Protein Fold Recognition Using Segmentation-Based Feature Extraction Model

Abdollah Dehzangi[1,2] and Abdul Sattar[1,2]

[1] Institute for Integrated and Intelligent Systems (IIIS), Griffith University,
Brisbane, Australia
[2] National ICT Australia (NICTA), Brisbane, Australia
{a.dehzangi,a.sattar}@griffith.edu.au

Abstract. Protein Fold recognition (PFR) is considered as an important step towards protein structure prediction. It also provides significant information about general functionality of a given protein. Despite all the efforts have been made, PFR still remains unsolved. It is shown that appropriately extracted features from the physicochemical-based attributes of the amino acids plays crucial role to address this problem. In this study, we explore 55 different physicochemical-based attributes using two novel feature extraction methods namely segmented distribution and segmented density. Then, by proposing an ensemble of different classifiers based on the AdaBoost.M1 and Support Vector Machine (SVM) classifiers which are diversely trained on different combinations of features extracted from these attributes, we outperform similar studies found in the literature for over 2% for the PFR task.

Keywords: Segmented distribution, Segmented Density, Physicochemical-based features, SVM, AdaBoost.M1, Ensemble of different Classifiers.

1 Introduction

Determining how a given protein is categorized to a fold based on its major secondary structure is called *Protein Fold Recognition (PFR)*. PFR is considered as an important step toward protein structure prediction. It also can provide significant information about general functionality of proteins. During the past few decades, a wide range of approaches proposed to solve PFR mainly based on classification techniques [1–4] as well as feature extraction methods [5–8]. Among the classification techniques used to tackle this problem, ensemble-based classifiers attained the best results for PFR [2, 9, 10]. They outperformed individual classifier used for this task which have driven the focus to these techniques [9–11].

Beside classification techniques used to approach PFR, feature extraction have also attained tremendous attention. Among the features being used for this task, physicochemical-based features (extracted based on physicochemical attributes of the amino acids (e.g. hydrophobicity)) showed promising results. It was shown that dissimilar to the other features used to tackle this problem (e.g. sequential

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 345–354, 2013.

based features which are extracted from the alphabetic sequence of the amino acids) physicochemical-based features maintain their discriminatory information when the sequential similarity rate is low. To the best of our knowledge, The impact of the widest range of physicochemical-based attributes for PFR was explored by Gromiha and his co-workers [7]. They explored 49 different physicochemical attributes of the amino acids using global density feature. Therefore, they extracted 49 features from these attributes. However, due to use of global density for feature extraction which just generate one global feature, they could not explore the local impact of the explored features properly.

To address this issue, later studies shifted their focus to explore fewer number of physicochemical-based attributes and instead, use more efficient feature extraction methods [5,12]. Recent studies shifted the focus to increase the number of attributes being explored as well as providing adequate local discriminatory information by categorizing amino acids into several subgroups based on the concept of alphabet reduction [8]. However, due to use of alphabet reduction, they discarded important information and could not appropriately enhance PFR. Furthermore, similar to their previous works, they tried to extract local information from the whole protein sequence as a single building block which failed to work properly, specially for large proteins.

In this study, we explore 55 different physicochemical-based attributes for PFR. To the best of our knowledge, most of these attributes have not been adequately explored for this task. We also propose two novel segmented base feature extraction methods which are aimed to provide more local discriminatory information than previously proposed approaches for PFR. We explore the impact of our propose approaches using four popular classification techniques namely, *AdaBoost.M1, SVM (SVM), Random Forest*, and *Naive Bayes* which have attained promising results for this task. In the final step, by proposing an ensemble of different classifiers (based on AdaBoost.M1 and SVM) which are diversely trained with the features extracted from a wide range of physicochemical-based attributes, we enhance the protein fold prediction accuracy for more than 2% better than similar studies found in the literature.

