Protein Structure Prediction Using an Associated Memory Hamiltonian and All-Atom Molecular Dynamics Simulations

Kijeong Kwac[*] and Peter G. Wolynes[†]

Department of Chemistry and Biochemistry, University of Texas at Austin, Austin, TX 78712, U.S.A.
[*]E-mail: kjkwac@hanmail.net
[†]Department of Chemistry and Biochemistry, University of California San Diego, La Jolla, CA 92093-0365, U.S.A.
Received July 28, 2008

We tried to predict the tertiary structure of the 63-residue-long alpha-helical protein, 1r69, from the amino acid sequence with the assumption that the locations of α-helical residues are known. We applied two approaches. One approach is to implement all-atom molecular dynamics (MD) simulations of segments of the target protein and use the snapshot structures of these simulations as memory sets in the associated memory Hamiltonian, which uses coarse-grained model of protein structure and describes the effect of solvent by a water-mediated long range interaction potential. The other approach is to implement all-atom MD simulations with implicit water model applying additional biasing potential functions to reduce the radius of gyration and induce the formation of secondary structures for the helical residues. In the coarse grained model of the associated memory Hamiltonian we tried two different sets of memory to see the effect of the local structural signals in the memory set. We found that the predicted results strongly depend on the structures used in the memory set. The predicted results from the associated memory Hamiltonian give a structure with RMSD value of 1.977 Å with respect to the native structure. The predicted results from the biased all atom MD simulation method give a structure with RMSD value of 2.971 Å.

Key Words : Protein folding, Associated memory hamiltonian, Molecular dynamics, Protein structure prediction

Introduction

Protein folding is a fascinating topic in the computational chemistry and biology. The problem of protein folding can be divided into two categories: (a) dynamics of the folding process and (b) prediction of the three dimensional structures of folded proteins.[1-7] These two topics are not completely separated and some insight in one field can benefit the other. In the field of protein structure prediction,[8,9] there are broadly three ways to approach the problem: homology modeling,[10,11] threading[12,13] and *ab initio* prediction of the structure using physicochemical principles. The former two methods, homology modeling and threading, use the knowledge of previously solved structures,[14] whereas ab initio folding uses the physical principles starting from the amino acid sequence. For the ab initio folding, potential energy function is proposed and global optimization is attempted or various sampling techniques are employed to search the configuration space.[15-26]

The structure of protein molecule has a hierarchy from one-dimensional amino acid sequences to secondary structures to tertiary folded structures.[27] So we can think that the problem of protein structure prediction can be tackled in a stepwise manner considering the structural hierarchy of protein molecules.[28-31] Especially there are many endeavors to predict the secondary structure from the amino acid sequence.[32-24] Also there are various investigations for the prediction of the loop structure for the segments of amino acid sequence as a part of a given protein molecule.[35] But the next step in the structural hierarchy of the protein structure is far more difficult than the previous steps because of the difficulty of modeling the interactions between the residues which are far apart in sequence. There are some recent advances in this direction such as the inclusion of water mediated long range interactions[36,25] into the associated memory (AM) Hamiltonia[22,37] in the description of protein folding and dynamics.

Another issue regarding the structural hierarchy from the secondary structure to the tertiary structure is the influence of the local structural signals in the folding into the 3-dimensional structures.[38-43] There are experimental observations that the segments of the peptide from the protein molecules preserve their structural propensity to the specific structure when they are part of the protein molecule.[44-47] Myers and coworkers[44] measured the helix propensities of the nonpolar amino acids for an α-helix in an intact protein and for a 17-residue peptide with a sequence identical to that of the α-helix in the protein. Their conclusion was that helix propensities can make equivalent energetic contribution in both peptide and proteins. Dyson and coworkers[45] have examined the conformational preferences of peptides derived from a simple four-helix bundle protein, myohemerythrin, in aqueous solution and found that the peptides corresponding to the helices of the folded protein all exhibit conformational preferences for nascent helix. But they also found that the peptide fragments derived from the β-sandwich protein plastocyanin are relatively devoid of secondary structure in aqueous solution.[46] Saven and Wolynes[48] have

implemented a theoretical study on the role that local conformational tendencies can have in guiding the folding of helical proteins using simple statistical models. Their calculation indicated that native turn and start-stop signals can be important in guiding the molecule toward the native structure and the presence of the native stabilization with local conformational signals acts to reduce the effective conformational entropy.

In this work we tried the protein structure prediction in the view point of the ab initio folding method. We assume that we know the exact information on the location of the helical residues and tried to predict the 3-dimensional structure only using that information. We tried two approaches: one is using the coarse-grained model described by the associated memory (AM) Hamiltonian with the recently developed water-mediated interaction potential for the residues in the long range in sequence, which will be denoted as AMW Hamiltonian in this work. The other is using the all-atom AMBER force field[49,50] with implicit solvent model[51,52] with biasing potentials for the dihedral angle of the helical residue and the radius of gyration of the molecule.

The information of the location of the helical residues is the only input in these two methods to predict the final structure of the target protein. In the first method of the AMW Hamiltonian, we need to provide a set of structures as a memory term in the Hamiltonian, which will be used to describe the short and medium range (in sequence) interactions of the residues.[22-24,53,54] In the previous calculations using the AM Hamiltonian, the memory terms are provided by a set of previously known protein structures. In the present work we implemented the all atom MD simulations of segments of target proteins and used the snapshot structures of these MD simulations as the memory set structures in the AM Hamiltonian. In the simulation of the segments of the target protein the information of locations of helical residues is used to set up the initial configuration in which the helical residues have a typical helical conformation. In the second method using all atom MD simulations of the whole molecule, the information of the location of the helical residues is used to bias the dihedral angles of the assigned helical residues to a typical helical conformation in the course of the simulation.

In the description of protein dynamics using the AM Hamiltonian, one of the important factors is how much helpful the structures in the memory terms of the Hamiltonian are in guiding the folding process. We found that the propensity of each residue for the specific secondary structures depends on the location of that residue within the segments in the MD simulation of the segments of the target protein. And the conformation of each residue in the memory set obtained from the MD simulation of the segments can be different from the conformations of the native state. In the present work we prepared two sets of segments to use in the memory term: One set consists of segments sequentially taken from the target protein without overlapping between them. The other set consists of the previous set and additional segments which has partial overlap with the previous segments. In this second set some residues have different conformational preferences as their locations within segments have changed. We have compared the results from these two memory sets and also compared them with the previous method of using a database of known protein structures. The objective of this work is (a) to investigate the influence of local structural signals expressed in the memory terms of the AMW Hamiltonian on the folding dynamics and the final predicted structure and (b) to access the performance of long range potential description in the AMW Hamiltonian scheme by comparing it with the implicit water model in the all atom simulation.

We obtained the improved results of prediction and the free energy profile for the memory term using the first segments set compared to the result using a database of known protein structures. In addition we obtained similarly good results of prediction for the second scheme of the biased all-atom MD simulation. In the next section we give briefly the formulation of the AMW Hamiltonian and the simulation methods. Then we summarize the results of the all-atom MD simulations of the segments of the target protein to get the memory set for using in the AM Hamiltonian, and we analyze the results of the simulated annealing run using the coarse-grained model of the AMW Hamiltonian. Next, we give the results of the biased all-atom MD simulation, which is the second scheme of the present work. In the final section we give concluding remarks.

