International Journal of
Molecular Sciences

Article

Prediction of Protein Structural Class Based on Gapped-Dipeptides and a Recursive Feature Selection Approach

Taigang Liu [1], Yufang Qin [1], Yongjie Wang [2,*] and Chunhua Wang [1,*]

Received: 20 November 2015; Accepted: 18 December 2015; Published: 24 December 2015
Academic Editor: Christo Z. Christov

[1] College of Information Technology, Shanghai Ocean University, Shanghai 201306, China; tgliu@shou.edu.cn (T.L.); yfqin@shou.edu.cn (Y.Q.)

[2] College of Food Science & Technology, Shanghai Ocean University, Shanghai 201306, China

* Correspondence: yjwang@shou.edu.cn (Y.W.); wchshou@163.com (C.W.);
Tel.: +86-21-6190-0505 (Y.W.); +86-21-6190-0624 (C.W.)

Abstract: The prior knowledge of protein structural class may offer useful clues on understanding its functionality as well as its tertiary structure. Though various significant efforts have been made to find a fast and effective computational approach to address this problem, it is still a challenging topic in the field of bioinformatics. The position-specific score matrix (PSSM) profile has been shown to provide a useful source of information for improving the prediction performance of protein structural class. However, this information has not been adequately explored. To this end, in this study, we present a feature extraction technique which is based on gapped-dipeptides composition computed directly from PSSM. Then, a careful feature selection technique is performed based on support vector machine-recursive feature elimination (SVM-RFE). These optimal features are selected to construct a final predictor. The results of jackknife tests on four working datasets show that our method obtains satisfactory prediction accuracies by extracting features solely based on PSSM and could serve as a very promising tool to predict protein structural class.

Keywords: feature selection; gapped-dipeptide; position-specific score matrix; protein structural class; recursive feature elimination; support vector machine

1. Introduction

Proteins can perform many biological functions within living organisms when they fold and take on a three-dimensional structure [1–4]. According to the concept of structural class introduced by Levitt and Chothia [5], proteins are divided into four major structural classes: all-α, all-β, α/β and α + β. The knowledge of protein structural class can provide important and useful information about a protein's three-dimensional structure and its functionality [6]. However, it is usually time-consuming and costly to determine the structure information of a protein by just relying on wet-bench experiments. On the other hand, sequence information has grown exponentially with the help of high-throughput sequencing techniques, which has made a huge gap between the sequence and structure space. Hence, there is a great need to explore bioinformatics prediction methods based on sequence data to fill this gap.

From the pattern recognition perspective, predicting protein structural class is usually described as a multi-class classification problem. During the past 30 years, various significant efforts have been made to solve this problem. These methods generally consist of two major steps: (1) protein sequence representation or feature extraction; (2) algorithm selection for classification. Many classification techniques have been proposed to perform the prediction of protein structural class such as neural network [7], support vector machine (SVM) [8–10], fuzzy *k*-nearest neighbor [11,12], fuzzy clustering [13], Bayesian classification [14], Logistic regression [15,16], rough sets [17], and ensembles of classifiers [18–22]. Among these algorithms, SVM has attained the best prediction performance for this task [9]. At the same time a wide range of sequence features have been used to reveal more discriminatory information for protein structural class, including amino acid composition (AAC) [23,24], pseudo-AAC [25–27], position-specific score matrix (PSSM) profile [28–31] and predicted secondary structure [32–34]. As a powerful feature extraction tool for analyzing DNA or protein sequences, pseudo-AAC has been widely applied to the field of bioinformatics [35–40].

Among the above sequence features, the most significant enhancements in prediction accuracy are based on the PSSM profile and predicted secondary structure. Since the prediction performance of protein secondary structure using PSIPRED software [41] crucially relies on PSSM, the PSSM profile provides more important and original discriminatory information for protein structural class prediction. Recently, several methods have been developed to extract the potential local and global information from PSSM such as AAC [31], dipeptide composition [31], auto covariance (AC) [30], and linear correlation coefficient [29]. However, the informative features encoded in PSSM have not been adequately explored due to limited prediction accuracy. This highlights the need for exploring more effective feature extraction techniques to represent protein sequences.

