Coordinate Refinement on All Atoms of the Protein Backbone with Support Vector Regression

Ding-Yao Huang[†], Chiou-Yi Hor[‡] and Chang-Biau Yang[†*]

[†]Department of Computer Science and Engineering,
National Sun Yat-sen University,
Kaohsiung 80424, Taiwan
[‡]Iron and Steel Research Development Department,
China Steel Corporation,
Kaohsiung 81233, Taiwan
[*]Corresponding author: cbyang@cse.nsysu.edu.tw

Abstract. For the past decades, many efforts have been made in the fields of protein structure prediction. Among these, the protein backbone reconstruction problem (PBRP) has attracted much attention. The goal of PBRP is to reconstruct the 3D coordinates of all atoms along the protein backbone for given a target protein sequence and its C_α coordinates. In order to improve the prediction accuracy, we attempt to refine the 3D coordinates of all backbone atoms by incorporating the state-of-the-art prediction softwares and support vector regression (SVR). We use the predicted coordinates of two excellent methods, PD2 and BBQ, as our feature candidates. Accordingly, we define more than 100 possible features. By means of the correlation analysis, we can identify several significant features deeply related to the prediction target. Then, a 5-fold cross validation is carried out to perform the experiments, in which the involved datasets range from CASP7 to CASP11. As the experimental results show, our method yields about 8% improvement in RMSD over PD2, which is the most accurate predictor for the problem.

Keywords: protein backbone, bioinformatics, three-dimensional coordinates, support vector regression, prediction

1 Introduction

Proteins are required for growth and development in the human body. A protein, called a polypeptide, is composed of a chain of amino acids. A series of amino acids are linked together by peptide chains to form a protein backbone. An *amino acid* is the fundamental unit of a protein. There are twenty kinds of standard amino acids and each kind of amino acid can be differentiated by its R group. Each protein has its own specific functions and unique structure which can cooperate with other proteins to achieve some required functionalities. Because protein structure and protein function are closely related, in order to find out these protein functions, most biologists adopt the approaches that predict protein tertiary structures by means of amino acid sequences or other related information.

There are two main approach types of the protein structure prediction. The first one is the experimental method, including X-ray diffraction and nuclear magnetic resonance (NMR)[11,17]. The other one is the computational method, including *homology modeling* [5], *folding recognition* [7], and *ab initio* [10]. As for these two categories, the former one requires a lot of time and cost while the latter one doesn't. Therefore, this motivates us to adopt the computational method to predict the 3D protein structure.

The *all-atom protein backbone reconstruction problem* (PBRP) is to utilize a protein sequence and its 3D coordinates of α-carbon (C_α) for predicting the 3D coordinates of all atoms (N, C and O atoms) on the protein backbone. There are several related studies which fall into this category, such as SABBAC [12], Wang's method [20], Chang's method [2], BBQ [4], Chen's method [3], Wu's method [21], PD2 [13] and so on. For a complete survey on the methods (or software) of the protein backbone prediction, one can refer to the thesis written by Yuan [22].

In order to improve the accuracy of protein structure prediction, we propose a new method to refine the 3D coordinates of all backbone atoms by means of *support vector regression* (SVR) [18,19]. Our prediction target is the differences of N and O atoms' coordinates between the predicted results of PD2 and the real 3D coordinates of PDB, and the differences of C atoms' coordinates between the predicted results of BBQ and the real 3D coordinates of PDB. Our training features are generated from the predicted results of PD2 and BBQ.

The experimental datasets range from CASP7 to CASP11, where CASP stands for the *Critical Assessment of Protein Structure Prediction* [14]. We perform a 5-fold cross validation experiment for performance evaluation. For each fold of validation, a CASP dataset is extracted for testing and the remaining CASP datasets are involved for training. Coordinates in the training datasets are first predicted by PD2 and BBQ. Then these predicted coordinates are compared with their real coordinates to produce the differences. The differences are the learning objective for the SVR. In the feature selection stage, we analyze the correlations between the objective value and available features, and then select the most representative features. To ease the training process, we partition the amino acids into twenty groups (twenty datasets) and then predict these dif-

ferences by each individual SVR. Finally, we combine our predicted differences, their corresponding predicted N and O coordinates with PD2, and predicted C coordinates with BBQ to export our predicted coordinates of N, C and O. The performance is evaluated by the RMSD values. The experimental results show that our prediction results yield about 8% improvement over the results predicted by PD2, which is the most accurate predictor for the problem[22].

The rest of this paper is organized as follows. In Section 2, we will introduce experimental datasets, root-mean-square deviation (RMSD) and features used in this paper. In Section 3, we will describe our proposed method in detail. In Section 4, we will present our experimental results. Finally, in Section 5, the conclusion will be given.