2 Datasets and Physicochemical-Based Attributes

In this study, two popular benchmarks namely EDD (extended version of the DD introduced by [5]), and TG (introduced by [13])are used. To be able to directly compare our results with the similar studies found in the literature, the EDD data set is used. We extract this data set from the latest version of the *Structural Classification of Proteins (SCOP 1.75)* consisting of 3418 proteins with less than 40% sequential similarities belonging to 27 fold used previously in DD (similar to [11]). We also use TG benchmark to be able to explore the impact of our proposed approaches when the sequential similarity rate is low. This dataset consists of 1612 proteins with less than 25% sequential similarities belonging to 30 folds. Similar to DD dataset which consists of two separate training and testing sets, we randomly separate the proteins in these datasets to training (3/5

Table 1. Names and number of the explored attributes in this study

No.	Attributes	No.	Attributes
1	Structure derived hydrophobicity value	29	Absolute entropy
2	Polarizability	30	Entropy of formation
3	Normalized frequency of α-helix	31	Buried and accessible molar fraction ratio
4	Normalized frequency of β-strand	32	Energy of transfer from inside to outside
5	Normalized frequency of βturn	33	Flexibility for one rigid residue
6	Hydrophobicity at ph 7.5 by HPLC	34	Side chain interaction parameter
7	Size	35	Side chain volume
8	Consensus normalized hydrophobicity scale	36	Hydropathy index
9	Hyd. index base on helix in membrane	37	Average surrounding hydrophobicity
10	Molecular weight	38	Average reduced distance for side chain
11	Hydrophobic parameter	39	Side chain orientation angle
12	Van Der Waals volume	40	Ave number of nearest neighbor in chain
13	Polarity (driven from amino acids)	41	Average Volume of surrounding residues
14	Volume	42	Hyd. scale (contact energy in 3D data)
15	Compressibility	43	Partition coefficient
16	Average long range contact energy	44	Average gain in surrounding hydrophobicity
17	Average medium range contact energy	45	Surrounding hydrophobicity in α-helix
18	Long range non bounded energy	46	Surrounding hydrophobicity in β-sheet
19	Mean RMS fluctuational displacement	47	Surrounding hydrophobicity in βturn
20	Refractive index	48	Surrounding hydrophobicity in folded form
21	Solvent accessible reduction	49	Average number of surrounding residues
22	Total non bounded energy	50	Membrane buried helix parameter
23	Unfolding entropy change of hydration	51	Mean fractional area loss (f)
24	Unfolding hydration heat capacity change	52	Flexibility
25	Retention coefficient (PH = 7.0)	53	Hydration potential (PH = 7.0)
26	Amino acids partition energy	54	Bulkiness
27	PKa-COOH	55	Polarity (driven from amino acids in proteins)
28	Hyd. value (driven from free amino acids)	-	-

of total proteins) and testing (2/5 of total proteins) to be able to simulate DD dataset's condition.

We also study 55 different physicochemical-based attributes as listed in Table 1 and explore their effectiveness on PFR. These attributes are taken from the APD database [14], and Gromiha and his co-workers study [7]. Our aim in this part is to explore the potential of each attribute to enhance PFR performance with respect to the feature extraction methods being used.

3 Physicochemical-Based Feature Extraction Approaches

In this study, we propose two novel feature extraction methods namely segmented-based density and segmented-based distribution. Our propose approaches are aimed to capture more local discriminatory information compared to previously proposed approaches [5]. These approaches are discussed in the following subsections.

3.1 Segmented Density

This method is mainly proposed to add more local discriminatory information based on the density of a given attribute. In this approach, we replace the amino acids in the original protein sequence (A_1, A_2, ..., A_L where L is the length of the protein) by the attribute values (R_1, R_2, ..., R_L) assigned to the amino acids (e.g. hydrophobicity). Then we segment the protein sequence and calculate the density for each segment. In this study, $K = 5$ segmentation factor is used due to its better performance compared to use of $K = 10$ and $K = 25$ explored experimentally. Hence, protein sequence divided to 20 segments. The expression for segmented density for each segment can be given as follows:

$$SD_{segmented\_density} = \frac{\sum_{i=1}^{D} R_i}{D}, \tag{1}$$

where D $(= L \times (5/100))$ is the length of each segment. Therefore, 20 segmented density features are extracted based on the given method. We also added the global density to these features to add global information to this feature set(20 + 1 features). The expression for global density is given as follows:

