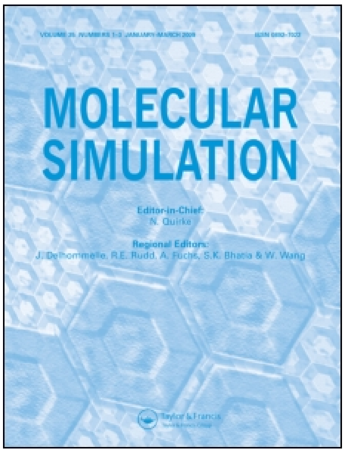


Molecular Simulation
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713644482

Refinement of comparative models of protein structure by using multicanonical molecular dynamics simulations

Ryuichiro Ishitani[a]; Tohru Terada[a]; Kentaro Shimizu[a]
[a] Agricultural Bioinformatics Research Unit, Graduate School of Agricultural and Life Sciences, The University of Tokyo, Tokyo, Japan

To cite this Article Ishitani, Ryuichiro , Terada, Tohru and Shimizu, Kentaro(2008) 'Refinement of comparative models of protein structure by using multicanonical molecular dynamics simulations', Molecular Simulation, 34: 3, 327 — 336
To link to this Article: DOI: 10.1080/08927020801930539
URL: http://dx.doi.org/10.1080/08927020801930539

Molecular Simulation
Vol. 34, No. 3, March 2008, 327–336

Taylor & Francis
Taylor & Francis Group

Refinement of comparative models of protein structure by using multicanonical molecular dynamics simulations

Ryuichiro Ishitani[a], Tohru Terada[a]* and Kentaro Shimizu[a,b]

[a]*Agricultural Bioinformatics Research Unit, Graduate School of Agricultural and Life Sciences, The University of Tokyo, Tokyo, Japan;* [b]*Department of Biotechnology, Graduate School of Agricultural and Life Sciences, The University of Tokyo, Tokyo, Japan*

(Received 26 July 2007; final version received 19 January 2008)

Comparative modelling is a powerful method that easily predicts a considerably accurate structure of a protein by using a template structure having a similar amino-acid sequence to the target protein. However, in the region where the amino-acid sequence is different between the target and the template, the predicted structure remains unreliable. In such a case, the model has to be refined. In the present study, we explored the possibility of a molecular dynamics-based method, using the human SAP Src Homology 2 (SH2) domain as the modelling target. The multicanonical method was used to alleviate the multiple-minima problem and the generalised Born/surface area model was used to reduce the computational cost. In addition, position restraints were imposed on the atoms in the reliable regions to avoid unnecessary conformational sampling. We analyzed the conformational distribution of the ligand-recognition loop of the domain and found that the most populated conformational clusters in the ensemble of the model agreed well with one of the two major clusters in the ensemble of the reference simulation starting from the crystal structure. This demonstrates that the current refinement method can significantly improve the accuracy of an unreliable region in a comparative model.

Keywords: comparative modelling; refinement; molecular dynamics; multicanonical method; implicit solvent model

1. Introduction

Comparative modelling (CM) is a powerful method for predicting the tertiary structure of a protein based on the sequence similarities with template proteins whose tertiary structures are already known [1]. Since this method can easily produce considerably accurate models, it has become quite popular in protein science and has been widely used for inferring the biological functions of structure-unknown proteins. Obvious from its way of generating models, the success of the CM method depends on whether a suitable template is available or not. In the regions where the sequences are highly conserved between the target and the template, the predicted structure is reliable, whereas in unconserved regions including insertions and gaps, the resulting structure tends to be unreliable. However, it is often the case that the amino-acid sequences of a protein family are locally diversified, causing functional diversity of, for example, ligand specificity. The bare CM method cannot provide a reliable model for the diversified region. In such cases, the models need to be refined.

Since the refinement can be viewed as searching for a global free-energy minimum in the conformational space of the model, energy-based approaches using molecular dynamics (MD) simulation are straightforward and therefore have been tried by several groups [2–4]. However, in spite of the simplicity of its principle, the MD-based method still remains a great challenge because it has a number of problems, i.e. the multiple-minima problem, a high computational cost and an inaccurate energy function. It is even said that MD generally leads to a model that is less like the experimental structure [5]. Therefore, in the Critical Assessment of Structure Prediction (CASP), an MD-based refinement was not actively performed for the CM models [6]. On the other hand, *ab initio* folding simulations have succeeded for several mini-proteins [7–11]. Encouraged by these successes, we explored the possibility of applying the MD-based refinement method in the present study.

Before applying the MD-based method, we have to cope with the aforementioned problems. The multiple-minima problem is caused by the large internal degrees of freedom and complicated intramolecular interactions of a protein. Due to this nature, the protein's energy surface is very rugged, with a large number of local energy-minimum states in which the trajectory of the

*Corresponding author. Email: tterada@iu.a.u-tokyo.ac.jp

ISSN 0892-7022 print/ISSN 1029-0435 online

DOI: 10.1080/08927020801930539
http://www.informaworld.com

conventional constant-temperature MD simulation can become easily trapped [12]. This trapping leads to convergence to a non-native structure that depends on the initial conditions. To alleviate this problem, we used a generalised ensemble method: the multicanonical MD method [13,14]. This method enables a random walk in the potential energy space so that the trajectory can easily escape from local minima. As a result, the multicanonical MD can efficiently sample a wider conformational space than the conventional MD can. Since the ensemble from the multicanonical MD simulation can be converted into a canonical one by reweighting [13–15], the global free-energy minimum state can be found by analyzing the canonical ensemble. This method has been applied to conformational sampling of various biomolecular systems and has been shown to be more efficient than the conventional MD [16–19].

The second problem, the large computational cost, can be divided into two sub-problems: the computational cost of each MD step and the length of simulation. To reduce the computational cost of each MD step, we used the generalised Born/surface area (GB/SA) model. This model is an implicit solvent model, and it can approximately calculate the solvation free energy more cheaply than the explicit water model can. Length of the simulation is mainly determined by the extent of the conformational space to be sampled. Since a wider conformational space is sampled in the multicanonical MD as described above, a longer simulation is required. In the CM models, one can distinguish reliable and unreliable regions based on the local sequence similarity between the target and the template and can estimate errors in the model coordinates. Therefore, we restricted the mobility of each amino acid according to its reliability: amino acids from the unreliable regions were free to move, while those from reliable regions were restrained to their initial position by using harmonic restraints. Such treatment can narrow down the conformational space and reduce the simulation length. In addition, the protein atoms do not experience the solvent's viscosity in the implicit solvent, which accelerates conformation transitions between local energy-minimum states. Therefore, the use of the GB/SA model can also reduce the simulation length.

