J. Chem. Phys. 120 (2004) 10837.

[13] K.R. MacKenzie, J.H. Prestegard, D.M. Engelman, Science 276 (1997) 131.

[14] J.N. Onuchic, Z. LutheySchulten, P.G. Wolynes, Annu. Rev. Phys. Chem. 48 (1997) 545.

[15] S.O. Smith, D. Song, S. Shekar, M. Groesbeek, M. Zilioz, S. Aimoto, Biochemistry 40 (2001) 6553.

Table 4.1: Various average properties of the structures classified by the principal component analysis, which were obtained at the temperature of 200 K by the REM simulation with the dielectric constant $\epsilon = 1.0$. The values after ± are the standard deviations. The averages were calculated from Eq. (2.12). The following abbreviations were used: the total potential energy E_{tot}, van der Waals energy E_{vdw}, electrostatic energy E_{c}, dihedral energy E_{t}, constraint energy E_{constr}, root-mean-square deviation RMSD, radius of gyration RGYR, interhelical crossing angle IHCA, and solvent accessible surface area SA. The interhelical crossing angle is defined as the angle between the principal axes of the moment of inertia for each helix. The entries under "200 K" stand for the results of the average over all the structures obtained at $T = 200$ K, and those under "NMR" stand for the results of the solution NMR experiments (Model 16 of PDB code 1AFO) [13]. The energy is in kcal/mol, distance is in Å, angle is in degrees, and area is in Å^2.

	Cluster 1	Cluster 2	Cluster 3	200 K	NMR
$< E_{\mathrm{tot}} >$	-1488.2 ± 2.7	-1489.1 ± 2.6	-1488.2 ± 2.6	-1489.3 ± 2.6	–
$< E_{\mathrm{v}} >$	-215.7 ± 2.5	-214.2 ± 2.3	-215.5 ± 2.4	-214.4 ± 2.3	–
$< E_{\mathrm{c}} >$	-1320.8 ± 1.6	-1322.0 ± 1.6	-1320.8 ± 1.7	-1321.8 ± 1.6	–
$< E_{\mathrm{t}} >$	13.0 ± 1.9	13.2 ± 2.0	12.9 ± 1.9	13.1 ± 1.9	–
$< E_{\mathrm{constr}} >$	2.06 ± 0.50	0.66 ± 0.49	2.02 ± 0.49	0.81 ± 0.65	–
< RMSD >	2.45 ± 0.10	0.73 ± 0.08	2.56 ± 0.10	0.94 ± 0.59	0.0
< RGYR >	10.47 ± 0.05	10.13 ± 0.02	10.46 ± 0.04	10.17 ± 0.11	10.00
< IHCA >	52.0 ± 1.7	50.7 ± 2.2	52.2 ± 1.8	50.9 ± 2.2	45.8
< SA >	3212.3 ± 25.2	3137.9 ± 29.8	3216.5 ± 25.4	3146.5 ± 38.7	3152.3

Table 4.2: Various average properties of the structures classified by the principal component analysis, which were obtained at the temperature of 200 K by the REM simulation with the dielectric constant $\epsilon = 4.0$. See the caption of Table 4.1.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
$< E_{\mathrm{tot}} >$	−500.9 ± 2.7	−498.9 ± 2.9	−498.2 ± 2.6	−499.1 ± 2.8	−500.6 ± 2.7
$< E_{\mathrm{vdw}} >$	−220.0 ± 2.2	−218.0 ± 2.3	−216.2 ± 2.2	−217.9 ± 2.3	−219.7 ± 2.2
$< E_{\mathrm{c}} >$	−326.9 ± 0.7	−327.9 ± 0.8	−328.0 ± 0.9	−328.0 ± 0.7	−326.9 ± 0.7
$< E_{\mathrm{t}} >$	11.9 ± 1.9	11.8 ± 1.7	12.1 ± 1.8	11.7 ± 1.7	11.9 ± 1.9
$< E_{\mathrm{constr}} >$	0.93 ± 0.42	1.95 ± 0.55	0.68 ± 0.51	1.92 ± 0.52	0.97 ± 0.47
< RMSD >	4.39 ± 0.07	2.37 ± 0.09	0.70 ± 0.07	2.51 ± 0.09	4.46 ± 0.07
< RGYR >	10.84 ± 0.03	10.44 ± 0.04	10.13 ± 0.03	10.45 ± 0.04	10.84 ± 0.03
< IHCA >	16.1 ± 1.3	51.3 ± 1.3	50.3 ± 2.0	51.3 ± 1.4	16.0 ± 1.4
< SA >	3097.5 ± 24.8	3208.1 ± 27.1	3130.6 ± 32.6	3201.0 ± 25.4	3098.8 ± 24.3
	Cluster 6	Cluster 7	Cluster 8	Cluster 9	200 K
$< E_{\mathrm{tot}} >$	−499.2 ± 2.8	−499.1 ± 2.8	−499.1 ± 3.1	−498.5 ± 2.4	−500.6 ± 2.8
$< E_{\mathrm{vdw}} >$	−219.9 ± 2.4	−219.9 ± 2.3	−217.7 ± 2.8	−217.9 ± 1.9	−219.6 ± 2.4
$< E_{\mathrm{c}} >$	−325.9 ± 0.8	−326.0 ± 0.7	−327.5 ± 0.7	−327.2 ± 0.7	−327.0 ± 0.9
$< E_{\mathrm{t}} >$	12.7 ± 1.9	12.8 ± 2.1	12.7 ± 2.1	12.8 ± 2.2	11.9 ± 1.9
$< E_{\mathrm{constr}} >$	0.76 ± 0.54	0.79 ± 0.49	0.22 ± 0.19	0.67 ± 0.53	0.98 ± 0.53
< RMSD >	5.76 ± 0.08	5.72 ± 0.09	3.49 ± 0.08	3.49 ± 0.05	4.20 ± 0.99
< RGYR >	10.59 ± 0.06	10.60 ± 0.05	10.29 ± 0.03	10.27 ± 0.02	10.75 ± 0.19
< IHCA >	20.7 ± 1.7	21.1 ± 1.9	38.3 ± 1.4	38.1 ± 1.4	20.9 ± 11.5
< SA >	3048.2 ± 29.9	3049.4 ± 30.6	3072.5 ± 34.3	3072.7 ± 24.6	3103.3 ± 41.6

Table 4.3: The interhelical distances (in Å) of the solid-state NMR structure [15], the solution NMR structure [13], the structures of Cluster 2 by the REM simulation with the dielectric constant $\epsilon = 1.0$, and the structures of Cluster 3 by the REM simulation with the dielectric constant $\epsilon = 4.0$.

pairs of atoms		solid-state NMR	solution NMR	Cluster 2	Cluster 3
				($\epsilon = 1.0$)	($\epsilon = 4.0$)
Gly79 C	Gly79 CA	4.1	4.7	4.3	4.3
Gly79 CA	Ile76 C	4.8	4.8	5.3	5.4
Gly83 C	Gly83 CA	4.3	5.1	4.8	4.9
Gly83 CA	Val80 C	4.2	4.3	4.3	4.4
Gly79 C	Val80 C	4.0	2.9	4.0	4.2
Gly83 C	Val84 C	4.0	3.7	3.7	4.0

Table 4.4: Free energy, internal energy, and entropy of the global- and local-minimum free energy states with the dielectric constant $\epsilon = 1.0$ and $\epsilon = 4.0$ at 200 K.

$\epsilon = 1.0$	Clusters 1 and 3	Cluster 2
ΔF (kcal/mol)	0.77	0.00
ΔU (kcal/mol)	0.90	0.00
$-T\Delta S$ (kcal/mol)	-0.13	0.00

$\epsilon = 4.0$	Clusters 1 and 5	Clusters 2 and 4	Cluster 3
ΔF (kcal/mol)	0.00	0.92	1.16
ΔU (kcal/mol)	0.00	1.75	2.60
$-T\Delta S$ (kcal/mol)	0.00	-0.83	-1.44

$\epsilon = 4.0$	Clusters 6 and 7	Clusters 8 and 9
ΔF (kcal/mol)	0.92	1.86
ΔU (kcal/mol)	1.67	2.07
$-T\Delta S$ (kcal/mol)	-0.75	-0.21

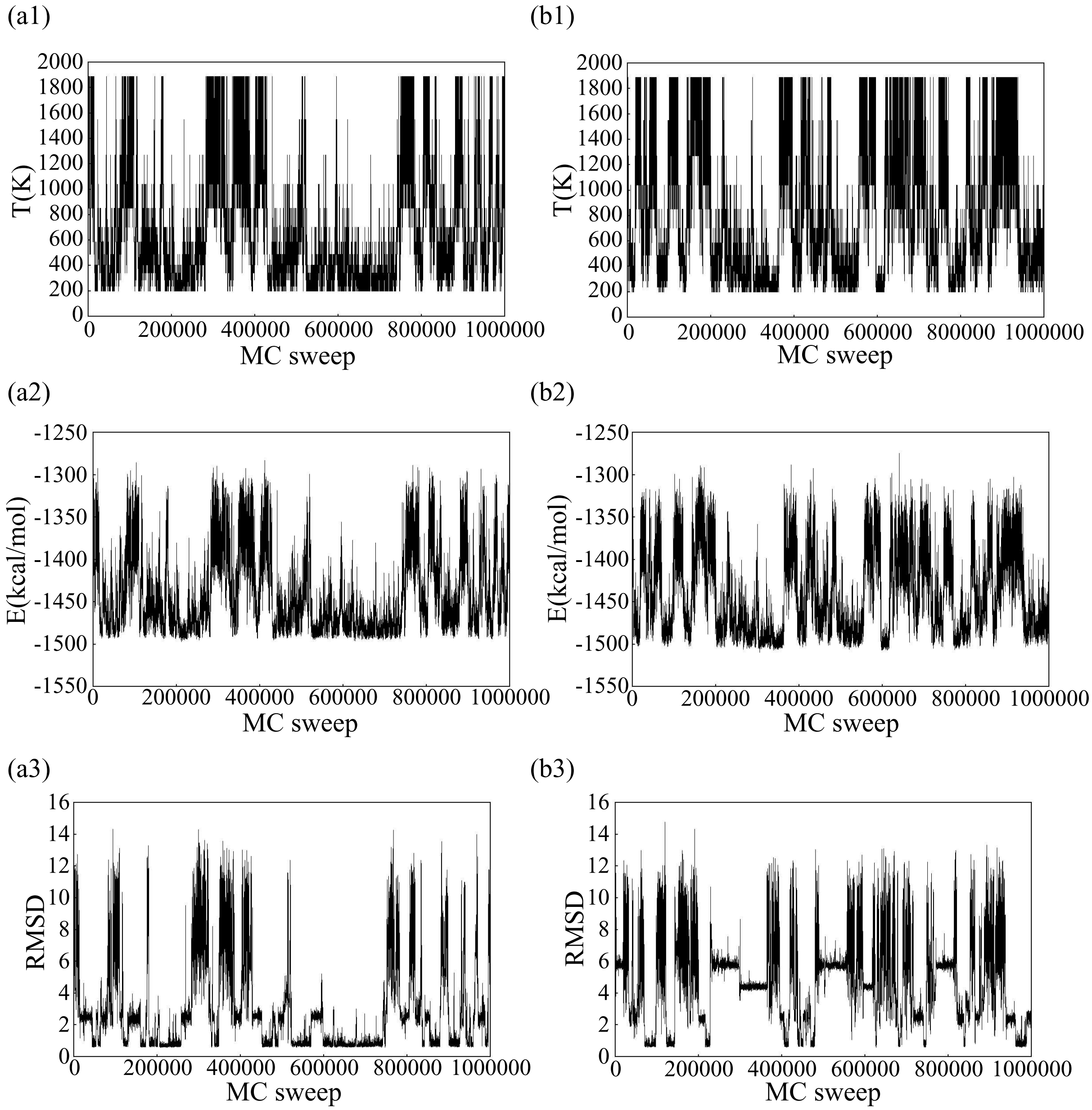

Figure 4.1: Time series of temperature exchange (a1), total potential energy (a2), and RMSD (in Å) of backbone atoms from the NMR structure (a3) for one of the replicas in the case of $\epsilon = 1.0$. (b1), (b2), and (b3) are the corresponding figures in the case of $\epsilon = 4.0$. The NMR structure is Model 16 of the PDB code 1AFO [13], which gave the smallest RMSD at low temperatures in the present simulation with $\epsilon = 1.0$. The smallest RMSD value for Model 16 was 0.6 Å, while that for other Models varied for 0.7 Å to 1.4 Å.

