Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 6802832, 10 pages
http://dx.doi.org/10.1155/2016/6802832

Research Article

ProFold: Protein Fold Classification with Additional Structural Features and a Novel Ensemble Classifier

Daozheng Chen,[1] Xiaoyu Tian,[1] Bo Zhou,[2] and Jun Gao[1]

[1] *College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China*
[2] *Shanghai University of Medicine & Health Sciences, Shanghai 201318, China*

Correspondence should be addressed to Bo Zhou; zhoub@sumhs.edu.cn and Jun Gao; jungao@shmtu.edu.cn

Received 1 June 2016; Revised 15 July 2016; Accepted 7 August 2016

Academic Editor: Dariusz Mrozek

Protein fold classification plays an important role in both protein functional analysis and drug design. The number of proteins in PDB is very large, but only a very small part is categorized and stored in the SCOPe database. Therefore, it is necessary to develop an efficient method for protein fold classification. In recent years, a variety of classification methods have been used in many protein fold classification studies. In this study, we propose a novel classification method called proFold. We import protein tertiary structure in the period of feature extraction and employ a novel ensemble strategy in the period of classifier training. Compared with existing similar ensemble classifiers using the same widely used dataset (DD-dataset), proFold achieves 76.2% overall accuracy. Another two commonly used datasets, EDD-dataset and TG-dataset, are also tested, of which the accuracies are 93.2% and 94.3%, higher than the existing methods. ProFold is available to the public as a web-server.

1. Introduction

Protein fold classification is a crucial problem in structural bioinformatics. Protein folding information is helpful in identifying the tertiary structure and functional information of a protein [1]. In recent years, many protein fold classification studies have been performed. The methods proposed by researchers can be roughly divided into two categories: one is template-based method [2–7], and the other is taxonomy-based method [8–15]. Recently, taxonomy-based methods have attracted more attention due to their relatively excellent performance.

The taxonomy-based method was proposed by Dubchak et al. [8, 9] in 1995 for the first time. Many taxonomy-based methods classify a query protein to a known folding type. This nonmanual label method contributes to the growth of the quantity of protein in Structural Classification of Proteins (SCOP) [16] and could narrow the gap between the number of proteins in SCOP and Protein Data Bank (PDB). In this paper, the taxonomy-based method is equivalent to the classification problem in machine learning. There are two significant problems in classification tasks: one is feature extraction, and the other is machine learning method.

In terms of feature extraction, most of the researchers extract multidimensional numerical feature vectors from amino acid sequences. In 1995, Dubchak et al. [8, 9] extracted global description of amino acid sequence for the first time. Since then, in order to improve the accuracy of classification, some researchers have put forward other feature extraction methods, such as pseudoamino acid composition [12, 17], pairwise frequency information [18], Position Specific Scoring Matrix (PSSM) [17], structural properties of amino acid residues and amino acid residue pairs [19], and hidden Markov model structural alphabet [20, 21]. Except for extracting features from amino acid sequence directly, some features are extracted from evolution information combining the functional domain and the sequential evolution information [22] and predicted secondary structure [14, 23, 24]. Although the classification accuracy can be improved after combining these features together [20, 25], it is still not good enough.

For protein fold classification, many classifiers have been used, such as neutral network (NNs) [8, 13], SVMs [10, 13, 18–21, 24, 26–33], k-nearest neighbors (k-NN) [12], probabilistic multiclass multikernel classifier [25], random forest [23, 34–37], rotation forest [38], and a variety of ensemble classifiers [11, 12, 14, 18, 22, 39–41].

Up to 28th April, 2016, PDB had 109850 protein structures (http://www.rcsb.org/pdb/home/home.do). However, Structural Classification of Proteins- extended (SCOPe) [42] only had 77439 PDB entries (http://scop.berkeley.edu/statistics/ver=2.06). Therefore, there still exists a great number of protein structures which do not have their structure classification labels in the SCOPe database. What is more, most protein structures in SCOPe are classified manually, so it requires a lot of manual labor. In this study, we start from the PDB file 3D structure studying the protein fold classification. In terms of feature extraction, we use a new feature extraction method, combining the existing methods of the global description of amino acid sequence [13], PSSM [43], and protein functional information [22] proposed by other researchers. The new feature extraction method extracts eight types of secondary structure states from PDB files by the Definition of Secondary Structure in Proteins (DSSP) software [44]. In terms of machine learning classifiers, we propose a novel ensemble strategy. With the new added feature extracted from DSSP and the novel ensemble strategy we propose, our method can achieve 1–3% higher accuracy than similar methods.

As demonstrated by a series of recent publications [45–55] in compliance with Chou's 5-step rule [56], to establish a really useful machine learning classifier for a biological system, we should follow the following five guidelines: (a) benchmark dataset construction or selection for training and testing the model; (b) extract features from the biological sequence samples with effective methods that can truly reflect their intrinsic correlation with the target to be predicted; (c) introduce or develop a powerful algorithm (or engine) to operate the classifier; (d) properly perform cross-validation tests and test on independent dataset to objectively evaluate the anticipated accuracy of the classifier; (e) establish a user-friendly web-server (http://binfo.shmtu.edu.cn/profold/) for the classifier that is accessible to the public. In the following, we are to describe how to deal with these steps one-by-one.

2. Materials and Methods

2.1. Data Sets. In this study, three benchmark datasets are used, respectively: (1) Ding and Dubchak (DD) [13], (2) Taguchi and Gromiha (TG) [58], and (3) Extended DD (EDD) [10]. DD-dataset was proposed by Ding and Dubchak in 2001 and modified by Shen and Chou in 2006 [12]. Since then, DD-dataset has been used in many protein fold classification studies [11, 18, 20–24, 26, 32–36, 38, 40, 57, 59]. There are 311 protein sequences in the training set and 386 protein sequences in the testing set with no two proteins having more than 35% of sequence identity. The protein sequences in DD-dataset were selected from 27 SCOP [35] folds comprehensively, which belong to different structural classes containing α, β, α/β, and $\alpha + \beta$.

