http://www.jstage.jst.go.jp/browse/biophysics

BIOPHYSICS

Vol. 8, pp. 127–138 (2012)
doi: 10.2142/biophysics.8.127

A scoring function based on solvation thermodynamics for protein structure prediction

Shiqiao Du[1], Yuichi Harano[2], Masahiro Kinoshita[3] and Minoru Sakurai[1]

[1]*Center for Biological Resources and Informatics, Tokyo Institute of Technology, Yokohama 226-8501, Japan*
[2]*Institute for Protein Research, Osaka University, Suita, Osaka 565-0871, Japan*
[3]*Institute of Advanced Energy, Kyoto University, Uji, Kyoto 611-0011, Japan*

Received June 27, 2012; accepted July 31, 2012

We predict protein structure using our recently developed free energy function for describing protein stability, which is focused on solvation thermodynamics. The function is combined with the current most reliable sampling methods, i.e., fragment assembly (FA) and comparative modeling (CM). The prediction is tested using 11 small proteins for which high-resolution crystal structures are available. For 8 of these proteins, sequence similarities are found in the database, and the prediction is performed with CM. Fairly accurate models with average Cα root mean square deviation (RMSD) ~ 2.0 Å are successfully obtained for all cases. For the rest of the target proteins, we perform the prediction following FA protocols. For 2 cases, we obtain predicted models with an RMSD ~ 3.0 Å as the best-scored structures. For the other case, the RMSD remains larger than 7 Å. For all the 11 target proteins, our scoring function identifies the experimentally determined native structure as the best structure. Starting from the predicted structure, replica exchange molecular dynamics is performed to further refine the structures. However, we are unable to improve its RMSD toward the experimental structure. The exhaustive sampling by coarse-grained normal mode analysis around the native structures reveals that our function has a linear correlation with RMSDs < 3.0 Å. These results suggest that the function is quite reliable for the protein structure prediction while the sampling method remains one of the major limiting factors in it. The aspects through which the methodology could further be improved are discussed.

Corresponding author: Yuichi Harano, Institute for Protein Research, Osaka University, Suita, Osaka 565-0871, Japan.
e-mail: yharano@protein.osaka-u.ac.jp

Key words: protein structure prediction, solvation thermodynamics, fragment assembly, homology modeling, replica-exchange molecular dynamics, coarse-grained normal mode analysis

Proteins have a specific three-dimensional atomic structure that is connected to their biological function. Determining the structure is essential for understanding the molecular mechanism of their function. At present, X-ray crystallography and nuclear magnetic resonance (NMR) are the experimental methods utilized to provide protein structures in atomic detail. However, the processes through which both methods determine the structure are time consuming and cannot keep pace with the identification of new protein sequences. Currently, over eight million protein sequences are deposited in the UniProtKB/TrEMBL database, but only approximately 50,000 of them are experimentally solved and stored in the Protein Data Bank (PDB). Therefore, a high demand has been placed on closing the gap between protein sequences and structures. Thus, the prediction of protein structure has become a major area of computational biology[1–8].

Knowledge-based prediction of protein structures is one of the most successful methods due to improvements of sequence alignment techniques and enrichment of annotated structures in database[9–11]. If the target protein shares sequence similarity with another protein in the database, homology modeling or comparative modeling (CM)[12–15] can construct its model structure from the experimental three-dimensional structure of the related homologous protein, which serves as a template. However, CM is limited when no sequence similarity is discerned in the database. When de novo protein structure is the prediction target, i.e., no templates are found in the database, the fragment assembly (FA) method proposed by Baker et al.[2,16–19] can be applied because it requires short templates of secondary structure fragments. However, the prediction accuracy of FA is limited to proteins with residue lengths shorter than approximately 100 due to its high computational cost[4]. These two heuristic methods, which dominate the Critical Assessment of Techniques for Protein Structure Prediction (CASP), often lack the ability to reliably identify near-native conformations if the target is beyond the limits of the application. Attempts have been made to improve the accuracy of predicted models built with such methods by combining physics-based approaches, in which the models are subjected to molecular dynamics (MD) simulation in an effort to improve their RMSD to the experimental structure. The MD approaches often show high performance in refining the structures generated by CM or FA[20,21] if the generated model is already in the near-native region (~1–3 Å in RMSD). Protein folding simulations using MD also demonstrate agreement with experimental investigations[22,23]. Although their applications still remain limited to small proteins or peptides due to the associated computational cost, such simulations have established the validity of physic-based approaches. However, as Zhang noted in his review[4], no atomic potential could distinguish the near-native structures from the more distant non-native structure because the energy of the best near-native structure was almost always higher than some of the non-native ones. This propensity of the physics-based energy functions is fatal when applying such energy functions, especially for modeling de novo protein structure. One reason for such a fault could be that the current treatment for the physics-based energy, which mainly considers interatomic interactions from a microscopic perspective[24–26], is inadequate for describing protein stability, which should be evaluated by thermodynamic quantities.

Recently, we developed a free energy function[27–31] for an all-atomic protein model to describe the protein folding thermodynamics. In our studies, we rely on the thermodynamics hypothesis[32], which states that proteins in their native configuration are in thermodynamic equilibrium with their solvent environment. Based on this paradigm, the native structure of a protein can be predicted as the global minimum of the free energy surface of the system, including protein and solvent[33]. Our previous study using statistical mechanics on the liquid state revealed that the solvation entropy is the key factor that stabilizes protein structure rather than direct intramolecular interaction within a protein[34]. The protein folding process, which is accompanied by significant conformational entropy loss, is originated from the translational entropy gain of environment water mainly due to the decrease in the excluded volume for water[27,35]. Based on those theoretical considerations, we have constructed a simple scoring function for an all-atomic description with minimal complexity of energy terms that still extracts the physical essence of the protein structure stability. In our previous studies, we found that this function succeeded in selecting native structures correctly from online-available decoy sets[30,31,35].

In this study, we investigate the availability of our scoring function in the typical prediction protocols (CM and FA) by checking whether there is an improvement of the accuracy in the protein structure prediction. Our scoring function is defined with atomic coordinates. Therefore, it can easily be applied to the structural refinement methods for an all-atomic molecular model such as replica exchange molecular dynamics (REMD)[36]. Furthermore, we also examine the prediction performance of the function in the near-native region (~1.0–3.0 Å in RMSD) by combining it with coarse-grained normal mode analysis (CGNMA), which is suitable for structural sampling around a fixed structure. By employing such a variety of methods, which cover a wide range of the conformational space, we can examine the performance of the function for both structural prediction and structural refinement.

Methods and Concept

Target proteins

To evaluate our prediction results, the targets for this study are selected among the proteins from previous CASPs (http://predictioncenter.org/). We also added some targets that have not appeared in previous CASPs, so the selected proteins have different residue lengths and various folding patterns of secondary structures. Furthermore, our previous works using the decoy sets showed that there are some limitations on selecting protein species due to physicochemical and technical reasons[30]. Thus, we selected target proteins according to following criteria. (1) The structures are obtained by X-ray crystallography. We excluded NMR structures due to their conformational ambiguity. NMR structures are usually deposited in the PDB as a collection of models that satisfy geometrical restrains from experiment. Typically, the models are composed of a structurally conserved core regions and variable region (loops and chain terminals). The variable region may cause structural difference as large as 5 Å in Cα RMSD among the models. The final structures are computationally determined through the energy optimization of the force fields. Therefore, the results may largely depend on both the optimization procedure and the force fields. Because our goal is to compare the native energy with the energies of the decoy structures, it is crucial to have one, well-defined native structure with the least bias in favor of other force fields. (2) The structure does not contain heme or metal ions that maintain the native fold. Such selections are justified by the fact that we cannot include crystallization partners in the calculation; most of the structures that co-crystallize with large partners are not suitable for prediction using physics-based scoring functions. In our theoretical treatment, the scoring function assumes that a

