PROTEINS: Structure, Function, and Bioinformatics 56:704–714 (2004)

# Prediction of Protein Tertiary Structure Using PROFESY, a Novel Method Based on Fragment Assembly and Conformational Space Annealing

**Julian Lee,[1–3] Seung-Yeon Kim,[1] Keehyoung Joo,[1,4] Ilsoo Kim,[1,4] and Jooyoung Lee[1]***

[1]*School of Computational Sciences, Korea Institute for Advanced Study, Seoul, Korea*
[2]*Department of Bioinformatics and Life Science, Soongsil University, Seoul, Korea*
[3]*BMDTIC and CAMDRC, Soongsil University, Seoul, Korea*
[4]*Department of Physics and Institute of Basic Science, Sungkyunkwan University, Suwon, Korea*

***ABSTRACT*** **A novel method for ab initio prediction of protein tertiary structures, PROFESY (PROFile Enumerating SYstem), is proposed. This method utilizes the secondary structure prediction information of a query sequence and the fragment assembly procedure based on global optimization. Fifteen-residue-long fragment libraries are constructed using the secondary structure prediction method PREDICT, and fragments in these libraries are assembled to generate full-length chains of a query protein. Tertiary structures of 50 to 100 conformations are obtained by minimizing an energy function for proteins, using the conformational space annealing method that enables one to sample diverse low-lying local minima of the energy. We apply PROFESY for benchmark tests to proteins with known structures to demonstrate its feasibility. In addition, we participated in CASP5 and applied PROFESY to four new-fold targets for blind prediction. The results are quite promising, despite the fact that PROFESY was in its early stages of development. In particular, PROFESY successfully provided us the best model-one structure for the target T0161. Proteins 2004;56:704–714.** © 2004 Wiley-Liss, Inc.

**Key words: protein folding; tertiary structure prediction; ab initio prediction; fragment assembly; global optimization.**

## INTRODUCTION

Determination of the unique tertiary (three-dimensional) structure of a protein from its amino-acid sequence alone is one of the most important and challenging problems in modern biology. The information on the tertiary structure of a protein is quite crucial in understanding the function and biological role of the protein. Currently, genome-sequencing projects are producing an unprecedented amount of linear amino-acid sequences. An exponential growth of protein sequence database in recent years by far outpaces the experimental determination of protein tertiary structures. Therefore, in the field of protein structure investigation, it becomes increasingly more popular to resort to computational methods as a complementary approach to the experimental structure determination. However, prediction of protein tertiary structures still remains as a long-standing challenge in computational biology.[1–3]

The most successful methods for protein structure prediction are the so-called knowledge-based methods such as comparative (or homology) modeling and fold recognition (or threading).[1–3] These methods make direct use of experimentally determined structures, for example, those in the protein data bank (PDB). When the amino-acid sequence of a target protein with an unknown structure is related to those of one or more proteins with known structures, the structures are similar. To find this relation, the first step in protein structure prediction is to find out if the sequence of the target protein is homologous to other sequences in a sequence database. Next, if homologous sequences are found, then multiple sequence alignment procedure is performed with these homologues plus the target sequence. The multiple sequence alignment defines a position-specific scoring matrix (PSSM), which facilitates the search of sequences that have weak homology with the target protein sequence.[4]

If there is an experimental structure (that is, a template) for a homologue with a relatively high sequence similarity (typically more than 30% of sequence identity), comparative modeling methods[3,5–16] are applied for predicting the tertiary structure of the target protein. In comparative modeling, the target sequence is aligned to template(s), and then all atom structures of the target protein are produced after filling in alignment gaps and properly orienting side chains.

If there exist no obvious homologues, fold recognition methods[17–27] can be used to search for a distant homologue or an analogous fold. In a fold recognition approach, the tertiary structure of the target protein is predicted by threading the target sequence through each of the structures in a database of known folds. Each sequence struc-

Grant sponsor: Basic Research Program of the Korea Science & Engineering Foundation; Grant numbers: R01-2003-000-11595-0 and R01-2003-000-10199-0.

*Correspondence to: Jooyoung Lee, School of Computational Sciences, Korea Institute for Advanced Study, 207-43 Cheongryangridong, Dongdaemun-gu, Seoul 130-722, Korea. E-mail: jlee@kias.re.kr

Received 26 August 2003; Accepted 2 February 2004

Published online 11 June 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20150

ture alignment is assessed by a specially designed sequence structure fitness function (often called a pseudo-energy function). The necessary condition for a reasonable performance of knowledge-based methods is that there should exist a sequence with a known structure that is related to the target sequence.

When homologous or weakly homologous sequences with known structures are not available, we turn to ab initio methods (or new fold methods).[3,28–40] The ab initio protein structure prediction is based on the thermodynamic hypothesis[41] that states that the native structure of a protein corresponds to the global minimum of its free energy for its physiological environment. However, ab initio methods based on the thermodynamic hypothesis can be truly successful only when both an accurate energy function and an efficient global-optimization method for searching the energy landscape are simultaneously available. Although much progress has been made in this field,[29–32] successful ab initio prediction still remains a challenging and unsolved problem.[1–3,28] For this reason, most ab initio methods use information on known structures to some degree. To refer to these methods, Moult et al.[2] have suggested to use the term "new fold methods" rather than using the traditional term "ab initio methods."

One of the popular trends in the new-fold methods is to determine the tertiary structure of a target protein by assembling fragments generated from the protein data bank (PDB). The effect of the short-range interactions are incorporated by using the fragments from the PDB, and only long-range interaction terms are included in the energy function, which are minimized in order to find conformations with optimal tertiary packing.[33–36] In this report, we introduce an approach based on fragment assembly, PROFESY (PROFile Enumerating SYstem). This method utilizes the information obtained from the secondary structure prediction method PREDICT (PRofile Enumeration DICTionary).[42] For a given protein sequence, PREDICT uses a sequence comparison method, PSI-BLAST,[4] to generate its profile, which defines patterns for its amino acid residues. Each pattern spanning fifteen residues is compared with those in the pattern database generated from the PDB, and the patterns close to the query pattern are selected to determine the secondary structure of the query residue. In order to construct the tertiary structure, we collect the backbone dihedral angles of these patterns, which constitute the fragment library of the residue under consideration. Tertiary structures of a given sequence can be generated from these libraries by fragment assembly.

The energy function we use includes the number of long-range hydrogen bonds, the radius of gyration, and the Lennard-Jones interactions for avoiding steric clashes. Replacement of fragments is carried out so that the energy function is locally minimized (see Methods). The global minimization of the energy function is performed by the conformational space annealing (CSA) method[43–45] that has played an integral role in the recent success of the hierarchical energy-based approach to protein structure prediction.[29–32] The CSA method enables one to sample diverse low-lying local minima of a given function. Therefore, PROFESY is in contrast to other fragment-based prediction methods[33–36] where only the simulated annealing (SA) method has been applied instead of more rigorous conformational search methods. Although the SA is easier to be implemented to a particular system, its sampling efficiency is far less powerful compared to CSA.[46,47] The PROFESY is unique in that the procedure of local energy minimization by fragment replacement is defined, so that a powerful global optimization method CSA can be readily utilized.

