Journal of the Korean Physical Society, Vol. 52, No. 1, January 2008, pp. 143∼151

Protein Structure Prediction Using the Hybrid Energy Function, Fragment Assembly and Double Optimization

Kwang-Hwi Cho and Julian Lee*

Department of Bioinformatics and Life Science, Soongsil University, Seoul 156-143 and Bioinformatics and Molecular Design Technology Innovation Center, Soongsil University, Seoul 156-743

Taek-Kyun Kim

Department of Bioinformatics and Life Science, Soongsil University, Seoul 156-743

(Received 31 October 2007)

We perform protein structure prediction by combining a hybrid energy function, fragment assembly, and double optimization. In the hybrid energy function, all the backbone atoms are described explicitly, but the side-chain is modeled as a few interaction centers in order to reduce computational costs. We reduce the search space by using a fragment assembly method, where the local structure of the backbone is obtained from a structural database using similarity of sequence features, and only the global tertiary packing of fragments is determined by minimizing the energy. The structure with the minimum energy is obtained using double optimization, where a combination of backbone fragments with minimum energy is obtained using the conformational space annealing (CSA) method, and the optimal side-chains for a given backbone structure are obtained using simulated annealing. We show the feasibility of our method by performing test predictions on two proteins, 1bdd and 1e0l, that belong to distinct structural classes.

PACS numbers: 87.14.Ee, 87.15.Aa, 87.15.Cc
Keywords: Protein folding, Protein structure prediction, Fragment assembly method

I. INTRODUCTION

Understanding folding of a protein into its three-dimensional structure from its amino-acid sequence, especially the prediction of the native structure, is a long-standing challenge in theoretical biophysics. The information on the native structure of a protein is quite crucial in understanding its biological function [1]. The most popular methods for protein structure prediction are knowledge-based methods such as comparative modeling and fold recognition [2–4]. In knowledge-based methods, there should exist a sequence with known structure that is related to the query sequence. When homologous or weakly homologous sequences with known structures are not available, we turn to physics-based methods [3, 5–17]. The physics-based structure prediction is based on the thermodynamic hypothesis [18] which states that the native tertiary structure of a protein corresponds to the global minimum of its free energy for its physiological environment. Therefore, in the physics-based prediction method, also called the energy-based method, the native structure of a protein is predicted by obtaining the conformation that minimizes the free energy. Since the physics-based method is based on fundamental principles of physics, the study of protein folding using this method provides us with valuable insight into not only the native structure but also the folding mechanism [19–21].

There are two main challenges to successful prediction of protein structures, the design of an accurate energy function with reasonable computational cost and the development of a powerful global optimization method. It is obvious that the calculation of the protein structure with an energy function that takes into account all the atomic degrees of freedom, although more accurate than that using a coarse-grained model, will take too much computational resource in order to produce meaningful results. Also, even if we are provided with an accurate free energy function, it is a nontrivial task to find the global minimum of such an energy function, because there is usually a huge number of local minima. Various global optimization methods have been developed to overcome this problem.

The fragment assembly method, which has been a popular trend in protein structure prediction [3, 9–17], addresses both of these issues to some extent. In this method, the local structures are collected from experimentally determined structures deposited in a structural database, such as the protein data bank (PDB), by using the similarity of sequence features. A set of candidates

*Corresponding Author: jul@ssu.ac.kr; Fax: +82-2-824-4383

for the local structures of individual parts of a protein is first constructed. Then, a combination of these fragments that minimizes the free energy is searched for.

Since the effects of local interactions are incorporated into the fragments, one needs to include only non-local interactions in the energy function during fragment assembly. (Local and non-local interactions in this work mean interactions between residues near and far in sequence, respectively). By eliminating the burden of accurate modelling of local interactions, computational costs are significantly reduced. Also, since the global minimum energy conformation is searched for among a finite (albeit large) number of conformations, the search space is drastically reduced, making it much easier to develop a suitable global optimization method.

In this work, we develop a protein structure prediction method based on fragment assembly combined with a hybrid energy function and double optimization. In our energy function, all the backbone atoms are described explicitly, but the side-chain atoms are approximated as a few interaction centers. The optimal side-chain conformation for a given backbone structure is obtained using simulated annealing (SA) [22], and the combination of fragments with minimal energy is obtained using a conformational space annealing (CSA) method [23–29]. We show the feasibility of our method by performing test predictions on two proteins, 1bdd and 1e0l, that belong to distinct structural classes.

II. METHOD

1. The Energy Function

The hybrid energy function, called ECEPP/SM, is derived from the all-atom potential energy ECEPP/3 [30] by coarse-graining the side-chain degrees of freedom whereas the backbone degrees of freedom are kept to atomic details. To elaborate, an all atom representation is used for the backbone atoms, except for C_α which is turned into a pseudo-atom that includes the effect of α hydrogen. C_β is also a pseudo-atom that includes the effect of β hydrogen. The pseudo-atom at the C_α position is classified as G and A, the pseudo-atoms for Glycine and other amino acids, respectively. C_β and β hydrogens are classified as B1 $\sim$ B9, depending on the amino acid type.