TG-dataset contains 30 SCOP folds and 1612 protein sequences with no two protein having more than 25% sequence identity.

EDD-dataset contains 27 SCOP folds, like DD-dataset. There are 3418 protein sequences with no protein having more than 40% sequence identity.

These three datasets can be downloaded directly from our website (http://binfo.shmtu.edu.cn/profold/benchmark.html).

2.2. Feature Extraction Method. With the rapid growth of biological sequences in the postgenomic age, one of the most important but also most difficult problems in computational biology is how to represent a biological sequence with a discrete model or a vector. Therefore Chou's PseAAC [60–62] was proposed. Encouraged by the successes of using PseAAC to deal with protein/peptide sequences, three web-servers [63–65] were developed for generating various feature vectors for DNA/RNA sequences. Particularly, recently a powerful web-server called Pse-in-One [66] has been established that can be used to generate any desired feature vectors for protein/peptide and DNA/RNA sequences according to the need of users' studies. Inspired by this, in this study, we extract four feature groups, including the DSSP feature, the amino acid composition and physicochemical properties (AAsCPP) feature, the PSSM feature, and the functional domain (FunD) composition feature. These feature extraction methods will be described as follows.

2.2.1. Definition of Secondary Structure in Proteins. The DSSP program was designed by Kabsch and Sander [44] and used to standardize protein secondary structure. The DSSP program works by calculating the most likely protein secondary structure given by the protein 3-dimensional structure. The specific principle of the DSSP program is calculating the H-bond energy between every two atoms by the atomic position in a PDB file, and then the most likely class of secondary structure for each residue can be determined by the best two H-bonds of each atom.

The DSSP feature extraction process is as follows. Firstly, DSSP entries are calculated from PDB entries by DSSP program. Secondly, the corresponding DSSP sequences from DSSP entries are obtained. DSSP sequence contains eight states (T, S, G, H, I, B, E, —), which can be divided into four groups, as shown in Table 1. Finally, according to the eight states and four groups, a 40D feature vector can be extracted from a DSSP sequence. The detail of the description and dimension of the features are shown in Table 2.

2.2.2. Amino Acids Composition and Physicochemical Properties. As effective features to describe a protein, the amino acid composition and physiochemical properties have reached good predict result, respectively [13, 34, 35]. Ding and Dubchak [13] tried to integrate the features for the first time and achieved a good result. Later, many other researchers proposed other feature integration methods. In 2013, Lin et al. [41] used a 188D feature vector combining amino acid composition and physiochemical properties. The 188D feature extraction method is also used in this paper.

The eight physiochemical properties of amino acids are hydrophobicity, van der Waals volume, polarity, polarizability, charge, surface tension, surface tension, and solvent accessibility. Different kinds of amino acids have different physiochemical properties so that they can be divided into three groups [13, 41], as shown in Table 3.

Table 1: The eight states of DSSP feature in four groups.

Eight-state SS	Code	Description	Four groups
3_{10} helix (G)	G	Helix-3	First
Alpha-helix (H)	H	Alpha helix	
pi-helix (I)	I	Helix-5	
Beta-strand (E)	E	Strand	Second
Beta-bridge (B)	B	Beta bridge	
Beta-turn (T)	T	Turn	Third
High curvature loop (S)	S	Bend	
Irregular (L)	—	Empty, no secondary structure assigned	Fourth

Table 2: The description and dimension of the DSSP feature.

Features description	Dimension
State composition	8
Group composition	4
Number of continuous states	8
Number of continuous groups	4
Number of continuous state compositions	8
Number of continuous group compositions	4
Alternate frequency between groups	4

The percentage composition of the 20 amino acids in the query protein forms a 20D feature vector. The group composition of amino acids (3D), the pairwise frequency between every two groups (3D), and the distribution pattern of constituents (where the first, 25%, 50%, 75%, and 100% of a given constituent are contained) (5 × 3D) from each physiochemical property are extracted. Therefore, we can get a 168D feature vector from a protein sequence according to the eight physiochemical properties. Adding up the 20D amino acid composition feature and the 168D physiochemical feature, we can get a 188D feature vector altogether. The name and the dimensions of the features are listed in Table 4.

2.2.3. Position Specific Scoring Matrix. PSSM is a relatively common feature. In addition to protein fold type classification research area, there are some studies on protein structural class prediction [67, 68] which used this feature. PSSM is derived from PSI-BLAST (Position Specific Iterative Basic Local Alignment Search Tool) [43] by taking the multiple sequence alignment of sequences in nonredundant protein sequence database (nrdb90) [69]. The iteration number is 3 and the cutoff E-value is 0.001. Two $L \times 20$ matrices can be obtained by PSI-BLAST, in which L represents the length of the query amino acid sequence, and 20 represents the 20 amino acids. One of the two matrices contains conservation scores of a given amino acid at a given position in sequence, and the other provides probability of occurrence of a given amino acid at a given position in the sequence. The PSSM feature is extracted from the former matrix. Suppose that the parameter in the matrix is S_{ij} ($i = 1, 2, \ldots, L$; $j = 1, 2, \ldots, 20$). Then the feature can be calculated by (1). That is to calculate the average value of each column in the matrix and form a 20D feature vector.

$$P_{\text{pssm}} = \left[\frac{\sum_{i=1}^{L} S_{i1}}{L}, \frac{\sum_{i=1}^{L} S_{i1}}{L}, \ldots, \frac{\sum_{i=1}^{L} S_{i20}}{L} \right]^{T}. \quad (1)$$

2.2.4. Functional Domain Composition. Proteins always contain some modules or domains, which involve different evolution resources and functions. Therefore, we can extract features in some FunD databases. There are some different FunD databases: SMART [70], Pfam [71], COG [72], KOG [72], and CDD [73]. In 2009, Shen and Chou [22] considered CDD as a relatively more complete functional domain database, and they used CDD to extract features. In this study, we used CDD (version 2.11), which co ntains 17402 common protein domains and families. Taking each of protein domains as a vector-base, we can extract a 17402D feature vector. Specific process is as follows. Firstly, use RPS-BLAST program [74] to compare the protein sequence with each of the 17402 domain sequences. Secondly, if the significance threshold value (expect value) is no more than 0.001, this component of the protein in the 17402D feature vector is assigned 1; otherwise, it is assigned 0. In this way, we can extract a 17402D feature vector, and each component of the feature can be either 1 or 0.