Methods

A. Associated Memory Energy Function. We used the AMW Hamiltonian to sample structures of the target protein.[22-25,53] It was originally developed by Friedrichs and Wolynes,[22] and is a coarse grained description with only C_α, C_β, and O atoms explicitly represented. The AM energy function consists of a backbone term E_{back} and interaction term E_{int} so that

$$E = E_{back} + E_{\text{int}}. \tag{1}$$

The backbone term, which describes the protein backbone, consists of several terms. For the detailed expression of each term in E_{back}, we refer to the previous publications.[23,24,53]

The non-bonded interactions between residues of the protein are supplied by the interaction term, E_{int}. The interactions described by E_{int} depend on the sequence separation $|i-j|$ between the residues i and j involved. Specifically, they are divided into three proximity classes, $x(|i-j|)$: x = short ($|i-j| < 5$), x = medium ($5 \le |i-j| \le 12$), and x = long ($|i-j| > 12$). Thus,

$$E_{\text{int}} = E_{short} + E_{med} + E_{long}. \tag{2}$$

The short- and medium-range interactions are treated by an associated memory energy functions:

$$E_{AM} = E_{short} + E_{med}$$

$$= -\frac{\varepsilon}{a}\sum_{\mu=1}^{N_{mem}}\sum_{j-12\le i\le j-3}\left\{\gamma[P_i, P_j, P_{i'}^{\mu}, P_{j'}^{\mu}, x(|i-j|)]\exp\left[\frac{(r_{ij}-r_{ij}^{\mu})^2}{2\sigma_{ij}^2}\right]\right\} \tag{3}$$

The unit of energy denoted by ε in Eq. (3) is defined as the native state interaction energy per contact:

$$\varepsilon = \frac{\left|E_{\text{int}}^{Native}\right|}{4N} \tag{4}$$

where N is the number of residues of the protein being considered. a is a dimensionless constant chosen so that Eq. (4) is satisfied. The sum over i and j runs over all possible pairs of C_α and C_β atoms ($C_\alpha - C_\alpha$, $C_\alpha - C_\beta$, $C_\beta - C_\alpha$, $C_\beta - C_\beta$) with sequence separation between 3 and 12, and r_{ij} is the distance between atoms i and j. The index μ runs over all N_{mem} memory proteins to which the sequence of the target protein has previously been aligned so that, for a given $i-j$ pair, there is a specific $i'-j'$ pair in some μth memory protein. The letter P_i represents the identity of the ith residue in the reduced four-letter (as opposed to 20-letter) code: hydrophilic (ala, gly, pro, ser, thr), hydrophobic (cys, ile, leu, met, phe, trp, tyr), acidic (asn, asp, gln, glu), and basic (arg, his, lys). Each term in the summation of Eq. (3) is a Gaussian well centered at the separation $r_{i'j'}^{\mu}$ of the corresponding memory atoms. The widths of the Gaussians are dependent upon the sequence separation $|i-j|$ such that $\sigma_{ij} = |i-j|^{0.15}$. The relative weights of these Gaussian wells are controlled by γ's, which depend on the identity of the residues matched between the target and memory proteins and the sequence separation $|i-j|$. In this work we used the γ values optimized using a set of α/β proteins in the previous work of Ref. 25.

In the previous studies using the AM energy function, this association between $i-j$ pair and $i'-j'$ pair of the memory has been done using a sequence-structure threading algorithm.[54] In the present work we use the snapshot structures taken from MD simulations of segments of the target protein as the set of memory proteins in the AMW Hamiltonian. An $i-j$ pair of the target protein is associated with the same $i-j$ pair in the memory protein which is a segment of the target protein. Thus, the Eq. (3) is effectively written as

$$E_{AM} = -\frac{\varepsilon}{a}\sum_{\mu=1}^{N_{mem}}\sum_{j-12\le i\le j-3}\left\{\gamma[P_i, P_j, x(|i-j|)]\exp\left[-\frac{(r_{ij}-r_{i'j'}^{\mu})^2}{2\sigma_{ij}^2}\right]\right\}, \tag{5}$$

where $\gamma[P_i, P_j, x(|i-j|)] \equiv \gamma[P_i, P_j, P_i, P_j, x(|i-j|)]$. Since P_i and P_j can have four different values and $x(|i-j|)$ is short or medium, there are effectively $4 \times 4 \times 2 = 32$ different γ parameters working in the associated memory terms.

The long-range interaction part E_{long} is not related to the memory proteins. This term is constructed to model a physically motivated, non-pairwise-additive model of water-mediated interactions.[36] E_{long} can be partitioned into three terms:

$$E_{long} = E_{contact} + E_{water} + E_{burial}. \tag{6}$$

$E_{contact}$ describes a potential well located between 4.5 Å and 6.5 Å to represent a direct contact between two residues. E_{water} gives a second well located between 6.5 Å and 9.5 Å, representing protein-mediated or water-mediated interactions. The switch between protein- and water-mediated potential is done by calculating the local density around each pair of residues. For detailed expressions for protein- and water-mediated potentials, we refer to ref. 36. The burial profile term, E_{burial}, is a many-body local density-based three well potential that describes amino acid preferences for a particular coordinate density.[25]

We carried out coarse-grained molecular dynamics simulations using AMW energy function with temperature quenching to search for low energy conformations. Temperature is reduced linearly from $\tilde{T} = 1.8$ to $\tilde{T} = 0.0$ in 720000 time steps, where $\tilde{T} = k_B T/\varepsilon$.

B. All-Atom MD Simulation of Segments of Protein. We implemented all-atom MD simulations for segments of the target protein to sample snapshot structures to be used for constructing a set of memory structures in the associated memory energy function. We have divided the 63-residue-long target protein 1r69 into three segments which are 21-residue long: segment A (residues 1-21), segment B (residues 22-42), segment C (residues 43-63). In addition, we take two segments which are overlapping with the previous segments: segment D (residues 11-31) and segment E (residues 32-52). The N-terminal and C-terminal of each segment are blocked by acetyl (ACE) and N-methyl (NME) group. We determined the location of the α-helical residues using the STRIDE program.[55] The residues 2-13, 17-24, 28-36, 44-52, and 56-61 are α-helical residues. Figure 1(a) shows the segment D starting from 11th residue as an example. For this segment the residues 11-13, 17-24, and 28-31 are α-helical residues. We only use the information of the locations of the alpha helical residues and do not use detailed values of dihedral angles. Thus we assign $\phi = 57°$ and $\psi = 47°$ for the alpha helical residues and $\phi = 180°$ and $\psi = 180°$ for the rest residues for the initial conformation of the MD simulation as shown in Figure 1(b). We modify the dihedral angles of non-helical residues slightly in the initial configurations if there is any steric clash between the side chain atoms.

We used the AMBER9 program package[56] to implement the simulation using ff03 force field[50] to describe the protein segments and implicit solvent model to describe water. We implemented MD simulation for the five segments at 300 K using the weak coupling algorithm of Bredensen *et al.*[57] Time step was 1 fs. For each segment, we implemented MD run for 24ns. For each MD trajectory, we take 60 snapshot structures with 400ps time interval between neighboring two snapshots. These snapshot structures will be used as memory structures in the associated memory energy function.

C. Biased All-Atom MD Simulation. In addition to the MD simulation of the segments of the target protein, we also implemented MD simulations of the whole protein molecule with biasing potentials with respect to the secondary structure and radius of gyration. The simulation protocols are similar to the segment MD simulation. We used the langevin dynamics for the temperature coupling. We have modified the Amber program package[56] to introduce two additional biases: secondary structure bias and the radius of gyration bias. For the secondary structure bias, we impose the follow-

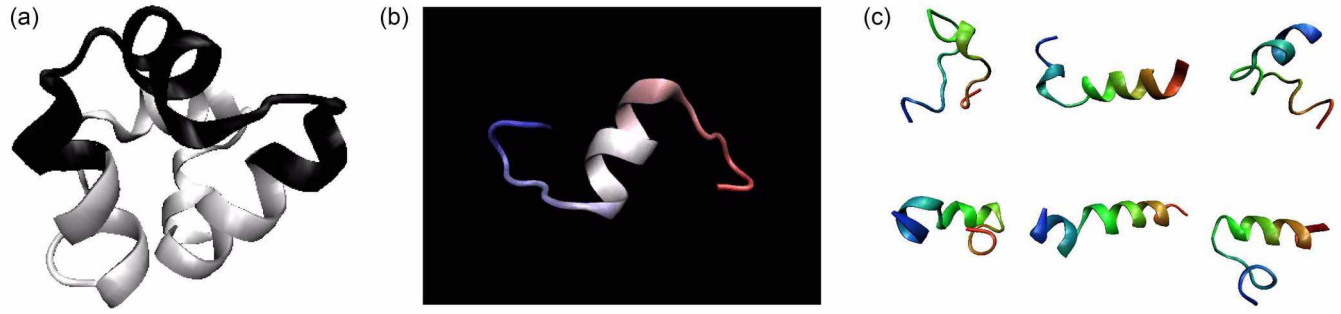

Figure 1. A segment starting from 11th residue of the target protein. (a) Location of the segment (black) in the protein 1r69 (b) The initial configuration for the segment MD simulation. (c) Snapshots sampled from the MD simulation trajectory for the segment.