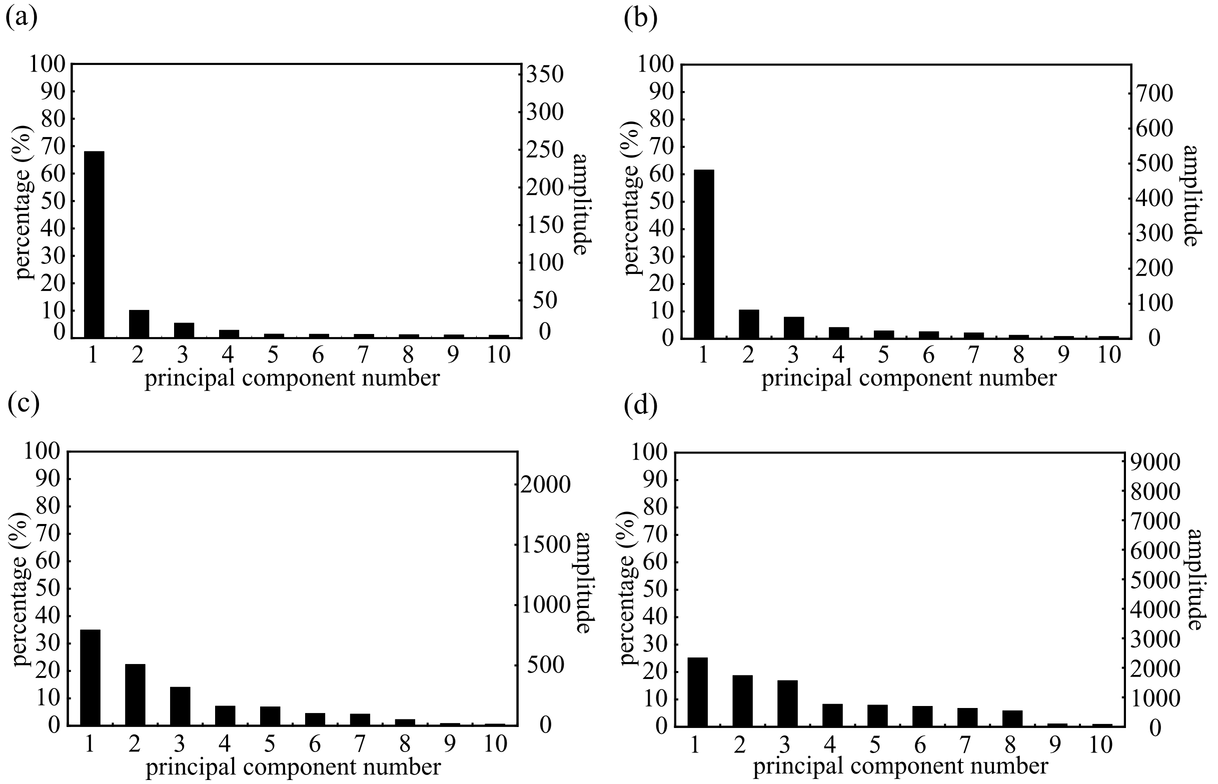

Figure 4.2: The percentage (on the left ordinate) and the amplitude (on the right ordinate) of the principal components from the structures in the replica-exchange simulation with the dielectric constant $\epsilon = 1.0$ at the following temperatures: 200 K (a), 342 K (b), 585 K (c), and 1888 K (d).

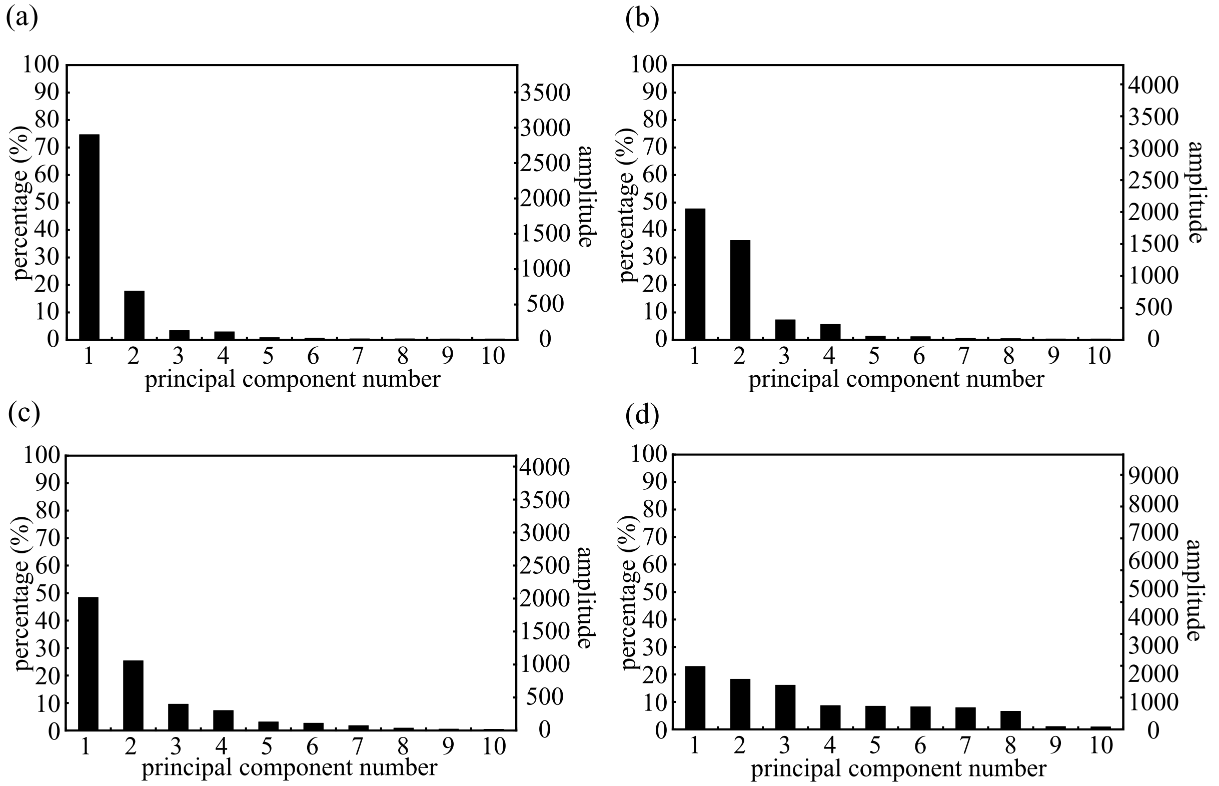

Figure 4.3: The percentage (on the left ordinate) and the amplitude (on the right ordinate) of the principal components from the structures in the replica-exchange simulation with the dielectric constant $\epsilon = 4.0$ at the following temperatures: 200 K (a), 342 K (b), 585 K (c), and 1888 K (d).

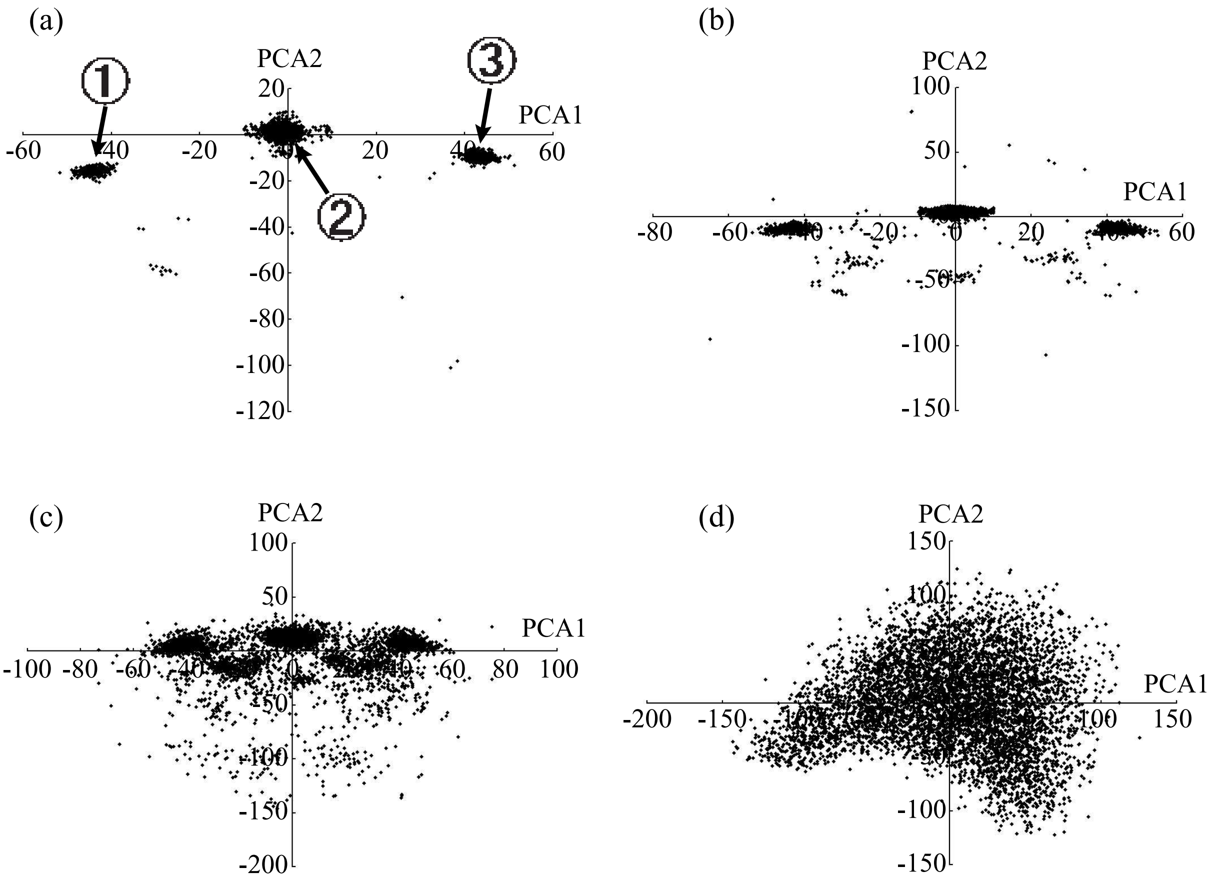

Figure 4.4: The projection of the sampled structures on the 1st and 2nd principal axes from the replica-exchange simulations with the dielectric constant $\epsilon = 1.0$ at the following temperatures: 200 K (a), 342 K (b), 585 K (c), and 1888 K (d). The circled numbers "1", "2", and "3" in (a) stand for Clusters 1, 2, and 3, respectively.