As benchmark tests, we applied PROFESY to proteins with known structures, and we found that the results are quite promising. In addition, we participated in the CASP5 experiment (http://predictioncenter.llnl.gov/casp5/) and applied PROFESY to four new-fold targets for blind prediction. The results are promising, despite the fact that PROFESY was in its very early stages of development. In particular, PROFESY has provided us the best model-one structure for the target T0161.

## METHODS

### Construction of Fragment Libraries

The fragment libraries used in PROFESY are constructed using the recently proposed secondary structure prediction method PREDICT.[42] For each residue of a query sequence, a window of size fifteen is considered, where the center of the window is located on the residue under consideration. The fragment library of this residue is the collection of 20 backbone structures of the corresponding 20 nearest patterns in the pattern database of PREDICT. After constructing fragment libraries for all residues of a query sequence, full-length chain conformations can be constructed by assembling fragments in these libraries. Since the only selection criterion for fragments in a library is the similarity of their profile patterns to that of the corresponding query residue, the amino acid composition of fragments does not agree with that of the query sequence. Therefore, at this stage of the current method, side-chains are not constructed and we cannot blindly add explicit solvation energy terms at an atomic level.

### Generation of Random Conformations

Random conformations are built from N- to C-terminal. Since the size of each fragment is 15 residues, we first consider fragment library corresponding to the eighth residue, and pick a fragment randomly from it. Next we randomly pick a second fragment from the library corresponding to the ninth residue. The first and second fragments have 14 overlapping residues. Among these residues, we inspect whether there is any residue whose dihedral angles are similar to each other in these two fragments. Two sets of dihedral angles ($\phi_1$, $\psi_1$) and ($\phi_2$, $\psi_2$) are considered to be similar to each other if either

$$|\phi_1 - \phi_2| \leq 30^\circ \text{ and } |\varphi_1 - \varphi_2| \leq 30^\circ \quad (1)$$

or

$$|\phi_1 - \phi_2| + |\varphi_1 - \varphi_2| \leq 45^\circ \quad (2)$$

If we find such a residue, then the second fragment is joined smoothly to the first one starting from this residue. If we cannot find such a residue, then another fragment is picked from the library, and this process is repeated until we find a fragment that can be joined smoothly to the first fragment. The third fragment is picked from the library corresponding to the tenth residue, and the whole process of picking and smoothly joining fragments continues until a full-length chain is constructed up to the C-terminal end. If, at any stage of the chain buildup process, we cannot find a fragment that can be joined smoothly to the previous one, then the previous fragment is replaced by another one in the corresponding library, and the process of fragment assembly is continued.

### Fragment Replacement and the Local Minimization of the Energy

A conformation is locally minimized with respect to the energy (see The Energy Function) by randomly selecting a residue and attempting to replace a part of the 15-residue-long fragment of the chain by another one in the corresponding library. A new fragment can be inserted smoothly to the existing chain if at least two residues *R1*, *R2* in this fragment (one from the N-terminal end and another one from the C-terminal end) have their dihedral angles similar to those of their neighboring fragments, where the criterion of the similarity is again the satisfaction of either Eq. 1 or Eq. 2. In this case, the part of the new fragment from the residue $R1$ to the residue $R2$ replaces the corresponding part of the existing chain to generate a conformation. If this conformation is lower in energy than the existing one, the former replaces the latter. This process is continued either for 10 $N$seq times, where $N$seq is the length of the protein, or until the update attempts fail for $N$seq consecutive times, whichever is encountered first.

### Global Search Using Conformational Space Annealing Method

Low-lying local minimum-energy conformations are obtained by a powerful global optimization algorithm, a conformational space annealing (CSA) method[43–45] that has played an integral role in the recent success of the ab initio energy-based approach for protein structure prediction.[29–32] The uniqueness of the CSA method lies in the way it controls the diversity of the conformations in the bank. In order to efficiently find the global minimum without getting trapped in local minima, it is important to sample wide regions of the conformational space with less emphasis on obtaining low-energy conformations in early stages. We gradually shift the emphasis from maintaining the diversity of the sampling to obtaining low-energy conformations. For this, we introduce an annealing parameter $D$cut (a cutoff distance in the conformational space) that plays the role of temperature in simulated annealing, hence the name "conformational space annealing." The diversity of sampling is directly controlled in CSA by introducing a distance measure $D$(A,B) between two conformations A and B and comparing it with $D$cut. As a CSA run proceeds, the value of $D$cut is slowly reduced just as in the simulated annealing.

**TABLE I. PROFESY Prediction Results for Proteins With Known Structures[†]**

| | | | | | |
|---|---|---|---|---|---|
| betanova | | | | | |
| Protein | | | | | |
| Model | 1 | 2 | 3 | 4 | 5 |
| RMSD (Å) | 3.6 | 3.1 | 3.2 | 5.7 | 6.8 |
| 1 fsd | | | | | |
| Protein | | | | | |
| Model | 1 | 2 | 3 | 4 | 5 |
| RMSD (Å) | 4.2 | 4.7 | 4.6 | 4.0 | 6.6 |
| 1 bdd | | | | | |
| Protein | | | | | |
| Model | 1 | 2 | 3 | 4 | 5 |
| RMSD (Å) | 8.9 | 9.0 | 4.4 | 7.5 | 4.8 |
| 1bk2 | | | | | |
| Protein | | | | | |
| Model | 1 | 2 | 3 | 4 | 5 |
| RMSD (Å) | 2.3 | 2.5 | 5.0 | 3.7 | 3.6 |

[†]The model numbers are the ranks of these conformations in terms of the score function.

Here, we briefly mention how a CSA run proceeds. We first randomly generate a certain number of initial conformations (for example, 100) whose energies are subsequently minimized by the fragment replacement procedure described earlier. We call the set of these conformations the *first bank*. We make a copy of the first bank and call it the *bank*. The conformations in the bank are updated in later stages, whereas those in the first bank are kept unchanged. Also, the number of conformations in the bank is kept unchanged when the bank is updated. We then choose a certain number of conformations (seeds) from the bank and perturb them by replacing parts by the corresponding parts of conformations randomly chosen from the first bank or the bank. The energies of these conformations are subsequently minimized in order to obtain new trial conformations that can be used to update the bank. A newly obtained local minimum-energy conformation $\alpha$ is compared with those in the bank to decide how the bank should be updated. One first finds the conformation A in the bank that is the closest to the conformation $\alpha$ with the distance $D(\alpha,A)$. If $D(\alpha,A) < D$cut, the conformation $\alpha$ is considered as being more or less similar to the conformation $A$. In this case, the conformation with the lower energy among $A$ and $\alpha$ is kept in the bank and the other one is discarded. However, if $D(\alpha,A) > D$cut, the conformation $\alpha$ is regarded as being distinct from all the other conformations in the bank. Therefore, the conformation with the highest energy among the bank conformations plus the conformation $\alpha$ is discarded and the rest are kept in the bank.

