ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/51157821

De novo protein structure prediction by dynamic fragment assembly and conformational space annealing

ARTICLE *in* PROTEINS STRUCTURE FUNCTION AND BIOINFORMATICS · AUGUST 2011

Impact Factor: 2.63 · DOI: 10.1002/prot.23059 · Source: PubMed

CITATIONS
15

READS
136

6 AUTHORS, INCLUDING:

Juyong Lee
National Institutes of Health
22 PUBLICATIONS **154** CITATIONS
SEE PROFILE

Masaki Sasai
Nagoya University
95 PUBLICATIONS **1,767** CITATIONS
SEE PROFILE

Jooyoung Lee
Korea Institute for Advanced Study
120 PUBLICATIONS **3,890** CITATIONS
SEE PROFILE

Available from: Jooyoung Lee
Retrieved on: 11 January 2016

PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

De novo protein structure prediction by dynamic fragment assembly and conformational space annealing

Juyong Lee,[1,2] Jinhyuk Lee,[2] Takeshi N. Sasaki,[3] Masaki Sasai,[2,4] Chaok Seok,[1,2]* and Jooyoung Lee[2]*

[1] Department of Chemistry, Seoul National University, Seoul 151-742, Korea
[2] Center for In Silico Protein Science, School of Computational Sciences, Korea Institute for Advanced Study, Seoul 130-722, Korea
[3] Department of Human Informatics, Aichi Shukutoku University, Aichi 480-1197, Japan
[4] Department of Applied Physics, Nagoya University, Nagoya 464-8603, Japan

ABSTRACT

***Ab initio* protein structure prediction is a challenging problem that requires both an accurate energetic representation of a protein structure and an efficient conformational sampling method for successful protein modeling. In this article, we present an *ab initio* structure prediction method which combines a recently suggested novel way of fragment assembly, dynamic fragment assembly (DFA) and conformational space annealing (CSA) algorithm. In DFA, model structures are scored by continuous functions constructed based on short- and long-range structural restraint information from a fragment library. Here, DFA is represented by the full-atom model by CHARMM with the addition of the empirical potential of DFIRE. The relative contributions between various energy terms are optimized using linear programming. The conformational sampling was carried out with CSA algorithm, which can find low energy conformations more efficiently than simulated annealing used in the existing DFA study. The newly introduced DFA energy function and CSA sampling algorithm are implemented into CHARMM. Test results on 30 small single-domain proteins and 13 template-free modeling targets of the 8th Critical Assessment of protein Structure Prediction show that the current method provides comparable and complementary prediction results to existing top methods.**

Proteins 2011; 79:2403–2417.
© 2011 Wiley-Liss, Inc.

Key words: template-free modeling; *ab initio* protein structure prediction; fragment assembly; energy parameter optimization; conformational space annealing.

INTRODUCTION

Ab initio prediction of the native structure of a protein from its sequence information is one of the most challenging problems of computational biophysics. During the last decade, there has been dramatic extension of our understanding and methodology on protein structure prediction. Among numerous algorithms suggested, fragment assembly approaches have been shown to be quite efficient and successful. Often, they are coupled with coarse-grained representations of proteins and Monte Carlo (MC) sampling methods.[1,2] Coarse-grained protein models are used to deal with large proteins by reducing computational load. MC methods are easy to implement and can be used with various potential energy functions.

However, these approaches have limitations. MC sampling is often trapped in a local energy minimum state during conformational search. To alleviate this problem, various advanced MC sampling methods have been suggested including generalized-ensemble,[3] replica exchange,[4] multicanonical-ensemble,[5,6] and parallel hyperbolic MC.[7] Final protein models by coarse-grained representations tend to include multiple steric clashes or noncanonical bond lengths and bond angles even after all-atom chain rebuilding and refinement procedures.[8]

Typical protein structure prediction methods utilizing fragments of existing PDB structures generate the whole chain model by assembling fragments from a preprocessed library. Structural variation is achieved by substituting parts of the whole chain with fragments of fixed conformations. Recently, Sasaki and Sasai[9,10] proposed a variation of the fragment assembly method where a fragment-based potential energy function was used. In this new way of

Additional Supporting Information may be found in the online version of this article.
Grant sponsor: Creative Research Initiatives, MEST/KOSEF (Center for *In Silico* Protein Science); Grant number: 2010-0000718; Grant sponsor: Japan Society for the Promotion of Science; Grant number: 20244068
*Correspondence to: Jooyoung Lee, Center for *In Silico* Protein Science, School of Computational Sciences, Korea Institute for Advanced Study, Seoul 130-722, Korea. E-mail: jlee@kias.re.kr or Chaok Seok, Department of Chemistry, Seoul National University, Seoul 151-742, Korea.
E-mail: chaok@snu.ac.kr.
Received 17 December 2010; Revised 24 March 2011; Accepted 12 April 2011
Published online 20 April 2011 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.23059

fragment assembly, protein is represented by a trace of Cα atoms, and statistical tendency of local and global conformational information from fragments is converted into an energy function. The role of fragments is fulfilled by establishing this energy function, which we call dynamic fragment assembly (DFA). Sasaki and Sasai[9,10] used simulated annealing (SA) Langevin dynamics sampling method to generate protein models.

Although local conformations are highly constrained to those of fragments in usual fragment assembly methods, conformational sampling in DFA is much less restrained and any fluctuation is allowed. In other words, conformational space is expanded in DFA relative to the conventional fragment assembly method, which may require more rigorous conformational sampling than the MC method. Additionally, with analytically differentiable form of DFA energy, more efficient local energy minimization methods such as steepest descent, conjugated gradient, and quasi-Newton methods,[11] can be readily applied during conformational sampling, which is not possible with conventional discrete fragment assembly methods.

Executing efficient sampling of the conformational space of proteins is a very hard problem. The difficulty arises from the fact that the conformational space of a reasonable-size protein constitutes a quite rugged energy landscape with many local minima separated by large energy barriers. Many sampling/optimization methods have been proposed to attack this difficult problem, and they all have advantages and disadvantages. Recently, an efficient global optimization method called conformational space annealing (CSA) has been successfully applied to various problems[12–17] including template-based protein modeling[18] and conformational sampling of various macromolecules with the CHARMM package.[19]

In this work, we combine CSA and DFA to obtain low-energy conformations of small single-domain proteins. We propose a full-atom version of DFA, which we call CHARMM-DFA. There are two advantages of the full-atom model over coarse-grained models. First, in the full-atom model, proper stereochemistry is maintained and severe steric clashes are avoided automatically at the level of the energy function. The other is that the full-atom model allows one to use additional energy terms such as the distance-scaled finite ideal-gas reference state statistical energy, called as DFIRE energy, and the effective energy function 1 (EEF1) solvation energy.[20,21] All energy terms are prepared to be analytically differentiable so that local energy minimization can be executed with the energy gradient.

Another important aspect of successful protein modeling is to optimize the parameters of the energy function, so that efficient conformational sampling of the energy function can naturally lead to native-like structures of a protein. To the best of our knowledge, except a few multiobjective approaches,[22–24] combination of various energy terms in a linear fashion is widely adopted as the total energy function. Parameter optimization of a linear objective function can be carried out by seeking the optimal relative weights between energy terms, and the most straightforward way is to perform a grid search with a set of many pregenerated decoy structures. However, when dealing with a differentiable energy function, the decoys which are local minima of a parameter set are not local minima of another parameter set. Moreover, overall shape of the energy landscape can be significantly altered upon the variation of the parameters. In this study, we employed local as well as global CSA searches to generate low-energy local minimum decoys including native-like ones.[25,26] Sampling by local CSA is limited to low-energy native-like conformations while no such restraint is applied to global CSA. The energy gap between the lowest-energy conformations from global CSA and local CSA is considered for parameter optimization and linear approximation of the total energy upon small variation of parameters[25,26] is used for efficient parameter optimization. The parameter optimization was carried out on 30 small single-domain proteins. To check the transferability of the optimized parameter set, additional tests were carried out on targets of the 8th Critical Assessment of protein Structure Prediction (CASP8).[27]

The results on the benchmark set show that CSA can sample lower energy conformations than SA with an equivalent amount of computational resources. Prediction accuracy is improved by using the full-atom modeltogether with empirical potentials. Tests on the free-modeling targets of the CASP8 show that predictions made by the current method are comparable and quite complementary to top performing template-free modeling (FM) methods.

MATERIALS AND METHODS

Model representation

In this study, proteins are represented either by Cα atoms or by full atoms including all hydrogens. In the coarse-grained model, Cα atoms are concatenated by harmonic springs with equilibrium length of 3.8 Å. The topology of the full-atom model is generated according to the representations of the CHARMM22 force field.[28]

Fragment selection

To perform a fair evaluation of our protocol for *ab initio* protein structure prediction, a 9-residue fragment library was constructed by carefully removing redundancy. The initial fragment library was built from 5495 nonredundant (NR) protein structures prepared by the PISCES server[29] with the sequence identity cutoff of 30%, the experimental resolution cutoff of 2.0 Å and the R-factor cutoff of 0.25 dated on July 25, 2009. For proper benchmarking, additional homologous proteins were removed from the library either if their sequence identities are over 20% with a target protein sequence or if their homology can be detected by the standard PSI-

BLAST search[30] with the E-value greater than 0.05. For CASP8 FM targets, the proteins released after May 1, 2008, were also filtered out from the library. The sequence profile of each selected protein was calculated by PSI-BLAST based on the NR-sequence database with the E-value cutoff of 0.001.