Since there is no perfect energy function, the third problem, its inaccuracy, becomes inevitable. The use of the GB/SA model also causes an error in the calculation of the solvent free energy. In this study, we therefore focused on evaluating the effect of the enhanced conformational sampling on refinement. A multicanonical MD simulation was performed on the crystal structure as a reference. The ensembles from the simulations on the CM models were compared with the one from the reference simulation rather than with the crystal structure itself to cancel the effect of the energy-function error.

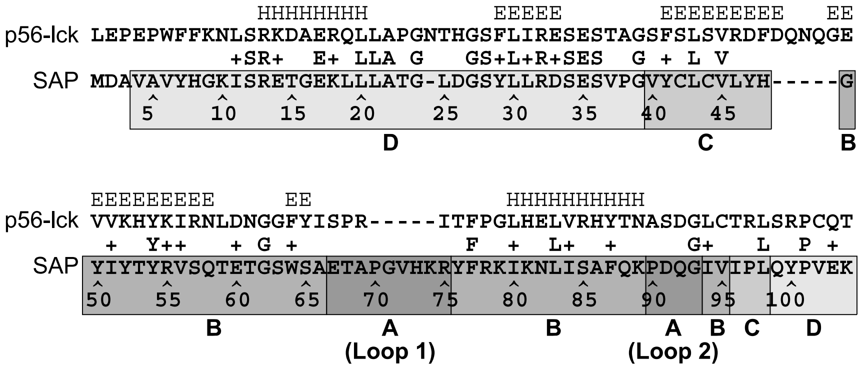
We took these measures against the problems and applied the MD simulation to CM-model refinement. Based on the refinement results that we got, we evaluated the strengths and weaknesses of the MD-based method.

2. Methods

2.1 Comparative modelling

We chose the Src homology 2 (SH2) domain as a model protein family, because it has unconserved regions in its amino-acid sequence that are related to its functional diversity [20]. This domain has a characteristic fold with an antiparallel central β sheet flanked by two α helices and specifically binds a peptide containing a phosphorylated tyrosine. In the target peptide, the residues located immediately C-terminal to the phosphotyrosine provide selectivity for specific SH2 domains [21]. These residues interact with the SH2 residues from the central β sheet and from two loops (hereafter referred to as loops 1 and 2, see Figure 1), which are less conserved in the SH2 family.

We chose the SH2 domain of human SAP protein [22] as the modelling target. This protein has an insertion in loop 1 and the sequence of loop 2 is different from



```
                    HHHHHHHH         EEEEE       EEEEEEEEE   EE
p56-lck LEPEPWFFKNLSRKDAERQLLAPGNTHGSFLIRESESTAGSFSLSVRDFDQNQGE
                  +SR+  E+ LLA G   GS+L+R+SES  G + L V
SAP     MDAVAVYHGKISRETGEKLLLATG-LDGSYLLRDSESVPGVYCLCVLYH-----G
            ^    ^    ^    ^     ^    ^    ^    ^    ^
            5   10   15   20    25   30   35   40   45
                             D                      C         B

        EEEEEEEEE     EE              HHHHHHHHHH
p56-lck VVKHYKIRNLDNGGFYISPR-----ITFPGLHELVRHYTNASDGLCTRLSRPCQT
         +  Y++   + G +            F  +  L+  +     G+   L  P +
SAP     YIYTYRVSQTETGSWSAETAPGVHKRYFRKIKNLISAFQKPDQGIVIPLQYPVEK
        ^    ^    ^    ^    ^    ^    ^    ^    ^    ^    ^
        50   55   60   65   70   75   80   85   90   95  100
                 B           A           B        A  B  C   D
                          (Loop 1)            (Loop 2)
```

Figure 1. Amino-acid sequence alignment between p56-lck and SAP. The secondary structure of the p56-lck crystal structure (1LKK) is designated above the alignment. 'H' and 'E' stand for helix and β strand, respectively. Ranges of regions A, B, C and D are indicated with boxes.

other SH2 domains, such as Src. The overall sequence identity of the SAP SH2 domain to the Src SH2 domain was about 27%. Its tertiary structure has already been determined in a ligand-free form by X-ray crystallography (PDB ID: 1D1Z). As for the template, we selected the SH2 domain of human p56-lck protein (PDB ID: 1LKK) from those whose sequences are similar to Src (sequence identity was about 50%) because its structure was determined with the highest resolution (1.0 Å). Note that the template structure was determined in complex with a phosphotyrosyl peptide, while the prediction was made for the peptide free form. The alignment between the target and the template sequences (Figure 1) was obtained from a multiple sequence alignment calculated by the program ClustalW [23] for SH2 domains whose structures have been experimentally determined. Sequence identity between the target and the template was 23% in the alignment. Since the first three residues of the SAP SH2 domain were missing from the crystal structure, they were not included in the modelling. Fifty CM models were generated by using the program MODELLER version 6 [24]. The qualities of the CM structures were evaluated by using the program Verify3D [25]. The model having the highest Verify3D score was selected for further refinement.