$$D_{glob\_density} = \frac{\sum_{i=1}^{L} R_i}{L}. \tag{2}$$

3.2 Segmented Distribution

as it is shown in previous subsection, in segmented density, the segments has equal length. Therefore, the length of segments vary crucially relying on the length of proteins. In this section, we propose a novel feature extraction method based on the concept of segmented-based distribution. In this method, we first calculate the total sum of attribute values (e.g. hydrophobicity) over a given protein sequence which is equal to $T = \sum_{i=1}^{L} R_i$. Then starting from the left side of the protein sequence, we sum the attributes values of the first $I_k^{(l)}$ amino acids until reaching to $K\%$ of T $(T_{seg} \leq (T \times K)/100)$. Then we return the distribution feature of this segment as $I_k^{(l)}/L$. We repeat this procedure for *2K, 3K, ... ,* until reaching to $N \times K = 50$ and calculate the $I_{2k}^{(l)}, I_{2k}^{(l)}, ..., I_{50}^{(l)}$ and then return the $I_{2k}^{(1)}/L, I_{2k}^{(1)}/L, ..., I_{50}^{(1)}/L$ as the assigned distribution features, respectively. The same procedure is done from the right sight to calculate $I_{2k}^{(r)}, I_{2k}^{(r)}, ..., I_{50}^{(r)}$ and then return the $I_{2k}^{(r)}/L, I_{2k}^{(r)}/L, ..., I_{50}^{(r)}/L$ as the assigned distribution features, respectively. Therefore, totally $N_{feat} = 2 \times (50/K) = 100/K$ features are extracted based on a given K in this feature set. The distribution factor *(*K) is a parameter which is determined here experimentally. For this, three values of K (5, 10, and 25) are investigated. To this set of *100/K* distribution features, we add the global density feature to provide more global information. Therefore, we have a total of $N_{feat} + 1$ features. Thus there will be 21, 11, and 6 features for *K=5, 10* and *25*, respectively.

Our proposed physicochemical-based feature extraction methods have two main contributions. First, they provide more local discriminatory information compared to previously adopted methods [7]. Second, instead of categorizing amino acids based on a given attributes to sub groups (as it was adopted in [5]), they work directly with the attributes values assigned to the amino acids. Therefore, they avoid information loss due to alphabet reduction.

4 Classification Techniques

In this study, four classifiers namely, AdaBoost.M1, Naive Bayes, Random Forest, and SVM that attained promising results for PFR used to evaluate the performance of the explored attributes with respect to our proposed feature extraction methods [4, 5, 9]. These classifiers are briefly described as follows:

Naive Bayes: As a kind of a Baysian-Based learner is considered as one of the simplest classifiers yet attained promising results for different tasks as well as PFR [2]. Naive Bayes is based on the assumption of independency of the employed features from each other to calculate the posterior probability [2].

AdaBoost.M1: Is considered as the best-of-the-shelf meta-classifier introduced by [15]. The main idea of the AdaBoost.M1 is to sequentially (in I iterations) apply a base learner (also called weak learner which refer to a classifier that at least performs better than random guess) on the bootstrap samples of data, adjust the weight of misclassified samples, and enhance the performance in each step. In this study, Adaboost.M1 implemented in WEKA using C4.5 decision tree (number of base learners is set to 100 (*I=100*)) as its base learners is employed [16].

Random Forest: Is also considered as a kind of meta-learner which recently attracted tremendous attention specifically for PFR [4]. Random Forest is based on bagging approaches [17]. It applies a base learner independently on B different bootstrap sample of data using randomly selected subset of features. In this study, for the Random Forest (implemented in WEKA) the number of iteration is set to 100 (*k=100*) and random tree based on the gain ratio is used as its base learner [4].

Support Vector Machine: SVM is considered as the state-of-the-art classification techniques which also attained the best results for PFR [11]. It aims at minimizing the classification error by finding the *Maximal Marginal Hyperplane (MMH)* based on the concept of support vector theory. To find the appropriate support vector, it transforms the input data using the concept of kernel function. In this study, we use SVM with *Sequential Minimal Optimization (SMO)* as a kind of polynomial kernel (implemented in WEKA) which its kernel degree is set to one (*p=1*).

Note that we also used the *Ensemble of Different Classifiers (EDC)* that we proposed in our previous work [2] which attained promising results for similar studies [5, 8]. This classifier consists of five different classifiers (Adaboost.M1, LogitBoost, Naive Bayes, *Multi Layer Perceptron (MLP)*, and SVM) which are trained on the same set of features and combined using majority voting as its algebraic combiner. This classifier is used in this study to evaluated the performance of our proposed approaches. It also used as a tie breaker in the diversely trained ensemble of classifiers proposed in this study.

