Reconstruction of Protein Backbone with the α-Carbon Coordinates *

Jen-Hui Wang, Chang-Biau Yang† and Chiou-Ting Tseng
Department of Computer Science and Engineering
National Sun Yat-sen University, Kaohsiung, Taiwan
†cbyang@cse.nsysu.edu.tw

Abstract

Given an amino acid sequence with the α-carbon 3D coordinates on its backbone, the *all-atom protein backbone reconstruction problem* (PBRP) is to rebuild the 3D coordinates of all atoms (N, C and O atoms) on the backbone. In this paper, we propose a method for solving PBRP based on the homology modeling. First, we extract all consecutive 4-residue fragments from all protein structures in PDB. Each fragment is identified by its second, third and fourth residues. Thus, the fragments are classified into 8000 residue groups. In each residue group, the fragments with similar structures are clustered together. And one typical fragment is used to represent one cluster. These typical fragments form our fragment library. Then, we find out possible candidates in the fragment library to reconstruct the backbone of the target protein. To test the performance of our method, we use two testing sets of target proteins, one was proposed by Maupetit *et al.* [13] and the other is a subset extracted from CASP7. We compare the experimental results of our method with three previous works, MaxSprout, Adcock's method, and SABBAC proposed by Maupetit *et al.* The reconstruction accuracy of our method is comparable to these previous works. And the solution of our method is more stable than the previous works in most target proteins.

Key words: protein, backbone, reconstruction, fragment, α-carbon.

1 Introduction

Much three-dimensional information for proteins has been collected in *Protein Data Bank* (PDB) [3]. However, some important proteins are only confined to the coarse grained model [10]. In other words, they have only α-carbon coordinates. And many fields which use the coarse grained model to describe protein structures right now should be expanded into the all-atom model to have higher accuracy, such as protein fold generation, and modeling of experimental data obtained at low resolution. Because the coarse grained model only offers partial information of the residues and their rough positions in 3D space, we need to devise efficient algorithms to reconstruct protein backbone atoms with known α-carbon coordinates. Modeling from atomic coordinates and all-atom protein reconstruction has been studied extensively, including improving the low resolution models from crystallography, *ab initio* (or *de novo*) folding computation, or comparing protein conformations to reconstruct the all-atom model. The full protein generation can usually be divided into two parts, backbone coordinates prediction and side chain positioning. Hsin *et al.* [7] presented much information of side chain positioning. In this paper, we shall discuss only the prediction of coordinates of all atoms on the protein backbone. For a given amino acid sequence with the α-carbon 3D coordinates on the backbone, the *all-atom protein backbone reconstruction problem* (PBRP) is to rebuild the 3D coordinates of all atoms (N, C and O atoms) on the backbone. Note that PBRP does not involve the information of atoms on the side chain.

In general, the approaches for PBRP can be classified into two types: exploitation of small fragment similarity to known protein structures [1, 6, 14, 13, 8] and minimization of local molecular energy [9, 11]. There are also some approaches combining the above two types of methods. The first type of method is usually to utilize a fragment library extracted from known protein structures to assemble fragments by using energy-based, homology-based or geometry criteria to generate a polypeptide chain that is optimal and consistent with α-carbon trace. Milik *et al.* [14] exploited the statistics of known protein structures to generate atom positions and to reconstruct the all-atom protein backbone. Iwata *et al.* [8] designed a method to analyze and to select fragments compatible with favored regions on the Ramachandran map. Methods of the other type often use molecular dynamics or Monte Carlo simulations to reconstruct and to refine backbone structure through standard molecular mechanics forcefields. For example, Kazmierkiewicz *et al.*[9] employed geometry criteria of peptide groups and polypeptide chains through the refinement of Monte

*This research work was partially supported by the National Science Council of Taiwan under NSC-95-2221-E-110-102.

Carlo simulations to generate a complete protein backbone.