2 Preliminaries

2.1 Performance Evaluation

Critical Assessment of Protein Structure Prediction (CASP) [14] is an international competition held every two years since 1994. The main goal of CASP is to evaluate the capabilities of the methods for identifying three-dimensional structure of the protein from its amino acid sequence. In order to assess the performance of a method, CASP examines the predicted 3D structures in many different ways, such as the accuracy of a model, accuracy of a quaternary structure, and so on. Because our research also focuses on the 3D structure prediction, we use CASP datasets to perform our experiments.

Root-mean-square deviation (RMSD) [8,9] is an evaluation method of molecular modeling which computes the average distance between the predicted values and the ground truths.

$$RMSD = \sqrt{\frac{1}{l}\sum_{i=1}^{l}(X_i^A - X_i^B)^2}, \tag{1}$$

where X_i^A and X_i^B denote the coordinates of the ith atom on the backbone in the proteins A and B, respectively, and l denotes the length of the proteins. Generally, a lower RMSD indicates the higher similarity, which means that the predicted coordinates is close to the real ones. RMSD has been widely used in structural biology. In this paper, we also use RMSD to evaluate the quality of the prediction models.

2.2 Datasets

In bioinformatics, several methods have been proposed for solving the protein backbone reconstruction problem in the past decades. Table 1 shows the RMSD comparison of the prediction results obtained from SABBAC [12], PULCHRA [16], Chang's method [2], BBQ [4] and PD2 [13], where the experimental datasets include CASP7, CASP8 and CASP9. These methods assume that the coordinates

of C_α are given in the target protein. As one can see, PD2 and BBQ are superior to the others for the protein backbone reconstruction. Because PD2 and BBQ show some diversity in their individual results, this motivate us derive our features based on both tools. Moreover, from the individual N, C and O predicted results shown in Table 2, we find that PD2 is better at predicting N and O atoms than BBQ while its prediction performance is not as good as BBQ for C atoms. With this observation, we define the objective values of our prediction as follows.

- O_x-difference: $O_x(PD2) - O_x(real)$, where $O_x(PD2)$ represents the x coordinate of O atom predicted by PD2, and $O_x(real)$ represents the real x coordinate of O atom in PDB.
- O_y-difference: $O_y(PD2) - O_y(real)$, for y coordinate.
- O_z-difference: $O_z(PD2) - O_z(real)$, for z coordinate.
- N_x-difference: $N_x(PD2) - N_x(real)$.
- N_y-difference: $N_y(PD2) - N_y(real)$.
- N_z-difference: $N_z(PD2) - N_z(real)$.
- C_x-difference: $C_x(BBQ) - C_x(real)$, where $C_x(BBQ)$ represents the x coordinate of C atom predicted by BBQ.
- C_y-difference: $C_y(BBQ) - O_y(real)$.
- C_z-difference: $C_z(BBQ) - O_z(real)$.

Table 1. The average RMSDs of SABBAC, PULCHRA, Chang's method, BBQ and PD2 in CASP7, CASP8 and CASP9, where the top-two ranking ones are underlined.

Dataset	Method				
	SABBAC	PULCHRA	Chang	BBQ	PD2
CASP7	0.5676	0.4705	0.4108	0.3632	0.3335
CASP8	0.4934	0.5951	0.4888	0.4412	0.4030
CASP9	0.5551	0.5989	0.4406	0.4344	0.3635

In this paper, we first define several kinds of features, all of which are obtained from the predicted results of PD2 and BBQ. Next, to find out the important features used in the prediction, we calculate the correlation of each objective value and each feature. These selected features are involved to train *support vector regression* (SVR) models. Then, we use the SVR models to predict these objective values. Finally, these objective values (coordinate differences) are combined with the predicted results of PD2 and BBQ together to build our predicted coordinates of the atoms on the protein backbone.

2.3 Feature Generation and Feature Selection

This subsection describes the features we use to build the SVR models. All features are extracted within a fragment, as illustrated in Figure 1. That is,

Table 2. The average RMSDs of individual atoms (N, C and O atoms) of BBQ and PD2 in CASP7 through CASP11.

Dataset	Method					
	BBQ			PD2		
	N	C	O	N	C	O
CASP7	0.2401	0.2395	0.6347	0.1897	0.2416	0.5879
CASP8	0.3775	0.2668	0.7051	0.2375	0.3057	0.6968
CASP9	0.4229	0.2569	0.6632	0.2462	0.2864	0.6115
CASP10	0.2429	0.2444	0.6513	0.2042	0.2298	0.5991
CASP11	0.2992	0.2310	0.6252	0.2197	0.2525	0.5994
Arithmetic mean	0.3165	0.2477	0.6449	0.2195	0.2632	0.6189
Weighted arithmetic mean	0.3205	0.2477	0.6482	0.2198	0.2646	0.6179

once the prediction target, like the C atom, is determined, we define a window around this atom. We say this atom and its surrounding atoms constitute a fragment. The features required by SVR are calculated as follows.