ing potential function of ϕ and ψ angles for the helical residues:

$$V(x) = \frac{1}{2}A \cdot \sin^2\left(\frac{x - x_0}{2}\right), \quad (7)$$

where x is ϕ or ψ. We select $A = 10$ kcal/mol and $x_0 = 57°$ for ϕ and 47° for ψ. We impose this biasing potential only on the helical residues. With these parameter values, the helical residues under this potential have formed the helical structures in the early phase of the simulation. The formation of α helical structures is completed within 100 ps from the start of the simulation. For the biasing potential for the radius of gyration, we use an one-sided harmonic potential: V_{Rg} = (10.0 kcal/mol/A^2)$(R_g(t) - R_g^0(t))^2$. We change the location of the minimum of this one-sided harmonic potential linearly with respect to time from 5 $R_g^{predict}$ to 0.95 $R_g^{predict}$ where $R_g^{predict} = 2.2N^{0.38}$. The length of the simulation is 2ns. From the analysis of the MD trajectory we found that, before the value of R_g^0 becomes about 3.5 $R_g^{predict}$, the protein molecule does not feel the biasing potential of R_g. But after R_g^0 is reduced to values less than about 3.5 $R_g^{predict}$, the protein molecule feels the biasing force to reduce its radius of gyration and is forced to fold. This folding is different from the usual protein folding in that the helical residues are already almost formed. Therefore the interaction between the already formed secondary structures is an important factor in this simulation.

We tried two versions of the implicit water model of Generalized Born implemented in the AMBER program packages.[51,52] The two different model can be selected by setting the igb = 5 or igb = 7 in the input file of the AMBER program, so we denote the two versions of the implicit solvent model with igb5 and igb7. In our biased simulation, we treat the dihedral angle motion of non-helical residues by the amber ff03 force field.[50] We also tried a modified version of ff03 force field, in which the potential parameters for the backbone ϕ and ψ are set to be zero for non-helical residues.[58]

Results and Discussion

A. Segment All-Atom MD Simulations. As an example, Figure 1(c) represents several snapshot structures of the segment D starting from 11th residues during the all-atom MD simulations, which are 0.4, 4.4, 12.4, 16.4, 20.4 ns configurations obtained from the 24 ns MD trajectory. Fig. S1 (supporting information) shows the distribution of ϕ and ψ angles obtained from the MD trajectory along with the values of the dihedral angles of the native structure. Figure S2 (supporting information) shows the difference between the native angle and the most probable value of the dihedral angle distribution.

(a) Helical residues in the ABC segment set: Forty five residues out of 63 residues are helical residues. In the distributions of dihedral angles for these residues, except for two cases of residue 6 and 24, most populated conformations are similar to the conformations of the native state. The distributions of conformation of the residue 6 and 24 sample both helical and extended regions but the major peaks are located in the P_{II} and β region, respectively.[59]

(b) Non-helical residues in the ABC segment set: The remaining 19 non-helical residues correspond to the turn and coil regions in the native structure. For eleven residues (residue 1, 14, 16, 37, 38, 40, 42, 43, 53, 54), the most populated region is near the native state conformation. Three residues (residue 15, 26, 27) have the distribution whose major peak is at the α_R conformation while the corresponding native state structures are the extended β or P_{II} conformations. There are five GLY residues which are α_L conformation in the native structure. Among these, three residues (residue 14, 37, 53) have conformations similar to the native state. Two GLY residues (residue 25, 62) have α_R conformation.

(c) Helical redisues in the DE segment set: The distributions for the most helical residues are similar to the case of segments A, B, and C. We can see that there is a tendency that the residues which are near the end of a segment have broader distributions of dihedral angle conformation compared to the case when the same residue is located in the middle of a segment. The five residues (19, 20, 21, 22, 23) have small contributions from the extended conformation in the ABC segment set, but the distributions of these residues are nearly exclusively α_R region in the DE segment set. The four residues (31, 32, 51, 52) have mainly the α_R conformation in the ABC segment set while the extended conformations are also sampled in the DE segment set. We observe drastic changes in the dihedral angle distributions for two residues (residue 36, 44) between the ABC segment set and

the DE segment set. The most populated conformation of these two residues is α_R conformation in the ABC segment set and P_{II} conformation in the DE segment set. Residue 24 is α_R conformation in the native state, but the population of α_R structure is significantly reduced in the DE segment set.

(d) Non-helical residues in the DE segment set: Most of these non-helical residues have significantly different distributions of conformations between the ABC segment set and the DE segment set. The Ramachandran plots for the distributions of dihedral angles of several non-helical residues in the MD simulation of both segment sets are shown in Figure S3 (supporting information). Among the 13 non-helical residues, five residues (14, 15, 16, 25, 37) are less native-like conformation in the DE segment set than in the ABC segment set. Three residues (26, 39, 40) have more native-like conformations in the DE segment set.

The number of residues whose dihedral angle distribution contain both helical and non-helical regions increases from 28 residues in the three segments case (ABC) to 37 residues in the five segments case (ABC+DE).

B. Protein Structure Prediction with AMW Hamiltonian. We have implemented the annealing MD runs with AMW Hamiltonian whose memory set consisting of the snapshot structures of segments described in the previous subsection. Firstly, we tried two different initial conditions: a collapsed structure and randomly generated extended structures. For the case of the collapsed initial structure, we used the same initial configuration for 20 annealing runs with different random seed to initiate the dynamics. For the case of extended initial configurations, our program for the annealing run generates 20 different initial structures and different random seed to start the dynamics. Secondly, we tried two different sets of memory proteins for the AMW Hamiltonian. The one is the set of memory which consists of snapshot structures taken from MD trajectories of the three segments (ABC) and the other is the set of memory which consists of structures taken from MD trajectories of five segments (ABC+DE).

We use the Q values to describe the resemblance of the predicted structure with the native structure, defined as[25,53]

$$Q = \frac{2}{(N-1)(N-2)} \sum_{i<j-1} \exp[-(r_{ij} - r_{ij}^{N})^2 / 2\sigma_{ij}^2], \quad (8)$$

which measures the resemblance between the predicted and native structure by calculating the average of the Gaussian weights as functions of the difference in the pair distances between the two structures. The structure with $Q = 1$ corresponds to the native state and $Q = 0$ means totally unfolded state. Usually the conformations having Q values near 0.6 correspond to the RMSD value near 2 Å. Figure 2 shows the Q values and RMSD values for the final structures obtained by the annealing MD runs with the AMW Hamiltonian starting from the random extended initial structure. Although data is not shown here, the calculations starting with a collapsed structure as initial conditions gave similar results. Thus the dependence on the starting structure of the present

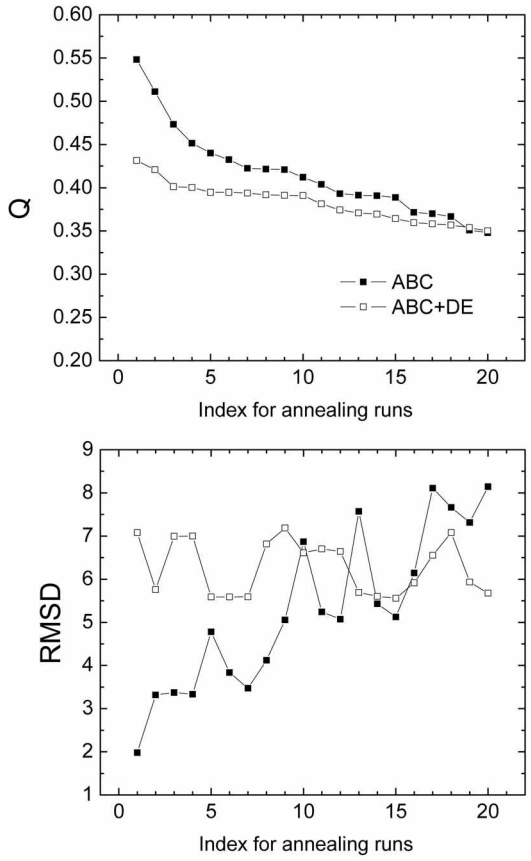

Figure 2. Plot of Q values and RMSD for the final structures resulting from the simulated annealing runs with the random extended initial conformation. The indexes in the x-axis are sorted with respect to the descending Q values.