In this study, we introduce a feature extraction approach based on gapped-dipeptides (*i.e.*, two residues separated by one or more positions) composition (GapDPC) to further explore more discriminatory information solely from the PSSM profile. The processes of our method are as follows. First, the PSSM profile of a protein is transformed into a fix-length feature vector by extracting GapDPC. Then, a recursive feature selection approach is applied to reduce feature redundancy and optimal features are input to an SVM classifier to conduct the prediction. Finally, validation results on four working datasets indicate that our method presents outstanding improvements in prediction accuracies compared with other existing methods.

2. Results and Discussion

2.1. Parameter Selection

Preliminary test results indicate that the length of the shortest sequence in the dataset is 10. By integrating GapDPC with different gapped distances, the value of parameter G is set to eight in this study, which results in 3600 features for each protein sequence. Then, these features are ranked based on their relevance to sample classification by support vector machine-recursive feature elimination (SVM-RFE). To explore the impact of selected feature dimensions on prediction performance, we calculate the overall accuracies for top K features using five-fold cross-validation, where K = 10, 20, 30, ... , 500. The results are shown in Figure 1. As can be seen, the overall accuracies for the 1189 and 25PDB datasets achieve a maximum value when K increases to 460. Thus, the top 460 features are selected to further compute the accuracies for two low-similarity datasets by jackknife tests. Similarly, the top 110 features are adopted for two small datasets, Z277 and Z498, due to their high accuracies. The results of jackknife tests on four datasets are listed in Table 1.

Table 1. Prediction performances on four datasets by our method.

Dataset	Accuracy (%)					Matthews Correlation Coefficient (MCC)			
	All-α	All-β	α/β	α + β	Overall	All-α	All-β	α/β	α + β
Z277	97.1	98.4	97.5	96.9	97.5	0.96	0.98	0.97	0.96
Z498	98.1	100	98.5	97.7	98.6	0.96	1	0.98	0.98
1189	94.2	93.2	92.5	83.0	90.9	0.89	0.91	0.89	0.82
25PDB	94.8	92.3	87.0	86.4	90.3	0.88	0.89	0.87	0.84

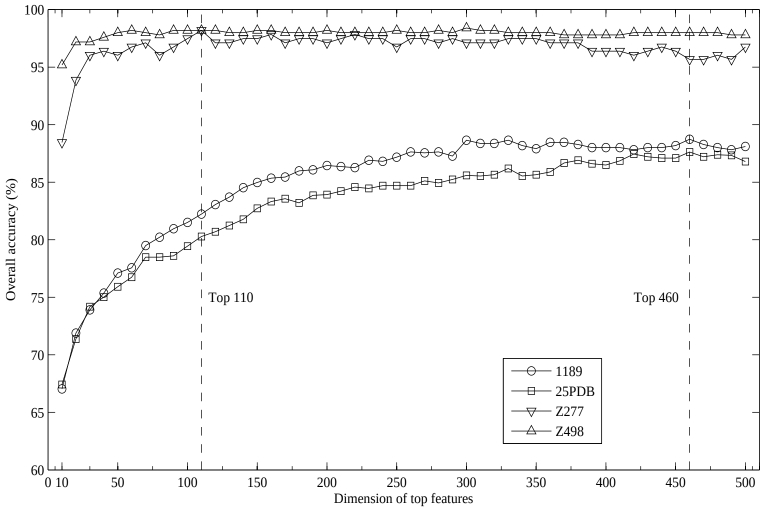

Figure 1. This graph shows how different top *K* features affect the overall accuracies.

2.2. Performance Comparison with Existing Methods

In order to evaluate the effectiveness of the proposed method, we first compare it with the other existing methods based on the Z277 and Z498 datasets. The results from the jackknife tests are summarized in Tables 2 and 3.

Table 2. Comparison of different methods by the jackknife test for the Z277 dataset.