Table 1 Properties of the 11 proteins used to test the prediction protocol and the score

PDB	N_{res}	Description	SCOP	Resolution
1whz	70	Hypothetical protein	α+β	1.52
1ttz	75	Unknown Function	α/β	2.11
1ptf	87	Phosphotransferase	α+β	1.60
2he4	90	Unknown Function	not classified	1.45
1s12	94	Unknown Function	α+β	2.00
2hd3	96	Ethanolamine Utilization Protein	all β	2.40
2ivy	101	Hypothetical protein	α+β	1.40
1tr0	106	Plant Protein	α+β	1.80
3dcx	117	Unknown Function	all β	2.00
2hng	127	Hypothetical protein	α+β	1.63
1hka	158	Transferase	α+β	1.50

The items listed include the PDB entry name, total number of residues (N_{res}), description of the biological and source of the protein, SCOP secondary structure class, experimentally determined resolution in Å.

protein is in water as a solvent at infinite dilution. (3) The benchmark proteins have less than 200 residues. This criterion is set simply for convenience in testing the prototype of our prediction protocol and scoring function. The structural properties of the proteins are listed in Table 1. Based on the criteria and the purpose of study described above, the number of target proteins ends up limited to 11. However, the prediction performance of our scoring function can be evaluated by exhaustive structural sampling with a wide variety of methods, as described in the following section.

Prediction protocol

Our prediction protocol consists of two phases: model generation and selection. For model generation, we use either CM or FA to generate a number of candidates for evaluation. If the sequence similarity is found after the alignment procedure using PSI-BLAST[13], we use CM. Otherwise, FA is employed for generating model structures. Those prediction methods do not give an all-atomic coordinates of a protein. Thus, energy minimization using molecular mechanics (MM) is performed for every generated structure to remove unrealistic steric overlaps between atoms before the evaluation with our scoring function. For further structural refinement, conformational sampling is performed using REMD starting from the model structure obtained from the CM or FA.

During model selection, the energies for all the generated models are calculated with our scoring function called "F_{solv}," which is based on solvation thermodynamics, and the model that gives the lowest value is picked up as a predicted structure from a given set of generated decoy structures. The details of the prediction protocol and the score function are described in the following subsections.

Model Generation

Here, the procedure of model generation is briefly summarized. The amino acid sequence of a target protein is first queried against the NCBI-NR database by PSI-BLAST, and the profile matrix is saved. The profile matrix is used to find homologous sequences from the PDBAA database, which only contains the sequences of known structures. Thus, during the first step, the sequence profile is constructed from the above database, and the templates are consequently searched for using the profile obtained from the sequence database of known structures. Both the NCBI-NR and the PDBAA databases are available in the NCBI repository (ftp://ftp.ncbi.nih.gov/blast/db/). When we perform CM, multiple targets are simultaneously treated if identified by PSI-BLAST. The multiple sequence alignment is constructed with CLUSTALW[37]. Although MODELLER[9] can handle multiple templates to generate the model structures, we limit the number of templates to three[38]. When good templates are found (alignment covers more than 80% of the target sequence), we use MODELLER to carry out CM. In this study, we try to compare the prediction performance with the result of previous CASPs. As long as the current database is used, however, it is inevitable that structure or sequence information leakage occurs. Any templates whose sequences have extremely small *E*-value ($<1.0^{-5}$) are omitted, so the refereed database is as similar to the corresponding CASP as possible. To obtain structural diversity for generated models, we imposed relatively large perturbation on the initial template by randomly moving the *x*-*y*-*z* coordinates of protein atoms. The range of the movement is within ±3.0 Å. Then, simulated annealing MD is carried out starting from those perturbed structures, so any steric overlaps between atoms of a modeled protein can be removed. If no suitable template is found, we use ROSETTA[16] to generate model structures according to the standard FA protocol, where 3 and 9 residue fragments are prepared. For each length and position of fragments, 200 fragments are selected from the torsion library of the ROSETTA package based on sequence information. Starting from the fully extended structure as an initial structure, a model structure is built by inserting fragments iteratively for 10,000 times. For each target protein, the number of structures generated by CM and FA are 3,000 and 30,000, respectively.

Conformational sampling by REMD

We also test whether REMD can be applied for a further structural refinement when combined with F_{solv}. As a structural sampling method, REMD simulation is performed starting from a model structure (with RMSD to native approximately 2.0 Å) generated from CM or FA. A detailed description of the REMD procedure can be found elsewhere[3,36,39,40]. Here, the procedure is briefly summarized as follows. We use the AMBER10 package[41], and the conditions and parameters for REMD followed a previous study by Zhu et al.[40] except that we use the AMBER99SB force field. The protein is first solvated with TIP3P water molecules, and energy minimization is performed. Then, the system is copied to 24 replicas, and their temperatures are set between 280 and 320 K. Each replica is equilibrated by 100 ps MD simulations at each corresponding temperature. The SHAKE algorithm[42] is used to freeze all bonds involving hydrogen atoms. The temperature is controlled with the Langevin temperature control algorithm[43]. The equilibrated protein models are subjected to 5 ns REMD at constant (N, V, T) with exchanges attempted every 1 ps. We retrieve atomic coordinates from the trajectory at every exchange trial. The conformations corresponding to the three lowest temperatures (i.e., T = 280.0, 281.6, and 283.3 K) are subjected to the evaluation with F_{solv}. Thus, the total number of evaluated structures is approximately 15,000.

Near-native models (NNMs)

To evaluate the propensity of our scoring function for a slight structural change from the native fold, we also performed CGNMA for the Cα coordinate determined by X-ray crystallography. Hereafter, a structure with Cα RMSD < 3.0 Å from the native is referred to as a NNM. To generate NNMs, we first performed CGNMA[44] for the native structure. It only needs a single cut-off parameter. If the distance of paired Cα atoms is within that value, these two atoms are connected by a spring. By diagonalizing the Hessian matrix of this model, the vibrational frequencies ω_i and direction vectors $\boldsymbol{v}_i$ (i=1 to 3l-6) are obtained as the eigenvalues and eigenvectors, respectively. Here, l corresponds to the residue number of the protein, and $\boldsymbol{v}_i$ is normalized to have a unit norm. Direction vectors are further weighted with their frequencies

$$v_i' = v_i \times \frac{3l}{\omega_i} .$$

This transformation ensures that the ratio of norms of weighted directions is proportional to the ratio of their thermal fluctuation amplitudes. Because ~90% of the vibrational amplitude is dominated by the first tenth modes, we can randomly perturb the Cα coordinates of the native structure along the weighted vibrational directions as follows:

$$X = X_0 + \sum_{i=1}^{10} \varepsilon_i \times v_i' ,$$

where X and X_0 are the Cα coordinates of the perturbed and native structure, respectively. ε_i is a random number from the uniform distribution, i.e., $Unif(-d,d)$, where d is adjusted, so the final models to have Cα RMSD < 3.0 Å. In this approach, we can sample the conformational space uniformly around the native fold because the eigenvectors correspond to the vibrational directions around it.

Local structural correction

In the CGNMA or the FA-based prediction method, all structures are only described with the position of Cαs. For such coarse-grained models, we reconstruct all-atomic structure with the PULCHRA software package[45]. Before calculating each energy value, steric overlaps between atoms are removed by energy minimizations with the AMBER99SB force field. To reduce computational cost, the energy calculations are made in vacuo. Because the main purpose of energy minimization is to remove unrealistic steric overlaps between the atoms in a protein, the calculation of the electrostatic interaction is also omitted. The convergence criterion for the structural optimization is set at 0.5 kcal/mol/Å in the RMS gradient.

