Prediction of Protein Backbone Structure by Preference Classification with SVM *

Kai-Yu Chen[#], Chang-Biau Yang[#1] and Kuo-Si Huang[&]
[#] National Sun Yat-sen University, Kaohsiung, Taiwan
[&] National Kaohsiung Marine University, Kaohsiung, Taiwan
[1] Corresponding author: cbyang@cse.nsysu.edu.tw

Abstract

Given the primary sequence of a protein and its α-carbon coordinates, the all-atom protein backbone reconstruction problem (PBRP) is to reconstruct the 3D coordinates of all atoms, including N, C, and O atoms on the backbone. A variety of methods for solving PBRP have been proposed, such as Adcock's method, SABBAC, BBQ, and Chang's methods. In this paper, we involve BBQ (Backbone Building from Quadrilaterals) and Chang's method as our candidate prediction tools. Then, we apply a tool preference classification with support vector machine (SVM) to determine which tool is more suitable for solving PBRP. According to the preference classification result, a proper prediction tool, either BBQ or Chang's method, is used to construct the atom of the target protein. Thus, after combining the results of all atoms, the backbone structure of the target protein is reconstructed. The three data sets of our experiments were extracted from CASP7, CASP8, and CASP9, which consists of 29, 24, and 55 proteins, respectively. The data sets contain only the proteins composed of standard amino acids. We improve the average RMSDs of BBQ results from 0.3955 to 0.3835 in CASP7, from 0.4437 to 0.4313 in CASP8, and from 0.4133 to 0.3691 in CASP9.

Key words: bioinformatics, protein, backbone prediction, RMSD, SVM

1 Introduction

A protein is composed of a series of amino acids. Amino acids can be divided into two groups, standard and nonstandard. There are twenty different kinds of standard amino acids, such as Glycine, Alnine, and Valine. Scientists categorize the rest of amino acids into the nonstandard group. In 1955, Frederick Sanger determined the sequence of amino acids in insulin (James, 1993). After a few years, Max Perutz and Sir John Cowdery Kendrew determined the 3D (three-dimensional) structure of a protein by using X-ray crystallographic. The first protein whose 3D structure was reconstructed is myoglobin.

*This research work was partially supported by the National Science Council of Taiwan under contract NSC-100-2221-E-110-050.

Nowadays, the 3D coordinates of a protein can be built by X-ray crystallographic or nuclear magnetic resonance (NMR), but plenty of time and cost are required. Therefore, two computational methods, *ab initio* and *homology modeling*, were proposed for predicting the 3D protein structure (Zhang, 2008).

The *ab initio* method considers the interactions between the amino acids and the liquors to determine the stable status of each molecule. The *homology method* builds the templates with known protein structures and then produces a model based on the alignments between the target protein and the templates. The proper templates can be determined by the similarities of their conformations (Holm & Sander, 1991; Kazmierkiewicz, Liwo, & Scheraga, 2002; Simons, Kooperberg, Huang, & Baker, 1997; Samudrala & Moult, 1998).

The *all-atom protein backbone reconstruction problem* (PBRP) is to reconstruct the 3D coordinates of all atoms (N, C, and O) on the backbone for a protein whose amino acid sequence and α-carbon coordinates are given. Wang *et al.* built a 4-residue fragment library to reconstruct all atoms on the backbone (Wang, Yang, & Tseng, 2007). Chang *et al.* proposed a method to refine the positions of oxygen atoms on the backbone of a protein (Chang, Yang, & Ann, 2009). Yen *et al.* further proposed a tool preference classification by choosing either SABBAC (Maupetit, Gautier, & Tuffery, 2006) or Chang's method (Chang et al., 2009) to predict the protein structure for PBRP (Yen, Yang, & Ann, 2010). The method proposed by Yen *et al.* tries to find a way to select possibly better predicting method between SABBAC and Chang's method for improving the accuracy. Another method for PBRP, named BBQ (Backbone Building from Quadrilaterals) (Gront, Kmiecik, & Kolinski, 2007), considers a fragment with four contiguous amino acids as quadrilateral. Then, it calculates C_α (α-carbon) trace to choose proper quadrilaterals as the prediction results.

Our goal is to improve the accuracy of structure prediction for PBRP. Since the performance of BBQ is better than that of SABBAC, we consider BBQ and Chang's method as our candidate prediction tool, instead of SABBAC. We apply the tool preference classification individually to determine the preferred prediction tool, either BBQ or Chang's method, for each atom. Based on the preferred tools, we get better accuracy for predicting the 3D coordinates of N, C, and O atoms. The preference classification tool we use is the *support vector machine* (SVM), with which the backbone structure of a protein can be reconstructed.

