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, 804 Taiwan

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 major atoms (N, C and O atoms) on the backbone. In this paper, we first build a 4-residue fragment library extracted from PDB. Then, to solve PBRP, we search for fragments with similar structures based on the inner distances. To test the performance of our method, we use two testing sets of target proteins, one was proposed by Maupetit *et al.* 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. These previous works contain complicated energy computation, while our method does not. Thus, our method requires much less execution time than the previous works.

Keywords: bioinformatics, protein structure, backbone reconstruction, α-carbon, RMSD

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. Many applications need the information of all atom coordinates. Modeling from atomic coordinates and all-atom protein reconstruction have been extensively studied, including improving the low resolution models from crystallography and *ab initio* (or *de novo*) folding computation, and 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 a method for side chain positioning. In this paper, we shall discuss only the prediction of all major atom coordinates on the protein backbone. For a given amino acid sequence with the α-carbon 3D coordinates, the *all-atom protein backbone reconstruction problem* (PBRP) is to rebuild the 3D coordinates of all major atoms (N, C and O atoms) on the backbone. Note that PBRP does not involve the prediction 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, 8, 13, 14] and minimization

Received April 19, 2008; revised May 19, 2009; accepted July 10, 2009.
Communicated by Nancy M. Amato.

[*] This research work was partially supported by the National Science Council of Taiwan under Contract NSC-95-2221-E-110-102. A preliminary version of this paper was presented at the National Computer Symposium, Taichung, Taiwan, Dec. 20-21, 2007.

[+] Corresponding author: cbyang@cse.nsysu.edu.tw.

of local molecular energy [9, 11]. There are also some approaches combining the above two types of methods. A method of the first type 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 backbone. Iwata *et al.* [8] designed a method to analyze and to select fragments compatible with favored regions on the Ramachandran map. A method of the second type often uses 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 Carlo simulations to generate a complete protein backbone.

In this paper, we shall propose a rapid and effective method for generating the all-atom backbone from the coarse grained model with known α-carbon coordinates. Our method is based on the homology modeling method to establish a fragment library and to predict all major 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 the field of structure biology, *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]. The simplest manner to determine the similarity of two three-dimensional protein structures is to superimpose them, which is the idea of RMSD. The formula of RMSD between two given proteins in 3D space is given as follows.

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

where $x_i^A - x_i^B$ means the distance between the ith pair of corresponding atoms in proteins A and B, and n represents the number of pairs to be calculated. The small value of RMSD implies the coordinate difference between them is small, which means the two given structures are similar. Usually, the coordinate of α-carbon is used to represent the 3D position of one residue in one protein. When RMSD is used to measure the structure similarity of two proteins on the residue level, the two input sequences are their α-carbon coordinates. Note that the two input sequences should be of the same length when RMSD is applied.

Distance RMSD (*DRMSD*) [2, 15], a variation of RMSD, is another way often used for measuring the similarity of two protein structures. 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} (d_{ij}^{B}) denotes the distance between the ith and jth atoms of protein $A(B)$. Fig. 1 shows the concept diagrammatically. The same property also holds that the smaller the DRMSD value is, the more similar the two protein structures are.

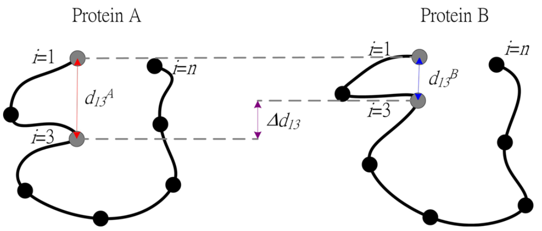


Fig. 1. The concept of DRMSD calculation.

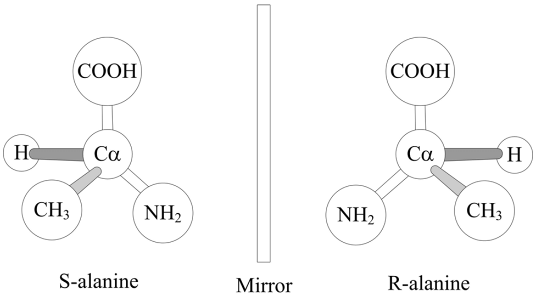


Fig. 2. The chiral enantiomers of alanine. Their DRMSD is equal to 0, though their structures are different.