Method	**Prediction Accuracy (%)**				
	All-α	**All-β**	α/β	α + β	**Overall**
Neural network [7]	68.6	85.2	86.4	56.9	74.7
Component coupled [23]	84.3	82.0	81.5	67.7	79.1
LogitBoost [19]	81.4	88.5	92.6	72.3	84.1
IGA-SVM [10]	84.3	88.5	92.6	70.7	84.5
CWT-PCA-SVM [27]	85.7	90.2	87.7	80.1	85.9
Markov-SVM [42]	90.0	85.2	86.4	81.5	85.9
SVM fusion [21]	85.7	90.2	93.8	80.0	87.7
AAC-PSSM-AC [30]	88.6	95.1	97.5	81.5	91.0
Our method	97.1	98.4	97.5	96.9	97.5

Table 3. Comparison of different methods by the jackknife test for the Z498 dataset.

Method	**Prediction Accuracy (%)**				
	All-α	**All-β**	α/β	α + β	**Overall**
Neural network [7]	86.0	96.0	88.2	86.0	89.2
Component-coupled [23]	93.5	88.9	90.4	84.5	89.2
SVM fusion [21]	99.1	96.0	80.9	91.5	91.4
Markov-SVM [42]	91.6	94.4	96.3	91.5	93.6
IGA-SVM [10]	96.3	93.6	97.8	89.2	94.2
LogitBoost [19]	92.6	96.0	97.1	93.0	94.8
CWT-PCA-SVM [27]	94.4	96.8	97.0	92.3	95.2
AAC-PSSM-AC [30]	94.4	96.8	97.8	93.8	95.8
Our method	98.1	100	98.5	97.7	98.6

As is shown, our method obtains the overall accuracies of 97.5% and 98.6% on these two datasets, which are better than the other classifiers including neural network [7], component-coupled [23], LogitBoost [19], AAC-PSSM-AC [30] and SVM-based methods [10,21,27,42]. It is worth noting that the AAC-PSSM-AC algorithm, which extracts AAC and AC features solely from the PSSM profile to represent a protein, also attains the second best prediction performance. This illustrates that the PSSM profile indeed provides important and useful discriminatory information for predicting protein structural class. In addition, we notice that the total accuracies of our method are higher than those of the LogitBoost and SVM fusion classifiers, which incorporate many weak classifiers to construct a strong classifier. This suggests that designing better sequence representations is more important than exploring more complex classifiers.

To explore the impact of sequence similarity on the performance of our method, we make comparisons with other competing prediction methods against two low-similarity datasets (*i.e.*, 1189 and 25PDB). The high prediction accuracies of these methods are mainly due to extracting features from the PSSM profile as well as the predicted secondary structure information. The approaches based on PSSM include AADP-PSSM [31], AAC-PSSM-AC [30], Comb_11,10,6 [22], LCC-PSSM [29] and PSSM-SPINE-S [34]. The approaches based on the predicted secondary structure include SCPRED [9], RKS-PPSC [43], MODAS [33], and PSSM-SPINE-S [34]. The results by jackknife tests are listed in Tables 4 and 5.

Table 4. Performance comparison of different methods on the 1189 dataset.

Method	Prediction Accuracy (%)				
	All-α	**All-β**	α/β	α + β	**Overall**
AADP-PSSM [31]	69.1	83.7	85.6	35.7	70.7
AAC-PSSM-AC [30]	80.7	86.4	81.4	45.2	74.6
Comb_11,10,6 [1] [22]	80.2	83.6	85.4	44.6	74.8
SCPRED [9]	89.1	86.7	89.6	53.8	80.6
LCC-PSSM [29]	89.2	88.8	85.6	58.5	81.2
RKS-PPSC [43]	89.2	86.7	82.6	65.6	81.3
MODAS [33]	92.3	87.1	87.9	65.4	83.5
PSSM-SPINE-S [34]	98.2	91.5	83.8	72.2	86.3
Our method	94.2	93.2	92.5	83.0	90.9

[1] The result is evaluated using 10-fold cross-validation test.

Table 5. Performance comparison of different methods on the 25PDB dataset.

