

Development of an *ab initio* Protein Structure Prediction System ABLE

Takashi Ishida¹ **Takeshi Nishimura**^{1,2} **Makoto Nozaki**¹
tak@bi.a.u-tokyo.ac.jp takeshi@bi.a.u-tokyo.ac.jp no@bi.a.u-tokyo.ac.jp

Tsuyoshi Inoue¹ **Tohru Terada**¹
ino@bi.a.u-tokyo.ac.jp tterada@bi.a.u-tokyo.ac.jp

Shugo Nakamura¹ **Kentaro Shimizu**¹
shugo@bi.a.u-tokyo.ac.jp shimizu@bi.a.u-tokyo.ac.jp

¹ Department of Biotechnology, Graduate School of Agricultural and Life Sciences, The University of Tokyo, 1-1-1 Yayoi, Bunkyo-ku, Tokyo 113-8657, Japan

² Media Center, Faculty of Letters, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

Abstract

An *ab initio* protein structure prediction system called ABLE is described. It is based on the fragment assembly method, which consists of two steps: dividing a target sequence into overlapping subsequences (fragments) of short length and assigning a local structure to each fragment; and generating models by assembling the local structures and selecting the models with low potential energy. One of the most important problems in conventional fragment assembly methods is the difficulty of selecting native-like structures by energy minimization only. ABLE thus employs a structural clustering method to select the native-like models from among the generated models. By applying the unit-vector root mean square distance (URMS) as a measure of structure similarity, we achieve more robust, effective structural clustering. When no enough clusters of good quality are obtained, ABLE runs the energy minimization procedure again by incorporating structural restraint conditions obtained from the consensus substructures in the previously generated models. This approach is based on our observation that there is a high probability that the consensus substructures of the generated models have native-like structures. Another feature of ABLE is that in assigning local structures to fragments, it assigns mainchain dihedral angles (ϕ, ψ) to the central residue of each fragment according to a probability distribution map built from candidate sequences similar to each fragment. This enables the system to generate appropriate local structures that may not already exist in a protein structure database. We applied our system to 25 small proteins and obtain near-native folds for more than half of them. We also demonstrate the performance of our structural clustering method, which can be applied to other protein structure prediction systems.

Keywords: protein structure prediction, fragment assembly method, protein folding, knowledge-based potential, structural clustering

1 Introduction

Protein structure prediction is one of the most important problems in structural biology. A number of methods for protein structure prediction have been proposed. They are divided into two classes. In the first class of methods, including threading and homology modeling, models are constructed based on at least one known structure used as templates, whose sequences are related to the target sequence (i.e., the sequence to be modeled). The templates for modeling are searched by sequence comparison methods, such as PSI-BLAST, or by sequence-structure alignment methods. These methods are limited because it is difficult to predict novel protein folds with them.

In the second class of methods, *ab initio* methods, the structure is predicted only from the sequences, without relying on the similarity at the fold level between the sequence to be modeled and any known structures. As a result, these methods can be used to predict novel protein folds that have not yet been experimentally determined. *Ab initio* methods start from the assumption that the native state of a protein has a global minimum free energy [1]. However, it is very difficult to obtain a structure in its native state, because this requires a large-scale search of the conformational space. Thus, a knowledge-based potential or database-derived potential is often used. In these potentials, various free energy parameters are related to statistics obtained from a database of high-resolution proteins structures; the potential function consists of terms reflecting the averaged structural properties of the proteins in the database.

The fragment assembly method is one of the most successful *ab initio* methods, as shown by Baker's group [3]. It is based on the assumption that short sequence segments are restricted to the local structures adopted by the most closely related sequences in the protein structure database [12]. The method consists of two steps: (1) dividing a target sequence into overlapping subsequences (fragments) of short length and assigning a local structure to each fragment, and (2) constructing models by assembling the local structures and selecting the models with low potential energy.