DRMSD would face the problem of *chirality* or *isomer*. Fig. 2 shows a common example that the amino acid alanine has two probable forms, the S-alanine and the R-alanine, which are the mirrored-images, and they are called *chiral enantiomers*. The DRMSD value between R-alanine and S-alanine is zero. But in fact, the structures of these two isomers are not the same, and the mistake does not happen in RMSD.

3. THE ESTABLISHMENT OF 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 of the protein chain. The extracted information includes the coordinates of α-carbon and N, C, O on the backbone. We represent the local conformation of one 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 relative to the *center 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 means the α-carbon at the center residue. Fig. 3 shows one fragment diagrammatically, where $d_{i,j}$ denotes the distance between the α-carbons of the ith and jth residues. The shadowed circle denotes the center α-carbon. Some previous methods, pro posed by Milik *et al.* [14], Gront *et al.* [5] and Iwata *et al.* [8], concluded that if the structures of two fragments are similar, their atomic position distribution of the center residues would be similar. We will apply this conclusion to our method. The main difference between our method and the previous methods is that we consider the impact of residue type on atomic coordinate distribution, and we use DRMSD to measure the fragment similarity.

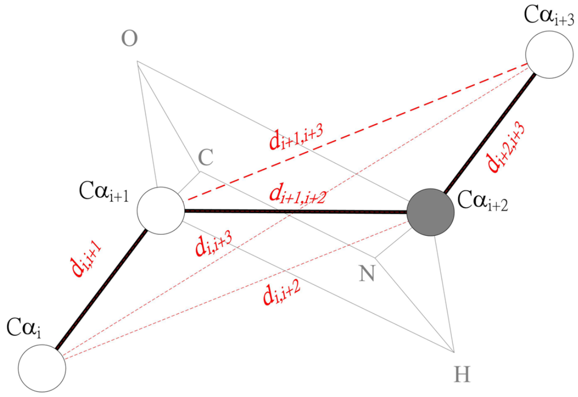


Fig. 3. The sketch of one fragment (four consecutive residues) with its six inner distances.

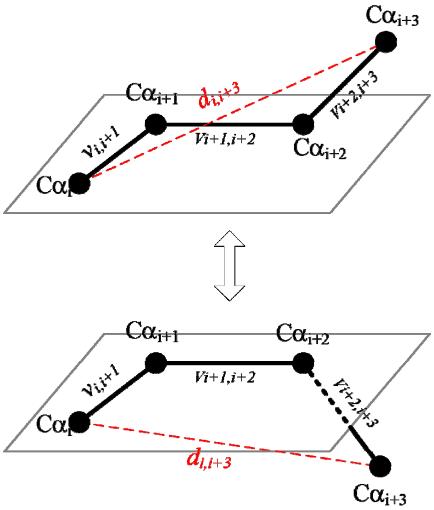


Fig. 4. The chirality of one fragment.

3.2 Chirality and Similarity

Unfortunately, owing to the chirality problem, it is not enough to represent the fragment with only six inner distances. As Fig. 4 indicates, one fragment is made up of four successive α-carbons, and there are three virtual axes connecting every two successive α-carbons. The last α-carbon may be either 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:

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

where $d_{i,j}$ denotes the distance between the ith and jth α-carbons, and $v_{i,j}$ denotes the 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 A and B are chiral enantiomers to each other, their structural similarity should be extremely low. We need not calculate the DRMSD and assign their similarity to 0 if A and B are of opposite sign. If they are not chiral enantiomers, we define their similarity function as the inverse of DRMSD of A and B, where DRMSD is calculated as follows,

$$d(A,B) = \frac{1}{4}\sqrt{\sum_{i=1}^{4}\sum_{j=1}^{4}(d_{ij}^{A} - d_{ij}^{B})^2}.$$