For evaluating the performance of our proposed method for PBRP, the data sets of CASP (Critical Assessment of Protein Structure Prediction) are used in this paper. CASP is a competition of protein structure prediction which takes place every two years from 1994. The data sets we use were extracted from the target proteins of 7th (CASP7), 8th (CASP8) and 9th (CASP9) competitions of CASP. In our experiments, we test 29, 24, and 55 proteins, consisting of only standard amino acids, in CASP7, CASP8, and CASP9, respectively. We improve the average RMSDs of BBQ results from 0.3955 to 0.3835 in CASP7, from 0.4437 to 0.4313 in CASP8, and from 0.4133 to 0.3691 in CASP9. The experimental results show that the accuracy of our method is superior to SABBAC, Chang's method, Yen's method and BBQ.

The rest of this paper is organized as follows. Section 2 introduces proteins and their properties, a method for measuring the similarity of two formations of a protein,

support vector machine (SVM) and some previous works. Then in Section 3, our method for PBRP is presented in detail. Section 4 shows the experimental results of self-tests and independent tests of our method and the comparisons with some previous methods. Finally, the conclusion and future work of this paper are given in Section 5.

2 Previous Work

The *protein backbone reconstruction problem* (PBRP) is to reconstruct all atoms on the protein backbone for a given protein sequence whose 3D coordinates of α-carbons are known. SABBAC is an on-line service devoted to protein backbone reconstruction from α-carbon trace (Maupetit et al., 2006). SABBAC applies the encoding of the protein trace in a hidden Markov model-derived structural alphabet to assemble the fragments. Then, they use a greedy algorithm to search a combination of fragments. The scoring function inspired from the OPEP force field (Santini, Wei, Mousseau, & Derreumaux, 2003) derives the search. The scoring function is defined as follows:

$$E = E_{SC,SC} + E_{HB} + E_{VdW} + E_{PhiP} + E_{BB} + E_{Trans}. \quad (1)$$

In Equation 1, $E_{SC,SC}$, E_{HB}, E_{VdW}, E_{PhiP}, E_{BB}, and E_{Trans} denote the side-chain and side-chain interaction, the hydrogen bonds, the van der Waals force, the phi (ϕ) positive contribution, the valence angle distortion of α-carbon, and a pseudo energy term related to the transitions between consecutive fragments, respectively. SABBAC consumes much execution time when the target protein is large. The SABBAC service is also available via the Internet (Maupetit, Gautier, & Tuffery, n.d.).

The BBQ (Backbone Building from Quadrilaterals) method (Gront et al., 2007) is an algorithm for protein backbone reconstruction that comprises high computational efficiency and accuracy. It considers a fragment with four contiguous amino acids as quadrilateral. Distances of C_α atoms are also described as the quadrilateral. Therefore, three distances form a three-dimensional look-up table in which average positions of N, C, and O atoms measured in a local coordinate system are acquired from the known PDB structures.

Wang *et al.* solved the all-atom PBRP by using the homology modeling technique (Wang et al., 2007). In the beginning, it takes every four consecutive residues as a fragment in the protein sequence, and extracts the fragments from all protein structures in PDB (Protein Data Bank) (Berman et al., 2000). These fragments are collected as the fragment library. Each fragment is identified by its second, third and fourth residues and it can be classified into 8000 residue groups. The fragments in each residue group are clustered based on their structures. In a cluster, one typical fragment is chosen as the delegate to represent this cluster. While reconstructing, it seeks out suitable candidates from the fragment library.

PULCHRA is a fast and robust method for the reconstruction of all-atom protein models starting from a reduced protein representation (Rotkiewicz & Skolnick, 2008). The three stages of PULCHRA are described as follows. First, PULCHRA optimizes α-carbon positions. Then, it reconstructs and optimizes backbone and side-chain. Finally, it reconstructs hydrogen atoms and output the result.

Table 1: The win frequencies of various methods in each data set.

Data set	SABBAC	BBQ	PULCHRA	Chang	Total
CASP7	6	12	2	9	29
CASP8	4	13	2	5	24
CASP9	7	32	1	15	55

Chang *et al.* (2009) found that some results of Wang *et al.* (2007) are not so good. They discovered that the average RMSD of O atom is usually worse than N and C atoms. They measured the lengths and angles of peptide bonds so that the energy function can be adjusted to a simpler one. According to the energy function, the position of each O atom can be found independently. For improving the accuracy, they also proposed the two-phase refinement strategy.