The first step is very effective for restricting the scope of the conformational space to be searched in the second step. One important problem in the first step, however, is that if appropriate fragments cannot be found or do not exist in the database, the correct fold cannot be obtained, even if the energy function is accurate and the energy minimization in the second step is efficient. To solve this problem, we propose a new method of assigning local structures. Our method assigns mainchain dihedral angles (ϕ, ψ) to the central residue of each fragment according to a probability distribution map built from candidate sequences "similar to" each fragment, whereas Baker *et al.* [12] select certain "similar" sequences and assign dihedral angles to all the residues in those sequences. Our method is more flexible because dihedral angles are assigned individually to each residue. In addition, we construct the probability distribution map from candidate sequences so as to restrict the scope of the conformational space, unlike methods using an ordinary Ramachandran map.

In the second step, it is very difficult to select native-like models based only on energy minimization, because the knowledge-based potential is not accurate enough to ensure that its minimum structure corresponds to a native state (Fig. 1). Shortle and Baker [10] first proposed using structural clustering to select native-like models. Their method searches for the largest cluster of structurally related low-energy conformations, instead of focusing on the lowest energy conformation. Structural clustering has also been used in several other protein structure prediction systems [2, 3]. Most of these systems use the root-mean-square deviation (RMSD) or its variation as a measure of structure similarity (i.e., an index of the distance between structures). However, using the RMSD causes a problem in that when the distribution of candidate structures is broad, clustering is apt to fail. This is because the RMSD value may become large even when only small parts of the structures are predicted to be different. Thus, we adopted the unit-vector root mean square distance (URMS) [6] as a measure of structure similarity. This is a normalized measure that is almost independent of the protein size and robust with respect to differences between small portions of structures. In addition, we introduce the cluster density, a measure for evaluating the quality of the clusters. We select the centers of each cluster with a density higher than a threshold value for the final predicted models. If not enough clusters with sufficient quality are obtained, we try to run a new energy minimization procedure incorporating distance restraints between the C_α atoms, which are obtained from the consensus substructures in the models generated by the previous minimization procedure. This technique is based on the high probability that the consensus substructures of the generated models have near-native structures. This is another original feature of our prediction method.

The rest of this paper describes the details of our method and the results obtained by applying it to 25 small proteins. Our clustering method is also applicable to other *ab initio* protein structure prediction systems that generate a number of candidate structures. Here, we present results obtained

by applying our clustering method to Baker’s Rosetta decoys [13].

2 Methods

First, we utilize only the mainchain as a prediction target. We then independently predict the sidechains after mainchain prediction. We use a representation of the mainchain dihedral angles ϕ and ψ to reduce the number of parameters by fixing the bond lengths and the angles of the peptide bonds to specific ideal values (C_α -C = 1.53 Å, C-O = 1.24 Å, C-N = 1.32 Å, N- C_α = 1.47 Å, C_α -C-O = 121°, C_α -C-N = 114°, O-C-N = 125°, C-N- C_α = 123°, N- C_α -C = 110°). If calculated sidechain interactions are required in the minimization process, the C_β atom represents all the sidechain atoms. For glycine, a virtual C_β atom is introduced since it does not have a C_β atom.

The procedure for protein structure prediction in ABLE consists of the following steps:

1. Assignment of local structures – A target sequence is divided into overlapping fragments of short length and a local structure is assigned to each fragment.
2. Model generation – Models are constructed by assembling the local structures, and they are evaluated by using knowledge-based potential functions. Monte Carlo simulated annealing is performed to select the models with low potential energy.
3. Structural clustering – The selected models are clustered according to the measure of structural similarity in order to further select native-like models.
4. Iterative prediction (optional) – A new simulation with distance restraints between C_α atoms is executed if enough clusters of good quality are not obtained.