5 Results and Discussion

To explore the effectiveness of the proposed approaches in this study, we first extract corresponding features to our proposed feature extraction methods for all 55 physicochemical-based attributes explored in this study. Therefore, for a given attribute, a feature group consisting of 21 features is extracted using segmented-based density method and three feature groups consisting of 5, 11, and 21 features are extracted using segmented-based distribution with three different distribution factors explored in this study (*K = 25, 10* and *5*, respectively). We then applied Adaboost.M1, Random Forest, Rotation Forest, and SVM to each feature group. Therefore, for a given attribute, 16 different experiments have been conducted (four classifiers applied to four extracted feature groups).

From the achieve results, we first explore the effectiveness of segmentation factor on the segmented-based distribution method. This experiment is conducted in the following manner. We calculate the average and maximum prediction accuracies achieved for each classifier used in this step with respect to the segmentation factor used in the segmentation-based distribution method for all of the 55 attributes. For example, first we apply SVM to the feature groups extracted from all 55 attributes using segmentation-based distribution method with *K = 5* separately. Then, we calculate the average and maximum prediction accuracies for all of the 55 achieved results. In the similar manner, SVM is used to feature groups extracted from all 55 attributes using segmentation-based distribution method with *K = 10* and then for *K = 25* separately. Then again, we calculate the average and maximum prediction accuracies for all of the 55 achieved results with respect to *K = 10* and again for all of the 55 achieved results with respect to *K = 25*. In result, 12 maximum and average prediction accuracies are calculated (four average and four maximum prediction accuracies corresponding to three variation of segmentation-based distribution method for SVM, Naive Bayes, AdaBoost.M1, and Random Forest).

In continuation, for a given classifier, we subtract maximum and average values calculated using segmented-based distribution with *K=25* feature extraction method from the average and maximum values calculated using segmented-based distribution with *K=10* as well as *K=5*. the results achieved in this step are shown in Table 2. As it is shown in this table, by adding just few features by adjusting segmentation factor from 25% to 5%, for the average, up to 6.9% for EDD dataset and 8.3% for TG dataset prediction enhancements and for the maximum, up to 12.3% for the EDD dataset and 11.5% for the TG dataset prediction enhancements are achieved. Similarly, by adjusting the distribution factor from 25% to 10%, for the average up to 4.8% for EDD dataset and 5.7% for TG dataset prediction enhancements and for the maximum, up to 7.9% for the EDD dataset and 10.2% for the TG dataset prediction enhancements are achieved. These results highlights the effectiveness of our proposed feature extraction methods with respect to the number of extracted features. Note that the performance of Naive Bayes is not improved due to the correlation of the extracted features and therefore is not explored in this part.

Next, we have generate eight different feature sets consisting of combination of features extracted from different attributes using our proposed feature extraction methods in the following two steps. We first study the performance of a given classifier, based on the employed feature extraction method (explored

Table 2. Comparison of the achieved results (%) using Adaboos.M1, Random Forest, and SVM to evaluate the enhancement achieved considering the segmentation-based distribution approach

	EDD		EDD	
AdaBoost.M1	From 25% to 5%	From 25% to 10%	From 25% to 5%	From 25% to 10%
Average	8.3	5.7	6.6	4.3
Maximum	11.3	10.2	10.3	7.9
Random Forest	From 25% to 5%	From 25% to 10%	From 25% to 5%	From 25% to 10%
Average	7.8	5.6	6.9	4.8
Maximum	11.5	9.3	12.2	7.3
SVM	From 25% to 5%	From 25% to 10%	From 25% to 5%	From 25% to 10%
Average	3.8	2.1	3.6	2.1
Maximum	7.1	6.5	6.9	7.1

on the TG dataset). And then, based on each classifier, two feature sets are constructed in the way that each feature set consists of features extracted using similar feature extraction method with the best performances (totally eight combinations). These feature sets have been constructed in the manner to maintain the number of employed features small. In the following paragraph, attributes as well as feature extraction method used to build each of our eight feature sets are explained. For simplicity, we refer to each attribute by its number as in Table 1.