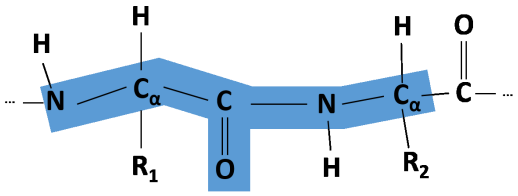

Fig. 1. The fragment of $L_C = 6$ consecutive atoms on the protein backbone, where the C atom is assigned to be the fragment center.

Coordinate: The predicted x, y and z coordinates of N, C, O and N_{next} atoms are obtained from BBQ and PD2, denoted as $N_x(BBQ)$, $N_y(BBQ)$, $C_z(PD2)$, etc. The real coordinates of two C_αs are also involved as the features. Thus, there are totally 30 features.

Coordinate difference: The coordinate differences are calculated from the predicted x, y and z coordinates of N, C, O and N_{next} atoms. Only the difference of each corresponding pair of coordinates is computed, such as $N_x(PD2) - N_x(BBQ)$, $N_y(PD2) - N_y(BBQ)$, etc. Thus, 12 features are obtained.

Euclidean distance: The Euclidean distance measures the amount of space between the two predicted positions, obtained from BBQ and PD2, of the

same atom on the Euclidean space. The formula for calculating Euclidean distance is given as follows.

$$d(x_{p_i}, x_{b_i}, y_{p_i}, y_{b_i}, z_{p_i}, z_{b_i}) = \sqrt{(x_{p_i} - x_{b_i})^2 + (y_{p_i} - y_{b_i})^2 + (z_{p_i} - z_{b_i})^2} \quad (2)$$

In Equation 2, x_{p_i}, y_{p_i}, z_{p_i} are the 3D coordinates of the ith atom on the protein backbone predicted by PD2 and x_{b_i}, y_{b_i}, z_{b_i} are the 3D coordinates of the ith atom on the protein backbone predicted by BBQ.

Bond length: Two adjacent atoms form a bond on the protein backbone. There are five bond lengths in a fragment, including N-C_α, C_α-C, C=O, C-N_{next}, and N_{next}-$C_{\alpha_{next}}$. The bond length is calculated in terms of the Euclidean distance. Because bond lengths associated with BBQ and PD2 are calculated separately, it follows that 10 features are obtained.

Bond length difference: Each bond length difference is derived by the two corresponding bond lengths, predicted from the two methods.

Bond angle: Since three atoms form an angle, the bond angle can thus be obtained by *law of cosines*. Consequently, for a fragment with $L_C = 6$, we can get five different angles, including N-C_α-C, C_α-C=O, C_α-C-N_{next}, O=C-N_{next}, and C-N_{next}-$C_{\alpha_{next}}$.

Bond angle difference: Each bond angle difference is calculated from the two corresponding predicted bond angles of two methods.

Torsion angle: The torsion angle is computed by four consecutive atoms on the main chain. In addition to ϕ(C_{prev}-N-C_α-C), ψ(N-C_α-C-N) and ω(C_α-C-N_{next}-$C_{\alpha_{next}}$), we also choose the other features in our fragment, including N-C_α-C=O, O=C-N_{next}-$C_{\alpha_{next}}$, C_α-C=O-N_{next}, C_α-C-N_{next}=O (planes of C_α-C-N_{next} and C -N_{next}=O), N_{next}-C-C_α=O and C-N_{next}-$C_{\alpha_{next}}$-C_{next}. Since PD2 and BBQ are used for the computation of torsion angles, 18 features are obtained.

Torsion angle difference: This feature is obtained from the torsion angles. We compute the differences from the torsion angles obtained by PD2 and BBQ.

So far, we have defined nine kinds of features. Table 3 shows the feature names and their sizes.

The above feature extraction method is performed around the C atom in a fragment-by-fragment manner, as shown in Figure 1. We assign the C atom as the fragment center because this arrangement is most suitable for predicting both C and O atoms. Nevertheless, one may wonder whether this is also suitable when N atom is served as the prediction target. Therefore, we assign the N atom as the fragment center and perform another experiment. The experimental results (not shown in this paper) exhibit that the prediction accuracies of N atoms with the C-center window and those with the N-center window only show little difference. Thus, we still use the C-center window for predicting N atoms here.

Since the performance of models depends heavily on the selected features, we have to consider which one is relevant to the coordinate prediction. In order

Table 3. The names and sizes of all feature subsets.

Feature index	Feature name	Size
F1	Coordinate	30
F2	Coordinate difference	12
F3	Euclidean distance	4
F4	Bond length	10
F5	Bond length difference	5
F6	Bond angle	10
F7	Bond angle difference	5
F8	Torsion angle	18
F9	Torsion angle difference	9
Total		103

to identify important features, we calculate the Pearson's correlation coefficient between the objective value and each feature value. For a given feature, its correlation coefficient with the objective variable is given in Equation 3.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2)}}, \tag{3}$$

where n denotes the number of data elements, x_i denotes the ith element of data instances (x_1, x_2, ..., x_n), y_i denotes the ith instance of the objective values (y_1, y_2, ...,y_n), $\bar{x}$ and $\bar{y}$ represent the means of x and y, respectively.