Scoring function and molecular models for proteins and water

In this subsection, the basic concept of our scoring function F_{solv} is explained, and the molecular models for a protein and water molecules are summarized. For more detailed information, our earlier publications should be referred[27–31,35]. Because F_{solv} must be evaluated for a large number of different protein structures generated in accordance with the procedures described above, the F_{solv} per structure should be calculated with low computational cost. The essential roles of water should be accounted for as well. We have developed a judicious method meeting both of these requirements as explained in the following.

Here, we consider the energetics of the system in which a protein with a fixed structure is immersed in water at infinite dilution. The free energy of the system W can be expressed as the sum of the protein intramolecular energy E_{intra} and the solvation free energy μ: $W=E_{\mathrm{intra}}+\mu$. μ consists of the hydration energy E_{hyd} and the hydration entropy S: $\mu=E_{\mathrm{hyd}}-TS$. Then, $W=E_{\mathrm{intra}}+E_{\mathrm{hyd}}-TS$, where T is the absolute temperature. μ is the same irrespective of the protein insertion condition, isobaric or isochoric, and the isochoric condition is much more convenient from a theoretical perspective. "$E_{\mathrm{intra}}+E_{\mathrm{hyd}}$" and S are calculated separately as described below.

A recent study on hydration thermodynamics of proteins has revealed that S changes by less than 5% even when the protein–water electrostatic interactions, which are quite strong, are completely eliminated[46]. This result is quite reasonable because S is determined primarily by the excluded volume effect. Based on the great advantage that S is not significantly influenced by the protein-water interaction potentials, we can model a protein as a set of fused hard spheres. An explicit molecular model for solvent (i.e., not a dielectric continuum model) must be employed when calculating S. Moreover, the details of the polyatomic structure, which have substantially large effects on S, should be taken into account on the atomic level. By utilizing a hybrid of the angle-dependent integral equation theory (an elaborate statistical-mechanical theory for molecular liquids) and the morphometric approach, we have made the evaluation of S possible with minor computational effort, as described later in this subsection. A multipolar model is employed for the water molecule[47], and the effect of the molecular polarizability is taken into account using the self-consistent mean field theory. At the theoretical level, the many-body induced interactions are reduced to pairwise additive potentials involving an effective dipole moment.

"$E_{\text{intra}}+E_{\text{hyd}}$" is evaluated by choosing a fully extended structure as the reference structure. The structure possesses the maximum number of hydrogen bonds with water molecules and no intramolecular hydrogen bonds. We refer to "$E_{\text{intra}}+E_{\text{hyd}}$" as the total dehydration penalty that occurs upon the transition to a more compact structure, in which intramolecular hydrogen bonds are not formed completely. Let ξ denote "$E_{\text{intra}}+E_{\text{hyd}}$". Then, F_{solv} is given by

$$F_{\text{solv}} = -TS + \xi.$$

The hydration entropy, which is strongly dependent on details of the protein polyatomic structure, is calculated using a hybrid of the angle-dependent integral equation theory[48–52], a statistical-mechanical theory for molecular liquids, and the morphometric approach[53].

In the angle-dependent integral equation theory, the roles of water as a molecular ensemble are fully considered. The water-water and solute-water orientational correlations are taken into account in a complete manner. Then, the multipolar model[47] is employed for mimicking a water molecule. The solute-solvent (water) pair correlation function can be denoted by g_{UV} (r, θ, ϕ, χ) where r is the distance from the center of the solute, (θ, ϕ) represents the orientation of the dipole-moment of vector water (θ is the angle between the vector and the solute-water axis), and χ describes the rotation around the dipole-moment vector. The Morita-Hiroike formula for calculating the hydration free energy μ is written as

$$\mu = \frac{k_B T \rho_S}{8\pi^2} \iiiint 4\pi \Big[\frac{1}{2}\{h_{\text{UV}}(r,\theta,\phi,\chi)\}^2 - \frac{1}{2} h_{\text{UV}}(r,\theta,\phi,\chi) c_{\text{UV}}(r,\theta,\phi,\chi) - c_{\text{UV}}(r,\theta,\phi,\chi) \Big] r^2 \sin\theta \, dr d\theta d\phi d\chi,$$

where h_{UV} ($= g_{\text{UV}} - 1$) and c_{UV} are the total and direct correlation functions, respectively. The integral range is $[0, \infty]$ for r, $[0, \pi]$ for θ, and $[0, 2\pi]$ for ϕ and χ. k_{B} and ρ_{S} are the Boltzmann constant and the number density of solvent water, respectively. As mentioned earlier in this subsection, S is considered under an isochoric condition and calculated through the thermodynamic relation

$$S = -\left(\frac{\partial \mu}{\partial T}\right).$$

The temperature derivatives are numerically evaluated from

$$\left(\frac{\partial \mu}{\partial T}\right) = \frac{\mu(T+\delta T)+\mu(T-\delta T)}{2\delta T},$$

where $dT = 5$ K is adopted.

Although employing this theory is computationally expensive, the calculation of hydration entropy for a protein is quite rapid combined and remains in quantitative agreement when combined with the morphological thermodynamics[53]. Using such a combined method, the computation time required per structure is ~0.1 sec on the standard workstation. In the morphometric approach, a hydration quantity such as S is expressed by the linear combination of only four geometric measures of a solute molecule:

$$S = c_1 V + c_2 A + c_3 C + c_4 X,$$

where V, A, C and X corresponds to exclude volume, solvent accessible surface area, integrated mean and Gaussian curvature, respectively[53]. In our approach, the solute shape enters S only via the four geometric measures. Therefore, the four coefficients (c_1–c_4) can be determined through simple geometry. They are determined in advance from the values of S for hard-sphere solutes with various diameters (d_{U}: $0 \le d_{\text{U}} \le 10 d_{\text{V}}$, where d_{V} denotes the diameter of a water molecule) immersed in our model water. The angle-dependent integral equation theory is employed in the calculation. The four coefficients are determined by the least square fitting applied to the following equation for hard-sphere solutes:

$$S = c_1(4\pi R^3/3) + c_2(4\pi R^2) + c_3(4\pi R) + c_4(4\pi) \text{ and}$$

$$R = (d_{\text{U}} + d_{\text{V}})/2.$$

As explained in the text, we can model a protein structure as a set of fused hard spheres whose diameters are the corresponding Lennard-Jones (LJ) parameters taken from the AMBER99SB. Once determined, the four coefficients can be used to calculate S for a protein with any structure. It is obtained just by calculating the four geometric measures for each protein structure using the Connolly algorithm[54]. The x-y-z coordinates of the protein atoms, which characterize each structure on the atomic level, are used as part of the input data for calculating the four geometric measures.

In contrast, to calculate the total dehydration penalty ξ, the protein-water interaction potentials play essential roles. The ξ is evaluated by choosing a fully extended structure, which possesses the maximum number of hydrogen bonds with water molecules and no intramolecular hydrogen bonds, as the reference structure. Compared to the fully extended structure with $\xi = 0$, in a more compact structure some proton donors and acceptors (e.g., N and O, respectively) are buried in the interior after breaking the hydrogen bonds with water molecules (CO ... W, NH ... W, etc.). When a donor and an acceptor are buried in the interior after the hydrogen bonds with water molecules are broken, we impose no penalty if they form an intramolecular hydrogen bond. On the other hand, when a donor or an acceptor is buried and no intramolecular hydrogen bond formed, we impose the penalty of $7k_BT_0$ (T_0=298 K). This value is based on the result obtained by a molecular dynamics simulation performed for hydrogen-bond formation between two formamide molecules in a nonpolar liquid[55]. We examine all the donors and acceptors for backbone-backbone, backbone-side chain, and side chain-side chain intramolecular hydrogen bonds and calculate ξ. It is necessary to determine if each donor and acceptor is buried. The water-accessible surface area is calculated for each of them by means of Connolly's algorithm[54]. If the surface area is smaller than a threshold value A_0, the donor or acceptor is considered buried. A_0 is set at 0.001 Å^2. To determine if an intramolecular hydrogen bond is formed, we use the criteria proposed by McDonald and Thornton[56]. Per protein structure, the computation of the ξ is also finished rapidly. It is now apparent that no dehydration penalty is considered for the nonpolar groups of a protein. The break of hydrogen bonds with water molecules is more serious and forms a principal component of the total dehydration penalty when they are not compensated by intramolecular hydrogen bonding.

