PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Refinement of protein termini in template-based modeling using conformational space annealing

Hahnbeom Park,[1] Junsu Ko,[1] Keehyoung Joo,[2] Julian Lee,[3] Chaok Seok,[1]* and Jooyoung Lee[2]*

[1] Department of Chemistry, Seoul National University, Seoul 151-747, Republic of Korea
[2] School of Computational Sciences and Center for In Silico Protein Science, Korea Institute for Advanced Study, Seoul 130-722, Republic of Korea
[3] Department of Bioinformatics and Life Science, Soongsil University, Seoul 156-743, Republic of Korea

ABSTRACT

The rapid increase in the number of experimentally determined protein structures in recent years enables us to obtain more reliable protein tertiary structure models than ever by template-based modeling. However, refinement of template-based models beyond the limit available from the best templates is still needed for understanding protein function in atomic detail. In this work, we develop a new method for protein terminus modeling that can be applied to refinement of models with unreliable terminus structures. The energy function for terminus modeling consists of both physics-based and knowledge-based potential terms with carefully optimized relative weights. Effective sampling of both the framework and terminus is performed using the conformational space annealing technique. This method has been tested on a set of termini derived from a nonredundant structure database and two sets of termini from the CASP8 targets. The performance of the terminus modeling method is significantly improved over our previous method that does not employ terminus refinement. It is also comparable or superior to the best server methods tested in CASP8. The success of the current approach suggests that similar strategy may be applied to other types of refinement problems such as loop modeling or secondary structure rearrangement.

Proteins 2011; 79:2725–2734.

Key words: protein structure prediction; protein structure refinement; protein terminus modeling; CASP; fragment assembly.

INTRODUCTION

The proportion of proteins that template-based modeling can generate reasonably accurate models is rapidly increasing due to the continuous growth of both structure and sequence databases. In many cases, however, we need more accurate protein structure models than possible with the current state-of-the-art modeling techniques for practical applications including structure-based drug design and identification of molecular mechanisms behind biological functions. For this reason, subsequent refinement of protein three-dimensional models generated by template-based methods is particularly important.[1] For example, in recent CASP (Critical Assessments of techniques for protein Structure Prediction) experiments refinement methods have been evaluated both by considering the extent of model quality improvement over the best available templates for template-based modeling (TBM) targets[2,3] and by considering the degree of improvement over provided initial models in the "refinement category."[4]

Various refinement strategies have been proposed to improve homology models, and they can be categorized into three types. First, a group of studies have focused on refining local structures to obtain physically more realistic local geometry. In this type of approach, atomic positions are modified with minimal change in the backbone structure by fixing errors in stereochemistry, by relieving steric clashes, or by maximizing side chain packing and/or hydrogen bond interactions.[5–7] This type of refinement methods is useful when the backbone structure is highly accurate.

Second, refinement methods that reorganize the overall structure have also been reported.[8–12] In this type of approach, movements of backbone atoms from the initial model obtained from template proteins are attempted. Search methods such as local minimization, Monte Carlo, or

Additional Supporting Information may be found in the online version of this article.
Grant sponsor: Creative Research Initiatives (Center for In Silico Protein Science) of MEST/KOSEF; Grant number: 2009-0063610; Grant sponsor: MEST/KOSEF (to C.S.); Grant number: 305-20100007; Grant sponsor: Center for Marine Natural Products and Drug Discovery (CMDD), Ministry of Land, Transport, and Maritime Affairs (to H.P.)
*Correspondence to: Chaok Seok, Department of Chemistry, Seoul National University, Seoul 151-747, Republic of Korea. E-mail: chaok@snu.ac.kr; Jooyoung Lee, School of Computational Sciences, Korea Institute for Advanced Study, Seoul 130-722, Republic of Korea. E-mail: jlee@kias.re.kr.
Received 6 April 2011; Revised 17 May 2011; Accepted 27 May 2011
Published online 16 June 2011 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.23101

molecular dynamics simulations have been applied to optimize various energy functions. Several studies have shown successful refinements over starting models. The most successful cases are those in which initial structures for refinement are moderately different from the native structures.[8,13,14] However, refining near-native models better than the best templates with this type of approach still remains a challenge.

Lastly, refinement techniques concentrating on regions of large variability among templates have been intensively investigated.[15–17] Such variable local regions often contribute to the functional specificity of the target protein. The target regions for refinement typically correspond to insertions in sequence alignments or to the regions for which inconsistent structural information is provided from various templates. Once such regions are identified, refinement efforts are focused on these variable local regions.

In this article, we report a new development that belongs to the third type of the approaches mentioned above. We first predict unreliable local regions that are relatively less accurate than the rest of the model structure. We then focus on refining unreliable "termini" based on our previous work on template-based model building.[18] Modeling of protein terminus is important since protein termini are known to be involved in a variety of biological processes due to their sequence diversity.[19,20] A new energy function is developed to refine unreliable termini that do not have proper template information. While considering the conformational spaces of both the template and terminus simultaneously, low energy conformations are searched utilizing the global optimization method of conformational space annealing.[21] The proposed terminus refinement method was tested on a set of 16 termini derived from a nonredundant structure database and two sets of protein termini from the CASP8 targets each containing 16 and 15 termini. The results of the current method show significant improvement over our previous method, the template-based modeling method of LEE-server in CASP8, which did not employ terminus refinement. Although the ULR detection step resulted in application of the current method to the particular cases in which LEE-server was not successful, it is notable that the results are comparable to Zhang-server, whose performance was assessed to be the best, but statistically indistinguishable from that of LEE-server.[22]

METHODS

Prediction of unreliable local regions

To refine a given protein model, we first identify local regions that are relatively less accurate than the rest of the model. In template-based modeling (TBM), these unreliable local regions (ULRs) often occur in the regions for which templates do not provide adequate structural information. To detect ULRs, we employ a method similar in spirit to model-consensus methods, which have been widely used to assess the quality of local structures of protein models.[23,24]

We predict ULRs based on the multiple sequence alignment (MSA) between a target protein and its templates used for generation of the model. First, a sufficient number of initial models, say 100, are built for a given MSA by using MODELLER.[25] The root-mean-square fluctuations (RMSFs) of each model from the average of the 100 models are then calculated at the residue level. Up to this point, the procedure is similar to typical model-consensus methods. A unique feature of our method is that the criterion to determine ULRs is adjusted depending on the overall quality of the model. A measure of the overall model quality, called base_dev, is defined as the average RMSF of the residues corresponding to the lowest 40% of residual RMSF. The residues with RMSF $> Max\ (S_{\mathrm{cut}} \times \text{base\_dev}, 0.6)$ and any single residues sandwiched by such residues are considered as candidate residues that may belong to ULRs. The value of S_{cut} was set to 2.5. The lower bound of 0.6 was introduced to prevent over-prediction of ULRs for highly accurate protein models. Finally, stretches of three or more consecutive candidate residues are predicted as ULRs.