In this paper, we shall propose a rapid and effective method for generating the full atom protein backbone from the coarse grained model with known α-carbon coordinates. It is based on the homology modeling method to establish a fragment library and to predict atomic coordinates at each residue of protein backbone by the structure similarity. By the experimental results, the prediction accuracy is comparable to the previous methods.

The organization of this paper is as follows. In Section 2, we will introduce two methods for measuring the similarity of protein structures. In Section 3, we will propose the way that we construct our fragment library. Section 4 shows our method for reconstructing protein backbone. Next, Section 5 will show the experimental results of our method. The conclusion of this paper will be given in Section 6.

2 Preliminary

In biological field, *root mean square deviation* (*RMSD*) or *coordinate root mean square deviation* (*CRMSD*) is the method used most frequently to measure the similarity of two protein structures [12, 16]. So far as immediate point of view, the simplest manner to determine the similarity of two protein three-dimensional structures is to superimpose them, which is the idea of RMSD. RMSD is calculated by the average distance between the corresponding pair of residues of two given proteins in 3D space. The small value of RMSD implies the coordinate difference between them is small, which means the two given protein structures are similar. The formula of RMSD is given as follows:

$$RMSD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i^A - x_i^B)^2},$$

where $\sum_{i=1}^{n}(x_i^A - x_i^B)^2$ means the sum of square of distance between the ith pair of residues in proteins A and B, and n represents the length of the proteins. Notice that the lengths of the two proteins must be equal.

Distance RMSD (*DRMSD*) [2, 15], a variation of RMSD, is another way often used for measuring the similarity of two protein structures. It adopts a strategy different from RMSD. The same characteristic as RMSD is that the smaller the DRMSD value is, the more similar the two protein structures are. The formula of DRMSD is given as follows:

$$DRMSD = \frac{1}{n}\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}(d_{ij}^A - d_{ij}^B)^2},$$

where the d_{ij}^A denotes the distance between the ith and jth residues of protein A, d_{ij}^B denotes the distance between the ith and jth residues of protein B, n represents the length of two proteins.

DRMSD would face the problem of *chirality* or *isomer*. Take a common example, the amino acid alanine has two probable forms–S-alanine and R-alanine, which are the mirrored-images, and they are called *chiral enantiomers*, as shown in Figure 1. The DRMSD value between R-alanine and S-alanine is zero. In fact, the structures of these two isomers are not the same, and the mistake does not happen in RMSD.

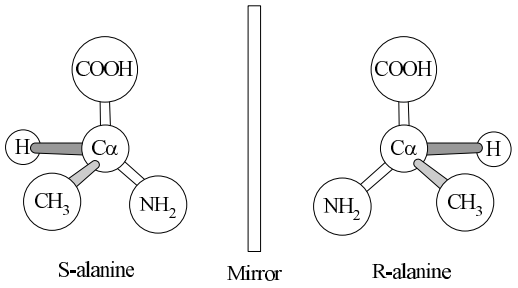

Figure 1: The chiral enantiomers of alanine.

3 Our Fragment Library

3.1 Conformation of Fragment

In our method for solving PBRP, we first create a fragment library extracted from PDB files, where each fragment consists of four successive α-carbons at the protein chain. We represent the local conformation of the fragment with six inner distances which are the distances between all pairs of the four α-carbons. We also record the local coordinates of N, C and O atoms on the backbone, which are relative to the α-carbon at the *center residue* of the fragment. The center residue is defined as the third residue of the fragment, and the center α-carbon is the α-carbon at the center residue. We can represent one fragment diagrammatically as shown in Figure 2, where $d_{i,j}$ denotes the distance between the α-carbons of the ith and jth residues. The shadow circle denotes the center α-carbon. It is believed that if the structures of two fragments are similar, the atomic position distribution of the center residue of the two fragments would be similar. Both Milik *et al.* [14] and Iwata *et al.* [8] proposed the similar strategies. The main difference of our method with the above two methods is that we consider the impact of residue type on atomic coordinate distribution, and we use DRMSD to determine the similarity of fragments.

