Generation of a flexible loop structural ensemble and its application to induced-fit structural changes following ligand binding

Yukihisa S. Watanabe^{1,4}, Yoshifumi Fukunishi² and Haruki Nakamura^{2,3}

¹Japan Biological Information Research Center (JBIRC), Japan Biological Informatics Consortium (JBIC), 2-41-6 Aomi, Koto-ku, Tokyo 135-0064, Japan

²Biological Information Research Center (BIRC), National Institute of Advanced Industrial Science and Technology (AIST), 2-41-6 Aomi, Koto-ku, Tokyo 135-0064, Japan

³Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka 565-0871, Japan

⁴Present address: Drug Discovery Research Laboratories, Sanwa Kagaku Kenkyusho Co. Ltd., 363 Shiosaki, Hokusei-cho, Inabe, Mie 511-0406, Japan

Received 19 July, 2005; accepted 5 December, 2005

Molecular recognition is often mediated by flexible loops that have widely fluctuating structures and are sometimes disordered, but that form particular complex structures following ligand binding. In fact, many loop structures found in the PDB database are too flexible to be determined precisely. A new loop modeling method was therefore developed using force-biased multicanonical molecular dynamics with the implicit solvent model to generate an ensemble of putative loop structures with low free energy values. The method was then used to create ensembles for several flexible loops that were compared with the corresponding NMR and X-ray structures. The induced-fit structural change of dihydrofolate reductase (DHFR) was also predicted from a structural ensemble of ligand-free M20 loop conformations and successive docking simulations.

Key words: docking simulation, flexible loop modeling, force-biased multicanonical molecular dynamics, induced-fit, structural ensemble

The elucidation of the relationship between the three-dimensional (3D) structure of a biological macromolecule

and its function remains an essential theme in structural biological research. Flexible loops play an important role in molecular functions, including those pertaining to enzymatic activity, macromolecule-macromolecule binding and ligand recognition^{1–4}. In many cases the flexible loops adopt particular rigid conformations when they function, such as in the case of ligand binding. However, in their ligand-free state, the structures are too flexible to be precisely and unambiguously identified due to the low electron densities derived from X-ray crystallography or the few Nuclear Overhauser Effect (NOE) cross peaks derived from NMR spectroscopy. In Fig. 1, two flexible loops of *bovine* ribonuclease A (RNase A) are shown with the 32-loop conformations observed by NMR (PDB code, 2AAS)⁵. Here, the diameter of the pipe model for the backbone of RNase A corresponds to the root-mean square deviations (RMSDs) of the backbone atoms, obtained from the 32 distance-geometry conformations.

Flexible loops in proteins form several different 3D structures with similar free energies, as well as short peptides that possess a number of local free energy minima^{6,7}. To build 3D conformations of the loops, many modeling approaches have been developed. These include the application of database searches and use of *ab initio* techniques.

The database search method is based on searches for loop fragments in the PDB database for the corresponding loops^{8–11}. A previous study reported that loops longer than

Corresponding authors: Yukihisa S. Watanabe, e-mail: yswatanabe@jbirc.aist.go.jp; Haruki Nakamura, e-mail: harukin@protein.osaka-u.ac.jp

Figure 1 Structure of RNase A: A pipe model of RNase A (PDB code, 2AAS)⁵ drawn using MOLMOL⁵⁴. The diameter of the pipe for the RNase A backbone is proportional to the RMSDs of the backbone atoms, obtained from the 32 distance-geometry conformations. The green and blue pipes represent Loop R8 with 8 residues and Loop R12 with 12 residues, respectively.

seven residues were not precisely modeled by this method¹². Furthermore, a loop conformation depends on its spatial neighborhood conditions, such as the residues surrounding the loop, ligands, solvents, dielectric constant, ions, lipids, temperature and crystallization conditions¹³. Thus, the reliability of the loop structures derived from database searches has been limited.

As an alternative, over the past 35 years *ab initio* loop prediction methods have been developed using various techniques^{14–16}. In particular, broad conformational sampling and clustering have been performed in loop modeling, and resulted in attainment of the global and local free energy minima^{17,18}. For the calculations, structural ensembles were observed as putative flexible loop conformations. This approach is very promising in the area of structured determination^{19,20}. However, achieving broad conformational sampling using these methods was problematic, and high-throughput prediction was never performed.

It is now widely understood that proteins can undergo conformational changes associated with ligand binding, ligand recognition or signal transduction^{1,2}. Such “induced-fit” or “induced-folding” changes are common mechanisms associated with molecular recognition, but only single complexed conformations are detected by NMR or X-ray crystallography. The NOE signals and X-ray diffraction data should reflect the structural variation of such flexible regions. However, the conventional analyses produce only one or a few major structures. Even if weak NOE signals or weak electron densities that may correspond to other minor

structures are observed, those signals tend to be ignored as inconsistent or unreliable data. In addition, when the X-ray diffractions are observed at low temperature using a cryo-stage, the probability of the major structure is emphasized. Recent studies showed that a structural ensemble could be obtained by using new modeling schemes in which multiple structural templates are allowed^{19,20}. These modeling schemes are expected to be used for understanding the dynamics of flexible regions.

The basis of structural changes should be the very small free energy difference between the structure with and without a ligand. Namely, it is considered that the ensemble of ligand-free protein loop conformations could include the ligand-bound loop conformation — although not the most common one — in equilibrium with the ligand-free conformations. Following ligand binding, the flexible loop should form a particular conformation that possesses a lower free energy than any of the ligand-free loop conformations in the prepared ensemble. Based on this assumption, the use of *ab initio* loop modeling to generate the free energy landscape is a promising way to predict the structural change induced after conformational sampling and clustering.

The M20 loop of *E. coli* dihydrofolate reductase (DHFR) is a good example of induced-fit structural changes following ligand binding. The M20 loop from residue 9 to 24 lies over the active site Met 20. Conformations of both ligand-bound and ligand-free M20 loops have been investigated by a variety of approaches, including X-ray crystallography^{21,22}, NMR^{3,23,24}, circular dichroism spectroscopy^{25,26} and theoretical calculations^{27,28}. The conformational change of the M20 loop occurs in a time range of picoseconds to nanoseconds, during which the loop assumes three conformations, referred to as the *closed* (PDB code, 1RX2), *open* (1RA2) and *occluded* (1RX5) conformations^{3,22}. These three loop conformations form the type III’ hairpin, the irregular structure, and the one-turn 3_{10} helix, respectively. In contrast, X-ray crystallographic analysis of the ligand-free M20 loop of DHFR (PDB code, 5DFR) appeared to show a disordered conformation where the atomic location of the 5 residues from 16 to 20 of the M20 loop could not be identified. This perhaps reflected the dynamic characteristic of the M20 loop^{3,22}.