2.3. The Proposed Ensemble Classifier. In this study, we propose a novel ensemble strategy which includes 5 individual steps. Step 1: 10 widely used machine learning classifiers, LMT [75], RandomForest [34], LibSVM [76], SimpleLogistic [75], RotationForest [38], SMO [77], NaiveBayes [78], RandomTree [79], FT [80], and SimpleCart [81], are selected, and a 5-fold cross validation is implemented on the DD-dataset. Step 2: the classifier with the highest accuracy in each feature group is chosen. Step 3: corresponding models by training each feature group with the chosen classifier are selected. The four models are DSSP classification model, AAsCPP classification model, PSSM classification model, and FunD classification model. Detailed process is shown in Figure 1. Step 4: features from the test dataset are extracted and the classification result P_{ij} by calculating the corresponding models is obtained, i represents a kind of classification model ranging from 1 to 4, and j represents a kind of fold index, ranging from 1 to the total number of the fold classes (e.g.,

Table 3: The 20 amino acids divided into 3 groups according to their physiochemical properties.

Physicochemical property	The 1st group	The 2nd group	The 3rd group
Hydrophobicity	RKEDQN	GASTPHY	CVLIMFW
Van der Waals volume	GASCTPD	NVEQIL	MHKFRYW
Polarity	LIFWCMVF	PATGS	HQRKNED
Polarizability	GASDT	CPNVEQIL	KMHFRYW
Charge	KR	ANCQGHILMFPSTWYV	DE
Surface tension	GQDNAHR	KTSEC	ILMFPWYV
Secondary structure	EALMQKRH	VIYCWFT	GNPSD
Solvent accessibility	ALFCGIVW	RKQEND	MPSTHY

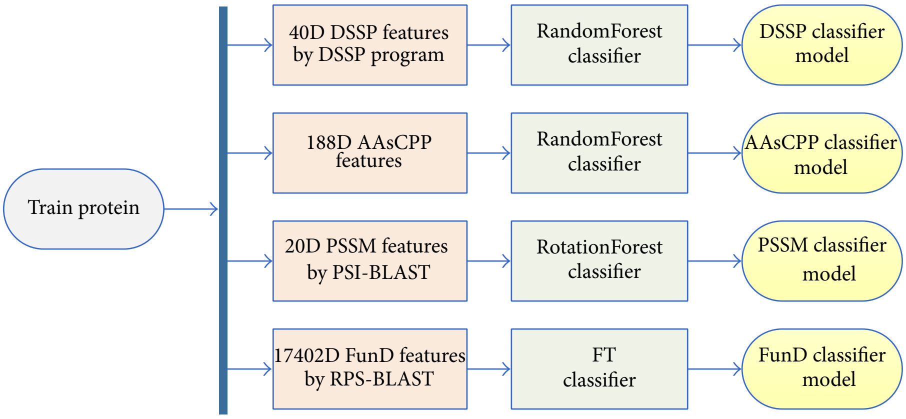

Figure 1: The training process of the four feature groups through the corresponding classifier.

Table 4: The name and the dimension of the amino acids composition and physiochemical features.

Feature name	Dimension
Amino acids composition	20
Hydrophobicity	21
Van der Waals volume	21
Polarity	21
Polarizability	21
Charge	21
Surface tension	21
Secondary structure	21
Solvent accessibility	21

the value of j ranges from 1 to 27 on DD-dataset). Step 5: the average of the probabilities of the four models in each fold class is calculated. The fold class with the highest probability will be chosen as the classification result. Detailed process is shown in Figure 2.

The machine learning tool we used is WEKA (Waikato Environment for Knowledge Analysis) [56], a collection of machine learning classifiers for data mining tasks based on Java.

2.4. Measurement. In this study, the standard Q percentage accuracy is used to test the effect of the proposed classification method, which helped us to compare our result with other researchers' results [12, 13, 34]. The definition of the standard Q percentage accuracy is described in

$$\begin{aligned} N &= n_1 + n_2 + \cdots + n_i + \cdots + n_k, \\ C &= c_1 + c_2 + \cdots + c_i + \cdots + c_k, \\ Q &= \frac{C}{N}, \end{aligned} \tag{2}$$

where n_i represents the number of the proteins which belong to class i, c_i represents the correct number in n_i test data, c_i/n_i represents the classification accuracy of class i, k represents the total number of classes, N represents the total number of tests, C represents the total number of the correct classified data, and Q represents the classification accuracy.

3. Results and Discussion

3.1. Performance of ProFold. In order to test the performance of proFold, we first select the widely used DD-dataset for evaluation. The overall accuracy is 76.2%. Comparison with existing ensemble learning methods on DD-dataset is shown in Table 5. From Table 5, we can see that the accuracy of the other methods are under 75%, and the accuracy of our method is 3% higher than PFPA (2015) [40], which is the best one in the other methods.

