Protein Backbone Reconstruction with Tool Preference Classification for Standard and Nonstandard Proteins *

Hsin-Fang Wu[1], Chang-Biau Yang[1]†, Chiou-Yi Hor[1],
Yung-Hsing Peng[2] and Kuo-Tsung Tseng[3]

[1]Department of Computer Science and Engineering
National Sun Yat-sen University, Kaohsiung, Taiwan
[2]Innovative DigiTech-Enabled Applications & Services Institute
Institute for Information Industry, Kaohsiung, Taiwan
[3]Department of Information Management
Fooyin University, Kaohsiung, Taiwan

Abstract

Given a protein sequence and the C_α coordinates on its backbone, the all-atom protein backbone reconstruction problem (PBRP) is to reconstruct the backbone by its 3D coordinates of N, C and O atoms. In the past few decades, many methods have been proposed for solving PBRP. Related research reveals that if proper prediction tools are selected to build the 3D coordinates of the desired atoms, the RMSD may be improved. In this paper, we propose a method for solving PBRP, performing tool preference classification on each atom of the residue, where the classification model is generated by support vector machine (SVM). We rebuild the backbone by combining the prediction results of all atoms in all residues. The data sets used in our experiments are CASP7, CASP8, and CASP9, which contain 65, 52, and 63 proteins, respectively. These data sets contain nonstandard amino acids along with standard ones. The RMSDs we achieve are 0.3496 in CASP7, 0.3084 in CASP8, and 0.3286 in CASP9.

*This research was partially supported by the National Science Council of Taiwan under contract NSC 100-2221-E-242-003. This research was also partially supported by the "Project Digital Convergence Service Open Platform" of the Institute for Information Industry which is subsidized by the Ministry of Economy Affairs of the Republic of China.

†Corresponding author. E-mail: cbyang@cse.nsysu.edu.tw (Chang-Biau Yang).

Keywords: Protein, Backbone Reconstruction, Classification

1 Introduction

For a given protein sequence and its α-carbon coordinates, the all-atom *protein backbone reconstruction problem* (PBRP) is that of reconstructing the 3D coordinates of major atoms (N, C, and O) on the backbone. Several methods were proposed for solving PBRP, such as SABBAC [17], Wang's method [22], Chang's method [3], BBQ (backbone building from quadrilaterals) [12] and Chen's method [5].

SABBAC is a famous online service for rebuilding the protein backbone from α-carbon trace. It selects assembly of fragments by encoding the protein trace with the structural alphabets derived by a hidden Markov model. Wang *et al.* [22] proposed an effective method based on homology modeling method to rebuild the full atom protein backbone with known α-carbon coordinates. According to the literature, the method is faster than SABBAC. Based on Wang's method, Chang *et al.* [3] modified the energy function and adopted a two-phase refinement method to refine the positions of O atoms. The method not only achieves higher accuracy than Wang's method, but is also faster than SABBAC. BBQ is another algorithm for PBRP, which is also prominent in efficiency and accuracy. This method manages to derive a custom database of quadrilaterals by extract-

ing quadrilaterals from PDB files and computing average positions of atoms C, N and O. Since each of the above methods has its individual merit and weakness, Chen *et al.* [5] thus applied the tool preference classification strategy to determine which tool is the more suitable one for predicting each atom in the protein. In this paper, we employ Chen's method as an initial stage for prediction. We apply the prediction tool, either BBQ or Chang's method, for predicting atoms in each residue in order to achieve higher accuracy.

The rest of this paper is organized as follows. In Section 2, we will describe some previous works. In Section 3, we propose our method for protein backbone reconstruction. In Section 4, we show the experimental results of our method. The conclusion of this paper is given in Section 5.

2 Previous Work

Over the past few years, several studies have been published for solving the PBRP, including SABBAC [17], Wang's method [22], Chang's method [3], BBQ [12], and Chen's method [5]. In the following, we briefly describe these methods.