For each 9-residue sliding window of a target protein sequence, its sequence profile was compared with those in the library. The correlation coefficient was used to identify the best fragments in the library. Top ranked fragments selected for the jth 9-residue-sliding window of the target sequence are denoted by $F_i(j)$, with $i = 1,\ldots,N^{\text{fragment}}$. N^{fragment} is set to 20 or 40 depending on the nature of each DFA energy term.[9,10]

Energy functions

Dynamic fragment assembly energy

The novel idea of Sasaki and Sasai for fragment assembly is first to extract much structural information stored in the fragment library and then to transfer the information into a residue-position specific energy terms. DFA energy function is derived from the statistical tendency of the fragment library, and it can be expressed as follows:

$$e_{\text{DFA}} = w_{\text{DFA\_dist}}\, e_{\text{DFA\_dist}} + w_{\text{DFA\_angle}}\, e_{\text{DFA\_angle}} + w_{\text{DFA\_nn}}\, e_{\text{DFA\_nn}} + w_{\text{DFA\_beta}}\, e_{\text{DFA\_beta}},$$

where $e_{\text{DFA\_dist}}$ and $e_{\text{DFA\_angle}}$ are to assimilate the local structure of a model to its corresponding fragments. $e_{\text{DFA\_nn}}$ represents the preferred packing environment around each residue by the number of neighboring Cα atoms from the fragment library, and $e_{\text{DFA\_beta}}$ represents the tendency of beta-sheet formation estimated from predicted contact information.[31] In this work, we used the identical energy functions as in Ref. 9, except for $e_{\text{DFA\_nn}}$ which is transformed into a continuous function from its original form of discrete one. Detailed description on DFA energy components are presented in the Supporting Information.

Full-atom DFA energy

For full-atom simulations, a number of physics-based energy terms and the all-atom statistical potential of DFIRE[21] were incorporated with DFA to improve the accuracy of structure prediction. They are introduced to capture physical properties that cannot be properly reflected in the fragment-based energy function. The complete form of the full-atom DFA energy function can be expressed as follows:

$$E = e_{\text{DFA}} + w_{\text{DFIRE}}\, e_{\text{DFIRE}} + e_{\text{physics}} = w_{\text{DFA\_dist}}\, e_{\text{DFA\_dist}} + w_{\text{DFA\_angle}}\, e_{\text{DFA\_angle}} + w_{\text{DFA\_nn}}\, e_{\text{DFA\_nn}} + w_{\text{DFA\_beta}}\, e_{\text{DFA\_beta}} + w_{\text{DFIRE}}\, e_{\text{DFIRE}} + e_{\text{physics}}$$

After primitive calculations of trials and errors, initial relative weight of each energy term to e_{physics} is set as $w_{\text{DFA\_dist}} = 20$, $w_{\text{DFA\_angle}} = 20$, $w_{\text{DFA\_nn}} = 20$, $w_{\text{DFA\_beta}} = 20$, and $w_{\text{DFIRE}} = 10$.

Physics-based energy terms

A number of physics-based potentials are introduced to take into account of physical properties of amino acids, such as maintaining proper bond angles and bond lengths, and avoiding steric clashes. The stereochemistry of CHARMM22 topology is used.[28,32] Van der Waals energy, electrostatic energy, and EEF1 implicit solvation energy with the CHARMM22 parameters are also included.[20] EEF1 solvation model has been effectively used in various areas of protein studies.[33–36] Throughout this study, the weights of physics-based energy terms are kept fixed at 1. Now,

$$e_{\text{physics}} = e_{\text{stereochemistry}} + e_{\text{elec}} + e_{\text{EEF1}} = e_{\text{bond}} + e_{\text{angle}} + e_{\text{torsion}} + e_{\text{improper}} + e_{\text{vdW}} + e_{\text{elec}} + e_{\text{EEF1}}.$$

DFIRE all-atom statistical potential term

DFIRE statistical potential is shown to be effective to identify the native structure of a protein out of many decoys.[21,37] Successful applications include loop modeling[38] and remodeling of terminal regions.[39] In DFIRE, a total of 167 residue-type specific heavy atoms are considered to measure pairwise inter-atomic distances, which are grouped into 25 discrete bins. In this study, to make the potential term analytically differentiable, a cubic spline smoothing algorithm is used.

Conformational space annealing (CSA)

CSA[15,40–43] is a global optimization algorithm which can be viewed as a generalized genetic algorithm.[44] The CSA has been successfully applied to a variety of optimization problems[12,15–17] including recent success in high-accuracy template-based modeling.[8,45] CSA searches a wide range of conformational space in the early stages of simulation and narrows down the range of conformational search as the simulation progresses. The diversity of conformational sampling is controlled by the value of distance cutoff, D_{cut}, which determines the similarity of two given conformations.

The CSA starts with a predetermined number of randomly generated and energy minimized protein conformations. In this study, 50 conformations were generated initially. This initial bank of conformations which is called "firstbank," represents the entire conformational space of the local minima and is kept intact throughout the simulation to provide moves for random perturbation during the search to achieve diverse sampling. Initially, the current bank is identical to "firstbank."

A number of conformations, 20 in this study, are selected from the bank as seed conformations to generate trial conformations. Trial conformations are generated by

performing crossover between internal coordinates of a seed and those of a randomly selected conformation either from the current bank or from the firstbank. In this study, for each seed, we have generated 20 trial conformations which amount to a total of 400 conformations. All 400 trial conformations are energy minimized, and this is the most time-consuming part of the algorithm. As 400 energy-minimization calculations are independent from each other, CSA algorithm can be easily executed in a highly parallelized fashion with a good scaling performance in the number of CPUs.[46]

All energy-minimized trial conformations are compared with those in the current bank, and the bank is updated according to the following procedures. If the energy of the trial conformation, s, is higher than the maximum energy of the bank, it is discarded. If not, the closest conformation to s, say t, in the bank is identified by calculating distances between s and all bank conformations. If the distance, $d(s, t)$ is closer than D_{cut}, then s is treated as rather similar to t. In this case, if the energy of s is lower than that of t, s replaces t. Otherwise, s is discarded. If the distance, $d(s, t)$ is larger than D_{cut}, the trial conformation s forms a new cluster itself, and it replaces the highest energy conformation in the bank. As mentioned above, in the subsequent stages of CSA, by reducing the value of D_{cut} in a slow fashion, diverse low-energy conformations can be maintained in the final bank conformations. The value of D_{cut} is initially set to the half of the average inter-distance of the firstbank, $D_{ave}/2$, and it is gradually reduced to $D_{ave}/5$ Additional details of the algorithm can be found elsewhere.[15,41,47] Three kinds of distance measure, RMSD, Manhattan distance with torsion angles and TMscore[48] were tested, and TMscore provided the most diverse distribution of sampling. Throughout this article, TMscore is used as the distance measure.

When CSA procedure is completed, the final bank contains a number of diverse low-energy local-minimum conformations, 50 in this study. In the CASP,[27] template-FM targets are evaluated considering up to five models. Accordingly, we have selected five models as follows; the lowest energy conformation is selected as the first model and the other four representative conformations are selected to maximize the sum of inter-distances between the final five conformations.

Local CSA procedure to sample low-energy native-like conformations

To devise the potential function to guide sampled conformations toward native-like structures, one should compare the energies of native-like and non-native conformations. We assume that relevant low-energy basins of a given potential function can be efficiently sampled by the CSA procedure described above, which we call global CSA. To sample low-energy native-like conformations, a restrained conformational search procedure that we call local CSA is introduced.[26] During the local CSA, only native-like conformations whose RMSD value is below a preset cutoff value are sampled. It should be noted that direct copy of the native conformation typically results in a rather high-energy structure, and our intention is to generate low-energy conformations in the native-like basin.

Local CSA differs from the global CSA in two aspects. (1) The *firstbank* of local CSA is initially prepared to contain the native backbone structure with randomized sidechain conformations rather than using completely randomized structures as in the global CSA. Local energy minimizations of these structures are performed to generate the *firstbank* of local CSA. (2) All trial conformations are prescreened to select only near-native structures. A near-native structure is defined by the RMSD cutoff of 3.5 Å from the native structure. If the RMSD value of a trial conformation is over the cutoff value, it is discarded regardless of its energy value during the bank update procedure.

It should be noted that the firstbank of the local CSA may contain non-native conformations whose RMSD values are over the cutoff. However, the prescreening procedure drives the simulation so that only native-like conformations remain in the bank after a number of update procedures. These two aspects allow us to keep only native-like conformations in the bank, and the local CSA procedure provides a number of low-energy near-native conformations.

Linear approximation of conformational energy

One of the difficult aspects of parameter optimization lies in that local energy minimum conformations obtained by a set of parameters are, in general, no longer energy minima of another set. That is, the energy landscape varies with parameters. To obtain relevant energy landscape after the parameters are altered, conformations sampled from previous optimization steps should be reminimized, and repeated energy reminimization requires considerable computational resources. However, it is known that if changes of parameters are small, the energy landscape of local energy minima can be well approximated by a linear equation.[25,26] We consider a local energy minimum structure $\vec{x}$, from $E(\vec{w}; \vec{x})$. Upon the perturbation of $\vec{w}' = \vec{w} + \delta\vec{w}$, the structure of the local minimum changes to $\vec{x}' = \vec{x} + \delta\vec{x} = \vec{x} + \delta\vec{w}\frac{\partial\vec{x}}{\partial\vec{w}} + (\delta\vec{w})^2\frac{\partial^2\vec{x}}{\partial\vec{w}^2} + \cdots$. When $\delta\vec{w}$ is small, up to the linear order of $\delta\vec{w}$, the new minimum energy value can be estimated as $E(\vec{w}', \vec{x}') \cong E(\vec{w}; \vec{x}) + \delta\vec{w}\frac{\partial E}{\partial\vec{w}} + \delta\vec{w}\frac{\partial\vec{x}}{\partial\vec{w}}\frac{\partial E}{\partial\vec{x}}$. Since $\frac{\partial E}{\partial\vec{x}} = 0$, and E is linear in $\vec{w}$, the local-minimum energy value of the new parameter set can be written as $E^{new} = E^{old} + \sum_i (w_i^{old} - w_i^{new})e_i^{old}$.