3 The Coordinate Difference Prediction Method

In order to improve the predicted results, we adopt SVR to predict the x-difference, y-difference and z-difference of each of N, C and O atoms on the backbone of a target protein. Then, these differences are combined with the predicted results of BBQ and PD2 to yield our predicted coordinates. Our coordinate difference prediction procedure is described as follows.

Algorithm: The Coordinate Difference Prediction Method.

Input: 1. One training set T, containing the predicted coordinates obtained from PD2 and BBQ and the real coordinates in PDB.
2. One target protein, containing the predicted coordinates of PD2 and BBQ, along with real C_α coordinates.

Output: The predicted coordinates of N, C and O atoms along the target protein backbone.

Step 1 (Extract features): Partition the residues of all proteins in T into 20 groups, corresponding to 20 types of standard amino acids. Calculate the 103 feature values associated with each kind of residue, defined in the previous section.

Step 2 (Perform correlation analysis): For each of the nine objective values (O_x-difference, O_y-difference, O_z-difference, N_x-difference, etc.) in T, calculate the Pearson's correlation coefficient between each feature value and the objective value. Since the p-value represents the confidence level associated with its correlation coefficient, we thus can adopt a thresholding method to identify significant features.

Step 3 (Predict the difference by SVR): For each kind of objective values and amino acid groups, we use the selected features to train an SVR model. Thus, 180 models (20 kinds of residues, 3 kinds of atoms, 3D coordinates) are obtained. Then, these models are invoked to perform prediction of the target protein based on the residue and atom types.

Step 4 (Combine the predicted difference with PD2/BBQ): Combine the predicted differences with their corresponding predicted positions obtained by BBQ and PD2 to generate the final coordinates.

Step 5 (Merge all residues together): Bring the predicted coordinates of all residues together to reconstruct the 3D positions of all atoms (N, C and O atoms) on the target protein backbone.

The flow chart is shown in Figure 2.

4 Experimental Results

4.1 Experimental Procedures

For evaluating the performance of our method, we adopt CASP7, CASP8, CASP9, CASP10 and CASP11 as the experimental datasets, which contain 65, 52, 63, 39 and 55 proteins, respectively. We use only the information of chain A of proteins to carry out the experiments. If there is no chain A, the next chain is used. All features are scaled into the range of $[a,b]$ = [-1, 1] by Equation 4.

$$\frac{x_i - x_{min}}{x_{max} - x_{min}}(b - a) + a, \tag{4}$$

where x_i denotes the value of a certain feature of the ith training data element, x_{max} is the maximum value in the feature, x_{min} is the minimum value, a and b are the lower and upper bounds of the range, respectively.

In the 5-fold cross validation, each CASP is selected as the testing dataset for one time. Once a CASP is determined for testing, the remaining CASPs serve as the training datasets. For example, if we select CASP7 as the testing dataset, then the rest ones, from CASP8 to CASP11, are used as the training dataset. The testing procedure is performed for each CASP dataset.

As mentioned in the previous section, in our algorithm, Step 1 generates the feature candidates and Step 2 calculates correlation coefficients of these features and the objective values. From the correlation analysis, we find that some of the features are indeed significant.

Figures 3 shows the correlations of the 103 features and the O_x-differences of all 20 amino acids in CASP7 (excluding the protein 2G3V due to the leave-one-protein-out scheme in this example). Next, we count the number of the

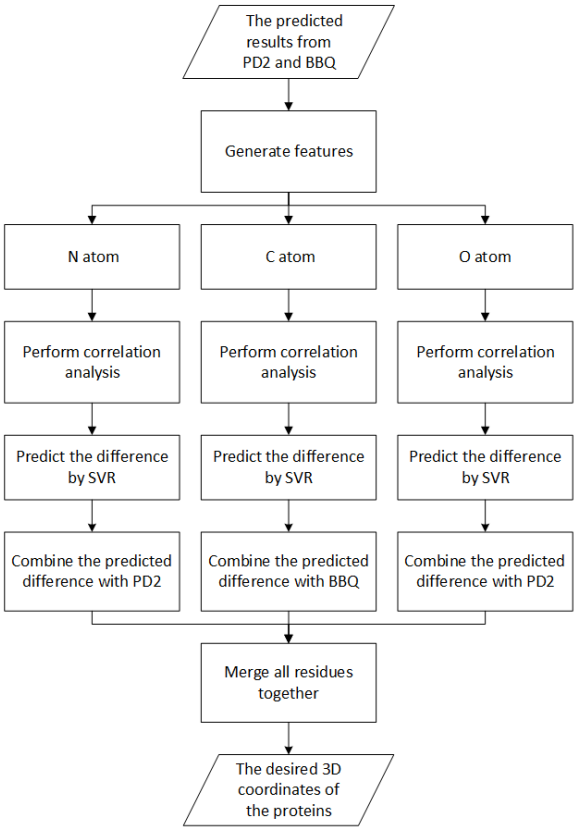

Fig. 2. The flow chart for predicting the 3D structure of a protein.