2.1 SABBAC

SABBAC [17] is an on-line service committed to protein backbone reconstruction from α-carbon trace. First, it encodes the α-carbon trace in the hidden Markov model for generating the collection of fragments. 155 fragments are used to describe the 27 letters of the alphabet; each fragment is assigned one letter to describe its conformation. SABBAC follows the procedure of Milik *et al.* [18] to calculate the coordinates from three consecutive α-carbons. Then it uses a greedy method to search for the optimal combination of fragments. To guide the search, they use the scoring function of the OPEP force field [21]. The execution time of SABBAC is known to range from a few seconds to tens of minutes according to the length of the protein sequence. SABBAC can be accessed at the website http://bioserv.rpbs.jussieu.fr/cgi-bin/SABBAC.

2.2 Wang's Method

Wang *et al.* [22] proposed a method to solve the PBRP based on homology modeling. First, they extract all consecutive four-residue fragments from all proteins in PDB. Suppose the length of protein is L, they obtain $L-3$ fragments. The fragments can be classified into 8000 residue groups by identifying each fragment with its second, third and fourth residues. The fragments with similar structures are clustered into the same residue group, and one typical fragment is chosen to be the representative in each cluster. These fragments form the fragment library. Then, for each target fragment, Wang *et al.* find its most similar typical fragment in the residue group with DRMSD as the measurement. They rotate the typical fragment to superimpose the target fragment, and calculate their coordinates at the center residue of the target fragment until all target fragments' coordinates are found. The achieved accuracy is comparable to most previous works.

2.3 Chang's Method

Chang *et al.* [3] found that the result of Wang *et al.* [22] can be improved by refining the 3D coordinates of O atoms. They computed the initial coordinates of N, C and O atoms by Wang's method, then tried to refine the O atoms based on the energy function modified from the AMBER force field [6]. They found that the average energy of a real protein backbone structure is smaller than that of the predicted one. They defined coarse moving scope as the boundary of the cube centered at the initial O position, and defined resolution as the number of grid points on each side of the cube. Each grid point represents one candidate position of the predicted O atom. The entire procedure is divided into two phases, thus it is named the two-phase refinement method. In the first phase, they selected several candidate positions, each of which serves as the center of one fine cube bounded by the fine moving scope. Then in the second phase, they examined all possible positions in order to find the structure with minimal energy in each fine cube. For each candidate position, a scoring function which only considers O atoms as the bonded potential energy is calculated. The lower the score is, the better the position is. They compared their experimental results with MaxSprout [13], Adcock's method [1], SABBAC [17], and Wang's method [22]. More than half of their results are better than those in the previous works. In addition, the execution efficiency is better than SABBAC.

2.4 BBQ

The BBQ (backbone building from quadrilaterals) method [12] is an algorithm for PBRP which comprises high computational efficiency and accuracy. First, it defines two coordinate systems, *R*-coordinate and *L*-coordinate. *R*-coordinate is used to define the protein fragment of four amino acids, and *L*-coordinate is used to define a local Cartesian coordinate system with the given C_α as the center. In this method, a consecutive fragment of four α-carbons is defined as a quadrilateral. BBQ keeps the quadrilaterals with R-factor below 50, and then it calculates the local coordinates that form the central peptide plate between the second and third C_α atoms. BBQ also discretizes the continuous space described by *R*-coordinates. In the three dimensional look-up table, it holds 22,680 different quadrilaterals. BBQ computes average positions for the N, C, and O atoms for each state. In some rare cases, it cannot find a specific combination of *R*-coordinates from the training set; BBQ will inspect the neighborhood of a given element of the grid, and obtain proper coordinates of N, C and O atoms from the look-up table. Gront *et al.* [12] compared their results with another four PBRP tools and showed that among these five tools, BBQ is the most accurate. Although other existing algorithms are relatively fast, they are significantly less accurate than BBQ.