Yen *et al.* (2010) observed the results from Chang *et al.* (2009) and SABBAC (Maupetit et al., 2006), and they discovered that the predictions of Chang's method are not always better than SABBAC, and vice versa. If one can exactly select the better predicting method between SABBAC and Chang's method for each protein, the result of prediction would be improved. Hence Yen *et al.* proposed a method which employs the SVM as the classifier to build a decision model, which can guide us to determine which tool is better for predicting the structure of the target protein.

3 Our Method

In the method proposed by Yen *et al.* (2010), Chang's method and SABBAC are employed as the candidate prediction tools. There are some other methods (PULCHRA and BBQ) for predicting protein structure. And, we did some primitive experiments on the data sets of CASP7, CASP8 and CASP9 for evaluating these methods. Table 1 shows that BBQ and Chang's method can be thought as better prediction tools among these four methods, because that their frequencies of winning predictions are greater than the other two. It can be easily seen that none of these two selected tools is always the best method for predicting 3D structures of proteins.

In order to predict the structure of protein more accurately, we further observe the results from BBQ and Chang's predictions. Comparing the predictions of N-atoms, C-atoms and O-atoms, we realize that each of these two prediction tools has its own superiority on prediction. Hence, we modify Yen's method by considering N, C, and O atoms individually. We come out with a new prediction model, named *atom classifier*. We apply the atom classifiers on a given protein sequence whose α-carbon coordinates are known. After assembling the results from different atom classifiers, we can rebuild the 3D coordinates of all atoms (N, C, and O) on the backbone. The flowcharts of our method and the atom classifier are shown in Figures 1 and 2, respectively.