2.2 MD simulations

Multicanonical MD simulations were performed on the model and the crystal structure of the SAP SH2 domain. All the MD simulations were carried out under identical conditions by using a modified version of the SANDER module of the AMBER 6.0 suite [26] on a PC cluster equipped with Intel Xeon 2.8 GHz processors. The detailed conditions are as follows. A modified version of the AMBER parm99 force field parameter was used [11]. Solvation free energy was approximately calculated with the GB/SA model [27]. The parameters of the GB model were taken from Tsui and Case [28]. The dielectric constant of the solute interior was 1.0 and that of the solvent was 78.5. Salts were not included in the system. The value of the surface tension used in the SA model was 20.9 J mol^{-1} Å^{-2}. Non-bonded interactions and effective Born radii were calculated with a cutoff radius of 20 Å. Hydrogen-containing groups including XH$_n$ groups (X = C, N and O, and $n = 1$, 2 and 3) and aromatic rings were treated as rigid bodies to allow a longer time step (2 fs) for integration with negligible error [29]. Since the imidazole groups of the three His residues were all exposed to the solvent in the model, their protonation states were arbitrarily chosen to be the one protonated at N$^{\epsilon}$. Each system was first equilibrated by using a 1-ns constant temperature MD simulation at 300 K. A multicanonical MD simulation was then performed for 25 ns at 500 K with a multicanonical potential function that was determined so as to yield a flat energy distribution covering an energy region corresponding to 290–500 K by using a method described elsewhere [30]. Snapshot structures were recorded at every 0.1 ps. The conformational ensemble thus obtained was converted into a canonical one at 300 K by using a reweighting formula [13–15].

During the MD simulations, residues from reliable regions in the CM model were restrained to their initial positions to avoid unnecessary conformational sampling in these regions and to minimise the extent of the conformational space to be sampled. The reliable and unreliable regions were determined based on the sequence alignment (Figure 1) and the template structure. As mentioned above, the sequence of the target has an insertion (residues 70–74) in loop 1. In addition, loop 1 was expected to come very close to the N-terminal part of loop 2 from the template structure. Based on these observations, we defined residues 67–75 and 90–93 as the unreliable region (region A). The positions of these residues were not restrained. The remaining part of the target sequence was further divided into three regions, B, C and D, according to the distance from the residues in loop 1. Note that the boundaries of the regions were determined so that the residues in the same secondary-structure element were classified into the same region. Region B was composed of residues 49–66, 76–89 and 94–95 that were close to the residues of region A. The C$^{\alpha}$ atoms in this region were weakly restrained to their initial positions with harmonic potentials with a force constant of 0.418 kJ mol^{-1} Å^{-2}. The side-chain atoms were free to move in this region. Region C (residues 40–48 and 96–98) was composed of residues close to region B. Distances between non-hydrogen atoms in region C and those of loop 1 were greater than 10 Å. The positions of the non-hydrogen atoms in this region were restrained with a weak force constant of 0.418 kJ mol^{-1} Å^{-2}. The other residues (4–39 and 99–104) were included in region D and were restrained with a moderate force constant of 4.18 kJ mol^{-1} Å^{-2}. The same restraints were imposed on the corresponding atoms in the simulations starting from the crystal structure.

2.3 Principal component analysis

In order to visualise the conformational distributions in the ensemble generated by the multicanonical MD simulations, we performed a principal component analysis (PCA) of the deviations of the C$^{\alpha}$ atoms from the average [31]. In the calculations, we used a combined ensemble composed of ensembles from the two simulations with an equal weight for each snapshot structure (i.e. without the reweighting operation). The average structure was calculated for the combined ensemble by fitting the C$^{\alpha}$ atoms whose RMSD values were less than 1.5 Å. The variance-covariance matrix was then calculated for the C$^{\alpha}$ atoms of residues 66–76 from the deviations from the average, and the principal axes were calculated by diagonalising the matrix. The conformational distributions along the principal axes were calculated by separately projecting the trajectories of the model and reference simulations onto the principal axes.

2.4 Cluster analysis

The conformational distributions were also analyzed by using an RMSD-based clustering method [31]. In this analysis, each ensemble was reweighted at 300 K. The structures within a cutoff distance (1.0 Å) from a cluster centre as measured by the RMSD of the C$^{\alpha}$ and C$^{\beta}$ (if present) atoms of residues 66–76 were grouped into the cluster. The cluster centre was initially set to a structure selected from the ensemble, and after the grouping around the centre, a new cluster centre was calculated as the weighted average of the member. The grouping was repeated until the cluster centre converged. The probability of each cluster's existence was calculated as the sum of the weights of structures in the cluster. A representative structure of a cluster was defined as the one nearest to the cluster centre.

3. Results

3.1 Comparative modelling and MD simulations

Using the sequence alignment shown in Figure 1 and the template structure of the human p56-lck SH2 domain (PDB ID: 1LKK) [32], we generated 50 model structures of the human SAP SH2 domain. The quality of the model structures was evaluated by using the Verify3D program, and the scores were between 27.65 and 44.97. Figure 2 shows the structure of the model superimposed on the crystal structure of the human SAP SH2 domain (PDB ID: 1D1Z). The C$^{\alpha}$ RMSDs of region A and the other regions (regions B, C and D) between the model and the crystal structure were 7.0 Å and 1.4 Å, respectively (M$_{\mathrm{initial}}$ in Table 1). Multicanonical MD simulations were performed for the model, and for the crystal structure as a reference. All the simulations resulted in flat energy distributions covering an energy region corresponding to temperatures between 290 K and 500 K, which indicates that random walks in the energy space occurred. Figure 3 shows superpositions of snapshot structures every 2.5 ns during the multicanonical MD simulations. The conformations of loops 1 and 2 corresponding to region A were diversified in all simulations, indicating the conformational sampling was enhanced in this region. The distributions of structures in the other regions became more confined further away from region A, because of the stepwise position restraints. As can be seen from Figure 3, the conformations of loop 1 were particularly diversified. Therefore, in the following sections, we analyze the results of the simulations by paying special attention to the conformational distributions of loop 1.

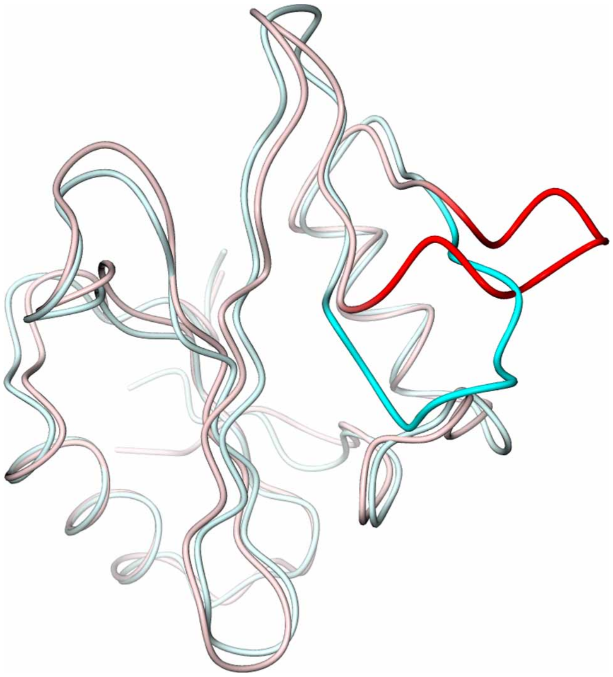


Figure 2. Backbone structures of the model (red) superimposed on that of the crystal structure of the human SAP (PDB ID: 1D1Z) (cyan) in tube representation. All structure representations were created with CueMol (http://www.cuemol.org/).