2.5 Chen's Method

Chen *et al.* [5] proposed a method that utilizes a tool preference classification to determine which prediction tool is more suitable for predicting the coordinates of N, C and O atoms in every protein. The method first splits the atoms of the target protein into three parts: N, C and O atoms. In each part, the method chooses the most appropriate tool to be used, which is either BBQ [12] or Chang's method [3]. Chen's method adopts nine features, including hydrophobicity, van der Waals volume, polarity, polarizability, size, charge, molecular weight, isoelectric point, and accessible surface area. Each of these nine features can be further divided into three groups. For example, the size feature can be divided into tiny, small and normal, and the charge can be divided into positive, neutral and negative type. In addition, Chen *et al.* consider three descriptors which are composition, transition, and distribution proposed by Dubchak *et al.* [7], and they combine these descriptors with the above features. In order to obtain better feature combinations, they divide each feature vector into two parts, the former one with 6 elements and the latter one with 15 elements. They reorganize these feature sets by crossover and extension operations. Finally, the results of SVM are used to determine the suitable tool for predicting the coordinates of a specific atom, and the coordinates of all atoms are combined to form the backbone.

3 Our Prediction Method

In the method of Chen *et al.* [5], Chang's method and BBQ are selected as the preference prediction tools. They compared the prediction performance of SABBAC, PULCHRA [20], Chang's method, and BBQ with the experiments on CASP7, CASP8 and CASP9 [19]. They found that Chang's method and BBQ are two most competent algorithms in terms of RMSDs. In order to improve the accuracy of protein structure prediction, in this paper we attempt to perform preference prediction on a residue-by-residue manner.

3.1 A New Method for Preference Tool Selection

Because Chen's method works only for standard amino acids, we transform nonstandard amino acids to standard ones according to the residue substitution table derived from Ligand Expo [10]. Our method for preference tool selection is described as follows:

Input: A protein sequence with its C_α coordinates.

Output: The complete protein backbone, including the coordinates of N, C, and O atoms of all residues.

Preprocess: For a protein containing nonstandard amino acids, transform each nonstandard amino acid into standard one according to the residue substitution table. Then, generate the FASTA file of the protein sequence, and obtain the PSSM file by invoking PSI-BLAST. Finally, perform the process of feature extraction (see Section 3.2).

Step 1: Divide the input protein chain into smaller groups in terms of the user-specified

atom parameter and amino acid type parameter. For example, if the specified atom parameter contains three elements (N, C, O), and the amino acid type parameter contains twenty standard amino acids, then each protein will be divided into sixty groups.

Step 2: For each atom in the protein, use SVM with the specified feature parameter to select the preferred prediction tool.

Step 3: Use the selected prediction tool to predict the 3D coordinate of each atom on the backbone.

Step 4: Combine prediction results of all atoms and output the 3D coordinates of the backbone of the target protein.

The PSSM file is obtained by invoking PSI-BLAST with a FASTA format file as the input. In addition, instead of performing tool selection in a protein-wise manner, our tool selection is based on each residue. The flow chart of our method is shown in Figure 1. The input is a protein sequence, along with the atom parameter $p_1 \in$\{N, C, O\}, the amino acid type parameter $p_2 \in$\{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}, and the feature parameter $p_3 \in \{F_1, F_2, F_3, F_4, F_5, F_6, F_7, F_8, F_9, F_{10}, F_{11}, F_{12}, F_{13}\}$, where F_i is defined in Table 1 (See Section 3.2). For example, $(p_1, p_2, p_3) =$ (C, R, F_5) means that for atom C in amino acid R, we select the prediction tool by SVM with the hydrophobic values around the residue as features.

