An Effective Tool Preference Selection Method for Protein Structure Prediction with SVM [*]

Hsin-Wei Yen, Chang-Biau Yang[1], and Hsing-Yen Ann
Department of Computer Science and Engineering
National Sun Yat-sen University, Kaohsiung, Taiwan 80424
[1]Corresponding author: cbyang@cse.nsysu.edu.tw

Abstract

Prediction of protein structure is a problem of great interest in bioinformatics. Many studies have been devoted to this issue, such as Adcock's method, MaxSprout, SABBAC and Chang's method. In this paper, we combine the power of two outstanding tools, SABBAC and Chang's method. Based on SVM, we propose a tool preference classification method for determining which tool is potentially the better one used to predict the structure of a target protein. We design a heuristic method to select the better feature set combination for SVM. We test our method on the proteins with standard amino acids in CASP7 dataset, which contains 30 protein sequences. The experiment results show that our method has 9.19% and 5.37% RMSD improvement against SABBAC and Chang's result, respectively. Our method can also be applied to other effective prediction methods developed in the future.

1 Introduction

A protein consists of a linear chain of amino acids which include 20 standard amino acids and some nonstandard amino acids. Each protein has its own functions, and its linear chain has to be folded correctly to reveal its corresponding functions. This is why the analysis of the 3D conformations of proteins is so important. People usually use X-ray crystallography and NMR (nuclear magnetic resonance) [10] to determine the structures of proteins, but both of them are time-consuming and costly. To make structure analysis easier, researchers try to determine the 3D-structures from protein sequences with computers, and then only use X-ray crystallography or NMR to reveal the details which they are really interested in.

There are two main methods to predict the positions of atoms in a protein sequence, *homology modeling* and *ab initio*. The homology modeling method uses the known structures to construct a template database, and use the geometry similarity of protein fragments to determine the positions of atoms. On the other hand, ab initio does not need to collect the existing information of proteins. Some ab initio methods base on the force field between molecules, which consider the energy between atoms in the amino acid, such as hydrogen bonds, van der Waals force, electrostatic force, etc. The conformation with minimum energy is considered as the proper positions of atoms, because the structure is the most stable when the energy is minimized.

In this paper, we consider the all-atom *protein backbone reconstruction problem* (PBRP). Given the 3D coordinates of the α-carbons (C_α) in protein and its sequence, the PBRP is to rebuild the 3D coordinates of all major atoms, including N, C and O, on the backbone. Wang [11] proposed an algorithm to reconstruct the atoms of protein backbone with the homology modeling method. However, the prediction accuracy of oxygen (O) atoms in amino acid is obviously lower other atoms. Later, Chang [2] proposed a method to further refine the O atoms based on Wang's result. SABBAC [9] is another software tool for predicting the all-atom position. Chang's method and SABBAC utilize different template knowledge, so neither one of them dominates the other in all protein sequence prediction.

In this paper, we propose a preference classification strategy to determine which tool, either Chang's method or SABBAC, is a better predictor for the given protein sequence. For a given protein sequence, if we can choose the most suitable tool for predicting, we can reduce the RMSD (root

[*] This research work was partially supported by the National Science Council of Taiwan under contract NSC-98-2221-E-110-062 .

mean square deviation) of the predicted structure. The rest of this paper is organized as follows. In Section 2, we review some results of the previous works. Our method which uses SVM (support vector machine) to choose the suitable software tool is given in Section 3. In Section 4, we will illustrate the experiment results. Finally, we will give a conclusion in Section 5.

2 Previous Results

SABBAC [9] is one of famous methods for solving the all-atom PBRP, and it provides online service on the Internet. They used *hidden Markov model* (HMM) to derive structure alphabet. The experimental result shows that SABBAC is more accurate than the other previous research. However, if the protein size (protein length) is large, the required calculation time may increase up to more than ten minutes.

Wang *et al.* [11] proposed another method which is based on the homology modeling to solve all-atom PBRP. They use the protein information in PDB to build the fragment library. All consecutive four-residue fragments from the structures of all proteins in PDB are extracted. Each fragment is identified by its second, third and fourth residues. Therefore, the fragments are classified into 8000 residue groups. The fragments of a residue are further clustered into several clusters based on their structures, including the six inner distances between C_α atoms of the four amino acids. This clustering strategy can greatly reduce the required time of prediction. Even for a large protein with length several hundreds, the calculation can be done in about ten or fewer seconds.