2.1 Assignment of Local Structures

To assign local structures to each fragment, ABLE searches a protein structure database for sequences "similar to" each fragment. As the structure database, we use NCBI’s nrpdb (non-redundant PDB chain set) [14] with a p-value of 10^{-7} , and we remove proteins determined by NMR, proteins with low resolution ($> 2.5\text{Å}$), membrane proteins, proteins with only C_α atoms, and proteins with missing residues. In ABLE, the sequence similarity score between the target fragment and the subsequences in the protein structure database is defined as

$$S_{ij} = \lambda \sum_i^N \sum_j^{20} \sum_k^{20} M(a_k, a_j) \cdot P_{tpl}(i, a_j) \cdot P_{tgt}(i, a_k) \\ + (1 - \lambda) \sum_i^N \sum_s^3 ssconf(i, s) \cdot \delta(ss_{i,s}^{tpl}, ss_{i,s}^{tgt}),$$

where $P(i, a_j)$ is the probability that the i th residue is amino acid type j ; N is the sequence length (normally we use $N = 9$); \mathbf{M} is the BLOSUM62 scoring matrix [4]; λ is a weight parameter used to combine the sequence similarity and the consensus of the secondary structures (generally we use $\lambda = 0.5$); $ssconf$ is a confidence value for secondary structure prediction (for lower $ssconf$ values, the factor of secondary structure prediction is decreased); and $ss_{i,s}$ is the secondary structure type of the i th residue. All matrix elements are normalized so that they take a value from zero to one. We use PSIPRED [5] to predict the secondary structures for the target structure. For the sequences whose similarity scores are higher than the threshold, ABLE creates a local map: each (ϕ, ψ) value of the central residue is clamped to one range within a discrete range of dihedral angle values, and the appearances of each range are counted. Then, ABLE calculates the probability distribution map

from the appearance frequencies. The quality of the map is improved by narrowing the discrete range, but this causes shortage of data in each range. Therefore, the map is smoothed by being convoluted with a two-dimensional Gaussian function. This allows probabilistic generation of local structures (dihedral angles) that may not exist in the database.

2.2 Model Generation

2.2.1 Scoring Functions

We designed knowledge-based potential functions to seek the most probable structure for a protein given its amino acid sequence and a large number of sequences with known structures in the protein structure database. These functions are designed to return better score for native-like structures and are composed with potential terms indicating certain physicochemical aspects of the proteins. All the potential terms are weighted and summed up to generate a score function. Thus, our score function is

$$E_{total} = w_{burial}E_{burial} + w_{pho}E_{pho} + w_{pair}E_{pair} + w_{hb}E_{hb} \\ + w_{ss}E_{ss} + w_{VDW}E_{VDW} + w_{compact}E_{compact}$$

Burial potential A force causing hydrophobic residues to be buried in the protein core and polar residues to be exposed to solvents is one of the major driving forces folding proteins. The potential term represents this force as the number of residues around a particular noticed residue. A large number indicates that the residue is buried, while a small number indicates that it is exposed. Here, we use the following probability calculated from the database.

$$E_{burial} = -\log\left(\prod_i P(aa_i|E_i)\right), \quad (1)$$

where aa_i means residue i is of amino acid type aa , and E_i is the number of residues around residue i smaller than 10 Å.

Convergence of hydrophobic residues Although the burial potential represents the burial tendency of hydrophobic residues effectively, it permits a small multiple hydrophobic core. This term is designed to collect the hydrophobic residues into one hydrophobic core more effectively, and it is defined as

$$E_{pho} = \sum_{i < j, h_i < 0, h_j < 0} r_{ij} h_i h_j, \quad (2)$$

where r_{ij} is the distance between the C_β atoms of hydrophobic residues i and j , and h_i and h_j are the hydrophobic scores (defined by Kyte & Doolittle [9]) for residues i and j , respectively.

Pair potential This term indicates the affinity between residues. We use a similar term defined by Simons *et al.* [11]

Hydrogen bond This term is a potential proportional to the number of hydrogen bonds. The electrostatic potential is calculated from the distance between CO and HN, and we assume that there is a hydrogen bond if this potential is less than a threshold of -0.5 kcal/mol. The electrostatic potential is defined as

$$E_{electrostatic} = Q\left(\frac{1}{r_{ON}} + \frac{1}{r_{CH}} - \frac{1}{r_{OH}} - \frac{1}{r_{CN}}\right), \quad (3)$$

where $Q = 0.42 \times 0.20 \times 332$. In addition, the hydrogen bond potential is defined as

$$E_{hb} = -(N_{hb} + c_{hb_{adj}}N_{hb_{adj}}), \quad (4)$$

where N_{hb} is the number of hydrogen bonds, and the term $c_{hb_{adj}}N_{hb_{adj}}$ is a bonus reflecting the number of adjacent hydrogen bond pairs in the primary sequence, with $c_{hb_{adj}}$ equal to 0.3.