3.2 Conformational ensemble of the reference simulation

The ensemble from the reference simulation was classified into clusters by using the C$^{\alpha}$ RMSD of residues 66–76 (i.e. loop 1 plus one N- and C-terminal residues) as the distance measure. Two major clusters were obtained, and their probabilities of existence *P* were 52% and 40% at 300 K (Figure 4). Hereafter, these clusters are referred to as clusters R$_1$ and R$_2$ ('R' stands for 'Reference'). The structures of both clusters were significantly different from the crystal structure in region A. Specifically, residues 71–75 adopted a helical structure in both clusters (Figure 4), forming hydrogen bonds between backbone carbonyl and amide groups of Gly71 and Lys74 in the R$_1$ structure and between those of Val72 and Arg75 in the R$_2$ structure. This is different from the β-strand-like conformation in the crystal structure. Since these residues make close contact with a neighbouring molecule in the crystal, the difference

Table 1. C^{α} RMSD values between structures.

			RMSD from references (Å)		
Structures	Fitting[a]	Calculation[b]	Crystal structure (1D1Z)	R_1	R_2
$M_{initial}$	B, C, D	B, C, D	1.4	–	–
	B, C, D	A	7.0	–	–
	B, C, D	Loop 1	8.2	–	–
	Loop 1	Loop 1	3.0	–	–
$M_{modloop}$	B, C, D	B, C, D	1.4	1.6	1.6
	B, C, D	A	6.7	7.7	7.0
	B, C, D	Loop 1	7.9	9.0	8.1
	Loop 1	Loop 1	2.2	3.1	2.9
M_1	B, C, D	B, C, D	1.5	1.5	1.6
	B, C, D	A	4.7	4.3	3.9
	B, C, D	Loop 1	5.4	4.9	4.2
	Loop 1	Loop 1	3.6	1.8	3.0
M_2	B, C, D	B, C, D	1.4	1.5	1.5
	B, C, D	A	4.7	4.1	2.9
	B, C, D	Loop 1	5.4	4.5	3.3
	Loop 1	Loop 1	2.3	1.4	2.2

[a] C^{α} atoms in the regions listed here were used for root-mean-square fitting. [b] After the fitting, C^{α} atoms in the regions listed here were used for the RMSD calculation.

may be ascribed to the relaxation from the crystal environment. In addition, the difference is also due in part to the error of the GB/SA model, because this model tends to overestimate the stability of hydrogen bonds, giving a propensity toward a helical structure [18,33]. In R_1, the side chain of Thr68 loses interactions with other residues and is exposed to the solvent, whereas the interactions between Thr68 and the surrounding residues observed in the crystal structure are preserved in the second cluster (Figure 4). Due to this difference, the conformation of loop 1 of R_1 is more open than that of R_2 and the crystal structure. The 'open' conformation of loop 1 might be similar to the one adopted when the protein accepts its binding peptide.

3.3 Refinement of model

We performed PCA on the C^{α} atoms of residues 66–76 to compare the conformational distribution of the model simulation with that of the reference simulation. Figure 5(a) and (b) show contour plots of the probability distributions of unreweighted and reweighted ensembles of the model, respectively, as a function of the first and

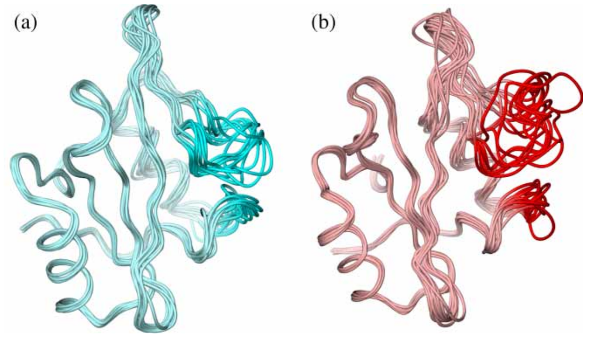


Figure 3. Superpositions of backbone structures every 2.5 ns during the multicanonical MD simulations, starting from the crystal structure (a) and the model (b).

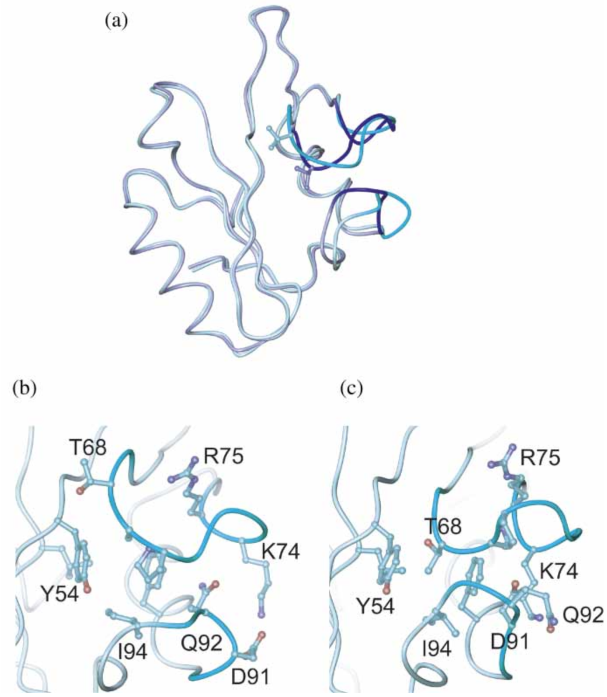


Figure 4. Backbone structures of representatives from clusters R_1 (cyan) and R_2 (blue) (a). Non-hydrogen side-chain atoms of Thr68 are shown in ball-and-stick representation. Close-up views of the peptide binding sites in the R_1 (b) and R_2 (c) structures with non-hydrogen side-chain atoms in ball-and-stick representation.