Results and Discussion

Predicted models by CM and FA

There are two generally accepted conditions that need to be fulfilled for any given potential to be suitable for structure prediction or refinement. The potential must score the native structure as the lowest in energy value, and the potential energy must be correlated with native-likeness (e.g., RMSD or TM score) to drive the conformational search in the direction of the native structure. In Figure 1, the scattered plot of the score is shown. The value of F_{solv} normalized with respect to the number of residues is plotted against the Cα-RMSD for all the generated structures of a targeted protein sequence. For all cases, the energy of the native structure is lower than that of any other decoy structures. As observed in Figure 1, the native-decoy energy gap, that is the difference between native energy and lowest decoy energy, seems to depend on the proteins and methodologies used for the predictions. The energy gap created by the sampling with FA; in the case of 1whz, 1ttz and 1s12, is wider than that with CM. As summarized in Table 2, FA based on our protocol could not sample the model structures below 2.44 Å in RMSD of Cα. On the contrary, CM based on our protocol gave better predicting results than FA. In the case of 1ptf, the best-modeled structure obtained by CM has an RMSD of 0.93 Å and is closest to the native. Therefore, the energy gap observed occurs because the FA sampling is unable to better reach the structural region around the native compared to the CM sampling.

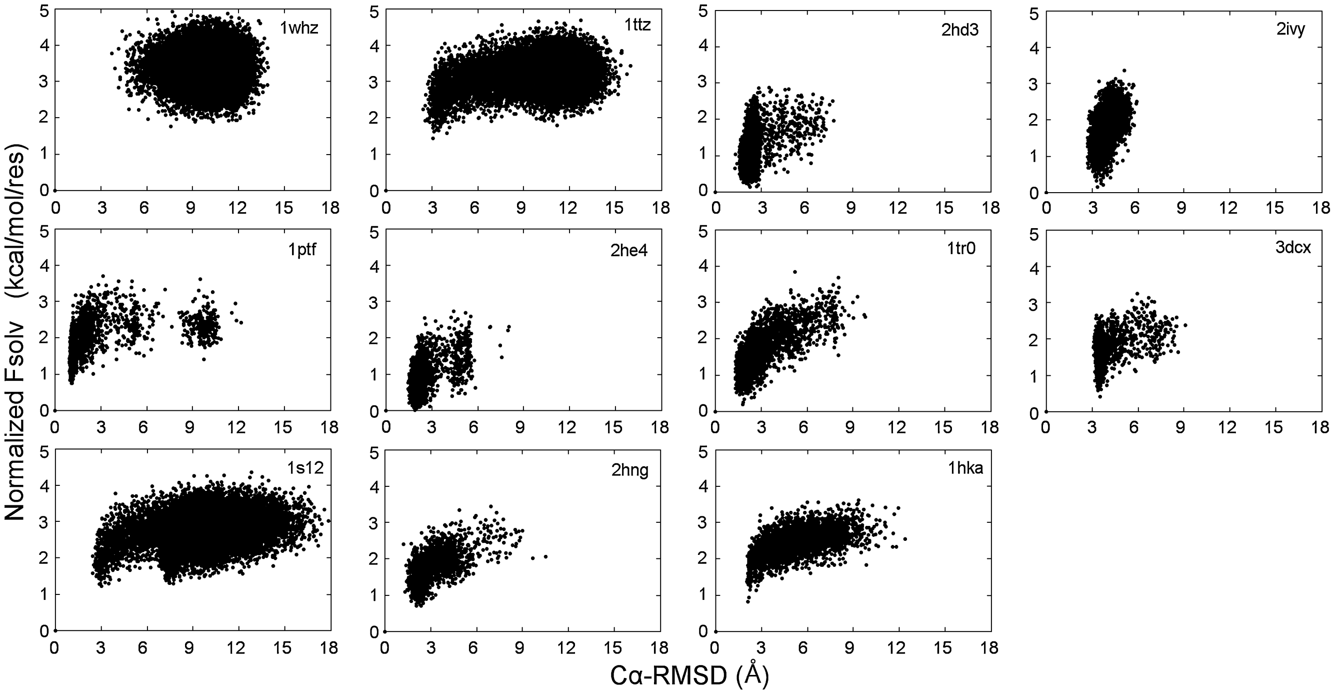

Figure 1 The plot of F_{solv} as a function of Cα-RMSD for the generated models. All targeted proteins are presented. The X-axis is the RMSD of the decoy structures from the native. The Y-axis is the corresponding normalized F_{solv} value.

Table 2 RMSD of model structures

PDB	Method	Closest decoy	F_{solv} selected	AMBER99 selected	Best in CASPs
1whz	FA	3.70	7.56	11.67	1.58
1ttz	FA	2.45	3.14	4.32	–
1ptf	CM	0.93	1.10	1.07	–
2he4*	CM	1.46	1.95	2.45	0.73
1s12*	FA	2.44	3.09	8.00	2.08
2hd3*	CM	1.30	2.20	2.53	3.78
2ivy	CM	2.71	3.35	3.51	–
1tr0	CM	1.30	1.82	1.98	2.08
3dcx	CM	3.09	3.49	3.51	–
2hng*	CM	1.23	2.06	4.64	5.38
1hka*	CM	2.03	2.11	2.11	6.03

The item listed include the PDB entry name, the method used to predict, minimum RMSD that can be found in the generated models, RMSD of selected by F_{solv}, best RMSD result in previous CASPs (if available). We put asterisk (*) if the model was generated with any structural data either as fragments or templates which had not been published at the time of the corresponding CASP round. The publish date of the protein structures are obtained from the RCSB PDB web page (http://www.rcsb.org/pdb/home/home.do).

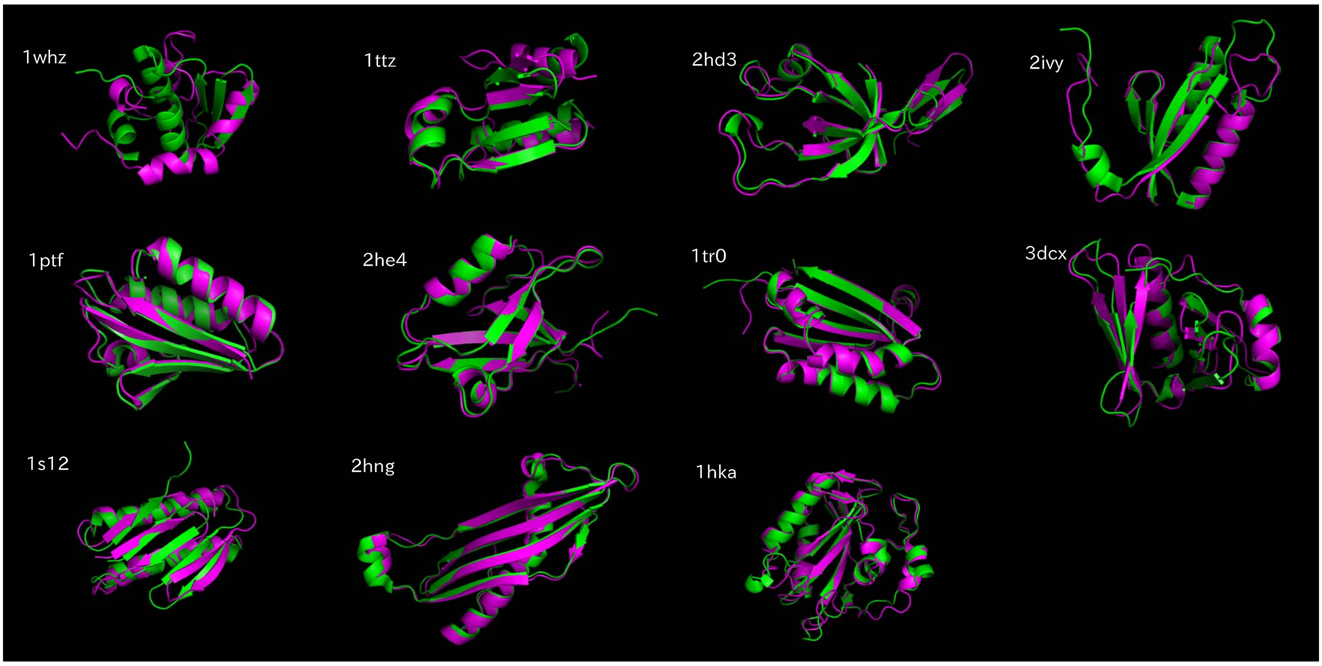

Figure 2 The superposition of the native and the predicted structure for all proteins. The native and predicted structures are respectively colored in green and purple.