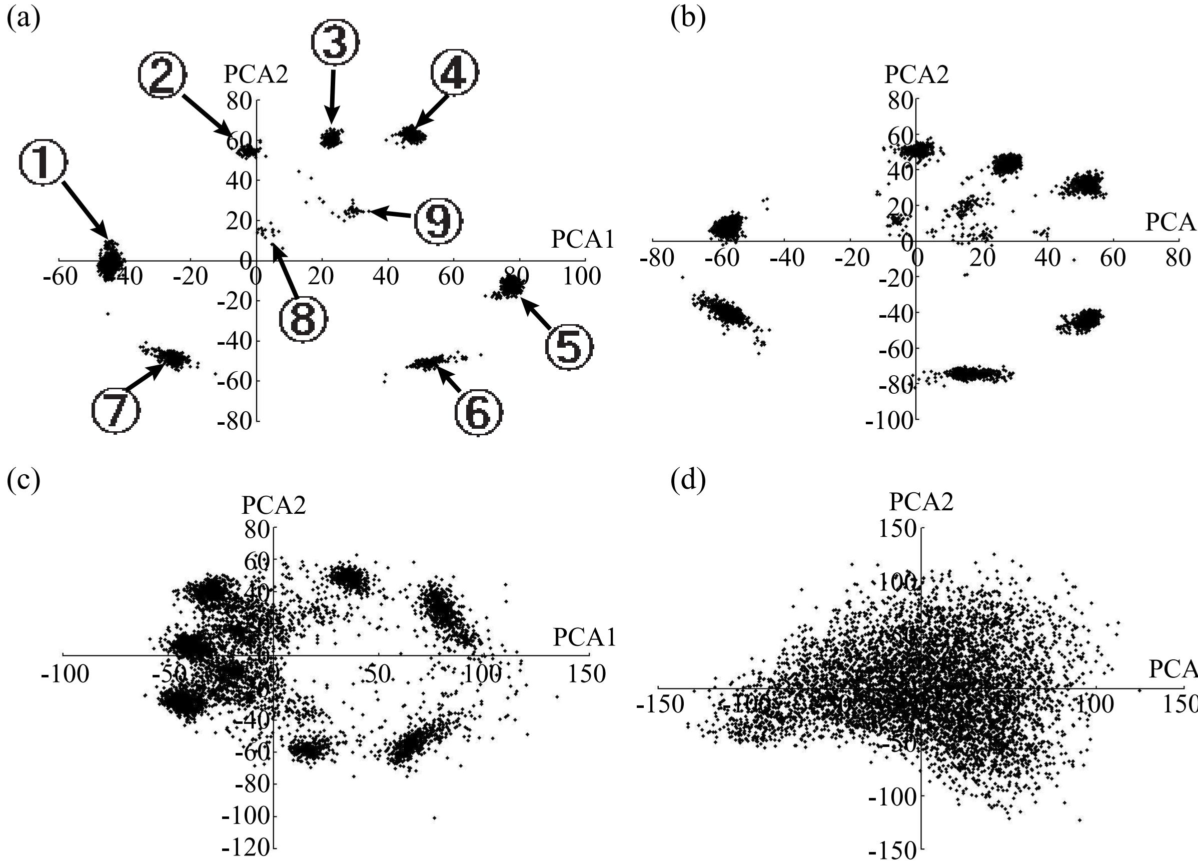

Figure 4.5: The projection of the sampled structures on the 1st and 2nd principal axes from the replica-exchange simulations with the dielectric constant $\epsilon = 4.0$ at the following temperatures: 200 K (a), 342 K (b), 585 K (c), and 1888 K (d). The circled numbers "1", "2", $\cdots$, "9" in (a) stand for Clusters 1, 2, $\cdots$, 9, respectively.

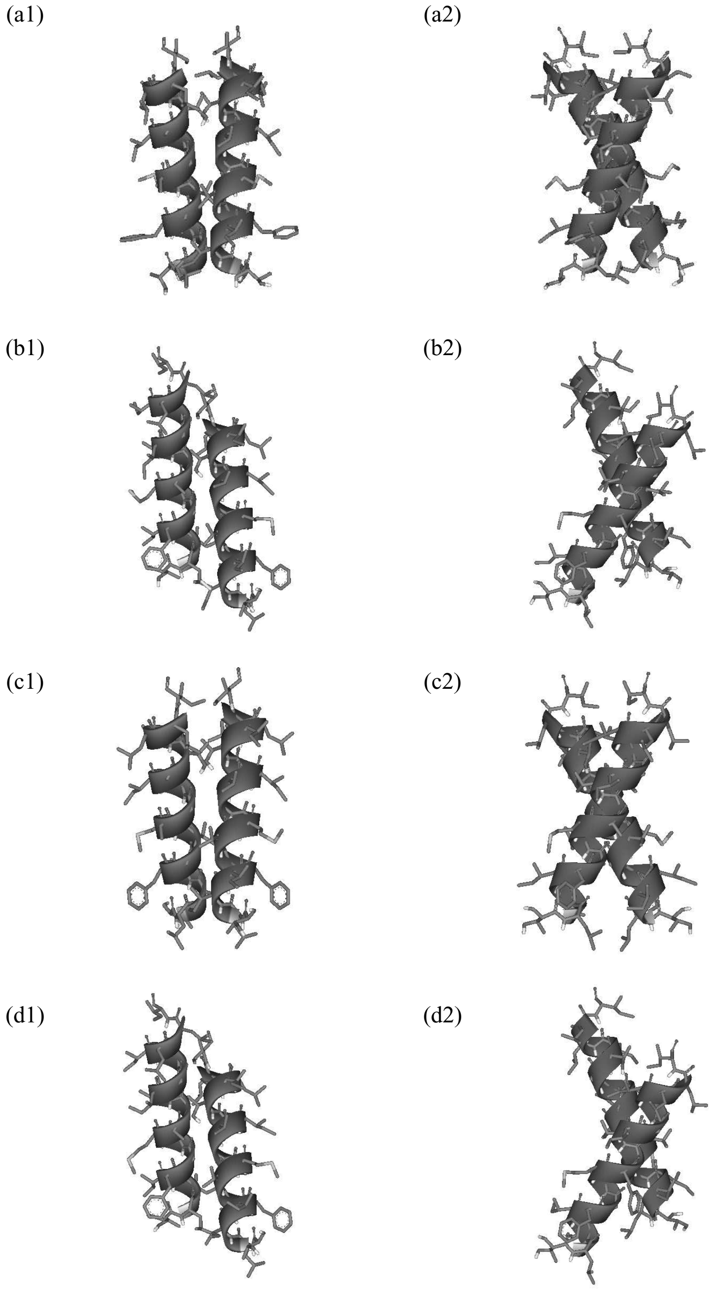

Figure 4.6: The NMR structure (Model 16 of the PDB code 1AFO) and typical cluster structures of the principal component analysis at the lowest temperature (200 K) from the REM simulation with the dielectric constant $\epsilon = 1.0$. (a1) and (a2) are the same structure viewed from different angles. Similarly, (b1) and (b2), (c1) and (c2), and (d1) and (d2) are the same structures viewed from different angles, respectively. (a) is the NMR structure. (b), (c), and (d) are the typical configurations of Cluster 1, Cluster 2, and Cluster 3, respectively. The figures were created with Viewer Lite.

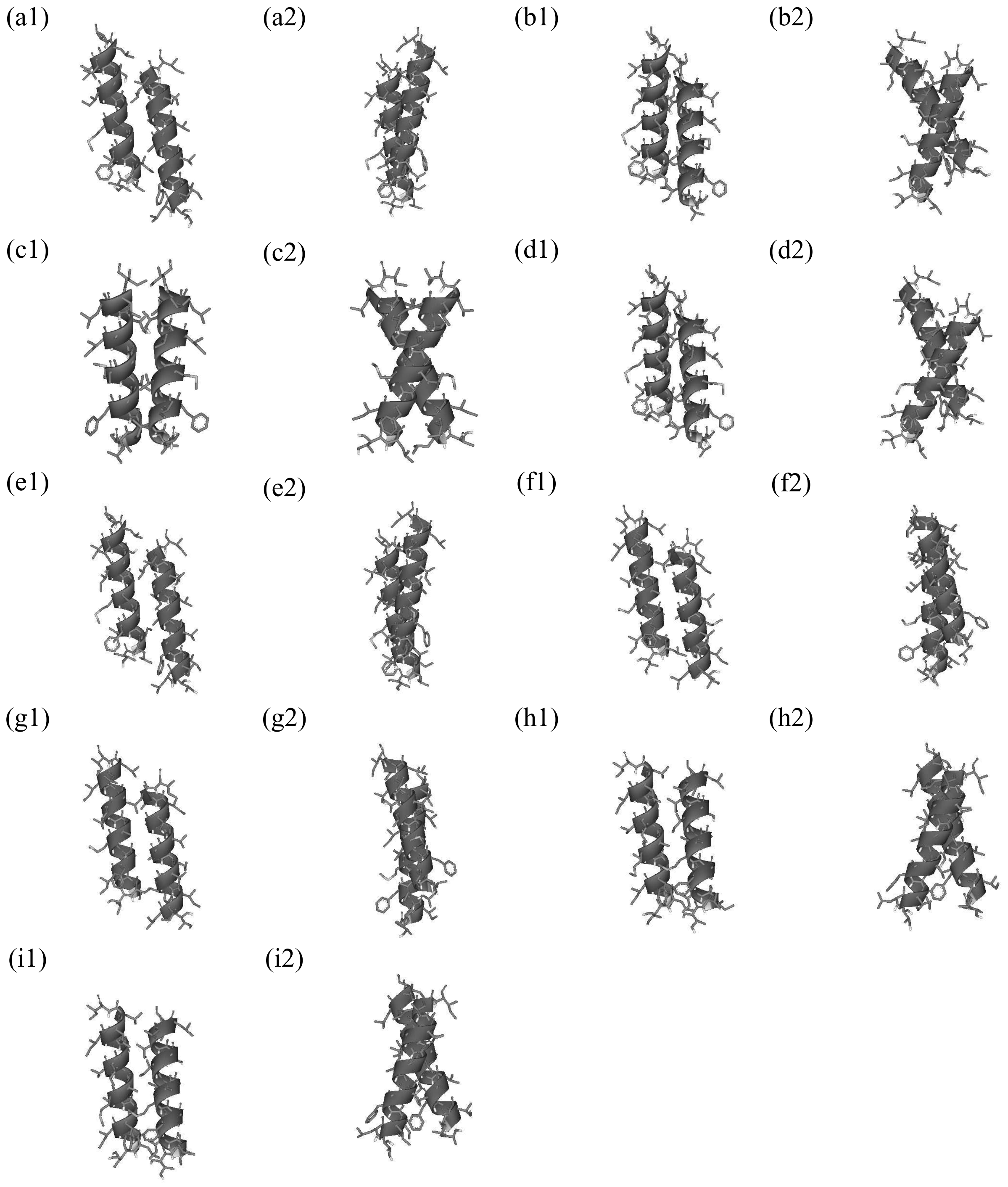

Figure 4.7: Typical cluster structures of the principal component analysis at the lowest temperature (200 K) from the REM simulation with the dielectric constant $\epsilon = 4.0$. (a1) and (a2) are the same structure viewed from different angles. (a), (b), (c), (d), (e), (f), (g), (h), and (i) are the typical configurations of Cluster 1, Cluster 2, Cluster 3, Cluster 4, Cluster 5, Cluster 6, Cluster 7, Cluster 8, and Cluster 9, respectively. The figures were created with Viewer Lite.