In this paper, we describe a new loop modeling method combining force-biased multicanonical molecular dynamics, FBMcMD²⁹, with the implicit solvent model applying the generalized Born (GB) with the solvent accessible surface area model (GBSA)^{30,31}. Although use of the implicit solvent model tended to result in an overestimation of the effect of intra-molecular hydrogen bonds, all of the free energy landscapes generated for short peptides with the explicit water model agreed well with our previous study³². In general, use of the multicanonical molecular dynamics (McMD) algorithm based on a non-Boltzmann probability can eliminate trapping at local minima and generate wide conformational sampling to obtain free energy landscapes^{33,34}.

This approach is the most extensively used of all the generalized ensemble methods. Because use of FBMcMD can generate an ensemble of loop conformations automatically, many putative loop conformations have been easily obtained in this manner^{29,32}. In the present study, these methods were used to generate the structural models of the two loops in Fig. 1, which were compared with the NMR and X-ray structures. Finally, ligand-free M20 loop conformations of DHFR were modeled in the same way to reveal the mechanism pertaining to induced-fit.

Materials and methods

Loop R8 in RNase A

The modeled region was composed of residues 58–74 and 106–123 in RNase A (a total of 35 residues), which included the flexible loop region from residue 64 to 71 and the surrounding residues adjacent to the flexible loop region. In choosing the surrounding residues, we first picked the residues inside a sphere with a radius of 20 Å and center at the average position of two junctions at residue 63 to 64 and residue 71 to 72 between the flexible loop region and the protein core. Then, examining the X-ray crystal structure, we removed the residues that could not interact with the flexible loop residues due to an obstruction by the backbone structure that forms the secondary structures. The flexible loop was initially replaced with a different loop as the initial fragment conformation derived from a PDB database search using the program FRGMNT³⁵. Here, the peptide fragment from residue 107 to 114 in the bilin-binding protein from *Pieris Brassicae* (PDB code, 1BBP) was used. The other atomic coordinates were taken from those of the X-ray crystal structure (8RAT)³⁶. The RMSD value of the initial backbone heavy atoms of the flexible loop was 8.37 Å against that of the X-ray crystal structure. The N-terminus and C-terminus of the individual peptide chains were capped with acetyl and N-methyl groups, respectively. During the subsequent calculations, the flexible loop region consisting of the above 8 residues was completely free without any restraints, and the atomic coordinates of the other 27 residues were restrained to their initial coordinates by a restraint harmonic potential with the force constant $k=0.1$ kcal/mol/Å².

Loop R12 in RNase A

The modeled region was composed of residues 5–36, 43–51, 79–85 and 96–104 (a total of 57 residues). The flexible region was from residue 13 to 24, and the adjacent surrounding residues were chosen in the same manner as for Loop R8. The initial loop structure was built using the peptide fragment of residue 81 to 92 in aspartic protease from *Rous sarcoma retrovirus* (2RSP). The RMSD value of the initial backbone heavy atoms of the flexible loop was 7.12 Å against that of the X-ray crystal structure. During the calculations, the above 12 residues in the flexible region

were completely free without any restraints, and the atomic coordinates of the other 45 residues were restrained in the same manner as those for Loop R8.

Loop M20 in DHFR

The M20 loop consists of 16 residues (residues 9 to 24), with residues 14 to 25 being very flexible. The modeled region was then composed of residues 6–28, 44–50, 94–97, 113–125 and 145–153 (a total of 56 residues). In the apo-DHFR crystal structure (PDB code, 5DFR)³⁷, there are no atom coordinates for the flexible region from residue 16 to 20. The initial loop structure was built using the peptide fragment of residue 69 to 80 in yeast cytochrome c peroxidase (2CYP), in the same manner as that for Loops R8 and R12. The backbone RMSD value of the flexible loop was 5.32 Å against that in the closed form (1RX2)²², 5.69 Å against that of the open form (1RA2)²², and 4.77 Å against that of the occluded form (1RX5)²². The other atomic coordinates were taken from those of the apo-form crystal structure without a ligand or cofactor. During the calculations, the above 12 residues making up the flexible loop region were completely free without any restraints, and the atomic coordinates of the other 44 residues were restrained in the same manner as those for Loop R8.

Simulation methods

FBMcMD with the weighted histogram technique was applied to the conformational sampling to reveal the free energy landscape of the canonical ensemble at 300 K. We will describe the FBMcMD algorithm in brief; a more detailed description is available in a previous study²⁹. With the FBMcMD approach, the trajectory is generated by performing constant-temperature MD at an arbitrarily chosen reference temperature $\beta_0=1/k_B T_0$, with force scaling as

$$\dot{\mathbf{p}}_i = \nu^k(E) \mathbf{f}_i, \quad (1)$$

where $\mathbf{p}_i$ and $\mathbf{f}_i$ correspond to the momentum and force acting on the atom i with a potential energy E , respectively. Here, $\nu^k(E)$ is the force scaling function that is iteratively refined using the update scheme of

$$\nu^{k+1}(E) = \nu^k(E) + 1/\beta_0 (\partial \ln P^k(E)/\partial E), \quad (2)$$

where $P^k(E)$ is the probability distribution at the k th iteration step. The advantage of FBMcMD is that the simulation can be done automatically through repeated brief multi-canonical samplings of N_k MD steps with the dynamic update scheme of Eq. (2). Since the weights of each iteration step are slightly different, all simulation data have been joined to estimate the density of states $\Omega(E)$ using a weighted histogram analysis³⁸ as

$$\Omega(E) = (\sum_k H_k(E)) / (\sum_k N_k Z_k^{-1}(E) \exp(-\beta_0 \alpha^k(E))), \quad (3)$$

where $H_k(E)$ and N_k correspond to the energy histogram and number of samplings at the k th iteration, respectively. Here,

$\alpha^k(E) = \int_{-\infty}^E v^k(E') dE'$ and $Z_k(E)$ is the relative partition function.