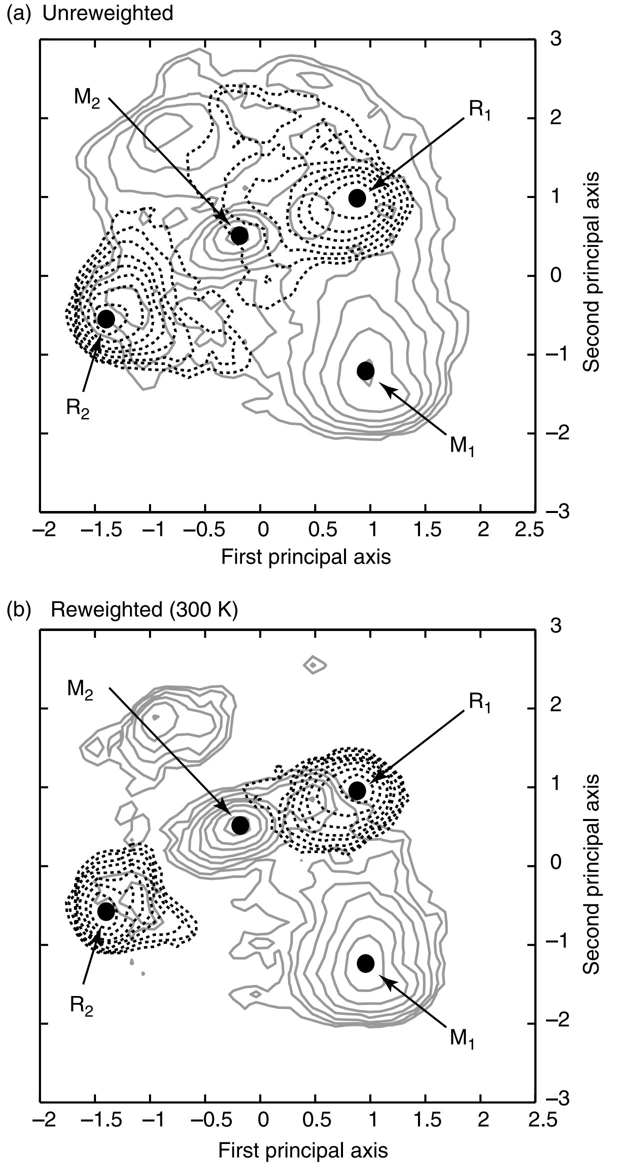


Figure 5. Projections of unreweighted (a) and reweighted (b) ensembles derived from multicanonical MD simulations onto the plane of the first and second principal axes. Gray solid lines show contour plots of probability distributions of the model simulation. The probability distributions of the reference simulation are plotted with dashed lines for comparison. Positions of representative structures of clusters R$_1$, R$_2$, M$_1$ and M$_2$ are indicated with filled circles.

second principal coordinates. The plots of the probability distributions of the reference simulation are overlaid for comparison. The first two principal axes explain 61% of the total variance in the unreweighted, combined ensemble. The unreweighted ensemble of the model simulation overlapped that of the reference simulation, which indicates that the multicanonical MD sampled similar conformational spaces irrespective of the initial structure. Actually, we found a structure in the ensemble of the model simulation that was quite similar to one of the structures in the ensemble of the reference simulation (Figure 6). The C$^{\alpha}$ atoms of residues 66–76 of the structures can be superimposed with an RMSD value of 0.39 Å. On the other hand, the reweighted distribution of the model simulation was quite different from that of the reference simulation: The distribution of the ensemble of the reference simulation was composed of two peaks that correspond to clusters R$_1$ (open conformation) and R$_2$ (closed conformation) described above, whereas the ensemble of the model simulation had one major and two minor peaks at different positions.

To further examine the conformations in the ensemble of the model simulation, we performed conformational clustering on the reweighted ensemble and compared the representative structures of the model and reference simulations. The clusters were ranked according to their populations, and the top two clusters ($P = 44$ and 27%) occupied more than half of the whole ensemble. Hereafter, the largest and second largest clusters are referred to as cluster M$_1$ and M$_2$, respectively. The representative structures of the two clusters were closer to that of R$_2$ than that of R$_1$ in terms of the C$^{\alpha}$ RMSD of loop 1 calculated by fitting the C$^{\alpha}$ atoms in region B–D (Table 1). For comparison, the initial model structure was submitted to an automated loop-modelling server, ModLoop [34], to obtain a refined structure for loops 1 and 2 (M$_{\mathrm{modloop}}$ in Table 1). Since the C$^{\alpha}$ RMSDs of loop 1 of M$_1$ and M$_2$ (3.3–5.4 Å) were better than those of M$_{\mathrm{modloop}}$ (7.9–9.0 Å), the structure of loop 1 was significantly improved by using the current refinement method.

The structure of residues 67–69 of cluster M$_1$ matched those of R$_2$ for both the backbone and side-chain atoms (Figure 7). Although the side chain of Thr68, one of the key residues for the peptide recognition [22], was exposed to the solvent in the initial structure of the model, it became partially buried in M$_1$, and recovered the interaction with the side chains of Tyr54 and Ile94 observed in the structure of R$_2$ and in the crystal structure (Figure 7B). However, in the structure of M$_1$, there was a critical difference from that of R$_2$. The conformation of residues 72–75, especially the arrangement of the side chains of Lys74 and Arg75, was quite different from that of R$_2$ as well as from the crystal structure. The side chain of Lys74 was pointed toward loop 2 and formed a hydrogen bond with the side chain of Asp91 in R$_2$ (Figure 7(d)), whereas the neighbouring residues, Arg75 and Gln92, substituted for Lys74 and Asp91 in M$_1$ (Figure 7(e)). This discrepancy in the hydrogen-bonding residues is the primary reason for the difference in the conformation of the C-terminal part of loop 1, which also affects the conformation of loop 2. Since the first principal coordinate is mainly related to the conformation of the C-terminal part of loop 1, the difference