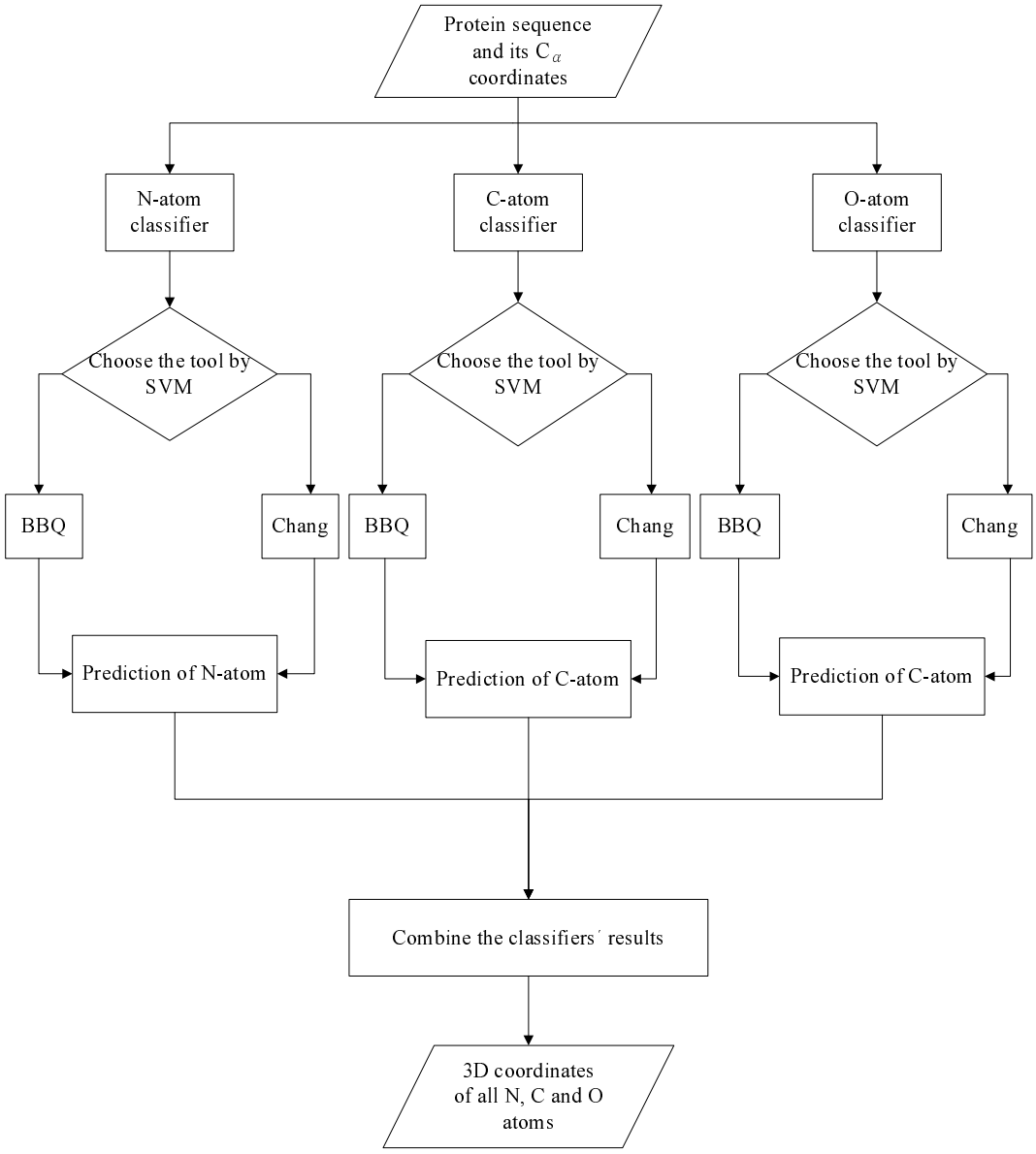

Figure 1: The flow chart of selecting better predictors for predicting the 3D structure of a protein.

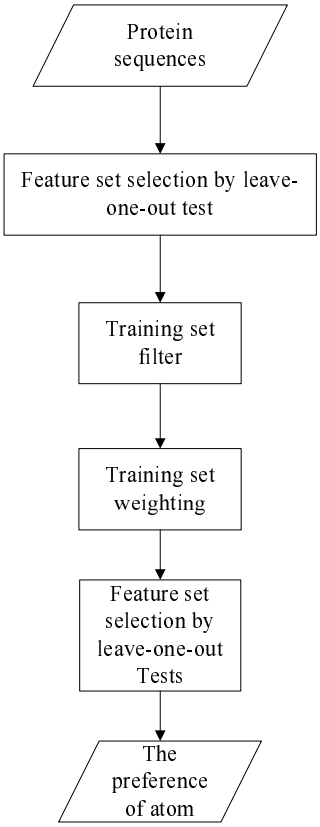

Figure 2: The flow chart of the atom classifier.

Table 2: The amino acid distribution of each feature (Huang et al., 2003; Li et al., 2006; Yu et al., 2006).

Feature	Group 1	Group 2	Group 3
Hydrophobicity (H)	(Polar) R, K, E, D, Q, N	(Neutral) G, A, S, T, P, H, Y	(Hydrophobic) C, V, L, I, M, F, W
Van der Waals volume (V)	(Positive) G, A, S, C, T, P, D	(Neural) N, V, E, Q, I, L	(Negative) M, H, K, F, R, Y, W
Polarity (P)	(Polarity value 4.9-6.2) L, I, F, W, C, M, V, Y	(Polarity value 8.0-9.2) P, A, T, G, S	(Polarity value 10.4-13.0) H, Q, R, K, N, E, D
Polarizability (Z)	(Value 0-1.08) G, A, S, D, T	(Value 0.128-120.186) C, P, N, V, E, Q, I, L	(Value 0.219-0.409) K, M, H, F, R, Y, W
Size (G)	(Tiny) G, C, S, A	(Small) T, D, N, P, V	(Normal) I, L, M, F, Y, W, K, R, H, E, Q
Charge (N)	(Positive) K, R	(Neutral) H, Q, N, S, C, T, Y, W, P, A, G, V, I, L, M, F	(Negative) D, E
Molecular weight (W)	(Small) G, A, S, P, V, T, C	(Middle) I, L, N, D	(Large) Q, K, E, M, H, F, R, Y, W
Isoelectric point (I)	(Negative) D, C, E	(Neural) N, F, Q, T, Y, S, M, W, V, G, L, A, I, P, H	(Positive) R, K
Accessible surface area (S)	(Buried) G, A, S, C, T, P, D	(Exposed) V, N, L, I, Q, M	(Intermediate) E, H, K, F, R, Y, W

3.1 Feature Extraction of Atom Classifier

In order to make the SVM classification more accurate, it is necessary to find out good features for organizing a good feature set. Once the inappropriate feature set is used, the noise would make the classification result incorrect. Our universal feature set includes nine features, *hydrophobicity* (H), *van der Waals volume* (V), *polarity* (P), *polarizability* (Z), *size* (G), *charge* (N), *molecular weight* (W), *isoelectric point* (I), and *accessible surface area* (S). According to the physical characteristic, each of these features can be divided into three groups (Huang, Lin, & Pal, 2003; Li et al., 2006; Yu, Cao, Cai, Shi, & Li, 2006). For example, hydrophobicity contains three groups: polar, neutral, and hydrophobic. Table 2 shows the detail of each feature.

