# Protein Tertiary Structure Prediction Method based on Fragment Assembly

Julian Lee[1], Seung-Yeon Kim[2], Keehyoung Joo[2], Ilsoo Kim[2], and Jooyoung Lee[2*]

[1]Department of Bioinformatics and Life Science,
Bioinformatics and Molecular Design Technology Innovation Center ,
and Computer Aided Molecular Design Research Center Soongsil University, Seoul 156-743, Korea
[2] School of Computational Sciences, Korea Institute for Advanced Study, Seoul 130-722, Korea

*To whom correspondence should be addressed. E-mail: jlee@kias.re.kr

## Abstract

A novel method for *ab initio* prediction of protein tertiary structures, PROFESY (PROFile Enumerating SYstem), is introduced. This method utilizes secondary structure prediction information and fragment assembly. The secondary structure prediction of proteins is performed with the PREDICT method which uses PSI-BLAST to generate profiles and a distance measure in the pattern space. In order to predict the tertiary structure of a protein sequence, we assemble fragments in the fragment library constructed as a byproduct of PREDICT. The tertiary structure is obtained by minimizing the potential energy using the conformational space annealing method which enables one to sample diverse low lying minima of the energy function. We apply PROFESY for prediction of some proteins with known structures, which shows good performances. We also participated in CASP5 and applied PROFESY to new fold targets for blind predictions. The results were quite promising, despite the fact that PROFESY was in its early stage of development. In particular, the PROFESY result is the best for the hardest target T0161.

## Introduction

Understanding how a protein folds into the unique tertiary structure (the three-dimensional structure) of its native state solely from its sequence information is a great challenge in modern science. In particular, determination of protein tertiary structures from amino-acid sequences alone is one of the most important problems in molecular biology. Determining the tertiary structure of a protein is very important in understanding the function and biological role of the protein. Currently, genome-sequencing projects are producing a great 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 that provides high-resolution structure information for some proteins. 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. Computational prediction of protein tertiary structures will provide structure information on many proteins whose structures are not be determined experimentally. However, prediction of protein tertiary structures based on sequence information alone is a long-standing challenge in computational molecular biology [1-3].

The most successful methods for protein structure prediction have been 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 the unknown structure is related to that of one or more proteins with the known structures, the structures will also be similar. Therefore, the first step in protein structure prediction is to identify if the sequence of the target protein is homologous to other sequences in the sequence databases. Next, if homologies are found, then a multiple sequence alignment is generated for the homologues of the target sequence. If there is an experimental structure (that is, a template) for a homologue, comparative modeling methods are applied for predicting the tertiary structure of the target protein. In comparative modeling [1,4-14], the target sequence is aligned on to template(s), and then an all-atom structure of the target protein is produced after filling in the alignment's gaps and orienting side chains. If there is no obvious homologue, fold recognition methods are used to search for distant homologue or an analogous fold. In fold recognition from amino acid sequence [15-25], the tertiary structure of the target protein is predicted by threading the target sequence through each of the structures in a database of already known folds. Each sequence-structure alignment is assessed by a designed sequence-structure fitness function (usually a pseudoenergy function), not by sequence similarity. The main disadvantage of knowledge-based methods is that there must be a sequence with known structure

This work was supported by grant No. R01-2003-000-10199-0 (Julian Lee) and No. R01-2003-000-11595-0 (Jooyoung Lee) from the Basic Research Program of the Korea Science \& Engineering Foundation.

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) to predict the tertiary structure of a target protein [1,26-38]. *Ab initio* protein structure prediction is based on the thermodynamic hypothesis[39] which states that the native structure of a protein corresponds to the global minimum of its free energy in a given environment. *Ab initio* methods based on the thermodynamic hypothesis will be truly successful if there are both an accurate energy function and an efficient global-optimization method for searching the resultant energy landscape at the same time. There are a few *ab initio* methods that are solely based on potential energy and global-optimization methods [29-31, 33]. However, most of *ab initio* methods use information on known structures to some degree. That is why Moult *et al.* [40] have suggested that the term "*new fold* methods" should replace the traditional term "*ab initio* methods". Currently, *ab initio* methods based solely on potential energy are not so successful as those based directly or indirectly on available structural information [1-3, 32].