To assess the accuracy of the ULR prediction method, we define the "true" ULRs for a protein model as the regions of three or more consecutive residues that have Cα deviations from the native structure above $q_{\mathrm{cut}} = 1 + 0.08 \times (100 - \text{GDT-TS})$ Å after superposition of the protein model to its native structure. GDT-TS is a standard measure used in the CASP for assessment of global model quality and defined as the percentage of the aligned residues within 1, 2, 4, and 8 Å divided by 4.[22] For a highly accurate model, q_{cut} approaches 1 Å. For a model with GDT-TS = 80, which is a rough boundary to define high-accuracy TBM targets in the CASP, q_{cut} becomes 2.6 Å. An example of true ULRs and predicted ULRs is illustrated in Figure 1.

The energy function for unreliable termini

The ULR prediction method introduced above can be applied to identify any types of ULRs such as loops or termini. In this work, we focus on developing an energy function specific for "terminus" ULRs.

The energy function we adopt for protein terminus modeling is expressed as a weighted sum of both physics-based and knowledge-based potential terms as follows:

$$E = E_{\mathrm{bonded}} + E_{\mathrm{soft-sphere}} + w_1 E_{\mathrm{DFIRE}} + w_2 E_{\mathrm{dDFIRE+}} + w_3 E_{\mathrm{neighbor}}. \tag{1}$$

The first two terms E_{bonded} and $E_{\text{soft-sphere}}$ are physics-based potential terms used in MODELLER[25] to maintain the proper stereochemistry of proteins. E_{bonded} refers to

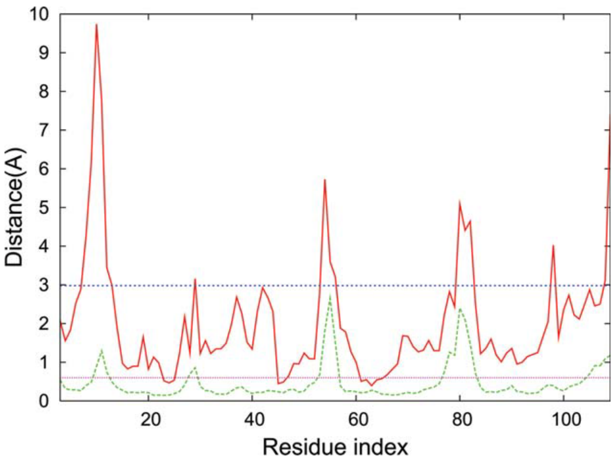

Figure 1

True ULRs of a model (GDT-TS = 75.23) for the CASP8 target T0415-D1 are defined as the three regions (residues 8–13, 54–56, 80–82) for which stretches of three or more residues have Cα deviations from the native (shown as the red solid line) greater than q_{cut} = 2.98 Å (indicated as the blue horizontal line). The predicted ULRs are the four regions of three or more consecutive residues (residues 10–12, 53–56, 77–83, 105–109) in which RMSFs of 100 models from the average model (shown as the green dotted line) are above (the pink horizontal line). The value of base_dev is 0.51 for this example. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

the bonded energy with modified CHARMM22 parameters,[26] and it consists of the energy terms for bond lengths, bond angles, torsion angles, and improper torsion angles. $E_{\text{soft-sphere}}$ refers to the soft-sphere potential that prevents steric clashes. The weights of these two terms are fixed at unity, and the other three weights (w_1,w_2,w_3) are optimized relative to these terms. The component E_{DFIRE} is the DFIRE statistical potential,[27,28] and $E_{\mathrm{dDFIRE+}}$ represents the additional orientation-dependent terms of dDFIRE.[29] The original dDFIRE potential is $E_{\mathrm{DFIRE}} + E_{\mathrm{dDFIRE+}}$, but we treat the weights w_1 and w_2 independently here. The knowledge-based potentials DFIRE and dDFIRE were shown to be efficient in discriminating native structures from decoys in recent studies,[30,31] but it was suggested that the implicit treatment of solvation effect in DFIRE[32] may have limited the performance. In this study, we incorporate the additional energy term E_{neighbor}, called "neighbor energy", for more proper consideration of the solvation effect.

The neighbor energy E_{neighbor} effectively estimates the solvation free energy as a function of solvent accessibility of each residue measured by the number of its neighboring residues. This kind of energy has already been applied to protein structure prediction with various functional forms,[33–35] and we adopt a functional form similar to that proposed by Sasaki *et al.*[33] In addition, we introduce a modification by categorizing the 20 amino acids into three types, hydrophobic ($t = h$), polar ($t = p$), and aromatic ($t = a$), and consider the number of neighbors for each type. In this way, hydrophobicity of local chemical environment is considered in addition to solvent accessibility. The numbers of neighboring residues in four spherical shells, 0–4 Å ($k = 1$), 4–6 Å ($k = 2$), 6–8 Å ($k = 3$), and 8–10 Å ($k = 4$) from the Cα atom of a given residue, are considered following Sasaki *et al.*[33]

Another important feature of the neighbor energy is that sequence-specific information derived from a fragment library is utilized.[33] For each residue position, protein "fragments" of nine residues similar in local sequence features to the position are derived from the structure database.[36] A conformation has favorable neighbor energy if the hydrophobic environment of each residue is similar to that of the corresponding residues in the fragment library. The neighbor energy is expressed as a sum of the contributions $E_{t,jk}$ estimated from the number of neighboring amino acids of type t in the kth shell around the jth residue, $N_{t,jk}$, as follows:

$$E_{\mathrm{neighbor}} = \sum_{jk} (E_{h,jk} + E_{a,jk}), \tag{2}$$

$$E_{t,jk} = -k_{\mathrm{B}}T \log \left[\frac{\sum_i \exp\{-c_{ij}(N_{t,jk} - (N'_{t,ijk} + \alpha_t N'_{ijk}))^2\}}{\sum_i c_{ij}} \right]. \tag{3}$$

The contribution from polar neighboring residues ($t = p$) is not included in Eq. (2) because it did not improve the overall results. $N'_{t,ijk}$ is the number of neighboring residues of type t in the kth shell around the jth residue for the ith fragment. c_{ij} is the similarity between the local sequence features of the ith fragment centered at the jth residue and the target protein. α_t is a coefficient used to enforce hydrophobic packing in proportion to the total number of neighboring residues, N'_{ijk}, found in the fragment. These parameters are set to $(\alpha_h, \alpha_a) = (1, 1/3)$.