Atoms beyond the C_β position in the side chain are reduced to a single pseudo-atom for an amino acid containing the C_γ atom and to two pseudo-atoms for an amino acid with branched side-chains (Ile, Thr, Val), or with long side-chains (Tyr). These side-chain pseudo-atoms are also classified into various types. The simplified models of the side-chains and the classification of the pseudo-atoms are shown in Figure 1 for a selected set of amino acids.

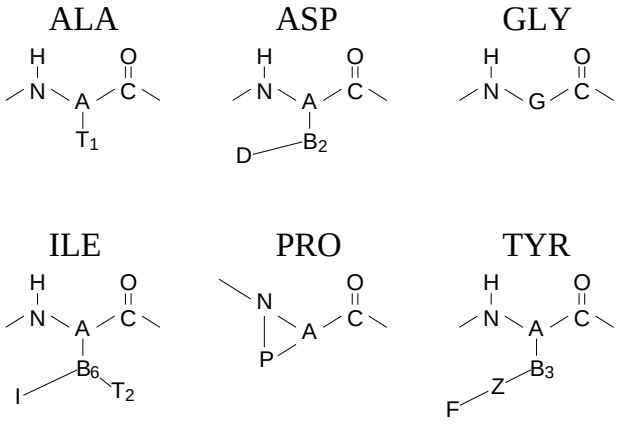

Fig. 1. Classification of the atom types in the simplified side-chain representation used in the hybrid energy function for a selected set of amino acids.

In terms of the pseudo-atom coordinates, the hybrid energy function (ECEPP/SM) takes the functional form

$$E = \sum_{i<j} \left(\frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right) + \sum_{i<j} 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \sum_{i<j} 4\eta_{ij} \left[\left(\frac{\rho_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\rho_{ij}}{r_{ij}} \right)^{10} \right] + w \cdot SME(\{\mathbf{S_a}\}, \{\mathbf{r_i}\}), \quad (1)$$

where $\mathbf{r_i}$ and $\mathbf{S_a}$ represent the coordinates of the backbone and the side-chain atoms respectively, with $r_{ij} \equiv |\mathbf{r_i} - \mathbf{r_j}|$. The first term represents the Coulomb electrostatic interaction, with q and ϵ_0, respectively, being the atomic charge and the dielectric constant. The second term and the third term represent van der Waals and hydrogen bonding interactions, respectively, with non-bonding parameters ε_{ij}, η_{ij}, σ_{ij} and ρ_{ij}. The first, the second, and the third terms are intended for the interactions between all-atom degrees of freedom, such as H, N, C and O in the backbone. The parameters of these terms were taken from ECEPP/3 without any modification. The last term

$$SME(\{\mathbf{S_a}, \mathbf{r_j}\}) = \sum_b \left[\sum_j E_{bs}(|\mathbf{S_b} - \mathbf{r_j}|) + \sum_c E_{ss}(|\mathbf{S_b} - \mathbf{S_c}|) \right] \quad (2)$$

accounts for the interaction involving side-chain pseudo atoms, where $\sum_j E_{bs}(|\mathbf{S_b} - \mathbf{r_j}|)$ and $\sum_c E_{ss}(|\mathbf{S_b} - \mathbf{S_c}|)$ represent the interactions of the side-chain at $\mathbf{S_b}$ with the backbone and with the other side-chain atoms, respectively. These terms do not have analytic forms, but their values are stored as tables for each 0.01 Å bin of $|\mathbf{S_b} - \mathbf{r_j}|$ and $|\mathbf{S_b} - \mathbf{S_c}|$. The energy values are derived by examining the dependence of the average side-chain energy on

these distances and depend on the types of pseudo-atoms involved (Figure 1). There is an ambiguity in the relative weight w of the last term compared to the rest, and it should be optimized using training sets of proteins with known structures so that native-like conformations are produced as lowest energy conformations. In this work, no attempt was made to optimize the weight parameter w; rather, it was set to an arbitrary value of 1.0 for most of the simulations. The details of the atom-type classification and the parameter derivation will be presented elsewhere [31].

Although it is impractical to consider all the orientations of a side-chain due to the huge size of the conformational space in the case of the all-atom model, where several dihedral angles are present as the degrees of freedom, one can drastically reduce the sampling space by using the simplified model for the side-chain because there is only one side-chain dihedral angle, χ, for each residue.

2. Fragment Library

In addition to approximating side-chains by using pseudo-atoms, the sampling space for the backbone degrees of freedom can also be drastically reduced by using a fragment library, which is a set of the ten most probable conformations of the local neighborhood for each residue. The fragments are selected from a reference database of non-redundant proteins, constructed by clustering ASTRAL SCOP (version 1.63) set [32] so that no two proteins in the database have more than 25 % sequence identity with each other. The resulting database consists of 4362 protein chains.

For fair benchmark tests, proteins in the reference database that are homologous to the query protein are removed. A BLAST search of the query sequence against the reference database is performed, and any protein chain whose local alignments have sequence identity of 70 % or more to the query sequence length is removed.