The $D$cut is reduced, and seeds are selected from the bank conformations that are not used as seeds yet, to generate new trial conformations. When all the conformations in the bank are used as seeds, one round of iteration is completed. We remove the record of bank conformations having been used as seeds, and start a new round of iteration. After these steps are repeated for a preset

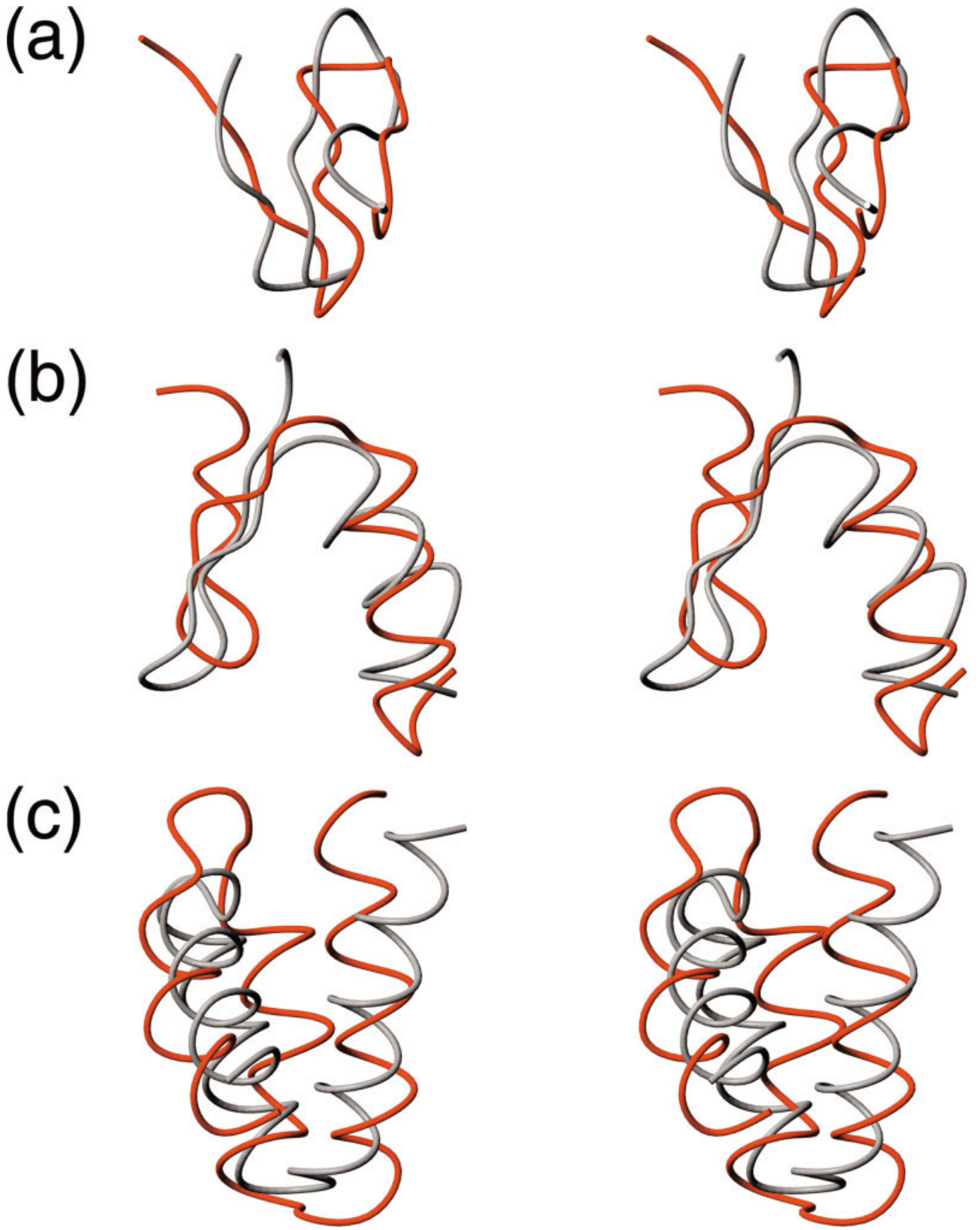


Fig. 1. The superposition of α-carbon traces of PROFESY results (gray) with their native structures (red). They are the closest conformation among the five prediction candidates, and also the closest conformations among the 100 bank conformations sampled from CSA except for 1bk2 (see text). The results are shown for (**a**) betanova, (**b**) 1fsd, (**c**) 1bdd, and (**d**) 1bk2. Prepared with the program MOLMOL.[61]


Figure 1. (Continued)

number of iterations, we conclude that our procedure has reached a deadlock. When this happens, we enlarge the search space by adding more random conformations into the bank. We repeat the whole procedure until a conformation with lower energy than a preset value is found.

### The Energy Function

The energy function used for the global optimization (and for local minimization) is given by $E = E\text{vdw} - 100N\text{hb}$ when the radius of gyration $R\text{g}$ is below the radius cutoff $R\text{cut}$ and $E = R\text{g}$ otherwise. Here, $E$vdw is the Lennard-Jones 6-12 van der Waals energy of the CHARMM forcefield[48] in the TINKER package (http://dasher.wustl.edu/tinker/), introduced in order to avoid steric clashes. $N$hb is the number of hydrogen bonds between residues, which are at least five residues apart in sequence. It should be noted that short-range hydrogen bonds are already favored by the fact that they are present in α-helical fragments. Therefore, it is not necessary to include short-range hydrogen bond energy terms. A hydrogen bond is assumed to exist when an amide hydrogen atom and a carboxyl oxygen atom are placed within 2.24 Å from each other. We used the value of radius cutoff[86] $R\text{cut} = (3N\text{seq}/0.026)^{1/3}/1.2$.

We also used an additional solvent accessible surface area solvation energy term[55] to the CHARMM forcefield for CASP5 target T0129. Since proteins do not have side-chains in our method, blindly adding the solvation term to our models had a disastrous effect of exposing naked hydrophilic backbone atoms to the surface of a protein. We realized this only after submitting models for T0129 and did not use this solvation term for the other targets.

### Clustering and Ranking Conformations for Structure Prediction

The CASP allows predictors to submit up to five models as prediction. Therefore, we select five distinct low-lying local minimum-energy conformations by grouping the final bank conformations into five clusters and choosing their representative conformations. We have used our own source code for this purpose, where the $k$-mean clustering algorithm[49] was implemented. To choose the representative conformation for each cluster, all bank conformations are ranked according to a score function based on the exposed volume with reduced radius independent Gaussian sphere approximation.[50] This score function is different from the energy function used in the CSA search, and favors the burial of hydrophobic residues and the exposure of hydrophilic residues. This score function is introduced to complement the weakness of our procedure, that the effect of solvent-side chain interaction is not incorporated into the energy function used in the conformational search by CSA. Throughout this report, the function used in the conformational search by CSA is called "energy," and the one used for ranking the final conformations obtained by CSA is called "score."

For each cluster, the conformation with the best score is chosen as its representative. However, for the CASP5 targets T0129, T0161, and T0162, the conformations at the centers of clusters are selected instead.