3.2 Chirality

Unfortunately, owing to the chirality problem, it is not enough to represent the fragment by only six inner distances. As Figure 3 indicates, the fragment is made up

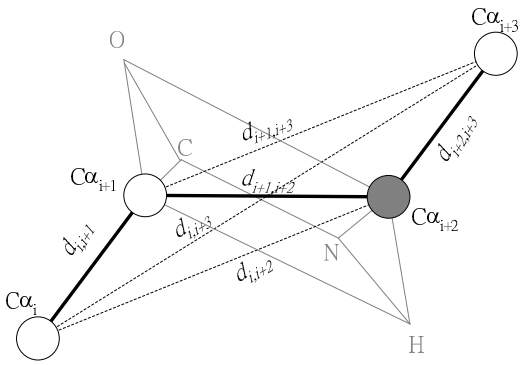

Figure 2: The sketch of one fragment.

of four successive α-carbons, there are three virtual axes connecting every two successive α-carbons. The last α-carbon may be above or under the plane formed by the three preceding α-carbons. This leads to chiral enantiomers. Because the fragment should contain enough information to distinguish chirality, thus Milik *et al.*[14] defined a formula for deciding the chirality as follows:

$$d_{i,i+3} = \chi |v_{i,i+1} + v_{i+1,i+2} + v_{i+2,i+3}|,$$
$$\chi = sign[(v_{i,i+1} \times v_{i+1,i+2}) \cdot v_{i+2,i+3}],$$

where $d_{i,j}$ denotes the distance between the ith and jth α-carbons, and $v_{i,j}$ denotes the 3D coordinate vector from the ith α-carbon to the jth α-carbon. The two different types of chiral enantiomers are determined by the sign of $d_{i,i+3}$. That is, if two fragments are chiral enantiomers to each other, their structural similarity should be extremely low. If they are not chiral enantiomers, we can use DRMSD to calculate their similarity.

3.3 Residue Group

Considering the impact of residue type on atomic coordinate distribution, we process the center residue of a fragment with its two neighboring residues together. Since there are 20 types of amino acids, we assign one unique integer ranging from 1 to 20 to each type of residue. A three-tuple (i, j, k), $1 \leq i, j, k \leq 20$ is used to represent a combination of three consecutive residues. In other words, we have the following combinations: (1,1,1), (1,1,2), (1,1,3) ..., (20,20,20). Accordingly, there are 8000 combinations of three successive residues, and 1 through 8000 are used to represent these combinations. A simple formula [4] can be used to get the combination ranking number as follows:

$$20^2 * (i-1) + 20^1 * (j-1) + 20^0 * (k).$$

For example, the ranking number of (1,1,1) and (20,20,20) are 1 and 8000, respectively. We classify and store the fragments of four consecutive residues extracted from PDB files into 8000 files according to the combination ranking number of their second, third and fourth residues. And we call each file as a *residue group*.

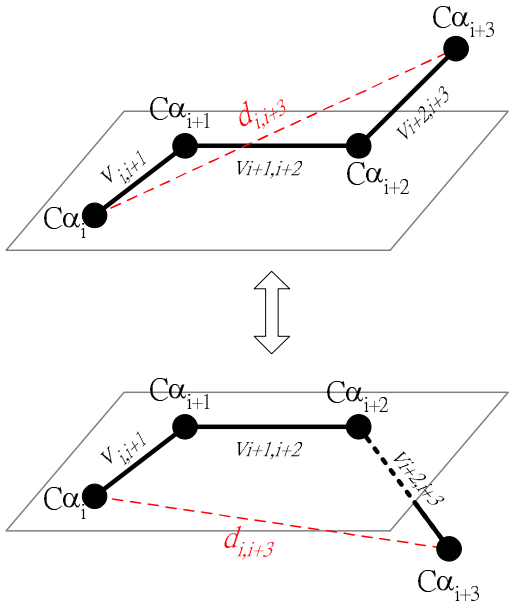

Figure 3: The chirality of one fragment.