The fragment selection is performed using the similarity between sequence features. Instead of comparing raw sequences directly, PSI-BLAST [33] profiles that contains evolutionary information are generated. For a given segment of the query sequence, ten fragments with similar sequence profiles are selected from the structural database, by using k-nearest neighbor method [34–37], to obtain the fragment library. The fragment set is the same as the one used in Ref. 17.

3. Conformational Sampling

The goal of the current study is to find the combination of fragments, as well as the side-chain orientation, that minimizes the energy function. It should be noted that the sampling space of the backbone is a finite discrete space of fragment combinations whereas that of the side-chains is a continuous space of dihedral angles χ. In this work, we perform a double optimization, where the conformational space annealing (CSA) method [23–29] is used for generating low-energy back-bone conformations, and the simulated annealing (SA) method [22] is used for obtaining the optimal side-chain orientation for a given backbone conformation.

The CSA method is based on the genetic algorithm in which a population of conformations, called a bank, is considered at a time. The initial bank is constructed by random generation of conformations and subsequent local minimizations. The local minimization in the context of fragment assembly means the Monte Carlo simulation at T = 0. That is, one of the fragments forming the structure is randomly selected and then replaced by another fragment in the library, which is also randomly selected. Then, the new conformation is accepted only when its energy is lower than the previous one.

The bank is updated by generating trial conformations. First, seeds are selected from the bank, and trial conformations are generated from the seeds by replacing parts of the seed conformations with those of other randomly selected bank conformations and by performing local minimizations. The CSA method has a parameter $D_{\rm cut}$ that controls the diversity of the bank, which decreases as the algorithm proceeds. To elaborate, a trial conformation a is compared with bank conformations, and the conformation A is selected from the bank, which is the closest to the conformation a with respect to a suitable distance measure $D(a, A)$. If $D(a, A) < D_{\rm cut}$, the conformation a is considered as being more or less similar to the conformation A. The conformation with the lower energy is kept in the bank, and the other one is discarded. However, if $D(a, A) > D_{\rm cut}$, the conformation a is regarded as being distinct from any other conformation in the bank. Therefore, a is compared with the bank conformation with the highest energy, and again, the conformation with the lower energy is kept in the bank, and the other one is discarded. The diversity of the bank is maintained for a large value of $D_{\rm cut}$, and the low-energy properties of the conformations are emphasized for a small value of $D_{\rm cut}$. The algorithm stops when all the conformation left in the bank are used as seeds.

The energy calculated for a given conformation in the CSA method always includes the optimal side-chain energy. That is, whenever a backbone conformation is newly constructed, a simulated annealing (SA) is performed to obtain the optimal side-chain orientation. In the SA method, several side-chain angles are randomly selected and randomly perturbed within given bounds, and the newly generated conformation is accepted or rejected according to the Metropolis criterion at a given temperature. The temperature is slowly reduced at each Monte Carlo step in order to obtain the side-chain conformation with minimum energy.

Table 1. The Energy and RMSD values of the lowest energy bank conformation (LEBC) of 1e0l and the smallest RMSD value found among the bank for various values of the SA parameters. The result of the simulation without side-chain ($w = 0.0$) is also shown. The unit for both the temperature and the energy is kcal/mol.

PDB ID (length)	$T_{\rm factor}$	$MC_{\rm step}$	$T_{\rm init}$	Energy (LEBC)	RMSD (LEBC)	RMSD (smallest)
	0.9	10	0.5	–166.17	4.19	1.48
	0.9	10	1.0	–167.25	3.96	1.42
	0.9	10	1.5	–167.21	4.21	1.58
	0.9	20	0.5	–170.11	4.08	1.41
	0.9	20	$1.0 \sim 1.5$	–163.11	4.11	1.23
	0.9	30	0.5	–163.98	4.32	1.17
	0.9	30	$1.0 \sim 1.5$	–170.24	3.80	1.43
1E0L (25) ($w = 1.0$)	0.5	10	$0.5 \sim 1.0$	–181.32	3.77	1.38
	0.5	10	1.5	–166.95	2.35	1.48
	0.5	20	$0.5 \sim 1.0$	–170.91	3.77	1.16
	0.5	20	1.5	–165.52	4.19	1.41
	0.5	30	$0.5 \sim 1.0$	–181.31	3.83	1.11
	0.5	30	1.5	–165.76	2.37	1.16
	0.3	10	$0.5 \sim 1.5$	–181.32	3.77	1.38
	0.3	20	$0.5 \sim 1.5$	–168.78	3.96	1.79
	0.3	30	$0.5 \sim 1.5$	–156.52	3.89	1.39
1E0L (25) ($w = 0.0$)	–	0	–	–5.94	4.43	1.22

4. Clustering and the Final Models

If the energy function in Eq. (1) is the accurate free-energy function describing the protein, we can simply consider the bank conformation with the lowest energy as the predicted model for the protein structure. However, there can be inaccuracies in various parameters describing the energy function, and since the structure of the lowest energy bank conformation (LEBC) may depend sensitively on these parameters, suboptimal conformations are also important for protein structure prediction.

Therefore, usually a clustering of the final bank conformations is performed, and the representative conformation for each cluster is selected to obtain multiple candidates for the native structure. By clustering the final bank conformations, one can include the effect of conformational entropy for the global structure [38], which may not be fully incorporated in fragment assembly combined with the energy function in Eq. (1).