Method	Prediction Accuracy (%)				
	All-α	**All-β**	α/β	α + β	**Overall**
AADP-PSSM [31]	83.3	78.1	76.3	54.4	72.9
AAC-PSSM-AC [30]	85.3	81.7	73.7	55.3	74.1
Comb_11,10,6 [1] [22]	86.1	80.8	80.6	60.1	76.7
LCC-PSSM [29]	91.7	80.8	79.8	64.0	79.0
SCPRED [9]	92.6	80.1	74.0	71.0	79.7
MODAS [33]	92.3	83.7	81.2	68.3	81.4
RKS-PPSC [43]	92.8	83.3	85.8	70.1	82.9
PSSM-SPINE-S [34]	96.8	93.7	90.1	87.0	92.2
Our method	94.8	92.3	87.0	86.4	90.3

[1] The result is evaluated using 10-fold cross-validation test.

For the 1189 dataset, the proposed method outperforms all other methods listed in Table 4, with an accuracy of 90.9%. It is also shown that studies which relied on predicted secondary structure to enhance the accuracy could not reach a result too much better than 80%. This may be due to the limited accuracy (about 80%) of the predicted secondary structure by PSIPRED. Referring to Table 5, the overall accuracy of our method achieves 90.3% for the 25PDB dataset, which is higher than those of other methods except for PSSM-SPINE-S. It should be pointed out that PSSM-SPINE-S combines PSSM features with secondary structure features extracted from the SPINE-X [44] to improve the performance. This indicates that predicted secondary structure information plays an important complementary role for predicting protein structural class. However, the proposed representation also attains satisfactory performance when only the PSSM profile is employed.

From the above comparisons, our method shows substantial improvements for the prediction of protein structural class. This could be attributed to the informative feature extraction technique based on GapDPC computed directly from PSSM and selected optimal features by SVM-RFE.

3. Materials and Methods

3.1. Datasets

Two datasets (*i.e.*, Z277 and Z498) constructed by Zhou [23] are first used to evaluate the proposed method, and they contain 277 and 498 protein domains, respectively. Despite the relatively small size of these two datasets, they were widely used in many studies. To explore the impact of the proposed method on the low-similarity datasets, another two datasets, 1189 [14] and 25PDB [15], are also studied separately. The first one consists of 1092 protein domains with sequence similarity less than 40% and the second one includes 1673 protein domains with sequence similarity lower than 25%. The detailed compositions of four datasets are listed in Table 6.

Table 6. The compositions of four datasets adopted in this study.

Dataset	All-α	All-β	α/β	α + β	Total
Z277	70	61	81	65	277
Z498	107	126	136	129	498
1189	223	294	334	241	1092
25PDB	443	443	346	441	1673

3.2. Protein Sequence Representation

Previous successful applications of PSSM profile illustrate that evolutionary information is more informative than sequence itself [28,30]. In this section, a simple sequence representation which combines PSSM profile and the concept of GapDPC is developed for the proposed prediction method.

The profile of each sequence is generated by running PSI-BLAST program [45] against the NCBI's non-redundant (NR) database with three iterations and a cutoff *E*-value of 0.001. The (*i*, *j*)th entry of the resulting matrix represents the probability of amino acid type *j* occurring at the *i*th position of the query sequence. The PSSM elements are mapped to the range of (0, 1) by the following sigmoid function:

$$f\left(x\right) = \frac{1}{1+e^{-x}} \tag{1}$$

where x is the original PSSM value.

For convenience, let us denote

$$P = \left(P_1,\ P_2, \ldots, P_{20}\right) \tag{2}$$

as the PSSM of the query sequence S, where

$$P_j = \left(p_{1,j},\ p_{2,j}, \ldots, p_{L,j}\right)^T\ \left(j = 1, 2, \ldots, 20\right) \tag{3}$$

L is the length of the query sequence S, and T is the transpose operator.