3.4 Clustering

There are many four-residue fragments in one residue group. The information involved in each fragment includes the residue group number, the six inner distances with chiral information, and the local coordinates of N, C and O atoms at the center residue. Based on our observation, two fragments with similar conformation have similar distribution on the relative atomic positions to the center residue. Only few of them have obvious discrepancy. Therefore, we cluster similar fragments together and take one typical fragment to represent all other fragments in the cluster. Because the fragments are similar, we can take any fragment as the typical fragment and it would not affect the performance much. The advantage of clustering is not only eliminating unnecessary similar structures to reduce the volume of fragment library, but also could accelerate the execution efficiency of the program.

We use the six inner distances with chiral information to calculate DRMSD and determine the similarity between the fragment structures. Based on empiricism, we cluster the fragments of a group with a very greedy method. We first take randomly a fragment which has not been clustered as the representative of a new cluster. Then, we examine each unexamined fragment f. If the DRMSD between f and some existing cluster c is less than 0.15 Å, then f is put into cluster c. If there exists no such cluster, then f creates another new cluster. After clustering, we set the local coordinates of N, C and O atoms at the center residues as the average local coor-

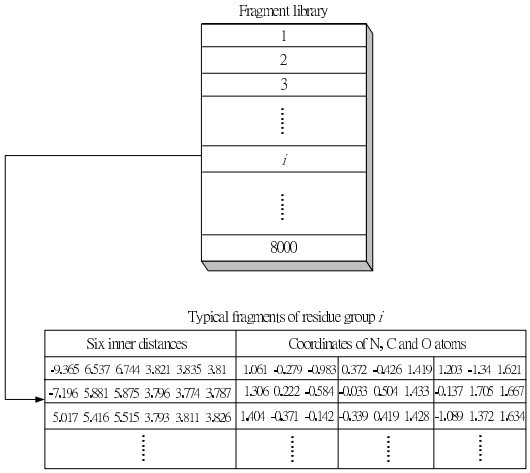

Figure 4: The schema of the fragment library.

dinates of N, C and O atoms at the center residue of all fragments in the same cluster. Figure 4 shows the organization of our fragment library. Each entry of the residue group in the fragment library records the information of a typical fragment.

4 Overview of Our Method

The input data of PBRP is a protein chain with residue sequence and α-carbon coordinates represented in PDB format. We divide the protein chain into sliding windows of four residues. Suppose that the length of input protein chain is L, we obtain $L-3$ fragments. The information associated with each fragment is the six inner distances with chiral information and the group number g decided by its second, third and fourth residues. Table 1 shows an example to illustrate several fragments after the input protein chain is divided.

After the input protein chain is divided into a set of target fragments, for each target fragment f, we search in residue group g in the library to find the most similar typical fragment j in the group. The similarity of two fragments is measured by DRMSD of their six inner distances with chiral information. Then, we rotate j into j' to superimpose the fragment f and we assign the local atomic coordinates of the center residue of j' to the center residue of the target fragment.

However, our method still has some unavoidable shortcomings. The most obvious one is unable to predict the atomic positions at the first two and the last one residues of the protein chain. The problem of the second residue can be solved by using the same method to create fragment library that takes the second residue as the center residue. For the atomic coordinates at the two terminal residues, we calculate their approximate values by a simple heuristic. We append a virtual residue on each terminal of the protein chain. Then all calculation is done similarly.

5 Experimental Results

Our method is implemented on a PC with AMD Athlon^{TM} 1.67 GHZ processor and 512 MB RAM. The operating system is Microsoft Windows XP Professional Version 2002 Service Pack 2.