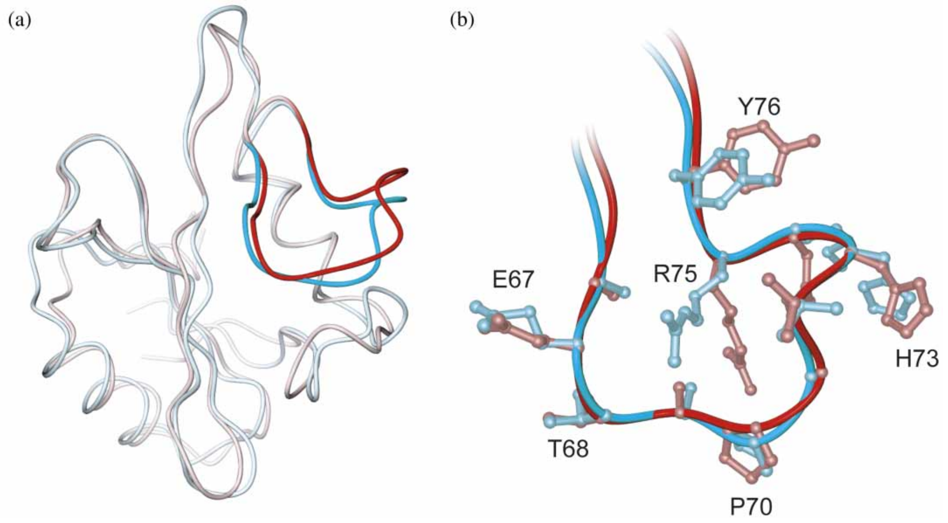


Figure 6. (a) Backbone structures of snapshots from the model (red) and reference (cyan) simulations that have the smallest pairwise C^{α}-RMSD value calculated for residues 66–76. (b) Close-up view of the loop-1 region. Non-hydrogen side-chain atoms of residues 66–76 are also shown in ball-and-stick representation.

in the first principal coordinates of M_1 and R_2 can be explained by the conformational difference of this part.

In contrast, the structure of M_2 matched that of R_2 in their C-terminal parts of loop 1 (Figure 7(c)). In this structure, a 'native' salt bridge was formed between Lys74 and Asp91 (Figure 7(f)). In addition, Asp91 formed another salt bridge with Arg75, which does not exist in R_2 and M_1. The $\chi 1$ angle of Tyr54 of M_2 was different from

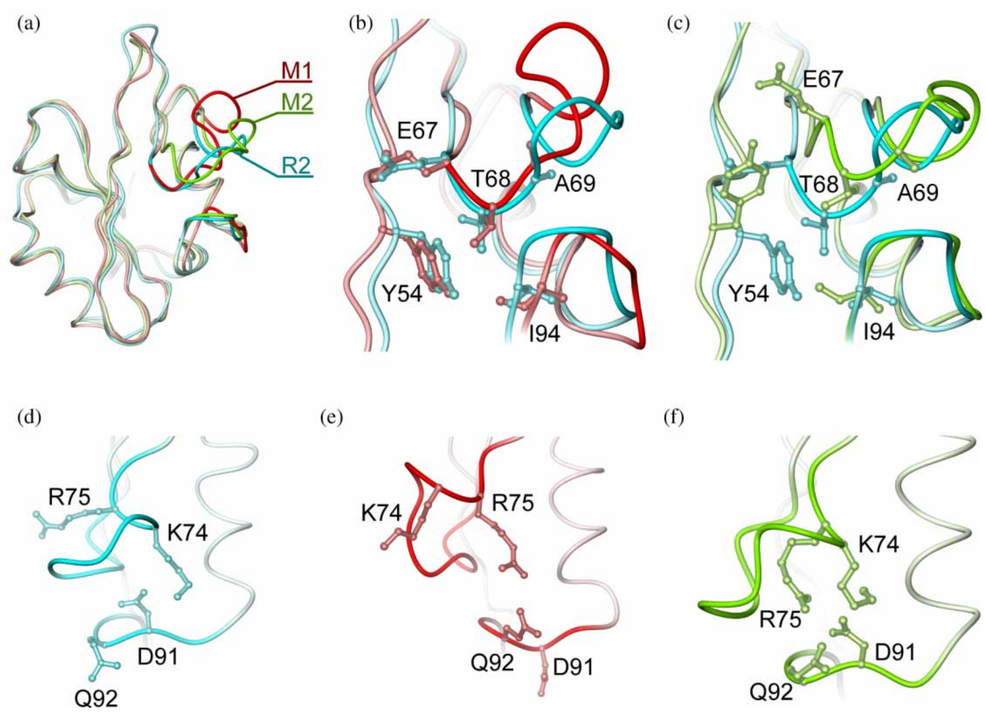


Figure 7. Backbone structures of representatives of clusters M_1 (red) and M_2 (green) superimposed on representative of cluster R_2 (cyan) (a) and their close-up views around the loop-1 region with non-hydrogen side-chain atoms of Tyr54, Glu67, Thr68, Ala69 and Ile94 in ball-and-stick representation ((b) and (c)). Panel b shows the representative structures of clusters M_1 and R_2, whereas panel c shows those of clusters M_2 and R_2. Close-up views from a different direction of the representative structures of clusters R_2 (d), M_1 (e) and M_2 (f). Non-hydrogen side-chain atoms of Lys74, Arg75, Asp91 and Gln92 are shown in ball-and-stick representation.

that of R_2 and M_1, being rotated so as to point the side chain to the solvent (Figure 7(c)). As a result, the side chain of Thr68 that interacts with Tyr54 shifted outward, and the conformation of the N-terminal part of loop 1 was more open than those of R_2 and M_1. This difference is related to the difference in the second principal coordinate of M_2 and R_2.