In this paper, we introduce a novel method for *ab initio* prediction of protein tertiary structure, PROFESY (PROFile Enumerating SYstem). This method utilizes secondary structure prediction information and fragment assembly. The secondary structure prediction of proteins is performed using the method PREDICT (PRofile Enumeration DICTionary), recently developed by Joo *et al.* [41]. For a given protein sequence, PREDICT uses a sequence-comparison method, PSI-BLAST[42], to generate profiles which define patterns for amino acid residues. Each pattern is compared with those in the pattern database generated from the PDB, and the patterns close to the query pattern is selected to determine the secondary structure of the query residue. In order to construct the tertiary structure, we also collect the backbone dihedral angles along with these patterns. These constitute a library of fragments for a given protein sequence. In order to obtain the optimal tertiary packing of these fragments, we define an energy function based on the number of long-range hydrogen bonds, the radius of gyration, and the inter-residue Lennard-Jones interactions to avoid steric clashes. Replacement of fragments by the ones in the library is carried out so that the energy function is locally minimized. The global minimization for the energy function is performed by the conformational space annealing method (CSA) [43-45] that enables one to sample diverse low-lying minima of the energy function. The resulting three-dimensional structure of the global-minimum conformation is used as a prediction for the tertiary structure of the target protein.

## Methods (Materials and Methods/ Systems and Methods etc)

### Construction of Fragment Libraries

The fragment library used in PROFESY is constructed as a by-product of the novel method of secondary structure prediction PREDICT, which was developed recently by Joo *et al.* (2003).

For each residue of the query protein, a window of size 15 is constructed, whose center is located on the residue under consideration. The fragment library is the collection of twenty fragments corresponding to the twenty nearest patterns to that of the center residue. A conformation is constructed by assembling the fragments in these libraries, and the conformations with low energies are obtained by the CSA method (See Methods). The energy function (See Methods) we minimize consists of Lennard-Jones type potential that is introduced to prevent the steric clashes, and the term that favors the formation of hydrogen bonds. When the radius of gyration exceeds a cutoff $R_{cut}$, only the radius of gyration is minimized. It should be noted that since the only selection criteria for the fragments in the fragment library are the similarity of their pattern with that of the query residue, their amino compositions do not have to be the same as the query protein. Therefore, in our method, the conformation we construct from the fragment assembly does not have the side-chains, and we cannot use an explicit solvation energy term.

**Generation of Random Conformations**

A random conformation is built from N- to C-terminal. Since the size of each fragment is fifteen residues, the first fragment is centered on the eighth residue. Therefore we first randomly pick a fragment from the fragment library corresponding to the eighth residue. Next we randomly pick a second fragment from the library corresponding to the ninth residue. The first and second fragments have fourteen overlapping residues. Among these residues, we inspect whether there is any residue whose dihedral angles have the same value for these two fragments. The dihedral angles are considered to have the same value if the backbone dihedral angles $\phi$ and $\psi$ are within 30 and 45 degrees, respectively. 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 can 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 the fragments continues until the whole chain is constructed up to the C-terminal. If at any stage 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 repeated.

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

A conformation is minimized with respect to energy by randomly choosing a residue and attempting to replace it by another one in the corresponding library. If the fragment has at least two residues whose dihedral angles agree with the neighboring fragments, then it can be joined smoothly to these neighboring fragments, similar to the case of random conformation generation. The dihedral angles are regarded as having the same value if $\phi$ and $\psi$ are within 30 and 45 degrees. If the resulting conformation is lower in energy, we accept the new conformation. This

process is continued $10N_{seq}$ times, where $N_{seq}$ is the length of the protein, or until the update attempt fails for $N_{seq}$ times, whichever is encountered first.

**Global Search Using Conformational Space Annealing Method**

The low-lying local minimum-energy conformations are obtained by a powerful global optimization algorithm, conformational space annealing (CSA) method [43-45]. 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 a local minimum, it is important to sample wide regions of the phase 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 in the bank. As in simulated annealing, we introduce an annealing parameter $D_{cut}$ that plays the role of temperature in simulated 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 $D_{cut}$. The value of $D_{cut}$ is slowly reduced just as in simulated annealing as a CSA run proceeds.