Secondary structure packing This term represents the packing of secondary structural elements, such as β sheets, and α helix packing. We again use a similar term defined by Simons *et al.*[11] The secondary structural elements are transformed into simple vectors, and the term is calculated from the angles and distances between vector pairs.

VdW force There is no overlap between atoms in native structures because of the VdW force. This term is the sum of the penalties for the steric crash in the protein. This penalty is applied if the distance between the C_α atoms is less than 3.0 Å.

Compactness Most native conformations are very compact. We use the radius of gyration as a term representing the compactness of the protein. We calculated the minimum radius of gyration for a protein with n residues from the database. We then set up the following equation by approximating an exponential distribution of the differences between the radii of gyration of the proteins with less than 200 residues and the minimum radius of gyration:

$$E_{compact} = -\log(0.34 \exp(-0.34(R_g - (0.064n + 10.5))))), \quad (5)$$

where n is the number of residues, and R_g is the radius of gyration.

Weight factors of each term The various terms in the scoring function are not completely independent of each other. Therefore, it is difficult to determine the weight factors of these terms. To do so, we create a non-redundant decoy including 10 small proteins and optimize it by a simulated annealing method to minimize the Z-score for near-native structures, which are defined as those with a RMSD of less than 6.5 Å from the native structure.

2.2.2 Searching Conformational Spaces

To search the conformational spaces, we use two types of structure transition. One is a move to change a dihedral angle based on the probability density map, while the other is just a fragment replacement. Employing these two structure transitions, ABLE searches for low-energy conformations by applying a simple Monte Carlo simulated annealing method. It takes about five minutes to generate one model with about 100 residues on a 2.4-GHz Intel Pentium 4 processor. Usually, we generate 1000 models per prediction target.

2.3 Structural Clustering

Here, we use a novel clustering technique to evaluate the generated models and select an appropriate model as a final prediction. We adopted the URMS as a distance measure for clustering, while most previous approaches used the RMSD. For calculating URMS, C_α backbone of protein is transformed into sequence of unit vectors in the direction from i th C_α to $i + 1$ th C_α . Then, all of the unit-vectors are placed at origin. URMS is the RMS distance for the two sequences of these unit-vectors. Our clustering method is based on Kohonen's self-organizing map (SOM)[8]. The quality of the resulting cluster is evaluated in terms of the cluster density, defined as

$$D = \frac{n^2}{N \sum_i d_i}, \quad (6)$$

where d_i denotes the URMS distance between the cluster center and cluster member i , n is the number of cluster members, and N is the total number of structures generated. A large value of D means that the cluster has a certain number of members and that those members have similar structures.

There is an inverse correlation between the cluster density and the RMSD from the native structure of the cluster center (Fig. 2). Thus, whether the prediction succeeds or not is judged from the

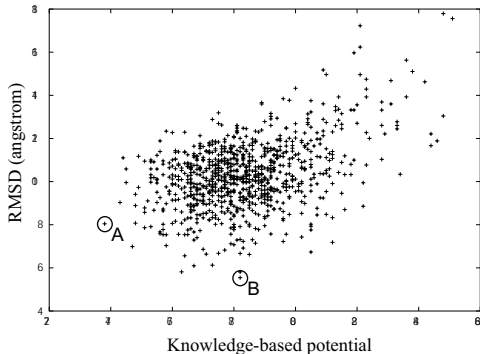


Figure 1: Correlation between the knowledge-based potential and the RMSD from the native structure for the predicted structures of 1erv. The model with the lowest score (A) does not correspond to the best RMSD model (B).