Parameter optimization by linear programming

The energy gap between the lowest energy from global CSA and the lowest energy from native-like conformation generated by local CSA is calculated for each protein in the benchmark set. The sum of thus obtained energy gaps is used for parameter optimization. Because each protein has a different energy scale, i.e., the DFA energy is approximately proportional to the length of a protein, the energy gap is normalized with respect to the lowest energy of each protein. The change of the sum of normalized energy gaps during the parameter optimization can be written as follows:

$$\Delta e_{\text{gap}} = \sum_k \left\{ e_{\text{gap},k}\left(\left\{w_i^{\text{new}}\right\}\right) - e_{\text{gap},k}\left(\left\{w_i^{\text{old}}\right\}\right)\right\} / e_k^{\text{lowest global}}\left(\left\{w_i^{\text{old}}\right\}\right)$$
$$= \sum_k \sum_i \left(e_{k,i}^{\text{lowest global}} - e_{k,i}^{\text{lowest local}}\right)\left(w_i^{\text{new}} - w_i^{\text{old}}\right) / e_k^{\text{lowest global}}\left(\left\{w_i^{\text{old}}\right\}\right),$$

where $e_{k,i}^{\text{lowestglobal}}$ and $e_{k,i}^{\text{lowestlocal}}$ are the ith energy term of the lowest energy sampled by the global and local CSA of the kth protein. To find an optimized parameter set, we have imposed an additional constraint that no energy gap from the benchmark proteins should deteriorate. This condition is expressed as $\Delta e_{\text{gap},k} = \sum_i \left(e_{k,i}^{\text{lowestglobal}} - e_{k,i}^{\text{lowestlocal}}\right)\left(w_i^{\text{new}} - w_i^{\text{old}}\right) \geq 0$ for all k. In summary, the parameter optimization problem is set to maximize the sum of energy gaps for 30 benchmark proteins while satisfying 30 inequality constraints. This class of problem can be solved with a linear programming method, and we have used "PuLP," a python linear programming module.[49]

Benchmark sets

To compare the performance of the current approach with existing results, we have selected 30 small single domain proteins suitable for *ab initio* modeling studies, 20 from I-TASSER study[2] and 10 from ROSETTA study.[1]

Table I
TMscores of the Lowest-Energy Conformations by CHARMM-DFA/CSA and Coarse-Grained DFA Simulations are Shown

	ID	N_{res}	%α[a]	%β[b]	DFA/CSA[c]	DFA/SA[d]	CHARMM-DFA/CSA[e]
Rosetta set	1csp_A	67	4	55	0.321	0.343	0.250
	1di2_A	69	46	33	0.638	0.375	0.534
	1r69_A	61	63	0	0.815	0.837	0.890
	1shf_A	59	5	45	0.280	0.283	0.283
	d1dtja_	74	39	27	0.590	0.707	0.817
	d1mkya3	81	32	24	0.332	0.278	0.314
	d1mlaa2	70	34	37	0.357	0.374	0.413
	d1o2fb_	77	38	27	0.323	0.332	0.344
	d1tiga_	88	35	35	0.518	0.578	0.617
	d2reba2	60	61	20	0.406	0.415	0.362
I-TASSER set	1ah9_A	63	4	44	0.245	0.263	0.229
	1aoy_A	65	49	12	0.659	0.650	0.648
	1b4b_A	71	38	35	0.389	0.411	0.692
	1bq9_A	53	16	26	0.336	0.317	0.431
	1fo5_A	85	24	24	0.487	0.474	0.504
	1gjx_A	77	0	29	0.229	0.340	0.227
	1gpt_A	47	23	36	0.292	0.312	0.317
	1hbk_A	89	62	0	0.266	0.283	0.274
	1itp_A	68	33	36	0.495	0.507	0.472
	1kjs_A	74	60	0	0.339	0.336	0.317
	1kvi_A	68	35	25	0.610	0.625	0.408
	1ne3_A	56	0	46	0.225	0.188	0.201
	1of9_A	77	68	0	0.313	0.311	0.452
	1pgx_A	59	23	47	0.439	0.310	0.305
	1sro_A	71	5	39	0.204	0.209	0.302
	1ten_A	87	0	55	0.295	0.357	0.273
	1tfi_A	47	0	36	0.412	0.418	0.460
	1vcc_A	76	14	34	0.338	0.411	0.368
	2cr7_A	60	80	0	0.541	0.502	0.404
	2f3n_A	65	70	0	0.521	0.510	0.547
	Average				0.407	0.409	0.422
	Number of cases with TMscore > 0.4				13	13	15

[a]Percentage of helical residues.
[b]Percentage of extended residues.
[c]The lowest energy conformation from the 100 independent runs.
[d]The lowest energy conformation from the 400 independent runs.
[e]Result obtained by a single run.

Proteins of chain length less than 90 residues are selected. The selected benchmark set is also used as a training set for parameter optimization of the full-atom DFA energy. It should be noted that one of our benchmark protein, 1di2, is overlapped with the training set of BETAPRO.

CASP8 test set

To validate the generality of the current methodology and the transferability of optimized weight parameters, we have tested the current protocol against 14 template-FM targets of CASP8.[50]

RESULTS AND DISCUSSION

Sampling efficiency of CSA

To investigate if CSA allows us to sample low energy conformations more efficiently than SA Langevin MD,[9,10] the energy landscapes obtained from the two methods with the initial energy parameter set are compared. For each protein in the benchmark set, the final conformations of independent 400 SA simulations and energy minimum conformations of 100 CSA simulations started with separate random numbers are compared. The numbers of simulations, 400 and 100, were chosen to obtain equivalent numbers of energy evaluations for the two methods.

It is found that, for "all" 30 proteins in the benchmark set listed in Table I, CSA provides lower energy conformations than SA Langevin MD (data not shown). Two typical energy landscapes are shown in Figure 1, where models' TMscores from the native structure are plotted along with their DFA energies. For the case of 1aoyA, as shown in Figure 1 A, DFA energy and TMscore are well correlated and lower energy conformations generated by CSA are more native-like than those from SA. However, sampling lower energy conformations does not necessarily results in more accurate protein models, as illustrated in Figure 1(B). For d1dtjaa, although the overall correlation between DFA energy and TMscore from SA is quite desirable, CSA found lower energy conformations, near TMscore 0.6.

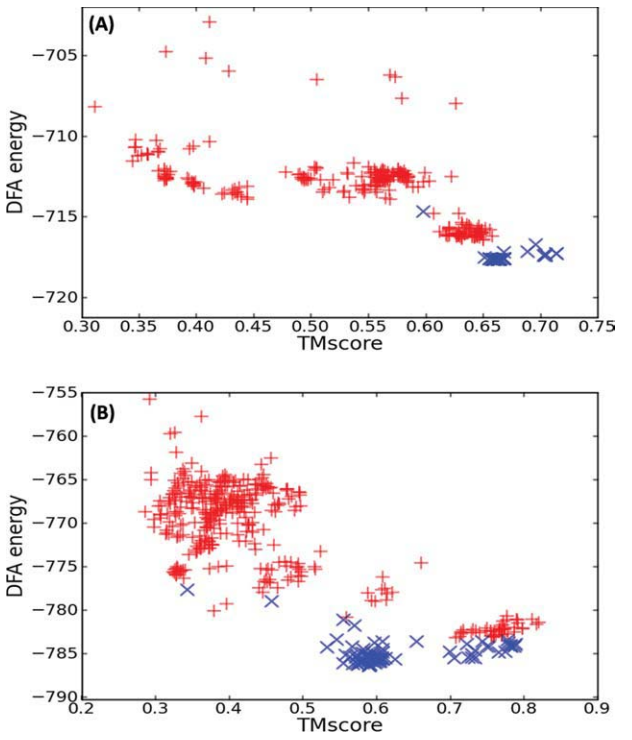

Figure 1

Energy landscapes of 1aoyA (**A**) and d1dtja_ (**B**) are shown. Red plus symbols represent the final snapshots from 400 SA Langevin MD simulations. Blue cross symbols represent the energy minimum conformations obtained from 100 CSA runs.