In the previous study, Dubchak *et al.* (Dubchak, Muchnik, Holbrook, & Kim, 1995) use three descriptors, composition (C), transition (T), and distribution (D), to describe the global composition of amino acid property in a protein. Composition (C) denotes the percentage of amino acids of a certain property. Transition (T) is used to indicate the change frequency of a property along the protein sequence. Distribution (D) represents the percentage of the sequence length corresponding to the positions of 1%, 25%, 50%, 75%, and 100% of amino acids for a property (Huang et al., 2003; Yu et al., 2006).

Figure 3 shows an example to illustrate how the feature set is extracted. This figure takes account of hydrophobicity (H) as an example. The length of this sequence is 25 and the counts of three residue groups polar (P), neutral (N) and Hydrophobic (H) are 6, 11 and 8, respectively. The sequence of composition C-descriptor is 0.24 (6/25), 0.44 (11/25), and 0.32 (8/25). The sequence of T-descriptor is 0.08 (2/24), 0.375 (9/24), and 0.08 (2/24). The sequence of D-descriptor for group P is 0.24 (6/25), 0.4 (10/25), 0.44 (11/25), 0.56 (14/25), and 0.6 (15/25). In the same way, the sequence of D-descriptor for

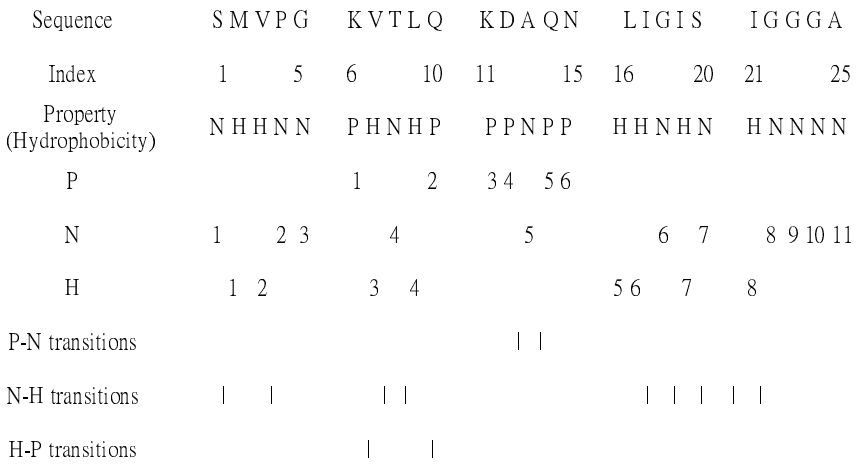

Figure 3: An example for illustrating the extraction of the feature sets.

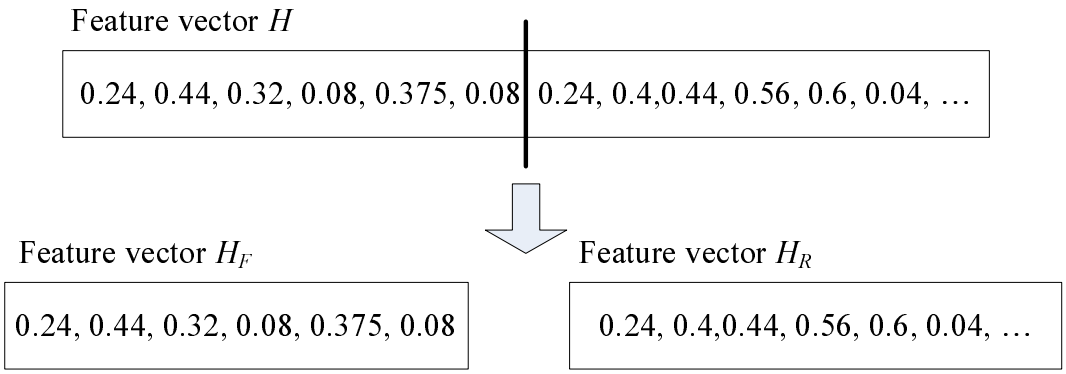

Figure 4: An example for illustrating that the feature vector H is divided into two parts, H_F and H_R.