We emphasize that all energy components described above are differentiable, and a gradient-based quasi-Newton method L-BFGS[37] was used for efficient local energy minimization. In particular, a continuous, analytic version of dDFIRE was devised by interpolating discrete regions of the original potential using a cubic-spline method.[38] To further improve the efficiency of energy evaluation, the distance cut-off value for interacting atom pairs in dDFIRE was reduced from the original value of 15 Å to 10 Å.

Energy parameter optimization for terminus modeling

The energy parameters (w_1,w_2,w_3) introduced in Eq. (1) were optimized by requiring that native-like conformations are energetically more favored than non-native decoy conformations for a training set of protein termini.

The training set was constructed by selecting protein termini in a non-redundant structure database (called Set 1) whose sequences are not aligned to any sequences in a larger second structure database (called Set 2). Set 1 and Set 2 consist of X-ray structures from PISCES[39] with maximum mutual sequence identity of 25% and 90% and resolution better than 1.6 Å and 2.5 Å, respectively. An in-house method FoldFinder[40] was used for sequence alignment. After filtering out protein termini at interchain interfaces or longer than 30 residues, the final training set of 27 protein termini was obtained. The training set is listed in Supporting Information, Table S1.

Next, an ensemble of conformations that contains both native-like and non-native conformations was generated for each protein terminus in the training set. To cover the conformational space effectively, conformations were built by using three separate methods: (1) assembly of fragments from the fragment library without using the crystal structure information,[36] (2) large perturbations to the crystal structure, and (3) small perturbations to the crystal structure. During generation of conformations, only the terminus was sampled, and the non-ULR regions, called framework, were fixed at the native structure. The numbers of initial conformations generated by the three methods for a single terminus are (1) 2000, (2) 2000, and (3) 300. The conformations were then refined by MD simulations with simulated annealing (SA), followed by local minimization, as in MODELLER. After the refinement, redundant decoys were removed by using 1.0 Å RMSD cut-off. A set of random energy parameters within preset bounds was used to generate each conformation to avoid bias to particular initial weight parameters and to cover the conformational space more broadly.

The energy parameters were optimized by minimizing the following objective function

$$F = \langle Z_{\mathrm{nat}} \rangle \times \langle \mathrm{Corr} \rangle, \tag{4}$$

where Z_{nat} is the *Z*-score of the average energy of the 100 conformations closest to the native structure in the energy distribution of the conformational ensemble, Corr is the Pearson correlation coefficient between the energy and *S*-score,[23] and $\langle\rangle$ denotes average over the 27 training termini. *S*-score is a measure of conformational deviation from the native defined as

$$S - \mathrm{score} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{1+(d_i/d_0)^2}, \tag{5}$$

where *i* is the residue index, *n* is the number of residues in the terminus, d_i is the deviation of the *i*th residue from the native. The parameter d_0 is set to 5 Å to consider the effect of large structural diversity of termini. *S*-score is used instead of the frequently used root mean square deviation (RMSD) because RMSD tends to exaggerate incorrectness in local structure and has strong size dependency.[23] Expressing the objective function for energy parameter optimization as a multiplication of two terms as in Eq. (4) was inspired by the work of Zhang *et al.*[34] The energy parameters (w_1,w_2,w_3) were determined by a grid-search starting from a coarse grid and subsequently focusing on finer grids near promising regions.

Test sets for ULR terminus modeling

We first selected 16 protein termini from Set 2 used for energy parameter training, after eliminating 11 termini that interact with other chains in the crystal structures. It should be noted that the decoy training discussed above was performed on Set 1. We call this set 'PISCES test set' (listed in Supporting Information, Table S2). This set can be used to assess the performance of the terminus modeling procedure when the framework structure is exact because the framework structure is fixed at the native during terminus modeling for this set.

We also selected unreliable protein termini from the CASP8 targets. Two test sets of CASP8 targets were generated, by predicting terminus ULRs using our ULR prediction method and by collecting the true ULRs using the native structure. The first set, called "predicted ULR set," consists of 16 targets (listed in Supporting Information, Table S3). The modeling results for these targets can be compared in a fair manner to the methods of other prediction groups in CASP8 because the whole procedure including prediction of ULRs was executed in a blind fashion. The second set, referred to as "assigned ULR set," was obtained by applying the definition of ULR to our models submitted in CASP8 as LEE group. This set consists of 15 template-based modeling (TBM) targets (listed in Supporting Information, Table S4). Eight targets are also included in the predicted ULR set and out of which four share the terminus regions to be modeled. We can evaluate the accuracy of the energy function and the efficiency of the sampling method from the results on this set.

Conformational search by global optimization

Low energy structures of the terminus ULRs were searched for using an extended version of MODELLERCSA.[18] MODELLERCSA is a method that builds homology models using conformational space annealing (CSA)[21] by optimizing the MODELLER energy derived from template structures and a multiple sequence alignment of the template and target sequences.[41] The energy terms for terminus ULR as expressed in Eq. (1) were implemented into MODELLERCSA, which we call MOD-ULR-CSA. Compared with the previous version of MODELLERCSA where local energy minimization was carried out using the MODELLER program as a black box and

Table I
Prediction Accuracy and Coverage for a Simple Sequence-Based ULR Prediction Method and the Model-Consensus ULR Prediction Method are Shown

			No offset allowed		Two-residue offset allowed[a]	
Prediction method	Parameter in the method		Accuracy	Coverage	Accuracy	Coverage
Sequence-based ULR prediction	Cut-off for the fraction of the aligned templates	0	0.830	0.082	0.925	0.088
		0.2	0.694	0.121	0.790	0.131
		0.5	0.594	0.224	0.699	0.261
Model-consensus ULR prediction method	S_{cut}	1.5	0.534	0.501	0.658	0.564
		2.0	0.585	0.472	0.720	0.537
		2.5	0.619	0.434	0.759	0.499
		3.0	0.642	0.396	0.780	0.453

[a]Those residues that are incorrectly predicted as ULR but are within two residues from the actual ULR are excluded from the statistics.