Our method has been tested on two sets of proteins. The first testing set, containing 32 proteins, is referred to the experimental results of Maupetit *et al.* [13] which was used for comparing several previous works [6, 1, 13]. The second testing set is a subset of CASP7 targets which has 104 proteins originally, but only 94 proteins can be found in PDB, among them 28 proteins are fragmented chains which are not considered here. The protein structures of the 66 proteins are extracted from PDB as our input. We divide the second testing set into two parts based on whether the protein contains nonstandard residues or not. Thus, Table 3 consists of 30 proteins which only contains standard residues, and Table 4 consists of 36 proteins containing nonstandard residues. The predicted results are compared with the real crystallographic structures extracted from PDB by calculating their RMSD. In this paper, we assume that the input is a single protein chain. But in the testing sets, some proteins have several chains. If it is not indicated specifically, we adopt chain A of the protein. Moreover, for fair comparison of various methods, we eliminate the proteins that are already in the testing set from our library. The atoms involved in the RMSD calculation of the main chain are different from the previous works. One of the previous works considered the N, C, O and β-carbon [14], and another considered N, H, α-carbon, β-carbon, C and O atoms [1]. However, in this paper, we consider only N, α-carbon, C and O atoms since we do not predict the β-carbon position on the side chain. Thus, the RMSD values may have a little difference from expected, but it does not influence the whole results with a wide margin.

Table 2 shows the experimental results of our method and previous works on the first testing set. In the table, "♭ ", "♮ "and "♯ "denote the previous methods MaxSprout [6], Adcock's method [1] and SABBAC (proposed by Maupetit *et al.*) [13], respectively. If the RMSD measurement of one testing protein with our method is better than a certain previous method, it is marked with the corresponding symbol. For example, for protein 5NLL, our method outperforms both MaxSprout and Adcock's method, thus its row is marked by "♭ "and "♮ ". In the last two rows, "mean "represents the average of the RMSD values of all testing proteins obtained by the method, the standard deviation is also calculated similarly. In this testing set, the length of protein chain 1VB5B marked

Table 1: An example of fragments obtained by dividing the input protein chain.

Six inner distances						Residue group number
-5.17	5.48	5.28	3.79	3.08	3.81	5000
10.463	6.662	7.312	3.854	3.739	3.872	6886
-8.757	6.821	5.47	3.793	3.799	3.83	7142
-9.673	7.236	6.693	3.828	3.824	3.837	1732
-9.034	6.727	6.228	3.828	3.795	3.796	3277
8.283	5.64	6.931	3.811	3.774	3.784	4934
⋮	⋮	⋮	⋮	⋮	⋮	⋮

with "* "is different from those in previous works due to possible different chain definition. Its length which we measure is 275 residues. The proteins in the first half of Table 2 are often used for testing in many previous literatures [14, 8, 1, 9, 6, 5]. And the latter half is the subset of recent newcomers of PDB, proposed by Maupetit *et al.* [13] . Among the 32 proteins in the first testing set, we have 20 and 15 solutions superior to MaxSprout and SABBAC, respectively. In addition, 2 solutions of our method and SABBAC have equal accuracy. For the average and standard deviation of RMSD values, Adcock's method is not discussed owing to the fewer number of samples. The average of our method is better than MaxSprout, and it is equal to that of SABBAC . Besides, the standard deviation of our method is almost equal to that of SABBAC. Furthermore, we can easily see that the most results of our method on the first half are better than SABBAC. However, we lose to SABBAC in the latter half. We infer that the results are caused by their homologous proteins are few in PDB since they are PDB newcomers. In summary, our method is comparable to the previous works in the first testing set, some of results are better and the others are worse.