The average TMscores of the lowest energy conformations of 30 benchmark proteins by CSA and SA are 0.407 and 0.409, respectively, which means that there is no statistically significant difference between the two results. Apparently, sampling lower energy conformations alone does not guarantee to generate more native-like conformations, and this is simply due to the imperfectness of

Table II

The Effect of Potential Energy Terms on the Secondary Structure and Solvent Accessibility of Proteins are Shown

		SS accuracy(%)				ACC accuracy(%)[a]		
	Average TMscore	Helix	Strand	Coil	Total	Buried	Exposed	ACC correlation[b]
S1[c]	0.399	76.9	63.6	82.0	74.8	56.0	85.6	0.632
S2[d]	0.409	89.9	71.6	74.5	78.6	50.4	90.1	0.681
CHARMM-DFA[e]	0.422	91.9	73.3	72.1	78.8	60.6	88.2	0.713
PSIPRED		87.7	84.3	81.4	84.0			
ACCpro						87.2	72.5	0.762

[a]A residue is defined as buried when its solvent accessible area is less than 25% of the fully exposed state.
[b]Predicted solvent accessible area was obtained by interpolating 20 ACCPRO calculations with varying cutoff values.
[c]$e_{S1} = e_{DFA} + e_{stereochemistry}$
[d]$e_{S2} = e_{S1} + e_{elec} + e_{EEF1}$
[e]$e_{CHARMM\text{-}DFA} = e_{S2} + e_{DFIRE}$

the DFA energy function and/or the coarse-grained representation of proteins.

Improvement of prediction accuracy with full-atom representation

In this section, we consider full-atom representation of proteins for possible structure modeling improvement. We intend to achieve proper balance in a protein model between the residue-position-specific DFA potential and general physics-based and statistical potentials. Maintaining such balance may help us to generate better protein models while taking advantage of the prediction power generated by fragments. If the weight of DFA potential is too weak compared with the physics-based and statistical potentials, our approach would be close to the pure *ab initio* modeling, which does not use any fragments or templates. On the other hand, if the relative influence of DFA potential is rather strong, the results would be similar to the previous coarse-grained DFA simulation. The initial weight parameters were set as follows: keeping the physics-based terms fixed, DFA potential is increased 20 folds and DFIRE potential is increased 10 folds from their original scale.

We have added physics-based terms of CHARMM force field and DFIRE statistical potential to the DFA potential, which we denote as the CHARMM-DFA model. CSA sampling was performed with CHARMM-DFA. Table I shows the simulation results of CHARMM-DFA/CSA with the initial weight parameter set along with the two coarse-grained ones, DFA/CSA and DFA/SA. DFA/CSA and DFA/SA simulations were performed 100 and 400 times, respectively, while CHARMM-DFA/CSA results are obtained from a single CSA run. The overall model quality is improved by integrating the physics-based energy terms and full-atom representation of the poly-peptide chain. The average TMscore of the lowest-energy conformations increased to 0.422 from 0.407 and 0.409. It should be noted that the average TMscore itself is somewhat misleading since protein models with TMscore < 0.3 can be considered as meaningless.[51] Considering only meaningful modeling cases, say TMscore > 0.4, 15 proteins were successfully

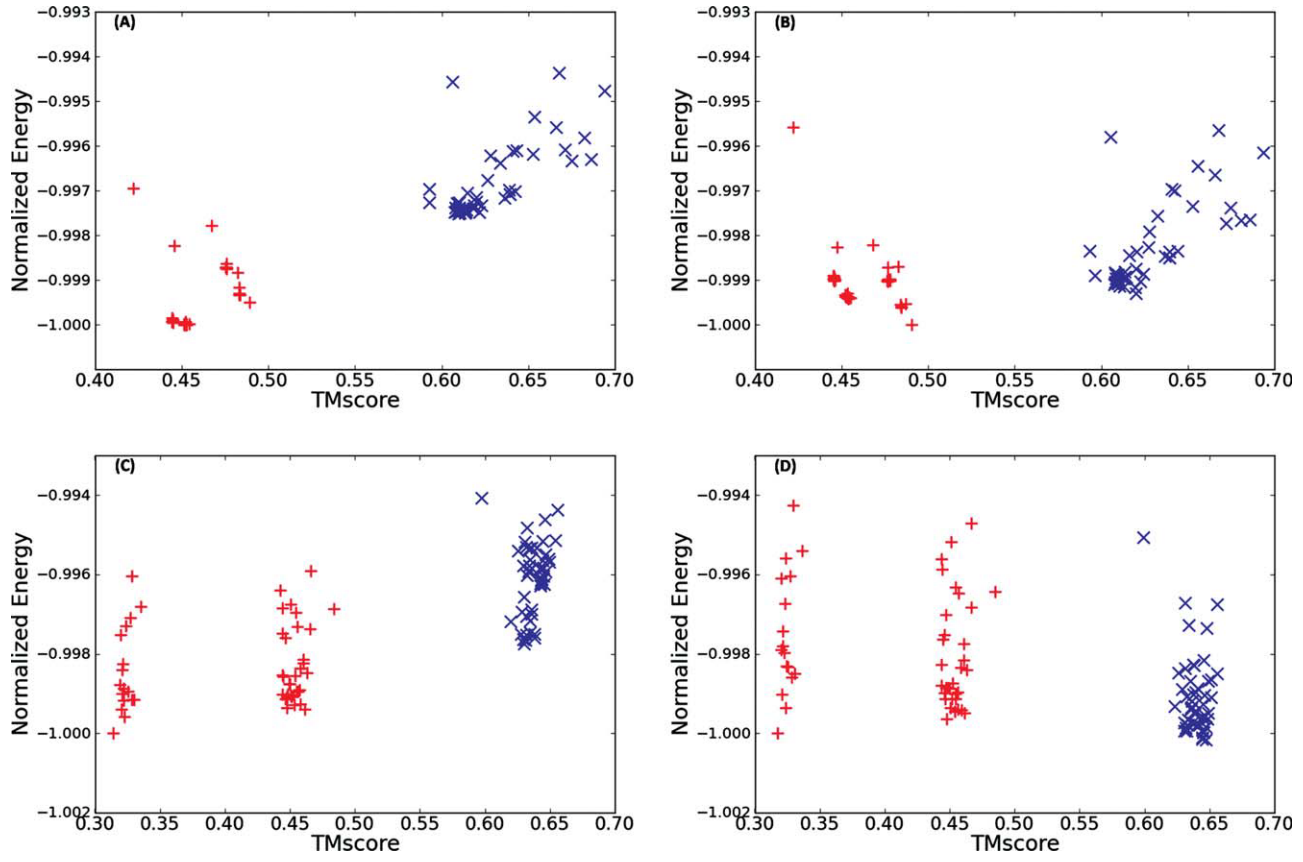

Figure 2

Changes of the energy landscapes by optimization of weight parameters are displayed. (**A**) and (**C**) represent the energy landscapes of 1of9_A and d1mkya3 with the original parameters, respectively. Red plus symbols are generated from the global CSA runs and blue cross symbols from the local CSA. The new landscapes after the parameter optimization are depicted in (**B**) and (**D**). Energy values are normalized with the lowest energy value from the global CSA to compare the relative change of energy gap. Therefore, scales of (A) and (B), (C) and (D) can be considered the same.

predicted with CHARMM-DFA/CSA, which is improved from 13.

Full-atom details with additional potential term lead to the increase of computational cost. The average wall-clock times for a single coarse-grained DFA/CSA run and CHARMM-DFA/CSA run were about 1 h and 15 h with 8 AMD Opteron 275 processors, respectively. However, it should be noted that the result of the coarse-grained DFA/CSA is taken from the lowest energy model out of 100 runs while that of CHARMM-DFA/CSA is from a single run.

Incorporation of full atom details to DFA potential will certainly introduce additional ruggedness in the energy landscape. To investigate if CHARMM-DFA/CSA simulations are critically trapped in local basins far from the coarse-grained basins, correlation coefficients between DFA/CSA, DFA/SA, and CHARMM-DFA/CSA results in Table I are calculated. If the correlation coefficients are high, the overall energy landscape of the potential can be considered to be governed by the DFA potential and addition of physics-based and statistical potentials provide additional positive influence to the final model structure. The correlation coefficients between DFA/CSA, DFA/SA, and CHARMM-DFA/CSA are 0.813 and 0.840, respectively, indicating that increased ruggedness in the energy landscape did not significantly affect the sampling efficiency.

It should be noted that there is an issue of double-counting from various potential terms. However, all available potential functions have their own limitations and their effects can compensate for each other to form more ideal energy landscape. Among the DFA potentials, the effect of long range interactions, the number of neighboring residues and beta-formation terms may overlap with that of electrostatic and solvation terms. Therefore, given the possible double-counting in energy terms,

Table III
Cα-RMSD Values (Å) of CHARMM-DFA Models with Initial and Optimized Parameters are Shown Along with ROSETTA and I-TASSER Results