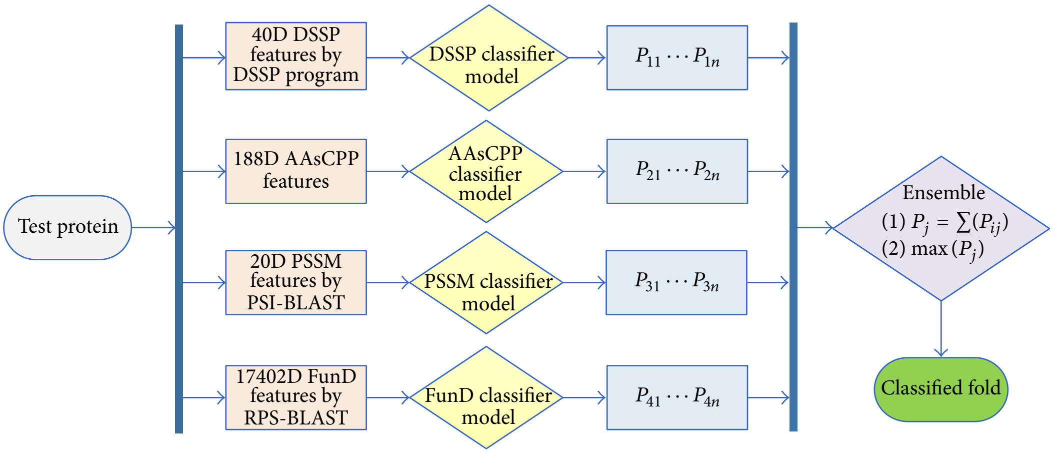

Figure 2: The ensemble process of calculating the test data through the models.

Table 5: Comparison with existing ensemble learning methods on DD-dataset.

Methods	References	Overall accuracy (%)
PFP-Pred	[12]	62.1
GAOEC	[11]	64.7
ThePFP-FunDSeqE	[22]	70.5
Dehzangi et al.	[34]	62.7
Dehzangi et al.	[38]	62.4
MarFold	[18]	71.7
PFP-RFSM	[35]	73.7
Feng and Hu	[36]	70.2
Feng et al.	[23]	70.8
PFPA	[40]	73.6
proFold (the proposed method)	*This paper*	*76.2*

Table 6: Comparison with the different methods on EDD-dataset by 10-fold cross validation.

Methods	References	Overall accuracy (%)
Paliwal et al.	[29]	90.6
Paliwal et al.	[30]	86.2
Dehzangi et al.	[31]	88.2
HMMFold	[32]	86.0
Saini et al.	[33]	89.9
Lyons et al.	[21]	92.9
proFold (the proposed method)	*This paper*	*93.2*

Table 7: Comparison with the different methods on TG-dataset by 10-fold cross validation.

Methods	References	Overall accuracy (%)
Paliwal et al.	[29]	77.0
Paliwal et al.	[30]	73.3
Dehzangi et al.	[31]	73.8
HMMFold	[32]	93.8
Saini et al.	[33]	74.5
NiRecor	[57]	84.6
Lyons et al.	[21]	85.6
proFold (the proposed method)	*This paper*	*94.3*

In order to further evaluate the performance of proFold, we also select another two large scale datasets: EDD-dataset and TG-dataset. Training and testing dataset are not clearly distinguished in the two datasets, so a k-fold cross validation is implemented on them.

We calculated the classification accuracy of EDD-dataset by 10-fold cross validation for 10 times and compared the result with other methods. The results are shown in Table 6. We can see from the table that only the accuracies of Paliwal et al. and Lyons et al. are more than 90%, which are lower than that of proFold. The result showed that the advantage of proFold is obvious when larger scale datasets are used for validation.

Regarding TG-dataset, we also took experiments by 10-fold cross validation for 10 times and compared the results with other methods. The results are shown in Table 7. We can see from the table that HMMFold (2015) method achieved the highest accuracy, which is 93.8%. The accuracy of our method is 94.3%, which is higher than HMMFold. TG-dataset has threefold classes more than DD-dataset and its scale is twice larger than DD-dataset. The result showed that the advantage of proFold is obvious when the dataset with more fold classes is tested.

3.2. Performance of the Proposed Ensemble Classifier. In the field of protein fold classification, many researchers used ensemble learning methods [11, 18, 22, 23, 34–36, 38, 46, 51, 54, 79, 82–89]. The specific process of those ensemble strategies is as follows. (1) Integrate all features. (2) Select several basic classifiers for training. (3) Propose an ensemble classifier

Table 8: The accuracy of 5-fold cross validation on the features extracted from DD-dataset using 10 basic classifiers.

Feature groups	Basic classifiers	Fivefold CV accuracy (%)
DSSP	LMT	43.0
	RandomForest*	51.3
	LibSVM	46.4
	SimpleLogistic	43.0
	RotationForest	49.7
	SMO	36.4
	NaiveBayes	43.4
	RandomTree	32.8
	FT	42.4
	SimpleCart	37.7
AAsCPC	LMT	32.5
	RandomForest*	35.4
	LibSVM	34.4
	SimpleLogistic	32.5
	RotationForest	27.7
	SMO	34.4
	NaiveBayes	28.3
	RandomTree	11.6
	FT	34.4
	SimpleCart	20.6
PSSM	LMT	56.3
	RandomForest	53.7
	LibSVM	57.2
	SimpleLogistic	55.9
	RotationForest*	56.1
	SMO	30.2
	NaiveBayes	42.4
	RandomTree	29.6
	FT	49.5
	SimpleCart	33.4
FunD	LMT	42.1
	RandomForest	43.1
	LibSVM	21.2
	SimpleLogistic	43.1
	RotationForest	41.8
	SMO	38.9
	NaiveBayes	38.3
	RandomTree	39.9
	FT*	44.1
	SimpleCart	34.7

*The basic classifier of each feature group with the highest accuracy.

according to the classification result probability of each basic classifier. In this study, we find that the redundancies of the features will influence the performance of those methods. Therefore, we propose a novel ensemble strategy.