Once $\Omega(E)$ is determined, the canonical ensemble at an arbitrary temperature is reconstructed as $P_o(E, T) = \Omega(E) \exp(-\beta E)$. Furthermore, various thermodynamic quantities of O can also be determined as

$$\langle O \rangle_T = \int dE O(E) P_o(E, T). \quad (4)$$

Instead of the explicit water model for describing the solvent, the GBSA implicit solvent model³⁰ was applied, using the Hawkins' formulation as the GB part³¹ and the conventional surface area calculation with a probe radius of 1.4 Å^{39,40}. The other GBSA parameters were the same as those used in our previous paper³².

The FBMcMD simulation with the GBSA implicit solvent model was performed with the AMBER force field parm 96⁴¹ using the in-house program *cosgene* included in the *prestox* ver 2.0 program suite⁴². The *cosgene* program provides molecular mechanics and molecular dynamics simulations. The unit time step of the MD simulation was 1.0 fs with the SHAKE method⁴³. The cut-off distance for non-bonded interactions was 20 Å. The temperature was controlled by the constant-temperature method⁴⁴. MD simulations of Loops R8, R12, and M20 were carried out with 30×10^6 , 80×10^6 , and 100×10^6 steps, respectively. For other parameters, the same values were used as in our previous report for short free peptides³².

In all cases, flat energy distributions were obtained from 250 K to 700 K, and re-weighted canonical ensembles at 300 K were obtained. The potential energy and all atomic coordinates of the system were monitored at each step and every 500 steps, respectively. The principal component analysis (PCA) method was applied as described in our previous paper¹⁸. All MD calculations were carried out on a Compaq ES40 cluster, and structural analyses were made on a Sun Fire 3800.

These modeled loops were verified by the in-house ligand-protein docking simulation program *sievgene* included in the *prestox* ver. 2.0 program suite. The program used the docking algorithm in combination with the rigid-protein and flexible-ligand technique. Details of the docking algorithm and the computation are given elsewhere⁴⁵; a brief summary is provided here.

The scoring function of the docking method is based on the rough shape of a protein surface to decrease structural noise. The conventional potential function is applied to the outer region of the receptor protein, and a smooth virtual function is calculated for the inner region of the receptor protein. Assuming that at least three ligand atoms come into contact with the protein surface, a geometrical hash method is used for protein-ligand conformation searching. This method was applied to the 132 known protein-ligand complexes, and correctly predicted approximately 50% of these

complex conformations within the 2 Å RMSD. In the current study, the number of conformers generated in docking was limited to 100 for each ligand.

Only single bonds of ligands were flexible and they were allowed to rotate every 120 degrees. In contrast, the protein structure was rigid. The search area of the ligand docking was limited to the area around the original ligand coordinates of the protein-ligand complex structure PDB code, 1RX2, where the shortest distance between the docked atoms and the original ligand atoms should be less than 5 Å.

The atomic charges of the ligands were determined by the restricted electrostatic point charge (RESP) procedure using HF/6-31G*-level quantum chemical calculations with the program GAMESS⁴⁶. Structure optimization after docking simulation was performed by potential energy minimization using the program *cosgene* in the *prestox* ver 2.0 program suite⁴² with the AMBER force field parm 96⁴¹.

The program suite *prestox* that includes the *cosgene* and *sievgene* programs is written in FORTRAN90. Information concerning the availability and distribution of the program can be obtained from <http://www.jbic.or.jp/> or by E-mail to prestox@jbic.or.jp.

Results and Discussion

Structural ensemble of the loop conformations

The residues that form the flexible loop, including the surrounding residues, were first extracted from the entire protein molecule. For the initial conformation of the flexible loop, an appropriate conformation was generated by searching the PDB database as described in the Materials and Methods section. Following this, the FBMcMD calculation with the implicit solvent model was performed for the system, where the positions of the backbone atoms except the flexible loop region were restrained by harmonic forces.

The potential energy trajectory of a multicanonical simulation is shown in Fig. 2A for the 8-residue flexible loop system in RNase A designated as Loop R8 in Fig. 1. Loop R8 is composed of flexible loop residues 64 to 71, and 27 other surrounding residues. The potential energy was initially ca. -290 kcal/mol, and gradually increased to cover a wide energy range by random-walking in the energy space. Figure 2B shows the energy distribution of the multicanonical ensemble for this flexible loop system obtained by the current FBMcMD method, with the energy distributions for the re-weighted canonical ensembles at 300 K and 700 K. The flat energy distribution in Fig. 2B indicates that the current method succeeded in generating the multicanonical ensemble.

In previous studies, the simulated annealing (SA) method was applied to loop modeling as a structural search method^{16,47}. An advantage of our method over the SA method is that it can construct a canonical ensemble of loop structures. Another advantage is that, with our method, we can confirm that the structural search is completed when the flat

Figure 2 (A) The trajectory of the potential energy of Loop R8 as determined by the FBMCMC method with the GBSA model. (B) Energy distribution of the multicanonical ensemble (solid line). The dotted curve and dashed curve represent the canonical ensembles calculated by the re-weighting method at 300 K and 700 K, respectively.

energy distribution is observed, as shown in Fig. 2B. In the SA method, in contrast, it is difficult to confirm the completion of the structural search.

Given that the re-weighted canonical ensemble at 300 K provides the putative loop conformation in the native protein, this ensemble was further examined by clustering the conformations using a principal component analysis (PCA) and Ward's method. Here, 1000 model structures were randomly selected from the canonical ensembles at 300 K. Also, each structure was assigned to several clusters by the Ward's clustering method. In this clustering, all the principal components were used to calculate the Euclidean distance among the structures.