Here, we shortly mention how a CSA run proceeds. We first randomly generate a certain number of initial conformations (for example, 100) whose energy is subsequently minimized by fragment replacement 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. Then the energies of these conformations are subsequently minimized in order to obtain the new trial conformations that can be used to update the bank. A new local energy-minimum 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 which 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 from 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 any other conformation in the bank. Therefore, the conformation with the highest energy among the bank conformations and 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. All these steps are repeated for a given number of

iterations. After a preset 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 and repeat the whole procedure until the stopping criterion is met.

**The Energy Function**

The The energy function used for the global optimization is given by $E = E_{vdw} - 100\ N_{hb}$ when the radius of gyration $R_g$ is below the radius cutoff $R_{cut}$ and $E = R_g$ otherwise. Here, $E_{vdw}$ is the Lennard-Jones 6-12 van der Waals energy of the CHARMM forcefield [46] introduced in order to avoid the steric clashes. $N_{hb}$ is the number of hydrogen bonds between residues, which are at least five residues apart in sequence since the hydrogen bonding term favors alpha helices. The hydrogen bond is assumed to exist when the position of a hydrogen atom and an oxygen atom is within 5 Å. The relative weight 100 in the potential energy is totally arbitrary. We used the value of radius cutoff $R_{cut} = (3N_{seq}/0.026\pi)^{1/3}/1.2$ [27], but in a very early stage of the development of our method, somewhat larger value of $R_{cut} = (3N_{seq}/0.026\pi)^{1/3}$ was used. .

**Clustering and Ranking Conformations for Structure Prediction**

The CASP allows the predictors to submit up to five models as predictions. Therefore, we have to pick five distinct low-lying local minimum-energy conformations. We cluster the bank conformations, choose the best five clusters, and pick up a representative conformation for each cluster.

The clusters are ordered with respect to the energies of the representative conformations, and top five clusters are chosen from the bottom, regardless of their sizes. For each cluster, the conformation with the lowest score is chosen as the representative. This score is based on burial of hydrophobic residues and exposure of hydrophilic residues, where the reduced radius independent Gaussian sphere (RRIGS) approximation was used for exposed volume [47].

## Results

To test the performance of PROFESY, we applied it both to the prediction of the tertiary structures of some proteins with known structures, and to the blind prediction of some proteins from the recent CASP5 targets (http://predictioncenter.llnl.gov/casp5/). Since the application to the CASP5 targets had to be performed within a deadline, a relatively primitive version of our protocol was applied, whereas for the proteins with known structures we applied an improved one. In particular, $R_{cut} = (3N_{seq}/0.026\pi)^{1/3}/1.2$ was used for proteins with known structures and CASP5 target T0181, but for CASP5 targets T0129, T0161 and T0162, a relatively large value of $R_{cut} = (3N_{seq}/0.026\pi)^{1/3}$ was used, and the hydrogen bond term was absent in the energy altogether.

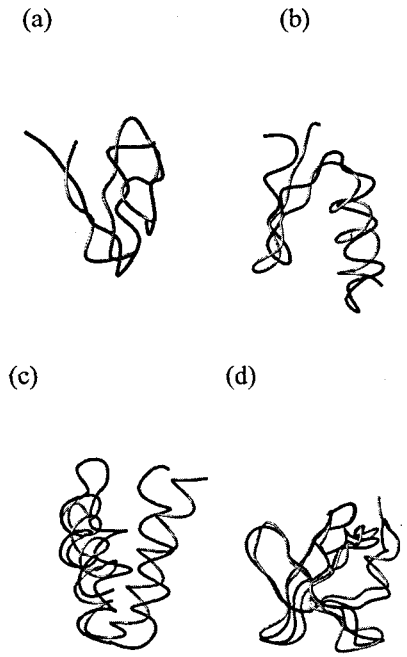


**Figure 1.** The superposition of backbone α-carbon traces of PROFESY predictions (grey) with those of the native structures (red), for proteins with known structures. The results are shown for (a) betanova, (b) 1fsd, (c) 1bdd, and (d) 1bk2.