Chang *et al.* [2] considered the potential energy to improve Wang's method. They analyzed Wang's results and found that the RMSD of O atoms is much higher than the other two atoms, N and C. Based on the AMBER force field [3], Chang *et al.* defined a simplified potential energy function and proposed a method to refine the O atoms' positions from Wang's results. Besides, to reduce the required time for searching the best position, they proposed a *two-phase refinement method.* In the first phase, the searching domain is divided into coarse grid points. After some better candidates have been found, these coarse grid points are further divided into finer grid points. As a result, the accuracy of the prediction on O atoms can be improved. And the time used is in the same order as Wang's method even if the protein size is large.

3 The Preference Classification Method

For a given protein sequence and its coordinates of α-carbon atoms, we propose a preference classification method to determine whether Chang's method or SABBAC is better for solving the PBRP. Our classification method is based on the SVM method. Figure 1 illustrates the flow chart of our method. The input of our method is the 3D coordinates of C_α in proteins and the sequences composed by 20 amino acids, and the output is the 3D coordinates of all main backbone atoms (N, C, O), predicted by either SABBAC or Chang's method.

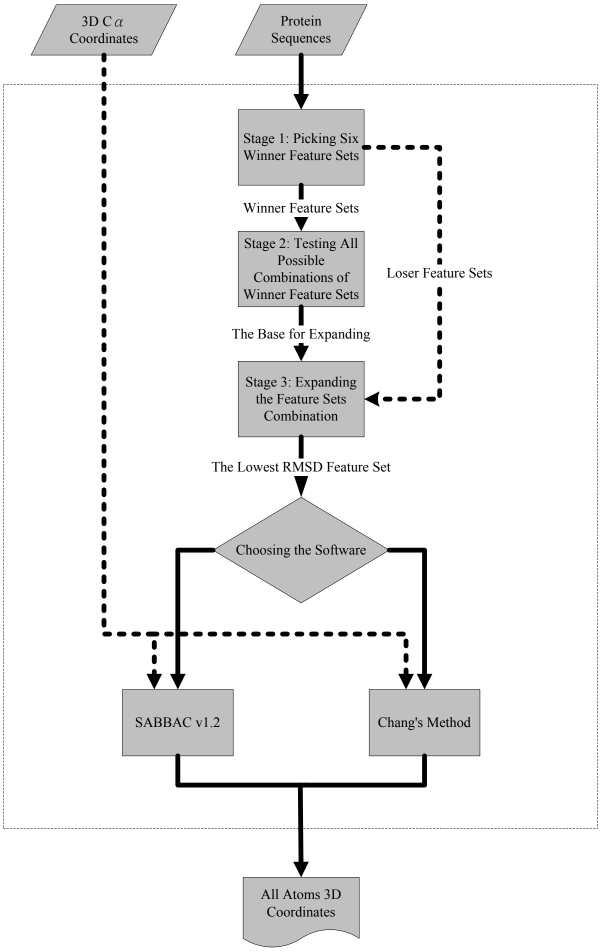
3D C α Coordinates
Protein Sequences
Stage 1: Picking Six Winner Feature Sets
Winner Feature Sets
Stage 2: Testing All Possible Combinations of Winner Feature Sets
Loser Feature Sets
The Base for Expanding
Stage 3: Expanding the Feature Sets Combination
The Lowest RMSD Feature Set
Choosing the Software
SABBAC v1.2
Chang's Method
All Atoms 3D Coordinates

Figure 1: The flow chart of the preference classification method.

3.1 The Features of SVM

SVM is a powerful and high-performance tool for classification problems. Many packages have been developed, and LIBSVM [1] is one of the packages with high classification accuracy. We use the grid search for tuning the parameters of SVM which achieve the best performance and avoid over-fitting or under-fitting. And then we import the test dataset to compute the accuracy.

Several feature sets, C, H, V, P and Z, are usually used to solve proteins classification problems [5, 6, 7]. C represents the *composition* of 20 amino acids, H represents the *Hydrophobicibility* , V represents the *Normalized van der Walls volume*, P represents the *Polarity*, and Z represents the *Polarizability.* In previous studies, these feature sets benefit the protein fold recognition. However, obtained from our primitive experiments, these feature sets cannot work well in the preference classification.