CSA was performed at a script level, computational efficiency has improved by integrating CSA with the ULR energy into one program.

The ULR region and the rest of protein, called framework, are modeled simultaneously for the two CASP8 test sets. During the modeling procedure, the new energy terms are applied to intra-ULR and ULR-framework interactions, while MODELLER restraints derived from templates are applied only to intra-framework interactions. This allows structural changes of the framework region due to ULR-framework interactions. Exceptionally for the PISCES test set, the framework is fixed at the native structure by strong restraints. The procedure of MODULR-CSA follows that of MODELLERCSA with identical CSA parameters except for two differences: (1) preparation of the "first bank" and (2) the inclusion of additional CSA "crossover operators."

The first bank of CSA provides a source of conformational diversity that is exploited in CSA iterations. To achieve sufficient structural diversity of the framework, 100 separate conformations were first generated using MODELLER, as in MODELLERCSA. To introduce additional structural diversity to terminus, 20 terminus conformations were generated by fragment assembly for each framework model. The resulting 2000 structures were refined using MD with simulated annealing and subsequently clustered into 100 groups by K-means clustering. The 100 cluster center structures were selected as the first bank.

CSA generates trial conformations by crossovers and mutations as in genetic algorithms. For effective sampling of the ULR regions, additional operators specific for ULRs were introduced. The relative frequency of the applications of the standard and new operators was set to 5:7 in this work.

RESULTS AND DISCUSSION

Prediction of unreliable local regions

CASP experiments make it possible to assess performances of state-of-the-art protein structure prediction techniques in a blind fashion. We analyzed the quality of the TBM models submitted during CASP8 for 48 high-accuracy TBM targets and 104 regular TBM targets for their accuracies in local regions (See Supporting Information, Figs. S1 and S2 for details). There are many examples of unreliably modeled regions even when the best models (not the models of the best group) are considered. For example, about 30% of the residues in the best models in the TBM category deviate by more than 2.5 Å from the native structures. The proportion of unreliably modeled terminus residues is about the same if the 20 residues from the N- and C-terminus residue are considered as terminus residues. In this work, we aim to improve unreliably modeled protein termini.

We applied the ULR prediction method to the CASP8 targets and compared the results with those obtained by a simple sequence-based method which selects the regions that are not aligned to templates. In the sequence-based method, a residue is predicted as a ULR residue if the fraction of the aligned templates at the residue position is less than a preset cut-off value. We used the multiple sequence alignments (MSA) of LEE group for the CASP8 targets. With the sequence-based method, accuracy of ULR prediction is quite high (92.5%), but coverage is rather low (8.8%) if the cut-off value of 0 is used even with two-residue offset (see Table I). Accuracy is defined as the percentage of the correctly predicted ULR residues out of all predicted residues, and coverage is defined as the percentage of the correctly predicted ULR residues out of the true ULR residues. The accuracy decreases to 69.9% and the coverage increases to 26.1% if the cut-off value is increased to 0.5.

The current model-consensus ULR prediction method is superior to the simple sequence-based method with the accuracy of 75.9% and the coverage of 49.9% when $S_{\mathrm{cut}} = 2.5$ and two-residue offset is used, as presented in Table I. This is due to the fact that the model-consensus method takes account of structural diversity among templates, which is neglected in the sequence-based method. For the modeling tests presented below, we used $S_{\mathrm{cut}} = 2.5$, emphasizing accuracy over coverage considering the current status of the modeling accuracy.

Table II

The Average Correlation Coefficient and the Average *Z*-Score (Inside Parenthesis) for the Three Energy Terms (E_{DFIRE}, $E_{\mathrm{dDFIRE+}}$, and E_{neighbor}), for the Optimized Energy with only Two Terms (E_{DFIRE} and $E_{\mathrm{dDFIRE+}}$, E_{DFIRE} and E_{neighbor}, $E_{\mathrm{dDFIRE+}}$ and E_{neighbor}), and for the Total Energy (*E*) with the Optimized Parameters are Shown

		Corr.[a] (*Z*-score[b])						
Energy		E_{DFIRE}	$E_{\mathrm{dDFIRE+}}$	E_{neighbor}	E_{DFIRE} and $E_{\mathrm{dDFIRE+}}$	E_{DFIRE} and E_{neighbor}	$E_{\mathrm{dDFIRE+}}$ and E_{neighbor}	E
SS[c]	Helix (15)	0.411 (−1.728)	0.207 (−1.094)	0.272 (−0.828)	0.475 (−1.896)	0.466 (−1.746)	0.346 (−1.202)	0.478 (−1.906)
	Strand (5)	0.342 (−1.501)	0.090 (−0.758)	0.344 (−1.031)	0.429 (−1.767)	0.469 (−1.782)	0.366 (−1.200)	0.440 (−1.789)
	Coil (7)	0.211 (−0.904)	−0.014 (−0.196)	0.212 (−0.472)	0.240 (−0.939)	0.301 (−1.057)	0.231 (−0.638)	0.269 (−0.983)
Overall[d] (27)		0.339 (−1.442)	0.120 (−0.766)	0.267 (−0.756)	0.397 (−1.588)	0.416 (−1.515)	0.331 (−0.987)	0.409 (−1.611)

[c] The average correlation coefficient between the energy and *S*-score, a measure of structural deviation from the native, for training set decoys is shown.

[d] The average *Z*-score of the average energy of the 100 conformations closest to the native structure in the energy distribution of the training set decoys is shown.

[c] The results are shown for the three secondary structure types of the termini. The number in the parenthesis for each secondary structure type is the number of training termini of the type.

[d] The overall results are averaged over the whole training terminus set.

Energy parameter optimization for terminus modeling

We determine the three weight factors for the energy components, DFIRE energy E_{DFIRE}, additional orientation-dependent dDFIRE terms $E_{\mathrm{dDFIRE+}}$, and the neighbor energy E_{neighbor}. The objective function for parameter optimization was designed to increase the correlation between total energy and structural deviation from the native and to lower the energy of native-like structures relative to the non-native ones, as described in the Methods section.