In this work, we grouped the final conformations into $k = 5$ clusters using the k-means clustering algorithm [39]. The choice for the number of clusters, k, is rather arbitrary. We chose it to be five only because in CASP(http://predictioncenter.gc.ucdavis.edu/), the competition for computational prediction of protein structures, allows up to five models to be submitted as multiple candidates for a protein structure, so $k = 5$ should be used in such a case. The center of a cluster was considered as the representative conformation of that cluster.

III. RESULTS

In order to test the feasibility of our method, we performed test predictions on two proteins, FBP28WW domain from mus musculus (PDB ID:1e0l) and staphylococcus aureus protein A, immunoglobulin-binding B domain (PDB ID:1bdd). They are of length 37 and 60, but after unstructured tail regions are removed, regions of length 25 (residue 6-30) and 46 (residue 10-55) remain to be modeled. 1e0l and 1bdd belong to distinct structural classes, all-α and all-β proteins, respectively. The parameters for the CSA runs were set to the default values used in previous works [14–17]: 50 bank conformations, 10 seeds to be selected, and 30 trial conformations to be generated for each seed. The relative weight w in (1) was set to 1.0 unless stated otherwise.

The values of the energy and the backbone root-mean-square deviation (RMSD) of the LEBC, along with the smallest value of RMSD found among the bank conformations, are displayed in Tables 1 and 2 for various values of the SA parameters. In the tables, $T_{\rm init}$ is the initial temperature, $T_{\rm factor}$ is the factor by which the temperature is multiplied after each Monte Carlo step, and $MC_{\rm step}$ is the number of Monte Carlo steps per side-chain minimization. The Boltzmann constant is absorbed into the temperature, so kcal/mol is used as the unit for both energy and temperature.

As can be seen from the tables, the results do not depend much on the parameters. We see that although the lowest energy bank conformation has a relatively large value of RMSD, there are native-like conformations in

Table 2. The Energy and RMSD values of the lowest energy bank conformation (LEBC) of 1bdd and the smallest RMSD value found among the bank for various values of the SA parameters. The results of the simulations with $w = 0.0$ (no side-chain) and $w = 0.5$ are also shown. The unit for both temperature and energy is kcal/mol.

PDB ID (length)	$T_{\rm factor}$	$MC_{\rm step}$	$T_{\rm init}$	Energy (LEBC)	RMSD (LEBC)	RMSD (smallest)
	0.9	10	$0.5 \sim 1.5$	–178.74	8.05	3.13
	0.9	20	$0.5 \sim 1.5$	–183.29	8.01	3.07
	0.9	30	$0.5 \sim 1.5$	–191.29	7.98	3.08
	0.5	10	$0.5 \sim 1.5$	–183.31	7.42	2.91
	0.5	20	0.5	–182.36	8.14	3.16
1BDD (46) ($w = 1.0$)	0.5	20	$1.0 \sim 1.5$	–183.29	8.01	3.04
	0.5	30	0.5	–192.92	7.94	3.02
	0.5	30	$1.0 \sim 1.5$	–192.92	7.94	3.08
	0.3	10	$0.5 \sim 1.5$	–191.29	7.97	2.91
	0.3	20	$0.5 \sim 1.5$	–182.36	8.14	3.14
	0.3	30	$0.5 \sim 1.5$	–192.92	7.94	3.02
1BDD (46) ($w = 0.5$)	0.5	30	0.5	–70.8	8.10	3.04
1BDD (46) ($w = 0.0$)	–	0	–	17.8	13.01	11.99

Table 3. For the five clusters of each protein, the energy and RMSD values of the center conformations are shown, along with the sizes of the clusters. The results of the simulations without side-chain are also shown.

	1E0L (25)					1BDD (46)				
Size	18	16	9	5	2	17	11	10	10	2
Energy	–166.87	–157.16	–180.08	–147.60	–160.22	–170.02	–129.09	–171.69	–167.69	–143.76
RMSD	4.10	4.04	3.90	4.46	4.10	3.99	6.46	3.97	8.01	7.94
	1E0L (25) (no side-chain)					1BDD (46) (no side-chain)				
Size	21	15	8	3	3	19	11	9	8	3
Energy	0.0086	–5.94	0.45	–2.78	0.34	22.4	20.0	20.0	20.7	19.2
RMSD	5.77	4.43	3.85	5.80	5.53	15.9	13.2	16.6	17.0	14.1

the final bank, which is promising because clustering is performed to produce the final models, instead of using the LEBC.

The RMSD and the energy values of the final bank conformations are plotted for 1e0l and 1bdd for $T_{\rm init} = 0.5$ kcal/mol, $MC_{\rm step} = 30$ and $T_{\rm step} = 0.5$ kcal/mol in Figure 2. Even before performing explicit clustering, we see that in addition to the conformation populated near the LEBC, there is a more native-like group. The lowest energy conformation of this sub-optimal group has energy $= -165.53$ kcal/mol and RMSD $= 2.44$ Å for 1e0l and energy $= -171.69$ kcal/mol and RMSD $= 3.97$ Å for 1bdd.