AMW annealing runs is not conspicuous. When we analyze the trajectory of the annealing runs, we found that the initial extended structures with the radius of gyration, R_g = 25-30 Å collapse to structures with R_g = 11-12 Å within 3000 steps of the total 720000 steps. This fast collapse in the annealing run is the main reason of the weak dependence on the starting configuration.

A notable feature in Figure 2 is that the memory from the MD trajectories of ABC segment set gives higher Q values than the case of the ABC+DE segment set. As observed in the Figure S3 (supporting information), the probability for the helical residues to sample extended configuration is increasing in the case of using ABC+DE set of segments.

The best Q value among the predicted structures is 0.5482 and its RMSD is 1.977 Å. This structure is obtained from the AMW MD run from the ABC segment set and the random extended initial structure. We show the best Q value structure in the Figures 3(a).

Figure S4 (supporting information) shows the best Q values for the whole trajectories of each annealing runs. We saved 240 structures equally spaced in annealing step for each MD runs and selected the structure with best Q values for each annealing trajectory. We observed that there exists a correlation between the best Q value of the trajectory and the Q values of the final structure of the same trajectory. The

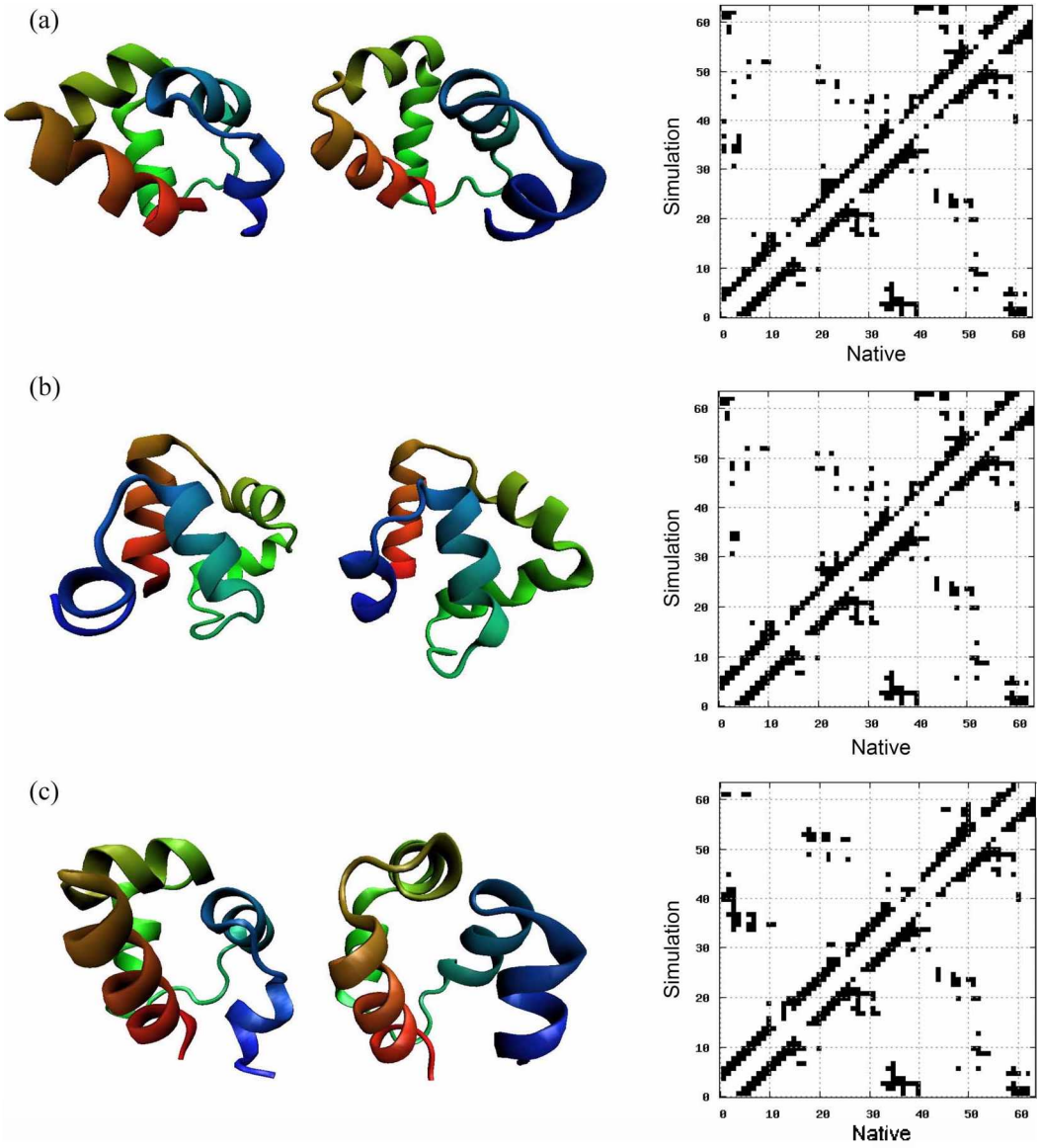

Figure 3. Ordered as the native (left), the predicted structure (middle), and the contact map (right). (a) Predicted structures with highest Q values among the final structures of the annealing runs with ABC memory set and starting from random extended configurations. Q = 0.5482, RMSD = 1.9770 Å (b) The structure with highest Q values in the entire trajectory, Q = 0.5629, RMSD = 1.916 Å. (c) The structure with highest Q values among the results of the biased all-atom MD simulations, Q = 0.515, RMSD = 2.971 Å.

best Q structure for the whole trajectory and the best Q structure among the final predicted structure are discovered in the same trajectory. The best value of Q for the whole trajectory is 0.5629 and its RMSD from the native structure is 1.916 Å. We show the structure in the Figure 3(b), which is the snapshot structure of the 642000th annealing step in the trajectory.

C. Analysis of the Folding Trajectories with the AMW Hamiltonian. To compare the effect of the different memory sets used in the AMW Hamiltonian, we calculated the Q values for the parts of the protein molecule. The target protein 1r69 has five helices. We denote the five helices as H1(2-13), H2(17-24), H3(28-36), H4(44-52) and H5(56-61) and the four non-helical regions between them as C1(14-16), C2(25-27), C3(37-43), and C4(53-55) (residue numbers in the parenthesis). Figure 4 shows the Q values calculated for each helical and non-helical regions of the protein molecule using the final structures of the simulated annealing runs. Figure 4(a) and 4(b) show the results of the ABC memory set and ABC+DE memory set, respectively. For comparison we also implemented 20 simulated annealing runs using the memory set consisting of 36 known protein structures given in the Ref. 53. The result using this database memory set is given in Figure 4(c).