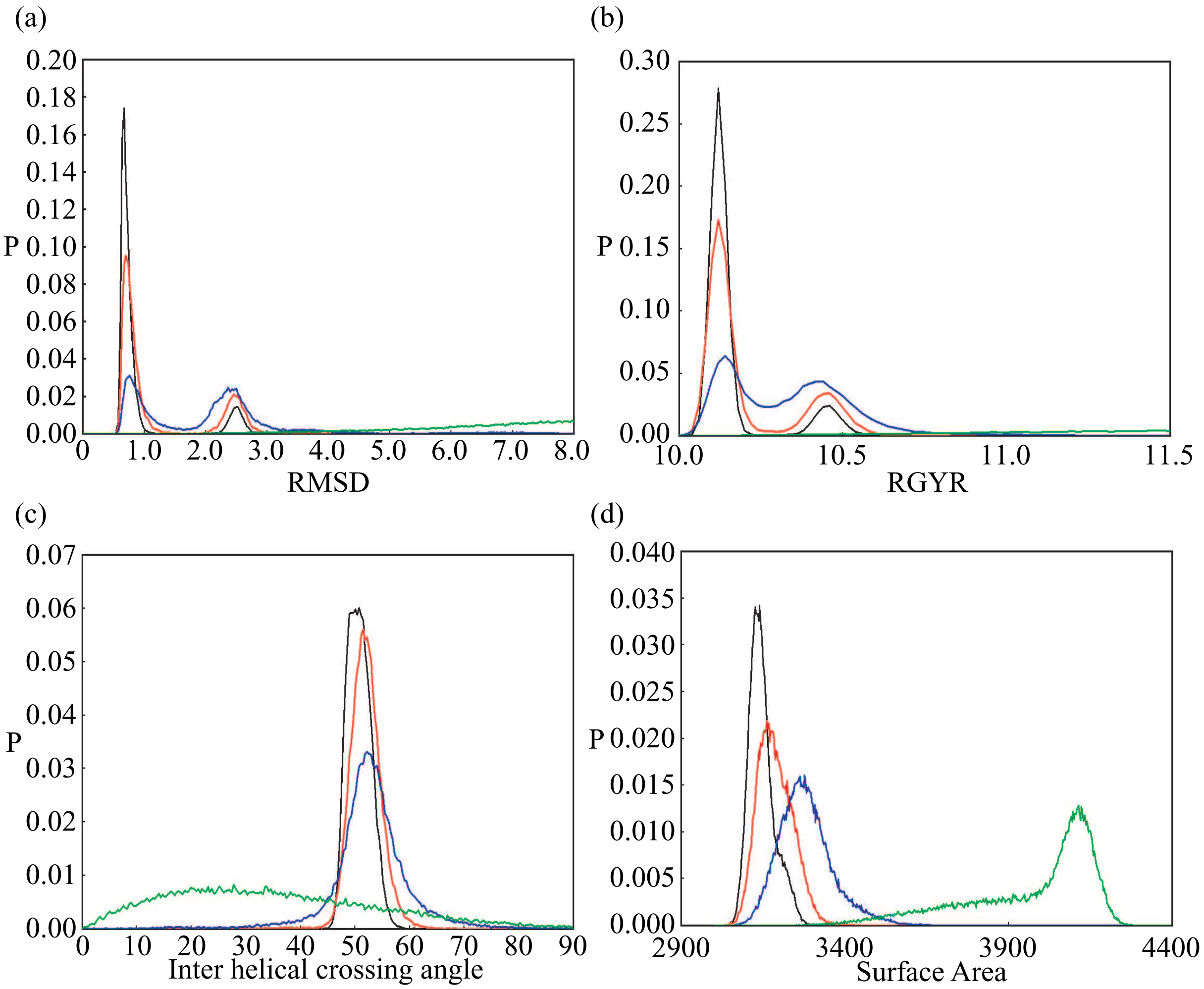

Figure 4.8: The probability distributions of RMSD (in Å) from the native structure (a), radius of gyration, RGYR, (in Å) (b), interhelical crossing angle (in degrees) (c), and surface area (in Å^2) (d) obtained from the replica-exchange MC simulation with the dielectric constant $\epsilon = 1.0$ at the following temperatures: 200 K (black), 342 K (red), 585 K (blue), and 1888 K (green).

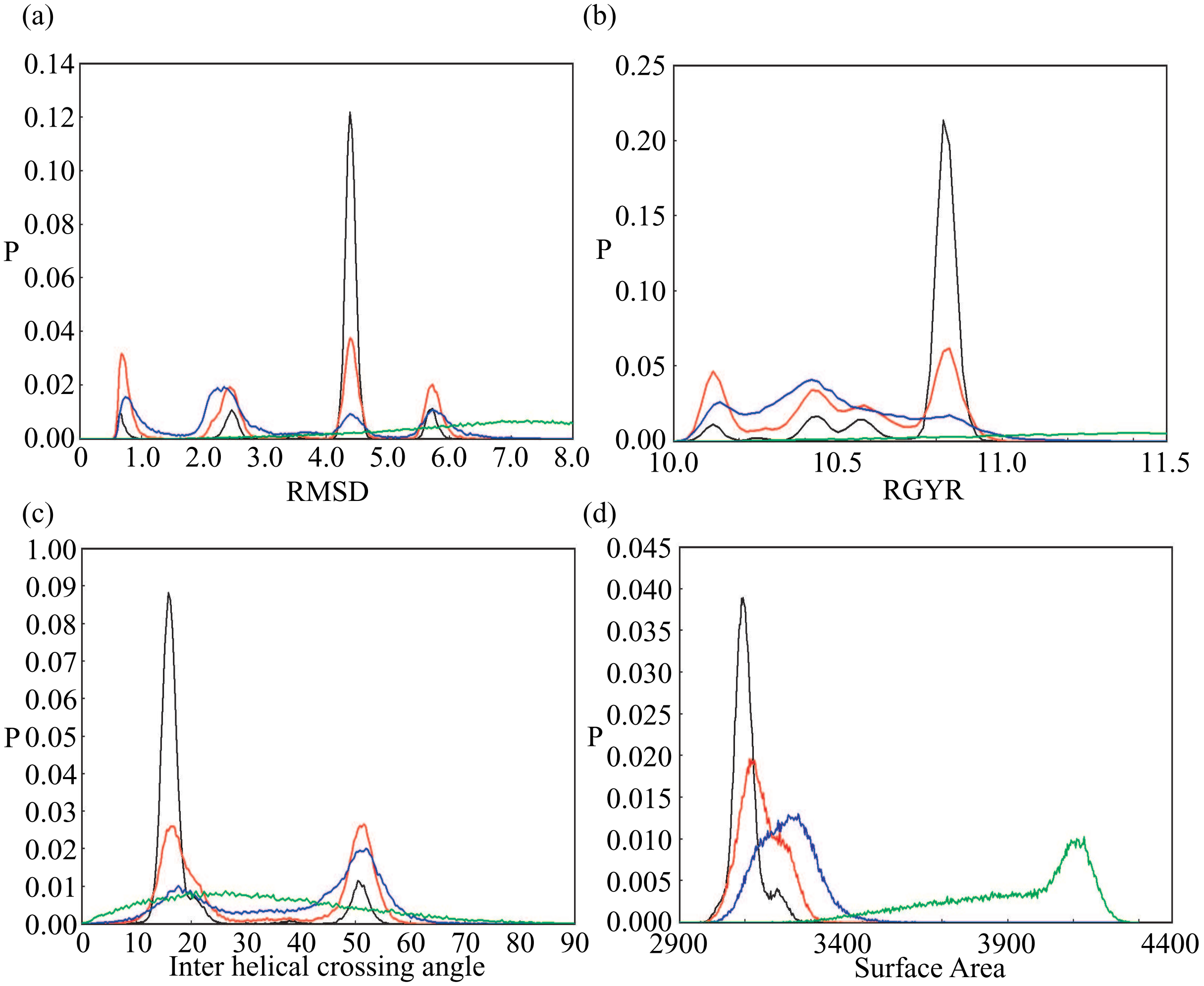

Figure 4.9: The probability distributions of RMSD (in Å) from the native structure (a), radius of gyration, RGYR, (in Å) (b), interhelical crossing angle (in degrees) (c), and surface area (in Å^2) (d) obtained from the replica-exchange MC simulation with the dielectric constant $\epsilon = 4.0$ at the following temperatures: 200 K (black), 342 K (red), 585 K (blue), and 1888 K (green).

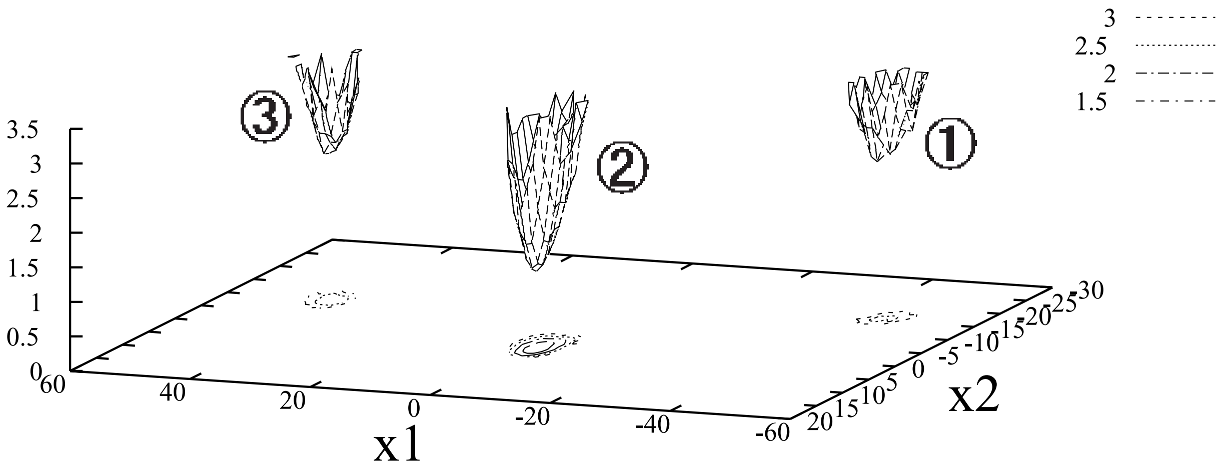

Figure 4.10: The free energy landscape $P(x_1, x_2)$ calculated from the replica-exchange MC simulation at 200 K with the dielectric constant $\epsilon = 1.0$. x_1 and x_2 correspond to the first principal component and the second principal component, respectively. The circled numbers "1", "2", and "3" stand for Clusters 1, 2, and 3, respectively.

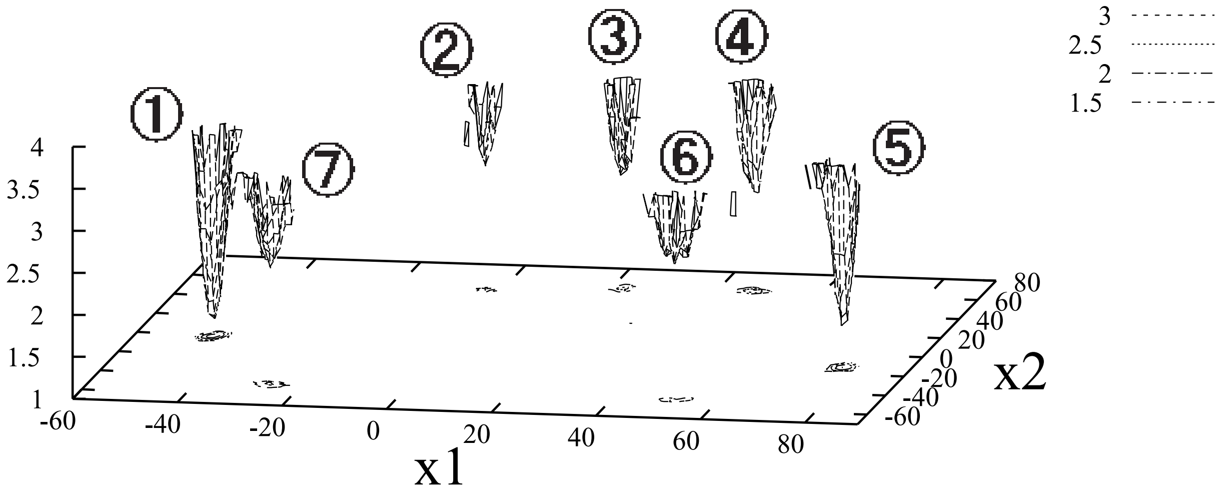

Figure 4.11: The free energy landscape $P(x_1, x_2)$ calculated from the replica-exchange MC simulation at 200 K with the dielectric constant $\epsilon = 4.0$. x_1 and x_2 correspond to the first principal component and the second principal component, respectively. The circled numbers “1”, “2”, $\cdots$, “7” in (a) stand for Clusters 1, 2, $\cdots$, 7, respectively. The valleys from Clusters 8 and 9 are not seen clearly because the numbers of samples are too small.