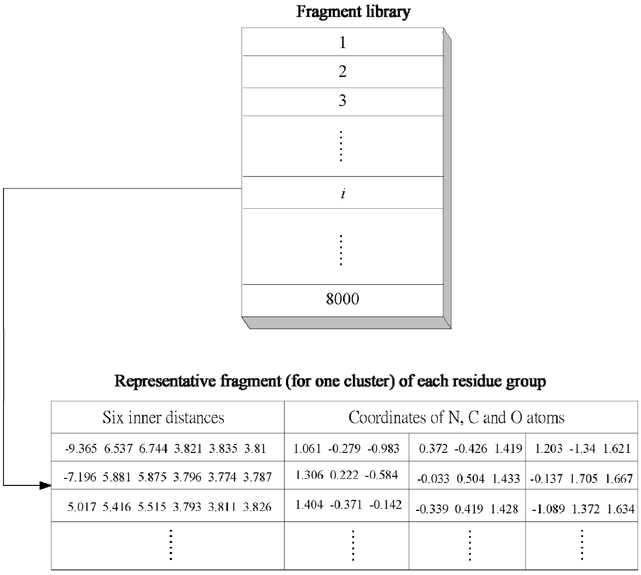
3.3 Residue Group and Clustering

Since one fragment consists of 4 consecutive amino acids and there are 20 kinds of standard amino acids, there are totally 20^4 = 160,000 distinct fragments. This amount is very huge and some of the fragments may contain too few samples extracted from PDB. Thus, we define a *residue group* as the combination of the second, third and fourth residues of one fragment. Then, the total amount of distinct residue groups in our fragment library is 20^3 = 8000. Note that there are many four-residue fragments in one residue group.

The information involved in each fragment includes the residue group ID number, the six inner distances with chiral information, and the local coordinates of N, C and O atoms at the center (third) residue. Based on previous observation [5, 8, 14], two fragments with similar conformation have similar distribution on the relative atomic positions in the center residue. Only few of them have obvious discrepancy. Therefore, we cluster similar fragments in a residue group together and take one representative fragment to represent all other fragments in the cluster. Because the fragments are similar, we can take any fragment as the representative 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 of the program.

We cluster the fragments of a residue group with a very greedy method based on the similarity. First, we randomly take a fragment as the representative fragment of the first cluster. Then, for every other fragment *f*, we calculate the DRMSD with the representative fragment of each existing cluster. If there exists a cluster *c* with DRMSD less than 0.15 Å, then *f* is put into cluster *c*. Otherwise, we create a new cluster with *f* as the representative fragment.

After clustering, we set the local coordinates of N, C and O atoms at the center residue of the representative fragment as the average local coordinates of N, C and O atoms at the center residues of all fragments in the same cluster. Fig. 5 shows an example of the organization of our fragment library, where we attach the chirality sign of the fragment to the first distance. Each entry of the residue group in the fragment library records the information of the representative fragment of one cluster. Totally we have 1,127,090 clusters (representative fragments) in the 8000 residue groups and the distribution histogram of the numbers of clusters in all residue groups is shown in Fig. 6. For example, there are 850 groups that have 80 to 99 clusters.


Fragment library
1
2
3
⋮
i
⋮
8000
Representative fragment (for one cluster) of each residue group

Six inner distances	Coordinates of N, C and O atoms		
-9.365 6.537 6.744 3.821 3.835 3.81	1.061 -0.279 -0.983	0.372 -0.426 1.419	1.203 -1.34 1.621
-7.196 5.881 5.875 3.796 3.774 3.787	1.306 0.222 -0.584	-0.033 0.504 1.433	-0.137 1.705 1.667
5.017 5.416 5.515 3.793 3.811 3.826	1.404 -0.371 -0.142	-0.339 0.419 1.428	-1.089 1.372 1.634
⋮	⋮	⋮	⋮

Fig. 5. The schema of the fragment library. We first classify the fragments into 8000 3-residue groups and then in each group we do further clustering according to the similarity function.

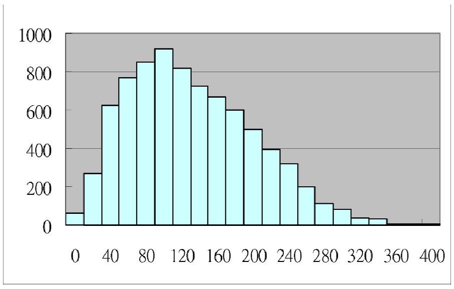


Fig. 6. The distribution histogram of the numbers of clusters in 8000 residue groups.

4. THE RECONSTRUCTION METHOD OF ALL-ATOM BACKBONE

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 $L - 3$ target fragments. The information associated with each target fragment is the six inner distances with chiral information and the residue group ID number g which is decided by its second, third and fourth residues. Table 1 shows an example to illustrate a series of target fragments after the input protein chain is divided.

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

Six Inner Distances						Reside 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	3.786	4934
⋮	⋮	⋮	⋮	⋮	⋮	⋮

After the input protein chain is divided into $L - 3$ target fragments, for each target fragment k, we search for the most similar representative fragment f in residue group g. Then, we rotate f into f' to superimpose onto the fragment k and we assign the local atomic coordinates of the center residue of f' to the center residue of the target fragment k.