## RESULTS AND DISCUSSION

To test the performance of PROFESY, we have applied it both to the calculation of the tertiary structures of proteins with known structures, and to the blind prediction of proteins from the recent CASP5 targets (http://predictioncenter.llnl.gov/casp5/; group ID: 531). Since the application to the CASP5 targets had to be performed within a deadline and since the procedures of PROFESY were being developed during the CASP5 experiments, relatively primitive versions of our protocol were applied to CASP5 targets, whereas for the proteins with known structures, we applied an improved one. In particular, the hydrogen bond term was absent in the energy for CASP5 targets of T0129, T0161, and T0162.

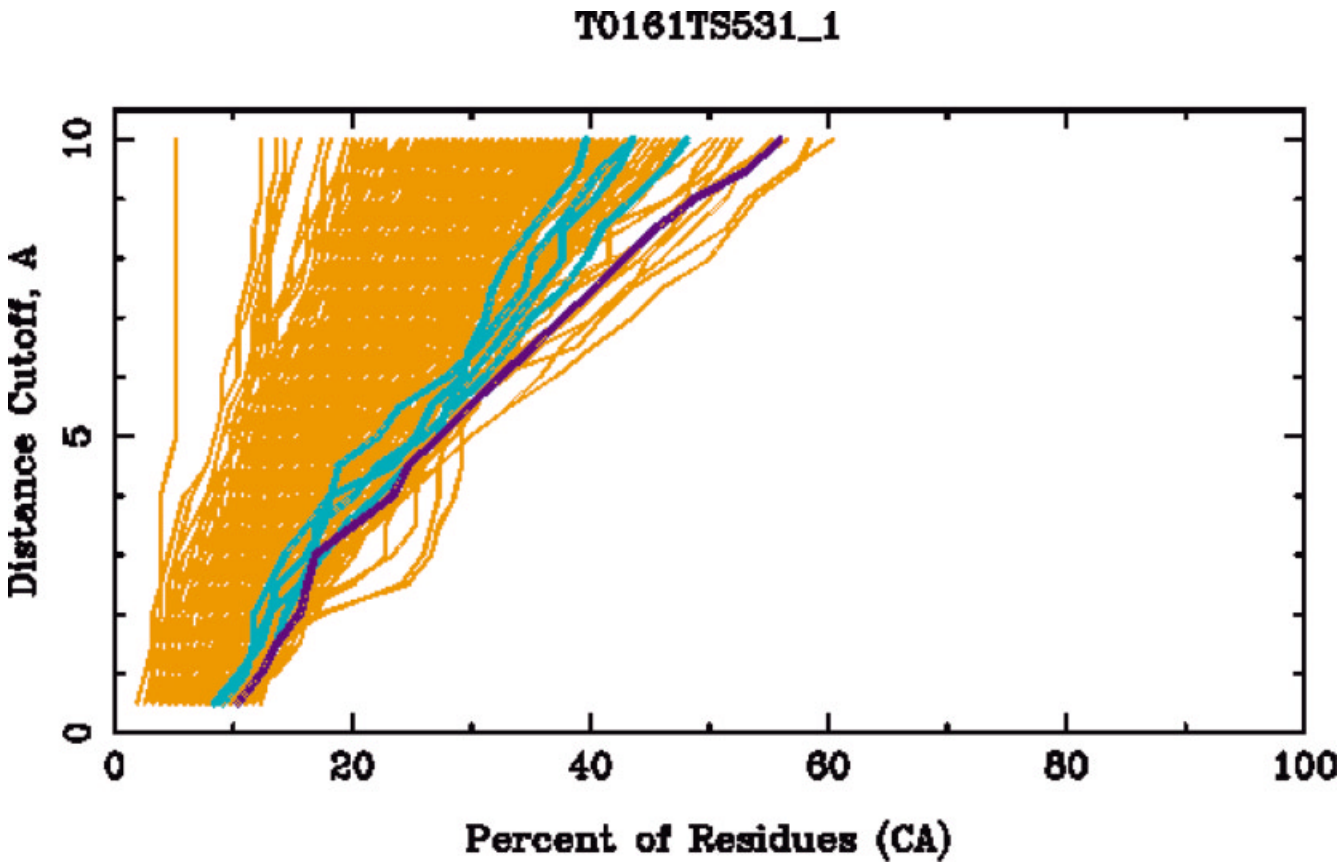


Fig. 2. The maximum number of residues that can be superposed with the native structure (horizontal axis) versus the distance cutoff for the superposition, in angstrom (vertical axis), for the CASP5 target T0161 (http://predictioncenter.llnl.gov/casp5/pubResultS/CASP_PLOTS/GDT/T0161TS531_1.html). The blue and cyan lines are the results for model one and the other four models predicted by PROFESY, respectively, whereas the orange lines are from the other predictors. The PROFESY model-one structure ranks as the third overall. When only model ones are considered, it is the best (http://predictioncenter.llnl.gov/casp5/pubResultS/CASP_BROWSER/DATA.html/3d_T0161.html).

### Test Results on Proteins With Known Structures

We first discuss the performance of the PROFESY for proteins with known structures. They are betanova (20 residues),[51] 1fsd (28 residues),[52] the fragment B of staphylococcal protein A (PDB ID 1bdd, 46 residues),[53] and A-Spectrin Sh3 Domain D48G Mutant (PDB ID 1bk2, 57 residues).[54] The α-carbon root-mean-square deviation (RMSD) of the representative conformation of each cluster from the native structure is shown in Table I. However, for 1bdd, the conformation with the best score, model one, has quite a large RMSD value of 8.9 Å, due to the fact that the parameters in the score function were determined by crude guesses without optimization. The closest predictions among the five candidates are of RMSD values 3.1, 4.0, 4.4, and 2.3 Å from their native structures for betanova, 1fsd, 1bdd, and 1bk2, respectively. These structures are also the closest conformations among the final bank conformations (100 of them in this work) sampled by the CSA search, except for 1bk2 where a structure with the RMSD value of 1.1 Å exists in the final bank. The α-carbon traces of these models are compared with those of the native structures in Figure 1. As seen from Figure 1, the predictions are quite close to their native structures, demonstrating the performance of PROFESY.