Chapter 5

Structure Prediction of Bacteriorhodopsin

H. Kokubo and Y. Okamoto, "Self-assembly of transmembrane helices of bacteriorhodopsin by a replica-exchange Monte Carlo simulation," *Chemical Physics Letters* **392**, 168-175 (2004).

5.1 Introduction

According to Anfinsen's dogma [1], the native structures of globular proteins correspond to the global-minimum free-energy states. The protein folding is thus governed by the principles of physical chemistry. We expect the same for membrane proteins, although the environment is anisotropic, inhomogeneous, and complicated; membrane protein structures are determined by the interactions of the peptide chains with each other, the lipid bilayer hydrocarbon core, the bilayer interface, and with water molecules (for a review, see [2]). Transmembrane regions of membrane proteins are often composed of either α-helices or β-sheets. In this chapter, we consider the former case only. It is considered that membrane protein folding energetics have three main aspects: protein binding in bilayer interfaces, transmembrane helix insertion, and helix associations. In this chapter we address the issue of the final stage of the structure formation, namely helix associations.

The two-stage model was proposed for the structure formation of membrane proteins which are composed of several transmembrane helices [3]. In the two-stage model, individual helices of a membrane protein are postulated to be stable separately as domains in a lipid bilayer and then side-to-side helix association is driven, resulting in the native structure. The transmembrane helices of membrane proteins are considered to assemble by themselves into functional proteins. Some experimental evidences show that this is true for bacteriorhodopsin [4], lactose permease [5], rhodopsin [6], and the red cell anion exchanger protein [7]. Therefore, it is reasonable to assume that processes of helix formation and positioning can be predicted separately and that helices behave as independent stable objects. Taking this knowledge into consideration, the efforts for the search in the conformational space are greatly reduced in our simulations.

In this thesis we proposed a prediction method [8, 9, 10] for helical membrane protein structures by the replica-exchange method (REM) [11]-[14]. REM can sample a wide configurational space without getting trapped in local-minimum free energy states and we can find stable structures at low temperatures (for a review, see Ref. [15]).

As discussed in detail in Chapters 3 and 4, we already applied our method to the structure prediction of glycophorin A transmembrane dimer as a test case and the prediction was successful. We assumed that the native structure is determined by the helix-helix interactions and so we neglected the rest of the protein such as loop regions and the surrounding lipids and water molecules [8, 9, 10]. In this chapter we target bacteriorhodopsin. Bacteriorhodopsin consists of seven transmembrane helices and is one of G protein-coupled receptors (GPCRs). GPCRs play key roles for our sense of vision, smell, etc. They are common targets for drug design. GPCR is also one of the smallest molecular machines and their mechanisms of functions are yet to be understood. Therefore, clarifying an atomistic-level mechanism of structure stability of GPCRs is very important from the physical and biological standpoints.

Here, we comment on the previous research of structure predictions of GPCRs by simulations. In Ref. [16] Vaidehi *et al.* target prediction of structure of GPCRs. However they used the experimental electron density map, and they did not sample wide confomational space. Their simulation is only the equilibrium simulation around the initial conformation. On the other hand, we sample wide conformational space by shuffling helices in our approach. In Refs. [17, 18] Suwa *et al.* assumed that membrane proteins are stabilized mostly by polar interactions and developed their original coarse-grained force field based on only polar interactions. This assumption is based on the early hypothesis, but as is discussed in Ref. [2] this does not appear to be correct from the modern knowledge, that is, the interiors of membrane proteins are similar to those of soluble proteins. Therefore their coarse-grained force field will not be versatile because we also have to estimate van der Waals energy, torsion energy, and so on. In fact our previous study showed that not only electrostatic energy but also van der Waals energy and torsion energy are important for the structure stability [8]-[10]. In Ref. [18] Hirokawa *et al.* proposed the triangle lattice model for predicting membrane protein structures. However, we think that it is generally impossible to estimate the energy correctly by such a corase-grained model. We showed that the energy difference between the structures which have different helix conformations is within several kcal/mol [8]-[10]. Triangle lattice model is obviously too coarse to estimate these differences. We therefore use the versatile force field with the atomistic details.

We examine in this chapter whether the native structure is recovered from random initial helix configurations by a REM simulation.

In Sec. 5.2 the computational details of the REM simulation for bacteriorhodopsin are explained. In Sec. 5.3 the results and discussion are presented.

5.2 Computational Details

We now explain our computational details. At first seven helix structures of transmembrane regions are extracted from the PDB structure (PDB code: 1C3W). Seven transmembrane helices are named A, B, C, D, E, F, and G from the N-terminus. The number of amino acids of Helices A, B, C, D, E, F, and G is 23, 22, 23, 23, 23, 25, and 24, respectively (the total number is 163), and their sequences are EWIWLALGTALMGLGTLYFLVKG, KFYAITTLVPAIAFTMYLSMLL, IYWARYADWLFTTPLLLLDLALL, QGTILALVGADGIMIGTGLVGAL, RFVWWAISTAAMLYILYVLFFGF, TFKVLRNVTVVLWSAYPVVWLIGSE, and LNIETLLFMVLDVSAKVGFGLILL. The N and C termini of each helix were blocked with acetyl and N-methyl groups, respectively. We consider that the inclusion of atomistic details of side chains is essential for estimating the energy balance accurately because helices are tightly packed, and thus we use a standard force field such as CHARMM param19 parameter set (polar hydrogen model) [19, 20] for the potential energy of the system. No cutoff was introduced to the non-bonded energy terms. The computer code based on the CHARMM macromolecular mechanics program [21] was used and REM was implemented in it. REM has already been explained in subsection 2.3.

Each helix structure was minimized subject to harmonic restraints on all the heavy atoms. We treat the backbone of the α-helices as rigid body and fix the backbone structures of helices. Only side-chain structures are made flexible. Each helix also has the freedom of translation and rotation. This is introduced following the two-stage model, in which each helix is stable as a domain and the native configurations are built mainly by the interactions between helices. We believe that the flexibility of side chains is also important because membrane proteins are very tightly packed and the packed structures are searched by varying side-chain structures. In principle, we can also use molecular dynamics method, but we employ Monte Carlo (MC) algorithm here. We update configurations with rigid translations and rigid rotations of each α-helix and torsion rotations of side chains. There are $2N_{\mathrm{H}} + N_{\mathrm{D}}$ kinds of MC moves, where N_{H} is the total number of transmembrane helices in the protein, and N_{D} is the total number of dihedral angles in the side chains of N_{H} helices. The first term corresponds to the rigid translation and rigid rotation of the helices and the second to the dihedral-angle rotations in the side chains. One MC step in this chapter is defined to be an update of one of these degrees of freedom which is accepted or rejected according to the Metropolis criterion.

We add the three simple harmonic constraints described in Sec. 2.2 to the original potential energy in order to make conformational sampling efficiency better and mimic the effects of membrane boundary surfaces.

The values for our simulation of bacteriorhodopsin are set to $N_{\mathrm{H}} = 7$, $k_1 = 1.0$ (kcal/mol)/Å^2, and $d_{i,i+1} = 20.0$ Å in Eq. (2.2), $k_2 = 1.0$ (kcal/mol)/Å^2, $z_0^{\mathrm{L}} = 0.0$ Å, $z_0^{\mathrm{U}} = 31.5$ Å, and $d^{\mathrm{U}} = d^{\mathrm{L}} = 2.0$ Å in Eq. (2.4) , and $k_3 = 0.05$ (kcal/mol)/Å^2 and $d_{\mathrm{C}_\alpha} = 100$ Å in Eq. (2.5).

In summary, only the transmembrane helices are used in our simulations, and loop regions of membrane proteins as well as lipid and water environment are neglected. Our assumptions are that a role of water is to push the hydrophobic transmembrane regions of membrane proteins into the lipid bilayer and that a major role of lipids is to prepare a hydrophobic environment and construct helix structures in the transmembrane regions. Loop regions of membrane proteins are often outside the membrane and do not directly affect the structure of transmembrane regions [4, 22]. Due to the difference in surface shapes of helices and lipids, the stabilization energy for helix-helix packing will be larger than that for helix-lipid packing. Therefore, water, lipids, and loop-region amino acids are not treated explicitly in our simulations, although the features of membrane boundaries are taken into account by the constraint conditions in Eq. (2.4). We examine whether a structure similar to the native one can be obtained by solely helix-helix interactions or not.

We performed a REM MC simulation of 168,000,000 MC steps. Every simulation was performed with the dielectric constant $\epsilon = 1.0$. In Refs. [8, 9, 10] we showed that $\epsilon = 1.0$ is more appropriate to use in the structure prediction, although $\epsilon = 4.0$ is the value close to that for the lipid environment. This is because almost no lipid molecules can exist between helices; the value $\epsilon = 4.0$ underestimates the electrostatic effects. We used the following 32 temperatures: 200, 218, 238, 260, 284, 310, 338, 369, 410, 455, 505, 561, 623, 691, 768, 853, 947, 1052, 1125, 1202, 1285, 1374, 1469, 1642, 1835, 2051, 2293, 2679, 3132, 3660, 4278, and 5000 K. This temperature distribution was chosen so that all the acceptance ratios are almost uniform and sufficiently large (> 10 %) for computational efficiency [14, 15]. The highest temperature was chosen sufficiently high so that no trapping in local-minimum-energy states occurs. Replica exchange was attempted once at every 50 MC steps.

The initial configurations for the REM simulation are the random structures prepared in the following way. At first a MC simulation of 5,000,000 MC steps was performed at a sufficiently high temperature ($T = 8000$ K) so that no trapping in local-minimum-energy states occurred, starting from the native structure. We picked out 32 initial structures for 32 replicas out of the trajectory from 1,000,000 MC steps to 4,200,000 MC steps at even intervals (every 100,000 MC steps). We show four of the initial configurations of the REM simulation in Fig. 5.1. We see that they are indeed very different from the native one (in Fig. 5.4(a) below). The helices of the initial structures are not packed and disjointed from each other, and the helix configurations and relative angles are very different from the native ones. Note that the initial configurations do not include the information of the native structure in this way.

5.3 Results and Discussion

5.3.1 Performances of the Replica-Exchange Simulation

In Fig. 5.2 we show the "time series" of various quantities from the present REM simulation. In Fig. 5.2(a) the time series of temperature exchange for the chosen four replicas (in Fig. 5.1) are shown. We observe random walks in the temperature space between the lowest and highest temperatures. Other replicas perform random walks similarly. In Fig. 5.2(b) the corresponding time series of the total potential energy are shown. We see that random walks in the potential energy space between low and high energy regions are also realized. Note that there is a strong correlation between the behaviors in Figs. 5.2(a) and 5.2(b), as there should. All these results confirm that the present REM simulation has been properly performed.

We now study how widely the configurational space was sampled during the present simulation. We plot the time series of the root-mean-square deviation (RMSD) with respect to all C_α atoms from the experimental structure (PDB code: 1C3W) in Fig. 5.2(c). When the temperature becomes high, RMSD takes a large value, and when the temperature becomes low, RMSD takes a small value. This implies that the simulation sampled a wide conformational space. In particular, Replica 14 (red curve) has a remarkably small RMSD at relatively low temperatures after 140,000,000 MC steps (the smallest value of Replica 14 in Fig. 5.2(c) is 4.42 Å). This means that the structures close to the native one were sampled as one of stable structures at low temperatures by Replica 14.

In Fig. 5.3 typical snapshots of Replica 14 from the REM simulation are shown. In Fig. 5.3(a) the helix configuration is different from the native one (see Fig. 5.4(a) below). In particular, Helix G is trapped in the center. As the simulation proceeds, the temperature becomes high and then drops to low values by the replica exchange process, and the same helix configuration ("topology") as the native one is finally obtained in Fig 5.3(f). These figures confirm that our simulations indeed sampled a wide configurational space. We see that the REM simulation performs random walks not only in energy space but also in conformational space and that they do not get trapped in one of a huge number of local-minimum-energy states.