However, our method still has some unavoidable shortcomings. The most obvious one is that we cannot accurately 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 build another new 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 with a simple heuristic. We append a virtual residue to each terminal of the protein chain, where the virtual residue is the same as the terminal residue. Then all calculations are done similarly.

5. EXPERIMENTAL RESULTS

Our reconstruction 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 data 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 [1, 6, 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 remaining 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 contains only standard residues, and Table 4 consists of 36 proteins containing standard and 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, C, O, α-carbon, and β-carbon atoms [1]. However, in this paper, we consider only N, C, O and α-carbon 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 with "*" is different from those in previous works due to possible difference in chain definition. Its length which we measure is 275 residues. The proteins in the first half of Table 2 were often used for testing in many previous literatures [1, 4, 6, 8, 9, 14]. And the latter half is a 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 most of the results of our method on the first half are better than SABBAC. However, we lose to SABBAC in the latter half. One of the possible reasons is that these PDB newcomers may have few proteins with similar structures in PDB.

As Table 2 shows, our method is comparable to the previous works, especially SABBAC, in the first testing set. Our method only perform searching in our fragment library, it does not involve any complicated energy computation. However, SABBAC refines the initial predicted result by performing energy computation. Thus, the execution time of our method is much less than that of SABBAC. With rough statistics, our program spends about 3 to 4 seconds for reconstructing the backbone of one protein with 100 residues.

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$^{\flat}$	Adcock's method$^{\maltese}$	SABBAC$^{\sharp}$	Our method
4PTI	58	0.44	0.51	0.53	$0.42^{\flat\maltese\sharp}$
5CPA	307	-	0.48	0.41	$0.34^{\maltese\sharp}$
5NLL	138	0.46	0.42	0.37	$0.39^{\flat\maltese}$
2CTS	437	0.45	0.37	0.4	$0.34^{\flat\maltese\sharp}$
1TIM	247	0.6	0.56	0.59	$0.54^{\flat\maltese\sharp}$
111M	154	0.42	0.31	0.29	$0.26^{\flat\maltese\sharp}$
1IGD	61	0.44	0.34	0.36	$0.36^{\flat}$
1OMD	107	0.41	0.39	0.35	$0.39^{\flat}$
2LYM	129	0.44	0.32	0.38	$0.29^{\flat\maltese\sharp}$
2PCY	99	0.54	0.48	0.42	$0.33^{\flat\maltese\sharp}$
1CTF	68	0.73	0.41	0.43	$0.42^{\flat\sharp}$
1SEMA	58	0.34	0.5	0.48	$0.45^{\maltese\sharp}$
1UBQ	76	0.38	0.37	0.35	$0.37^{\flat}$
2MHR	118	0.54	0.33	0.5	$0.39^{\flat\sharp}$
2OZ9	104	0.42	0.24	0.3	$0.22^{\flat\maltese\sharp}$
PDB newcomers subset					
1PXZA	346	0.54	-	0.55	$0.53^{\flat\sharp}$
1RKIA	101	0.44	-	0.58	$0.5^{\sharp}$
1S7LA	177	0.36	-	0.29	0.38
1T70A	255	0.5	-	0.42	$0.48^{\flat}$
1TXOA	235	0.38	-	0.41	0.44
1V0ED	666	0.45	-	0.48	$0.4^{\flat\sharp}$
1V7BA	175	0.41	-	0.3	$0.37^{\flat}$
1VB5B	(255/275)*	0.42	-	0.34	$0.41^{\flat}$
1VKCA	149	0.33	-	0.28	0.37
1VR4A	103	0.59	-	0.47	$0.47^{\flat}$
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	$0.55^{\flat\sharp}$
Mean		0.45	0.4	0.41	0.41
Standard Deviation		0.09	0.09	0.09	0.08

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 Tables 3 and 4, where 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 solutions 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. Again, we can see that the predicted accuracy of our method is comparable to that of SABBAC, though our method does not perform any energy computation.

Among the 36 proteins in Table 4 that contain standard and nonstandard residues, we have better solutions in only 5 proteins and equal solutions in one protein. The mean

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

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. There are two reasons that our method does not outperform SABBAC. First, our fragment library does not contain any information of nonstandard residues, so we use the average coordinates of the 20 possible fragments as the coordinates of the nonstandard residue. Second, SABBAC achieves the high accuracy with the complicated computation of energy functions.