### Blind Prediction on CASP5 Targets

In order to obtain the performance of PROFESY in blind tests on sequences with unknown structures, we participated in CASP5 (group ID: 531) and applied the PROFESY procedure to four new-fold targets, and obtained rather promising results. According to the CASP5 assessment, which was performed after the experimental structures of target proteins had been provided, there are five new-fold targets, which are T0129 (HI0187, H. influenzae, 182 residues), T0149_2 (domain 2 of yjiA, *Escherichia coli*, residues 203-318), T0161 (HI1480, *H. influenzae*, 156 residues), T0162_3 (domain 3 of 286-residue protein F actin capping protein α1 subunit, chicken, residues 114–281), and T0181 (Hypothetical protein YHR087w, *S. cerevisiae*, 111 residues). Among them, the target T0161 is the most difficult one to predict according to the CASP5 evaluators, having no homologues of any kind, even in sequence databases. For multi-domain proteins, it would be ideal that we first split the proteins into domains and calculate the structure of each domain separately. However, we could not implement this procedure in time for CASP5, and we applied our procedure to the whole sequences of target proteins. Since rigorous conformational search of proteins over 300 residues was not feasible due to

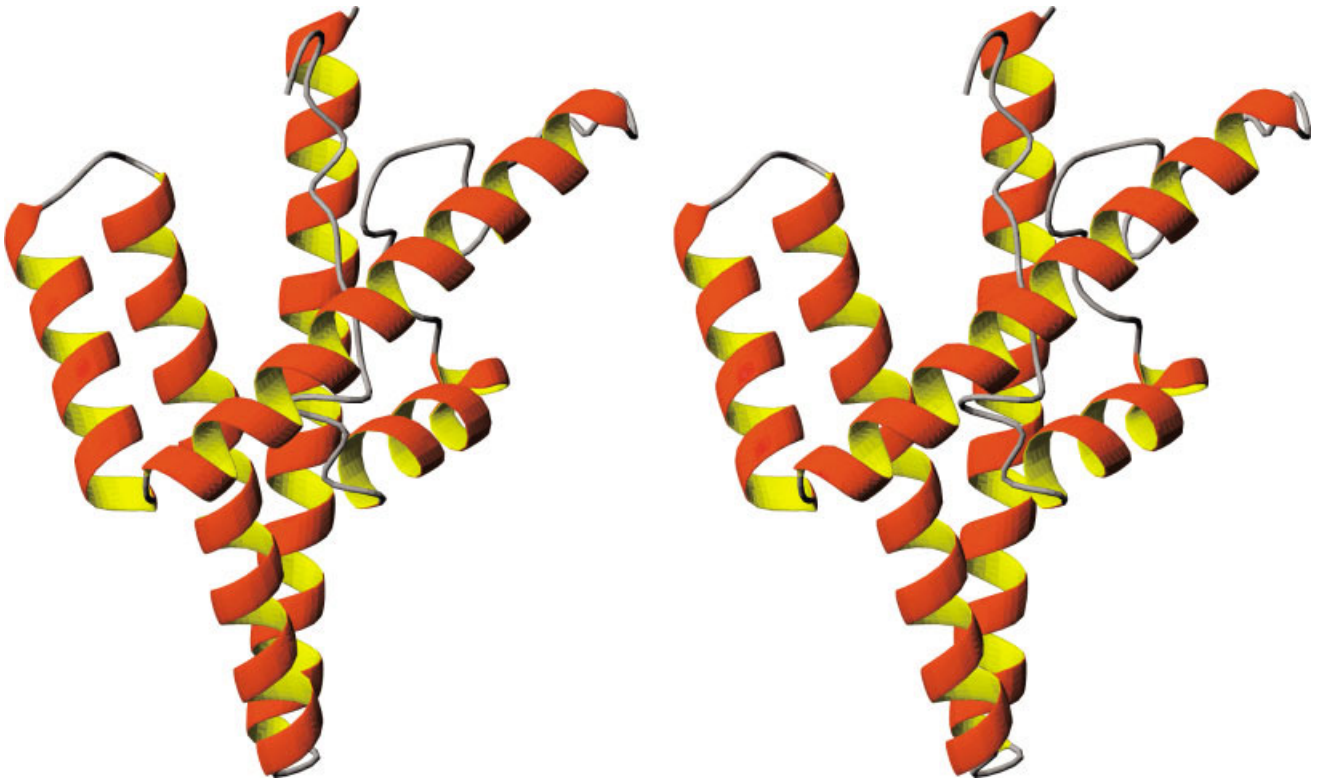

Fig. 3. The conformation from PROFESY (model one) for the CASP5 target T0161 is shown. When only model ones are considered, it is the best. The native structure is not shown because it is not yet publicly available. Prepared with the program MOLMOL.[61]

practical reasons, we did not attempt to calculate the structures of 318-residue protein T0149. Even though we submitted models for the remaining four new-fold targets, our method has illustrated quite promising performance as a whole, as shown by the evaluation at the CASP5 meeting (http://predictioncenter.llnl.gov/casp5/; group ID: 531).

The results for the target T0161 are especially promising, as shown in Figure 2, where the maximum number of residues that can be superposed with the native structure is plotted as a function of the cutoff defining the superposition. Among the five models we submitted, the model one is closest to the native structure, and it ranks as the third among all the models submitted by predictors, which is about one thousand models. If only the model ones are considered, our model one is the best (http://predictioncenter.llnl.gov/casp5/pubResultS/CASP_BROWSER/DATA.html/3d_T0161.html). However, this result is far from satisfactory in an absolute sense; the $C^{\alpha}$ RMSD of the whole chain is 16.5 Å. The conformation of the model one is shown in Figure 3. The native structure is not shown since it is not yet publicly available. The native secondary structures of residues 15–18, 115, 116, 119, 120, 123–128, 146–148 are assigned as extended according to the three-state classification. In our prediction, they appear as extended segments that are not close enough to be paired. In fact, we need long-range hydrogen bonds between fragments in order to obtain conformations with β-strands, which were absent in the energy terms used for this target.

We also obtained relatively good results for the target T0162. The results for the target T0129 are not as good, probably due to the fact that we included an additional solvent accessible surface area solvation term[55] in the energy function used in the CSA search, only for this target (see Methods). The results for T0181 are not as good as others. In this case, most of the bank conformations were similar to each other after the CSA search was terminated. We think that since the energy terms we used during the CSA run were rather incomplete due to the fact that they did not incorporate the effect of hydrophobic burial and hydrophilic exposure of side chains, most of the good conformations were removed in the early stages of the CSA calculation.

## PDB CAFASP

We have performed additional test runs on eight PDB CAFASP targets (http://bioinfo.pl/PDBCafasp/). We have selected targets that are monomers, since our methods do not yet include the effect of intermolecular interactions. We also restricted our study to the proteins less than 110 amino acid long, since rigorous application of PROFESY to larger proteins requires a huge amount of computational resources. In addition, we selected proteins with e-values for both PSI-BLAST and BLAST greater than 0.1, in order to concentrate on low sequence-homology targets. They are 1kkg (108aa), 1kwi (101aa), 1mfw (107aa), 1ny4 (82aa), 1o7b (98aa), 1owt (66aa), 1qxf (66aa), and 1ucp (91aa). The results are shown in Figures 4 and 5, where the minimum