					Initial		Optimized		ROSETTA		I-TASSER	
	ID	N_{res}	%α	%β	Lowest	Best of five	Lowest	Best of five	Round2[a]	Best of five l argest cluster[b]	First cluster[c]	Best of top five cluster
Rosetta set	1csp_A	67	4	55	10.5	9.8	7.9	7.8	4.7	5.1	2.1	2.1
	1di2_A	69	46	33	3.6	3.0	3.7	3.7	2.6	1.9	2.3	2.3
	1r69_A	61	63	0	1.1	1.1	1.2	1.2	2.1	1.2	1.9	1.9
	1shf_A	59	5	45	11.4	11.3	5.1	4.0	10.8	10.9	1.7	1.7
	d1dtja_	74	39	27	2.4	2.3	2.9	2.9	1.2	1.8	1.9	1.7
	d1mkya3	81	32	24	10	6.3	3.6	3.4	6.3	3.7	5.2	4.5
	d1mlaa2	70	34	37	7.1	6.7	7.1	6.0	8.7	7.2	2.8	2.7
	d1o2fb_	77	38	27	11.6	10.8	5.3	5.3	N/A	10.3	7.1	5.2
	d1tiga_	88	35	35	3.5	3.5	3.3	3.3	4.1	3.5	7.7	4.4
	d2reba2	60	61	20	9.6	9.6	8.2	7.9	1.2	2.1	5.6	4.7
I-TASSER set	1ah9_A	63	4	44	11.8	11.8	12.1	12.1			4.3	2.8
	1aoy_A	65	49	12	3.4	2.6	3.0	2.8			4.5	2.7
	1b4b_A	71	38	35	2.5	2.5	9.9	1.9			6.4	5.6
	1bq9_A	53	16	26	6.4	6.2	6.4	4.1			7.3	5.0
	1fo5_A	85	24	24	5.0	4.6	4.2	4.1			3.8	3.8
	1gjx_A	77	0	29	13.9	13.8	12.7	11.7			6.9	5.6
	1gpt_A	47	23	36	8.4	7.4	8.8	7.1			5.2	3.8
	1hbk_A	89	62	0	15.2	12.7	7.1	6.8			3.5	3.5
	1itp_A	68	33	36	8.1	6.5	8.9	7.5			10.9	4.5
	1kjs_A	74	60	0	12.6	12	11.5	11.2			8.5	5.7
	1kvi_A	68	35	25	5.1	5.1	3.0	3.0			2.0	2.0
	1ne3_A	56	0	46	9.3	9.2	8.9	8.3			4.6	4.6
	1of9_A	77	68	0	8.7	5.4	7.7	7.4			3.6	3.6
	1pgx_A	59	23	47	11.5	11.3	4.4	4.4			3.1	3.1
	1sro_A	71	5	39	11.5	11.5	11.4	11.4			3.4	3.0
	1ten_A	87	0	55	10.8	10.8	9.3	8.7			1.6	1.6
	1tfi_A	47	0	36	3.8	3.8	3.1	3.1			4.6	4.0
	1vcc_A	76	14	34	7.8	7.7	9.7	9.7			5.7	5.7
	2cr7_A	60	80	0	8.1	5.4	4.4	2.5			4.5	2.6
	2f3n_A	65	70	0	3.2	2.7	3.5	3.2			1.8	1.8
Average[d]		69	32	28	7.9 (0.422)	7.2 (0.449)	6.6 (0.453)	5.9 (0.493)			4.5 (0.577)	3.5 (0.613)
Rosetta set only[e]					7.1 (0.482)	6.5 (0.515)	4.8 (0.541)	4.5 (0.571)	4.6	4.8	3.8 (0.627)	3.1 (0.663)
# of good predictions (RMSD<4.0Å)	All set				8	8	9	12			14	18
	Rosetta set only				4	4	5	6	4	6	6	6

[a]High-resolution models refined from low-resolution models.
[b]Centers of the five largest clusters from low-resolution models.
[c]The cluster with the highest conformational density.
[e]Average TMscores are shown in parentheses when available.

we intend to obtain their optimal relative weights to generate more accurate model structures.

Effects of physics-based terms and DFIRE on protein structure prediction

To identify the effects of newly added potential energy terms, electrostatic, EEF1 solvation, and DFIRE statistical energies, we performed two additional simulations on the benchmark set of 30 proteins in Table I. The first simulation, S1, was performed with stereochemistry terms added to DFA terms, $e_{S1} = e_{DFA} + e_{stereochemistry}$. In the second simulation, S2, polar energies represented by electrostatic and EEF1 solvation energies were additionally added, $e_{S2} = e_{S1} + e_{elec} + e_{EEF1}$. In Table II, the effects of electrostatic, EEF1 solvation, and DFIRE terms on structure prediction is displayed with emphasis on the secondary structure and solvation states. The average accuracies of secondary structure and solvent accessible area prediction servers, PSIPRED,[52] and ACCpro,[53] are also shown together for comparison. Average TMscores improve, although not significantly, as additional energy terms are introduced. With respect to the secondary structure, both polar and DFIRE terms facilitate helix and strand formation. When polar terms, i.e., electrostatic and EEF1 solvation terms are introduced, prediction accuracies of helical and strand regions improve by 13.0% and 8.0%, respectively. Especially, the helical region is predicted better than widely used PSIPRED. On the other hand, the accuracy of the coil region deteriorates as the potential terms are added, indicating that polar terms favor helix and strand structures over coil structures. Similar changes are observed when DFIRE term is added. Accuracies of helices and strands improve by 2.0% and 1.6%, respectively, while that of coils worsens by 2.4%.

Regarding solvent accessibility, EEF1 and DFIRE terms seem to play opposite roles. The EEF1 term drives some of the buried native residues to be exposed while DFIRE potential helps the residues to be buried. Interestingly, although the EEF1 term sometimes worsens the accuracy of the buried residues, it improves the correlation between accessible surface areas of corresponding residues from the native and a model as shown in Table II. When both polar and DFIRE potentials are used, the best results were obtained. When we compare CHARMM-DFA with S1, the accuracies of solvent accessibility of buried and exposed residues improve from 56.0% to 60.6% and from 85.6% to 88.2%, respectively, and the correlation of solvent accessible area reaches to 0.713. The relatively poor values of solvent accessibilities of buried residues in our models suggest that our models are packed less optimally than native structures. Therefore, to further improve the quality of models, one may consider rather stringent ways to incorporate the solvent accessibility information into the model.

Improvement from parameter optimization

As described in the Materials and Methods section, there are five weight parameters to be optimized in the current full-atom DFA potential energy function, w_{DFA_dist}, w_{DFA_angle}, w_{DFA_nn}, w_{DFA_beta}, and w_{DFIRE}. The

Table IV
Results on the Template-Free Modeling Targets of CASP8 are Shown. Numbers in boldface correspond to the highest TMscores out of all predictions of CASP8

				CHARMM-DFA					
				Best of five models					
ID	N_{res}	%α	%β	Initial	Optimized	BEST[a]	MUFOLD-MD	BAKER-ROBETTA	Zhang-server
T0397-D1	82	9	39	0.230	0.242	0.337	0.310	0.244	0.262
T0405-D1	72	81	0	0.339	0.345	0.611	0.354	0.352	0.373
T0405-D2	208	38	25	0.241	0.227	0.443	**0.443**	0.321	0.300
T0416-D2	57	75	0	**0.719**	**0.648**	0.612	0.612	0.565	0.528
T0443-D1	66	79	0	**0.622**	**0.610**	0.543	0.543	0.441	0.468
T0443-D2	61	30	36	0.184	0.185	0.416	0.332	0.337	0.351
T0460-D1	81	40	27	0.252	0.211	0.589	0.387	0.343	0.316
T0465-D1	98	52	11	0.315	**0.412**	0.371	0.293	0.347	0.363
T0476-D1	86	35	12	0.299	0.325	0.463	0.431	0.381	0.398
T0482-D1	71	28	45	0.287	0.253	0.688	0.376	0.570	0.446
T0496-D1	120	40	21	0.347	0.269	0.360	0.279	0.343	0.317
T0510-D3	44	18	14	0.315	0.283	0.407	**0.407**	0.268	0.25
T0513-D2	69	35	29	0.587	**0.645**	0.635	0.600	0.635	0.507
Average				0.364	0.358	0.498	0.413	0.396	0.375
TMscore > 0.4[b]				0.441	0.440		0.468	0.449	0.409
No. of best models[c]				2	4		2	0	0

[a]Best TMscores out of all submitted models including human and server predictions.
[b]Average TMscore of the 7 targets for which meaningful predictions of TMscore > 0.4 are obtained by any of the five methods (CHARMM-DFA/CSA with initial and optimized parameters, MUFOLD-MD, BAKER-ROBETTA, and Zhang-Server).
[c]Number of the best predictions including human and server predictions in CASP8.

rationale for the current parameter optimization procedure is to make the energies of native-like conformations as low as possible relative to the energies of the non-native-like conformations. This is accomplished by maximizing the sum of normalized energy gaps over all benchmark cases while satisfying the constraint that the energy gap of each protein does not deteriorate. The definitions of energy gap and normalized energy gap are described in the Materials and Methods section. This condition can be formulated as follows:

$$\text{Maximize } e_{\text{gap,total}} = \sum_{k=1}^{N_b} \Delta e_{\text{gap},k}$$

$$\text{subject to } \Delta e_{\text{gap},k} = \sum_i \left(e_{k,i}^{\text{lowest global}} - e_{k,i}^{\text{lowest local}} \right) \left(w_i^{\text{new}} - w_i^{\text{old}} \right) / e_{k,i}^{\text{lowest global}} \geq 0 \ \forall k,$$

where N_b is the number of benchmark proteins, 28 in this study. Two proteins, where near native structures were already found from the global CSA, 1r69 and d1dtja_, were not considered in the parameter optimization step. Maximizing a term which is expressed as the linear sum of several terms, subject to linear inequality constraints can be solved with the well-established mathematical method called linear programming. In this study, we used "PulP," the python linear programming package to optimize our problem.

The optimized parameters are $w_{\text{DFA\_dist}} = 13.3$, $w_{\text{DFA\_angle}} = 22.45$, $w_{\text{DFA\_nn}} = 26.48$, $w_{\text{DFA\_beta}} = 13.3$, and $w_{\text{DFIRE}} = 12.0$.