Since the structural class of a protein is closely related to its dipeptide composition (DPC) [31], we first extend the concept of traditional DPC from the primary sequence to the PSSM. DPC is defined as a 400-dimentional vector:

$$X = \left(x_{1,1},\ \ldots, x_{1,20}, x_{2,1},\ \ldots, x_{2,20}, \ldots, x_{20,1},\ \ldots, x_{20,20}\right) \tag{4}$$

where

$$x_{i,j} = \sum_{k=1}^{L-1} p_{k,i} \times p_{k+1,j} \ (1 \leqslant i, j \leqslant 20) \tag{5}$$

As we all know, sequence-order information is as important as its residue composition in a protein sequence. To partially reflect the local sequence-order effect, GapDPC is introduced to explore the long-range correlation between two residues separated by one or more positions, which can be calculated by

$$y_{i,j,g} = \sum_{k=1}^{L-g-1} p_{k,i} \times p_{k+g+1,j} \ (1 \leqslant i, j \leqslant 20) \tag{6}$$

where g is the distance between amino acid i and amino acid j. Note that GapDPC is reduced to DPC when g is equal to 0.

These elements of the three-dimensional matrix $y_{i,j,g}$, which correspond to the frequencies of PSSM-based gapped-dipeptides, are used to represent the given query sequence. We generate PSSM-based GapDPC for g = 0, 1, 2, . . . , G, which results in 400*(G + 1) features for each sequence.

3.3. Recursive Feature Selection

After running the proposed feature extraction technique, all protein sequences with different length are converted into numerical feature vectors with the same dimension. In order to decrease feature redundancy and reduce computation cost, we introduce a recursive feature selection approach to rank the features according to their importance. Support vector machine-recursive feature elimination (SVM-RFE), which was originally carried out on gene selection for cancer classification by Guyon and his co-workers [46], has been proven to be an effective tool for dimensionality reduction in the field of pattern recognition. The process is conducted as follows. First, all the feature vectors of proteins for each dataset are trained using SVM with a linear kernel. Then, the features are ranked with decreasing order according to their weights which reflect the relevance to prediction of protein structural class. Finally, top K features with the most relevant ranks are selected to represent each protein sequence.

3.4. Support Vector Machine

SVM, which is first introduced by Vapnik [47], is considered as the state-of-the-art machine learning algorithm for classification. It maps the input data into higher dimensional feature space using the kernel function and then finds an optimal hyper-plane to separate a given set of labeled data. Among a lot of classification algorithms used for prediction of protein structural class, SVM has shown the best prediction accuracies [9]. In this work, the SVM classifier implemented by the LIBSVM software (Chang and Lin, Taipei, Taiwan) [48] is employed to perform the prediction. Though LIBSVM provides four basic kernel functions, *i.e.*, linear, polynomial, radial basis function (RBF) and Gaussian, RBF kernel is adopted here due to its better performance than other kernel functions. The cost parameter C and the width parameter γ are optimized based on the grid search algorithm implemented in the LIBSVM software.

3.5. Cross-Validation and Performance Evaluation

In this study, the jackknife test is adopted to evaluate the prediction performance of our method. Although the jackknife test is time-consuming, it is considered more objective than other cross-validation methods (e.g., independent dataset test and sub-sampling test) [49]. The basic idea behind the jackknife test lies in systematically calculating the statistic estimate, leaving out each sample from a dataset and then finding the average of these calculations. To evaluate the performance of our predictor, the accuracy, overall accuracy and Matthews correlation coefficient (MCC) are adopted as the comparative measures. They are defined by the following formulas:

$$Accuracy_j = \frac{TP_j}{TP_j + FN_j} = \frac{TP_j}{|C_j|} \tag{7}$$

$$MCC_j = \frac{TP_j \times TN_j - FP_j \times FN_j}{\sqrt{(TP_j + FP_j)\,(TP_j + FN_j)\,(TN_j + FP_j)\,(TN_j + FN_j)}} \tag{8}$$

$$Overall\ accuracy = \frac{\sum_j TP_j}{\sum_j |C_j|} \tag{9}$$

where TP_j, TN_j, FP_j, FN_j, and $|C_j|$ are the number of true positives, true negatives, false positives, false negatives, and proteins in the structural class C_j, respectively.