The RMSDs of the best structures evaluated by our scoring function are listed in Table 2. The best results in the previous CASPs are also listed. Compared to the previous prediction results, our function provides fairly accurate model structures. In Figure 2, the structures predicted by our protocol and scoring function are overlaid with the corresponding experimentally determined structures. Although we cannot make a direct comparison with the previous CASP results because the database we used in this study has been already updated, the more accurate results are obtained by the standard prediction methods combined with F_{solv} for the cases of 1hka, 2hng, 1tr0 and 2hd3. Some structures have lower energy than the closest structure in RMSD for every case. However, the cases of 1ptf, 1hka, 1tr0 and 1ttz can be regarded as successful because F_{solv} selects fairly good models when the structural diversity of their decoys is taken into consideration. In those cases, F_{solv} exhibits a significant correlation with RMSD. On the other hand, when native-like conformations are hardly sampled, i.e., in the case of 1whz, F_{solv} does not select a relatively closer model from among the whole decoy structures. In such cases, there is no observed correlation of the score with RMSD. In both cases, there seems to be a barrier against achieving a structure similar to native structures with RMSD < 2.0 Å. One explanation for this barrier could be the bias for generated structures due to the difference in prediction methodologies employed here. The fact that all the structures for the evaluation are generated with the help of MM to avoid steric overlap between atoms in a protein may also explain this discrepancy. Such structures are not always optimized in terms of F_{solv}.

Generally, CM can generate quite native-like structures (RMSD < 3.0 Å). 1ptf is one of the most successful cases. The generated model that is closest to the native has an RMSD of 0.93 Å, and the model selected by F_{solv} has an RMSD of 1.11 Å. Obviously, the structural differences between the native structure and those two decoys are mainly in the loop domain. The α-helix and β-sheet elements are arranged properly in the tertiary structure of the predicted decoy. Therefore, F_{solv} correctly select the native-fold pattern from the other decoys. The structural diversity of decoys generated by FA is significantly larger than that of decoys generated by CM. FA covers the structures that have RMSDs of 3 to 18 Å. In the case of 1s12, two basins appear in the F_{solv}-RMSD profile. One basin is around an RMSD of 3.0 Å, and the other is around an RMSD of 7.0 Å. The energy gap between the two seems to be very small; thus, it is hard to distinguish them. However, for 1s12, the lowest energy in the former basin is lower than that of the latter, and we are able to retrieve a structure with an RMSD of 3.1 Å as a predicted structure. For 1whz, conformations with RMSDs < 3.7 Å are not sampled, and F_{solv} fails to select the best decoy and instead selects the structure with 7.56 Å in RMSD. Judging from the fact that F_{solv} still identifies the native as the lowest energy structure and that the sampling is not covering the near-native region (RMSD < 3.0 Å), we expect that F_{solv} can select more accurate structures if more aggressive sampling is possible in the near-native region like other successful cases.

It is worthwhile to compare our scoring function with the representative all-atomic potential. As described in the methods section, all the structures are generated through optimization with the AMBER99SB force field. Thus, it is fair to compare the performance of our scoring function with that of the AMBER99SB force field and the generalized Born surface area (BGSA) solvent model because no additional modification to the generated structures is necessary. In Figure 3, the score values of the AMBER99SB/GBSA are plotted against Cα-RMSD for all the target proteins. Except for 2he4, the experimentally determined native structures are obtained as the lowest energy. The RMSD of the best-scored structure is listed in Table 2. Among the structures generated with CM, the AMBER99SB/GBSA identified the structure that is relatively close to the native as the best-scored one, showing good performance along with our scoring function. However, in the structures generated with FA, which covers a wide range of conformational space, a more distant non-native structure is detected as the lowest energy. For 1whiz, the RMSD of the best-scored structure is 11.67 Å. The profile shows that the AMBER99SB/GBSA also cannot capture the funnel-like shape toward the native structure. This failure is due to insufficient structural

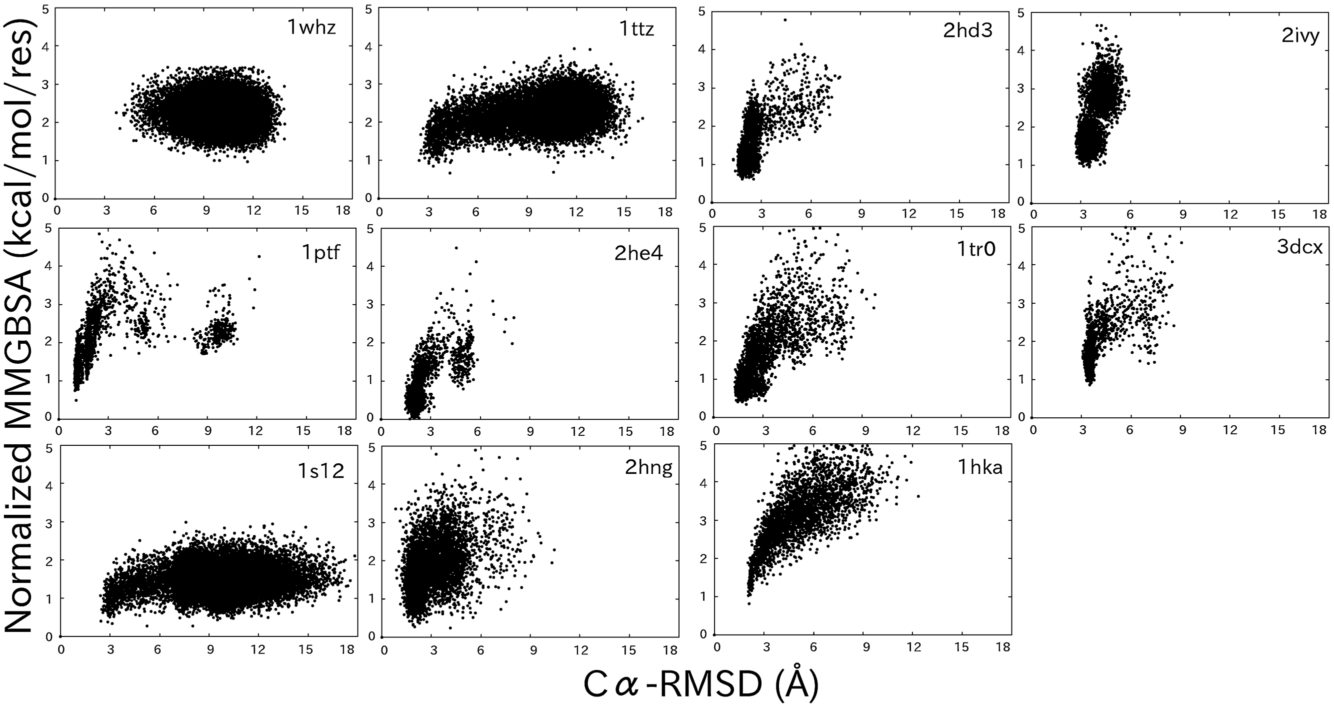

Figure 3 The plot of AMBER99SB/GBSA energy for generated models as a function of the RMSD. The X-axis is the Cα-RMSD of the decoy structures from the native. The Y-axis is the corresponding normalized score value. The values are normalized against the score of the native structure.