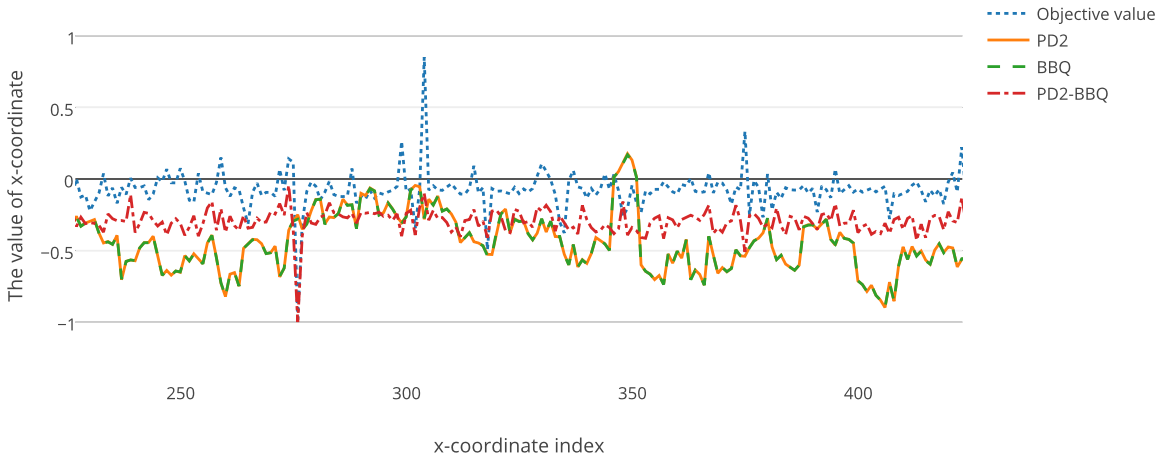

Fig. 3. The x-coordinates of O atoms of amino acid Ala in CASP7, obtained from various sources. PD2 and BBQ represent the predicted $O_x(PD2)$ and $O_x(BBQ)$, respectively. PD2-BBQ represents $O_x(PD2) - O_x(BBQ)$. The objective value is O_x-difference $(O_x(PD2) - O_x(real))$.

significance, which means that p-value (the complement of confidence level) is less than 0.05. Note that maximum count of significances is 20, since there are 20 types of amino acids. Finally, each feature is examined by some criteria to determine whether to be selected as a training feature.

Figures 4 shows the correlations of the 103 features and the O_x-differences of all 20 amino acids obtained from CASP8 to CASP11. We count the number of significances, which means that p-value (the complement of confidence level) is less than 0.05. The value of 0.05 is a widely adopted standard cut-off. It denotes that the test shows strong evidence against the null hypothesis that no correlation between the objective value and the feature value. Note that maximum count of significances is 20, because there are 20 types of amino acids. Here, if one of the following criteria is satisfied, a feature is considered as a significant one, and it is selected for training.

1. The feature has correlation value greater than or equal to 0.15 and the number of its significance counts is greater than or equal to 15.
2. The feature has correlation value greater than or equal to 0.4 and the number of its significance counts is greater than or equal to 10.

Furthermore, Figure 5 illustrates the count of significances of the 103 features with respect to O_x-differences obtained from CASP8 to CASP11. According to our criteria, we can extract three significant features (indices 34, 37 and 40), which are the predicted differences $C_x(PD2) - C_x(BBQ)$, $O_x(PD2) - O_x(BBQ)$ and $N_{next,x}(PD2) - N_{next,x}(BBQ)$, respectively. That is to say, these features highly correlate with our target O_x-differences. Similar results are revealed for y-coordinates and z-coordinates. All significant features over all CASPs are listed in Table 4.

4.2 Performance Comparison and Significance Tests

In this paper, we use the selected features to train the SVR models. We adopt LibSVM [1] to perform our experiments. In addition, we use the RBF (Radial Basis Function) kernel and set the three SVR parameters (cost, tube and hyper-parameter) to the default values.

The performance of our experimental results is evaluated by RMSD. Table 5 shows the RMSDs of individual N, C and O atoms of BBQ, PD2 and our 5-fold cross validation method. The improvement of O prediction over PD2 is about 7.7%. And the improvement of C prediction over PD2 is about 13.2%, which combines the gain from the BBQ over PD2 (about 6.4%) and the gain from ours over BBQ (about 7.8%). Table 6 shows the overall RMSDs of BBQ, PD2 and our method. Our method yields 8.03% improvement over PD2.