**Test results on proteins with known structures**

We applied the PROFESY for the tertiary structure prediction of proteins with known structures for benchmark test, most of them being in the PDB. They are the designed proteins betanova (20 residues), 1fsd (28 residues), the fragment B of staphylococcal protein A (PDB ID 1bdd, 46 residues), and A-Spectrin Sh3 Domain D48G Mutant (PDB ID 1bk2, 57 residues).

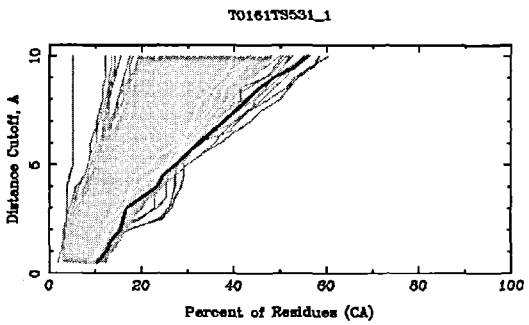


**Figure 2.** The maximum numbers of residues that can be superposed with those of the native structures (horizontal axis), along with the cutoffs defining the superposition (vertical axis), for CASP5 target T0161 (PDB ID 1MW5). The blue and cyan lines are the results for the model 1 and the other four models predicted by PROFESY, respectively, whereas the orange lines are the ones predicted by other predictors.

The best predictions for betanova, 1fsd, 1bdd, and 1bk2 have RMSDs of 3.1, 4.0, 4.4, and 2.3 Å, respectively. The backbone α-carbon traces of these models are compared with those of the native structures in Figure 1.

The CSA search was terminated after 47500, 49900, 46900, and 80500 conformations were locally minimized, for betanova, 1fsd, 1bdd, and 1bk2, respectively. The energy used as the selection criterion for the top 5 clusters are hydrophobic burial and hydrophilic exposure terms that are not used during the conformation search, as mentioned in the Method section.

**Blind Prediction on CASP5 Targets**

In order to obtain the performance of ROFESY in a blind test on completely unknown sequences, we applied the PROFESY procedure on the

CASP5 targets. There were five new fold targets, which were T0129 (HI0187, H. influenzae, 182 residues), T0149_2 (domain 2 of yjiA, E.coli, residues 203-318), T0161 (HI1480, H. influenzae, 156 residues), T0162_3 (domain 3 of 286-residue protein F-actin capping protein alpha-1 subunit, chicken, residues 114-281), and T0181 (Hypothetical protein YHR087w, S. cerevisiae, 111 residues). Among them, the target T0161 is the hardest to predict, as assessed by the CASP5 evaluators, with no homologues of any kind, even in the sequence databases. For multi-domain proteins we first have to split the protein into separate domains and apply our method to each of them, but we could not implement this procedure in time for the CASP5. Therefore we applied our method to the multi-domain protein T0162 as a whole. Since 318-residue protein T0149 was too large for computing its structure in time, we did not attempt to predict its structure. Even though we submitted the results only for the remaining four new-fold targets, our method showed good performance as a whole, as shown by the evaluation at the CASP5 meeting (http://predictioncenter.llnl.gov/casp5/; group name: 531).

In particular, the results for the target T0161 are much better than other those of other predictors, as shown by the graph 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. As shown in Figure 2, the model 1 is closest to the native structure among the five models we submitted. As can be seen from Figure 2, this model rank as the third among all the models submitted by predictors, which is about one thousand models. If only the first models are compared, our model is the best (http://predictioncenter.llnl.gov/casp5/pubResultS/CASP_BROWSER/DATA.html/3d_T0161.html). The native structure has short beta strands (residues 15-18, 115, 116, 119, 120, 123-128, 146-148) whereas our prediction consists only of alpha helices. The result for T0162 is also one of the best results. The result for the target T0129 is not as good, and we think the reason is that for this target, we also added the SASA solvation term [48] to the CHARMM forcefield [46] in the TINKER package (http://dasher.wustl.edu/tinker/). Since a protein does not have side-chains in our method, naively adding the solvation term to our model had a disastrous effect of exposing hydrophilic backbone atoms to the solvent. We realized this fact after submitting the predictions for T0129 and did not use the solvation terms for other targets. The results for T0181 are not as good as expected, although we applied improved version of our protocol. In fact, most of the bank conformations are similar after the CSA search terminates. We think that, since the energy terms we used during the CSA runs are incomplete in that it does not incorporate the effect of hydrophobic burial and hydrophilic exposure, the most of the good conformations are removed in the early stages of the CSA runs.