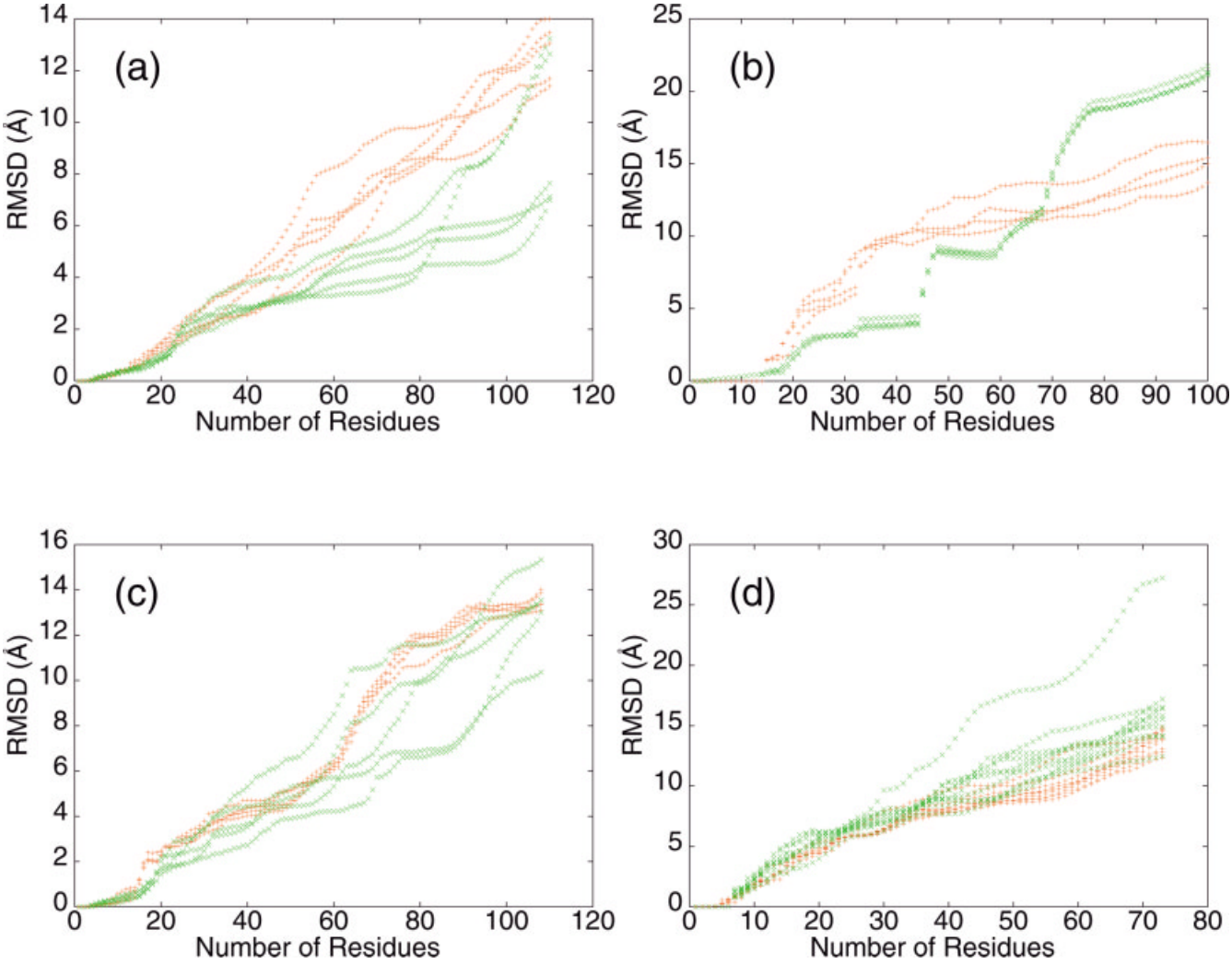


Fig. 4. The minimum value of RMSD in angstrom (vertical) versus the number of contiguous residues superposed with the native structure (horizontal). We compare the results of PROFESY (red) with those of the state-of-the-art method ROBETTA (green). We see that PROFESY shows a promising performance. The results are shown for (**a**) 1kkg (5 models), (**b**) 1kwi (4 models), (**c**) 1mfw (5 models), and (**d**) 1ny4 (10 models).

value of RMSD is plotted as a function of the number of continuous residues superposed with the native structure. We compare the results with those of ROBETTA.[56] We chose the same number of models as the ROBETTA results, being five for 1kkg and 1mfw, four for 1kwi and 1o7b, and ten for the rest. Although ROBETTA outperforms PROFESY in six out of eight cases, the overall performance of PROFESY is quite promising, considering the fact that the energy and score functions used in PROFESY are still in their early stages of development.

## CONCLUSION

In this work, we have introduced a novel method PROFESY for the prediction of protein tertiary structure, based on the fragment assembly and the rigorous conformational search by the conformational space annealing method. We applied this method to four new-fold targets in CASP5 for blind tests, which clearly demonstrates the promising performance of PROFESY. The PROFESY is also applied to proteins with known structures where 2–4 Å structures from the native structures are obtained. Although the method is in its early stages of development, these results illustrate quite a promising performance of PROFESY. The method is still under development and there exists much room for further improvement.

First of all, due to the fact that our models do not have side-chains, we could not use all-atom solvation terms directly. We have to incorporate solventside chain interaction terms indirectly. This goal may be achieved by building $C^{\beta}$atoms and introducing a pairwise interaction term between them, whose strength depends on the types of amino acids. We have not implemented this term directly in the energy function used in the CSA procedure, but used them only in the score function to select five best conformations from the final bank. In the case of the target T0181, the absence of solvent side-chain interaction had a disastrous effect that good conformations were removed from the bank during the early stages of the CSA procedure. We