group N is 0.04 (1/25), 0.2 (5/25), 0.72 (18/25), 0.92 (23/25), and 1 (25/25). For group H, the sequence of D-descriptor is 0.08 (2/25), 0.12 (3/25), 0.36 (9/25), 0.68 (17/25), and 0.84 (21/25). Finally, we combine above elements and get the feature vector of 21 elements: {0.24, 0.44, 0.32, 0.08, 0.375, 0.08, 0.24, 0.4, 0.44, 0.56, 0.6, 0.04, 0.2, 0.72, 0.92, 1, 0.08, 0.12, 0.36, 0.68, 0.84}. Then, the vector is divided into two parts. The first part includes the former 6 elements and the other one contains the rest 15 elements. For instance, H is the feature vector of 21 elements. The first (former) part and the other (rear) part are denoted as H_F and H_R, where $|H_F| = 6$ and $|H_R| = 15$. Figure 4 shows an example that the feature vector is divided into the former and rear parts.

3.2 Feature Set Reorganization

First, all feature sets are initialized by self-tests for selecting better feature subsets. Then, these feature sets are reorganized by using *crossover* and *extension* operations. In the crossover operation, given two feature vectors X and Y, we exchange X_F and Y_F. Then, we get two new feature vectors $X_F Y_R$ and $Y_F X_R$. The extension operation is to add X_F or X_R to Y, and to add Y_F or Y_R to X, we get $X_F Y_F Y_R$, $X_R Y_F Y_R$, $X_F X_R Y_F$ and $X_F X_R Y_R$.

Table 3: An example for setting filter and weight.

PDB ID	BBQ	Chang	RMSD difference
2H57	0.1791	0.2917	0.1126
2HB6	0.1678	0.2155	0.0447

3.3 Training Set Filtering and Weighting

Generally speaking, every method for PBRP wishes that the RMSD of prediction result is as small as possible. If we know which prediction tool should be used, we can improve the overall RMSD. When the RMSD difference of the prediction results of these two tools on a protein prediction is high, we can improve the overall RMSD significantly if we choose the better tool correctly to do prediction. If the RMSD exceeds the filter (a threshold value), we will duplicate some clones of this sequence into training set to increase the weight of this sequence. On the other hand, when the RMSD difference is less than the filter, it will not be critical if the selected prediction tool is not the better one. Yen's method indicated that once filter is set to 0.05, the results will be better.

Then, we consider the RMSD differences of N, C and O atoms individually between BBQ and Chang's method. We weight each protein sequence by the following equation:

$$\alpha \times \frac{\text{RMSD difference}}{0.01}, \tag{2}$$

where parameter α will be set in the self-test. Table 3 shows an example for training set filtering and weighting. The RMSD difference of N-atom in protein 2HB6 is 0.0447, and the difference is less than filter. This protein is removed from the training set. The RMSD difference of N-atoms in protein 2H57 is 0.1126, then additional 22 clones of the sequence will be added into the training set when $\alpha = 2$. After the additional clones are added, the total number of clones of 2H57 in the set is 23.

3.4 The Classification of SVM

In the atom classification step, we examine the reorganized feature sets by leave-one-out test. Then, the feature set with smaller RMSD is used in the final atom classifier. We use the results of SVM as the tool preference to determinate which tool is the better one for predicting the structure of a specific atom. Then, we combine the predicted results of N, C, and O atoms to construct the all-atom 3D coordinates of a protein.

4 Experimental Results

CASP (Critical Assessment of Protein Structure Prediction) (Moult, Fidelis, Kryshtafovych, Rost, & Tramontano, 2009) is a contest for making the techniques of protein structure prediction improved and finding out which features could be important. It takes place every two years from 1994 and CASP9 was held in 2010. The number of target sequences in CASP7 is 107. In these sequences, 66 proteins are contained in the server part and the

Table 4: The RMSDs of our twelve experiments and the comparison of various methods.