**Discussions**

In this work, we have introduced PROFESY, which is a novel method for prediction of protein tertiary structure based on pattern matching and fragment assembly. We applied a primitive

version of this method to the CASP5 new-fold targets for blind tests, and also slightly improved version to some proteins with known structures. Although the method is in its early stage of development, the results show excellent performances. The method is still incomplete and there is much room for improvements.

First of all, due to the fact that our model does not have the side chains, we cannot use the all-atom salvation term directly, and we have to incorporate solvation effect indirectly by using the term favoring the burial of hydrophobic residues and exposure and hydrophilic ones. We did not implement this term directly in the CSA procedure, but used them only in the final stage where we chose best five conformations among the final bank conformations. As can be shown from the poor performance for target T0181, this can have disastrous effect in that the conformations which are low in *true* energy get removed from the bank during the CSA procedure. We will have to incorporate this indirect solvation term into the energy used in the CSA.

Secondly, the relative weights of various energy terms are totally arbitrary. We have to optimize the values of these parameters using the proteins with known structures, in such a way our method predicts the correct native structure for as many proteins as possible, using the optimized parameters.

These improvements and the tests results will be reported elsewhere.

## Acknowledgements

This work was supported by grant R01-2003-000-10199-0 (Julian Lee) and No. R01-2003-000-11595-0 (Jooyoung Lee) from the Basic Research Program of the Korea Science & Engineering Foundation.

## References

[1] Baker, D. & Sali, A. (2001). Protein structure prediction and structural genomics. Science 294, 93-96.

[2] Moult, J., Hubbard, T., Fidelis, K. & Pedersen, J. T. (1999). Critical assessment of methods of protein structure prediction (CASP): round III. *Proteins: Struct. Funct. Genet.* **37 Suppl. 3**, 2-6.

[3] Moult, J., Fidelis, K., Zemla, A. & Hubbard, T. (2001). Critical assessment of methods of protein structure prediction (CASP): round IV. *Proteins: Struct. Funct. Genet.* **45 Suppl. 5**, 2-7.

[4] Bates, P. A., Kelley intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. Proteins: Struct. Funct. Genet. 45 Suppl. 5, 39-46.

[5] Blundell, T. L., Sibanda, B. L., Sternberg, M. J. E. & Thornton, J. M. (1987). Knowledge-based prediction of protein structures and the design of novel molecules. Nature 326, 347-352.

[6] Bower, M., Cohen, F. E. & Dunbrack, R. L. (1997). Sidechain prediction from a backbone-dependent rotamer library: A new tool for homology modeling. J. Mol. Biol. 267, 1268-1282.

[7] Browne, W. J., North, A. C. T., Phillips, D. C., Brew, K., Vanaman, T. C. & Hill, R. L. (1969). A possible three-dimensional structure of bovine alpha-lactalbumin based

on that of hen's egg-white lysozyme. J. Mol. Biol. 42, 65-86.

[8] Fiser, A., Do, R. K. G. & Sali, A. (2000). Mode, L. A., MacCallum, R. M. & Sternberg, M. J. E. (2001). Enhancement of protein modeling by human ling of loops in protein structures. Protein Sci. 9, 1753-1773.

[9] Greer, J. (1981). Comparative model-building of the mammalian serine proteases. J. Mol. Biol. 153, 1027-1042.

[10] Havel, T. F. & Snow, M. E. (1991). A new method for building protein conformations from sequence alignments with homologues of known structure. J. Mol. Biol. 217, 1-7.