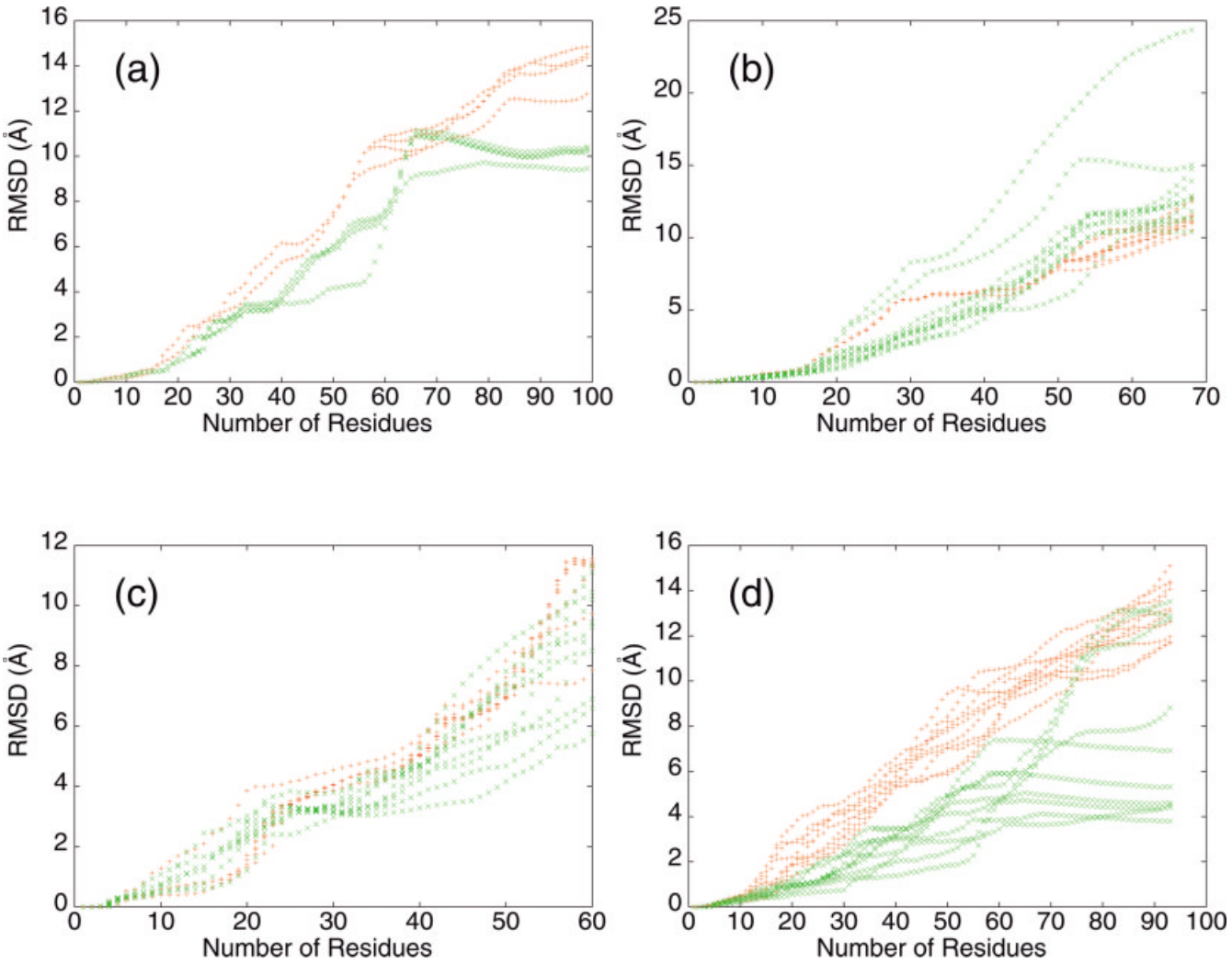


Fig. 5. The minimum value of RMSD in angstrom (vertical) versus the number of contiguous residues superposed with the native structure (horizontal). We compare the results of PROFESY (red) with those of the state-of-the-art method ROBETTA (green). We see that PROFESY shows a promising performance. The results are shown for (**a**) 1o7b (4 models), (**b**) 1owt (10 models), (**c**) 1qxf (10 models), and (**d**) 1ucp (10 models).

will have to incorporate the indirect solvation term into the energy used in the CSA.

Secondly, the relative weights of various energy terms were set in a totally arbitrary fashion. We have to optimize the values of these parameters using proteins with known structures, in such a way that our method predicts correct native structures for as many proteins as possible with optimized parameters.[57–60]

Admittedly, despite recent rapid progresses of fragment-based methods, the performance of ab initio protein structure prediction is currently far behind that of knowledge-based methods such as comparative modeling, which are the methods of choice when target proteins are homologous to those in the PDB. We believe that one of the main limitations of current fragment-based methods in the literature is that the conformational sampling is performed by algorithms that can be rather inefficient for complicated systems. PROFESY is the method that incorporates a powerful global optimization algorithm, CSA method, into a fragment-based method. It shows promising performances *despite the fact that the energy and score functions are extremely crude*. This suggests that by combining an accurate energy function and the CSA search method, one may develop a powerful structure prediction method for new-fold targets, bringing us one step closer to biologically useful applications.

## ACKNOWLEDGMENTS

This work was supported by grant R01-2003-000-11595-0 (Jooyoung Lee) and R01-2003-000-10199-0 (Julian Lee) from the Basic Research Program of the Korea Science & Engineering Foundation.

## REFERENCES

1. Moult J, Hubbard T, Fidelis K, Pedersen JT. Critical assessment of methods of protein structure prediction (CASP): round III. Proteins 1999;(Suppl)3:2–6.