In order to understand whether these improvements are significant, we adopt *significance tests* for model comparison. For each testing CASP dataset, we use the remaining CASP datasets to build prediction models. Then, each sequence in the testing CASP dataset is feed into our built models and the benchmark tool, like PD2 or BBQ, for the structure prediction. These prediction results are

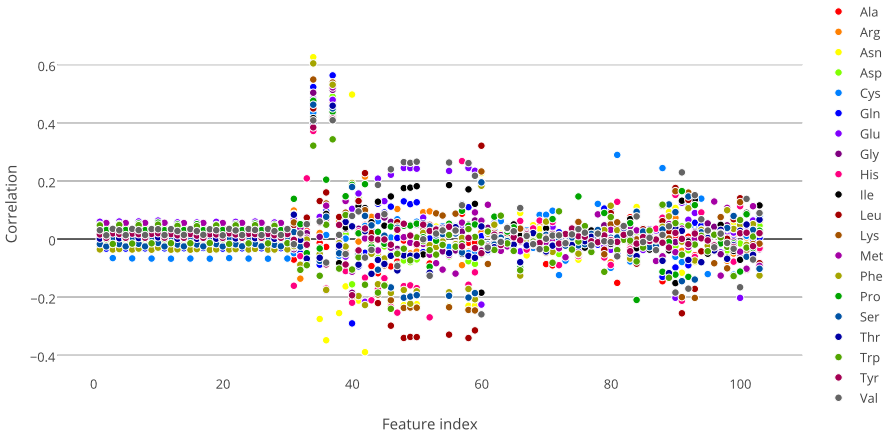

Fig. 4. The correlations of 103 features and O_x-differences of 20 amino acids in the training set consisting of CASP8 to CASP11.

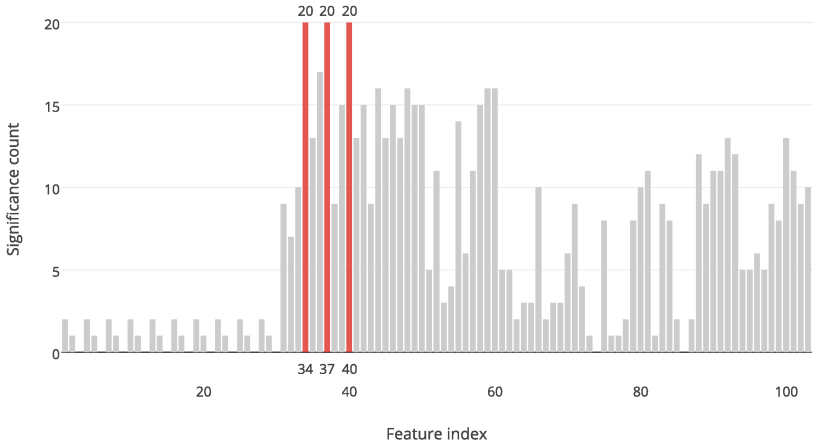

Fig. 5. The counts of significant occurrences in various features with respect to O_x-differences in the training set consisting of CASP8 to CASP11.

Table 4. The significant features selected in the nine objective values of the 5-fold cross validation experiments. The bold underlined one means that the feature is chosen in all training sets.

N_x	N_y	N_z	C_x	C_y	C_z	O_x	O_y	O_z
31, 32, **33**, **43**, 47, **52**, **57**.	31, **32**, 33, 43, 47, 52, 57.	31, **32**, **33**, **43**, **47**, **52**, **57**.	**34**, **37**, **40**.	**35**, **38**. 41, 60.	**36**, **39**, **42**. 55, 60.	**34**, **37**. 40.	**35**, **38**. 41.	**36**, **39**, **42**, 46, 48, 55, 60.

Table 5. The average RMSDs of individual atoms (N, C and O) of BBQ, PD2 and our method in CASP7, CASP8, CASP9, CASP10 and CASP11 datasets. Here, the percentage inside parentheses means the improvement over PD2.

Dataset	Method								
	BBQ			PD2			Our Method		
	N	C	O	N	C	O	N	C	O
CASP7	0.2401	0.2395	0.6347	0.1897	0.2416	0.5879	0.1853	0.2209(8.57%)	0.5443(7.42%)
CASP8	0.3775	0.2668	0.7051	0.2375	0.3057	0.6968	0.2336	0.2486(18.68%)	0.6415(7.94%)
CASP9	0.4229	0.2569	0.6632	0.2462	0.2864	0.6115	0.2434	0.2367(17.35%)	0.5672(7.24%)
CASP10	0.2429	0.2444	0.6513	0.2042	0.2298	0.5991	0.2015	0.2227(3.09%)	0.5515(7.95%)
CASP11	0.2992	0.2310	0.6252	0.2197	0.2525	0.5994	0.2144	0.2127(15.76%)	0.5497(8.29%)
Arithmetic mean	0.3165	0.2477	0.6449	0.2195	0.2632	0.6189	0.2157	0.2283(12.69%)	0.5708(7.77%)
Weighted arithmetic mean	0.3205	0.2477	0.6482	0.2198	0.2646	0.6179	0.2161	0.2284(13.17%)	0.5701(7.73%)

Table 6. The RMSDs of all atoms of BBQ, PD2 and our method for CASP7, CASP8, CASP9, CASP10 and CASP11 datasets.