In order to verify the performance of our method further, we use the second testing set to compare our method with SABBAC. For SABBAC, we use the online server to obtain the results, and then calculate the RMSD values between their results and the real structures. We do not compare our results with MaxSprout and Adcock's method, since we do not get their software programs. The comparison results are shown in Table 3 and Table 4, the RMSD values of our method marked by underlines are smaller than SABBAC. Among the 30 proteins in Table 3, that contain only standard residues, we have better solutions in 12 proteins and equal solution in 2 proteins. The mean of our method is almost equal to that of SABBAC, and the standard deviation of our method is less than that of SABBAC, which indicates that our method is more stable than SABBAC. Among the 36 proteins in Table 4, that contain nonstandard residues, we have better solutions in only 5 proteins and equal solution in one protein. The mean of our method is worse comparing to SABBAC, but the standard deviation is less than that of SABBAC. According to the observation of the second testing set, though our method might not be superior to SABBAC in the accuracy of reconstructing protein backbone, but the solutions are more stable than SABBAC. Broadly speaking, our method needs further improvement on backbone prediction of the proteins which include nonstandard residues.

For the execution time, because it does not need complicated calculation in the process of finding solutions, it does not spend much time, With rough statistics, our program spends about 3 to 4 seconds for reconstructing the backbone of one protein with 100 residues.

6 Conclusion

In this paper, we propose a method for solving PBRP based on the homology modeling. In addition to the structure similarity, the type of residue is also considered to have large impact on the atomic coordinates. The size of our fragment library is about only 100 MB. For a given target protein, we find out possible candidate fragments with similar structures in the library to reconstruct all atoms on the backbone of the target protein. The experimental results show the reconstruction accuracy of our method is comparable to previous works. And the solution of our method is more stable than previous works in most target proteins.

The future work may include improvement on the prediction accuracy and the execution efficiency. We are also going to set up an online service of our method, which will provide a public software for reconstructing all atoms on the backbone of a protein. Considering the real condition, the drawback of our method is that it cannot deal with all kinds of protein structures, such as heterogens, fragmented chains, and unknown type of residues. We have to investigate and study further to solve these problems by extending our method.

References

[1] S. A. Adcock, "Peptide backbone reconstruction using dead-end elimination and a knowledge-based

Table 2: Comparison of our method with previous works on the first testing set.

Protein	Length	Main chain RMSD (Å)			
		Prior works			
PDB ID	Number of residues	MaxSprout[♭]	Adcock's method[♯]	SABBAC[♯]	Our method
4PTI	58	0.44	0.51	0.53	0.42[♭♯♯]
5CPA	307	-	0.48	0.41	0.34[♯♯]
5NLL	138	0.46	0.42	0.37	0.39[♯]
2CTS	437	0.45	0.37	0.4	0.34[♭♯♯]
1TIM	247	0.6	0.56	0.59	0.54[♭♯♯]
111M	154	0.42	0.31	0.29	0.26[♭♯♯]
1IGD	61	0.44	0.34	0.36	0.36[♭]
1OMD	107	0.41	0.39	0.35	0.39[♭]
2LYM	129	0.44	0.32	0.38	0.29[♭♯♯]
2PCY	99	0.54	0.48	0.42	0.33[♭♯♯]
1CTF	68	0.73	0.41	0.43	0.42[♭♯]
1SEMA	58	0.34	0.5	0.48	0.45[♯♯]
1UBQ	76	0.38	0.37	0.35	0.37[♭]
2MHR	118	0.54	0.33	0.5	0.39[♭♯]
2OZ9	104	0.42	0.24	0.3	0.22[♭♯♯]
PDB newcomers subset					
1PXZA	346	0.54	-	0.55	0.53[♭♯]
1RKIA	101	0.44	-	0.58	0.5 [♯]
1S7LA	177	0.36	-	0.29	0.38
1T70A	255	0.5	-	0.42	0.48[♭]
1TXOA	235	0.38	-	0.41	0.44
1V0ED	666	0.45	-	0.48	0.4 [♭♯]
1V7BA	175	0.41	-	0.3	0.37[♭]
1VB5B	(255/275)*	0.42	-	0.34	0.41[♭]
1VKCA	149	0.33	-	0.28	0.37
1VR4A	103	0.59	-	0.47	0.47[♭]
1VR9A	121	0.45	-	0.42	0.49
1WMHA	83	0.28	-	0.27	0.38
1WPBG	168	0.35	-	0.37	0.43
1WMIA	88	0.42	-	0.41	0.5
1X6JA	88	0.36	-	0.43	0.49
1XB9A	108	0.51	-	0.46	0.53
1XE0B	107	0.62	-	0.61	0.55[♭♯]
Mean		0.45	0.4	0.41	0.41
Standard Deviation		0.09	0.09	0.09	0.08