4. Discussion

Here, we consider the reason why the reweighted distributions of the model and the reference simulations did not match very well, despite that the PCA map of the unreweighted ensemble of the model sufficiently overlapped that of the reference simulation. Since only the coordinates of C^α atoms were considered in the PCA, the overlap in the PCA maps only indicates the backbone structures matched. The structures having the same PCA coordinates can have different side chain conformations and can therefore have different free energies. Since the structures of M_1 and M_2 were different from that of R_2 in the salt-bridge pattern and in the $\chi 1$ angle of Tyr54, respectively, we compared the salt-bridge distances and the $\chi 1$ angle between the model and reference simulations (Figure 8). In the reference simulation, a hydrogen bond was formed between Arg75 and Gln92 for a short time at the beginning of the simulation, and then a salt bridge was formed between Lys74 and Asp91 for most of the rest of the simulation time. Although the salt-bridge distance between Lys74 and Asp91 fluctuated, a transition to a different salt-bridge state was not observed after the Lys74–Asp91 salt bridge was established. In contrast, the Arg75–Asp91 salt bridge was also observed in the model simulation, as was the Arg75–Gln92 hydrogen bond and the Lys74–Asp91 salt bridge. However, a state in which only the Lys74–Asp91 salt bridge is formed was not observed in the model simulation. Therefore, the discrepancy in the reweighted distributions must be due to the failure in sampling all of the possible interaction combinations. To sample all of them, it would be necessary to make the transition to different interaction modes occur more frequently. Although a higher simulation temperature can improve sampling efficiency, it greatly enlarges the conformational space to be sampled. The extension of the simulation time might not be helpful either, because a transition to different interaction mode was not observed in the reference simulation that was extended to 40 ns. A practical solution would be to use umbrella potentials that weaken the hydrogen-bond and salt-bridge interactions.

As for the $\chi 1$ angle of Tyr54, the state of the $\chi 1$ angle of 180°, which corresponds to the M_2 structure, was rarely sampled in the reference simulation. Since Tyr54 is in region B, the position of its C^α atom was weakly restrained. Because the $\chi 1$ angle of His73, an unrestrained residue in region A, changed more frequently, the position restraint probably interfered with the sampling of the $\chi 1$ angle of Tyr54. It may be worth trying other types of restraints, such as dihedral angle restraints.

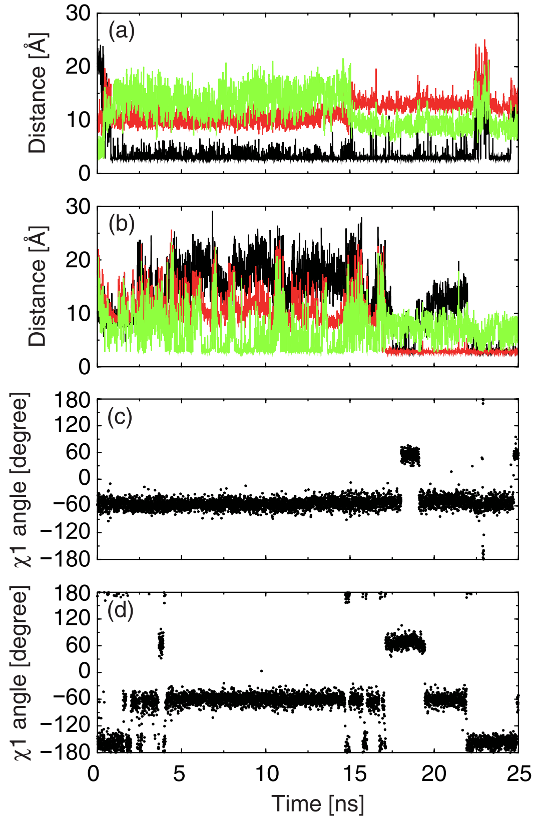


Figure 8. Time evolution of the distances between Lys74 N^ζ and Asp91 O^δ (black), between Arg75 N^ϵ or N^η and Asp91 O^δ (red), and between Arg75 N^ϵ or N^η and Gln92 O^ϵ (green) during the reference (a) and model (b) simulations. The smallest value among the distances between possible atom pairs (e.g. Lys74 N^ζ–Asp91 $O^{\delta 1}$ and Lys74 N^ζ–Asp91 $O^{\delta 2}$) is shown at each time. Time evolution of the $\chi 1$ angle of Tyr54 during the reference (c) and model (d) simulations.

5. Conclusions

We evaluated the usefulness of the multicanonical MD simulation in the refinement of the CM model. We used the multicanonical MD method and the GB/SA model to cope with the multiple minima and the large computational cost problems, respectively. In addition, position restraints were imposed on the atoms in the reliable regions to narrow down the conformational space to be sampled. The ensemble from the simulation of the model was compared with the one obtained from a reference simulation starting from the crystal structure to cancel the errors in the energy function and the GB/SA model. The structures of the most and the second populated clusters in the ensemble of the model simulation were close to the structure of one of the two major clusters in the ensemble of the reference simulation. Therefore, we concluded that the current refinement method can significantly improve the accuracy of an unreliable region in a CM model. However, there were critical differences in the side-chain structures, resulting in discrepancies in the conformational distributions in the reweighted ensembles. To achieve higher precision, the sampling efficiency of the side-chain conformations will have to be improved. The sampling was hampered by the strong hydrogen-bond and salt-bridge interactions and by the position restraints imposed on the atoms in the reliable region. Therefore, practical solutions of this problem may involve imposing umbrella potentials that break interactions that are too strong or devising restraints on the backbone atoms of the reliable region that do not affect the conformational sampling of side chains.

Very recently, Chen and Brooks III reported that an MD-based method using a short (3–5 ns) replica exchange MD simulation and the GB/SA model could refine the protein model from CASP7 refinement targets [35]. Although their objective was to improve the overall agreement of the model structures with the experimental structures and it is different from ours to refine the unreliable region, our study and theirs together demonstrate that the MD-based method is now reliable and useful for structure refinement, if it is properly combined with an enhanced conformational sampling method.

Acknowledgements

This work was supported in part by Grants-in-Aid for Scientific Research on Priority Areas (no. 16014204) from the Ministry of Education, Culture, Sports, Science and Technology of Japan and by a grant from the Ministry of Agriculture, Forestry and Fisheries of Japan (Rice Genome Project) to K.S. This work was also supported by the Program for Promotion of Basic Research Activities for Innovative Biosciences (PROBRAIN) in Japan.

References

[1] M.A. Marti-Renom et al., *Comparative protein structure modeling of genes and genomes*, Annu. Rev. Biophys. Biomol. Struct. 29 (2000), p. 291.

[2] M.R. Lee et al., *Molecular dynamics in the endgame of protein structure prediction*, J. Mol. Biol. 313 (2001), p. 417.