From the plot of Figure 4(a) and 4(b), we can see that there is no difference in the Q values for the parts from C2 to H5 between the two employed memory sets. Also in both cases Q values for the C2 segment are around 0.2, meaning that the predicted structure of the C2 region is very different from the native structure. As shown in Figure S3 (supporting

information), the dihedral angle conformations for these residues are quite different from the native state in both the memory sets and these have direct effects on the final predicted structures. Among the five helices the H2 helix shows

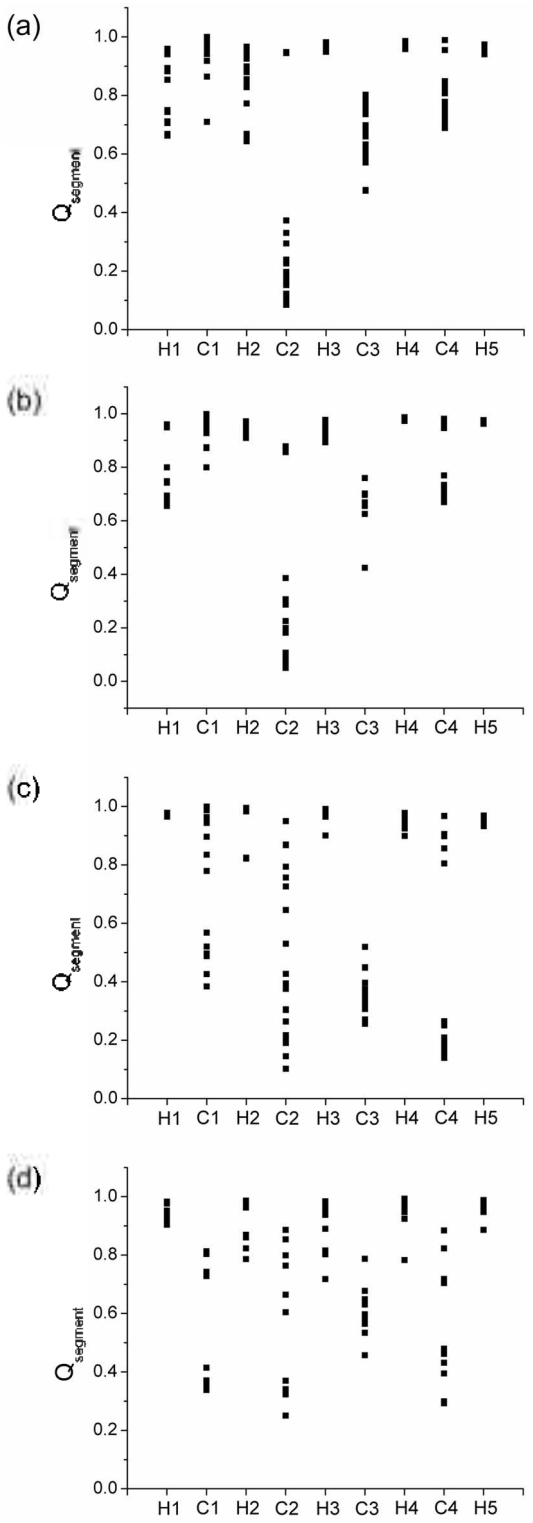

Figure 4. Plot of Q values for the helix and non-helix region of the target protein for the final structures of annealing runs. (a) ABC memory set. (b) ABC+DE memory set (c) knowledge-based memory consisted of 36 proteins. (d) All-atom biased MD simulation.

difference between the two memory sets as shown in Figure 4(a) and 4(b). One notable feature in the Figures 4(a) and 4(b) is that the distributions of the Q values are not broad even for the non-helical regions. This fact is contrasted with the situation using the database memory shown in Figure 4(c), where the non-helical regions of C1-4 have very broad distributions of Q values. We can see that the final structures are quite strongly influenced by the structures in the memory set in the present prediction scheme.

Figure S5 (supporting information) shows the plot of the Q values for the helical and non-helical regions for the two trajectories of the simulated annealing runs, each using the different memory sets and giving the second best Q values for the given memory sets, as an example. We can see that the helices H5, H4 and H3 form in the earlier phase of the trajectory even though the intervening non-helical regions, C4, C3, C2, are not well formed to the native structure. And in this earlier phase of the trajectory there is little difference between the two memory sets. The helix H2 forms in the later phase of the trajectory. From this point the trajectories become different between the two memory sets. We observed these features for more than half of the trajectories in the total AMW MD runs. The trajectory of H1 shows that this helix forms in the later phase of the trajectory. By inspecting

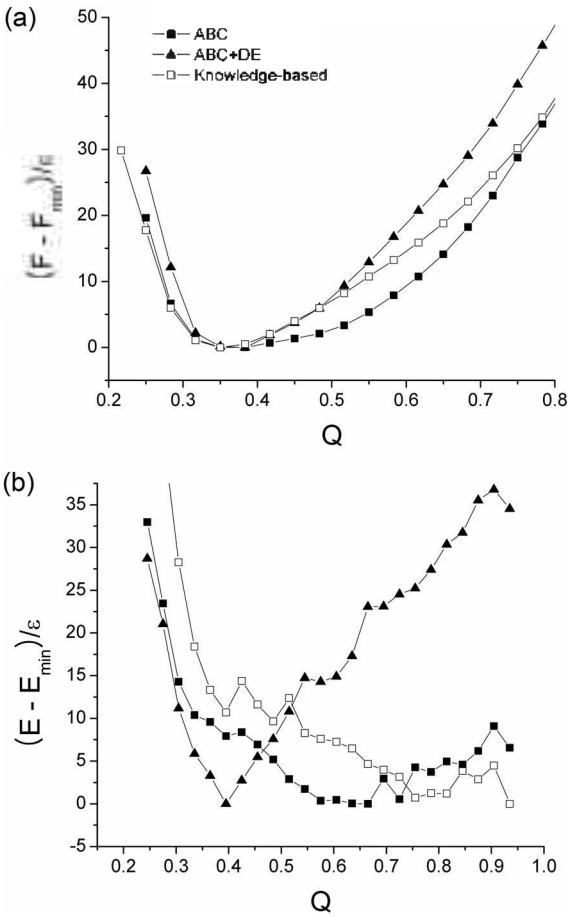

Figure 5. (a) Plot of free energy calculated from the umbrella sampling runs with respect to the Q values. (b) Plot of average potential energy calculated from the structures of the umbrella sampling runs.

all trajectories we found that there exists trajectories which fails to form the complete H1 helix, thus giving the relatively broad distribution in the final structures as shown in the Figures 4(a) and (b).

D. Free Energy Profiles for the *Q* coordinates. We calculated the free energy profile with respect to the Q coordinate using the umbrella sampling with the constraining potential of the form, $V_i(Q) = 5000\varepsilon(Q-Q_i)^4$, where we select the location of Q_i as equally spaced 21 points from 0 to 1 with 0.05 interval. In each Q window we implemented 720000 steps of AMW MD runs at T = 1.0 and get 240 snapshot structures at 3000 step interval. Figure 5 shows the free energy profile and the average potential energy with respect to the Q values. We can see that the difference between the memory sets in the AMW Hamiltonian gives large difference in free energy and potential energy profile. We also plotted the free energy for the case of memory set using the database of 36 known protein structures. In comparison with the database memory set, the memory set from the ABC segments gives much improved result and the memory set from the ABC+DE segments gives worse result, as shown in Figure 5(a). Figure 5(b) shows the average potential energy plot as a function of the Q values. The ABC+DE memory set gives very unfavorable values for the higher Q values and its minimum is near $Q = 0.4$. The plot for the database memory set nearly monotonically decreasing as the Q value increases and gives its lowest values for the Q values greater than 0.9, which is quite desirable feature. The ABC memory set, which gives most favorable free energy profile among the three memory sets, also gives decreasing behavior as Q value increases, but the energy slowly increases as the Q values are greater than 0.7. Considering the fact that the Q values greater than 0.5 give the structures with the RMSD value of near 2 Å in the practical calculation of the structure prediction, the energy plot of the ABC memory set gives most smoothly funneled feature in the range of Q values from 0.2 to 0.6 among the three memory sets and this might be one reason why this memory set gives most successful prediction result.