Each feature in C is composition percentage of one amino acid, so there are 20 features in C. In each of the other feature sets (HVPZ), 20 amino acids are divided into three groups by their properties, as shown in the first four rows of Table 1.

To improve accuracy of preference classification, we add three feature sets E, N and A [8], which represent *size, charge*, and *aliphaticity or aromaticity.* Besides, in the feature selection process, the features in one set are considered to be independent. We separate one original feature set into three parts with coding rules except C, because there is only one coding rule in C. The features in one original set are regrouped according to the rule of *composition, transition,* and *distribution.* For example, H is separated into three parts: H, I, J; P is separated into P, Q, R. Finally, we have 22 feature sets. Table 2 shows all new feature sets and the number of features in each set.

Here is an example, shown in Figure 2, to explain the coding scheme of 3, 3 and 15 features in P, Q and R, respectively. The 3 feature values in P are 0.4, 0.28 and 0.32. The first is 0.4 (10/25) because 10 amino acids are contained in Group 1 and the sequence length is 25. Similarly, we get 0.28 (7/25) and 0.32 (8/25) for Group 2 and Group 3, respectively. The set Q includes 0.42, 0.16 and 0.25. 10 ”1-to-2” or ”2-to-1” transitions are represented by 0.42 (10/24) and there are 24 transitions in a sequence consisting of 25 amino acids. So the next 2 values are 0.17 (4/24) for Group 2 and 0.25 (6/24) for Group 3. In the set R, the first five values are derived from the locations of the first, 1/4, 1/2, 3/4, and the last amino acid in Group 1 over the length of the protein sequence. Thus, we get 0.04 (1/25), 0.12 (3/25), 0.56 (14/25), 0.76 (19/25), 1 (25/25) for Group 1, 0.08, 0.32, 0.52, 0.6, 0.96 for Group 2, and 0.16, 0.24, 0.36, 0.8, 0.88 for Group 3.

Table 1: Three groups of 20 amino acids in each coding scheme.

Coding Scheme	Group 1	Group 2	Group 3
Hydrophobicity (H)	R, K, E, D, Q, N	G, A, S, T, P, H, Y	C, V, L, I, M, F, W
Polarity (P)	L, I, F, W, C, M, V, Y	P, A, T, G, S	H, Q, R, K, N, E, D
Normalized van der Waals volume (V)	G, A, S, C, T, P, D	N, V, E, Q, I, L	M, H, K, F, R, Y, W
Polarizability (Z)	G, A, S, D, T	C, P, N, V, E, Q, I, L	K, M, H, F, R, Y, W
Size (E)	A, G, C, S	P, N, D, T, V	the others
Charge (N)	D, E	K, R, H	the others
Aliphatic or Aromatic (A)	V, I, L	Y, H, W, F	the others

Table 2: 22 feature sets with their sizes.

Original Set	Composition	Transition	Distribution
C (20)	C(20)		
H (21)	H (3)	I (3)	J (15)
V (21)	T (3)	U (3)	V (15)
P (21)	P (3)	Q (3)	R (15)
Z (21)	X (3)	Y (3)	Z (15)
E (21)	E (3)	F (3)	G (15)
N (21)	L (3)	M (3)	N (15)
A (21)	A (3)	B (3)	D (15)

```
                        5    10    15    20    25
Sequnce:            VSLNF KDPEA VRALT CTLLR EDFGL
Group  :            12131 33232 13212 12113 33121
Number of Group 1:  1 2 3       4  5  6 78    9 10
Number of Group 2:   1      2 3   4 5  6       7
Number of Group 3:     1  23 4   5        6 78
                    12131 33232 13212 12113 33121
 1-2/2-1 transitions:^^        ^   ^^^ ^^      ^^
 2-3/3-2 transitions:       ^^^   ^
 1-3/3-1 transitions:  ^^^       ^        ^   ^
```

Figure 2: A coding example of a sequence.

3.2 Our Classification Method

We propose a three-stage method to search for the effective feature sets for solving the preference classification problem. In Stage 1, we pick ρ winner feature sets by checking the accuracy of each combination in jackknife test, where the feature set combinations are the elements of the power set of the 3 sets derived from the same original feature set. For example, we divide P into P, Q and R, so 7 combinations P, Q, R, PQ, QR, PQ and PQR are tested in Stage 1. Totally 50 combinations, as shown in Table 3, are examined. Since

Table 3: 50 feature set combinations examined in Stage 1.