4. Conclusions

In this study, we combine gapped-dipeptides with SVM-RFE to predict protein structural class. In order to partly reflect the local sequence-order effect, the proposed method extracts features from gapped-dipeptides of various distances based on PSSM. These features are further ranked by SVM-RFE according to their importance and the optimal features are input to SVM classifiers to perform the prediction. Comparison with other existing techniques on four benchmark datasets indicates that our predictor is a useful tool to predict protein structural class and also shows the generality of the proposed method.

Acknowledgments: This work was partially supported by the Innovation Program of Shanghai Municipal Education Commission (No. 13YZ098), the National Natural Science Foundation of China (No. 41376135, 31570112) and the Doctoral Fund of Shanghai Ocean University.

Author Contributions: The manuscript was written through contributions of all authors. Taigang Liu designed the study and wrote the paper; Yufang Qin collected and analyzed the data; Chunhua Wang implemented the algorithm; Yongjie Wang revised the paper. All authors read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Leung, C.H.; Chan, D.S.; He, H.Z.; Cheng, Z.; Yang, H.; Ma, D.L. Luminescent detection of DNA-binding proteins. *Nucleic Acids Res.* **2012**, *40*, 941–955. [CrossRef] [PubMed]
2. Singh, R.; Brewer, M.K.; Mashburn, C.B.; Lou, D.; Bondada, V.; Graham, B.; Geddes, J.W. Calpain 5 is highly expressed in the central nervous system (CNS), carries dual nuclear localization signals, and is associated with nuclear promyelocytic leukemia protein bodies. *J. Biol. Chem.* **2014**, *289*, 19383–19394. [CrossRef] [PubMed]
3. Leung, K.H.; He, B.; Yang, C.; Leung, C.H.; Wang, H.M.; Ma, D.L. Development of an aptamer-based sensing platform for metal ions, proteins, and small molecules through terminal deoxynucleotidyl transferase induced G-quadruplex formation. *ACS Appl. Mater. Interfaces* **2015**, *7*, 24046–24052. [CrossRef] [PubMed]
4. Lin, S.; Gao, W.; Tian, Z.; Yang, C.; Lu, L.; Mergny, J.-L.; Leung, C.-H.; Ma, D.-L. Luminescence switch-on detection of protein tyrosine kinase-7 using a G-quadruplex-selective probe. *Chem. Sci.* **2015**, *6*, 4284–4290. [CrossRef]
5. Levitt, M.; Chothia, C. Structural patterns in globular proteins. *Nature* **1976**, *261*, 552–558. [CrossRef] [PubMed]
6. Chou, K.C. Progress in protein structural class prediction and its impact to bioinformatics and proteomics. *Curr. Protein Pept. Sci.* **2005**, *6*, 423–436. [CrossRef] [PubMed]
7. Cai, Y.D.; Zhou, G.P. Prediction of protein structural classes by neural network. *Biochimie* **2000**, *82*, 783–785. [CrossRef]
8. Cai, Y.D.; Liu, X.J.; Xu, X.; Zhou, G.P. Support vector machines for predicting protein structural class. *BMC Bioinformatics* **2001**, *2*, 3. [CrossRef] [PubMed]