E. Effect of collapse and local structural signals. The prediction results shown in the Figures 2-3 are from the trajectories in which the protein molecule takes a collapsed shape in the early stages of annealing runs. To investigate the correlation of the collapse and the local structure signals from the memory terms in the AMW Hamiltonian, we have performed the annealing runs in which we use the modified form of R_g bias potential,

$$V_{R_g} = 10\varepsilon(R_g(t)-R_g^0(t))^2 \text{ for } R_g^0(t) \le R_g(t) \le 2.5R_g^0(t), \quad (9)$$

$$= 0 \text{ otherwise,}$$

where the potential minimum $R_g^0(t)$ changes linearly from $R_g^{\max}$ to $R_g^{\min}$, as illustrated in Figure S6 (supporting information). In the present calculations $R_g^{\min} = 2.2N^{0.38}$ and $R_g^{\max} = 4R_g^{\min}$, where N is the number of residues. The conventional AMW run corresponds to the case of $R_g^0(t) = R_g^{\min}$ for all t. We modified the biasing potential such that

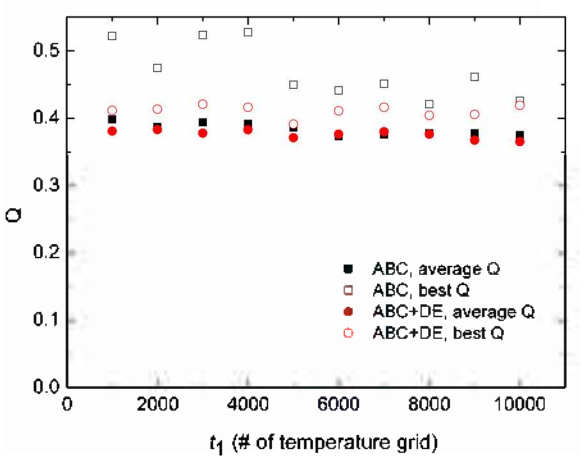

Figure 6. Average and highest Q values for the each 20 simulated annealing MD simulations with respect to t_1, which denote when the R_g bias potential is in full effect in the simulation.

$R_g^0(t)$ decreased linearly from $R_g^{\max}$ at $t = 0$ to $R_g^{\min}$ at $t = t_1$ and $R_g^0(t) = R_g^{\min}$ for $t \ge t_1$. The AMW MD run consists of 12000 grids of temperatures from 1.8 to 0.0 and each grid consists of 60 annealing steps. We tried different values of parameter t_1 from 1000th grid to 10000th grid and for each tried value of t_1 we implemented twenty annealing MD runs with randomly selected initial configurations of extended structures. We plotted in Figure 6 the average of the Q values of the final low temperature structures from the resulting trajectories along with the highest value of the Q among the final structures of the 20 runs. The average values of Q do not show any noticeable difference between the runs with different t_1 values. But the best values of Q shows that there is an advantage in the runs with the smaller values of t_1. In Figure 7 we plot the Q, RMSD and R_g for the trajectory of the best Q value structure when $t_1 = 1000$. With this value of t_1, the protein molecule collapsed in the early stage of the annealing run and search for the lower energy configuration within the collapsed state. Figures 7(a) and (c) show that the configuration right after the collapse has the Q value less than 3.5 but the molecule has enough energy to overcome the energy barriers between local minima and can do the search for the more stable configuration so that the molecule finally found the configuration with the Q value of 0.519 in the end of the annealing run. Figure 6 shows that the mechanism of folding with early collapse has slight advantage compared to the mechanism of searching broad R_g range and approximately $t_1 = 5000$ is the boundary dividing the dominant mechanisms of folding. This t_1 value corresponds to the temperature $T = 1.1$. If the molecule arrives at a collapsed state with incorrect packing when the temperature is below this temperature, it seems that it is not easy to correct the packing configuration of the collapsed state. Another noticeable feature in Figure 6 is difference between the results with the two different memory sets. The fact that the ABC+DE memory set have more diverse secondary structures seems to give energy landscape which is weakly biased to the stable native state compared to the case with the ABC memory set. In other words, the Hamiltonian with ABC memory set is

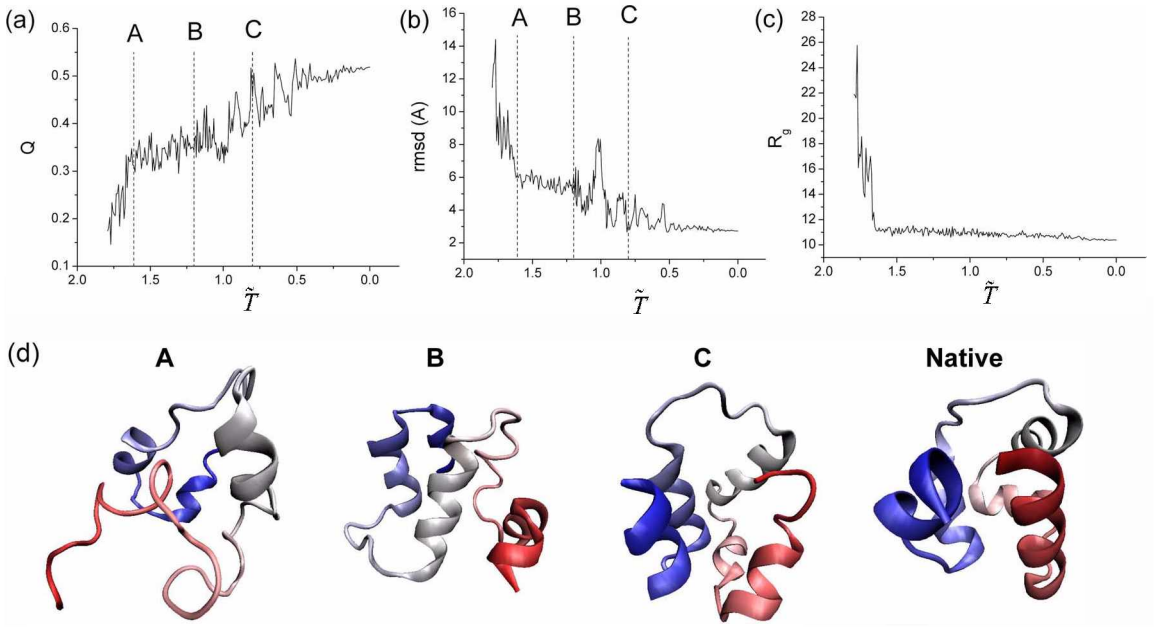

Figure 7. plots of (a) Q values, (b) RMSD, and (c) radius of gyration along the annealing trajectory with n_t = 1000 temperature grid (= 60000 timestep). (d) Snapshot structures corresponding to the vertical lines in (a) and (b). **A**: $\tilde{T}$ = 1.6126 (= 75000th time step), **B**: $\tilde{T}$ = 1.2002 (= 240000th time step), **C**: $\tilde{T}$ = 0.8001 (= 480000th time step). Color is from red to blue in the sequence order.

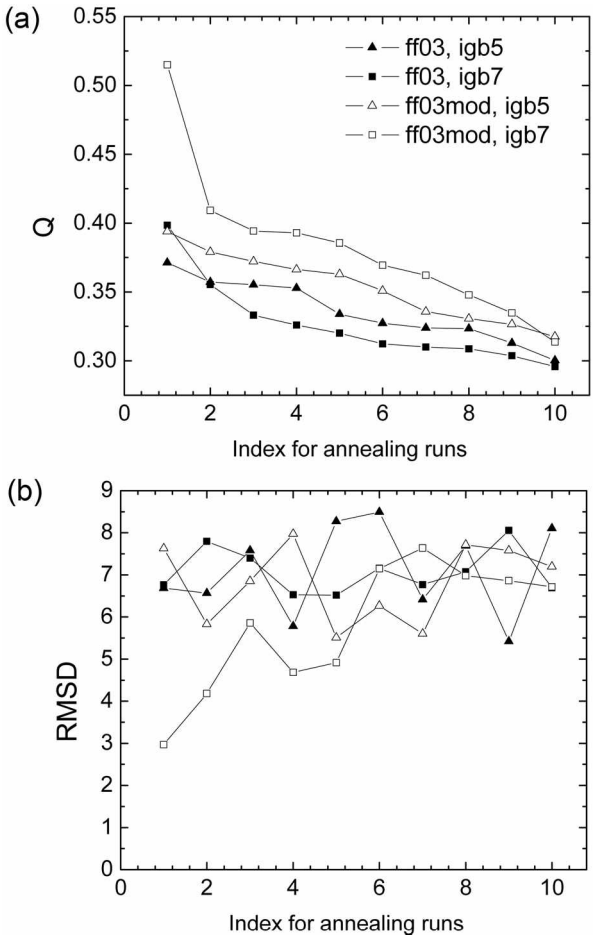

Figure 8. Plots of Q values and RMSD for the final structures from the biased all-atom MD simulations at 300 K. The indexes in the x-axis are sorted with respect to the descending Q values.

more capable of finding stable structures by rearrangements in the collapsed state.