[3] J.A. Flohil, G. Vriend, and H.J.C. Berendsen, *Completion and refinement of 3-D homology models with restricted molecular dynamics: Application to targets 47, 58, and 111 in the CASP modeling competition and posterior analysis*, Proteins 48 (2002), p. 593.

[4] B. Qian, A.R. Ortiz, and D. Baker, *Improvement of comparative model accuracy by free-energy optimization along principal components of natural structural variation*, Proc. Natl Acad. Sci. USA 101 (2004), p. 15346.

[5] P. Koehl and M. Levitt, *A brighter future for protein structure prediction*, Nat. Struct. Biol. 6 (1999), p. 108.

[6] M. Tress et al., *Assessment of predictions submitted for the CASP6 comparative modeling category*, Proteins 61(Suppl 7) (2005), p. 27.

[7] D. Satoh et al., *Folding free-energy landscape of a 10-residue mini-protein, chignolin*, FEBS Lett. 580 (2006), p. 3422.

[8] C.D. Snow et al., *How well can simulation predict protein folding kinetics and thermodynamics?*, Annu. Rev. Biophys. Biomolec. Struct. 34 (2005), p. 43.

[9] S. Chowdhury et al., Ab initio *folding simulation of the Trp-cage mini-protein approaches NMR resolution*, J. Mol. Biol. 327 (2003), p. 711.

[10] C.L. Brooks III, *Protein and peptide folding explored with molecular simulations*, Acc. Chem. Res. 35 (2002), p. 447.

[11] C. Simmerling, B. Strockbine, and A.E. Roitberg, *All-atom structure prediction and folding simulations of a stable protein*, J. Am. Chem. Soc. 124 (2002), p. 11258.

[12] A. Mitsutake, Y. Sugita, and Y. Okamoto, *Generalized-ensemble algorithms for molecular simulations of biopolymers*, Biopolymers (Pept. Sci.) 60 (2001), p. 96.

[13] U.H.E. Hansmann, Y. Okamoto, and F. Eisenmenger, *Molecular dynamics, Langevin and hybrid Monte Carlo simulations in a multicanonical ensemble*, Chem. Phys. Lett. 259 (1996), p. 321.

[14] N. Nakajima, H. Nakamura, and A. Kidera, *Multicanonical ensemble generated by molecular dynamics simulation for enhanced conformational sampling of peptides*, J. Phys. Chem. B 101 (1997), p. 817.

[15] A.M. Ferrenberg and R.H. Swendsen, *New Monte Carlo technique for studying phase transitions*, Phys. Rev. Lett. 61 (1988), p. 2635.

[16] N. Nakajima et al., *Flexible docking of a ligand peptide to a receptor protein by multicanonical molecular dynamics simulation*, Chem. Phys. Lett. 278 (1997), p. 297.

[17] H. Shirai et al., *Conformational sampling of CDR-H3 in antibodies by multicanonical molecular dynamics simulation*, J. Mol. Biol. 278 (1998), p. 481.

[18] T. Ishizuka et al., *Improvement of accuracy of free-energy landscapes of peptides calculated with generalized Born model by using numerical solutions of Poisson's equation*, Chem. Phys. Lett. 393 (2004), p. 546.

[19] R. Jono, K. Shimizu, and T. Terada, *A multicanonical* ab initio *molecular dynamics method: Application to conformation sampling of alanine tripeptide*, Chem. Phys. Lett. 432 (2006), p. 306.

[20] G. Waksman and J. Kuriyan, *Structure and specificity of the SH2 domain*, Cell S116 (2004), p. S45.

[21] T. Pawson and G.D. Gish, *SH2 and SH3 domains: From structure to function*, Cell 71 (1992), p. 359.

[22] F. Poy et al., *Crystal structures of the XLP protein SAP reveal a class of SH2 domains with extended, phosphotyrosine-independent sequence recognition*, Mol. Cell 4 (1999), p. 555.

[23] R. Chenna et al., *Multiple sequence alignment with the Clustal series of programs*, Nucleic Acids Res. 31 (2003), p. 3497.

[24] A. Sali and T.L. Blundell, *Comparative protein modeling by satisfaction of spatial restraints*, J. Mol. Biol. 234 (1993), p. 779.

[25] R. Luthy, J.U. Bowie, and D. Eisenberg, *Assessment of protein models with three-dimensional profiles*, Nature 356 (1992), p. 83.

[26] D.A. Case et al., AMBER. 6, University of California, San Francisco, CA, 1999.

[27] W.C. Still et al., *Semianalytical treatment of solvation for molecular mechanics and dynamics*, J. Am. Chem. Soc. 112 (1990), p. 6127.

[28] V. Tsui and D.A. Case, *Theory and applications of the generalized Born solvation model in macromolecular simulations*, Biopolymers (Nucleic Acid Sci.) 56 (2000), p. 275.

[29] T. Terada and A. Kidera, *Generalized form of the conserved quantity in constant-temperature molecular dynamics*, J. Chem. Phys. 116 (2002), p. 33.

[30] T. Terada, Y. Matsuo, and A. Kidera, *A method for evaluating multicanonical potential function without iterative refinement: Application to conformational sampling of a globular protein in water*, J. Chem. Phys. 118 (2003), p. 4306.

[31] O.M. Becker, *Conformational analysis*, in *Computational Biochemistry and Biophysics*, O.M. Becker, A.D. MacKerell Jr., B. Roux and M. Watanabe, eds., Marcel Dekker, New York, NY, 2001, pp. 69–90.

[32] L. Tong et al., *Crystal structures of the human p56(lck) SH2 domain in complex with two short phosphotyrosyl peptides at 1.0 Angstrom and 1.8 Angstrom resolution*, J. Mol. Biol. 256 (1996), p. 601.

[33] R. Zhou and B.J. Berne, *Can a continuum solvent model reproduce the free energy landscape of a beta-hairpin folding in water?*, Proc. Natl Acad. Sci. USA 99 (2002), p. 12777.

[34] A. Fiser and A. Sali, *ModLoop: Automated modeling of loops in protein structures*, Bioinformatics 19 (2003), p. 2500.

[35] J. Chen and C. L. Brooks, III, *Can molecular dynamics simulation provide high-resolution refinement of protein structure?*, Proteins 67 (2007), p. 922.