We took experiments on DD-dataset. Firstly, extract four feature groups which have been tested in 10 basic classifiers by cross validation. The detailed information of the test results is listed in Table 8. We can see from Table 8 that the best classifier is RandomForest using the DSSP feature group and AAsCPP feature group. The best classifiers are RotaionForest and FT when PSSM and FunD features are implemented, respectively. Secondly, train the four feature groups with corresponding basic classifiers and get four models. Finally, test the models on DD-dataset. The overall accuracy is 76.2%. Our method improves the accuracy effectively compared with other existing ensemble learning methods.

Table 9: Comparison with the different ensemble strategies on three datasets.

Datasets	The accuracy of traditional ensemble strategy (%)	The accuracy of this paper ensemble strategy (%)
DD	72.5	76.2
EDD	89.9	93.2
TG	91.7	94.3

In order to compare our ensemble strategy with the traditional ensemble strategy, we took experiments on the four feature groups with traditional ensemble strategy. (1) Integrate the four feature groups. (2) Train the models with RandomForest, RotationForest, and FT respectively. (3) Test the models on DD-dataset, EDD-dataset, and TG-dataset. The classification accuracy of our ensemble strategy has increased by 3% to 4%, as shown in Table 9. The result showed that our ensemble strategy has a better classification performance.

3.3. Accuracy Improvements with the DSSP Feature. In order to evaluate the influence on importing the DSSP feature, we calculated the classification accuracy of each fold class with and without the DSSP feature, respectively, using the DD-dataset. The accuracies are shown in Table 10. From the table, we can see that the accuracies of some fold classes, such as Fold number 2, number 4, number 6, number 12, number 23, and number 26, have increased obviously after importing the DSSP feature. The overall accuracy has increased from 71.3% to 76.2%. For example, the protein chain 1FAPB in DD-dataset was incorrectly classified into Fold number 5 before importing the DSSP feature, and it was reclassified into Fold number 4 correctly after importing the DSSP feature. The results showed that the DSSP feature has a significant effect on protein structure classification.

As we know that PDB files contain protein 3D structure information, we started from the PDB file of the protein in this study. The DSSP feature is extracted from the 3D structure in PDB and the 3D structure of a protein is more stable. Thus it explains why the DSSP feature has a significant effect on the protein structure classification.

4. Conclusion

In this study, we proposed a novel method called proFold. ProFold is an ensemble classifier combining the protein structural and functional information. In terms of feature extraction, we imported the DSSP feature into protein fold

Table 10: The accuracy of each fold class with and without the DSSP feature.

Fold number	The accuracy without the DSSP feature	The accuracy with the DSSP feature
1	100.0	100.0
2*	88.9	100.0
3*	55.0	60.0
4*	62.5	87.5
5	88.9	88.9
6*	66.7	77.8
7*	77.3	84.1
8	66.7	66.7
9	92.3	92.3
10	66.7	66.7
11	50.0	50.0
12*	47.4	68.4
13	100.0	100.0
14	50.0	50.0
15	100.0	100.0
16*	91.7	93.8
17*	83.3	91.7
18*	38.5	46.2
19	85.2	85.2
20	50.0	50.0
21	87.5	87.5
22	58.3	58.3
23*	57.1	71.4
24	100.0	100.0
25	25.0	25.0
26*	44.4	59.3
27*	92.6	96.3
Overall	71.3	76.2

*The fold class of which the accuracy has increased significantly after importing the DSSP feature.

classification for the first time. Experiments showed that the classification accuracy will increase by about 5% using the DD-dataset by importing the DSSP feature. In terms of classification method, we proposed a novel ensemble classifier and improved the classification accuracy with this method. The classification accuracies of proFold on DD-, EDD-, and TG-dataset are 76.2%, 93.2%, and 94.3%, respectively, which are higher than the existing similar methods. The results showed that proFold is a relatively better classifier.

Competing Interests

The authors declare that there are no competing interests regarding the publication of this paper.

Acknowledgments

This work was supported in part by grants from National Natural Science Foundation of China (Grant no. 61303099).

References

[1] H. S. Chan and K. A. Dill, "The protein folding problem," *Physics Today*, vol. 46, no. 2, pp. 24–32, 1993.

[2] D. T. Jones, W. R. Taylor, and J. M. Thornton, "A new approach to protein fold recognition," *Nature*, vol. 358, no. 6381, pp. 86–89, 1992.

[3] J. Xu, M. Li, D. Kim, and Y. Xu, "RAPTOR: optimal protein threading by linear programming," *Journal of Bioinformatics and Computational Biology*, vol. 1, no. 1, pp. 95–117, 2003.

[4] J. Söding, A. Biegert, and A. N. Lupas, "The HHpred interactive server for protein homology detection and structure prediction," *Nucleic Acids Research*, vol. 33, supplement 2, pp. W244–W248, 2005.

[5] J. Cheng and P. Baldi, "A machine learning information retrieval approach to protein fold recognition," *Bioinformatics*, vol. 22, no. 12, pp. 1456–1463, 2006.

[6] W. Zhang, S. Liu, and Y. Zhou, "SP^5: improving protein fold recognition by using torsion angle profiles and profile-based gap penalty model," *PLoS ONE*, vol. 3, no. 6, Article ID e2325, 2008.

[7] R.-X. Yan, J.-N. Si, C. Wang, and Z. Zhang, "DescFold: a web server for protein fold recognition," *BMC Bioinformatics*, vol. 10, article 416, 2009.

[8] 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, no. 19, pp. 8700–8704, 1995.

[9] I. Dubchak, I. Muchnik, C. Mayor, I. Dralyuk, and S.-H. Kim, "Recognition of a protein fold in the context of the SCOP classification," *Proteins: Structure, Function, and Bioinformatics*, vol. 35, no. 4, pp. 401–407, 1999.

[10] Q. Dong, S. Zhou, and J. Guan, "A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation," *Bioinformatics*, vol. 25, no. 20, pp. 2655–2662, 2009.

[11] X. Guo and X. Gao, "A novel hierarchical ensemble classifier for protein fold recognition," *Protein Engineering, Design and Selection*, vol. 21, no. 11, pp. 659–664, 2008.