The k-means clustering of these conformations was performed with $k = 5$. The size of these clusters, as well as the energy and the RMSD values of the representative conformations, are displayed in Table 3. The conformation with the smallest value of RMSD among the five models (the best model), the native structure, and the conformation with the smallest value of RMSD among the final bank conformations, are displayed in Figure 3 and 4 for the proteins 1e0l and 1bdd, respectively. We see that for 1e0l, the secondary structure of the model structure is incomplete, but that the overall global structure is similar to the native one. For 1bdd, the model structure has more tightly packed helices, but again, the overall structure is remarkably native-like.

To assess the importance of the side-chain interaction, we also performed simulations with $w = 0.0$, *i.e.*, without side-chains. Since the side-chain distinguishes distinct types of amino acids, for $w = 0.0$, the sequence plays only the role of selecting the fragments, and the non-local interaction is completely independent of the amino-acid sequence. Since there is no side-chain, SA is not necessary, and only the CSA search for the backbone conformation is performed.

The results of the simulations without side-chain are also displayed in Tables 1, 2 and 3. The RMSD and the energy values of the final bank conformations are plotted for 1e0l and 1bdd in Figure 5.

We note that the results for 1e0l are similar to those for $w = 1.0$ whereas for 1bdd there is no native-like conformation in the final bank. The results suggest that for an all-β protein such as 1e0l, the amino-acid sequence only plays a major role in determining the local structure, and the global tertiary structure is determined mainly

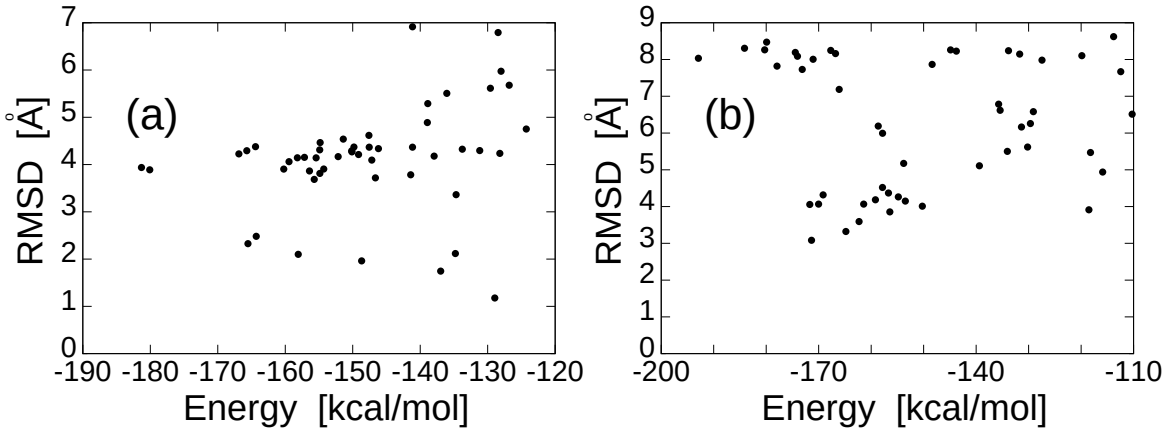

Fig. 2. RMSD and energy values of the final bank conformations for (a) 1e0l and (b) 1bdd.

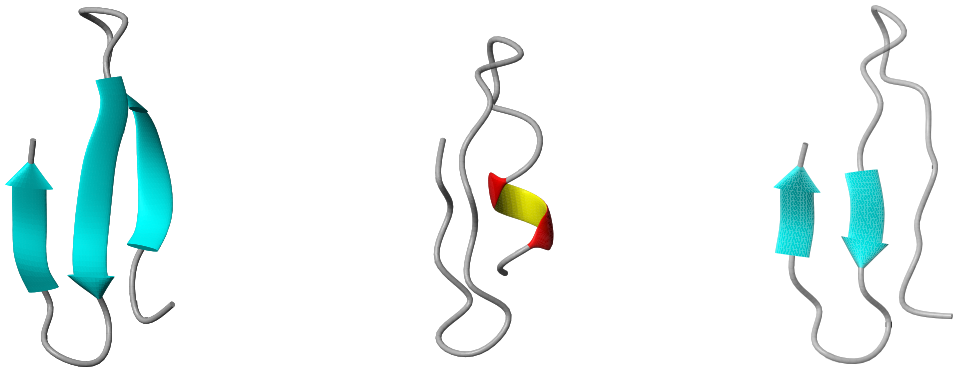

Fig. 3. (a) Native structure, (b) best model and (c) most native-like bank conformation for 1e0l. The figures are prepared with the program MOLMOL [40].