Application to modeling of R8 and R12 loops in RNase A

Figure 3A shows the results of the PCA analysis projected onto two-dimensional (2D) space with the first and second principal axes for the Loop R8 ensemble with the NMR (PDB code, 2AAS) and X-ray crystal (8RAT) structures. The contributions from the first, second, third, fourth and fifth axes correspond to 64.2%, 19.9%, 4.9%, 3.4% and 1.5% of the total components, respectively. Three different clusters, 'R8 α ', 'R8 β ' and 'R8 γ ', were clearly observed,

Figure 3 (A) PCA of the canonical ensemble of Loop R8 at 300 K, projected on the first and second principal axes. The blue, cyan, green, orange and red dots correspond to structures of the R8 α cluster, the R8 β cluster, the R8 γ cluster, the 32 NMR structures (2AAS) and a crystal structure (8RAT), respectively. The filled black circle, filled black triangle and filled black square correspond to center structures of the R8 α , R8 β and R8 γ clusters, respectively. (B) PCA for the canonical ensemble of Loop R12 at 300 K, projected on the first and second principal axes. The blue, cyan, orange, and red dots correspond to structures of the R12 α cluster, the R12 β cluster, 32 structures of NMR (2AAS), and a crystal structure (8RAT), respectively. The filled black circles and filled black triangles correspond to center structures of the R12 α and R12 β clusters, respectively.

Table 1 Structural differences among center structures and experimental structures for Loop R8 and R12 in RNase A

Loop R8 in RNase A			
	R8 α cluster	R8 β cluster	R8 γ cluster
<RMSDn> ¹⁾ (Å) of center structure	0.93	6.27	4.32
RMSDx ²⁾ (Å) of center structure	1.20	6.48	4.02
Conformation ratio (%)	38.8	35.4	25.8
Loop R12 in RNase A			
	R12 α cluster	R12 β cluster	
<RMSDn> ¹⁾ (Å) of center structure	2.00	7.75	
RMSDx ²⁾ (Å) of center structure	1.65	7.37	
Conformation ratio (%)	78.5	21.5	

¹⁾ The value is the average of 32 RMSDs, each of which is the RMSD of the heavy atoms in Loop R8 between the center structure of the cluster and one of the 32 NMR structures (PDB code, 2AAS).

²⁾ Center structures in the model ensemble vs. the X-ray crystal structure (8RAT).

and the major cluster, R8 α , coincided with the NMR and X-ray crystal structures with <RMSDn>=0.93 Å and RMSDx=1.20 Å, as shown in Table 1. Here, <RMSDn> is the average of the 32 RMSD values, each of which is the RMSD of the heavy atoms in Loop R8 between a current model structure and one of the 32 NMR structures (2AAS). RMSDx represents the RMSD value of the heavy atoms in Loop R8 between the X-ray crystal structure (8RAT) and the modeled structure.

Since our method provides an ensemble of loop conformations, it is useful to have a representative structure for each structural cluster in the ensemble. For that purpose, we initially calculated an average loop structure for each cluster using the backbone heavy atoms in the loop. Following this, a loop structure was selected as the ‘center structure’, which possesses the smallest RMSD value from the aforementioned average loop structure following structural superpositioning of the backbone atoms of the three preceding residues before the loop and the three residues after the loop. The center structures are shown in Fig. 3A. Using the center structure, we further calculated the RMSDc, which represents the RMSD value of the heavy atoms in Loop R8 between the center structure of each cluster and the model structure. The correlation between RMSDc and RMSDx in the structural ensemble for the R8 α cluster was very high, with a Pearson correlation coefficient of 0.892. For other clusters, there was little or no correlation. This suggests that the center structure of the R8 α cluster may be a good representative of the cluster. The three center structures of the individual clusters are shown in Fig. 4A with the X-ray crystal structure.

Figure 3B shows the same PCA analysis for the longer flexible loop system in RNase A designated as Loop R12, also shown in Fig. 1. Loop R12 is composed of 12 residues from 13 to 24, and 45 surrounding residues. The contributions from the first, second, third, fourth and fifth axes correspond to 50.9%, 20.3%, 9.7%, 3.6% and 3.0% of all

Figure 4 Stereo drawings of Loop R8 and R12 in RNase A. (A) Loop R8: The backbone of the X-ray crystal structure of bovine RNase A (8RAT) is shown in red. Blue, cyan and green backbones correspond to center structures of the R8 α , R8 β and R8 γ clusters, respectively. The gray backbone represents the structure with the smallest RMSDx (0.72 Å) in R8 α . (B) Loop R12: The backbone of the X-ray crystal structure of bovine RNase A (8RAT) is shown in red. Blue and cyan backbones correspond to center structures of the R12 α and R12 β clusters, respectively. The gray backbone represents the structure with the smallest RMSDx (1.50 Å) in R12 α .

the components, respectively. There are only two clusters, ‘R12 α ’ and ‘R12 β ’, and the major cluster, R12 α , also coincided with the experimental conformations with <RMSDn>=2.00 Å and RMSDx=1.65 Å, as shown in Table 1. Here, RMSD values were calculated for the heavy atoms in Loop R12, in the same manner as those for Loop R8. The correlation between RMSDc and RMSDx in the structural ensemble for the R12 α cluster was also very high, with a Pearson correlation coefficient of 0.965. The two center structures and the X-ray crystal structure are shown in Fig. 4B.

The two aforementioned flexible loops have been intensively investigated using several other modeling techniques^{16,48–51}, essentially by looking for the optimum loop conformation that has the lowest potential energy. For Loop R8, the 8-residue loop from 64 to 71 in RNase A, the modeled structures were similar to the X-ray crystal structure with small RMSD values of less than or approximately 1 Å. For Loop R12, the longer 12-residue loop from 13 to

Figure 5 PCA of the ensemble of Loop M20 in DHFR at 300 K, projected on the first and second principal axes. The blue, cyan, green and yellow dots correspond to the M20 α , M20 β , M20 γ and M20 δ clusters, respectively. The filled red circles, open red circles and red circles with a cross correspond to crystal structures of the *closed* form with NADP⁺ (PDB code, 1RX2), the *open* form with NADP⁺ (1RA2) and the *occluded* form without NADP⁺ (1RX5), respectively. The other 34 crystal structures are indicated by filled orange circles.