[12] H.-B. Shen and K.-C. Chou, "Ensemble classifier for protein fold pattern recognition," *Bioinformatics*, vol. 22, no. 14, pp. 1717–1722, 2006.

[13] C. H. Q. Ding and I. Dubchak, "Multi-class protein fold recognition using support vector machines and neural networks," *Bioinformatics*, vol. 17, no. 4, pp. 349–358, 2001.

[14] K. Chen and L. Kurgan, "PFRES: protein fold classification by using evolutionary information and predicted secondary structure," *Bioinformatics*, vol. 23, no. 21, pp. 2843–2850, 2007.

[15] L. Wei, M. Liao, Y. Gao, R. Ji, Z. He, and Q. Zou, "Improved and promising identificationof human microRNAs by incorporatinga high-quality negative set," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 11, no. 1, pp. 192–201, 2014.

[16] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, "SCOP: a structural classification of proteins database for the investigation of sequences and structures," *Journal of Molecular Biology*, vol. 247, no. 4, pp. 536–540, 1995.

[17] K.-C. Chou, "Prediction of protein cellular attributes using pseudo-amino acid composition," *Proteins: Structure, Function, and Bioinformatics*, vol. 43, no. 3, pp. 246–255, 2001.

[18] T. Yang, V. Kecman, L. Cao, C. Zhang, and J. Zhexue Huang, "Margin-based ensemble classifier for protein fold recognition," *Expert Systems with Applications*, vol. 38, no. 10, pp. 12348–12355, 2011.

[19] M. T. A. Shamim, M. Anwaruddin, and H. A. Nagarajaram, "Support Vector Machine-based classification of protein folds using the structural properties of amino acid residues and amino acid residue pairs," *Bioinformatics*, vol. 23, no. 24, pp. 3320–3327, 2007.

[20] P. Deschavanne and P. Tufféry, "Enhanced protein fold recognition using a structural alphabet," *Proteins: Structure, Function and Bioinformatics*, vol. 76, no. 1, pp. 129–137, 2009.

[21] J. Lyons, K. K. Paliwal, A. Dehzangi, R. Heffernan, T. Tsunoda, and A. Sharma, "Protein fold recognition using HMM–HMM alignment and dynamic programming," *Journal of Theoretical Biology*, vol. 393, pp. 67–74, 2016.

[22] H.-B. Shen and K.-C. Chou, "Predicting protein fold pattern with functional domain and sequential evolution information," *Journal of Theoretical Biology*, vol. 256, no. 3, pp. 441–446, 2009.

[23] Z. Feng, X. Hu, Z. Jiang, H. Song, and M. A. Ashraf, "The recognition of multi-class protein folds by adding average chemical shifts of secondary structure elements," *Saudi Journal of Biological Sciences*, vol. 23, no. 2, pp. 189–197, 2016.

[24] J.-Y. Yang and X. Chen, "Improving taxonomy-based protein fold recognition by using global and local features," *Proteins: Structure, Function, and Bioinformatics*, vol. 79, no. 7, pp. 2053–2064, 2011.

[25] T. Damoulas and M. A. Girolami, "Probabilistic multi-class multi-kernel learning: on protein fold recognition and remote homology detection," *Bioinformatics*, vol. 24, no. 10, pp. 1264–1270, 2008.

[26] W. Chmielnicki and K. Stąpor, "A hybrid discriminative/generative approach to protein fold recognition," *Neurocomputing*, vol. 75, no. 1, pp. 194–198, 2012.

[27] L. Liu, X.-Z. Hu, X.-X. Liu, Y. Wang, and S.-B. Li, "Predicting protein fold types by the general form of Chou's pseudo amino acid composition: approached from optimal feature extractions," *Protein and Peptide Letters*, vol. 19, no. 4, pp. 439–449, 2012.

[28] A. Sharma, J. Lyons, A. Dehzangi, and K. K. Paliwal, "A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition," *Journal of Theoretical Biology*, vol. 320, pp. 41–46, 2013.

[29] K. K. Paliwal, A. Sharma, J. Lyons, and A. Dehzangi, "Improving protein fold recognition using the amalgamation of evolutionary-based and structural based information," *BMC Bioinformatics*, vol. 15, supplement 16, p. S12, 2014.

[30] K. K. Paliwal, A. Sharma, J. Lyons, and A. Dehzangi, "A tri-gram based feature extraction technique using linear probabilities of position specific scoring matrix for protein fold recognition," *IEEE Transactions on Nanobioscience*, vol. 13, no. 1, pp. 44–50, 2014.

[31] A. Dehzangi, K. Paliwal, J. Lyons, A. Sharma, and A. Sattar, "A segmentation-based method to extract structural and evolutionary features for protein fold recognition," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 11, no. 3, pp. 510–519, 2014.

[32] J. Lyons, A. Dehzangi, R. Heffernan et al., "Advancing the accuracy of protein fold recognition by utilizing profiles from hidden markov models," *IEEE Transactions on NanoBioscience*, vol. 14, no. 7, pp. 761–772, 2015.

[33] H. Saini, G. Raicar, A. Sharma et al., "Probabilistic expression of spatially varied amino acid dimers into general form of Chou's pseudo amino acid composition for protein fold recognition," *Journal of Theoretical Biology*, vol. 380, pp. 291–298, 2015.

[34] A. Dehzangi, S. Phon-Amnuaisuk, and O. Dehzangi, "Using random forest for protein fold prediction problem: an empirical study," *Journal of Information Science and Engineering*, vol. 26, no. 6, pp. 1941–1956, 2010.

[35] J. Li, J. Wu, and K. Chen, "PFP-RFSM: protein fold prediction by using random forests and sequence motifs," *Journal of Biomedical Science and Engineering*, vol. 6, no. 12, pp. 1161–1170, 2013.