forcefield," *Journal of Computational Chemistry*, Vol. 25, pp. 16–27, 2004.

[2] A. K. Arakaki, Y. Zhang, and J. Skolnick, "Large-scale assessment of the utility of low-resolution protein structures for biochemical function assignment," *Bioinformatics*, Vol. 20, No. 7, pp. 1087–1096, 2004.

[3] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucleic Acids Research*, Vol. 28, No. 1, pp. 235–242, 2000.

[4] S. Dayalan, S. Bevinakoppa, and H. Schroder, "A dihedral angle database of short sub-sequences for protein structure prediction," *2nd Asia-Pacific Bioinformatics Conference (APBC)*, Dunedin, New Zealand, 2004.

[5] H. J. Feldman and C. W. V. Hogue, "A fast method to sample real protein conformational space," *PROTEINS: Structure, Function, and Genetics*, Vol. 39, pp. 112–131, 2000.

[6] L. Holm and C. Sander, "Database algorithm for generating protein backbone and side-chain co-ordinates from a c alpha trace application to model building and detection of co-ordinate errors," *Journal of Molecular Biology*, Vol. 218, pp. 183–194, 1991.

[7] J. L. Hsin, C. B. Yang, K. S. Huang, and C. N. Yang, "An ant colony optimization approach for the protein side chain packing problem," *Proc. of the 6th WSEAS International Conference on Microelectronics, Nanoelectronics, Optoelectronics*, Istanbul, Turkey, pp. 44–49, 2007.

[8] Y. Iwata, A. Kasuya, and S. Miyamoto, "An efficient method for reconstructing protein backbones from α-carbon coordinates," *Journal of Molecular Graphics and Modelling*, Vol. 21, pp. 119–128, 2002.

[9] R. Kazmierkiewicz, A. Liwo, and H. A. Scheraga, "Energy-based reconstruction of a protein backbone from its α-carbon trace by a Monte-

Table 3: Comparison of our method and SABBAC on the proteins that contain only standard residues of the second testing set.

Protein		Length	Main chain RMSD (Å)	
CASP7 ID	PDB ID	Number of residues	SABBAC	Our method
T0288	2GZV	91	0.62	0.41
T0293	2H00	225	0.52	0.55
T0295	2H1R	271	0.48	0.41
T0305	2H4V	278	0.58	0.4
T0308	2H57	165	0.38	0.4
T0313	2H58	316	0.48	0.38
T0307	2H5N	132	0.45	0.39
T0332	2HA8	159	0.33	0.43
T0318	2HB6	489	0.4	0.42
T0350	2HC5	117	0.4	0.5
T0317	2HCM	159	0.51	0.47
T0345	2HE3	185	0.41	0.49
T0340	2HE4	90	0.34	0.46
T0346	2HE9	172	0.46	0.41
T0335	2HEP	42	0.64	0.63
T0353	2HFQ	85	0.42	0.55
T0349	2HFV	97	0.6	0.71
T0314	2HG6	106	0.43	0.46
T0351	2HG7	60	0.37	0.5
T0327	2HGC	78	0.61	0.49
T0357	2HI6	132	0.45	0.52
T0363	2HJ1	77	0.6	0.44
T0358	2HJJ	66	0.53	0.61
T0372	2HQY	298	0.38	0.41
T0385	2IB0	142	0.25	0.33
T0338	2IVX	256	0.23	0.32
T0377	2IVY	88	0.41	0.41
T0319	2J6A	136	0.29	0.42
T0302	2JM5	134	0.45	0.42
T0334	2OAL	527	0.35	0.31
Mean			0.44	0.45
Standard Deviation			0.11	0.09

Carlo method," *Journal of Computational Chemistry*, Vol. 23, pp. 715–723, 2002.