24, most of the modeled structures differed from the X-ray crystal structure with RMSD values larger than 2 Å. These results are comparable to our current values of $\langle \text{RMSD}_n \rangle$ and RMAD_x for the clusters of Loop R8 α and R12 α as indicated in Table 1. Our method does not predict a single model structure, but creates a structural ensemble consisting of structures that are stable at a given temperature from the free energy landscape. Recently, Lindorff-Larsen et al. showed that protein structures should be determined simultaneously with their dynamics by demonstrating the considerable conformational heterogeneity of the human ubiquitin molecule as determined by the NMR and X-ray methodologies²⁰. This approach agrees well with our modeling concept. In particular, overlaps are always observed between one of the clusters in the structural ensemble of the current method and those derived from NMR analysis.

Induced-fit of a flexible loop in DHFR

Loop modeling of the DHFR M20 loop

We generated a structural ensemble of the M20 loop system designated as Loop M20, starting from the X-ray crystal structure of apo-form DHFR (5DFR). Loop M20 is composed of the most flexible loop region from residue 14 to 25 with the surrounding 44 amino acid residues, as indicated in the Materials and Methods section.

From the canonical ensemble of Loop M20 at 300 K, 2000 structures were randomly selected, and the structures in the ensemble were subjected to PCA with several crystal structures. Figure 5 shows the results of the PCA projected

onto 2D space with the first and second principal axes. The contributions from the first, second, third, fourth and fifth axes correspond to 46.1%, 16.2%, 8.7%, 5.6% and 4.8% of all the components, respectively. Following Ward's method, the ensemble was separated by the four clusters M20 α , M20 β , M20 γ and M20 δ , as shown in Fig. 5. The probability of the conformations in the M20 α , M20 β , M20 γ and M20 δ clusters is 65.4%, 21.3%, 10.3% and 3.1%, respectively.

The center structures of the four clusters and the *closed* loop crystal structures including the cofactor, Nicotinamide Adenine Dinucleotide Phosphate (NADP⁺), and the ligand are shown in Figs. 6A and B. The backbone of the center structures of M20 α and M20 γ penetrated the ligand pocket. The backbone of the M20 β center structure made slight contact with the ligand, similar to the case of the M20 *closed* loop that contacted the ligand. The M20 δ center structure is located far from the active site, and did not interact with the ligand at all. The RMSD_x values of the four center structures for the heavy atoms in Loop M20 are listed in Table 2.

As shown in Fig. 5, the conformation of the M20 β cluster is widely distributed on the 2D plane by the first and the second principle axes. We attempted to search for the experimental ligand-bound loop conformations in the M20 β cluster. Consequently, we found that one of the model structures in M20 β shown in Fig. 7A (designated as M20 β 1) is comparable to the *closed* loop of the X-ray crystal structure (1RX2) with a RMSD value of 1.72 Å. In the same way, the M20 β 2 loop shown in Fig. 7B and the M20 β 3 loop shown in Fig. 7C are very similar to the *open* and *occluded* loops of

(A)

(B)

Figure 6 (A) Structural overview of DHFR: The magenta pipe (residues 1–13 and 26–149) of DHFR (1RX2) was drawn using INSIGHT II (Accelrys Inc., San Diego, CA, USA). The thick red pipe, dotted orange sphere and dotted gray sphere correspond to residues 14–25 of the M20 loop of DHFR, NADP⁺ and folate, respectively. (B) The center structures for Loop M20: Stereo drawings of the ligand-free M20 loop and the *closed* M20 loop of ternary complex conformations. The blue, cyan, green, yellow and red pipes correspond to center structures of the M20 α , M20 β , M20 γ and M20 δ clusters, and the X-ray crystal structure (1RX2), respectively. The dotted orange sphere is NADP⁺, and the dotted gray sphere is folate. The loop of the X-ray crystal structure of the *closed* form is close to NADP⁺.

the X-ray crystal structures (1RA2 and 1RX5) with RMSD values of 2.12 Å and 3.10 Å, respectively. Here, the RMSD values were calculated between the heavy atoms of the modeled flexible loop regions from residue 14 to 25 and those in the X-ray crystal structures, where the X-ray crystal structures of DHFR backbones were superimposed on the backbone of the current template structure. Thus, in the current structural ensemble, we can find structures similar to several different loop conformations that were observed with the ligand in the X-ray crystal structures, even when the conformational search was made for the apo-form DHFR starting from a very different initial loop conformation.

Table 2 Structural differences among center structures and experimental structures for Loop M20 in DHFR

	Loop M20 clusters			
	α	β	γ	δ
RMSDx (Å) for closed loop (1RX2)	3.78	3.00	5.92	8.75
RMSDx (Å) for open loop (1RA2)	5.78	4.68	7.04	8.10
RMSDx (Å) for occluded loop (1RX5)	6.75	5.51	6.08	6.78
Conformation ratio (%)	65.4	21.3	10.3	3.1

Identification of the ligand-bound conformation based on a ligand-docking study

Our modeling method provided several clusters of loop conformations. However, the PCA and cluster analysis do not show which structure can bind the ligand, since a structure of the major cluster is not necessarily in a complex with the ligand. In an effort to identify the actual bound conformation of the flexible loop, we made further docking simulations with the ligand starting from the four center structures of the clusters, and then made a comparison of the complexed conformations.

First, the reliability of the application of the docking study to the current problem was examined. When the ligand-bound forms of the crystal DHFR structures including the observed loop conformations were used as the test templates, the experimental ternary complexes with the cofactor NADP⁺ (1RX2 and 1RA2) and the binary complex without NADP⁺ (1RX5) were well reproduced by the docking simulation, with the RMSD values of the ligand being 1.10 Å, 1.06 Å, and 1.63 Å, respectively. For 1RX2 and 1RA2, the ligand is folate, while for 1RX5, it is 5,10-dideazatetrahydrofolate. In the docking for 1RX2 and 1RA2, NADP⁺ was included in the model and the position of NADP⁺ was based on the coordinates of the X-ray crystal structure. Thus, the reproduced complexed structures show that our docking simulation should work well to build the complex models in the current system.

Next, we analyzed the docking of folate to the DHFR, whose M20 loop conformations were built with the cofactor NADP⁺ as described in the previous section. Sawaya et al. have suggested that NADP⁺ binds to DHFR, and then folate binds to the binary complex²². The positions of NADP⁺ in the model structures were estimated by superimposing the peptide backbone heavy atoms of the template apo-form protein from 5DFR on the backbone heavy atoms of 1RX2, which includes the NADP⁺ molecule. After optimization of binary complexed structures with NADP⁺ by energy minimizations to fix all of the heavy atoms of DHFR, the folate ligand was docked to the four modeled binary complexes, respectively. Table 3 shows the results of docking simulations with the docking scores and the RMSD values of the folate ligand against the two crystal structures, 1RX2 and 1RA2.