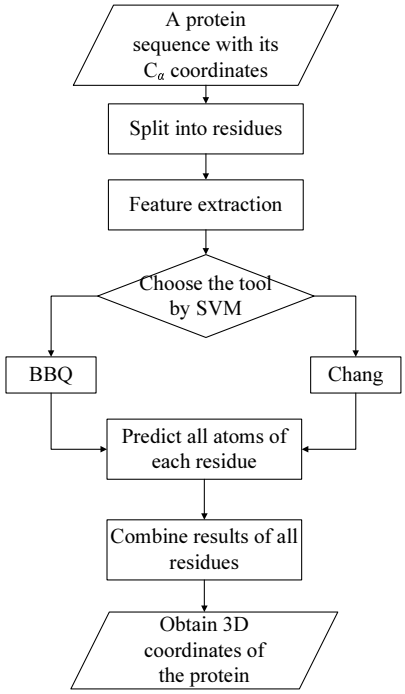

Figure 1: The flow chart of our method.

3.2 Feature Extraction

In order to obtain accurate result from the SVM model, effective features are required for training the classifier. In the following, we describe our approach for feature extraction. Let L be the length of protein, f_σ be the number of occurrences of amino acid σ in the protein, $ami(j)$ be the type of amino acid at position j, and $PSSM_{k,j}$ be the submatrix of size $(2k+1) \times 20$ in the PSSM matrix, where the residue at position j is considered as the center.

Our feature sets include thirteen feature subsets, and each of them is computed with respect to one residue.

1. The frequency index $\frac{j}{f_\sigma}$.
2. The normalized position $\frac{j}{L}$.
3. The central index $\frac{j}{f_{ami(j)}}$. This feature is an alternative to the frequency index.
4. PSSM for the central residue: $PSSM_{k,j}$ with k=12.
5. Hydrophobicity [16] of the residue.
6. Normalized van der waals volume [9] of the residue.
7. Polarity [11] of the residue.
8. Polarizability [4] of the residue.
9. Size [2] of the residue.
10. Charge [15] around the residue.
11. Molecular weight [8] of the residue.
12. Isoelectric point [23] of the residue.
13. Accessible surface area [14] of the residue.

For feature subsets F_5 to F_{13}, the half window size k is also set to 12. Table 1 shows the names and sizes of all feature subsets.

Table 1: The names and sizes of the feature subsets.

Feature	Description	Size
F_1	Frequency index	20
F_2	Normalized position	1
F_3	Central index	1
F_4	PSSM	500
F_5	Hydrophobic	25
F_6	Normalized van der waals volume	25
F_7	Polarity	25
F_8	Polarizability	25
F_9	Size	25
F_{10}	Charge	25
F_{11}	Molecular weight	25
F_{12}	Isoelectric point	25
F_{13}	Accessible surface area	25

4 Experimental Results

In this section, we will show our experimental results, and explain the procedure of our experiments in detail. The data sets of protein sequences we use are acquired from CASP7, CASP8 and CASP9. We adopt 65 proteins in CASP7, 52 proteins in CASP8, and 63 proteins in CASP9, where only chain A information is used. We perform nine experiments, including three self-tests and six independent tests. That is, we use CASPi as the training data sets and CASPj as the testing data sets, where $i, j \in \{7, 8, 9\}$.

Table 2 shows the average RMSDs of our method and other methods. We list five different feature combinations in Table 2, including $S_1 = \{F_1, F_2, F_4, F_5\}$, $S_2 = \{F_1, F_2, F_5\}$, $S_3 = \{F_1\}$, $S_4 = \{F_5\}$, and $S_5 = \{F_2, F_3, F_4, F_5, F_6, F_7, F_8, F_9, F_{10}, F_{11}, F_{12}, F_{13}\}$. The term "PAAR" (perfect for all atoms per residue) and "PIAR" (perfect for individual atoms per residue) in the table mean the proper tool is selected every time for predicting all atoms in each residue and individual atoms of each residue, respectively. In other words, "PAAR" and "PIAR" are the lower bounds of RMSDs achievable by tool preference classification. In Table 2, the lowest achieved RMSD is marked by an underline.