Figure 7(d) shows three snapshot structures obtained from the annealing MD simulation trajectory shown as the vertical lines in the Figure 7(a). Structure A shows the structure right after the collapse and structure B shows the structure when $\tilde{T}$ = 1.2002 which corresponds to 240000th annealing step. In these two structures the packing of the helices are quite different from that of the native state shown in Figure 7(d). If we number the helices in the sequence order, then the packing of the last three helices are in reversed order with respect to the native structure from the view point of the first helix. Structure C of Figure 7(d) shows the snapshot structure when $\tilde{T}$ = 0.8001 which corresponds to 480000th annealing step. The packing of the helices in the structure C is different from that of structure B and is more similar to that of the native structure. The Q values for the structures A, B and C are 0.3124, 0.3431 and 0.4568, respectively. This trajectory gives the final predicted structure with the Q value of 0.5190 and RMSD of 2.725 A.

F. Biased All-Atom MD in Implicit Solvent. We implemented all-atom MD simulation of the whole molecule with additional biasing potential functions to force folding of the target protein. In the calculation with AMW Hamiltonian, we used the bias potential terms such as the secondary structure bias and radius of gyration bias. Here we apply similar kind of biasing potentials to all-atom MD simulations of the whole molecule with the implicit solvent model. Thus this calculation can give a chance to compare between the long range interaction potential in the AMW Hamiltonian and the implicit solvent model in the all-atom MD in the perspective of protein structure prediction. We implemented 10 runs for each combination of simulation conditions at constant

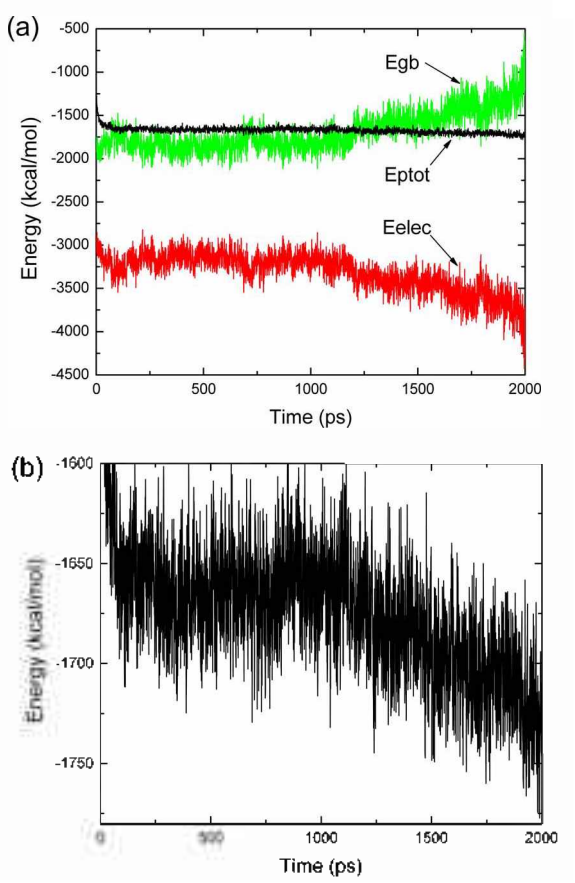

Figure 9. (a) Time dependence of the total energy, the electrostatic energy and the solvation energy. (b) Time dependence of the total energy shown in (a) with more narrow y-axis range.

temperature of 300 K and show the Q values and RMSD values in the Figure 8. The Q values and RMSD are slightly poor compared to the results of the MD runs with the AMW Hamiltonian. But we should take into account that the final structures corresponds to 300 K not 0 K in this calculation. We get the best predicted structure from the trajectory with the implicit solvent model of igb7 and the modified ff03 force field. The Q value of that structure is 0.515 and its RMSD from the native structure is 2.971 Å. Figure 3(c) shows the structure of the best prediction along with the native structure.

Figure 9 shows the time dependence of the energy terms along the MD simulation trajectory which gives the final structure shown in Figure 3(c). In this plot, the contributions from the biasing potentials are not included. In Figure 9(a) we plot the time dependence of the total energy, electrostatic energy, and the solvation energy described by the implicit solvent model. As the protein collapses due to the R_g bias potential, the electrostatic interaction energy becomes more and more favorable. But the solvation energy term behaves to counteract the favorable electrostatic interaction. The roles of these two energy terms are in the opposite direction as shown in the Figure 9(a). But the cancellation between the solvation and the electrostatic interaction among the various part of the protein molecule is not perfect, so the resulting total energy is stabilized as the protein folds as shown in Figure 9(b). We are not sure how much of the favorable electrostatic interactions are cancelled out by the solvation energy in the real situation. In general, the implicit solvent model tends to exaggerate the contribution of the solvation energy within the interior of the protein molecule. Thus we expect that the cancellation of the electrostatic interaction by the solvation energy will be less if the solvation model become more realistic, thus making the drop of the total energy shown in Figure 9(b) more steep, that is, making the energy landscape more funneled.

Concluding Remark

In this work we tried to predict the protein tertiary structure for the selected target protein 1r69 from the amino acid sequence under the assumption that we only know the location of the helical residues without using any knowledge database.

We have implemented the simulated annealing MD simulation with the coarse grained model of AMW Hamiltonian by employing the memory sets which are obtained from the all-atom MD simulations of segments of the target protein. In the all-atom MD simulations of the segments of the proteins we found that dihedral angle conformations of some residues depend on the location within the segment. We prepared two kinds of memory sets, one of which consists of non-overlapping three segments and the other of which consists of overlapping five segments. These two memory sets are different in the magnitude and nature of the local signal especially for the loop and turn regions of the protein molecule. The results of the simulations with AMW Hamiltonian show that the Q values of predicted final structures of the AMW MD simulations strongly depend on the memory terms. The analysis of the individual annealing trajectories shows that the AMW Hamiltonian can describe the process of rearrangement of the packing of helices in the collapsed conformation and in this process the role of local signals from the memory set of the AMW energy function is in effect determining the quality of the final predicted structure.

In addition we also implemented the all-atom implicit-solvent MD simulation of the protein molecule with similar constraints as in the AMW Hamiltonian such as the bias potentials for the helical residues and radius of gyration. The process of folding in the all atom simulation is different from the AMW simulations in that the helical structures are made to form before collapse by the biasing potential. This fact could be one reason for the lower average Q values in the all atom simulations. It is very well likely that the speed of biasing potential seems too fast to properly sample the conformational space, so that it can be easily trapped in local minima before getting to the native state. We think one needs much longer simulation to observe the rearrangement of the packing of partially formed secondary structure in the all-atom R_g-annealing simulations.

One lesson from this study using the AMW Hamiltonian

is that, for the non-helical residues, memory set with only partially correct structures is better than the mixture of correct and wrong structures for the same residue in the preparation of the memory set to be used in the AMW Hamiltonian, and this is related to the problem of finding the optimal size of the memory set.

Acknowledgments. Kijeong Kwac thanks the Post-doctoral Fellowship Program (KRF-2005-214-C00207) of Korea Research Foundation by the Korean Government. He also thanks Chenghang Zong for useful discussions.