9. Kurgan, L.; Cios, K.; Chen, K. Scpred: Accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences. *BMC Bioinform.* **2008**, *9*, 226. [CrossRef] [PubMed]
10. Li, Z.C.; Zhou, X.B.; Lin, Y.R.; Zou, X.Y. Prediction of protein structure class by coupling improved genetic algorithm and support vector machine. *Amino Acids* **2008**, *35*, 581–590. [CrossRef] [PubMed]
11. Zheng, X.; Li, C.; Wang, J. An information-theoretic approach to the prediction of protein structural class. *J. Comput. Chem.* **2010**, *31*, 1201–1206. [CrossRef] [PubMed]
12. Zhang, T.L.; Ding, Y.S.; Chou, K.C. Prediction protein structural classes with pseudo-amino acid composition: Approximate entropy and hydrophobicity pattern. *J. Theor. Biol.* **2008**, *250*, 186–193. [CrossRef] [PubMed]
13. Shen, H.B.; Yang, J.; Liu, X.J.; Chou, K.C. Using supervised fuzzy clustering to predict protein structural classes. *Biochem. Biophys. Res. Commun.* **2005**, *334*, 577–581. [CrossRef] [PubMed]
14. Wang, Z.X.; Yuan, Z. How good is prediction of protein structural class by the component-coupled method? *Proteins* **2000**, *38*, 165–175. [CrossRef]
15. Kurgan, L.A.; Homaeian, L. Prediction of structural classes for protein sequences and domains—Impact of prediction algorithms, sequence representation and homology, and test procedures on accuracy. *Pattern Recogn.* **2006**, *39*, 2323–2343. [CrossRef]
16. Kurgan, L.; Chen, K. Prediction of protein structural class for the twilight zone sequences. *Biochem. Biophys. Res. Commun.* **2007**, *357*, 453–460. [CrossRef] [PubMed]
17. Cao, Y.F.; Liu, S.; Zhang, L.D.; Qin, J.; Wang, J.; Tang, K.X. Prediction of protein structural class with rough sets. *BMC Bioinform.* **2006**, *7*, 20. [CrossRef] [PubMed]
18. Cai, Y.D.; Feng, K.Y.; Lu, W.C.; Chou, K.C. Using logitboost classifier to predict protein structural classes. *J. Theor. Biol.* **2006**, *238*, 172–176. [CrossRef] [PubMed]
19. Feng, K.Y.; Cai, Y.D.; Chou, K.C. Boosting classifier for predicting protein domain structural class. *Biochem. Biophys. Res. Commun.* **2005**, *334*, 213–217. [CrossRef] [PubMed]
20. Chen, L.; Lu, L.; Feng, K.; Li, W.; Song, J.; Zheng, L.; Yuan, Y.; Zeng, Z.; Lu, W.; Cai, Y. Multiple classifier integration for the prediction of protein structural classes. *J. Comput. Chem.* **2009**, *30*, 2248–2254. [CrossRef] [PubMed]
21. Chen, C.; Zhou, X.; Tian, Y.; Zou, X.; Cai, P. Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network. *Anal. Biochem.* **2006**, *357*, 116–121. [CrossRef] [PubMed]
22. Dehzangi, A.; Paliwal, K.; Sharma, A.; Dehzangi, O.; Sattar, A. A combination of feature extraction methods with an ensemble of different classifiers for protein structural class prediction problem. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2013**, *10*, 564–575. [CrossRef] [PubMed]
23. Zhou, G.P. An intriguing controversy over protein structural class prediction. *J. Protein Chem.* **1998**, *17*, 729–738. [CrossRef] [PubMed]
24. Chou, K.C. A key driving force in determination of protein structural classes. *Biochem. Biophys. Res. Commun.* **1999**, *264*, 216–224. [CrossRef] [PubMed]
25. Chou, K.C. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins* **2001**, *43*, 246–255. [CrossRef] [PubMed]
26. Zhang, T.L.; Ding, Y.S. Using pseudo amino acid composition and binary-tree support vector machines to predict protein structural classes. *Amino Acids* **2007**, *33*, 623–629. [CrossRef] [PubMed]
27. Li, Z.C.; Zhou, X.B.; Dai, Z.; Zou, X.Y. Prediction of protein structural classes by Chou's pseudo amino acid composition: Approached using continuous wavelet transform and principal component analysis. *Amino Acids* **2009**, *37*, 415–425. [CrossRef] [PubMed]
28. Chen, K.; Kurgan, L.A.; Ruan, J.S. Prediction of protein structural class using novel evolutionary collocation-based sequence representation. *J. Comput. Chem.* **2008**, *29*, 1596–1604. [CrossRef] [PubMed]
29. Ding, S.; Yan, S.; Qi, S.; Li, Y.; Yao, Y. A protein structural classes prediction method based on PSI-BLAST profile. *J. Theor. Biol.* **2014**, *353*, 19–23. [CrossRef] [PubMed]
30. Liu, T.; Geng, X.; Zheng, X.; Li, R.; Wang, J. Accurate prediction of protein structural class using auto covariance transformation of PSI-BLAST profiles. *Amino Acids* **2012**, *42*, 2243–2249. [CrossRef] [PubMed]
31. Liu, T.; Zheng, X.; Wang, J. Prediction of protein structural class for low-similarity sequences using support vector machine and PSI-BLAST profile. *Biochimie* **2010**, *92*, 1330–1334. [CrossRef] [PubMed]