Cases I, V and IX are self-tests. In the self-test experiments, the proteins in the training set are fully identical to those in the testing set. The goal of these experiments is to determine whether the model we generate from the training set is appropriate or not. Cases II, III, IV, VI, VII and VIII are independent tests. In the independent test, proteins in the testing set do not appear in the training set. Both self and independent experiments help us identify whether the model is appropriate or not. In our experiment, we test different feature sets, and we find that some feature sets can achieve lower RMSDs than that of the method proposed by Chen *et al.*

For each atom, we calculate its RMSDs, and assign it a class label corresponding to the preferred software. The labels we use are Chang and BBQ. Table 3 shows the accuracies of our nine experiments on preference classification. The accuracy is calculated by the following equation:

$$Q = \frac{P}{N} = \frac{TP + TN}{TP + TN + FP + FN}. \quad (1)$$

In Equation 1, P denotes the total number of correct predictions, and N denotes the number of total predictions. TP, TN, FP, and FN represent the numbers of true positive, true negative, false positive, and false negative, respectively. Based on the above definition, the accuracy for selecting appropriate prediction tool is

$$Q_R = \frac{P_R}{N_R}, \quad (2)$$

where P_R represents the number of residues with correctly predicted preference in the testing set, and N_R means the total number of residues in the testing set. In Table 4, we show the detailed RMSDs of Chang, BBQ, and our method for CASP9. The detailed RMSDs for CASP7 and CASP8 are omitted for conciseness.

We find that if we can always select the correct tool to do the prediction for each individual atom, the average RMSDs would be 0.2485, 0.3078, and 0.2651 for CASP7, CASP8, and CASP9, respectively, which are the lower bounds of the RMSDs in our experiments. The best average RMSDs we achieve are 0.3496, 0.3084 and 0.3286 for self-test in our method. The difference of RMSDs between PIAR and our method shows that there are still rooms to improve our method. In these nine experiments, we have results better than Chen's method in Cases I, III, IV, V, VII, and IX. However, in Cases II, VI, and VIII, Chen's method yields better results. In summary, our method is better than Chen's for six out of nine experiments, which include three self tests and three independent tests. This implies that our method is at least as good as Chen's.

Table 2: The RMSDs of nine experiments with various methods.

Case	Train	Test	Chang	BBQ	Chen	S_1	S_2	S_3	S_4	S_5	PAAR	PIAR
I	CASP7	CASP7			0.3505	0.3553	0.3496	0.3580	0.3590	0.3499		
II	CASP8	CASP7	0.4108	0.3624	0.3589	0.3632	0.3646	0.3593	0.3611	0.3605	0.2625	0.2485
III	CASP9	CASP7			0.3609	0.3617	0.3642	0.3621	0.3646	0.3586		
IV	CASP7	CASP8			0.4558	0.4413	0.4422	0.4557	0.4432	0.4599		
V	CASP8	CASP8	0.4888	0.4584	0.4106	0.3084	0.4047	0.4474	0.4267	0.4185	0.3227	0.3078
VI	CASP9	CASP8			0.4187	0.4607	0.4542	0.4365	0.4394	0.4515		
VII	CASP7	CASP9			0.4127	0.4000	0.4082	0.4139	0.4091	0.4172		
VIII	CASP8	CASP9	0.4406	0.4280	0.3757	0.4246	0.4156	0.4083	0.4044	0.4248	0.2794	0.2651
IX	CASP9	CASP9			0.3693	0.3960	0.3993	0.3811	0.3286	0.4322		

Table 3: The accuracies of the preference classifications in our nine experiments.