[36] Z. Feng and X. Hu, "Recognition of 27-class protein folds by adding the interaction of segments and motif information," *BioMed Research International*, vol. 2014, Article ID 262850, 9 pages, 2014.

[37] L. Wei, M. Liao, X. Gao, and Q. Zou, "An improved protein structural classes prediction method by incorporating both sequence and structure information," *IEEE Transactions on Nanobioscience*, vol. 14, no. 4, pp. 339–349, 2015.

[38] A. Dehzangi, S. Phon-Amnuaisuk, M. Manafi, and S. Safa, "Using rotation forest for protein fold prediction problem: an empirical study," in *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*, C. Pizzuti, M. D. Ritchie, and M. Giacobini, Eds., vol. 6023 of *Lecture Notes in Computer Science*, pp. 217–227, Springer, Berlin, Germany, 2010.

[39] G. Bologna and R. D. Appel, "A comparison study on protein fold recognition," in *Proceedings of the 9th International Conference on Neural Information Processing (ICONIP '02)*, pp. 2492–2496, IEEE, Singapore, November 2002.

[40] L. Wei, M. Liao, X. Gao, and Q. Zou, "Enhanced protein fold prediction method through a novel feature extraction technique," *IEEE Transactions on NanoBioscience*, vol. 14, no. 6, pp. 649–659, 2015.

[41] C. Lin, Y. Zou, J. Qin et al., "Hierarchical classification of protein folds using a novel ensemble classifier," *PLoS ONE*, vol. 8, no. 2, article e56499, 2013.

[42] N. K. Fox, S. E. Brenner, and J.-M. Chandonia, "SCOPe: structural classification of proteins—extended, integrating SCOP and ASTRAL data and classification of new structures," *Nucleic Acids Research*, vol. 42, no. 1, pp. D304–D309, 2014.

[43] S. F. Altschul, T. L. Madden, A. A. Schäffer et al., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[44] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, no. 12, pp. 2577–2637, 1983.

[45] H. Lin, E.-Z. Deng, H. Ding, W. Chen, and K.-C. Chou, "IPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition," *Nucleic Acids Research*, vol. 42, no. 21, pp. 12961–12972, 2014.

[46] J. Jia, Z. Liu, X. Xiao, B. Liu, and K.-C. Chou, "iPPI-Esml: an ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC," *Journal of Theoretical Biology*, vol. 377, pp. 47–56, 2015.

[47] W.-R. Qiu, X. Xiao, and K.-C. Chou, "iRSpot-TNCPseAAC: identify recombination spots with trinucleotide composition

and pseudo amino acid components," *International Journal of Molecular Sciences*, vol. 15, no. 2, pp. 1746–1766, 2014.

[48] W. Chen, H. Ding, P. Feng, H. Lin, and K.-C. Chou, "iACP: a sequence-based tool for identifying anticancer peptides," *Oncotarget*, vol. 7, no. 13, pp. 16895–16909, 2016.

[49] J. Jia, Z. Liu, X. Xiao, B. Liu, and K.-C. Chou, "iSuc-PseOpt: identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset," *Analytical Biochemistry*, vol. 497, pp. 48–56, 2016.

[50] B. Liu, L. Fang, F. Liu, X. Wang, and K.-C. Chou, "IMiRNA-PseDPC: microRNA precursor identification with a pseudo distance-pair composition approach," *Journal of Biomolecular Structure and Dynamics*, vol. 34, no. 1, pp. 220–232, 2016.

[51] J. Jia, Z. Liu, X. Xiao, B. Liu, and K.-C. Chou, "pSuc-Lys: predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach," *Journal of Theoretical Biology*, vol. 394, pp. 223–230, 2016.

[52] B. Liu, L. Fang, R. Long, X. Lan, and K.-C. Chou, "iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition," *Bioinformatics*, vol. 32, no. 3, pp. 362–369, 2016.

[53] J. Jia, Z. Liu, X. Xiao, B. Liu, and K. Chou, "iCar-PseCp: identify carbonylation sites in proteins by Monto Carlo sampling and incorporating sequence coupled effects into general PseAAC," *Oncotarget*, vol. 7, no. 23, pp. 34558–34570, 2016.

[54] B. Liu, R. Long, and K.-C. Chou, "iDHS-EL: identifying DNase I hypersensitive sites by fusing three different modes of pseudo nucleotide composition into an ensemble learning framework," *Bioinformatics*, vol. 32, no. 16, pp. 2411–2418, 2016.

[55] W.-R. Qiu, B.-Q. Sun, X. Xiao, D. Xu, and K.-C. Chou, "iPhos-PseEvo: identifying human phosphorylated proteins by incorporating evolutionary information into general PseAAC via grey system theory," *Molecular Informatics*, 2016.

[56] K.-C. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 273, no. 1, pp. 236–247, 2011.

[57] N. J. Cheung, X.-M. Ding, and H.-B. Shen, "Protein folds recognized by an intelligent predictor based-on evolutionary and structural information," *Journal of Computational Chemistry*, vol. 37, no. 4, pp. 426–436, 2016.

[58] J. Lyons, N. Biswas, A. Sharma, A. Dehzangi, and K. K. Paliwal, "Protein fold recognition by alignment of amino acid residues using kernelized dynamic time warping," *Journal of Theoretical Biology*, vol. 354, pp. 137–145, 2014.

[59] K. Kavousi, B. Moshiri, M. Sadeghi, B. N. Araabi, and A. A. Moosavi-Movahedi, "A protein fold classifier formed by fusing different modes of pseudo amino acid composition via PSSM," *Computational Biology and Chemistry*, vol. 35, no. 1, pp. 1–9, 2011.

[60] P. Du, X. Wang, C. Xu, and Y. Gao, "PseAAC-Builder: a cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions," *Analytical Biochemistry*, vol. 425, no. 2, pp. 117–119, 2012.