32. Kurgan, L.A.; Zhang, T.; Zhang, H.; Shen, S.Y.; Ruan, J.S. Secondary structure-based assignment of the protein structural classes. *Amino Acids* **2008**, *35*, 551–564. [CrossRef] [PubMed]
33. Mizianty, M.J.; Kurgan, L. Modular prediction of protein structural classes from sequences of twilight-zone identity with predicting sequences. *BMC Bioinform.* **2009**, *10*, 414. [CrossRef] [PubMed]
34. Dehzangi, A.; Paliwal, K.; Lyons, J.; Sharma, A.; Sattar, A. Proposing a highly accurate protein structural class predictor using segmentation-based features. *BMC Genom.* **2014**, *15*, S2. [CrossRef] [PubMed]
35. Nieto, J.J.; Torres, A.; Georgiou, D.N.; Karakasidis, T.E. Fuzzy polynucleotide spaces and metrics. *Bull. Math. Biol.* **2006**, *68*, 703–725. [CrossRef] [PubMed]
36. Georgiou, D.N.; Karakasidis, T.E.; Nieto, J.J.; Torres, A. Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition. *J. Theor. Biol.* **2009**, *257*, 17–26. [CrossRef] [PubMed]
37. Georgiou, D.N.; Karakasidis, T.E.; Nieto, J.J.; Torres, A. A study of entropy/clarity of genetic sequences using metric spaces and fuzzy sets. *J. Theor. Biol.* **2010**, *267*, 95–105. [CrossRef] [PubMed]
38. Du, P.; Gu, S.; Jiao, Y. Pseaac-general: Fast building various modes of general form of Chou's pseudo-amino acid composition for large-scale protein datasets. *Int. J. Mol. Sci.* **2014**, *15*, 3495–3506. [CrossRef] [PubMed]
39. Qiu, W.R.; Xiao, X.; Chou, K.C. Irspot-tncpseaac: Identify recombination spots with trinucleotide composition and pseudo amino acid components. *Int. J. Mol. Sci.* **2014**, *15*, 1746–1766. [CrossRef] [PubMed]
40. Huang, Q.; You, Z.; Zhang, X.; Zhou, Y. Prediction of protein-protein interactions with clustered amino acids and weighted sparse representation. *Int. J. Mol. Sci.* **2015**, *16*, 10855–10869. [CrossRef] [PubMed]
41. Jones, D.T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **1999**, *292*, 195–202. [CrossRef] [PubMed]
42. Qin, Y.F.; Wang, C.H.; Yu, X.Q.; Zhu, J.; Liu, T.G.; Zheng, X.Q. Predicting protein structural class by incorporating patterns of over-represented *k*-mers into the general form of Chou's PseAAC. *Protein Pept. Lett.* **2012**, *19*, 388–397. [CrossRef] [PubMed]
43. Yang, J.Y.; Peng, Z.L.; Chen, X. Prediction of protein structural classes for low-homology sequences based on predicted secondary structure. *BMC Bioinform.* **2010**, *11*, S9. [CrossRef] [PubMed]
44. Faraggi, E.; Zhang, T.; Yang, Y.; Kurgan, L.; Zhou, Y. Spine x: Improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. *J. Comput. Chem.* **2012**, *33*, 259–267. [CrossRef] [PubMed]
45. Altschul, S.F.; Madden, T.L.; Schaffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped blast and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402. [CrossRef] [PubMed]
46. Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **2002**, *46*, 389–422. [CrossRef]
47. Vapnik, V. *Statistical Learning Theory*; Wiley: New York, NY, USA, 1998.
48. Chang, C.C.; Lin, C.J. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2001**. [CrossRef]
49. Chou, K.C.; Zhang, C.T. Prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol.* **1995**, *30*, 275–349. [CrossRef] [PubMed]