Before considering the above three parameters, the parameters inside the E_{neighbor} term were first determined to give improved correlation and *Z*-score for the identical training set of decoys (See Supporting Information, Table S5 for details). E_{neighbor} with the parameters, $(\alpha_h, \alpha_a) = (1, 1/3)$, shows the improved average correlation coefficient of 0.27, and the average *Z*-score of −0.76, compared with the corresponding values of 0.18 and −0.46 from the original form of neighbor energy proposed by Sasaki *et al.*[33]

By a grid search, the three weight parameters for the three energy components were determined as $(w_1, w_1, w_1) = (22.0, 12.0, 8.7)$. The average correlation coefficients and average *Z*-scores for each energy component and the total energy are displayed in Table II. Results for the secondary structure classes (helix, extended, and coil) are also shown. The DFIRE energy contributes the most to the correlation and *Z*-score, and the neighbor energy follows. The contribution from the additional dDFIRE term is the least, but still meaningful, and this is represented in the smaller weight of 12.0 relative to the DFIREs 22.0. The correlation and the *Z*-score improve the most for strands when all three terms are combined.

Energy landscapes for each energy component as well as for the total energy with the optimized parameters are shown in Figure 2 for a member of the training set, 2arz. We observe that the decoys cover the conformational space quite broadly. The energy landscapes show some degree of correlation for each of the three components, and the best correlation is achieved from the combined total energy using the optimized parameters.

Performance of MODULR-CSA on the PISCES test set

We first applied the MODULR-CSA method to the PISCES test set. As mentioned above, this set can be used to check the performance of the method when the frameworks are exact. The results are summarized in Table III. We have succeeded in generating terminus models with better than 10 Å accuracy in 11 of 16 cases. If the lowest energy structure is selected as the single answer, 6 out of 16 termini are modeled within 10 Å. Modeling worked particularly well for the termini of helical structure. Among the eight helical termini, six termini were sampled with better than 2 Å accuracy. Termini containing strands and coils were predicted poorly. We attribute this failure to the energy function rather than to the sampling method because more native-like terminus models sampled by CSA were not favored by their energy values.

Interestingly, several protein termini successfully modeled here were previously proposed to be involved in protein function. For example, synaptotagmin I C2B domain (PDB ID 1tjx) shows strikingly different C-terminus structure from its homologs, and this terminus structure is relevant to endocytosis.[42] N-terminus of 5-formyl tetrahydrofolate cycloligase (PDB ID 2jcb), having extended helical structure unlike its homologs, tightly interacts with ADP. Although the above examples are on already known cases, our approach can be potentially applied to other unknown cases, providing molecular-level understanding on function.

Performance of MODULR-CSA on termini of CASP8 targets

MODULR-CSA was also tested on the terminus ULRs selected from the CASP8 targets. For the framework

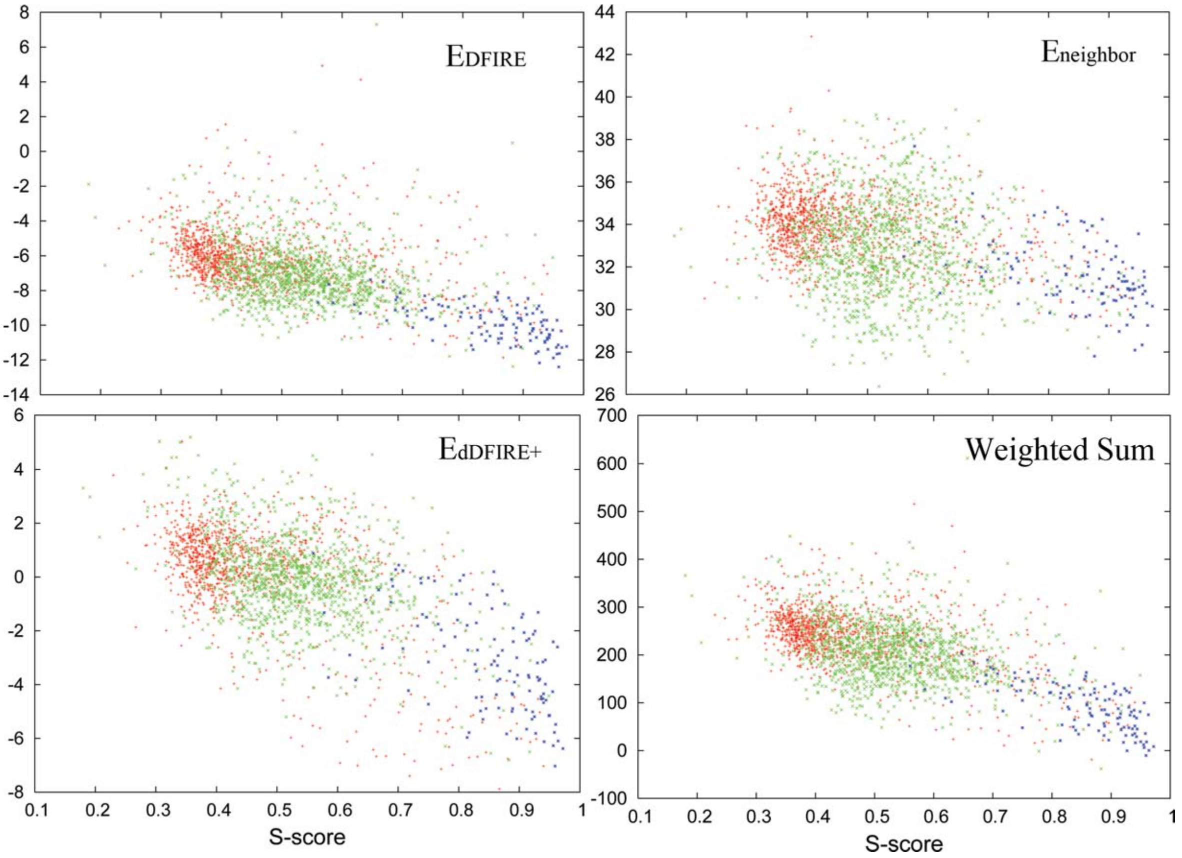

Figure 2

The energy landscape (plot of energy vs. *S*-score) for each of the three energy components, E_{DFIRE}, $E_{\mathrm{dDFIRE+}}$, and E_{neighbor}, and that for the total energy with the optimized weight parameters are shown for 2arz. Conformations are generated by fragment assembly (red), large perturbations to the native (green), and small perturbations to the native (blue). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

regions, the same MODELLER restraint energy used for homology modeling by LEE group in CASP8 was used. MODELLERCSA optimizes the restraint energy obtained from templates with CSA, while MODULR-CSA optimizes the energy function developed in this work with more intensive sampling on termini. Comparison between the results of MODELLERCSA and MODULR-CSA allows us to evaluate the performance of the new energy and the sampling method for termini because the same restraint energy for the framework is used.