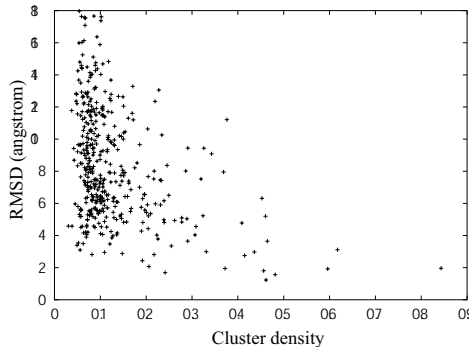


Figure 2: Correlation between the cluster density and the RMSD from the native structure of the cluster center for 25 proteins of Table 1.

maximum density, D_{max} , of all clusters. If $D_{max} > d$, the prediction is judged successful. We use a threshold of $d = 0.112$. This value was chosen because it enabled this method to select near-native structures whose RMSD with respect to a native structure was less than 6.5 Å from among 70% of the decoys generated by the ABLE system (Fig. 2). If the simulation succeeds, the centers of the top five clusters with density higher than the threshold are selected as the final models.

2.4 Iterative Prediction

If not enough clusters with sufficient quality are obtained, we run a new simulation with distance restraints between the C_α atoms, which are produced from the consensus substructures of the generated models. This approach is based on the observation that there is a high probability that most of the consensus substructures in the generated models will be similar to the native substructure. The degree of substructure consensus is evaluated in terms of the standard deviation of the distance between corresponding C_α atoms in the generated models. The distance restraints are given as

$$E_{restraint} = \sum_{i>j} e_{ij}, \tag{7}$$

$$e_{ij} = \begin{cases} |d_{ij} - \bar{d}_{ij}| - W & \text{if } |d_{ij} - \bar{d}_{ij}| > W \\ 0 & \text{otherwise.} \end{cases}, \tag{8}$$

where d_{ij} is the distance between C_α atoms in native structure, \bar{d}_{ij} is the median of the distances between C_α atoms in the generated models, and W is a sufficient margin width to avoid the effects of outliers. The margin width is defined as

$$W = 1.176\sigma + 0.303, \tag{9}$$

where σ is the standard deviation. The margin width is determined by following procedure. First, the errors between the native C_α distance and the C_α distance in the consensus substructure are divided by standard deviations into the bin with size 0.5 Å. Next, the higher ten percent of errors in each bin are removed. The margin width is determined by least mean-square approximation to the maximum errors of these bins (Fig. 4). The distance restraints are applied to the substructures for which the above standard deviation is less than 4.0 Å. This is because the correlation between the C_α distances in the consensus substructures of generated models and the native C_α distances becomes lower as the standard deviation is larger (Fig. 3). The above procedure of simulation, clustering, and distance restraint generation is then repeated until enough clusters of good quality are obtained.

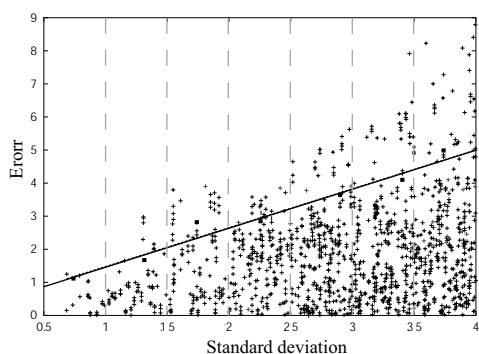


Figure 3: Correlation between the errors between the C_α distances in the consensus substructures of generated models and the native C_α distances, and the standard deviations of the C_α distances.

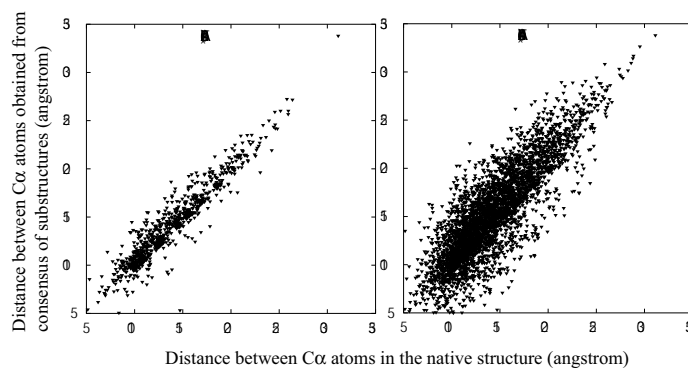


Figure 4: Comparison of the C_α distances in the consensus substructures of generated models and the native C_α distances. The left figure shows the C_α distances for a standard deviation threshold of 4.0 Å while the right one shows the distances for a threshold of 2.0 Å. The C_α distances obtained from pairs that were closer than 5 residues to the primary sequence or that were in the same secondary structure element have been removed.