For the M20 α and M20 γ center loop structures, the docking scores were rather high, -3.99 and -3.45 , respectively. The docking score is a measure of the binding energy, with

Figure 7 Stereo drawings of the modeled M20 loop conformations in DHFR. (A) The X-ray crystal structure of the M20 loop in the *closed* form (1RX2) is shown by a red backbone and orange side-chains. A blue backbone and cyan side-chains indicate the structure with the smallest RMSDx (1.72 Å) with 1RX2, designated as Loop M20β1 in the M20β cluster. (B) The X-ray crystal structure of the M20 loop in the *open* form (1RA2) is shown by a red backbone and orange side-chains. A blue backbone and cyan side-chains indicate the structure with the smallest RMSDx (2.11 Å) with 1RA2, designated as Loop M20β2 in the M20β cluster. (C) The X-ray crystal structure of the M20 loop in the occluded form (1RX5) is shown by a red backbone and orange side-chains. A blue backbone and cyan side-chains indicate the structure with the smallest RMSDx (3.10 Å) with 1RX5, designated as Loop M20β3 in the M20β cluster.

a lower score corresponding to higher affinity. Consequently, the high docking scores suggest low affinity of the folate ligand for the M20α and M20γ center loop structures. In addition, the docked structures show different binding schemes from the complexed X-ray crystal structures. In the *closed* M20 loop crystal structure (1RX2), four hydrogen bonds and a hydrophobic interaction between DHFR and the folate ligand are observed, in addition to hydrophobic interactions between NADP⁺ and the ligand (Fig. 8A). The *open* M20 loop crystal structure (1RA2) shows a binding

mode almost identical to that of the *closed* loop structure, but fewer interactions between the M20 loop and the ligand are observed (Fig. 8B)²². In contrast, the predicted ternary structure in Fig. 8C has quite a different hydrogen bond network between DHFR and folate from that of the *closed* and *open* loops. Here, the docking scheme of folate bound to DHFR with NADP⁺ is displayed for the M20α center structure. The corresponding RMSD values of the folate against the X-ray crystal structure in 1RX2 and 1RA2 were very large, 4.34 Å and 4.25 Å, respectively.

Table 3 Results of docking simulations for the DHFR model structures with the flexible Loop M20 in DHFR

Template structure	Docking score	RMSD (Å) ¹⁾
M20 α center structure	−3.99	4.34/4.25
M20 β center structure	−4.37	1.35/1.09
M20 γ center structure	−3.45	8.30/8.48
M20 δ center structure	−4.80	2.61/2.47

¹⁾ The RMSD values between the docked folate ligand in the center structures of the model ensemble and those in the *closed/open* M20 loop crystal structure (1RX2/1RA2), respectively.

The docking score for the M20 β center structure was very low, −4.37. In the docking scheme for the optimized complexed structure in Fig. 8D, four hydrogen bonds were reproduced between DHFR and folate as well as those in the *closed* and the *open* M20 loop crystal structures. The hydrophobic interactions between the ligand folate and the side-chains of Ile5, Ala6, Met20, Leu28, Phe31, Ile50, Leu54, and Ile94 were also reproduced. The RMSD values of the folate in 1RX2 and 1RA2 were small, 1.35 Å and 1.09 Å, respectively. Although the RMSD value of folate is smaller for the *open* M20 loop than for the *closed* loop, the Loop

M20 β center structure clearly resembles the *closed* loop, as indicated in Fig. 5 and Table 2. Thus, the current simulation could predict the complexed structure with the *closed* M20 loop.

In the case of the M20 δ center structure, although the docking score was the lowest, −4.80, Loop M20 δ extended toward the outside of the protein and the folate ligand would not be trapped by the M20 δ center structure.

In summary, the ensemble of the disordered M20 loop conformations of ligand-free DHFR includes four stable structural clusters. The loop conformation of the most stable cluster, M20 α , did not correspond to the ligand-bound conformation as suggested from the virtual docking simulation. In contrast, the second most stable cluster, M20 β , has the ability to bind the ligand. From the similarity between the experimental flexible loop conformations and the current structural ensembles in the case of Loop R8 and R12 in RNase A, the M20 α cluster may correspond to the major conformations of the ligand-free M20 loop in DHFR. Though each cluster is thought to be identified, the quantitative probabilities of each cluster could have some uncertainties,

Figure 8 The ligand-binding pocket of DHFR and the folate molecule generated from a docking simulation with NADP⁺ and structural optimization. The dotted lines correspond to intermolecular hydrogen bonds. Semi-circles represent the protein residues involved in the hydrophobic interaction. Numbers in italics represent the interatomic distance between the interacting heavy atom pairs (Å). (A) Binding pocket in the X-ray crystal structure of the *closed* form of the M20 loop in DHFR, with NADP⁺ and folate (1RX2). (B) Binding pocket in the X-ray crystal structure of the *open* form of the M20 loop in DHFR, with NADP⁺ and folate (1RA2). (C) Binding pocket in the docked structure between the center structure of the M20 α cluster and folate. (D) Binding pocket in the docked structure between the center structure of the M20 β cluster and folate.

because of the incompleteness of the force-field parameters^{52,53} and the implicit water model³².

Conclusion

The current loop modeling method reproduced the experimental structures with high accuracy without any heuristic scoring functions for the 8- and 12-residue loops in RNase A. A structural ensemble of the ligand-free M20 loop in DHFR was also generated, and one of the ensembles with stable structures in the free energy landscape corresponded to the DHFR in a complex with the folate ligand. Docking simulations for the center structure of the M20 loop clusters were performed, and the most probable ligand-bound conformation was ascertained from the docking score. Thus, in the M20 loop of DFHR, we confirmed that the ligand-bound structure resulting from induced-fit was generated as one of the conformations stable in the ligand-free state. The analysis presented in this report may provide the means of predicting induced-fit structural or induced-folding structural changes.