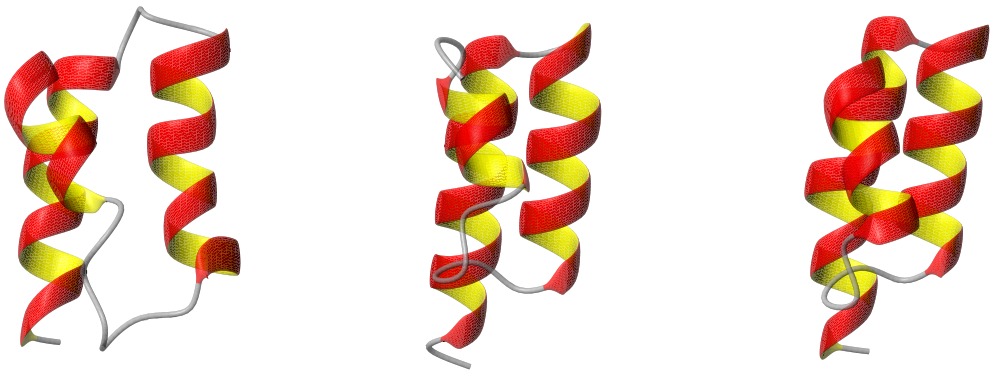

Fig. 4. (a) Native structure, (b) best model and (c) most native-like bank conformation for 1bdd.

by sequence-independent interactions, such as hydrogen bonding between extended fragments that results in a β-sheet structure. On the other hand, for an all α proteins such as 1bdd, non-local hydrophobic side-chain interactions are also important for native-like packing of helices. This can be confirmed by examining the structure with the smallest RMSD among the final bank (Figure 6), where the helices are formed, but not correctly packed into the three-helix bundle. These results are consistent with the conclusion of an earlier work [17], where CSA searches were performed for ten proteins from various structural classes, with the CHARMM energy function used for the back-bone interaction, and a simple contact energy between C_β positions used to mimic side-chain interactions.

One additional simulation was performed with $w = 0.5$

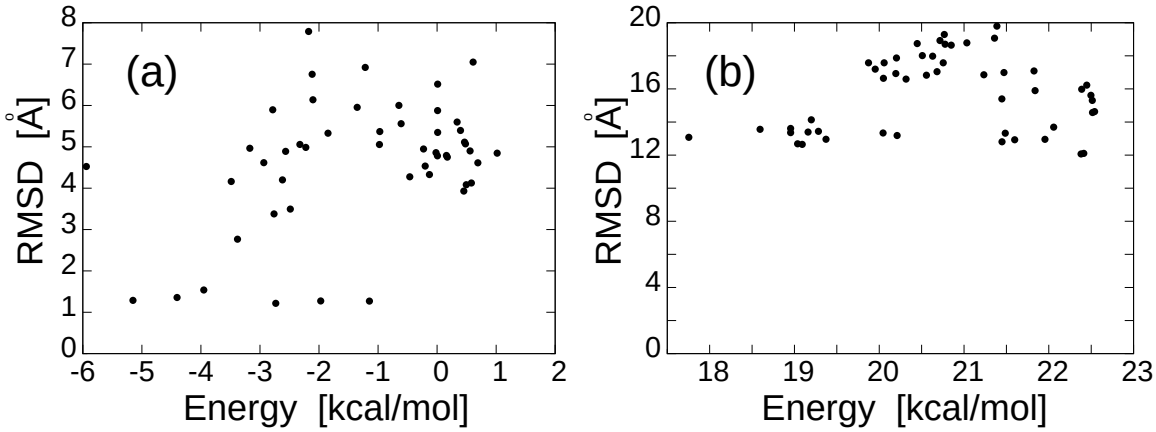

Fig. 5. RMSD and energy values of the final bank conformations for the simulations without side-chains for (a) 1e0l and (b) 1bdd.

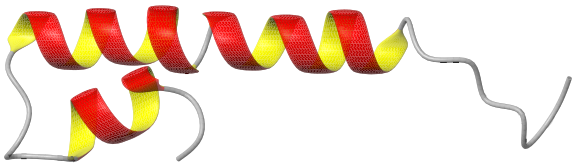

Fig. 6. Bank conformation of 1bdd with the smallest RMSD (11.99 Å) obtained from the simulation without side-chains.

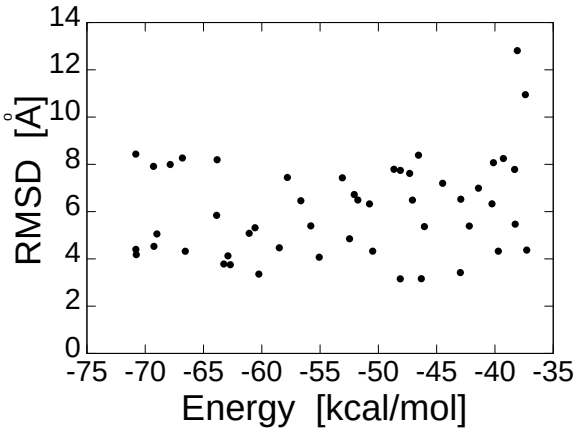

Fig. 7. RMSD and energy values of the final bank conformations of 1bdd for the simulation with $w = 0.5$.

for 1bdd, with $T_{\text{init}} = 0.5$ kcal/mol, $MC_{\text{step}} = 30$ and $T_{\text{step}} = 0.5$ kcal/mol, and the results are summarized in Table 2 and Figure 7, which are more or less similar to those with $w = 1.0$. If an optimal value of w is to be found, simulations for a training set consisting of a larger number of proteins will be necessary.