5.3.2 Comparison of the Structures Obtained by the Simulation and the Experimental Structure

In Fig. 5.4 the PDB structure, the smallest RMSD structure, and the global-minimum-energy structure obtained by the REM simulation are compared. The retinal molecule is included in the native PDB structure (Fig. 5.4(a)), but it was not used in our simulation. Nevertheless, the structure obtained by Replica 14 (Fig. 5.4(b)) has the same helix topology (relative helix configuration) as the native structure. This is consistent with the experimental implications that the retinal molecule is not essential for obtaining the native-like topology [22]. Note that the initial configuration of Replica 14 is very different from the native one (RMSD = 16.39 Å; see Fig. 5.1(b)). It is indeed remarkable that we could obtain a native-like structure from such a random initial configuration, even though we neglected loop regions, retinal, lipids, surrounding water molecules in our simulation. This suggests that the helix-helix interactions are the main driving force in the final stage of the structure formation of membrane proteins.

The structure with the smallest RMSD in Fig. 5.4(b) has a higher temperature ($T =$ 947 K) and a larger energy (-7425.0 kcal/mol) than the minimum-energy structure in Fig. 5.4(c) ($T = 200$ K, -7833.3 kcal/mol) which has a very different helix configuration from the native one. This is because Replica 14, after reaching the smallest RMSD at 140,825,000 MC steps, kept rather high temperatures. The energy of Replica 14 will approach the global-minimum value as the temperature decreases to the lowest value during the replica-exchange process, which did not happen by the present REM simulation; we need more MC steps. Hence, we performed a simulated annealing simulation of 40,000,000 MC steps from the structure with the smallest RMSD in Fig. 5.4(b). The temperature was decreased from $T = 947$ K to 200 K during the simulation and the obtained lowest-energy value was -7824.1 kcal/mol. In this simulation the conformation changed little (RMSD of this lowest-energy structure is 4.53 Å). This means that the structure with the smallest RMSD is one of stable structures at low temperatures.

In Table 5.1 we list the RMSD from the native structure with respect to C_α atoms of some pairs of helices in order to examine which parts resemble the native one closely. We see that the RMSD of pairs among Helices A, B, C, and G in the structure with the smallest RMSD are small and these parts resemble the native structure. Fig. 5.4(a) suggests that these helices have much less contact with the retinal molecule than other helices (Helices D, E, and F). Because we neglected the retinal molecule in our simulation, it is reasonable that we did not get close agreement with the native structure for Helices D, E, and F. The RMSD of pairs among helices in the global-minimum structure are rather large, and no partial structures are similar to the native ones.

In Fig. 5.5 the four pairs of helices (C-G, A-G, B-C, and B-G) that have the lowest RMSD values in Table 5.1 are depicted and compared with the corresponding native helix pairs. These four pairs of helices are shown in ascending order of RMSD. We confirm from this Figure that these partial structures are indeed very similar to the native ones. Note that the structures of the first two pairs of helices (C-G and A-G) obtained by the simulation are in remarkable agreement with those of the PDB structure including side-chain packing (see also the RMSD values with respect to all heavy atoms in Table 5.1).

Bibliography

[1] C.B. Anfinsen, Science 181 (1973) 223.

[2] S.H. White, W.C. Wimley, Annu. Rev. Biophys. Biomol. Struct. 28 (1999) 319.

[3] J.L. Popot, D.M. Engelman, Annu. Rev. Biochem. 69 (2000) 881.

[4] J.L. Popot, S.E. Getchman, D.M. Engelman, J. Mol. Biol. 198 (1987) 665.

[5] E. Bibi, H.R. Kaback, Proc. Natl. Acad. Sci. USA 87 (1990) 4325.

[6] K.D. Ridge, S.S.J. Lee, L.L. Yao, Proc. Natl. Acad. Sci. USA 92 (1995) 3204.

[7] J.D. Groves, M.J.A. Tanner, J. Biol. Chem. 270 (1995) 9097.

[8] H. Kokubo, Y. Okamoto, Chem. Phys. Lett. 383 (2004) 397.

[9] H. Kokubo, Y. Okamoto, J. Chem. Phys. 120 (2004) 10837.

[10] H. Kokubo, Y. Okamoto, J. Phys. Soc. Jpn, 73 (2004), in press.

[11] K. Hukushima, K. Nemoto, J. Phys. Soc. Jpn. 65 (1996) 1604.

[12] K. Hukushima, H. Takayama, K. Nemoto, Int. J. Mod. Phys. C 7 (1996) 337.

[13] C.J. Geyer, Proceedings of the 23rd Symposium on the Interface, edited by E. Keramidas (Interface Foundation, Fairfax Station, 1991) p. 156.

[14] Y. Sugita, Y. Okamoto, Chem. Phys. Lett. 314 (1999) 141.

[15] A. Mitsutake, Y. Sugita, Y. Okamoto, Biopolymers (Pept. Sci.) 60 (2001) 96.

[16] N. Vaidehi, W.B. Floriano, R. Trabanino, S.E. Hall, P. Freddolino, E.J. Choi, G. Zamanakos, W.A. Goddard III, Proc. Natl. Acad. Sci. USA 99 (2002) 12622.

[17] M. Suwa, T. Hirokawa, S. Mitaku, Proteins 22 (1995) 363.

[18] T. Hirokawa, J. Uechi, H. Sasamoto, M Suwa, S. Mitaku, Protein Eng. 13 (2000) 771.

[19] W.E. Reiher, III, Theoretical Studies of Hydrogen Bonding, Ph.D. Thesis, Department of Chemistry, Harvard University, Cambridge, MA, USA, 1985.

[20] E. Neria, S. Fischer, M. Karplus, J. Chem. Phys. 105 (1996) 1902.

[21] B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan, M. Karplus, J. Comput. Chem. 4 (1983) 187.

[22] T.W. Kahn, J.M. Sturtevant, D.M. Engelman, Biochemistry 31 (1992) 8829.

[23] R.A. Sayle, E.J. Milner-White, Trends Biochem. Sci. 20 (1995) 374.

Table 5.1: The RMSD (in Å) of pairs of helices from those of the corresponding PDB structure with respect to all C_α atoms and all heavy atoms. Transmembrane helices are named as A, B, C, D, E, F and G from the N terminus.

Pairs of helices	The smallest RMSD structure[a]		The global-minimum-energy structure[b]	
	RMSD (all C_α atoms)	RMSD (all heavy atoms)	RMSD (all C_α atoms)	RMSD (all heavy atoms)
A and B	2.12	2.99	4.07	5.93
B and C	1.35	2.49	3.74	5.27
C and D	2.63	3.20	4.19	4.42
D and E	3.84	4.47	7.83	9.06
E and F	4.06	5.24	4.66	6.54
F and G	4.14	4.83	2.91	3.64
A and G	0.73	1.78	5.58	6.50
B and G	1.87	2.62	8.53	10.08
C and F	4.50	5.33	3.54	4.57
C and G	0.32	1.69	6.27	7.37

[a] The structure is shown in Fig. 5.4(b). [b] The structure is shown in Fig. 5.4(c).

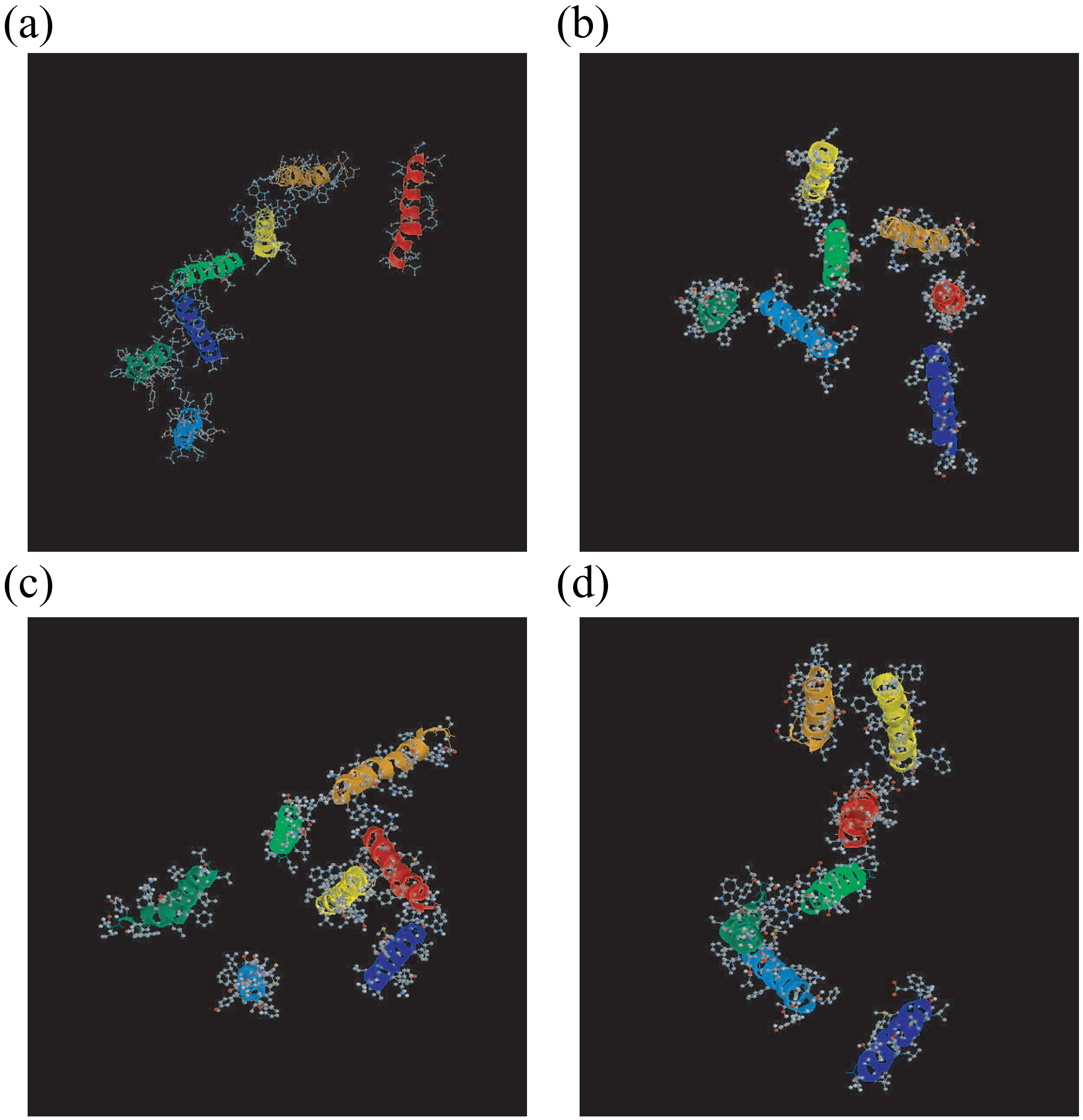

Figure 5.1: The initial structures of the REM simulation for Replica 6 (a), Replica 14 (b), Replica 19 (c), and Replica 25 (d). The RMSD from the native configuration is 23.14 Å (a), 16.39 Å (b), 17.41 Å (c), and 18.13 Å (d) with respect to all C_α atoms. The color of the helices from the N terminus is as follows: Helix A (blue), Helix B (aqua), Helix C (green), Helix D (yellow-green), Helix E (yellow), Helix F (orange), and Helix G (red). The figures were created with RasMol [23].