Acknowledgments

Y. S. W. is grateful to Drs. Jae Gil Kim (Boston Univ.) and Narutoshi Kamiya (Kobe Univ.) for comments and discussions. We thank M. S. Takashi Kurosawa (Hitachi East Japan Solutions, Ltd.), Satoshi Kubota (HEJS), and Yoshiaki Mikami (HEJS) for their technical support. This work was supported by grants from the New Energy and Industrial Technology Development Organization (NEDO), and the Ministry of Economy, Trade and Industry (METI) of Japan.

References

1. Spolar, R. S. & Record, M. T., Jr. Coupling of local folding to site-specific binding of proteins to DNA. *Science* **263**, 777–784 (1994).
2. Wright, P. E. & Dyson, H. J. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.* **293**, 321–331 (1999).
3. Osborne, M. J., Schnell, J., Benkovic, S. J., Dyson, H. J. & Wright, P. E. Backbone dynamics in dihydrofolate reductase complexes: role of loop flexibility in the catalytic mechanism. *Biochemistry* **40**, 9846–9859 (2001).
4. Dyson, H. J. & Wright, P. E. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* **6**, 197–208 (2005).
5. Santoro, J., Gonzalez, C., Bruix, M., Neira, J. L., Nieto, J. L., Herranz, J. & Rico, M. High-resolution three-dimensional structure of ribonuclease A in solution by nuclear magnetic resonance spectroscopy. *J. Mol. Biol.* **229**, 722–734 (1993).
6. Higo, J., Ito, N., Kuroda, M., Ono, S., Nakajima, N. & Nakamura, H. Energy landscape of a peptide consisting of alpha-helix, 3(10)-helix, beta-turn, beta-hairpin, and other disordered conformations. *Protein Sci.* **10**, 1160–1171 (2001).
7. Searle, M. S. & Ciani, B. Design of beta-sheet systems for understanding the thermodynamics and kinetics of protein folding. *Curr. Opin. Struct. Biol.* **14**, 458–464 (2004).
8. Greer, J. Model for haptoglobin heavy chain based upon structural homology. *Proc. Natl. Acad. Sci. USA* **77**, 3393–3397 (1980).
9. Ring, C. S., Kneller, D. G., Langridge, R. & Cohen, F. E. Taxonomy and conformational analysis of loops in proteins. *J. Mol. Biol.* **224**, 685–699 (1992).
10. Tramontano, A. & Lesk, A. M. Common features of the conformations of antigen-binding loops in immunoglobulins and application to modeling loop conformations. *Proteins* **13**, 231–245 (1992).
11. Sudarsanam, S., DuBose, R. F., March, C. J. & Srinivasan, S. Modeling protein loops using a phi i + 1, psi i dimer database. *Protein Sci.* **4**, 1412–1420 (1995).
12. Fidelis, K., Stern, P. S., Bacon, D. & Moult, J. Comparison of systematic search and database methods for constructing segments of protein structure. *Protein Eng.* **7**, 953–960 (1994).
13. Minor, D. L., Jr. & Kim, P. S. Context-dependent secondary structure formation of a designed protein sequence. *Nature* **380**, 730–734 (1996).
14. Go, N. & Scheraga, H. A. Ring closure and local conformational deformations of chain molecules. *Macromolecules* **3**, 178–187 (1970).
15. Brucoleri, R. E. & Karplus, M. Prediction of the folding of short polypeptide segments by uniform conformational sampling. *Biopolymers* **26**, 137–168 (1987).
16. Fiser, A., Do, R. K. & Sali, A. Modeling of loops in protein structures. *Protein Sci.* **9**, 1753–1773 (2000).
17. Shirai, H., Nakajima, N., Higo, J., Kidera, A. & Nakamura, H. Conformational sampling of CDR-H3 in antibodies by multicannonical molecular dynamics simulation. *J. Mol. Biol.* **278**, 481–496 (1998).
18. Watanabe, Y. S., Fukunishi, Y. & Nakamura, H. Modelling of third cytoplasmic loop of bovine rhodopsin by multicannonical molecular dynamics. *J. Mol. Graph. Model.* **23**, 59–68 (2004).
19. DePristo, M. A., de Bakker, P. I. & Blundell, T. L. Heterogeneity and inaccuracy in protein structures solved by X-ray crystallography. *Structure (Camb)* **12**, 831–838 (2004).
20. Lindorff-Larsen, K., Best, R. B., DePristo, M. A., Dobson, C. M. & Vendruscolo, M. Simultaneous determination of protein structure and dynamics. *Nature* **433**, 128–132 (2005).
21. Bystroff, C., Oatley, S. J. & Kraut, J. Crystal structures of Escherichia coli dihydrofolate reductase: the NADP⁺ holoenzyme and the folate. NADP⁺ ternary complex. Substrate binding and a model for the transition state. *Biochemistry* **29**, 3263–3277 (1990).
22. Sawaya, M. R. & Kraut, J. Loop and subdomain movements in the mechanism of Escherichia coli dihydrofolate reductase: crystallographic evidence. *Biochemistry* **36**, 586–603 (1997).
23. Falzone, C. J., Wright, P. E. & Benkovic, S. J. Dynamics of a flexible loop in dihydrofolate reductase from Escherichia coli and its implication for catalysis. *Biochemistry* **33**, 439–442 (1994).
24. Li, L., Falzone, C. J., Wright, P. E. & Benkovic, S. J. Functional role of a mobile loop of Escherichia coli dihydrofolate reductase in transition-state stabilization. *Biochemistry* **31**, 7826–7833 (1992).
25. Iwakura, M. & Nakamura, T. Effects of the length of a glycine linker connecting the N- and C-termini of a circularly permuted dihydrofolate reductase. *Protein Eng.* **11**, 707–713 (1998).
26. Ionescu, R. M., Smith, V. F., O'Neill, J. C., Jr. & Matthews, C. R. Multistate equilibrium unfolding of Escherichia coli dihydrofolate reductase: thermodynamic and spectroscopic description of the native, intermediate, and unfolded ensembles. *Biochemistry* **39**, 9540–9550 (2000).