The first and second combinations are extracted respectively based on the performance of the Adaboost.M1 classifier on the segmented-based distribution (with *K=10%*) (attribute numbers: 3, 4, 5, 14, 17, 26, 28, 30, 33, 41, 48 = 121 features) and the segmented-based density (with *K=5%*) feature extraction methods (attributes numbers: 1, 3, 4, 20, 54, 55 = 126 features). The third and forth are extracted based on the performances of the Random Forest classifier on the segmented-based density (with *K=5%*) (1, 3, 16, 17, 41, 55 = 126 features) and the segmented-based distribution (with *K=10%*) (3, 4, 5, 14, 16, 17, 26, 28, 30, 41, 44, 48 = 132 features) feature extraction approaches. The fifth and sixth combinations are extracted based on the performances of the SVM classifier on the segmented-based distribution (with *K=25%*) (1, 3, 4, 5, 17, 27, 29, 30, 31, 33, 35, 37, 38, 39, 40, 41, 44, 47, 48, 55 = 100 features) and the segmented-based distribution (with *K=5%*) (3, 5, 15,17, 30, 41, 44 = 147 features) feature extraction methods. Finally, the seventh and eighth are extracted based on the performances of the Naive Bayes classifier on the segmented-based distribution (with *K=25%*) (1, 3, 4, 5, 14, 16, 17, 27, 29, 30, 31, 32, 33, 37, 38, 39, 40, 41, 44, 47, 48, 55 = 110 features) and the segmented-based density (with *K=5%*) (3, 16, 17, 24, 33, 42 = 126 features) feature extraction methods. It is important to highlight that most of the attributes used to construct these feature sets have not been used or adequately explored for PFR. However, these attributes individually outperform most of the popular attributes used to tackle this problem (e.g. average long range contact energy (16), total non bounded energy (22), and mean fractional area loss (51)).

In continuation, composition of the amino acid feature group (the percentage of occurrence of the amino acids along the protein sequence divided by the length of proteins) as well as the length of the amino acids feature (which attained

Table 3. Results achieved (in percentage %) by using AdaBoost.M1 (Ada), Random Forest (RF), Naive Bayes (NA), SVM, and EDC for 15 feature vectors extracted from the combination of features are extracted in this step (for both EDD and TG datasets).

Datasets	TG					EDD				
Comb_Numb	Na	SVM	Ada	RF	EDC	Na	SVM	Ada	RF	EDC
Comb_1	32.0	40.5	39.1	35.5	**42.7**	38.3	47.4	46.3	41.4	**50.0**
Comb_2	34.1	36.0	38.2	35.1	41.2	39.8	44.6	44.4	38.6	49.2
Comb_3	32.2	36.8	39.0	34.4	41.3	39.4	43.9	45.3	39.7	49.2
Comb_4	34.9	39.3	38.5	37.7	42.4	38.0	44.1	46.2	41.5	47.0
Comb_5	32.7	37.9	38.6	37.9	40.5	37.2	45.0	44.9	42.6	46.4
Comb_6	28.4	40.9	36.9	33.0	39.8	35.8	47.9	44.0	41.6	49.8
Comb_7	30.8	38.3	39.1	36.1	41.2	37.4	44.7	45.1	40.8	47.5
Comb_8	33.0	34.4	37.5	33.6	38.5	42.5	44.2	45.2	38.9	48.6

good results in previous studies [3]) are added (20 + 1 features in total) to each extracted combination of feature groups (which for the rest of this study will be referred as comb_1 to comb_8 respectively). We then apply the employed classifiers in this study to each combination. The results are shown in Table 3. We also apply the *Ensemble of Different Classifiers (EDC)* proposed in [2] which attained the best results for similar studies found in the literature to the extracted combination of features. To compare our results with previous studies, we reproduce the results achieves using EDC to the features extracted in [8] (219 features), extracted features in [5] (125 features), and the 69D feature vector (the 49D feature vector extracted in [7] in addition to the composition of the amino acid feature group (49 + 20 = 69 features)). By reproducing this results we respectively achieve to 48.8%, 47.6%, and 40.7% prediction accuracies for the EDD dataset and 41.1%, 40.7%, and 33.0% for the TG dataset.