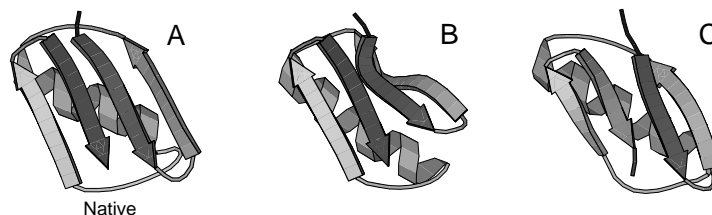


Figure 5: (A) Native structure of 1gb1; (B) The best RMSD model of 1gb1 selected by our clustering method; (C) The best RMSD model of all.

3 Results and Discussion

3.1 Prediction Results

We applied our prediction approach to 25 small proteins. The test set included seven α proteins, six β proteins, and 12 α/β proteins. None of the proteins had any homologues in the database used to generate the potentials and fragment candidates. The results of this blind test are summarized in Table 1. Here, we define a near-native structure as one whose RMSD from the native structure is less than 6.5 Å [7]. We could predict 11 proteins successfully on the first attempt, and two more folds were successfully predicted after refinement by iterative simulation with distance restraints. Especially small α and α/β proteins were successfully predicted. Prediction for 1gb1 was the most successful in this test set, and the prediction models are shown in Figure 5. The best RMSD model among all models almost corresponded to native structure and the model selected by our method was also very similar to native.

3.1.1 Results of Clustering

We compared the clustering method of Baker's group in CASP4 [3] (Column 5 in Table 1) with our clustering method (Column 4 in Table 1). On average, our method could select better folds than the

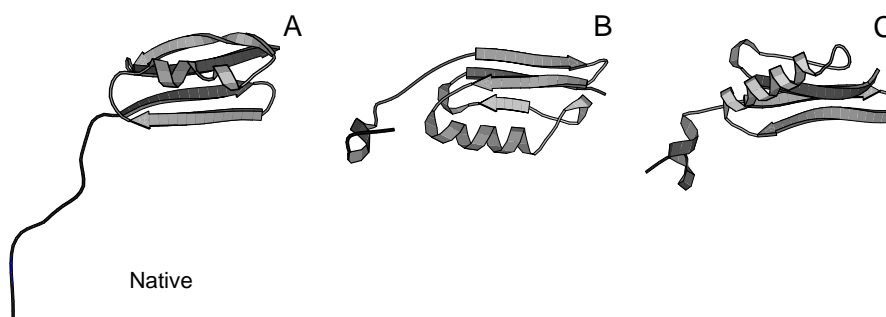


Figure 6: (A) Native structure of 2ptl; (B) The model of 2ptl selected by the clustering method based on RMSD; (C) The model of 2ptl selected by our clustering method.

clustering method in [3]. There were only five cases (among 25) in which our method performed worse than the clustering method in [3]. Additionally, in two of these cases the cluster density was lower than the threshold, and in three cases, the folds selected by our method also near-native folds. Figure 6 shows the typical case that our method was superior. There is a long disordered region at N-terminal in 2ptl and the small changes in this N-terminal region causes major influence on the RMSD value. Therefore, the clustering method based on RMSD could not capture the good consensus at core region because of the bad effect of this N-terminal region. In contrast, the URMS is robust for the changes in such region, and our method was able to capture the good consensus at core region and selected the better model.