C (20)		
HIJ (21)		
HI (6)	IJ (18)	HJ (18)
H (3)	I (3)	J (15)
TUV (21)		
TU (6)	UV (18)	TV (18)
T (3)	U (3)	V (15)
PQR (21)		
PQ (6)	QR (18)	PQ (18)
P (3)	Q (3)	R (15)
XYZ (21)		
XY (6)	YZ (18)	XZ (18)
X (3)	Y (3)	Z (15)
EFG (21)		
EF (6)	FG (18)	EG (18)
E (3)	F (3)	G (15)
LMN (21)		
LM (6)	MN (18)	LN (18)
L (3)	M (3)	N (15)
ABD (21)		
AB (6)	BD (18)	AD (18)
A (3)	B (3)	D (15)

these 50 combinations are from 22 feature sets, if one combination has high rank in the accuracy, the feature sets composed in the combination are also with high rank. For example, if the accuracies of the combinations are PQ, C, AB, BD, HI, $\cdots$, we will keep the top ρ *winner feature sets* P, Q, C, A, B, D, H, I $\cdots$ as the result of Stage 1. And the other $22 - \rho$ feature sets, called *loser feature sets*, will be used in Stage 3. In our experiments, we set $\rho = 6$.

Then, we test all possible combinations of the ρ winner feature sets in Stage 2. $2^{\rho} - 1$ combinations are examined in this stage. Improving accuracy is usually the main goal in classification problems; however, the aim of our preference classification is to decrease the RMSD of the predicted results with respect to real protein structures. So the classification accuracy is not the only criterion that we have to consider. The high accuracy does not always lead to the low average RMSD because the differences of RMSD in the two software tools for different proteins are not the same. Table 4 shows the RMSD of two software tools. Sometime, we correctly choose the software for several proteins but miss the proteins with large RMSD difference, the average RMSD will increase by the missed ones. High accuracy might result in low average RMSD, so we find the combinations with accuracy higher than 70% as candidates in Stage 2. We calculate the average RMSD of each candidate, and then rank these combinations by average RMSD. The combination with the lowest average RMSD is the base for Stage 3.

Table 4: The RMSD comparison of SABBAC and Chang's method on the proteins with standard amino acids in CASP7. μ denotes the average of RMSD. A value marked with "●" means the better result.

CASP7 ID	PDB ID	SABBAC v1.2	Chang's method
T0288	2GZV	0.621	0.383 ●
T0293	2H00	0.520 ●	0.529
T0295	2H1R	0.481	0.374 ●
T0305	2H4V	0.578	0.377 ●
T0308	2H57	0.383	0.359 ●
T0313	2H58	0.479	0.354 ●
T0307	2H5N	0.450	0.383 ●
T0332	2HA8	0.330 ●	0.390
T0318	2HB6	0.397	0.377 ●
T0350	2HC5	0.402 ●	0.468
T0317	2HCM	0.511	0.439 ●
T0345	2HE3	0.408 ●	0.445
T0340	2HE4	0.343 ●	0.443
T0346	2HE9	0.457	0.381 ●
T0335	2HEP	0.636 ●	0.657
T0353	2HFQ	0.423 ●	0.538
T0349	2HFV	0.603 ●	0.684
T0314	2HG6	0.432	0.407 ●
T0351	2HG7	0.373 ●	0.465
T0327	2HGC	0.611	0.494 ●
T0357	2HI6	0.451 ●	0.476
T0363	2HJ1	0.604	0.423 ●
T0358	2HJJ	0.534 ●	0.564
T0372	2HQY	0.377 ●	0.378
T0385	2IB0	0.252 ●	0.305
T0338	2IVX	0.231 ●	0.305
T0377	2IVY	0.415	0.375 ●
T0319	2J6A	0.293 ●	0.405
T0302	2JM5	0.450	0.391 ●
T0334	2OAL	0.351	0.275 ●
μ		0.446	0.428

In Stage 3, called the *greedy expansion stage*, we give the loser feature sets the second chance. Suppose the base from Stage 2 is XY and the loser feature sets are ABCD. We check the combinations XYA, XYB, XYC and XYD in the first round. The feature set combination with the lowest average RMSD will be the base for the next round. Suppose XYC results in the lowest average RMSD, it will become the new base. Then, we check XYCA, XYCB and XYCD in the second round. $22 - \rho$ rounds are executed in this stage, since there are $22 - \rho$ loser feature sets. When one loser feature set is added into the base, we test the new combination for the average RMSD, instead of the accuracy. It is time-consuming for calculating the average RMSD after jackknife test, but looking for the combination with the lowest RMSD is the only goal in this stage because we have found the combination with relatively high accuracy in Stage 2.