Using a single Intel Xeon CPU (2.4 GHz), the wall clock times for the simulations with side-chain were 36 $\sim$ 135 minutes for 1e0l, and 148 minutes $\sim$ 8 hours 21 minutes for 1bdd. Those for the simulations without

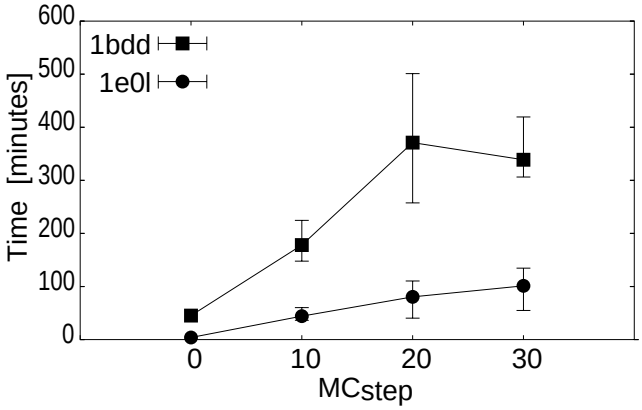

Fig. 8. Average wall clock times for simulations as functions of MC_{step} for 1e0l (filled circles) and 1bdd (filled boxes). The error bars indicate the ranges of the values. The simulations for $MC_{\text{step}} = 0$ were performed without side-chains.

side-chains were 4 minutes for 1e0l and 45 minutes for 1bdd. The average values of these simulation times for each value of MC_{step} are plotted in Figure 8, along with the ranges. It should be remembered that the simulation time not only depends on the SA parameter MC_{step} but also on the number of local minimizations performed in the CSA search for the backbone conformation. The latter quantity depends on the random selections of seed conformations, and the average wall clock time for 1bdd with $MC_{\text{step}} = 20$ is larger than that with $MC_{\text{step}} = 30$ only because the total number of trial conformations generated in the CSA searches happened to be larger, on average, for $MC_{\text{step}} = 20$.

The CSA algorithm can be easily adapted for parallel computation by dividing the workload of local minimizations among slave nodes [25]. We did not implement the parallel code in this work because we had to perform runs with various SA parameters, and that already required many CPUs without parallel computations. However, if one were to fix the SA parameters to a particular set

of values and run the algorithm for a few proteins, then parallel computation would certainly speed up the computation.

IV. DISCUSSIONS

In this work, we developed a protein structure prediction algorithm based on a hybrid energy function combined with fragment assembly and double optimization. We performed test predictions on two proteins, 1bdd and 1e0l, and obtained promising results. We produced five models for each protein by clustering the final bank of low-energy conformations and showed that at least one of the models had a native-like global structure.

By performing simulations without side-chains, we could see that the all-β protein 1e0l folds into a native structure only with sequence-independent backbone nonlocal interaction whereas for the all-α protein 1bdd, the sequence-dependent side-chain interaction is essential for the native-like packing of helices, in accordance with an earlier work using a simpler side-chain model [17].

The method presented in this work is by no means optimal, and there is room for improvement. As already mentioned, an optimal value for the weight parameter w in Eq. (3) should be determined. As for the double optimization, there is some arbitrariness in the manner the CSA and the SA methods are combined, and it is possible for the SA method to be used for the back-bone sampling and for the CSA method to be applied for the side-chain sampling, for example. Also, a better method for selecting the model structures from the final bank should be developed. Further refinements to the energy function and the optimization algorithm will be subjects of a future study.

ACKNOWLEDGMENTS

This work was supported by the Soongsil University Research Fund.