[10] T. Lezon, J. R. Banavar, and A. Maritan, "Recognition of coarse-grained protein tertiary structure," *PROTEINS: Structure, Function and Bioinformatics*, Vol. 55, pp. 536–547, 2004.

[11] A. Liwo, M. R. Pincus, R. J. Wawak, S. Rackovsky, and H. A. Scheraga, "Calculation of protein backbone geometry from {alpha}-carbon coordinates based on peptide-group dipole alignment," *Protein Science*, Vol. 2, pp. 1697–1714, 1993.

[12] V. N. Maiorov and G. M. Crippen, "Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins," *Journal of Molecular Biology*, Vol. 235, pp. 625–634, 1994.

[13] J. Maupetit, R. Gautier, and P. Tuffery, "SABBAC: online structural alphabet-based protein backbone reconstruction from alpha-carbon trace," *Nucleic Acids Research*, Vol. 34, pp. W147–W151, 2006.

[14] M. Milik, A. Kolinski, and J. Skolnick, "Algorithm for rapid reconstruction of protein backbone from alpha carbon coordinates," *Journal of Computational Chemistry*, Vol. 18, pp. 80–85, 1997.

[15] K. Nishikawa and T. Ooi, "Comparison of homologous tertiary structures of proteins," *Journal of Theoretical Biology*, Vol. 43, pp. 351–374, 1974.

[16] L. S. Reid and J. M. Thornton, "Rebuilding flavodoxin from Cα coordinates: a test study.," *Proteins*, Vol. 5, pp. 170–182, 1989.

Table 4: Comparison of our method and SABBAC on the proteins that contain nonstandard residues of the second testing set.

Protein		Length	Main chain RMSD (Å)	
CASP7 ID	PDB ID	Number of residues	SABBAC	Our method
T0287	2G3V	161	0.36	0.49
T0289	2GU2	307	0.4	0.43
T0315	2GZX	253	0.44	0.44
T0294	2H1S	321	0.43	0.4
T0304	2H28	109	0.45	0.5
T0309	2H4O	62	0.41	0.46
T0323	2H56	218	0.37	0.42
T0312	2H6L	140	0.49	0.42
T0328	2HAG	307	0.41	0.47
T0322	2HBO	142	0.45	0.53
T0306	2HD3	96	0.5	0.56
T0324	2HDO	207	0.36	0.44
T0348	2HF1	61	0.46	0.45
T0283	2HH6	112	0.32	0.35
T0329	2HI0	240	0.43	0.44
T0298	2HJS	334	0.35	0.41
T0369	2HKV	148	0.3	0.39
T0375	2HLZ	296	0.33	0.4
T0376	2HMC	314	0.3	0.4
T0383	2HNG	125	0.32	0.46
T0380	2HQ7	141	0.39	0.47
T0368	2HR2	156	0.26	0.36
T0367	2HSB	126	0.26	0.37
T0297	2HSJ	211	0.35	0.43
T0303	2HSZ	225	0.48	0.4
T0371	2HX1	284	0.39	0.46
T0362	2HX5	144	0.31	0.44
T0360	2HXJ	116	0.39	0.38
T0325	2I5I	261	0.37	0.48
T0342	2I5T	169	0.39	0.4
T0374	2I6C	160	0.39	0.45
T0382	2I9C	121	0.33	0.36
T0381	2IA2	250	0.28	0.45
T0370	2IAB	153	0.39	0.4
T0311	2ICP	87	0.2	0.29
T0354	2ID1	120	0.42	0.51
Mean			0.37	0 .43
Standard Deviation			0.07	0.05