Case	Training Set	Testing Set	Accuracy
I	CASP7	CASP7	61.05%
II	CASP8	CASP7	58.03%
III	CASP9	CASP7	57.94%
IV	CASP7	CASP8	58.27%
V	CASP8	CASP8	99.94%
VI	CASP9	CASP8	58.04%
VII	CASP7	CASP9	57.61%
VIII	CASP8	CASP9	57.22%
IX	CASP9	CASP9	81.60%

5 Conclusion

In this paper, we propose a method for protein backbone reconstruction, which performs tool preference classification on each residue in the target protein. The prediction tools we use are Chang's method and BBQ. We split the protein sequence into residues, and select the tool by means of SVM with various feature combinations. The backbone can be reconstructed by combining the prediction results from all residues. The achieved RMSDs for our method, Chen's method, and BBQ are 0.3496, 0.3505, and 0.3624 in CASP7, respectively. For CASP 8, the RMSDs for these three methods are 0.3084, 0.4106 and 0.4584. The RMSDs in CASP9 are 0.3286, 0.3693 and 0.4280. In nine experiments, we achieve better RMSDs in six cases, while the other three cases are worse than Chen's results.

Our future work will focus on how to improve the prediction accuracy. This can be achieved by several possible ways. First, we can try to devise other features which may facilitate the SVM classification. We observe that if we can improve the accuracy for predicting O atom, the RMSD can be reduced significantly. Next, instead of utilizing SVM to obtain discrete outputs, we may adopt SVR (support vector regression), which can export continuous outputs to predict the coordinates of N, C and O atoms directly. Finally, we may incorporate some other backbone prediction tools with RMSD lower than Chang's method and BBQ.

References

[1] S. A. Adcock, "Peptide backbone reconstruction using dead-end elimination and a knowledge-based forcefield," *Journal of Computational Chemistry*, Vol. 25, pp. 16–27, 2004.

[2] D. Brock and O. Mayo, *Biochemical genetics of man.* Academic Press, New York, 1972.

[3] 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.

[4] M. Charton and B. I. Charton, "The structural dependence of amino acid hydrophobicity parameters," *Journal of Theoretical Biology*, Vol. 99, pp. 629–644, 1982.

[5] 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.

[6] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, J. K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, 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*

Table 4: The RMSDs of Chang's method, BBQ, and our method in CASP9.