2. Moult J, Fidelis K, Zemla A, Hubbard T. Critical assessment of methods of protein structure prediction (CASP): round IV. Proteins 2001;(Suppl)5:2–7.
3. Baker D, Sali A. Protein structure prediction and structural genomics. Science 2001;294:93–96.
4. Altschul SF, Madden TL, Schaffer AA, Zhang JH, 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.
5. Browne WJ, North ACT, Phillips DC, Brew K, Vanaman TC, Hill RL. A possible three-dimensional structure of bovine alpha-lactalbumin based on that of hen's egg-white lysozyme. J Mol Biol 1969:42;65–86.
6. Greer J. Comparative model building of the mammalian serine proteases. J Mol Biol 1981;153:1027–1042.
7. Blundell TL, Sibanda BL, Sternberg MJE, Thornton JM. Knowledge-based prediction of protein structures and the design of novel molecules. Nature 1987;326:347–352.
8. Havel TF, Snow ME. A new method for building protein conformations from sequence alignments with homologues of known structure. J Mol Biol 1991;217;1–7.
9. Levitt M. Accurate modeling of protein conformation by automatic segment matching. J Mol Biol 1992;226:507–533.
10. Sali A, Blundell TL. Comparative protein modeling by satisfaction of spatial restraints, J Mol Biol 1993;234:779–815.
11. Bower M, Cohen FE, Dunbrack RL. Sidechain prediction from a backbonedependent rotamer library: a new tool for homology modeling. J Mol Biol 1997;267:1268–1282.
12. Yang AS, Honig B. Sequence to structure alignment in comparative modeling using PrISM. Proteins 1999;(Suppl)3:66–72.
13. Fiser A, Do RKG, Sali A. Modeling of loops in protein structures. Protein Sci 2000;9:1753–1773.
14. Kolinski A, Betancourt MR, Kihara D, Rotkiewicz P, Skolnick J. Generalized Comparative Modeling (GENECOMP): a combination of sequence comparison, threading, lattice and off lattice modeling for protein structure prediction and refinement. Proteins 2001;44:133–149.
15. Bates PA, Kelley LA, MacCallum RM, Sternberg MJE. Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. Proteins 2001; (Suppl)5:39–46.
16. Venclovas C. Comparative modeling of CASP4 target proteins: combining results of sequence search with three-dimensional structure assessment. Proteins 2001;(Suppl)5:47–54.
17. Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. Nature 1992;358:86–89.
18. Sippl MJ. Knowledge-based potentials for preteins. Curr Opin Struct Biol 1995;5;229–235.
19. Jones DT, Thornton JM. Potential energy functions for threading. Curr Opin Struct Biol 1996;6:210–216.
20. Sippl MJ, Flockner H. Threading thrills and threats. Structure 1996;4;15–19.
21. Bohm G. New approaches in molecular structure prediction. Biophys Chem 1996;59:1–32.
22. Jernigan RL, Bahar I. Structure-derived potentials and protein simulations. Curr Opin Struct Biol 1996;6:195–209.
23. Torda AE. Perspectives in protein-fold recognition. Curr Opin Struct Biol 1997;7:200–205.
24. Koretke KK, Russell RB, Lupas AN. Fold recognition from sequence comparisons. Proteins 2001;(Suppl)5:68–75.
25. Murzin AG, Bateman A. CASP2 knowledge-based approach to distant homology recognition and fold prediction in CASP4. Proteins 2001;(Suppl)5:76–85.
26. Karplus K, Karchin R, Barrett C, Tu S, Cline M, Diekhans M, Grate L, Casper J, Hughey R. What is the value added by human intervention in protein structure prediction? Proteins 2001; (Suppl)5:86–91.
27. Williams MG, Shirai H, Shi J, Nagendra HG, Mueller J, Mizuguchi K, Miguel RN, Lovell SC, Innis CA, Deane CM, Chen L, Campillo N, Burke DF, Blundell TL, de Bakker PIW. Sequence-structure homology recognition by iterative alignment refinement and comparative modeling. Proteins 2001;(Suppl)5:92–97.
28. Lesk AM, Conte LL, Hubbard TJP. Assessment of novel fold targets in CASP4: predictions of three-dimensional structures, secondary structures, and interresidue contacts. Proteins 2001; (Suppl)5:98–118.
29. Lee J, Liwo A, Scheraga HA. Energy-based de novo protein folding by conformational space annealing and an off-lattice united-residue force field: application to the 10-55 fragment of staphylococcal protein A and to apo calbindin D9K. Proc Natl Acad Sci 1999;96:2025–2030.
30. Lee J, Liwo A, Ripoll DR, Pillardy J, Scheraga HA. Calculation of protein conformation by global optimization of a potential energy function. Proteins 1999;(Suppl)3:204–208.
31. Lee J, Liwo A, Ripoll DR, Pillardy J, Saunders JA, Gibson KD, Scheraga HA. Hierarchical energy-based approach to protein-structure prediction: blind-test evaluation with CASP3 targets. Int J Quant Chem 2000;77:90–117.
32. Liwo A, Lee J, Ripoll DR, Pillardy J, Scheraga HA. Protein structure prediction by global optimization of a potential energy function. Proc Natl Acad Sci 1999;96:5482–5485.
33. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 1997;268:209–225.
34. Simons KT, Strauss C, Baker D. Prospects for ab initio protein structural genomics. J Mol Biol 2001;306:1191–1199.
35. Bonneau R, Tsai J, Ruczinski I, Chivian D, Rohl C, Strauss CEM, Baker D. Rosetta in CASP4: progress in ab initio protein structure prediction. Proteins 2001;(Suppl)5:119–126.
36. Jones DT. Predicting novel protein folds by using FRAGFOLD. Proteins 2001;(Suppl)5:127–132.
37. Standley DM, Eyrich VA, An Y, Pincus DL, Gunn JR, Friesner RA. Protein structure prediction using a combination of sequence-based alignment, constrained energy minimization, and structural alignment. Proteins 2001;(Suppl)5:133–139.
38. Xu D, Crawford OH, LoCascio PF, Xu Y. Application of PROSPECT in CASP4: characterizing protein structures with new folds. Proteins 2001;(Suppl)5:140–148.
39. Skolnick J, Kolinski A, Kihara D, Betancourt M, Rotkiewicz P, Boniecki M. Ab initio protein structure prediction via a combination of threading, lattice folding, clustering, and structure refinement. Proteins 2001;(Suppl)5:149–156.
40. Kihara D, Lu H, Kolinski A, Skolnick J. Touchstone: an ab initio protein structure prediction method that uses threading based tertiary restraints. Proc Natl Acad Sci 2001;98:10125–10130.
41. Anfinsen CB. Studies on the principles that govern the folding of protein chains. Science 1973;181:223–230.
42. Joo K, Lee J, Kim SY, Kim I, Lee SJ, Lee J. Profile-based nearest neighbor method for pattern recognition. J Korean Phys Soc 2004;44:599–604.
43. Lee J, Scheraga HA, Rackovsky S. New optimization method for conformational energy calculations on polypeptides: conformational space annealing. J Comp Chem 1997;18:1222–1232.
44. Lee J, Scheraga HA, Rackovsky S. Conformational analysis of the 20-residue membrane-bound portion of melittin by conformational space annealing. Biopolymers 1998;46:103–115.
45. Lee J, Scheraga HA. Conformational space annealing by parallel computations: extensive conformational search of met-enkephalin and the 20-residue membrane-bound portion of melittin. Int J Quant Chem 1999;75:255–265.
46. Kim SY, Lee SJ, Lee J. Conformational space annealing and an off-lattice frustrated model protein. J Chem Phys 2003;119:10274–10279.
47. Lee J, Lee IH, Lee J. Unbiased global optimization of Lennard-Jones clusters for N≤201 using conformational space annealing method. Phys Rev Lett 2003;91:0802011–0802014.
48. MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, JosephMcCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiorkiewicz-Kuczera J, Yin D, Karplus M. All atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B 1998;102:3586–3616.
49. McQueen J. Some methods for classification and analysis of multivariate observations. In: LeCam LM, Neyman J, editors. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Volume I, Statistics. University of California Press.
50. Auspurger JD, Scheraga H. An efficient, differentiable hydration potential for peptides and proteins. J Comp Chem 1996;17:1549–1558.
51. Kortemme T, Ramirez-Alvardo M, Serrano L. Design of a 20-

amino acid, three-stranded β-sheet protein. Science 1998;281:253–256.
52. Dahiyat BI, Mayo SL. De novo protein design: fully automated sequence selection. Science 1997;278:82–87.
53. Gouda H, Torigoe H, Saito A, Sato M, Arata Y, Shimada I. Three-dimensional solution structure of the B domain of staphylococcal protein A: comparisons of the solution and crystal structures. Biochemistry 1992;31:9665–9672.
54. Martinez JC, Pisabarro MT, Serrano L. Obligatory steps in protein folding and the conformational diversity of the transition state. Nat Struct Biol 1998;5:721–729.
55. Ooi T, Oobatake M, Nemethy G, Scheraga HA. Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. Proc Natl Acad Sci 1987;84:3086–3090.
56. Chivian D, Kim DE, Malmström L, Bradley P, Robertson T, Murphy P, Strauss C, Bonneau R, Rohl CA, Baker D. Automated prediction of CASP5 structures using the Robetta server. Proteins 2003;53:524–533.
57. Lee J, Ripoll DR, Czaplewski C, Pillardy J, Wedemeyer WJ, Scheraga HA. Optimization of parameters in macromolecular potential energy functions by conformational space annealing. J Phys Chem B 2001;105:7291–7298.
58. Pillardy J, Czaplewski C, Liwo A, Wedemeyer WJ, Lee J, Ripoll D, Arlukowicz P, Oldziej S, Arnautova YA, Scheraga HA. Development of physics-based energy functions that predict medium-resolution structures for proteins of the α, β, and α/β structural classes. J Phys Chem B 2001;105:7299–7311.
59. Lee J, Park K, Lee J. Full optimization of linear parameters of a united residue protein potential. J Phys Chem B 2002;106:11647–11657.
60. Liwo A, Arlukowicz P, Czaplewski C, Oldziej S, Pillardy J, Scheraga HA. A method for optimizing potential-energy functions by a hierarchical design of the potential-energy landscape: application to the UNRES force field. Proc Natl Acad Sci 2002;99:1937–1942.
61. Koradi R, Billeter M, and Wuthrich K. MOLMOL: a program for display and analysis of macromolecular structures. J Mol Graph 1996;14:51–55.