6. CONCLUSION

In this paper, we propose a method for solving PBRP based on the homology modeling. In addition to the structure similarity, the residue type is also considered to have large impact on the atomic coordinates. The size of our fragment library is about 100M bytes. For a given target protein, we find out possible candidate fragments with similar structures in the library to reconstruct all major atoms (N, C and O) on the backbone of the

Table 4. Comparison of our method and SABBAC on the proteins that contain non-standard 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

target protein. The experimental results show the reconstruction accuracy of our method is comparable to previous works which involves complicated computation of energy functions. 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. Considering the real condition, the drawbcak of our method is that it cannot deal with all kinds of protein structures, such as heterogens, fragmented chains, and unknown type of residues. These problems are still big challenges in the future.

REFERENCES

1. S. A. Adcock, "Peptide backbone reconstruction using dead-end elimination and a knowledge-based forcefield," *Journal of Computational Chemistry*, Vol. 25, 2004,

pp. 16-27.

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, 2004, pp. 1087-1096.
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, 2000, pp. 235-242.
4. H. J. Feldman and C. W. V. Hogue, "A fast method to sample real protein conformational space," *Proteins: Structure, Function, and Genetics*, Vol. 39, 2000, pp. 112-131.
5. D. Gront, S. Kmiecik, and A. Kolinski, "Backbone building from quadrilaterals: A fast and accurate algorithm for protein backbone reconstruction from alpha carbon coordinates," *Journal of Computational Chemistry*, Vol. 28, 2007, pp. 1593-1597.
6. L. Holm and C. Sander, "Database algorithm for generating protein backbone and side-chain coordinates from a c^{α} trace: Application to model building and detection of coordinate errors," *Journal of Molecular Biology*, Vol. 218, 1991, pp. 183-194.
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," in *Proceedings of the 6th WSEAS International Conference on Microelectronics, Nanoelectronics, Optoelectronics*, 2007, pp. 44-49.
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, 2002, pp. 119-128.
9. R. Kazmierkiewicz, A. Liwo, and H. A. Scheraga, "Energy-based reconstruction of a protein backbone from its α-carbon trace by a Monte-Carlo method," *Journal of Computational Chemistry*, Vol. 23, 2002, pp. 715-723.
10. T. Lezon, J. R. Banavar, and A. Maritan, "Recognition of coarse-grained protein tertiary structure," *PROTEINS: Structure, Function and Bioinformatics*, Vol. 55, 2004, pp. 536-547.
11. A. Liwo, M. R. Pincus, R. J. Wawak, S. Rackovsky, and H. A. Scheraga, "Calculation of protein backbone geometry from α-carbon coordinates based on peptide-group dipole alignment," *Protein Science*, Vol. 2, 1993, pp. 1697-1714.
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, 1994, pp. 625-634.
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, 2006, pp. W147-W151.
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, 1997, pp. 80-85.
15. K. Nishikawa and T. Ooi, "Comparison of homologous tertiary structures of proteins," *Journal of Theoretical Biology*, Vol. 43, 1974, pp. 351-374.
16. L. S. Reid and J. M. Thornton, "Rebuilding flavodoxin from Cα coordinates: a test study," *Proteins*, Vol. 5, 1989, pp. 170-182.

Jen-Hui Wang (王仁暉) received the B.S. degree in Computer Science and Engineering from Yuan Ze University, Taoyuan, Taiwan, in 2005, and the M.S. degree in Computer Science and Engineering from National Sun Yat-sen University, Kaohsiung, Taiwan, in 2007. He is now working in HTC Corporation. His research interests are bioinformatics and computer algorithms.

Chang-Biau Yang (楊昌彪) received the B.S. degree in Electronic Engineering from National Chiao Tung University, Hsinchu, Taiwan, in 1982, and the M.S. degree in Computer Science from National Tsing Hua University, Hsinchu, Taiwan, in 1984. Then, he received the Ph.D. degree in Computer Science from National Tsing Hua University in 1988. He is currently a professor in the Department of Computer Science and Engineering, National Sun Yat-sen University. His research interests include computer algorithms, interconnection networks, and bioinformatics.

Chiou-Ting Tseng (曾球庭) received the B.S. degree and M.S. degree in Computer Science and Engineering from National Sun Yat-sen University, Kaohsiung, Taiwan, in 2003 and 2006, respectively. He is currently a Ph.D. candidate of the Department of Computer Science and Engineering at the National Sun Yat-sen University. His research interests include computer algorithms, bioinformatics and sequence analysis.