Two representative examples of energy landscape before and after parameter optimization are shown in Figure 2. Figure 2(A/B,C/D) represent the energy landscapes of 1of9_A and d1mkya3 with the original/optimized parameters, respectively. We observe that the energy gap between the lowest energy from the global CSA and that of the local CSA, $\min(\{E_{\text{global\_CSA}}\}) - \min(\{E_{\text{local\_CSA}}\})$, improves for both cases by parameter optimization. It should be noted that the new energy landscapes with the optimized parameters are generated based on the decoy structures obtained with the old parameters, and the actual energy landscapes of the new parameter set not fully represented by these decoys. The energy landscape represented by a small number of decoys, which are the local minima generated from the old parameter set, may not cover all relevant low-energy basins, which can be searched only by extensive conformational sampling with the new parameter set. For more thorough parameter optimization, an iterative procedure is necessary, while decoy structures are accumulated.[26]

With the optimized parameter set, global CSA simulations were carried out again for the benchmark set. The results along with existing *ab initio* prediction studies are shown in Table III. It should be noted that our weight parameters are optimized with this benchmark set, and it can give beneficial effects to our results. The average TMscore of the best model among five representative models by the CHARMM-DFA method with optimized parameters improved from 0.449 to 0.493, in RMSD measure, from 7.2 Å to 5.9 Å. It is noticeable that most of reasonably predicted targets, approximately RMSD $\leq$ 4.0 Å, in the initial run are also predicted accurately with the optimized parameter set. This indicates that the restraint applied in the parameter optimization that no energy gap from the benchmark proteins deteriorate successfully worked.

While preserving the accuracy of the good models, a number of previously incorrectly modeled targets showed dramatic improvement as in 1shf_A, d1mkya3, d1o2fb_, and 1pgx_A. For 2cr7_A and 1kvi_A, highly accurate

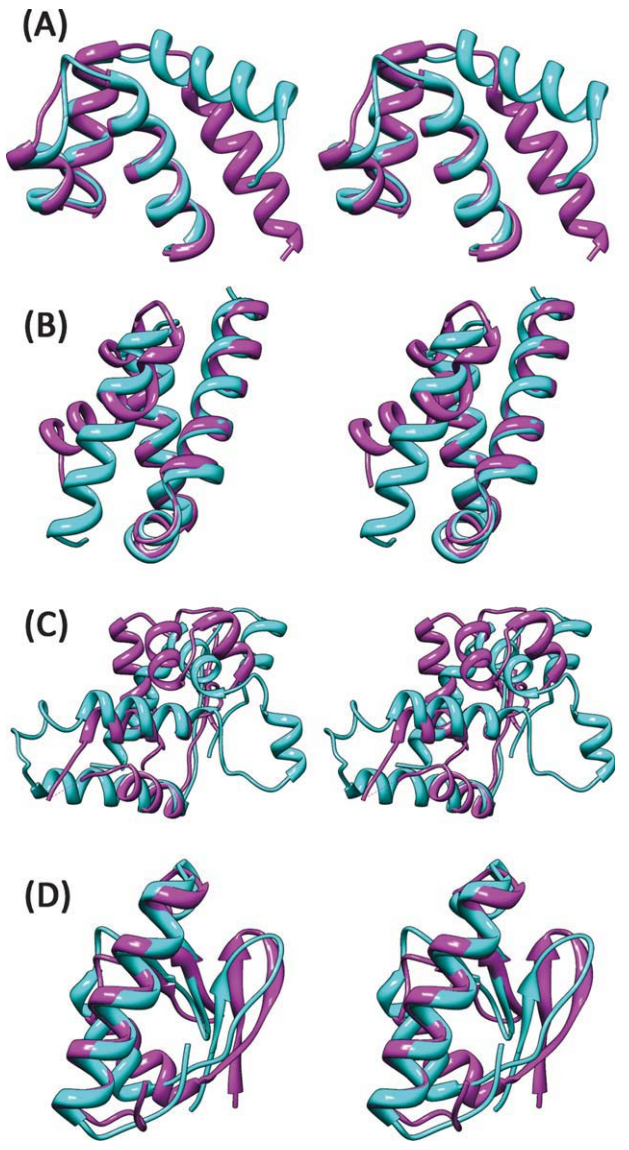

Figure 3

Superposition of the native (purple) and the best model out of 5 CHARMM-DFA/CSA models with the optimized parameters (cyan) is shown for (**A**) T0416-D2, (**B**) T0443-D1, (**C**) T0465-D1, and (**D**) T0513-D2.

models with RMSD $\leq$ 3.0 Å were obtained by the optimized parameter set. There are three targets, 1itp_A, 1of9_A, and 1vcc_A, whose RMSD values deteriorated over 1.0 Å. However, these targets were incorrectly modeled, i.e., RMSD $>$ 5 Å, with both the initial and the optimized parameter sets. Comparing our results with ROSETTA study,[1] the average of the best Cα-RMSD from five selected models (4.5 Å) is slightly better than that of ROSETTA (4.8 Å). Additionally, when we count the number of good predictions represented by RMSD $<$ 4.0 Å, the result is comparable with that of ROSETTA.

In comparison to I-TASSER study,[2] the average Cα-RMSD is worse by 2.4 Å, which corresponds to 0.12 TMscore difference. The number of good predictions represented by RMSD $<$ 4.0 Å made by I-TASSER and CHARMM-DFA are 18 and 12, respectively. One of the reasons for this difference comes from a number of rather wrong predictions made by CHARMM-DFA, whose RMSD values are over 10 Å. Another reason is that although highly homologous templates are removed from the template library of I-TASSER before *ab initio* modeling simulation, it is likely that structural homologs containing large sizes of fragments are still included in the library. On the other hand, in our method, protein models are all generated strictly based on 9-residue fragments. It should be noted that, in I-TASSER, additional template-based information, such as the long-range contact restraint is utilized.

Test with CASP8 template free modeling targets

To test the transferability of the optimized parameter set, we have performed CHARMM-DFA/CSA runs with the initial and the optimized parameter sets against the template-FM targets of CASP8.[50] The prediction results of the current method as well as those of the top three

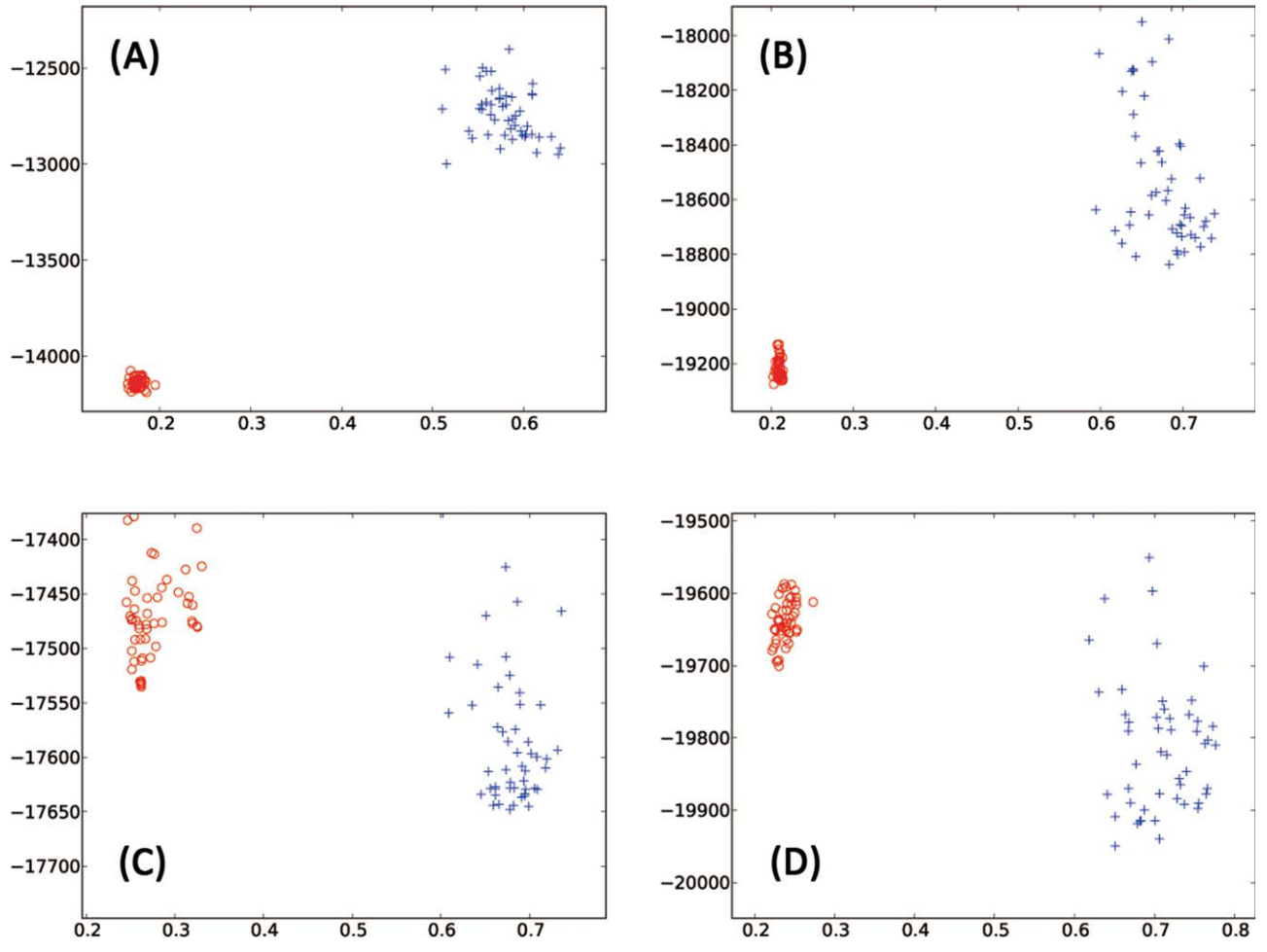

Figure 4

Energy landscapes of four incorrectly predicted proteins from CASP8 new fold targets are shown. The conformations obtained by the global CSA and the local CSA are shown as blue crosses and red circles, respectively. Near-native conformations are defined by the RMSD cutoff value of 3.5 Å during the local CSA. The *x*-axis represents the TMscore of models.