Table 5: The classification accuracies of the top 10 feature set combinations in Stage 1.

Combination	Accuracy	log(c)	log(γ)
N	80.00	8	-6
LN	76.67	8	-6
AB	73.33	6	-4
I	70.00	8	-2
P	70.00	2	6
D	70.00	2	-2
H	66.67	8	-2
Q	66.67	4	2
PQ	66.67	8	-8
Z	66.67	6	-6

4 Experimental Results

We test our preference classification method on CASP7. There are 30 protein sequences consisting of only standard amino acids. Table 4 shows the ID in CASP7, PDB and the RMSD in two prediction tools.

In Stage 1, 50 combinations, including C, and powers sets of HIJ, TUV, PQR, XYZ, EFG, LMN, and ABD are tested. N, LN, AB, I, P, D, H, Q, PQ, Z are the top 10 combinations with respect to accuracy, as shown in Table 5. In our experiments, we set $\rho = 6$. So we pick N, L, A, B, I, and P as the 6 winner feature sets, and the other 16 feature sets are the losers. The power sets of these 6 feature sets are tested in Stage 2. The combination with the lowest average RMSD is NB (0.410 $\mathring{A}$), as listed in Table 6. And then NB is chosen as the initial base of Stage 3. In the first round of Stage 3, we add the 16 loser feature sets into NB: NBC, NBH, NBJ, NBT, NBU, NBV, NBQ, NBR, NBX, NBY, NBZ, NBE, NBF, NBG, NBM and NBD. We get that the lowest average RMSD appears in NBT (0.405$\mathring{A}$). NBT becomes the new base for the next round, and then we test NBTC, NBTH, NBTJ, NBTU, NBTV, $\cdots$. The best RMSD of each round is shown in Table 7. In the table, we conclude that the best feature set combination found by our method is NBT, which improves RMSD to 0.405$\mathring{A}$. For a perfect classifier (100% classification), the average RMSD could be reduced to 0.399$\mathring{A}$, which is the upper bound of performance improvement of our method. The conclusion of our method and performance improvement are shown in Table 8. The average RMSD of our method is reduced to 0.405$\mathring{A}$, which is 9.19% and 5.37% improvement against SABBAC and Chang's result, respectively.

Table 6: The average RMSD and classification accuracies of the top 10 feature set combinations in Stage 2. log(c) and log(γ) denote the best parameter values for SVM.

Combination	Average RMSD	Accuracy	log(c)	log(γ)
NB	0.410	83.33	8	-8
NBP	0.414	80.00	8	-8
NBI	0.415	76.67	8	-6
AB	0.415	76.67	6	-4
NAB	0.416	80.00	8	-6
NLA	0.416	76.67	8	-8
N	0.416	80.00	8	-6
NL	0.418	73.33	8	-6
NLB	0.421	73.33	8	-6
NABI	0.422	76.67	8	-6

Table 7: The best RMSD in each round of Stage 3.

Round	Combination	Average RMSD
Round 0	NB	0.410
Round 1	NBT	0.405
Round 2	NBTE	0.406
Round 3	NBTEU	0.409
Round 4	NBTEUX	0.410
Round 5	NBTEUXG	0.412
Round 6	NBTEUXGF	0.415
Round 7	NBTEUXGFH	0.417
Round 8	NBTEUXGFHY	0.417
Round 9	NBTEUXGFHYZ	0.421
Round 10	NBTEUXGFHYZR	0.419
Round 11	NBTEUXGFHYZRV	0.417
Round 12	NBTEUXGFHYZRVM	0.421
Round 13	NBTEUXGFHYZRVMQ	0.427
Round 14	NBTEUXGFHYZRVMQD	0.428
Round 15	NBTEUXGFHYZRVMQDJ	0.427
Round 16	NBTEUXGFHYZRVMQDJC	0.428

5 Conclusions

SVM is an effective classifier used widely. If we want to get high classification accuracy, extracting more effective features for the classifier is necessary. Therefore, many studies, such as feature filter [4], and feature wrapper, are in progress. People try to find better features to improve the accuracy, but our goal is a little different. We want not only to improve the accuracy but also to decrease the average RMSD. For this reason, we propose a three-stage method to search the effective feature set combination for decreasing the RMSD in protein structure prediction. Our aim is to get the lowest RMSD for each protein in the dataset. We propose a new feature extraction scheme to get 22 new feature sets and then build the best feature combination with a heuristic method.