The results of MODELLERCSA and MODULR-CSA for the 16 termini of the predicted ULR set are compared in Table IV. MODULR-CSA provides three more termini within 3 Å RMSD, two more within 5 Å, and four more within 10 Å than MODELLERCSA. Since wrong orientation of a long terminus can cause a large RMSD value, we use RMSD <10 Å as a criterion for acceptable terminus orientation. Significant improvements are found in the terminus modeling results of T0414, T0457, T0462, and T0477. The orientations of these termini were wrongly predicted by MODELLERCSA, but they are fixed by the current procedure. Figure 3 illustrates the successful example of T0457. The termini of T0408 and T0449 show RMSDs of only 3.03 and 1.74 Å with MODELLERCSA, meaning that they are obviously mispredicted as ULRs. However, these termini are accurately modeled by MODULR-CSA, too.

The results for the 15 termini of the assigned ULR set are shown in Table V. This set includes more of difficult targets that were not selected by ULR prediction (7 out of 15 termini), and this detection failure is partly due to our choice to increase the accuracy of ULR prediction in sacrifice of the coverage. In addition to the targets included in the predicted ULR set, we succeeded in refining structures for three additional termini, those of T0423, T0501, and T0504. In particular, T0423 shows the possibility of high-accuracy refinement. This terminus was already in the right position with RMSD = 3.98 Å by MODELLERCSA. MODULR-CSA can further refine this terminus to the accuracy of RMSD = 1.24 Å. Termini of T0388 and T0414 in the assigned ULR set have different ranges of residues from those in the predicted

Table III

Terminus Modeling Results for the PISCES Test Set are Summarized

PDB ID	Length	SS[a]	Init[b] (Å)	CSA min[c] (Å)	CSA best[d] (Å)
2qml	16	H	4.51	3.87	1.48
2v9l	29	H	16.37	33.28	9.50
1eyh	22	H	7.48	2.14	1.02
2jcb	7	H	2.94	2.77	1.77
1qgi	20	H	14.65	25.63	10.30
1rtq	10	H	10.20	5.38	5.05
1tjx	10	H	4.02	1.60	1.50
1tua	17	H	6.91	13.94	1.96
2oa2N	20	H, C	7.79	7.83	7.51
2oa2C	13	H, C	15.07	18.90	18.47
1xdz	21	H, C	14.03	26.41	7.89
1tt8	20	H, C	19.09	11.96	10.28
2iuw	10	E	21.97	16.93	9.60
2gz4	24	E, C	13.64	12.07	10.68
1z6m	23	C	13.70	10.11	10.11
1j77	18	C	9.68	25.14	8.58

[a] Secondary structure elements of the termini are denoted as H, E, and C for helix, extended, and coil, respectively.

[b] RMSD of the lowest energy conformation among the initial 2000 conformations generated by fragment assembly followed by MD/SA is shown.

[c] RMSD of the lowest energy conformation among the 100 conformations in the CSA final bank is shown.

[d] The lowest RMSD value out of the 100 conformations in the CSA final bank is shown.

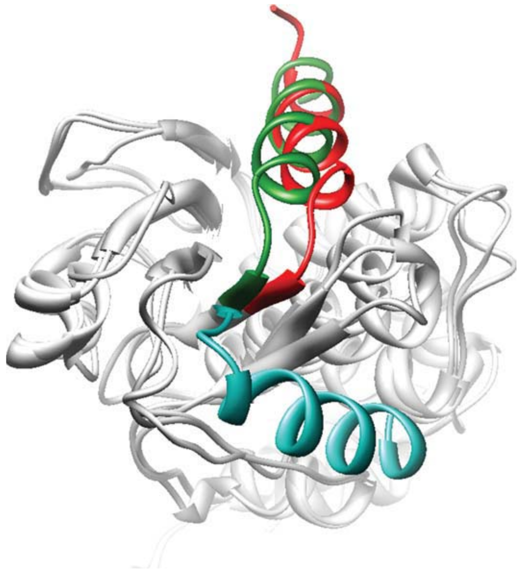

Figure 3

A successful example of terminus modeling is shown for a CASP8 target, T0457. The terminus is colored red for the native structure, cyan for the model generated by MODELLERCSA, and green for the model generated by MODULR-CSA. Even though the conformation of the framework region of the model slightly deviates from the native structure, the overall orientation of the terminus is recovered after the terminus ULR modeling.

ULR set, but MODULR-CSA results from both sets provided similar improvement of termini as shown in Tables IV and V.

Since we modeled the framework and termini simultaneously for the CASP8 targets, the framework structure can be affected by terminus modeling. We therefore assessed the effect of terminus modeling on the overall structure by measuring TM-score,[43] as presented in Supporting Information, Tables S6 and S7. A good correlation is found between the overall quality of the structure

Table IV

Comparison of the Terminus RMSD Values Obtained with and without Terminus Modeling for the Predicted ULRs is Shown

Target no.[a]	Length	SS[b]	MODELLERCSA[c] (Å)	MODULR-CSA[d] (Å)
T0388	17	H	4.80	2.44
T0395	36	H, C	32.82	33.51
T0408	16	H	3.03	3.45
T0412	9	H	4.77	2.46
T0414	21	H	24.74	5.48
T0434[e]	15	H	11.83	8.89
T0435	13	H	11.24	12.37
T0438	14	H	3.82	4.46
T0449	9	E	1.74	1.63
T0451	12	C	9.69	9.76
T0457[e]	19	H	20.84	3.48
T0462[e]	10	C	15.99	3.07
T0477[e]	12	H	18.75	2.22
T0479	9	C	3.55	3.63
T0485	24	H	8.73	8.77
T0509	16	H	3.94	12.88

[a] The CASP8 target number.

[b] The secondary structure of the terminus (H: helix, E: extended, C: coil).

[c] The global RMSD of the terminus of LEE model from CASP8.

[d] The global RMSD of the terminus after terminus refinement by MODULR-CSA.

[e] The targets for which the ULR regions are correctly predicted within one residue. For these targets, the same ULR regions are assigned to the assigned ULR set, so the modeling results are identical with those reported in Table V.