[61] D.-S. Cao, Q.-S. Xu, and Y.-Z. Liang, "Propy: a tool to generate various modes of Chou's PseAAC," *Bioinformatics*, vol. 29, no. 7, pp. 960–962, 2013.

[62] S.-X. Lin and J. Lapointe, "Theoretical and experimental biology in one—a symposium in honour of Professor Kuo-Chen Chou's 50th anniversary and Professor Richard Giegé's 40th anniversary of their scientific careers," *Journal of Biomedical Science and Engineering*, vol. 6, no. 4, pp. 435–442, 2013.

[63] W. Chen, T.-Y. Lei, D.-C. Jin, H. Lin, and K.-C. Chou, "PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition," *Analytical Biochemistry*, vol. 456, no. 1, pp. 53–60, 2014.

[64] W. Chen, X. Zhang, J. Brooker, H. Lin, L. Zhang, and K.-C. Chou, "PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions," *Bioinformatics*, vol. 31, no. 1, pp. 119–120, 2015.

[65] B. Liu, F. Liu, L. Fang, X. Wang, and K.-C. Chou, "RepDNA: a Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects," *Bioinformatics*, vol. 31, no. 8, pp. 1307–1309, 2015.

[66] B. Liu, F. Liu, X. Wang, J. Chen, L. Fang, and K.-C. Chou, "Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences," *Nucleic Acids Research*, vol. 43, no. 1, pp. W65–W71, 2015.

[67] L. Zhang, X. Zhao, and L. Kong, "Predict protein structural class for low-similarity sequences by evolutionary difference information into the general form of Chou's pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 355, pp. 105–110, 2014.

[68] S. Zhang, "Accurate prediction of protein structural classes by incorporating PSSS and PSSM into Chou's general PseAAC," *Chemometrics and Intelligent Laboratory Systems*, vol. 142, pp. 28–35, 2015.

[69] L. Holm and C. Sander, "Removing near-neighbour redundancy from large protein sequence collections," *Bioinformatics*, vol. 14, no. 5, pp. 423–429, 1998.

[70] I. Letunic, R. R. Copley, B. Pils, S. Pinkert, J. Schultz, and P. Bork, "SMART 5: domains in the context of genomes and networks," *Nucleic Acids Research*, vol. 34, supplement 1, pp. D257–D260, 2006.

[71] R. D. Finn, J. Mistry, B. Schuster-Böckler et al., "Pfam: clans, web tools and services," *Nucleic Acids Research*, vol. 34, supplement 1, pp. D247–D251, 2006.

[72] R. L. Tatusov, N. D. Fedorova, J. D. Jackson et al., "The COG database: an updated version includes eukaryotes," *BMC Bioinformatics*, vol. 4, article 41, 2003.

[73] A. Marchler-Bauer, J. B. Anderson, M. K. Derbyshire et al., "CDD: a conserved domain database for interactive domain family analysis," *Nucleic Acids Research*, vol. 35, no. 1, pp. D237–D240, 2007.

[74] A. A. Schäffer, L. Aravind, T. L. Madden et al., "Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements," *Nucleic Acids Research*, vol. 29, no. 14, pp. 2994–3005, 2001.

[75] N. Landwehr, M. Hall, and E. Frank, "Logistic model trees," *Machine Learning*, vol. 59, no. 1-2, pp. 161–205, 2005.

[76] 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, article 27, 2011.

[77] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy, "Improvements to Platt's SMO algorithm for SVM classifier design," *Neural Computation*, vol. 13, no. 3, pp. 637–649, 2001.

[78] G. H. John and P. Langley, "Estimating continuous distributions in Bayesian classifiers," in *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, pp. 338–345, Morgan Kaufmann, 1995.

[79] J. J. Rodríguez, L. I. Kuncheva, and C. J. Alonso, “Rotation forest: a new classifier ensemble method,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 10, pp. 1619–1630, 2006.

[80] J. Gama, “Functional trees,” *Machine Learning*, vol. 55, no. 3, pp. 219–250, 2004.

[81] A. Liaw and M. Wiener, “Classification and regression by randomForest,” *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[82] K.-C. Chou and H.-B. Shen, “Predicting protein subcellular location by fusing multiple classifiers,” *Journal of Cellular Biochemistry*, vol. 99, no. 2, pp. 517–527, 2006.

[83] K.-C. Chou and H.-B. Shen, “Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers,” *Journal of Proteome Research*, vol. 5, no. 8, pp. 1888–1897, 2006.

[84] K.-C. Chou and H.-B. Shen, “Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides,” *Biochemical and Biophysical Research Communications*, vol. 357, no. 3, pp. 633–640, 2007.

[85] H.-B. Shen and K.-C. Chou, “Gpos-PLoc: an ensemble classifier for predicting subcellular localization of Gram-positive bacterial proteins,” *Protein Engineering, Design and Selection*, vol. 20, no. 1, pp. 39–46, 2007.

[86] H.-B. Shen and K.-C. Chou, “Hum-mPLoc: an ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites,” *Biochemical and Biophysical Research Communications*, vol. 355, no. 4, pp. 1006–1011, 2007.

[87] H.-B. Shen, J. Yang, and K.-C. Chou, “Euk-PLoc: an ensemble classifier for large-scale eukaryotic protein subcellular location prediction,” *Amino Acids*, vol. 33, no. 1, pp. 57–67, 2007.

[88] H.-B. Shen and K.-C. Chou, “QuatIdent: a web server for identifying protein quaternary structural attribute by fusing functional domain and sequential evolution information,” *Journal of Proteome Research*, vol. 8, no. 3, pp. 1577–1584, 2009.

[89] J. Jia, Z. Liu, X. Xiao, B. Liu, and K.-C. Chou, “iPPBS-Opt: a sequence-based ensemble classifier for identifying protein-protein binding sites by optimizing imbalanced training datasets,” *Molecules*, vol. 21, no. 1, p. 95, 2016.