Most of time, SVM is faster than NN (Neural Network), but sometime it is still not fast enough. In our case, we have to check $2^{22} - 1$ combina-

Table 8: The performance improvement in percentage of average RMSD against Chang's method and SABBAC.

Dataset	Average RMSD	Against SABBAC / Perfect	Against Chang's method / Perfect
CASP7 standard	0.405	9.19% / 10.54%	5.37% / 6.78%

tions of feature sets if the real best combination is desired. But it is not easy to finish the all-combination job in limited time. Therefore, we design this heuristic method to solve the problem in acceptable time. When we set $\rho = 6$, in the worst case, our method need to check only $1+(2^3-1)\times7+(2^6-1)+\frac{16\times15}{2}$ combinations, which is only 0.0056% of the original combinations we have to test. By our method, the RMSD is reduced to 0.405$\mathring{A}$, which is 9.19% and 5.37% improvement against SABBAC and Chang's result, respectively. Thus, we believe our method is an efficient and effective method to reduce the RMSD of protein 3D structure prediction. If, in the future, some other prediction tools with lower RMSD are proposed, we can still apply our method to improve the performance a little.

References

[1] C. -C. Chang and C. -J. Lin, “LIBSVM: A library for support vector machines”, National Taiwan University, No. 1, Roosevelt Rd. Sec. 4, Taipei, Taiwan 106, ROC, 2001.

[2] H. Y. Chang and C. B. Yang and H. Y. Ann, “Refinement on O Atom Positions for Protein Backbone Prediction,” *2009 WSEAS International Conference on Biomedical Electronics and Biomedical Informatics (BEBI'09)*, pp. 99-104, 2009.

[3] W. D. Cornell and P. Cieplak and C. I. Bayly and I. R. Gould and K. M. Merz and Jr. and D. M. Ferguson and D. C. Spellmeyer and T. Fox and J. W. Caldwell and P. A. Kollman, “A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules”, *Journal of American Chemical Society* Vol. 117, pp. 5179-5197, 1995.

[4] M. Dash and H. Liu, “Feature Selection for Classification”, *Intelligent Data Analysis*, Vol. 1, pp. 131-156, 1997.

[5] C. H. Q. Ding and I. Dubchak, “Multi-class protein fold recognition using support vector machines and neural networks”, *Bioinformatics*, Vol. 17, No. 4, pp. 349-358, 2001.

[6] I. Dubchak and I. Muchnik and C. Mayor and I. Dralyuk and S. -H. Kim, “Recognition of a protein fold in the context of the SCOP classification”, *Proteins: Structure, Function, and Genetics*, Vol. 35, No. 4, pp. 401-407, 1999.

[7] J. -S. Lin, “An Effective Feature Selection for Protein Fold Recognition”, Department of Computer Science and Engineering, National Sun Yat-sen University, Taiwan, No. 70, Lienhai Rd., Kaohsiung 80424, Taiwan, R.O.C, 2007.

[8] Craig D.Livingstone and Geoffrey J.Barton, “Protein Sequence Alignments: a Strategy for the Hierarchical Analysis of Residue Conservation”, *CABIOS*, Vol. 9, pp. 745-756, 1993.

[9] J. Maupetit and 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.

[10] I. Ruczinski and C. Kooperberg and R. Bonneau and D. Baker, “Distributions of beta sheets in proteins with application to structure prediction”, *Proteins: Structure, Function, and Genetics*, Vol. 48”, No. 1, pp. 85-97, 2002.

[11] J. H. Wang and C. B. Yang and C. T. Tseng, “Reconstruction of Protein Backbone with the α-Carbon Coordinates”, *Journal of Information Science and Engineering*, 2009.