Dataset	Method			Improvement over PD2 %
	BBQ	PD2	Ours	
CASP7	0.3632	0.3335	0.3102	6.99%
CASP8	0.4412	0.4030	0.3672	8.88%
CASP9	0.4344	0.3635	0.3341	8.09%
CASP10	0.3726	0.3386	0.3159	6.70%
CASP11	0.3775	0.3491	0.3166	9.31%
Arithmetic mean	0.3978	0.3575	0.3288	7.99%
Weighted arithmetic mean	0.3986	0.3574	0.3286	8.03%

compared with the ground truths to calculate RMSDs. The RMSD difference ΔRMSD, obtained from the benchmark tool and our models, associated with each sequence is collected and used for the significance tests.

We adopt Wilcoxon signed-rank test [23,15] for performance comparison because it makes less assumptions, like normal distribution and similar variance (also known as *homoscedastic*), than *parametric* methods do. The confidence level of the interval is set to 0.95, which is a widely used parameter. The test gives two types of results: p-value and confidence interval. The p-value is the probability to reject the null hypothesis, which means that there exists no systematic difference in the RMSDs between prediction models. The smaller the p-value, the stronger the test rejects the null hypothesis. The significance can also be determined by checking whether the 95% confidence interval covers 0. If the entire confidence interval is above or below 0, it also suggests that there exists significant difference between prediction models. The test results are shown in Figure 6 and Figure 7. It shows that RMSDs associated with our models are significantly smaller. This further suggests that our prediction methods, which incorporates information from other prediction methods, can improve performance significantly.

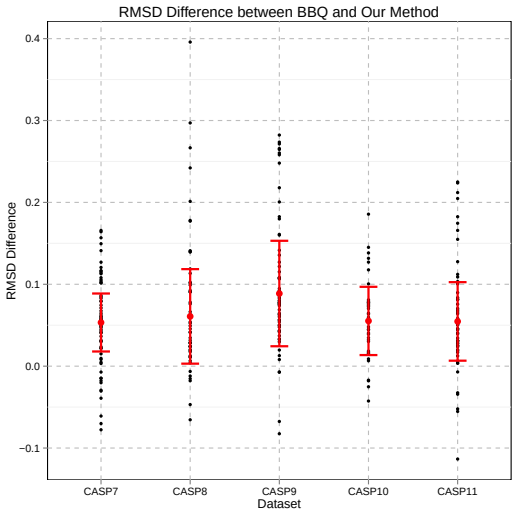

Fig. 6. The RMSD differences of BBQ and our method, where each dot is a RMSD difference associated with a sequence and the error bar denotes 95% confidence interval obtain form Wilcoxon tests.

5 Conclusion

In the past decades, lots of efforts have been devoted to the study of the protein backbone reconstruction problem. Until now, the methods, such as BBQ, PD2

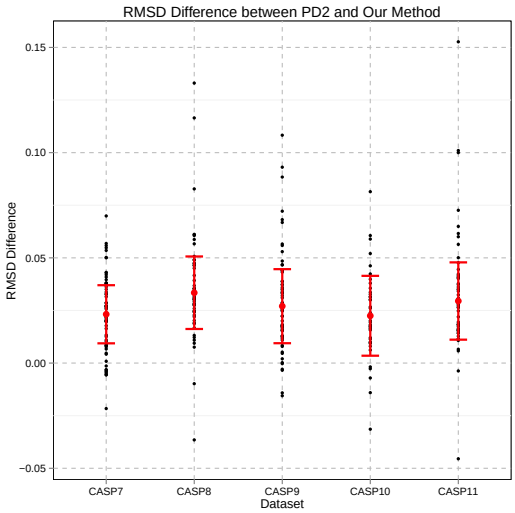

Fig. 7. The RMSD differences of PD2 and our method, where each dot is a RMSD difference associated with a sequence and the error bar denotes 95% confidence interval obtain form Wilcoxon tests.

and so on, have already been successfully developed and applied to the problem. Since these methods have their individual strengths and weakness, the prediction accuracy can thus be improved by taking advantage of their strengths.