CASP9 ID	PDB ID	Chang	BBQ	Case I, S_1	Case II, S_4	Case III, S_4	PAAR	PIAR
T0572	2KXY	0.4003	0.4719	0.4724	0.4734	0.3058	0.3064	0.2906
T0557	2KYY	0.5606	0.5002	0.4876	0.4943	0.3602	0.3522	0.3413
T0549	2KZV	0.4945	0.4991	0.4988	0.4993	0.3806	0.3523	0.3430
T0559	2L01	0.4128	0.3514	0.3450	0.3437	0.3181	0.3208	0.3135
T0560	2L02	0.5468	0.3823	0.3816	0.3859	0.3662	0.3486	0.3381
T0555	2L06	0.5165	0.3750	0.3623	0.3788	0.2946	0.2851	0.2717
T0538	2L09	0.4293	0.3309	0.3252	0.3245	0.2989	0.3093	0.2980
T0539	2L0B	0.5520	0.3846	0.3899	0.3808	0.2922	0.2987	0.2822
T0540	2L0D	0.5496	0.3953	0.3982	0.3959	0.3068	0.3194	0.2943
T0552	2L3B	0.4772	0.4214	0.4200	0.4225	0.3392	0.3035	0.2992
T0545	2L3F	0.4711	0.4051	0.4042	0.4029	0.2916	0.2516	0.2470
T0546	2L5Q	0.6070	0.4991	0.4996	0.5016	0.4063	0.3974	0.3901
T0577	2L7Q	0.6411	0.4278	0.4413	0.4288	0.3807	0.3716	0.3619
T0554	2L8V	0.5709	0.4604	0.4539	0.4620	0.3752	0.3723	0.3611
T0636	2X3O	0.4168	0.4468	0.4554	0.4517	0.3375	0.3359	0.3200
T0535	3MPX	0.4530	0.6218	0.6193	0.5001	0.3973	0.3324	0.3088
T0525	3MQZ	0.4045	0.3013	0.3055	0.3058	0.2271	0.2276	0.2201
T0527	3MR0	0.4358	0.5222	0.5310	0.5326	0.3317	0.3302	0.3155
T0521	3MSE	0.5005	0.6477	0.3671	0.6383	0.3493	0.3062	0.2975
T0532	3MWB	0.3828	0.3373	0.3468	0.3391	0.2596	0.2397	0.2362
T0524	3MWX	0.4703	0.3407	0.3500	0.3457	0.2511	0.2386	0.2339
T0533	3MX3	0.4101	0.3319	0.3240	0.3173	0.2427	0.2350	0.2285
T0536	3MXQ	0.4589	0.6396	0.6338	0.6310	0.2541	0.2458	0.2365
T0542	3N05	0.4244	0.4922	0.3841	0.4938	0.4127	0.2744	0.2694
T0528	3N0X	0.3851	0.2648	0.2745	0.2645	0.1968	0.1893	0.1831
T0635	3N1U	0.4224	0.3070	0.3121	0.3098	0.2421	0.2520	0.2245
T0587	3N2W	0.4144	0.7503	0.7498	0.7500	0.7118	0.2498	0.2379
T0634	3N53	0.5089	0.7011	0.6165	0.7009	0.4950	0.3807	0.3608
T0567	3N70	0.4112	0.4118	0.4129	0.4130	0.2776	0.2816	0.2668
T0585	3NE8	0.4048	0.6775	0.3381	0.6787	0.6398	0.2426	0.2263
T0589	3NET	0.4783	0.5258	0.4188	0.3856	0.4654	0.3296	0.3111
T0593	3NGW	0.3792	0.3162	0.2666	0.2715	0.1978	0.2267	0.1923
T0597	3NIE	0.4599	0.8070	0.4096	0.4134	0.7797	0.2989	0.2875
T0600	3NJA	0.4105	0.3644	0.3645	0.3649	0.2362	0.2359	0.2141
T0603	3NKD	0.4241	0.4282	0.3096	0.3092	0.2774	0.2774	0.2532
T0623	3NKH	0.4067	0.3584	0.3150	0.3137	0.2337	0.2541	0.2090
T0518	3NMB	0.4139	0.3847	0.3797	0.3799	0.2702	0.2652	0.2491
T0548	3NNQ	0.3403	0.3035	0.3351	0.3347	0.2147	0.2367	0.2060
T0611	3NNR	0.4323	0.4160	0.2593	0.4089	0.3882	0.2836	0.2113
T0570	3NO3	0.3889	0.3388	0.4062	0.4077	0.3298	0.2400	0.3224
T0516	3NO6	0.3250	0.2311	0.2398	0.2343	0.1942	0.2019	0.1819
T0565	3NPF	0.4131	0.3415	0.3431	0.3414	0.2667	0.2694	0.2558
T0530	3NPP	0.4312	0.4078	0.4090	0.4092	0.3275	0.3316	0.3075
T0615	3NQW	0.3902	0.3599	0.3596	0.3614	0.2668	0.2662	0.2559
T0620	3NR8	0.5272	0.4987	0.4963	0.4845	0.3830	0.3379	0.3191
T0591	3NRA	0.4044	0.6305	0.6303	0.6309	0.5588	0.2483	0.2306
T0522	3NRD	0.5521	0.4168	0.4107	0.4209	0.3588	0.3562	0.3392
T0575	3NRG	0.2438	0.1971	0.1941	0.1975	0.1683	0.1682	0.1659
T0617	3NRV	0.5022	0.6507	0.6401	0.5021	0.5790	0.3878	0.3299
T0632	3NWZ	0.4081	0.3306	0.3511	0.3544	0.2771	0.2651	0.2626
T0641	3NYI	0.3236	0.3374	0.3359	0.3367	0.2203	0.2301	0.2145
T0640	3NYW	0.4217	0.5794	0.3273	0.3162	0.5317	0.2552	0.2319
T0612	3O0L	0.3822	0.3296	0.3365	0.3339	0.2438	0.2554	0.2384
T0626	3O1L	0.4079	0.2837	0.3004	0.2828	0.2243	0.2148	0.1985
T0551	3OBH	0.4849	0.3096	0.3267	0.3276	0.2835	0.2589	0.2597
T0613	3OBI	0.3348	0.3320	0.3362	0.3358	0.2043	0.1970	0.1865
T0563	3ON7	0.4512	0.4127	0.4152	0.3589	0.3291	0.2675	0.2442
T0573	3OOX	0.4353	0.3392	0.3635	0.3664	0.3165	0.2807	0.2674
T0599	3OS6	0.4544	0.6924	0.6960	0.4050	0.2985	0.2882	0.2789
T0609	3OS7	0.4260	0.3733	0.3465	0.3628	0.2357	0.2485	0.2212
T0636	3P1T	0.4073	0.3497	0.3360	0.3353	0.2300	0.2265	0.2225
T0607	3PFE	0.3902	0.3268	0.3465	0.3297	0.2358	0.2179	0.2082
T0601	3QTD	0.3715	0.2907	0.2937	0.2912	0.2351	0.2290	0.2243
Average		0.4406	0.4280	0.4000	0.4044	0.3286	0.2794	0.2651