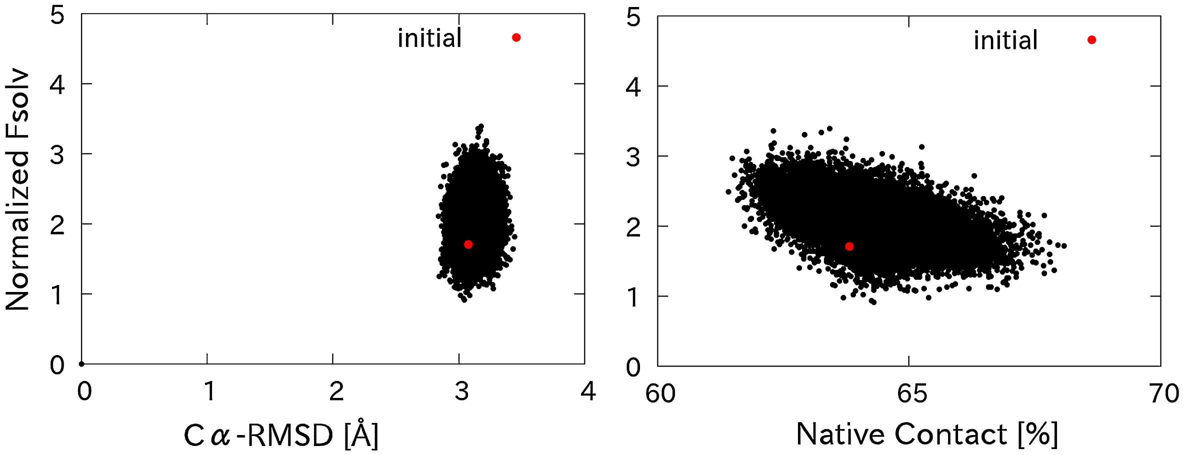

Figure 4 (a) The plot of F_{solv} as a function of the Cα-RMSD for models generated by REMD starting from the best predicted models (PDB ID: 1ttz). The initial structure is colored in red. (b) The plot of F_{solv} as a function of the value of native contact for the models generated by REMD. To compare profiles, normalization is done with the equation ($F_{\mathrm{solv\ (decoy)}} - F_{\mathrm{solv\ (native)}}$)/(residue length).

sampling. On the contrary, the funnel-like shape toward the native structure appears for 1ttz and 1s12, and the generated decoy structure approaches the native. Even for these two cases, the RMSDs of the best-scored structures are 4.32 and 8.00 Å, respectively. As described in the introduction, the drawback in the physics-based all-atomic potentials that Zhang noted[4] is observed for the decoy structures generated with our FA procedure. This observation is also consistent with our previous result that the all-atomic potential energy of the native structure is higher than that of a hypothetically folded single α-helix structure with the identical sequence[31]. This tendency of the scoring function is fatal in the ab initio prediction for the new-fold pattern of proteins.

Table 3 Spearman coefficient between RMSD and F_{solv} value for NNMs

PDB	Spearman coefficient of NNMs
1whz	0.72
1ttz	0.44
1ptf	0.74
2he4	0.17
1s12	0.55
2hd3	0.70
2ivy	0.16
1tr0	0.38
3dcx	0.47
2hng	0.23
1hka	0.43

Structural refinement with REMD

REMD is a well-established method for global structural sampling. In particular, the REMD-based protocol is effective for the refinement of high-quality models of small proteins[40]. Here, using REMD, we test the ability of our function to refine the structure generated from CM or FA. We chose 1ttz as a benchmark protein for this purpose. The initial structure, which is evaluated as the best scored in the CM, has a main chain structure with 3.14 Å for the RMSD. In Figure 4(a), the scatter plot of F_{solv} as a function of RMSD is shown for the structures in the REMD trajectory over a temperature range of 280–283 K. Approximately 15,000 structures are available for the evaluation. The decoy structure that is closest to the native is 2.84 Å away from the native structure; thus, the REMD procedure as reported in the previous study[40] successfully samples the structural space that is slightly closer to the native structure. However, F_{solv} did not have a linear correlation with the Cα-RMSD, and the best-selected structure has an RMSD of 3.04 Å and belongs to the middle region of the whole space sampled with REMD.

In Figure 4(b), the values of F_{solv} for the same structures generated by REMD are plotted against the native contact (NC). Unlike the plot of F_{solv} against the RMSD, F_{solv} has a linear correlation to NC. Here, the Spearman coefficient is employed as the estimator of these correlations. This coefficient measures the monotonicity of two variables, while the Pearson coefficient is generally used to measure their linearity. Because we want to know whether F_{solv} decreases along RMSD or NC, the Spearman coefficient is more appropriate. The Spearman coefficient for the plot of F_{solv} *vs* NC is 0.47, while that for the plot of F_{solv} *vs* RMSD is 0.11. This high correlation observed in the plot of F_{solv} *vs* NC indicates that F_{solv} can still find the native structure through more exhaustive structural sampling. In the near-native region, Cα-RMSD is not always the best estimator of the native-likeness. This result is also consistent with our previous study[57], which showed that the side-chain packing toward the native structure also constrains the whole secondary structure of the native protein.

Although the experimentally determined native structures have the lowest F_{solv} score among any other decoy structures generated here, F_{solv} combined with REMD with MM is not suitable for the structural refinement. As described in the

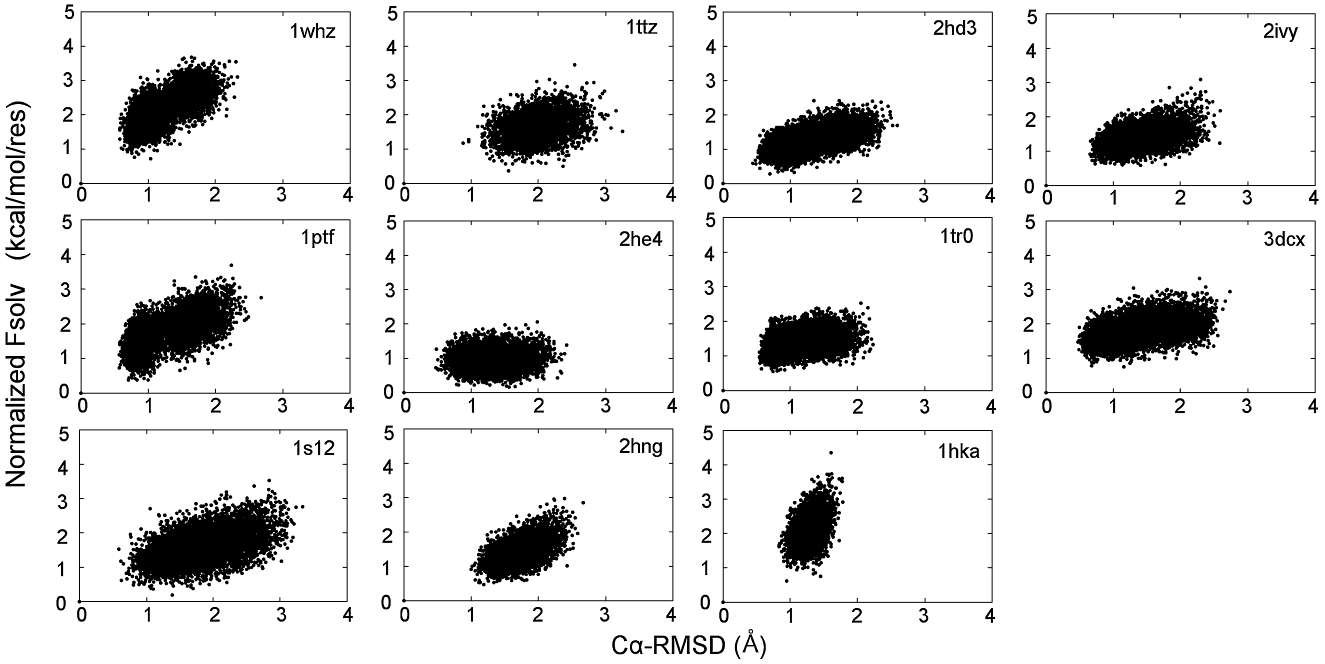

Figure 5 The plot of F_{solv} as a function of Cα-RMSD for NNMs. The values of axes are the same as in Figure 1.

methods section, structural sampling in the REMD is based on atomic forces defined with the MM energy function. Therefore, the structures generated by REMD in the near-native region could not be the best choice for the evaluation by F_{solv}. Applying F_{solv} for molecular dynamics through the formulation of a gradient for atomic positions should be a task for future study.