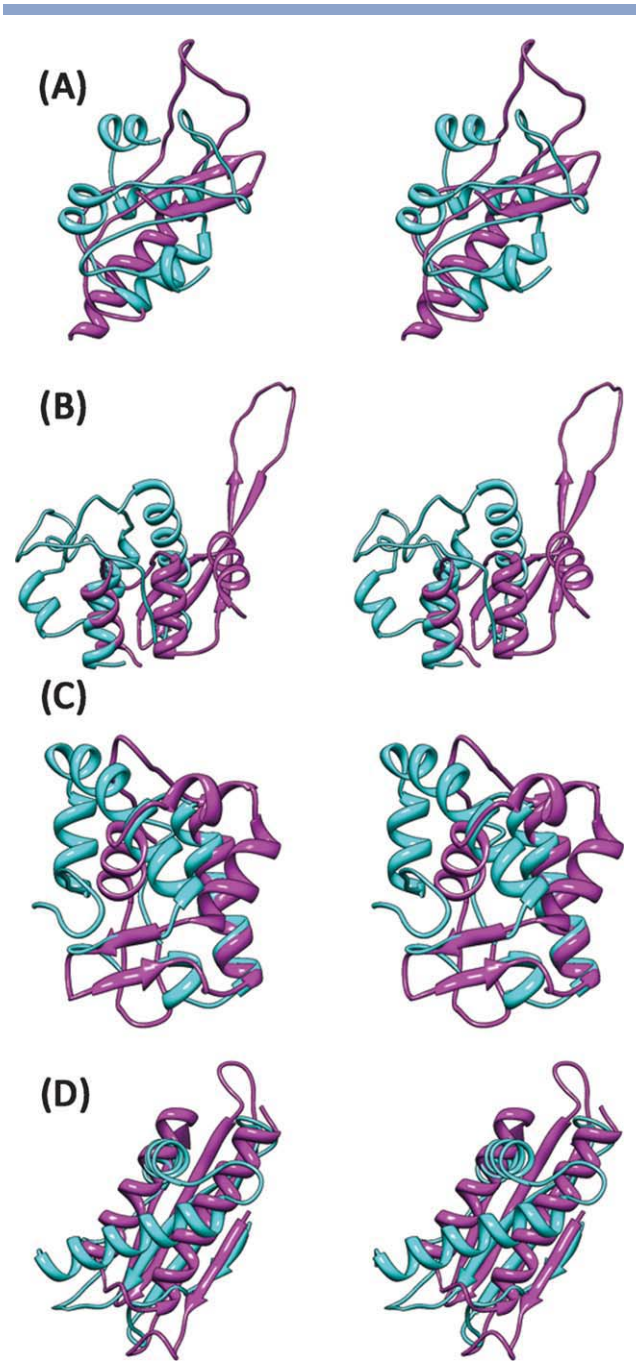

Figure 5

Superposition of the lowest energy structures from global-CSA (cyan) and local-CSA (purple) is shown for (**A**) T0443-D2, (**B**) T0460-D1, (**C**) T0476-D1, and (**D**) T0482-D1.

servers in the FM category of CASP8[34] are shown in Table IV. The servers are MUFOLD-MD, BAKER-ROBETTA, and Zhang-server. It is remarkable that predictions made by CHARMM-DFA/CSA with the optimized parameter set can provide the most accurate models for four targets out of 13 FM targets in CASP8. Comparisons of the models and native structures are shown in Figure 3. Apparently, the optimized parameters are not over-trained for the 30 benchmark proteins. Although the average TMscore from the optimized parameters is slightly worse than that from the initial parameters, the number of correct predictions (TMscore $>$ 0.4) increases from 3 to 4. Similarly, the number of best predictions (CHARMM-DFA/CSA $>$ BEST) increases from 2 to 4. Excluding targets with TMscore $<$ 0.4, two targets, T0513-D2 and T0465-D1, show significant improvements in TMscores by parameter optimization. For T0416-D2, TMscore of the best model decreased from 0.719 to 0.648, and the major discrepancy of the prediction comes from a small deviation of the N-terminal region. However, T0416-D2 corresponds to a large insertion in the middle of a template-based modeling target and the terminal structure could be improved if the distance information between two terminal ends were used, which was not incorporated here. Seven targets with a meaningful prediction (TMscore $>$ 0.4) out of five methods are marked by underlines in Table IV. The average TMscores of this study are comparable with the top 3 servers of CASP8, and all α- and mostly α-proteins are particularly well modeled by the current method.

To find out the direction for further improvement, we tried to identify whether the reason for a bad prediction is the potential function or insufficient conformational sampling. This question can be answered by comparing the potential energy values of near-native structures with the wrong predictions presented here. To obtain the near-native and low-energy conformations, local CSA simulations of four relatively worse cases compared with the other servers, T0443-D2, T0460-D1, T0476-D1, and T0482-D1, were carried out. It should be emphasized that using sophisticated method to sample near-native region, the local CSA procedure, is essential because simple minimization of a crystal native structure may not result in a sufficiently low-energy conformation with a reasonably small RMSD value. The energy landscapes generated with the conformations from the global and local CSA are shown in Figure 4. Structural comparisons between the lowest energy structures from the global and local CSA are displayed in Figure 5.

For T0443-D2 and T0460-D1, energies for near-native structures are higher than those from the global-CSA, which shows that there exists an energy issue. Bad prediction of T0443-D2 can be explained by domain selection problem. For simplicity, we used the official domain definition of CASP8 and treated it as a single domain. However, T0443-D2 consists of three domains and the others are template-based modeling targets. The long exposed loop region of T0443-D2, Figure 5(A), is actually not exposed considering the entire structure of T0443, and this region is folded and packed in the global-CSA result. This problem can be fixed if existence of the other domains is taken into account during the modeling. For T0460-D1, physics-based and DFA-beta

Table V
Energy Components of Two Badly Predicted Examples, T0476-D1 and T0482-D1 are Shown

Target		Total	Physics	DFIRE	DFA_dist	DFA_angle	DFA_nn	DFA_beta
T0476	Local[a]	−17647.72	−856.11	−1689.87	−7982.36	−1418.80	−5700.58	−7.93
	Global[a]	−17535.08	−860.14	−1616.43	−8023.80	−1394.44	−5640.27	−1.01
	$\Delta e_{Local\text{-}Global}$	112.65	−4.03	73.44	−41.44	24.36	60.31	6.92
T0482	Local[a]	−19949.46	−1236.85	−1644.03	−9092.45	−1494.58	−6481.55	−216.65
	Global[a]	−19700.70	−1150.60	−1548.45	−9129.31	−1503.62	−6368.72	−133.62
	$\Delta e_{Local\text{-}Global}$	248.76	86.25	95.58	−36.86	−9.04	112.84	83.03

[a]Local and Global correspond to the minimum energy conformations obtained by Local and Global CSA.

terms are favorable for near-native structures while DFIRE and the other DFA terms are unfavorable. This suggests that higher weight for DFA-beta term may be necessary for beta-sheet containing proteins.

For T0476-D1 and T0480-D1, the failure is apparently due to insufficient sampling, as shown in Figure 5(C,D). Native-like conformations generated by local-CSA are of lower energy than those of global-CSA. The breakdown of the energy component of T0476-D1 and T0480-D1 shown in Table V indicates that native-like conformations are stabilized by better DFIRE and DFA-neighbor terms while non-native global-CSA-generated conformations are favored by the DFA-distance term. Therefore, the deficiency of the sampling can be attributed to the premature straight-forward minimization of the DFA-distance term over more sophisticated combinatorial terms such as DFIRE and DFA-neighbor terms.

CONCLUSION

In this study, we have presented CHARMM-DFA/CSA approach for *ab initio* protein structure prediction which is an extension from the DFA coarse-grained *ab initio* modeling algorithm by Sasaki and Sasai.[9,10] First, we have combined the full-atom representation of CHARMM with DFA, so the basic elements of physics-based energy terms as well as the statistical potentials of DFIRE can be readily included in the energy function.

One of the motivations for this study is to establish a common ground to develop a force field suitable for *ab initio* protein structure prediction. For this purpose, the force field should be readily accessible to the community and provide reasonable prediction power for hard *ab initio* target proteins. In this article, we propose that this goal can be achieved to some extent by (1) converting the popular fragment assembly approach into differentiable residue-position-specific energy terms containing both local as well as global features of a fragment library, and (2) combining the DFA energy with the CHARMM force field. Inclusion of additional energy terms, such as DFIRE is rather straightforward. The next step to improve the energy function is to optimize many parameters included in each energy term. As the initial step toward this ambitious process, we have performed weight optimization of several independently introduced energy terms, namely four DFA terms and DFIRE term, while the CHARMM weights are fixed to unity. The parameter optimization executed in this work is neither complete nor rigorous, but it sheds light into the future direction for successful *ab initio* protein structure prediction.