Table V

Comparison of the Terminus ULR RMSD Values Obtained with and without Terminus Modeling for the Assigned ULR Set is Shown

Target no.[a]	Length	SS[b]	MODELLERCSA[c] (Å)	MODULR-CSA[d] (Å)
T0388	11	H	6.91	2.56
T0395	38	H	31.95	32.64
T0414	19	H	25.91	6.86
T0423	15	H	3.98	1.24
T0434[e]	16	H	11.83	8.89
T0435	10	H	12.81	12.56
T0457[e]	18	H	20.84	3.48
T0462[e]	10	C	15.99	3.07
T0477[e]	12	H	18.75	2.22
T0483N	9	C	15.21	16.59
T0483C	13	C	22.71	19.98
T0501	24	H	22.84	5.33
T0503	25	H, C	4.78	9.50
T0504	10	C	10.55	4.90
T0511	11	H	11.55	11.48

[a] The CASP8 target number.

[b] The secondary structure of the terminus ULR region (H: helix, E: extended, C: coil).

[c] The global RMSD values of the terminus of LEE model from CASP8.

[d] The global RMSD values of the terminus after terminus modeling with MODULR-CSA.

[e] The targets for which the ULR regions are correctly predicted within one residue. For these targets, the same ULR regions as in the predicted ULR set are assigned, so the modeling results are identical with those reported in Table IV.

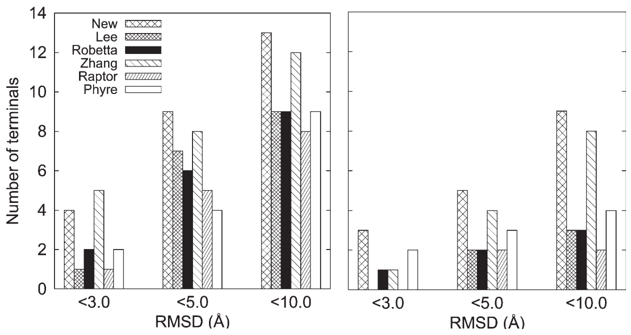

Figure 4

The numbers of termini modeled better than the RMSD values, 3 Å, 5 Å, and 10 Å, for the predicted ULR set (left) and the assigned ULR set (right) for several different methods are shown. The modeling results of MODELLERCSA (LEE) and MODULR-CSA (New) are compared with those of the top-ranking servers tested in CASP8.

(TM-score) and the local quality of the termini (RMSD). As terminus structures are improved, the overall model quality also tends to be improved. The average TM-score is improved by 0.94% and 2.22% from those obtained with MODELLERCSA for the predicted and the assigned terminus ULR set, respectively.

We have shown that MODULR-CSA improves the overall quality of a three-dimensional model containing terminus ULRs over MODELLERCSA. We also compare the results of MODULR-CSA with those of other methods tested during CASP8. We selected five top-ranked server methods, not human methods, because of the following reasons. First, the number of human targets for terminus modeling is insufficient, only 6 out of 16 for the predicted ULR set and 5 out of 15 in the assigned ULR set. Second, our method is fully automatic for the predicted ULR set, so our new method can be fairly compared with the results of the server groups.

The results of MODULR-CSA and those of the top-ranked server methods are compared in Figure 4. MODULR-CSA produces more models of <3 Å, <5 Å, and <10 Å than MODELLERCSA (labeled as "LEE" in the Fig. 4) for both the predicted and assigned ULR sets, as discussed above. MODULR-CSA shows the best performance for the assigned set when compared with the five top server methods. For the predicted ULR set, only the number of models <3.0 Å is one less than that obtained the Zhang server.

The performance of Zhang server is the closest to that of MODULR-CSA. Zhang server uses a method called i-TASSER,[14] which reassembles secondary structure segments guided by contact score extracted from homologous proteins and by their own energy function. When the contact score is unreliable, modeling relies on the energy. Even though the details are different, both MODULR-CSA and i-TASSER share common methodological features in that homology information and a general-purpose energy which can be used for ab initio modeling are combined together. This kind of efforts seems to contribute to higher-accuracy modeling of structurally variable regions in template-based modeling.

CONCLUSIONS

In this article, we have developed a method for improving unreliable termini in template-based models together with a method for predicting such regions. A simple combination of the two knowledge-based potentials, dDFIRE and neighbor energy, is shown to work quite well. The efficiency of the search method, conformational space annealing (CSA), has been demonstrated in many previous modeling examples.[18,40,41] The CSA method developed here allows efficient simultaneous sampling of the framework and the unreliable local regions. The performance of the new method, MODULR-CSA, in modeling unreliable terminus is comparable or superior to the best server methods when tested on the CASP8 targets. There is still a room for improvement, especially for terminus ULR containing strands and/or coils.

The current method has a potential applicability to other related studies. For example, it can be extended to the refinement of other regions of proteins, including loops and the packing of secondary structures. The method can also be applied to situations in which reliable template-based information and local structure variations should be considered simultaneously as in homology-based protein-protein or protein-ligand docking problems.

ACKNOWLEDGMENTS

The authors thank Dr. Yuedong Yang for providing them the source code of dDFIRE and Prof. Masaki Sasai for helpful discussions on energy function development.

REFERENCES

1. Qian B, Raman S, Das R, Bradley P, McCoy AJ, Read RJ, Baker D. High-resolution structure prediction and the crystallographic phase problem. Nature 2007;450:259–264.
2. Keedy DA, Williams CJ, Headd JJ, Arendall WB III, Chen VB, Kapral GJ, Gillespie RA, Block JN, Zemla A, Richardson DC, Richardson JS. The other 90% of the protein: assessment beyond the Cαs for CASP8 template-based and high-accuracy models. Proteins 2009;77:29–49.
3. Kopp J, Borddoli L, Battey JN, Kiefer F, Schwede T. Assessment of CASP7 predictions for template-based modeling targets. Proteins 2007;69:38–56.
4. MacCallum JL, Hua L, Schnieders MJ, Pande VS, Jacobson MP, Dill KA. Assessment of the protein-structure refinement category in CASP8. Proteins 2009;77:66–80.
5. Feig M, Rotkiewwicz P, Kolinski A, Skolnick J, Brooks CL, III. Accurate reconstruction of all-atom protein representations from side-chain-based low-resolution models. Proteins 2000;41:86–97.