Near-native models

In the case of 1ttz, our scoring function did not work well with REMD sampling based on MM for the purpose of refining the predicted structural model. However, F_{solv} identified the native structure as the lowest energy structure for all the target proteins. Thus, it is worthwhile to confirm whether this result is caused by conformational sampling or by scoring function itself, and we sampled the near-native region (RMSD < 3.0 Å) using CGNMA. The generated decoys by CM or FA are generally limited to the structural region with an RMSD > 2.0 Å. On the other hand, CGNMA is expected to generate models that maintain the initial backbone conformation. Here, the experimentally determined native structures are the initial conformation.

The scatter plots of F_{solv} against RMSD for 1ttz, which is used for the REMD study, are shown in Figure 5. Unlike the result from REMD (Fig. 4(a)), a significant correlation between RMSD and F_{solv} is observed. NNMs are the conformations that are sampled uniformly along the dominant vibrational directions around the native. Thus, if sampling can be performed sufficiently around the near-native region, F_{solv} can find the native-like structure from among the decoy structures.

We also test F_{solv} in NNMs of the other proteins used in this study. The Spearman coefficient is employed again as the estimator of correlation. For the proteins examined here, we find that F_{solv} has a significant correlation with RMSD (Spearman coefficient > ~0.4) except for 2he4. The observed coefficients also indicate that we can reach the native structure if structural sampling is sufficient around the native structure. Those results clearly show that F_{solv} has a funnel-like potential surface shape for scoring near-native conformations.

For 2he4, however, almost no significant correlation between energy and RMSD is observed (Spearman coefficient = 0.17). The NNMs for 2he4 generated by CGNMA accumulated in the same energy level as observed in Figure 5. As observed from the energy plot obtained by using decoy structures with CM (Fig. 1: 2he4), the energy gap between the decoy structure with the lowest energy and the native is extremely narrow, even though the lowest energy structure is the native structure. If the native structure itself was included in the group of decoy structures in NNMs, that is, if the native had some conformational ambiguity, we could say that F_{solv} was able to score the native correctly. As we noted in the methods section, our scoring function assumes that the protein is isolated in water at an infinite dilution. Therefore, this result suggests that another major factor also stabilized 2he4 when experimentally determining the structure, e.g., a co-solvent (ethanediol) or other protein domains as described in the PDB file.

Conclusion

We perform protein structure prediction using the free energy function based on solvation thermodynamics. If PSI-BLAST found sequence similarity, CM is employed for structural modeling. Then, we have successfully selected fairly accurate predicted models with an average RMSD ~2.0 Å for the small proteins we selected. In several cases, the results we obtain are better than those reported in the previous CASPs. If the target proteins have no sequence similarity in the database, FA, which is currently the most effective structural sampling method for de novo structures, is performed. Although the generated structure that is closest to the native in terms of RMSD does not always have the minimum in energy values within the models generated using FA, the native structures still receive the lowest scores. Therefore, from the viewpoint of selectivity, F_{solv} exhibits strong performance. However, there are still structural differences between the native and predicted models, and more aggressive sampling is expected to generate more accurate models. If the sampling succeeded as in the case of 1ptf, which covered the structural region below 1 Å in RMSD, F_{solv} could correctly identify the native structure.

As is shown in the NNM study, F_{solv} obtains a linear correlation with RMSD in the near-native region (RMSD < ~3 Å). By improving the sampling method to efficiently cover the region like the one created by the NNMs, F_{solv} is expected to successfully refine the more accurate native-like structures. Another point for improvement is how we can incorporate the effectiveness of F_{solv} into the sampling processes with coarse-grained molecular models, which are generally used for FA or CM. Thus, we are now investigating such models and protocols.

Acknowledgments

The development of computer program for the morphometric approach was done in collaboration with Roland Roth. This work was supported by the Program to Disseminate Tenure Tracking System and Grants-in-Aid for Scientific Research on Innovative Areas (No. 20118004 and 20118006) from the Ministry of Education, Culture, Sports, Science and Technology of Japan and by the Grand Challenges in Next-Generation Integrated Nanoscience, MEXT, Japan.

References

1. Baker, D. & Sali, A. Protein structure prediction and structural genomics. *Science* **294**, 93–96 (2001).
2. Simons, K. T., Strauss, C. & Baker, D. Prospects for ab initio protein structural genomics. *J. Mol. Biol.* **306**, 1191–1199 (2001).
3. Liwo, A., Czaplewski, C., Oldziej, S. & Scheraga, H. A. Computational techniques for efficient conformational sampling of proteins. *Curr. Opin. Struct. Biol.* **18**, 134–139 (2008).
4. Zhang, Y. Progress and challenges in protein structure prediction. *Curr. Opin. Struct. Biol.* **18**, 342–348 (2008).
5. Kryshtafovychand, A. & Fidelis, K. Protein structure prediction and model quality assessment. *Drug Discov. Today* **14**, 386–393 (2009).
6. Zhang, Y. Protein structure prediction: when is it useful? *Curr. Opin. Struct. Biol.* **19**, 145–155 (2009).
7. Floudas, C. A. Computational methods in protein structure prediction. *Biotechnol. Bioeng.* **97**, 207–213 (2007).
8. Ginalski, K., Grishin, N. V., Godzik, A. & Rychlewski, L. Practical lessons from protein structure prediction. *Nucleic Acids Res.* **33**, 1874–1891 (2005).
9. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
10. Eddy, S. R. Profile hidden Markov models. *Bioionformatics* **14**, 755–763 (1998).
11. Ohlson, T., Wallner, B. & Elofsson, A. Profile-prorofile methods provide improved fold-recognition: a study of different profile-profile alignment methods. *Proteins* **57**, 188–197 (2004).
12. Sali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815 (1993).
13. Ginalski, K. Comparative modeling for protein structure prediction. *Curr. Opin. Struct. Biol.* **16**, 172–177 (2006).
14. Martí-Renom, M. A., Stuart, A. C., Fiser, A., Sánchez, R., Melo, F. & Sali, A. Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 291–325 (2000).
15. Schwede, T., Kopp, J., Guex, N. & Peitsch, M. C. SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Res.* **31**, 3381–3385 (2003).
16. Simons, K. T., Kooperberg, C., Huang, E. & Baker, D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *J. Mol. Biol.* **268**, 209–225 (1997).
17. Bradley, P., Chivian, D., Meiler, J., Misura, K. M., Rohl, C. A., Schief, W. R., Wedemeyer, W. J., Schueler-Furman, O., Murphy, P., Schonbrun, J., Stauss, C. E. & Baker, D. Rosetta predictions in CASP5: successes, failures, and prospects for complete automation. *Proteins* **53**, 457–468 (2003).
18. Bujnicki, J. M. Protein-structure prediction by recombination of fragments. *Chem. Bio. Chem.* **7**, 19–27 (2006).
19. Chaudhury, S., Lyskov, S. & Gray, J. J. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioionformatics* **26**, 689–691 (2010).
20. Wroblewska, L. & Skolnick, J. Can a physics-based, all-atom potential find a protein's native structure among misfolded structures? I. large scale AMBER benchmarking. *J. Comput. Chem.* **28**, 2059–2066 (2007).
21. Wroblewska, L., Jagielska, A. & Skolnick, J. Development of a physics-based force field for the scoring and refinement of protein models. *Biophys. J.* **94**, 3227–3240 (2008).
22. Adcock, S. A. & McCammon, J. A. Molecular dynamics, survey of methods for simulating the activity of proteins. *Chem. Rev.* **106**, 1589–1615 (2006).
23. Zagrovic, B., Snow, C. D., Shirts, M. R. & Pande, V. S. Simulation of folding of a small alpha-helical protein in atomistic detail using worldwide-distributed computing. *J. Mol. Biol.* **323**, 927–937 (2002).
24. Duan, Y., Wu, C., Chowdhury, S., Lee, M. C., Xiong, G., Zhang, W., Yang, R., Cieplak, P., Luo, R., Lee, T., Caldwell, J., Wang, J. & Kollman, P. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.* **24**, 1999–2012 (2003).
25. Brooks, B. R., Brooks, C. L. 3rd., Mackerell, A. D. Jr.,