REFERENCES

[1] X. Shi, H. Xie, S. Zhang and B. Hao, J. Korean Phys. Soc. **50**, 118 (2007); H. Y. Kim, H. Y. Kang, J. W. Ryu, C. N. Yoon and S. K. Han, J. Korean Phys. Soc. **50**, 290 (2007); P. Holme, J. Korean Phys. Soc. **50**, 300 (2007); J. W. Ryu, H. Y. Kim, T. H. Kang, J. S. Yoo and J. S. Chung, J. Korean Phys. Soc. **51**, 1805 (2007).
[2] J. Moult, K. Fidelis, B. Rost, T. Hubbard and A. Tramontano, Proteins **61 (S7)**, 3 (2005).
[3] D. Baker and A. Sali, Science **294**, 93 (2001).
[4] J. Jung, H.-T. Moon and J. Lee, J. Korean Phys. Soc. **46**, 625 (2005).
[5] J. Lee, A. Liwo and H. A. Scheraga, Proc. Natl. Acad. Sci. U.S.A. **96**, 2025 (1999).
[6] J. Lee, A. Liwo, D. R. Ripoll, J. Pillardy and H. A. Scheraga, Proteins **37 (S3)**, 204 (1999).
[7] J. Lee, A. Liwo, D. R. Ripoll, J. Pillardy, J. A. Saunders, K. D. Gibson and H. A. Scheraga, Int. J. Quantum Chem. **77**, 90 (2000).
[8] S. Oldziej, C. Czaplewski, A. Liwo, M. Chinchio, M. Nanias, J. A. Vila, M. Khalili, Y. A. Arnautova, A. Jagielska, M. Makowski, H. D. Schafroth, R. Kaźmierkiewicz, D. R. Ripoll, J. Pillardy, J. A. Saunders, Y. K. Kang, K. D. Gibson and H. A. Scheraga, Proc. Natl. Acad. Sci. U.S.A. **102**, 7547 (2005).
[9] A. M. Lest, L. Lo Conte and T. J. P. Hubbard, Proteins **45 (S5)**, 98 (2001); P. Aloy, A. Stark, C. Hadley and R. B. Russell, Proteins **53 (S6)**, 436 (2003); J. J. Vincent, C. H. Tai, B. K. Sathyanarayana and B. Lee, Proteins **61 (S7)**, 67 (2005).
[10] K. T. Simons, C. Kooperberg, E. Huang and D. Baker, J. Mol. Biol. **268**, 209 (1997); C. Rohl, C. Strauss, K. Misura and D. Baker, Methods Enzymol. **383**, 66 (2004).
[11] D. T. Jones, Proteins **45 (S5)**, 127 (2001); D. T. Jones, K. Bryson, A. Coleman, L. J. McGuffin, M. I. Sadowski, J. S. Sodhi and J. J. Ward, Proteins **61 (S7)**, 143, (2005).
[12] G. Chikenji, Y. Fujitsuka and S. Takada, J. Chem. Phys. **119**, 6895 (2003); Y. Fujitsuka, G. Chikenji and S. Takada, Proteins **62**, 381 (2006).
[13] G. Chikenji, Y. Fujitsuka and S. Takada, Proc. Natl. Acad. Sci. U.S.A. **103**, 3141 (2006).
[14] J. Lee, S.-Y. Kim, K. Joo, I. Kim and J. Lee, Proteins **56**, 704 (2004).
[15] J. Lee, S.-Y. Kim and J. Lee, Biophys. Chem. **115**, 209 (2005).
[16] J. Lee, S.-Y. Kim and J. Lee, J. Korean Phys. Soc. **46**, 707 (2005).
[17] S-Y. Kim, W. Lee and J. Lee, J. Chem. Phys. **125**, 194908 (2006).
[18] C. B. Anfinsen, Science **181**, 223 (1973).
[19] S.-Y. Kim, J. Lee and J. Lee, J. Chem. Phys. **120**, 8271 (2004).
[20] S.-Y. Kim, J. Lee and J. Lee, J. Korean Phys. Soc. **44**, 594 (2004).
[21] I.-H. Lee, S.-Y. Kim and J. Lee, J. Korean Phys. Soc. **46**, 601 (2005).
[22] S. Kirkpatrick, C. D. Gelatt, Jr. and M. P. Vecchi, Science **220**, 671 (1983).
[23] J. Lee, H. A. Scheraga and S. Rackovsky, J. Comput. Chem. **18**, 1222 (1997).
[24] J. Lee, H. A. Scheraga and S. Rackovsky, Biopolymers **46**, 103 (1998).
[25] J. Lee and H. A. Scheraga, Int. J. Quantum Chem. **75**, 255 (1999).
[26] J. Lee, I.-H. Lee and J. Lee, Phys. Rev. Lett. **91**, 080201 (2003).
[27] S.-Y. Kim, S. J. Lee and J. Lee, J. Chem. Phys. **119**, 10274 (2003).
[28] Julian Lee, J. Korean Phys. Soc. **45**, 1450 (2004).
[29] S.-Y. Kim, S. B. Lee and J. Lee, Phys. Rev. E **72**, 011916 (2005).
[30] F. A. Momany, R. F. McGuire, A. E. Burgess and H. A. Scheraga, J. Phys. Chem. **79**, 2361 (1975).
[31] K.-H. Cho and B. Lee, *A Simplified Potential Energy*

Function for Ab-initio Protein Folding, in preparation.
[32] S. E. Brenner, P. Koehl and M. Levitt, Nucleic Acids Res. **28**, 254 (2000).
[33] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman, Nucleic Acids Res. **25**, 3389 (1997).
[34] K. Joo, J. Lee, S.-Y. Kim, I. Kim, S. J. Lee and J. Lee, J. Korean Phys. Soc. **44**, 599 (2004).
[35] K. Joo, I. Kim, J. Lee, S.-Y. Kim, S. J. Lee and J. Lee, J. Korean Phys. Soc. **45**, 1441 (2004).
[36] J. Sim, S.-Y. Kim and J. Lee, Bioinformatics **21**, 2844 (2005).
[37] S.-Y. Kim, J. Sim and J. Lee, Lecture Notes in Bioinformatics **4115**, 444 (2006).
[38] Z. Xiang Z, C. S. Soto and B. Honig, Proc. Natl. Acad. Sci. U.S.A. **99**, 7432 (2002); J. Lee and C. Seok, *A Statistical Rescoring Scheme for Protein-Ligand Docking: Consideration of Entropic Effect*, Proteins, in press.
[39] J. B. MacQueen, *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability* (Berkeley, University of California Press, CA, 1967), p. 281.
[40] R. Koradi, M. Billeter and K. Wuthrich, J. Mol. Graph. **14**, 51 (1996).