27. Cody, V., Chan, D., Galitsky, N., Rak, D., Luft, J.R., Pangborn, W., Queener, S.F., Laughton, C.A. & Stevens, M. F. Structural studies on bioactive compounds. 30. Crystal structure and molecular modeling studies on the Pneumocystis carinii dihydrofolate reductase cofactor complex with TAB, a highly selective antifolate. *Biochemistry* **39**, 3556–3564 (2000).
28. Rod, T. H., Radkiewicz, J. L. & Brooks, C. L., III. Correlated motion and the effect of distal mutations in dihydrofolate reductase. *Proc. Natl. Acad. Sci. USA* **100**, 6980–6985 (2003).
29. Kim, J. G., Fukunishi, Y., Kidera, A. & Nakamura, H. Determination of multicanonical weight based on a stochastic model of sampling dynamics. *Phys. Rev. E* **68**, 021110 (2003).
30. Still, W. C., Tempczyk, A., Hawley, R. C. & Hendrickson, T. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **112**, 6127–6129 (1990).
31. Hawkins, D. G., Cramer, J. C. & Truhlar, G. D. Parametrized models of aqueous free energies of solvation based on pairwise descreening of solute atomic charges from a dielectric medium. *J. Phys. Chem.* **100**, 19824–19839 (1996).
32. Watanabe, Y. S., Kim, J. G., Fukunishi, Y. & Nakamura, H. Free energy landscapes of small peptides in an implicit solvent model determined by force-biased multicanonical molecular dynamics simulation. *Chem. Phys. Lett.* **400**, 258–263 (2004).
33. Berg, B. A. & Neuhaus, T. Multicanonical ensemble: A new approach to simulate first-order phase transitions. *Phys. Rev. Lett.* **68**, 9–12 (1992).
34. Nakajima, N., Nakamura, H. & Kidera, A. Multicanonical ensemble generated by molecular dynamics simulation for enhanced conformational sampling of peptides. *J. Phys. Chem. B* **101**, 817–824 (1997).
35. Nakamura, H., Katayanagi, K., Morikawa, K. & Ikehara, M. Structural models of ribonuclease H domains in reverse transcriptases from retroviruses. *Nucleic Acids Res.* **19**, 1817–1823 (1991).
36. Tilton, R. F., Jr., Dewan, J. C. & Petsko, G. A. Effects of temperature on protein structure and dynamics: X-ray crystallographic studies of the protein ribonuclease-A at nine different temperatures from 98 to 320 K. *Biochemistry* **31**, 2469–2481 (1992).
37. Bystroff, C. & Kraut, J. Crystal structure of unliganded Escherichia coli dihydrofolate reductase. Ligand-induced conformational changes and cooperativity in binding. *Biochemistry* **30**, 2227–2239 (1991).
38. Ferrenberg, A. M. & Swendsen, R. H. Optimized monte carlo data analysis. *Phys. Rev. Lett.* **63**, 1195–1198 (1989).
39. Richmond, T. J. Solvent accessible surface area and excluded volume in proteins. Analytical equations for overlapping spheres and implications for the hydrophobic effect. *J. Mol. Biol.* **178**, 63–89 (1984).
40. Kinjo, A. R., Kidera, A., Nakamura, H. & Nishikawa, K. Physicochemical evaluation of protein folds predicted by threading. *Eur. Biophys. J.* **30**, 1–10 (2001).
41. Kollman, P. A., Dixon, R., Cornell, W., Fox, T., Chipot, C. & Pohorille, A. The development/application of a ‘minimalist’ organic/biochemical molecular mechanic force field using a combination of *ab initio* calculations and experimental data. in *Computer Simulations of Biomolecular Systems* (van Gunsteren, W. F., Weiner, P. K. & Wilkinson, A. J. eds.) vol. 3, pp. 83–96 (1997).
42. Fukunishi, Y., Mikami, Y. & Nakamura, H. The filling potential method: A method for estimating the free energy surface for protein-ligand docking. *J. Phys. Chem. B* **107**, 13201–13210 (2003).
43. Ryckaert, J. P., Ciccotti, G. & Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *J. Comput. Phys.* **23**, 327–341 (1977).
44. Evans, D. J. & Morriss, G. P. The isothermal/isobaric molecular dynamics ensemble. *Phys. Lett.* **98A**, 433–436 (1983).
45. Fukunishi, Y., Mikami, Y. & Nakamura, H. Similarities among receptor pockets and among compounds: Analysis and application to in silico ligand screening. *J. Mol. Graph. Model.* **24**, 34–45 (2005).
46. Schmidt, M. W., Baldridge, K. K., Boatz, J. A., Elbert, S. T., Gordon, M. S., Jensen, J. H., Koseki, S., Matsunaga, N., Nguyen, K. A., Su, S., Windus, T. L., Dupuis, M. & Montgomery, J. A. General atomic and molecular electronic structure system. *J. Comput. Chem.* **14**, 1347–1363 (1993).
47. Higo, J., Collura, V. & Garnier, J. Development of an extended simulated annealing method: application to the modeling of complementary determining regions of immunoglobulins. *Biopolymers* **32**, 33–43 (1992).
48. Rapp, C. S. & Friesner, R. A. Prediction of loop geometries using a generalized born model of solvation effects. *Proteins* **35**, 173–183 (1999).
49. Xiang, Z., Soto, C. S. & Honig, B. Evaluating conformational free energies: the colony energy and its application to the problem of loop prediction. *Proc. Natl. Acad. Sci. USA* **99**, 7432–7437 (2002).
50. Das, B. & Meirovitch, H. Solvation parameters for predicting the structure of surface loops in proteins: Transferability and entropic effects. *Proteins* **51**, 470–483 (2003).
51. Jacobson, M. P., Pincus, D. L., Rapp, C. S., Day, T. J., Honig, B., Shaw, D. E. & Friesner, R. A. A hierarchical approach to all-atom protein loop prediction. *Proteins* **55**, 351–367 (2004).
52. Ono, S., Nakajima, N., Higo, J. & Nakamura, H. Peptide free-energy profile is strongly dependent on the force field: Comparison of C96 and AMBER95. *J. Comput. Chem.* **21**, 748–762 (2000).
53. Mackerell, A. D., Jr., Feig, M. & Brooks, C. L., III. Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J. Comput. Chem.* **25**, 1400–1415 (2004).
54. Koradi, R., Billeter, M., and Wüthrich, K. MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graph.* **14**, 51–55, 29–32 (1996).