Nilsson, L., Petrella, R.J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A.R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., Kuczera, K., Lazaridis, T., Ma, J., Ovchinnikov, V., Paci, E., Pastor, R.W., Post, C.B., Pu, J.Z., Schaefer, M., Tidor, B., Venable, R.M., Woodcock, H.L., Wu, X., Yang, W., York, D.M. & Karplus, M. CHARMM: the biomolecular simulation program. *J. Comput. Chem.* **30**, 1545–1614 (2009).

26. Qiu, D., Shenkin, P.S., Hollinger, F.P. & Still, W.C. The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii. *J. Phys. Chem. A* **101**, 3005–3014 (1997).
27. Harano, Y. & Kinoshita, M. Large gain in translational entropy of water is a major driving force in protein folding. *Chem. Phys. Lett.* **399**, 342–348 (2004).
28. Harano, Y. & Kinoshita, M. Translational-entropy gain of solvent upon protein folding. *Biophys. J.* **89**, 2701–2710 (2005).
29. Harano, Y., Roth, R. & Kinoshita, M. On the energetics of protein folding in aqueous solution. *Chem. Phys. Lett.* **432**, 275–280 (2006).
30. Harano, Y., Roth, R., Sugita, Y., Ikeguchi, M. & Kinoshita, M. Physical basis for characterizing native structures of proteins. *Chem. Phys. Lett.* **437**, 112–116 (2007).
31. Yoshidome, T., Oda, K., Harano, Y., Roth, R., Sugita, Y., Ikeguchi, M. & Kinoshita, M. Free-energy function based on an all-atom model for proteins. *Proteins* **77**, 950–961 (2009).
32. Anfinsen, C.B. Principles that govern the folding of protein chains. *Science* **181**, 223–230 (1973).
33. Lazaridis, T. & Karplus, M. Effective energy functions for protein structure prediction. *Curr. Opin. Struct. Biol.* **10**, 139–145 (2000).
34. Imai, T., Harano, Y., Kinoshita, M., Kovalenko, A. & Hirata, F. Theoretical analysis on changes in thermodynamic quantities upon protein folding, Essential role of hydration. *J. Chem. Phys.* **126**, 225102 (2007).
35. Yasuda, S., Yoshidome, T., Harano, Y., Roth, R., Oshima, H., Oda, K., Sugita, Y., Ikeguchi, M. & Kinoshita, M. Free-Energy Function for Discriminating the Native Fold of a Protein from Misfolded Decoys. *Proteins* **79**, 2161–2171 (2011).
36. Sugita, Y. & Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **314**, 141–151 (1999).
37. Thompson, J.D., Higgins, D.G. & Gibson, T.J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680 (1994).
38. Larsson, P., Wallner, B., Lindahl, E. & Elofsson, A. Using multiple templates to improve quality of homology models in automated homology modeling. *Protein Sci.* **17**, 990-1002 (2008).
39. Mitsutake, A., Sugita, Y. & Okamoto, Y. Generalized-ensemble algorithms for molecular simulations of biopolymers. *Biopolymers* **60**, 96–123 (2001).
40. Zhu, J., Fan, H., Periole, X., Honig, B. & Mark, A.E. Refining homology models by combining replica-exchange molecular dynamics and statistical potentials. *Proteins* **72**, 1171–1188 (2008).
41. Case, D.A., Cheatham, T.E. 3rd., Darden, T., Gohlke, H., Luo, R., Merz, K.M. Jr., Onufriev, A., Simmerling, C., Wang, B. & Woods, R.J. The Amber biomolecular simulation programs. *J. Comput. Chem.* **26**, 1668–1688 (2005).
42. 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).
43. Uberuaga, B.P., Anghel, M. & Voter, A.F. Synchronization of trajectories in canonical molecular-dynamics simulations: observation, explanation, and exploitation. *J. Chem. Phys.* **120**, 6363–6374 (2004).
44. Tirion, M.M. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys. Rev. Lett.* **77**, 1905–1908 (1996).
45. Rotkiewicz, P. & Skolnick, J. Fast procedure for reconstruction of full-atom protein models from reduced representations. *J. Comput. Chem.* **29**, 1460–1465 (2008).
46. Imai, T., Harano, Y., Kinoshita, M., Kovalenko, A. & Hirata, F. Theoretical analysis on hydration thermodynamics of proteins. *J. Chem. Phys.* **125**, 024911 (2007).
47. Ren, P. & Ponder, J.W. Polarizable atomic multipole water model for molecular mechanics simulation. *J. Phys. Chem. B* **107**, 5933–5947 (2003).
48. Kusalik, P.G. & Patey, G.N. The solution of the reference hypernetted-chain approximation for water-like models. *Molecular Physics* **65**, 1105–1119 (1988).
49. Kinoshita, M. & Bérard, D.R. Analysis of the Bulk and Surface-Induced Structure of Electrolyte Solutions Using Integral Equation Theories. *J. Comput. Phys.* **124**, 230–241 (1996).
50. Kusalik, P.G. & Patey, G.N. On the molecular theory of aqueous electrolyte solutions. I. The solution of the RHNC approximation for models in finite concentration. *J. Chem. Phys.* **88**, 7715–7738 (1988).
51. Kinoshita, M. Molecular origin of the hydrophobic effect: analysis using the angle-dependent integral equation theory. *J. Chem. Phys.* **128**, 024507–024520 (2008).
52. Kinoshita, M. Water Structure and Phase Transition Near a Surface. *J. Sol. Chem.* **33**, 661–687 (2004).
53. Roth, R., Harano, Y. & Kinoshita, M. Morphometric approach to the solvation free energy of complex molecules. *Phys. Rev. Lett.* **97**, 078101 (2006).
54. Connolly, M.L. Solvent-accessible surfaces of proteins and nucleic acids. *Science* **221**, 709–713 (1983).
55. Snedon, S.F., Tobias, D.J. & Brooks, C.L. 3rd. Thermodynamics of amide hydrogen bond formation in polar and apolar solvents. *J. Mol. Biol.* **209**, 817–820 (1989)
56. McDonald, I.K. & Thornton, J.M. Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* **238**, 777–793 (1994).
57. Yasuda, S., Yoshidome, T., Oshima, H., Kodama, R., Harano, Y. & Kinoshita, M. Effects of side-chain packing on the formation of secondary structures in protein folding. *J. Chem. Phys.* **132**, 065105 (2010).