Supporting Information Available. Additional figures mentioned in the text (Figs. S1-S6) are available on request from the correspondence author.

References

1. Fersht, A. R. *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*; W. H. Freeman and Co.: New York, 1999.
2. Oleververg, M.; Wolynes, P. G. *Quarterly Rev. Biophys.* **2005**, *38*, 405.
3. Munoz, V. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 395.
4. Plotkin, S. S.; Onuchic, J. N. *Quarterly Rev. Biophys.* **2002**, *35*, 111, *ibid.* **2002**, *35*, 205.
5. Shea, J.-E.; Brooks III, C. L. *Annu. Rev. Phys. Chem.* **2001**, *52*, 499.
6. Scheraga, H. A. *Biopolymers* **2007**, *89*, 479.
7. Wolynes, P. G. *Phil. Trans. R. Soc. A* **2005**, *363*, 453.
8. Petrey, D.; Honig, B. *Mol. Cell* **2005**, *20*, 811.
9. Floudas, C. A. *Biotech. Bioeng.* **2007**, *97*, 207.
10. Sali, A.; Blundell, T. L. *J. Mol. Biol.* **1993**, *234*, 779.
11. Fiser, A.; Do, R. K. G.; Sali, A. *Prot. Sci.* **2000**, *9*, 1753.
12. Jones, D. T. *J. Mol. Biol.* **1999**, *287*, 797.
13. Skolnick, J.; Kihara, D.; Zhang, Y. *Proteins* **2004**, *56*, 502.
14. Zhang, Y.; Skolnick, J. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 1029.
15. Boas, F. E.; Harbury, P. B. *Curr. Opin. Struct. Biol.* **2007**, *17*, 199.
16. Quintilla, A.; Starikov, E.; Wenzel, W. *J. Chem. Theory Comput.* **2007**, *3*, 1183.
17. Verma, A.; Wenzel, W. *BMC Struct. Biol.* **2007**, *7*, 12.
18. Yang, J. S.; Chen, W. W.; Skolnick, J.; Shakhnovich, E. I. *Structure* **2007**, *15*, 53.
19. Summa, C. M.; Levitt, M.; Degrado, W. F. *J. Mol. Biol.* **2005**, *352*, 986.
20. Summa, C. M.; Levitt, M. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 3177.
21. Liwo, A.; Czaplewski, C.; Pillardy, J.; Scheraga, H. A. *J. Chem. Phys.* **2001**, *115*, 2323.
22. Freidrichs, M. S.; Wolynes, P. G. *Science* **1989**, *246*, 371.
23. Hardin, C.; Eastwood, M. P.; Luthey-Schulten, Z.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 14235.
24. Hardin, C.; Eastwood, M. P.; Prentiss, M. C.; Luthey-Schulten, Z.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 1679.
25. Zong, C.; Papoian, G. A.; Ulander, J.; Wolynes, P. G. *J. Am. Chem. Soc.* **2006**, *128*, 5168.
26. Prentiss, M. C.; Hardin, C.; Eastwood, M. P.; Zong, C.; Wolynes, P. G. *J. Chem. Theory Comput.* **2006**, *2*, 705.
27. Stryer, L. *Biochemistry*; W. H. Freeman and Company: New York, 1988.
28. Bryngelson, J. D.; Hopfield, J. J.; Southard Jr., S. N. *Tetrahedron Computer Methodology* **1990**, *3*, 129.
29. Krishna, M. M. G.; Maity, H.; Rumbley, J. N.; Lin, Y.; Englander, S. W. *J. Mol. Biol.* **2006**, *359*, 1410.
30. Ptitsyn, O. B. *Dokl. Nauk. SSSR* **1973**, *210*, 1213.
31. Baldwin, R. L.; Rose, G. D. *Trends Biochem. Sci.* **1999**, *24*, 26.
32. Cuff, J. A.; Clamp, M. E.; Siddiqui, A. S.; Finlay, M.; Barton, G. J. *Bioinformatics* **1998**, *14*, 892.
33. Pollastri, G.; McLysaght, A. *Bioinformatics* **2005**, *21*, 1719.
34. Rost, B.; Sander, C. *J. Mol. Biol.* **1993**, *232*, 584.
35. Jacobson, M. P.; Pincus, D. L.; Rapp, C. S.; Day, T. J. F.; Honig, B.; Shaw, D. E.; Friesner, R. A. *Proteins* **2004**, *55*, 351.
36. Papoian, G. A.; Ulander, J.; Eastwood, M. P.; Luthey-Schulten, Z.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 3352.
37. Goldstein, R. A.; Luthey-Schulten, Z. A.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 4918, 9029.
38. Aurora, R.; Creamer, T. P.; Srinivasan, R.; Rose, G. D. *J. Biol. Chem.* **1997**, *272*, 1413.
39. Scott, K. A.; Alonso, D. O. V.; Sato, S.; Fersht, A. R.; Daggett, V. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 2661.
40. Sharpe, T.; Jonsson, A. L.; Rutherford, T. J.; Daggett, V.; Fersht, A. R. *Prot. Sci.* **2007**, *16*, 2233.
41. Jayachandran, G.; Vishal, V.; Garcia, A. E.; Pande, V. S. *J. Struct. Biol.* **2007**, *157*, 491.
42. Kim, D. E.; Yi, Q.; Gladwin, S. T.; Goldberg, J. M.; Baker, D. *J. Mol. Biol.* **1998**, *284*, 807.
43. Charkrabarty, A.; Baldwin, R. L. *Adv. Protein Chem.* **1995**, *46*, 141.
44. Myers, J. K.; Pace, C. N.; Scholtz, J. M. *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 2833.
45. Dyson, H. J.; Merutka, G.; Waltho, J. P.; Lerner, R. A.; Wright, P. E. *J. Mol. Biol.* **1992**, *226*, 795.
46. Dyson, H. J.; Sayre, J. R.; Meruka, G.; Shin, H.-C.; Lerner, R. A.; Wright, P. E. *J. Mol. Biol.* **1992**, *226*, 819.
47. Jimenez, M. A.; Munoz, V.; Rico, M.; Serrano, L. *J. Mol. Biol.* **1994**, *242*, 487.
48. Saven, J. G.; Wolynes, P. G. *J. Mol. Biol.* **1996**, *257*, 199.
49. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, Jr., K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179.
50. Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G. M.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J. M.; Kollman, P. *J. Comput. Chem.* **2003**, *24*, 1999.
51. Onufriev, A.; Bashford, D.; Case, D. A. *Proteins* **2004**, *55*, 383.
52. Mongan, J.; Simmerling, C.; McCammon, J. A.; Case, D. A.; Onufriev, A. *J. Chem. Theory Comput.* **2007**, *3*, 156.
53. Eastwood, M. P.; Hardin, C.; Luthey-Schulten, Z.; Wolynes, P. G. *IBM J. Res. Dev.* **2001**, *45*, 475.
54. Koretke, K. K.; Luthey-Schulten, Z.; Wolynes, P. G. *Prot. Sci.* **1996**, *5*, 1043.
55. Andersen, C. A. F.; Palmer, A. G.; Brunak, S.; Rost, B. *Structure* **2002**, *10*, 175.
56. Case, D A.; Darden, T. A.; Cheatham III, T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, Jr., K. M.; Pearlman, D. A.; Crowley, M.; Walker, R. C.; Zhang, W.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Wong, K. F.; Paesani, F.; Wu, X.; Brozell, S.; Tsui, V.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Mathews, D. H.; Schafmeister, C.; Ross, W. S.; Kollman, P. A. *AMBER 9*; University of California: San Francisco, CA, 2006.
57. Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684.
58. Garcia, A. E.; Sanbonmatsu, K. Y. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 2782.
59. The definitions of the α_R, β, P_{II}, and α_L regions in the Ramachandran plot are the same as in the following paper: Kwac, K.; Lee, K.-K.; Han, J. B.; Oh, K.-I.; Cho, M. *J. Chem. Phys.* **2008**, *128*, 105106.