6. Krivov GG, Shapovalov MV, Dunbrack RL, Jr. Improved prediction of protein side-chain conformations with SCWRL4. Proteins 2009;77:778–795.
7. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen MY, Pieper U, Sali A. Comparative protein structure modeling using MODELLER. Curr Protoc Bioinformatics 2006;5.6:1–30.
8. Summa CM, Levitt M. Near-native structure refinement using in vacuo energy minimization. Proc Natl Acad Sci USA 2007;104:3177–3182.
9. Chopra G, Kalisman N, Levitt M. Consistent refinement of submitted models at CASP using a knowledge-based potential. Proteins 2010;78:2668–2678.
10. Fan H, Mark AE. Refinement of homology-based protein structures by molecular dynamics simulation techniques. Protein Sci 2004;13: 211–220.
11. Flohil JA, Vriend G, Berendsen HJC. Completion and refinement of 3-D homology models with restricted molecular dynamics: application to targets 47, 58, and 11 in the CASP modeling competition and posterior analysis. Proteins 2002;48:593–604.
12. Zhang Y. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics 2008;9:40.
13. Zhang Y. I-TASSER: fully automated protein structure prediction in CASP8. Proteins 2009;77:100–113.
14. Raman S, Vernon R, Thompson J, Tyka M, Sadreyev R, Pei J, Kim D, Kellogg E, DiMaio F, Lange O, Kinch L, Sheffler W, Kim B-H, Das R, Grishin NV, Baker D. Structure prediction for CASP8 with all-atom refinement using Rosetta. Proteins 2009;77:89–99.
15. Rohl CA, Strauss CEM, Chivian D, Baker D. Modeling structurally variable regions in homologous proteins with Rosetta. Proteins 2004;55:656–677.
16. Soto CS, Fasnacht M, Zhu J, Forrest L, Honig B. Loop modeling: sampling, filtering and scoring. Proteins 2008;70:834–843.
17. Jacobson MP, Pincus DL, Rapp CS, Day TJ, Honig B, Shaw DE, Friesner RA. A hierarchical approach to all-atom protein loop prediction. Proteins 2004;55:351–367.
18. Joo K, Lee J, Seo JH, Lee K, Kim BG, Lee J. All-atom chain-building by optimizing MODELLER energy function using conformational space annealing. Proteins 2009;75:1010–1023.
19. Polevoda B, Sherman F. N-terminal acetyltransferases and sequence requirements for N-terminal acetylation of eukaryotic proteins. J Mol Biol 2003;325:595–622.
20. Chung JJ, Shikano S, Hanyu Y, Li M. Functional diversity of protein C-termini: more than zipcoding? Trends Cell Biol 2002;12:146–150.
21. Lee J, Liwo A, Scheraga HA. Energy-based de novo protein folding by conformational space annealing and an off-lattice united-residue force field: application to the 10-55 fragment of staphylococcal protein A and apo calbindin d9k. Proc Natl Acad Sci USA 1999;96:2025–2030.
22. Cozzetto D, Kryshtafovych A, Krzysztof F, Moult J, Rost B, Tramontano A. Evaluation of template-based models in CASP8 with standard measures. Proteins 2009;77:18–28.
23. Wallner B, Elofsson A. Identification of correct regions in protein models using structural, alignment, and consensus information. Protein Sci 2006;15:900–913.
24. Wang Z, Eickholt J, Cheng J. MULTICOM: a multi-level combination approach to protein structure prediction and its assessments in CASP8. Bioinformatics 2010;26:882–888.
25. Sali A, Blundell T. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 1993;234:779–815.
26. MacKerell AD, Jr, Bashford D, Bellott M, Dunbrack RL, Jr, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE III, Roux B, Schienkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiorkiewicz-Kuczera J, Yin D, Karplus M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B 2002;102: 3586–3616.
27. Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 2002;11:2714–2726.
28. Yang Y, Zhou Y. Ab initio folding of terminal segments with secondary structures reveals the fine difference between two closely-related all-atom statistical energy functions. Protein Sci 2008;17: 1212–1219.
29. Yang Y, Zhou Y. Specific interactions for ab initio folding of protein termini with secondary structures. Proteins 2008;72:793–803.
30. Zhu J, Xie L, Honig B. Structural refinement of protein segments containing secondary structure elements: local sampling, knowledge-based potentials and clustering. Proteins 2006;65:463–479.
31. Zhou Y, Zhou H, Zhang C, Liu S. What is a desirable statistical energy function for proteins and how can it be obtained? Cell Biochem Biophys 2006;46:165–174.
32. Shen M, Sali A. Statistical potential for assessment and prediction of protein structures. Protein Sci 2006;15:2507–2525.
33. Sasaki TN, Cetin H, Sasai M. A coarse-grained Langevin molecular dynamics approach to de novo protein structure prediction. Biochem Biophys Res Commun 2008;369:500–506.
34. Zhang Y, Kolinski A, Skolnick J. TOUCHSTONE II: a new approach to ab initio protein structure prediction. Biophys J 2003;85:1145–1164.
35. Rohl CA, Strauss CEM, Misura KMS, Baker D. Proteins structure prediction using Rosetta. Methods Enzymol 2004;383:66–93.
36. Lee J, Lee D, Park H, Coutsias EA, Seok C. Protein loop modeling by using fragment assembly and analytical loop closure. Proteins 2010;78:3426–3436.
37. Liu D, Nocedal J. On the limited memory BFGS method for large scale optimization. Math Programming B 1989;45:503–528.
38. Press WH. Numerical recipes, 2nd ed. Cambridge: Cambridge University Press; 1992.
39. Wang G, Dunbrack RL, Jr. PISCES: a protein sequence culling server. Bioinformatics 2003;19:1589–1591.
40. Joo K, Lee J, Lee S, Seo JH, Lee SJ, Lee J. High accuracy template based modeling by global optimization. Proteins 2007;69:83–89.
41. Joo K, Lee J, Kim I, Lee SJ, Lee J. Multiple sequence alignment by conformational space annealing. Biophys J 2008;95:4813–4819.
42. Fernandez I, Arac D, Uback J, Gerber SH, Shin O, Gao Y, Anderson RGW, Südhof TC, Rizo J. Three-dimensional structure of the synaptotagmin 1 C_2B-domain: synaptotagmin 1 as a phospholipid binding machine. Neuron 2001;32:1057–1069.
43. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins 2004;57:702–710.