The direct application of the proposed method shows comparable prediction accuracy with existing top performing *ab initio* prediction programs, ROSETTA, I-TASSER, and MUFOLD. The optimized parameter set was not over-trained, and this was validated with tests on independent CASP8 template FM targets. The current approach is shown to provide complimentary results by generating best models for four out of 13 targets.

In summary, the proposed CHARMM-DFA/CSA method is showing comparable modeling accuracy with other existing methods for new-fold targets. Especially, all α- and mostly α-proteins are accurately modeled by the current method. However, there remains much room for improvement in both the accuracy of energy functions and the efficiency of the sampling methods.

By introducing an energy function for *ab initio* modeling with widely used CHARMM program, we hope that a broad range of researchers would be able to utilize DFA together with other facilities already implemented in the CHARMM. The current implementation can constitute an alternative way to study various biomolecular problems including the protein folding problem.

ACKNOWLEDGMENTS

We thank Korea Institute for Advanced Study for providing computing resources (KIAS Center for Advanced Computation) for this work.

REFERENCES

1. Bradley P, Misura K, Baker D. Toward high-resolution de novo structure prediction for small proteins. Science 2005;309:1868–1871.
2. Wu S, Skolnick J, Zhang Y. Ab initio modeling of small proteins by iterative TASSER simulations. BMC Biol 2007;5:17.
3. Hansmann U, Okamoto Y. New Monte Carlo algorithms for protein folding. Curr Opin Struct Biol 1999;9:177–183.
4. Sugita Y, Okamoto Y. Replica-exchange molecular dynamics method for protein folding. Chem Phys Lett 1999;314:141–151.
5. Chikenji G, Fujitsuka Y, Takada S. A reversible fragment assembly method for de novo protein structure prediction. J Chem Phys 2003;119:6895–6903.

6. Lee J. New Monte Carlo algorithm: entropic sampling. Phys Rev Lett 1993;71:211–214.
7. Zhang Y, Kihara D, Skolnick J. Local energy landscape flattening: parallel hyperbolic Monte Carlo sampling of protein folding. Proteins Struct Funct Bioinform 2002;48:192–201.
8. Keedy D, Williams C, Headd J, Arendall W, III, Chen V, Kapral G, Gillespie R, Block J, Zemla A, Richardson D. The other 90% of the protein: Assessment beyond the Cas for CASP8 template-based and high-accuracy models. Proteins Struct Funct Bioinform 2009;77:29–49.
9. Sasaki T, Cetin H, Sasai M. A coarse-grained Langevin molecular dynamics approach to de novo protein structure prediction. Biochem Biophys Res Commun 2011;369:500–506.
10. Sasaki T, Sasai M. A coarse-grained langevin molecular dynamics approach to protein structure reproduction. Chem Phys Lett 2005;402:102–106.
11. Nocedal J, Wright S. Numerical optimization. Springer Verlag; 1999.
12. Joo K, Lee J, Kim I, Lee S. Multiple sequence alignment by conformational space annealing. Biophys J 2008;95:4813–4819.
13. Kim S, Lee S, Lee J. Structure optimization by conformational space annealing in an off-lattice protein model. Phys Rev E 2005;72:11916.
14. Lee J, Kim S, Joo K, Kim I. Prediction of protein tertiary structure using PROFESY, a novel method based on fragment assembly and conformational space annealing. Protein Struct Funct Bioinform 2004;56:704–714.
15. Lee J, Lee I. Unbiased global optimization of Lennard-Jones clusters for N$\leq$201 using the conformational space annealing method. Phys Rev Lett 2003;91:80201.
16. Lee K, Czaplewski C, Kim S, Lee J. An efficient molecular docking using conformational space annealing. J Comput Chem 2005;26:78–87.
17. Lee K, Sim J, Lee J. Study of protein-protein interaction using conformational space annealing. Protein Struct Funct Bioinform 2005;60:257–262.
18. Joo K, Lee J, Lee S, Seo J, Lee S. High accuracy template based modeling by global optimization. Protein Struct Funct Bioinform 2007;69:83–89.
19. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem 1983;4:187–217.
20. Lazaridis T. Effective energy function for proteins in lipid membranes. Protein Struct Funct Bioinform 2003;52:176–192.
21. 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.
22. Day R, Zydallis J, Lamont G, Pachter R. Solving the protein structure prediction problem through a multiobjective genetic algorithm. Nanotechnology 2002;2:32–35.
23. Cutello V, Narzisi G, Nicosia G. A multi-objective evolutionary approach to the protein structure prediction problem. J R Soc Interface 2006;3:139–151.
24. Handl J, Lovell S, Knowles J. Investigations into the effect of multiobjectivization in protein structure prediction. Parallel Problem Solving from Nature (PPSN X) 2008:702–711.
25. Lee J, Ripoll D, Czaplewski C, Pillardy J, Wedemeyer W, Scheraga H. Optimization of parameters in macromolecular potential energy functions by conformational space annealing. J Phys Chem B 2001;105:7291–7298.
26. Lee J, Park K. Full optimization of linear parameters of a united residue protein potential. J Phys Chem B 2002;106:11647–11657.
27. Ben-David M, Noivirt-Brik O, Paz A, Prilusky J, Sussman J, Levy Y. Assessment of CASP8 structure predictions for template free targets. Protein Struct Funct Bioinform 2009;77:50–65.
28. MacKerell A, Jr, Bashford D, Bellott M, Dunbrack R, Jr, Evanseck J, Field M, Fischer S, Gao J, Guo H, Ha S. All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B 1998;102:3586–3616.
29. Wang G, Dunbrack R, Jr. PISCES: a protein sequence culling server. Bioinformatics 2003;19:1589–1591.
30. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
31. Cheng J, Baldi P. Three-stage prediction of protein {beta}-sheets by neural networks, alignments and graph algorithms. Bioinformatics 2005;21(Suppl 1):i75–i84.
32. Mackerell AD, Feig M, Brooks CL. Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. J Comput Chem 2004;25:1400–1415.
33. Gray J, Moughon S, Wang C, Schueler-Furman O, Kuhlman B, Rohl C, Baker D. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. J Mol Biol 2003;331:281–299.
34. Kuhlman B, Dantas G, Ireton G, Varani G, Stoddard B, Baker D. Design of a novel globular protein fold with atomic-level accuracy. Science 2003;302:1364–1368.
35. Rohl C, Strauss C, Misura K, Baker D. Protein structure prediction using Rosetta. Methods Enzymol 2004;383:66–93.
36. Eramian D, Shen M, Devos D, Melo F, Sali A, Marti-Renom M. A composite score for predicting errors in protein structure models. Protein Sci 2006;15:1653–1666.
37. Fasnacht M, Zhu J, Honig B. Local quality assessment in homology models using statistical potentials and support vector machines. Protein Sci 2007;16:1557–1568.
38. Soto C, Fasnacht M, Zhu J, Forrest L, Honig B. Loop modeling: Sampling, filtering, and scoring. Protein Struct Funct Bioinform 2008;70:834–843.
39. 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.
40. Lee J, Scheraga H, Rackovsky S. New optimization method for conformational energy calculations on polypeptides: conformational space annealing. J Comput Chem 1997;18:1222–1232.
41. Lee J, Scheraga HA, Rackovsky S. New optimization method for conformational energy calculations on polypeptides: conformational space annealing. J Comput Chem 1997;18:1222–1232.
42. Lee J, Liwo A, Scheraga H. 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 to apo calbindin D9K. Proc Natl Acad Sci USA 1999;96:2025–2030.
43. Lee J, Scheraga H. Conformational space annealing by parallel computations: extensive conformational search of Met-enkephalin and of the 20-residue membrane-bound portion of melittin. Int J Quantum Chem 1999;75:255–265.
44. Goldberg DE. Genetic algorithm. Reading: Addison-Wesley; 1989.
45. Read R, Chavali G. Assessment of CASP7 predictions in the high accuracy template-based modeling category. Protein Struct Funct Bioinform 2007;69(S8):27–37.
46. Lee J, Pillardy J, Czaplewski C, Arnautova Y, Ripoll D, Liwo A, Gibson K, Wawak R, Scheraga H. Efficient parallel algorithms in global optimization of potential energy functions for peptides, proteins, and crystals* 1. Comput Phys Commun 2000;128:399–411.
47. Lee J, Scheraga HA. Conformational space annealing by parallel computations: extensive conformational search of met-enkephalin and of the 20-residue membrane-bound portion of melittin. Int J Quantum Chem 1999;75:255–265.

48. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Protein Struct Funct Bioinform 2004;57:702–710.
49. Mitchell S. An introduction to pulp for Python programmers. Python Papers Monograph 2009;1:14.
50. Tress M, Ezkurdia I, Richardson J. Target domain definition and classification in CASP8. Protein Struct Funct Bioinform 2009;77(S9):10–17.
51. Xu J, Zhang Y. How significant is a protein structure similarity with TM-score = 0.5? Bioinformatics 2010;26:889–895.
52. McGuffin L, Bryson K, Jones D. The PSIPRED protein structure prediction server. Bioinformatics 2000;16:404–405.
53. Pollastri G, Baldi P, Fariselli P, Casadio R. Prediction of coordination number and relative solvent accessibility in proteins. Protein Struct Funct Bioinform 2002;47:142–153.
54. Zhang J, Wang Q, Barz B, He Z, Kosztin I, Shang Y, Xu D. MUFOLD: a new solution for protein 3D structure prediction. Protein Struct Funct Bioinform 2010;78:1137–1152.