Test for the Rosetta decoy set Our clustering method to evaluate models and select appropriate conformations is independent of model generation, and it can be applied to the models generated by other prediction systems. We thus applied the clustering method to the Rosetta decoy set [13]. This decoy set contains about 1000 models generated by the Rosetta prediction system for 92 small proteins, and the decoys for 55 proteins include near-native structures ($< 6.5 \text{ \AA}$). In the 41 cases (among 55), our method could select near-native structures. Here, as well, we compared the clustering method in [3] with our clustering method. Our method could select better folds than the clustering method in [3] in 60 cases (among 92), and average RMSD of models selected by our method is superior to it.

3.1.2 Results of Iterative Prediction

Most of the folds were refined with distance restraints in the final model (Column 8 in Table 1) selected by clustering, and only two cases (among nine) were worse than the result of the first prediction. Figure 7 shows the successful refinement of 1aa3. In the case of 1aa3, the 588 distance restraints were obtained from the generated models in the first prediction. The 333 of them were between C_α atoms on three α helices, and there were only 8 cases in which the errors between the native C_α distance and the C_α distance in the consensus substructures of generated models were beyond the margin width. Thus, the spatial relationship among the three helices was correct in most of the models generated by second prediction with the distance restraints. This refined the convergence of models and made it easier to select near-native folds. However, there were 62 distance restraints between C_α atoms on the β sheets, and in the 51 case, the errors were beyond the margin width. Therefore, most of the models had no correct substructure in the β sheet region, and the best RMSD value among all models became worse after prediction with distant restraints. In fact, the best RMSD among all models became worse after new simulation with distance restraints in half the cases. This is because the wrong distance

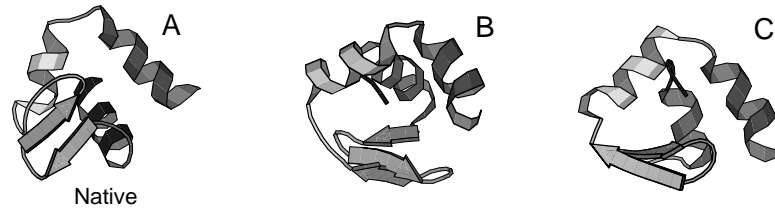


Figure 7: (A)Native structure of 1aa3; (B) The model of 1aa3 predicted first; (C)The model of 1aa3 predicted with distance restraints.

restraints could not remove perfectly. To remove these wrong distance restraints and determine the better margin width is one of the our future works.

Table 1: Summary of prediction results

id ¹	N_{aa} ²	Type	$Cl_u(\text{\AA})$ ³	$Cl_r(\text{\AA})$ ⁴	Best(\AA) ⁵	D_{max} ⁶	$Cl_{re}(\text{\AA})$ ⁷	Best $_{re}(\text{\AA})$ ⁸	$D_{re}(\text{\AA})$ ⁹	N_p ¹⁰
1aa3	63	α/β	7.39	7.59	3.93	0.090	5.83	4.75	0.113	1
1ag2	97	α/β	8.77	9.79	5.25	0.102	8.24	5.92	0.113	
1ah9	63	β	7.83	8.94	6.63	0.099	6.69	5.59	0.132	2
1crb	134	α/β	10.57	12.64	9.42	0.103	10.66	9.32	0.122	3
1csp	64	β	4.25	4.25	4.05	0.128				
1ctf	67	α/β	4.04	3.00	2.54	0.179				
1dro	122	α/β	12.37	12.16	8.67	0.071	- ¹¹	-	-	-
1erd	40	α	5.43	4.91	3.59	0.252				
1erv	105	α/β	7.13	7.51	5.63	0.104	8.25	7.27	0.114	1
1fk5	93	α	4.63	4.63	4.46	0.137				
1gb1	54	α/β	3.46	3.57	1.79	0.208				
1h40	69	α	4.61	4.91	3.59	0.252				
1hi2a	135	α/β	13.86	14.50	11.18	0.087	- ¹¹	-	-	-
1ngr	85	α	6.08	9.54	5.09	0.136				
1plc	99	β	12.19	12.65	8.80	0.091	10.13	8.91	0.116	2
1poh	85	α/β	6.03	6.03	5.21	0.135				
1rpo	61	α	3.86	3.40	2.57	0.287				
1stu	68	α/β	6.33	6.39	4.40	0.165				
1utg	70	α	8.13	8.44	7.61	0.301				
1vif	52	β	8.45	9.41	6.41	0.118				
1who	88	β	9.41	7.79	7.09	0.104	- ¹¹	-	-	-
1wiu	93	β	9.81	10.74	8.92	0.100	9.56	8.77	0.118	2
2mhr	118	α	8.59	8.59	5.30	0.111	6.23	5.57	0.122	1
2ptl	78	α/β	6.48	8.32	6.48	0.114				
5icb	72	α/β	5.14	5.33	3.35	0.217				
Average			7.39	7.88						

¹ PDB id.