[11] Kolinski, A., Betancourt, M. R., Kihara, D., Rotkiewicz, P. & Skolnick, J. (2001). Generalized Comparative Modeling (GENECOMP): a combination of sequence comparison, threading, lattice and off-lattice modeling for protein structure prediction and refinement. *Proteins: Struct. Funct. Genet.* **44**, 133-149.

[12] Levitt, M. (1992). Accurate modeling of protein conformation by automatic segment matching. *J. Mol. Biol.* **226**, 507-533.

[13] Sali, A.& Blundell, T. L. (1993). Comparative protein modeling by satisfaction of spatial restraints, *J. Mol. Biol.* **234,** 779-815.

[14] Venclovas, C. (2001). Comparative modeling of CASP4 target proteins: Combining results of sequence search with three-dimensional structure assessment. *Proteins: Struct. Funct. Genet.* **45 Suppl. 5**, 47-54.

[15] Bohm, G. (1996). New approaches in molecular structure prediction. Biophys. Chem. 59, 1-32.

[16] Jernigan, R. L. & Bahar, I. (1996). Structure-derived potentials and protein simulations. Curr. Opin. Struct. Biol. 6, 195-209.

[17] Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992). A new approach to protein fold recognition. Nature 358, 86-89.

[18] Jones, D. T. & Thornton, J. M. (1996). Potential energy functions for threading. Curr. Opin. Struct. Biol. 6, 210-216.

[19] Karplus, K., Karchin, R., Barrett, C., Tu, S., Cline, M., Diekhans, M., Grate, L., Casper, J. & Hughey, R. (2001). What is the value added by human intervention in protein structure prediction? Proteins: Struct. Funct. Genet. 45 Suppl. 5, 2-7.86-91

[20] Koretke, K. K., Russell, R. B. & Lupas, A. N. (2001). Fold recognition from sequence comparisons. *Proteins: Struct. Funct. Genet.* **45 Suppl. 5**, 2-7.68-75

[21] Murzin, A. G. & Bateman, A. (2001). CASP2 knowledge-based approach to distant homology recognition and fold prediction in CASP4. *Proteins: Struct. Funct. Genet.* **45 Suppl. 5**, 76-85.

[22] Sippl, M. J. (1995). Knowledge-based potentials for preteins. *Curr. Opin. Struct. Biol.* **5**, 229-235.

[23] Sippl, M. J. & Flockner, H. (1996). Threading thrills and threats. *Structure* **4**, 15-19.

[24] Torda, A. E. (1997). Perspectives in protein-fold recognition. *Curr. Opin. Struct. Biol.* **7**, 200-205.

[25] Williams, M. G., Shirai, H., Shi, J., Nagendra, H. G., Mueller, J., Mizuguchi, K., Miguel, R. N., Lovell, S. C., Innis, C. A., Deane, C. M., Chen, L., Campillo, N., Burke, D. F., Blundell, T. L. & de Bakker, P. I. W. (2001). Sequence-structure homology recognition by iterative alignment refinement and comparative modeling. *Proteins: Struct. Funct. Genet.* **45 Suppl. 5**, 92-97.

[26] Bonneau, R., Tsai, J., Ruczinski, I., Chivian, D., Rohl, C., Strauss, C. E. M. & Baker, D. (2001). Rosetta in CASP4: Progress in ab initio protein structure prediction. Proteins: Struct. Funct. Genet. 45 Suppl. 5, 119-126.

[27] Jones, D. T. (2001). Predicting novel protein folds by using FRAGFOLD. Proteins: Struct. Funct. Genet. 45 Suppl. 5, 127-132.

[28] Kihara, D., Lu, H., Kolinski, A. & Skolnick, J. (2001). Touchstone: an ab initio protein structure prediction method that uses threading-based tertiary restraints. *Proc. Natl. Acad. Sci.* **98,** 10125-10130.

[29] Lee, J., Liwo, A. & Scheraga, H. A. (1999b). 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* **96**, 2025-2030.

[30] Lee, J., Liwo, A., Ripoll, D. R., Pillardy, J. & Scheraga, H. A. (1999c). Calculation of protein conformation by global optimization of a potential energy function. *Proteins: Struct. Funct. Genet.* **Suppl. 3**, 204-208.