Chemical Society, Vol. 117, pp. 5179–5197, 1995.

[7] I. Dubchak, I. Muchnik, S. R. Holbrook, and S. H. Kim, "Prediction of protein folding class using global description of amino acid sequence," *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 92, pp. 8700–8704, 1995.

[8] G. D. Fasman, *Handbook of Biochemistry and Molecular Biology, 3rd edition: Proteins.* CRC Press; 3 edition (April 19, 1976), 1976.

[9] J.-L. Fauchere, M. Charton, L. B. Kier, A. Verloop, and V. Pliska, "Amino acid side chain parameters for correlation studies in biology and pharmacology," *International Journal of Peptide and Protein Research*, Vol. 32, pp. 269–278, 1988.

[10] Z. Feng, L. Chen, H. Maddula, O. Akcan, R. Oughtred, H. M. Berman, and J. Westbrook, "Ligand depot: a data warehouse for ligands bound to macromolecules," *Bioinformatics*, Vol. 20, pp. 2153–2155, 2004.

[11] R. Grantham, "Amino acid difference formula to help explain protein evolution," *Science*, Vol. 185, pp. 862–864, 1974.

[12] 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.

[13] 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.

[14] J. Janin, S. Wodak, M. Levitt, , and B. Maigret, "Conformation of amino acid side-chains in proteins," *Journal of Molecular Biology*, Vol. 125, pp. 357–386, 1978.

[15] P. Klein, M. Kanehisa, and C. DeLisi, "Prediction of protein function from sequence properties: Discriminant analysis of a data base," *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology*, Vol. 787, pp. 221–226, 1984.

[16] J. Kyte and R. F. Doolittle, "A simple method for displaying the hydropathic character of a protein," *Journal of Molecular Biology*, Vol. 157, pp. 105–132, 1982.

[17] 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.

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

[19] 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.

[20] 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.

[21] S. Santini, G. Wei, N. Mousseau, and P. Derreumaux, "Exploring the folding pathways of proteins through energy landscape sampling: Application to alzheimer's β-amyloid peptide," *Internet Electronic Journal of Molecular Design*, Vol. 2, No. 9, pp. 564–577, 2003.

[22] 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.

[23] J. Zimmerman, N. Eliezer, and R. Simha, "The characterization of amino acid sequences in proteins by statistical methods," *Journal of Theoretical Biology*, Vol. 21, pp. 170–201, 1968.