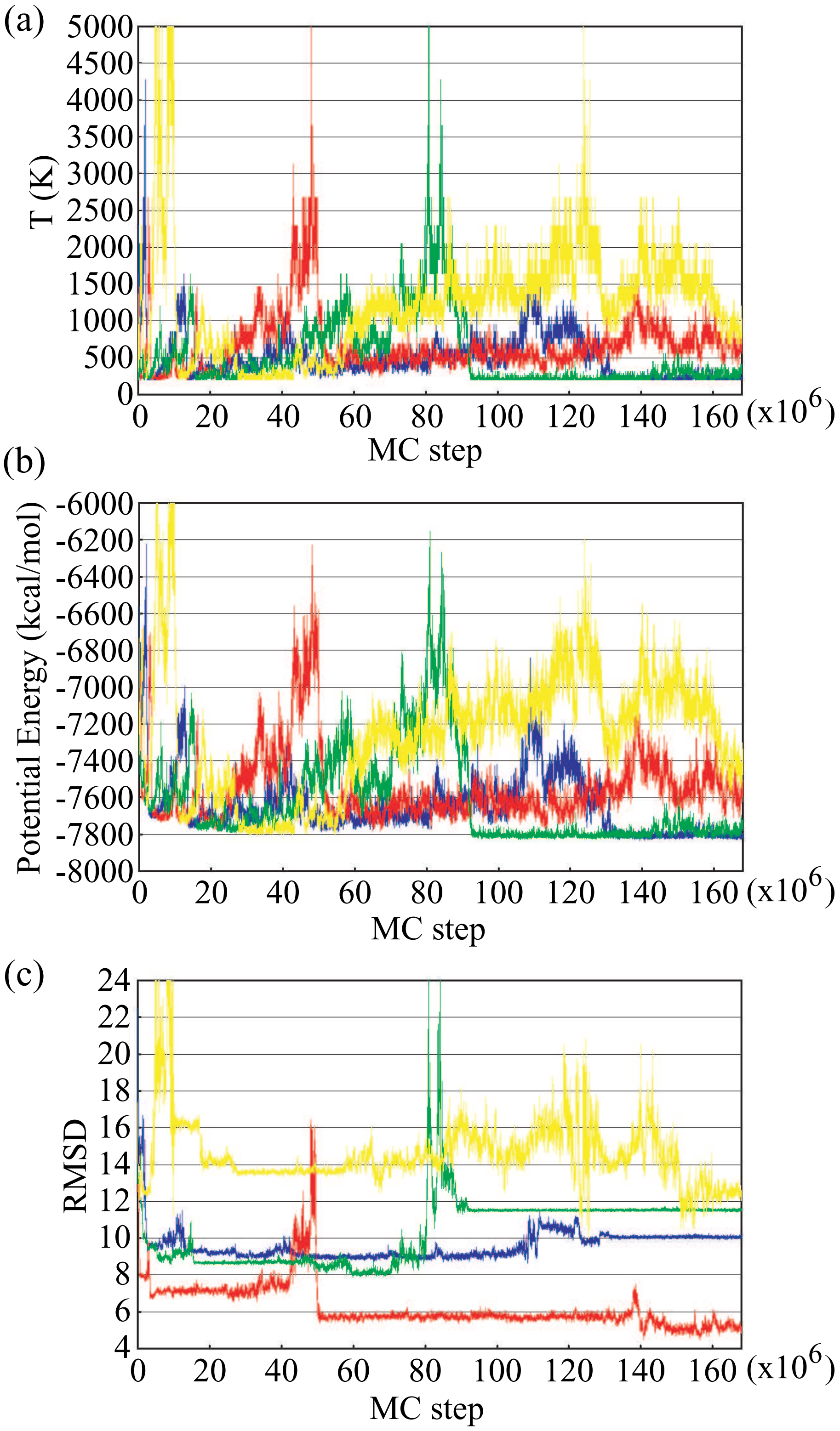

Figure 5.2: Time series of temperature exchange (a), total potential energy (b), and RMSD (in Å) with respect to all C_α atoms from the PDB structure (PDB code: 1C3W) for Replicas 6 (blue), 14 (red), 19 (green), and 25 (yellow). These four replicas correspond to those in Fig. 5.1.

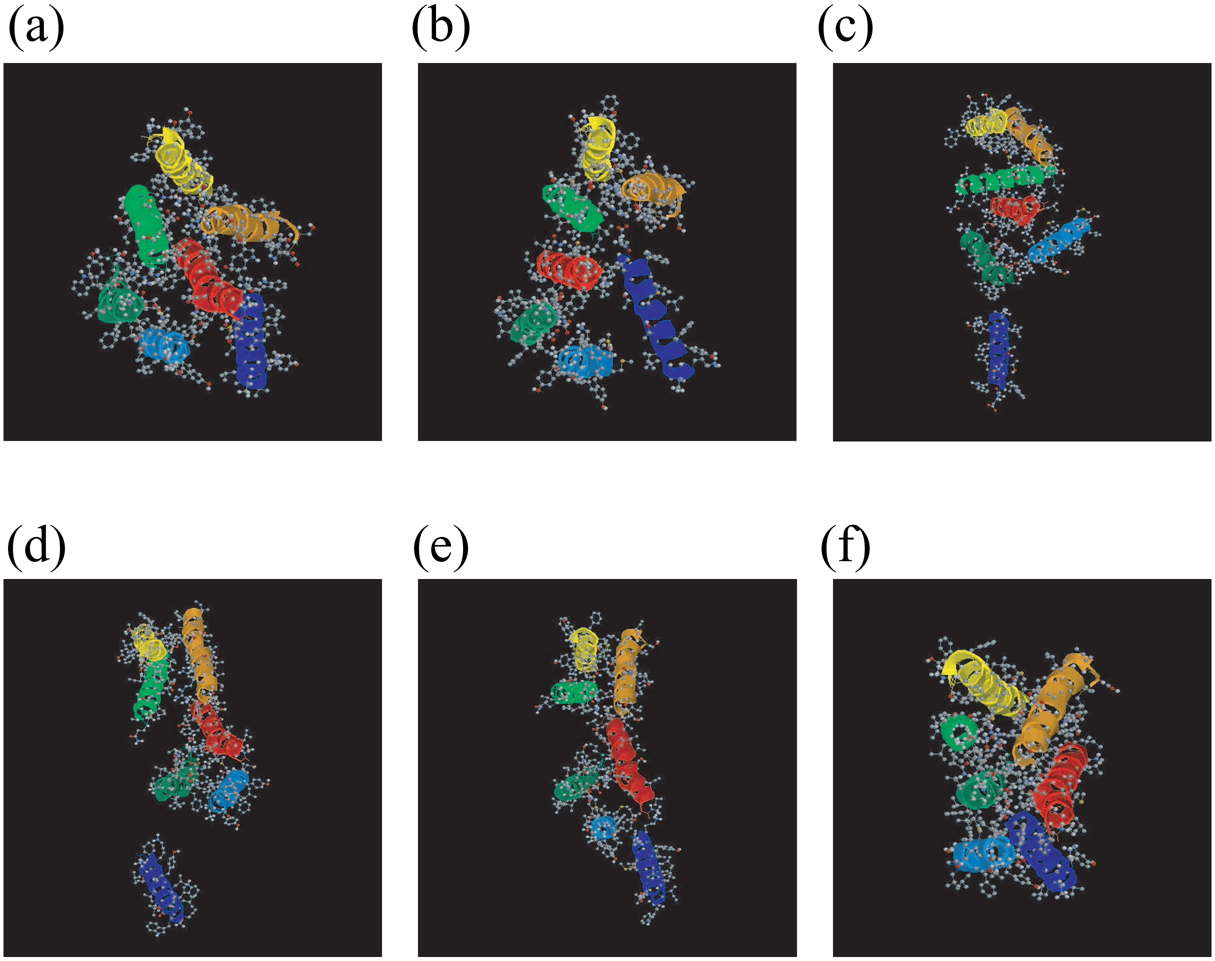

Figure 5.3: Typical snapshots from the REM simulation for Replica 14. The configurations were taken at the 43,146,000-th MC step (a), at the 47,664,000-th MC step (b), at the 48,155,000-th MC step (c), at the 48,822,000-th MC step (d), at the 49,500,000-th MC step (e), and at the 58,398,000-th MC step (f). The RMSD from the native configuration is 7.78 Å (a), 10.84 Å (b), 15.18 Å (c), 14.76 Å (d), 11.71 Å (e) and 5.72 Å (f) with respect to all C_α atoms. The corresponding temperatures are 3132 K (a), 2679 K (b), 3132 K (c), 3132 K (d), 2051 K (e), and 561 K (f). The color of the helices from the N terminus is as follows: Helix A (blue), Helix B (aqua), Helix C (green), Helix D (yellow-green), Helix E (yellow), Helix F (orange), and Helix G (red). The figures were created with RasMol [23].

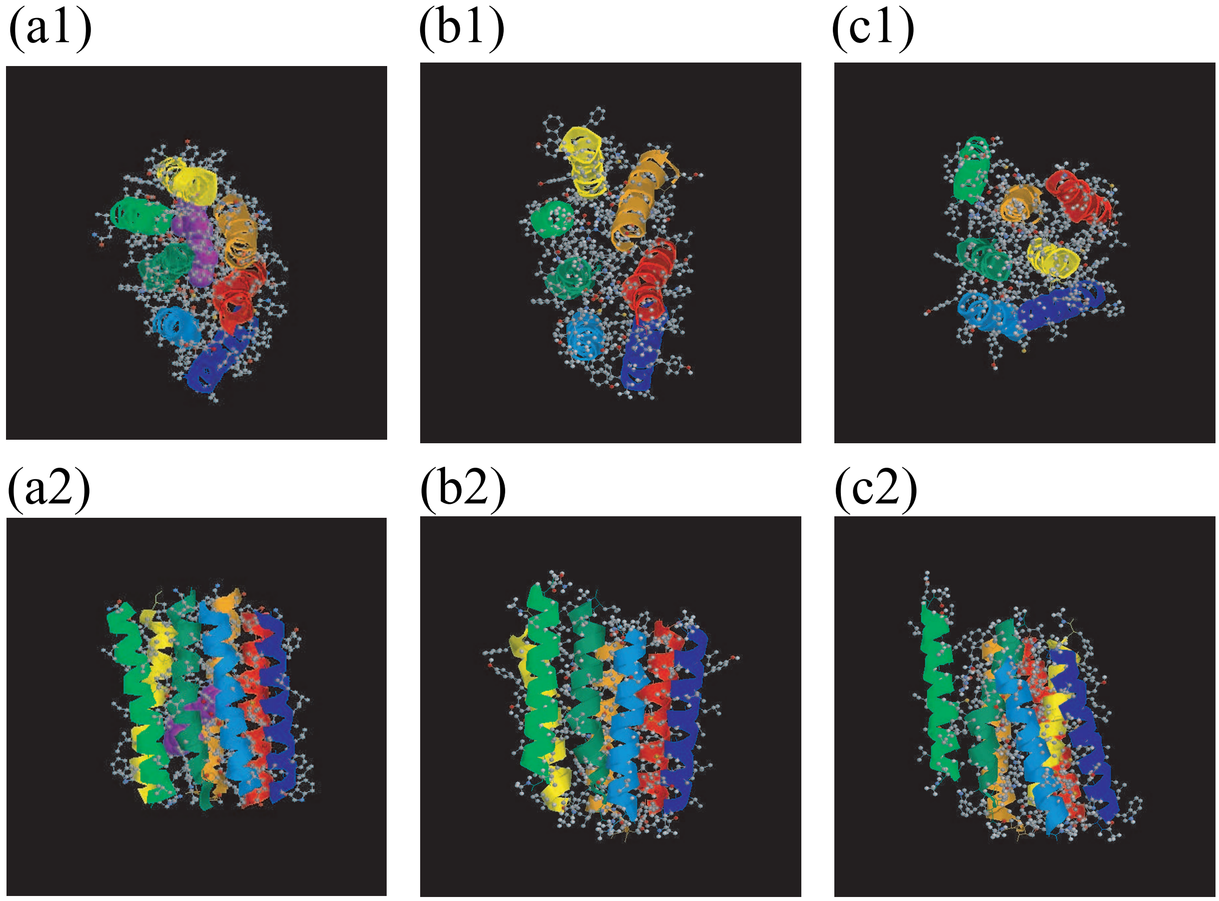

Figure 5.4: (a) The PDB structure (PDB code: 1C3W) with retinal. (b) The smallest RMSD configuration that was obtained by the REM simulation. (c) The global-minimum-energy configuration that was obtained by the REM simulation. (a1) and (a2), (b1) and (b2), and (c1) and (c2) are the same structures viewed from different angles (from top and from side), respectively. Purple-color atoms in (a) represent the retinal. (a) was drawn by eliminating the loop regions and lipids from the PDB file. The RMSD from the native configuration of (a) is 4.42 Å (b) and 10.06 Å (c) with respect to all C_α atoms. The color of the helices from the N terminus is as follows: Helix A (blue), Helix B (aqua), Helix C (green), Helix D (yellow-green), Helix E (yellow), Helix F (orange), and Helix G (red). The figures were created with RasMol [23].