[31] Lee, J., Liwo, A., Ripoll, D. R., Pillardy, J., Saunders, J. A., Gibson, K. D. & Scheraga, H. A. (2000). Hierarchical energy-based approach to protein-structure prediction: Blind-test evaluation with CASP3 targets. *Int. J. Quant. Chem.* **77**, 90-117.

[32] Lesk, A. M., Conte, L. L. & Hubbard, T. J. P. (2001). Assessment of novel fold targets in CASP4: Predictions of three-dimensional structures, secondary structures, and interresidue contacts. *Proteins: Struct. Funct. Genet.* **45 Suppl. 5**, 98-118.

[33] Liwo, A., Lee, J., Ripoll, D. R., Pillardy, J. & Scheraga, H. A. (1999). Protein structure prediction by global optimization of a potential energy function. *Proc. Natl. Acad. Sci. USA* **96**, 5482-5485.

[34] Simons, K. T., Kooperberg, C., Huang, E. & Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* **268**, 209-225.

[35] Simons, K. T., Strauss, C. & Baker, D. (2001). Prospects for ab initio protein structural genomics. *J. Mol. Biol.* **306**, 1191-1199.

[36] Skolnick, J., Kolinski, A., Kihara, D., Betancourt, M., Rotkiewicz, P. & Boniecki, M. (2001). Ab initio protein structure prediction via a combination of threading, lattice folding, clustering, and structure refinement. *Proteins: Struct. Funct. Genet.* **45 Suppl. 5**, 149-156.

[37] Standley, D. M., Eyrich, V. A., An, Y., Pincus, D. L., Gunn, J. R. & Friesner, R. A. (2001). Protein structure prediction using a combination of sequence-based alignment,

constrained energy minimization, and structural alignment. *Proteins: Struct. Funct. Genet.* **45 Suppl. 5**, 133-139.

[38] Xu, D., Crawford, O. H., LoCascio, P. F. & Xu, Y. (2001). Application of PROSPECT in CASP4: Characterizing protein structures with new folds. *Proteins: Struct. Funct. Genet.* **45 Suppl. 5**, 140-148.

[39] Anfinsen, C. B. (1973). Studies on the principles that govern the folding of protein chains. Science 181, 223-230.

[40] Moult, J., Fidelis, K., Zemla, A. & Hubbard, T. (2001). Critical assessment of methods of protein structure prediction (CASP): round IV. *Proteins: Struct. Funct. Genet.* **45 Suppl. 5**, 2-7.

[41] Joo, K., Kim, I., Lee, J., Kim, S.-Y., Lee, S. J. & Lee, J. (2003). Prediction of protein secondary structure using PREDICT, a novel method based on pattern matching. Journal of Korean Physical Society, accepted.

[42] Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J. H., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl. Acids. Res. 25, 3389-3402.

[43] Lee, J., Scheraga, H. A. & Rackovsky, S. (1997). New optimization method for conformational energy calculations on polypeptides: Conformational Space Annealing. J. Comp. Chem. 18, 1222-1232.

[44] Lee, J., Scheraga, H. A. & Rackovsky, S. (1998). Conformational analysis of the 20-residue membrane-bound portion of Melittin by Conformational Space Annealing. Biopolymers 46, 103-115.

[45] Lee, J. & Scheraga, H. A. (1999a). 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.* **75**, 255-265.

[46] MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F. T. K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D. T., Prodhom, B., Reiher, W. E., Roux, B., Schlenkrich, M., Smith, J. C., Stote, R., Straub, J., Watanabe, M., Wiórkiewicz-Kuczera, J., Yin, D. & Karplus, M. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **102**, 3586-3616.

[47] Auspurger, J. D. et al. (1996). An efficient, differentiable hydration potential for peptides and proteins. J. Comp. Chem. 17, 1549-1558.

[48] Ooi, T., Oobatake, M., Nemethy, G. & Scheraga, H. A. (1987). Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. *Proc. Natl. Acad. Sci. USA* **84**, 3086-3090.