² The number of residues.

³ The best RMSD among the top five cluster centers with our clustering method.

⁴ The best RMSD among the top five cluster centers with the clustering method of Baker's group in CASP4 [3].

⁵ The best RMSD among all models.

⁶ The maximum cluster density.

⁷ The best RMSD among the top five cluster centers with our clustering method after refining the models in a new simulation with distance restraints.

⁸ The best RMSD among all models after refinement with distance restraints.

⁹ The cluster density after refining the models in a new simulation with distance restraints.

¹⁰ The number of iterations.

¹¹ Iterative prediction was aborted because this target could not be converged upon in 5 attempts.

4 Conclusion

We are now analyzing the effects of each term of the knowledge-based potential in detail to enable our method to produce better sets of models. This is important for improving the quality of the results of structural clustering. The design of the potential function for relatively large proteins consisting of more than 100 residues is another topic for future work. A potential function considering the topologies of proteins, inter-domain interactions, and other factors will be necessary. We are also developing a parallel algorithm for our methods in order to speed up prediction. To obtain more precise modeling, sidechain modeling and structure refinement using the physical potential can be applied.

References

- [1] Anfinsen, C., Principles that govern the folding of protein chains, *Science*, 181:223–230, 1973.
- [2] Betancourt, M. and Skolnick, J., Finding the needle in a haystack: educing native folds from ambiguous ab initio protein structure predictions, *J. Compt. Chem.*, 22:339–353, 2001.
- [3] Bonneau, R., Tsai, J., Ruczinski, I., Chivian, D., Rohl, C., Strauss, C.E.M., and Baker, D., Rosetta in CASP4: progress in ab initio protein structure prediction, *Proteins*, Suppl 5:119–126, 2001.
- [4] Henikoff, S. and Henikoff, J., Amino acid substitution matrices from protein blocks, *Proc. Natl. Acad. Sci.*, 89:10915–10919, 1992.
- [5] Jones, D.T., Protein secondary structure prediction based on position-specific scoring matrices, *J. Mol. Biol.*, 292:195–202, 1999.
- [6] Kedem, K., Chew, P., and Elber, R., Unit-vector RMS(URMS) as a tool to analyze molecular dynamics trajectories, *Proteins*, 37:554–564, 1999.
- [7] Kihara, D., Kolniski, A., and Skolnick, J., Touchstone: an ab initio protein structure prediction method that uses threading-based tertiary restraints, *Proc. Natl. Acad. Sci.*, 98:10125–10130, 2001.
- [8] Kohonen, T., *Self-Organizing Maps*, Springer Series in Information Sciences, 1995.
- [9] Kyte, J. and Doolittle, R.F., A simple method for displaying the hydropathic character of a protein, *J. Mol. Biol.*, 157:105–132, 1982.
- [10] Shortle, D., Simons, K.T., and Baker, D., Clustering of low-energy conformation near the native structures of small proteins, *Proc. Natl. Acad. Sci.*, 95:11158–11162, 1998.
- [11] Simons, K.T., Ruczinski, I., Kooperberg, C., Fox, B.A., Bystroff, C., and Baker, D., Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins, *Proteins*, 34:82–95, 1999.
- [12] Simons, K.T., Kooperberg, C., Haung, E., and Baker, D., Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring function, *J. Mol. Biol.*, 286(2):209–225, 1999.
- [13] <http://depts.washington.edu/bakerpg/>.
- [14] <http://www.ncbi.nlm.nih.gov/Structure/VAST/nrpdb.html>.