In this paper, we propose an algorithm to refine the 3D coordinates of all atoms on a protein backbone with SVR. The objective values of our prediction is the differences between the predicted coordinates and the real ones. We first define a set of feature candidates extracted from the predicted coordinates of BBQ and PD2. It is well-known that the key factor to affect the prediction performance is the feature relevance. Thus, we perform the correlation analysis to identify significant features. The experimental datasets range from CASP7 to CASP11. As the experimental results show, the three most significant features for predicting the O_x-differences and C_x-differences are the differences of the predicted x-coordinates of PD2 and BBQ in C, O and N_{next} atoms. Similar results are exhibited for y-coordinates and z-coordinates. In summary, our method yields about 8% improvement in RMSD over PD2, which is the best previous predictor in this problem up to now.

Acknowledgments

This research work was partially supported by the Ministry of Science and Technology of Taiwan under contract MOST 104-2221-E-110-018-MY3. A preliminary version of this paper was presented at the 16th Industrial Conference on Data Mining [6]

References

1. C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, Vol. 2, No. 3, pp. 27:1–27:27, 2011.
2. H. Y. Chang, C. B. Yang, and H. Y. Ann, "Refinement on O atom positions for protein backbone prediction," *Proceedings of the 2nd WSEAS International Conference on Biomedical Electronics and Biomedical Informatics (BEBI '09)*, Moscow, Russia, pp. 99–104, 2009.
3. K. Y. Chen, C. B. Yang, and K. S. Huang, "Prediction of protein backbone structure by preference classification with SVM," *Proceedings of the 9th International Conference on Information Systems and Technology Management*, Sao Paulo, Brazil, pp. 1193–1206, 2012.
4. 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, pp. 1593–1597, 2007.
5. L. Holm and C. Sander, "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*, Vol. 21, No. 1, pp. 183–194, 1991.
6. D. Y. Huang, C. Y. Hor, and C. B. Yang, "Coordinate refinement on all atoms of the protein backbone with support vector regression," *16th Industrial Conference on Data Mining*, p. 1, 2016.
7. D. E. James U. Bowie, Roland Luthy, "A method to identify protein sequences that fold into a known three-dimensional structure," *Science*, Vol. 253, pp. 164–170, 1991.
8. W. Kabsch, "A solution for the best rotation to relate two sets of vectors," *Acta Crystallographica Section A*, Vol. 32, pp. 922–923, 1976.
9. W. Kabsch, "A discussion of the solution for the best rotation to relate two sets of vectors," *Acta Crystallographica Section A*, Vol. 34, pp. 827–828, 1978.
10. 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, pp. 715–723, 2002.
11. N. Krasnogor, W. E. Hart, J. Smith, and D. A. Pelta, "Protein structure prediction with evolutionary algorithms," *Proceedings of the Genetic and Evolutionary Compution Conference*, Orlando, USA, pp. 1596–1601, 1999.
12. 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.
13. B. L. Moore, L. A. Kelley, J. Barber, J. Murray, and J. T. MacDonald, "High-quality protein backbone reconstruction from alpha carbons using Gaussian mixture models," *Journal of Computational Chemistry*, Vol. 34, pp. 1881–1889, 2013.
14. J. Moult, K. Fidelis, A. Kryshtafovych, B. Rost, and A. Tramontano, "Critical assessment of methods of protein structure prediction (CASP) x Round IX," *Proteins*, Vol. 79, pp. 1–5, 2011.
15. R Development Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, 2008.
16. P. Rotkiewicz and J. Skolnick, "Fast procedure for reconstruction of full-atom protein models from reduced representations," *Journal of Computational Chemistry*, Vol. 29, pp. 1460–1465, 2008.

17. I. Ruczinski, C. Kooperberg, R. Bonneau, and D. Baker, "Distribution of beta sheets in proteins with application to structure prediction," *Proteins: Structure, Function, and Genetics*, Vol. 48, pp. 85–97, 2008.
18. A. J. Smola and B. Scholkopf, "A tutorial on support vector regression," *Statistics and Computing*, Vol. 14, pp. 199–222, 2004.
19. V. Vapnik, S. E. Golowich, and A. Smola, "Support vector method for function approximation, regression estimation, and signal processing," *Advances in Neural Information Processing Systems 9*, pp. 281–287, MIT Press, 1996.
20. J. H. Wang, C. B. Yang, and C. T. Tseng, "Reconstruction of Protein Backbone with the α-Carbon Coordinates," *Journal of Information Science and Engineering*, Vol. 26, No. 3, pp. 1107–1119, 2010.
21. H. F. Wu, C. B. Yang, C. Y. Hor, Y. H. Peng, and K. T. Tseng, "Protein backbone reconstruction with tool preference classification for standard and nonstandard proteins," *Proceedings of the 12th Conference on Information Technology and Applications in Outlying Islands*, Kingmen, Taiwan, pp. 175–182, 2013.
22. H. C. Yuan, "A survey of computational methods for protein structure prediction," *Master's Thesis*, National Sun Yat-sen University, Kaohsiung, Taiwan, July, 2015.
23. J. H. Zar, *Biostatistical Analysis*. Prentice Hall, 2009.