Test	Training	Test	Our method	Modified Yen	Yen	BBQ	Chang	PAA	PIA
I	CASP7	CASP7	0.3684	0.3762	0.4019	0.3955	0.4264	0.3624	0.3547
II	CASP8	CASP7	0.3882	0.3902	0.4191				
III	CASP9	CASP7	0.3849	0.3956	0.4265				
IV	CASP8+9	CASP7	0.3835	0.3956	0.4291				
V	CASP8	CASP8	0.4203	0.4358	0.4543	0.4437	0.5010	0.4321	0.4192
VI	CASP7	CASP8	0.4288	0.4437	0.4219				
VII	CASP9	CASP8	0.4260	0.4437	0.4852				
VIII	CASP7+9	CASP8	0.4313	0.4419	0.4866				
IX	CASP9	CASP9	0.3609	0.3837	0.4104	0.4133	0.4320	0.3768	0.3585
X	CASP7	CASP9	0.3839	0.4133	0.4384				
XI	CASP8	CASP9	0.3763	0.4133	0.4415				
XII	CASP7+8	CASP9	0.3691	0.4106	0.4302				

other 41 proteins are in the human part. Considering the 66 proteins in the server part, 37 sequences contain at least one nonstandard amino acid. Hence, we have 29 sequences in our training set, which consists of only standard amino acids. Similarly, there are 24 and 55 sequences in the server part containing only standard amino acids in CASP8 and CASP9, respectively. We perform twelve experiments, including three self-tests and nine independent tests. Table 4 describes the training set and testing set of each experiment. It also summarizes the average RMSD of each method. The term "Modified Yen" means that this paper re-performs Yen's method by replacing prediction tools with BBQ and Chang's method rather than SABBAC and Chang's method originally.

The term "PAA" (perfect for all atoms) in Table 4 means the proper tool can be exactly chosen for predicting all atoms, i.e., the best tool is applied to the prediction of all atoms in the whole target protein. One may note that the RMSD of PAA for a certain protein is the better RMSD between BBQ and Chang's method. Another term "PIA" (perfect for individual atoms) means the RMSD of each atom on the backbone (including N, C and O atoms) which is correctly predicted by either BBQ or Chang's method individually. Note again that if one atom (say, N) decides to employ a method as its classifier, then this classifier is invoked for the same kind of atom (N) in the whole target protein. It is clear that "PIA" forms the result bound of our method for PBRP.

Table 5 shows the average RMSD of each atom on the backbone. One may observe that the average RMSD of O-atom is always maximal. Our method is excellent on selecting the proper method for predicting O-atom. Therefore, RMSDs of our prediction are better than the previous methods.

5 Conclusion

For the all-atom protein backbone reconstruction problem (PBRP), Yen *et al.* (2010) recently proposed a tool preference classification method to predict the coordinates of all atoms on the backbone of a target protein by choosing either Chang's method or SABBAC. According to the preference, a better prediction tool is employed to do the coordinate prediction. This paper improves Yen's method by performing tool preference for N, C and O atoms separately. The candidate tools for us are Chang's method and

Table 5: The average RMSD of each atom on the backbone by BBQ and Chang's method for data sets CASP7, CASP8 and CASP9.

Data set	Method	N	C	O
CASP7	BBQ	0.2939	0.2581	0.6777
	Chang	0.2791	0.2887	0.7492
CASP8	BBQ	0.3240	0.2842	0.7573
	Chang	0.3072	0.3334	0.8913
CASP9	BBQ	0.3976	0.2454	0.6359
	Chang	0.2657	0.2954	0.7645

BBQ (Backbone Building from Quadrilaterals). The backbone can be reconstructed by combining the predictions of all atoms. The feature set contains nine feature vectors initially, and we employ a feature selection scheme to reduce the size of the feature set. The result is pretty close to perfect RMSD. In CASP7, the average RMSDs are 0.3684, 0.4019 and 0.3955 in our method, Yen's method and BBQ, respectively. In addition, the average RMSDs are 0.4203, 0.4543 and 0.4437 in CASP8 and 0.3609, 0.4155 and 0.4133 in CASP9, respectively. For more fair comparison, we compare the results of independent tests in our methods (experiments IV, VIII and XII). We still improve the average RMSDs of BBQ results from 0.3955 to 0.3835in CASP7, from 0.4437 to 0.4313 in CASP8, and from 0.4133 to 0.3691 in CASP9. The results of our method are better than other previous works. Comparing with Yen's method, we may conclude that predicting N, C and O atoms individually is better than using one method to predict all atoms.

In the future, the performance may be further improved through several possible ways. First, one may consider other predicting methods with higher accuracies to improve our method. According to the preference of these methods, a better method can be selected by voting. Second, our method is also affected by nonstandard amino acids because Chang's method is not very good on the proteins containing nonstandard amino acids. If we can find an excellent method that can predict proteins containing both standard and nonstandard amino acids, our method would be improved. Third, many methods for feature selection have been proposed (Huang et al., 2003; Muni, Pal, & Das, 2006; Sotoca & Pla, 2010). We may use these methods to choose the better features for improving the classification accuracy.