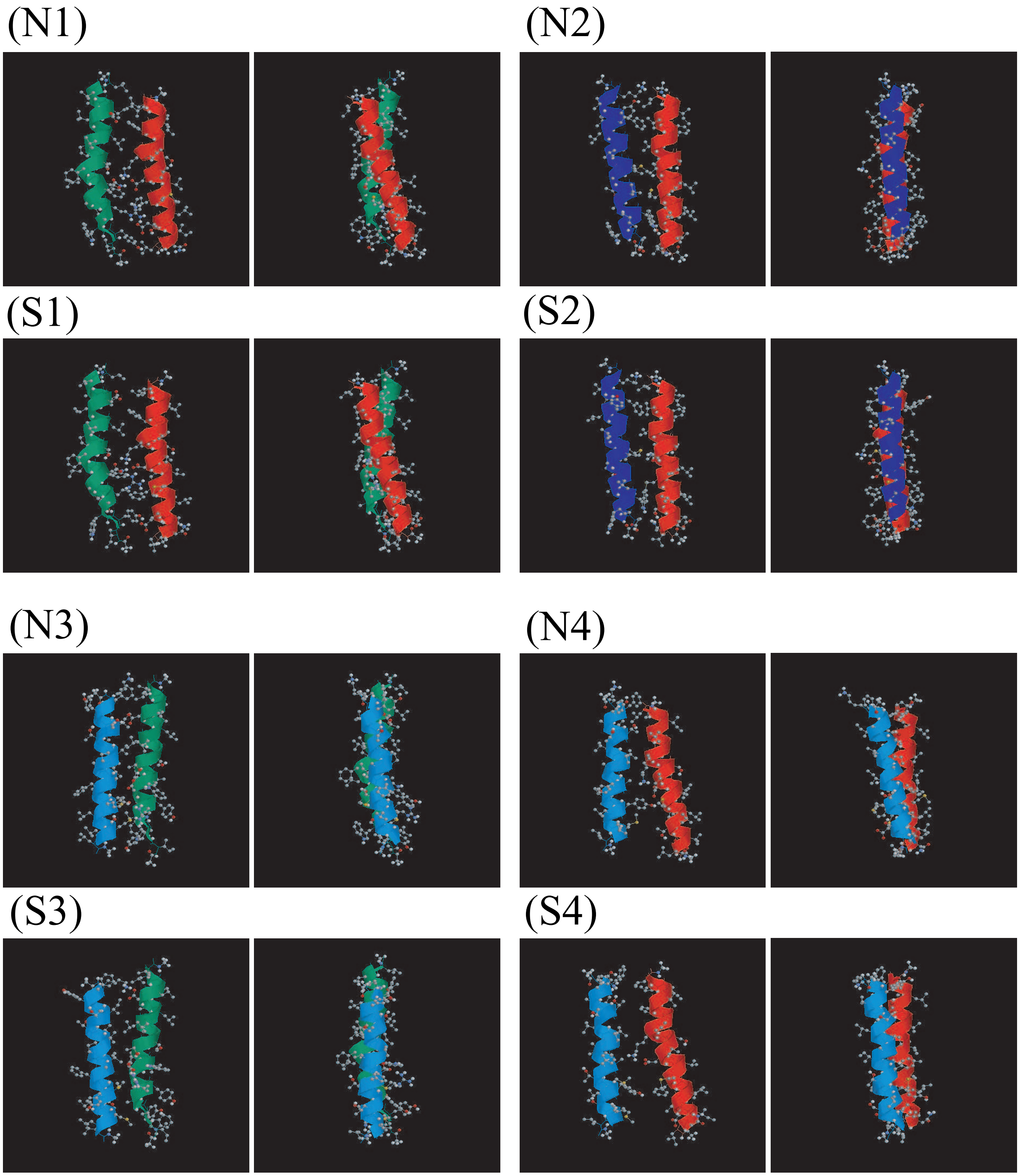

Figure 5.5: The structures of some pairs of helices in Figs. 5.4(a) and (b). Helix colors are the same as in Fig. 5.4. (N1), (N2), (N3), and (N4) were taken from the PDB structure (Fig. 5.4(a)). (S1), (S2), (S3), and (S4) were taken from the smallest RMSD structure (Fig. 5.4(b)) obtained from the REM simulation. (N1) and (S1) are Helices C and G, (N2) and (S2) are Helices A and G, (N3) and (S3) are Helices B and C, and (N4) and (S4) are Helices B and G. For each entry two figures of the same structure viewed from different angles are shown. The figures were created with RasMol [23].

Chapter 6

Conclusions

In this thesis we addressed the issue of predicting membrane protein structures by molecular simulations from the first principles.

In Chapter 2 we proposed a method for predicting helical membrane protein structures by computer simulations. Our method consists of two parts. In the first part, we obtain the amino-acid sequences of the transmembrane helix regions of the target protein from one of existing WWW servers. The precision of these programs in the WWW servers is at present about 85 %, but it is expected to be further improved. In the second part of our method, we perform a replica-exchange simulation of these transmembrane helices with atomistic details to obtain the global-minimum-energy state, which we identify as the predicted structure. Replica-exchange simulations can sample wide configurational space without getting trapped in local-minimum free energy states and we can find stable structures at low temperatures. In order to save computation time, we introduced rather bold approximations in the second part: Backbones are treated as rigid body (only side-chain structures are made flexible) and the rest of the protein such as loop regions and the surrounding lipids and water are neglected. This method was applied in Chapters 3, 4, and 5.

In Chapter 3 we tested our prediction method of membrane protein structures with glycophorin A transmembrane dimer and analyzed the predicted structures in detail. The structure obtained in the case for the dielectric constant $\epsilon = 1.0$ was in close agreement with the NMR experimental data, while that for $\epsilon = 4.0$ was more packed than the native one. The dielectric constant for a lipid system is closer to 4.0 than to 1.0. Our results imply that the helix-helix interaction is the main driving force for the native structure formation. It was found from the analysis of the average physical quantities as functions of temperature that the temperature variation of the electrostatic energy term was much smaller than the van der Waals and torsion energy terms and the contribution of the electrostatic energy to the stability was small on the average. However, only the case with $\epsilon = 1.0$ gave the native structure as the global-minimum-energy structure, although structures close to the native one were also sampled with $\epsilon = 4.0$ as one of the local-minimum-energy states. We saw that the predicted structure with $\epsilon = 4.0$ has smaller solvent accessible surface area than the native one and is more stabilized by van der Waals energy than in the case for $\epsilon = 1.0$. This finding suggests that although only four hydrophilic amino acids are included in 36 amino acids used in our simulations, the electrostatic energy term contributes to the stability of this membrane protein and forces the native structure to be a little less packed than the case with weakened electrostatic interactions ($\epsilon = 4.0$). In other words, the stability of the native structure is determined by the balance of the electrostatic term, van der Waals term, and torsion term, and the contribution of electrostatic energy is indeed important for correct predictions. We believe that the inclusion of atomistic details of side chains is important to estimate this balance accurately because transmembrane helices are usually tightly packed.

In Chapter 4 the effectiveness of our classification and prediction method for transmembrane helix configurations of membrane proteins by replica-exchange simulations was further examined with the glycophorin A transmembrane dimer. We studied in detail the low-energy structures of this system. We classified the obtained structures into clusters of similar structures by the principal component analysis. These clusters are identified as the global-minimum and local-minimum free energy states. Projecting the free energy onto the first two principal axes, we found that although this system has many degrees of freedom, this membrane protein system has rather simple free energy landscape and the sampled structures can be classified by only two principal components.

In the case of the dielectric constant $\epsilon = 1.0$, there exist only two local-minimum free energy states. The global-minimum free energy state in the case of $\epsilon = 1.0$ is very close to the structure of the NMR experiments and the prediction was successful. It turned out that the global-minimum free energy state is also the global-minimum potential energy state in the present system, and hence our method of membrane protein structure predictions by searching the global-minimum potential energy structure is justified.

In the case of $\epsilon = 4.0$ (the value close to that for the lipid environment), on the other hand, there exist five local-minimum free energy states and the native structure does not correspond to the global-minimum state but to a local-minimum one. The two cases differ only in the electrostatic interactions, and this implies that the native structure is determined by a subtle balance of the electrostatic term, van der Waals term, and torsion term in the potential energy function. Most amino acids of transmembrane regions consist of hydrophobic ones, but not only van der Waals interactions but also electrostatic interactions contribute to the structure stability of the native state. In order to have the right prediction, we have to use the appropriate electrostatic term. We interpreted the results that the choice of the dielectric constant $\epsilon = 1.0$ is more appropriate than $\epsilon = 4.0$, because transmembrane helices are tightly packed and there is almost no room for the lipid molecules to exist between helices.

The two states identified in the case of $\epsilon = 1.0$ are found to constitute a subset of the five states in the case of $\epsilon = 4.0$. This means that although the prediction in the case of $\epsilon = 4.0$ was not successful in the sense we did not get the native structure as the global-minimum free energy state, it was rather close to the right answer. The differences in free energy among local-minimum states are indeed small (less than 2.0 kcal/mol). Moreover, the potential energy function that was perfectly suitable in the present system of the glycophorin A dimer may not be as good for other larger membrane systems. We therefore refine our method for predicting transmembrane helix configurations as follows. Instead of identifying the global-minimum potential energy state as the predicted structure, we classify the obtained low-energy configurations into clusters of similar structures by the principal component analysis. We then identify all the local-minimum free energy states and compare them carefully. The native structure should correspond to one of them (the global-minimum free energy state being the most likely candidate). Our predictions can be made more narrowed down with confidence if we have some experimental information such as the distances between some atoms. Note that the detailed analysis in Chapters 4 and 5 were made possible bacause the useful sampling methods such as the replica-exchange method were used.

In Chapter 5 we examined whether the transmembrane helices of bacteriorhodopsin self-assemble into the native configuration by themselves by a replica-exchange MC simulation. Although we need a longer simulation time to confirm the results, one of the replicas sampled the same helix configuration as the native one, starting from random initial configurations. Our results imply that the helix-helix interactions are the dominant forces that determine the transmembrane structure of bacteriorhodopsin. In other words, the interactions between peptides are the most important for the structure formation, and the complicated interactions with hydrocarbon core of lipid bilayer, lipid interface, and water molecules are a secondary effect in the final stage of structure formation of membrane proteins.

We also examined which partial structures resemble the native ones. We found that those of Helices A, B, C, and G, with which the retinal molecule has less contact than the rest of the seven helices, are in remarkable agreement with the native ones including side-chain structures. The agreement was less impressive for the rest of the helices (Helices D, E, and F), because we neglected the retinal molecule in our simulation. However, the fact that we were still able to predict the correct native-like topology of helix configurations suggests that our prediction method can be widely applied to structure predictions of membrane proteins where no experimental information is available.

In the future we have to make our approximations better. For example, we should introduce some flexibility in the helix backbone structures because membrane proteins are not necessarily composed of ideal α-helices. The electrostatic interactions, in which we used the dielectric constant value of 1.0 or 4.0, can also be made more accurate so that some environmental effects including lipids may be taken into account. Our results support the two-stage model for the structure formation of membrane proteins. The applicability of our method to membrane proteins with transmembrane loop regions that interact with transmembrane helices remains to be established.