References

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., ... Bourne, P. E. (2000). The protein data bank. *Nucleic Acids Research*, *28*, 235-242.

Chang, H. Y., Yang, C. B., & Ann, H. Y. (2009). Refinement on o atom positions for protein backbone prediction. In *Proceedings of the 2nd WSEAS international conference on biomedical electronics and biomedical informatics.*

Dubchak, I., Muchnik, I., Holbrook, S. R., & Kim, S. H. (1995). Prediction of protein folding class using global description of amino acid sequence. In *Proceedings of the national academy of sciences of the united states of america* (Vol. 92, p. 8700-8704).

Gront, D., Kmiecik, S., & Kolinski, A. (2007). Backbone building from quadrilaterals: A fast and accurate algorithm for protein backbone reconstruction from alpha carbon coordinates. *Journal of Computational Chemistry*, *28*, 1593-1597.

Holm, L., & Sander, C. (1991). Database algorithm for generating protein backbone and side-chain coordinates from a c alpha trace application to model building and detection of coordinate errors. *Journal of Molecular Biology*, *21*(1), 183-194.

Huang, C.-D., Lin, C.-T., & Pal, N. R. (2003). Hierarchical learning architecture with automatic feature selection for multiclass protein fold classification. *IEEE Transactions on NanoBioscience*, *2*, 221-232.

James, L. K. (1993). *Nobel laureates in chemistry, 1901-1992.* Chemical Heritage Foundation (June 1, 1993).

Kazmierkiewicz, R., Liwo, A., & Scheraga, H. A. (2002). Energy-based reconstruction of a protein backbone from its α-carbon trace by a Monte-Carlo method. *Journal of Computational Chemistry*, *23*, 715-723.

Li, Z. R., Lin, H. H., Han, L. Y., Jiang, L., Chen, X., & Chen, Y. Z. (2006). PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Research*, *34*, W32-W37.

Maupetit, J., Gautier, R., & Tuffery, P. (n.d.). *SABBAC v1.2: Structural alphabet based protein backbone builder from alpha carbon trace.* http://bioserv.rpbs.jussieu.fr/cgi-bin/SABBAC.

Maupetit, J., Gautier, R., & Tuffery, P. (2006). SABBAC: online structural alphabet-based protein backbone reconstruction from alpha-carbon trace. *Nucleic Acids Research*, *34*, W147-W151.

Moult, J., Fidelis, K., Kryshtafovych, A., Rost, B., & Tramontano, A. (2009). Critical assessment of methods of protein structure prediction - Round VIII. *Proteins*, *77*, 1-4.

Muni, D. P., Pal, N. R., & Das, J. (2006). Genetic programming for simultaneous feature selection and classifier design. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, *36*, 106-117.

Rotkiewicz, P., & Skolnick, J. (2008). Fast procedure for reconstruction of full-atom protein models from reduced representations. *Journal of Computational Chemistry*, *29*, 1460-1465.

Samudrala, R., & Moult, J. (1998). An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *Journal of Molecular Biology*, *275*, 895-916.

Santini, S., Wei, G., Mousseau, N., & Derreumaux, P. (2003). Exploring the folding pathways of proteins through energy landscape sampling: Application to Alzheimer's β-amyloid peptide. *Internet Electronic Journal of Molecular Design*, *2*(9), 564-577.

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

Sotoca, J. M., & Pla, F. (2010). Supervised feature selection by clustering using conditional mutual information-based distances. *Pattern Recognition*, *43*, 2068-2081.

Wang, J. H., Yang, C. B., & Tseng, C. T. (2007). Reconstruction of protein backbone with the α-carbon coordinates. In *Proceedings of 2007 national computer symposium*

(p. 136-143). Taichung, Taiwan.

Yen, H. W., Yang, C. B., & Ann, H. Y. (2010). An effective tool preference selection method for protein structure prediction with SVM. In *Proceedings of the 27th workshop on combinatorial mathematics and computation theory* (p. 62-67). Taichung, Taiwan.

Yu, X., Cao, J., Cai, Y., Shi, T., & Li, Y. (2006). Predicting rRNA-, RNA-, and DNA-binding proteins from primary structure with support vector machines. *Journal of Theoretical Biology*, *240*, 175-184.

Zhang, Y. (2008). Progress and challenges in protein structure prediction. *Current Opinion in Structural Biology*, *18*, 342-348.