As it is shown in Table 3, by using EDC to Comb_1 we achieve to 50.0% and 42.7% prediction accuracy, up to 1.2% and 1.6% better than the best results reported in the literature for similar studies. These results are emphasize on the effectiveness of using features extracted from a wide range of physicochemical-based attributes. To explore even a further range of physicochemical-based attributes with respect to our proposed feature extraction methods, we propose *Ensemble of Diversely Trained AdaBoost.M1 and SVM Classifiers (EDTAS)* in the following manner. We first train two AdaBoost.M1 classifiers diversely trained with Comb_1 and Comb_3 feature vectors and two SVM classifiers diversely trained with Comb_1 and Comb_6 feature vectors. Then for a given test sample, we produce the output of the system using EDC classifier which is trained with Comb_1 feature vector as a tie breaker for two different cases. In the first case, when a fold reached to majority of the votes, it will be directly chosen as the output which out consideration of the EDC classifier. While, in case that a fold would not reach to the majority (two fold with two votes or four different folds with one vote each), the output of EDC will be directly chosen as the output of the system. The architecture of the EDTAS is shown in Figure 2.

Using EDTAS, we achieve up to 50.9% and 43.5% prediction accuracies, up to 2.1% and 2.4% better than previously reported results for the similar studies for the EDD and TG datasets, respectively. The results achieved in this study com-

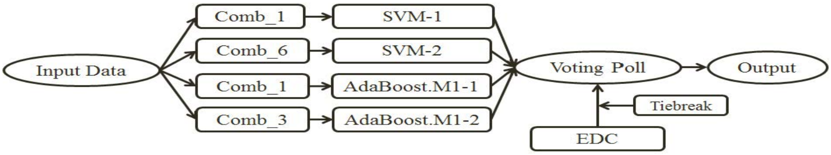

Fig. 1. The overall architecture of the EDTAS

pared to previous results found in the literature for similar studies are reported in Table 4. We also achieved up to 14.3% and 10.5% better prediction accuracy than using 69D feature vector which emphasize on the effectiveness of our proposed feature extraction methods to reveal more discriminatory information from a wide range of physicochemical-based attributes compared to previously used approaches found in the literature [7].

Table 4. The best results (in percentage %) achieved in this study compared to the best results found in the literature for the EDD and TG benchmarks respectively

Study	Attributes (Number of features)	Method	EDD (Results)	TG (Results)
[5]	Features proposed in [5] (126)	SVM	46.3	38.5
[3]	Features proposed in [5] (126)	Ada	44.7	36.4
[4]	Features proposed in [5] (126)	RF	42.9	37.1
[2]	Features proposed in [5] (126)	EDC	47.6	40.7
[8]	Features proposed in [8] (219)	SVM	47.3	40.1
[3]	Features proposed in [8] (219)	Ada	45.3	37.2
[4]	Features proposed in [8] (219)	RF	43.9	38.1
[2]	Features proposed in [8] (219)	EDC	48.8	41.1
[7]	69D (49+20)	SVM	36.6	33.0
This study	Comb_1 (202)	EDC	50.0	42.7
This study	Comb_2 (202)	EDC	49.2	41.3
This study	Comb_3 (202)	EDC	49.2	41.2
This study	Fused (202 for each classifier)	EDTAS	**50.9**	**43.5**

6 Conclusion

In this study we proposed two novel feature extraction methods namely segmented-based density and segmented-based distribution to reveal more local discriminatory information compared to similar approaches found in the literature. We also explored the effectiveness of 55 different physicochemical-based attributes that mostly have not been studied adequately for PFR. We evaluated our proposed approaches using five different classification techniques namely, Naive Bayes, Random Forest, AdaBoost.M1, SVM, and EDC. Then, we generate eight different combination of features extracted from a wide range of attributes based on the results of previous step. Finally, by proposing *Ensemble of Diversely Trained Adaboost.M1 and SVM (EDTAS)* we enhanced the protein fold prediction accuracy for more than 2% better than previously reported results for the similar studies found in the literature.

References

1. Kavousi, K., Moshiri, B., Sadeghi, M., Araabi, B.N., Moosavi-Movahedi, A.A.: A protein fold classifier formed by fusing different modes of pseudo amino acid composition via pssm. Computational Biology and Chemistry 35(1), 1–9 (2011)
2. Dehzangi, A., Phon Amnuaisuk, S., Ng, K.H., Mohandesi, E.: Protein Fold Prediction Problem Using Ensemble of Classifiers. In: Leung, C.S., Lee, M., Chan, J.H. (eds.) ICONIP 2009, Part II. LNCS, vol. 5864, pp. 503–511. Springer, Heidelberg (2009)
3. Krishnaraj, Y., Reddy, C.K.: Boosting methods for protein fold recognition: An empirical comparison. In: Proceedings of the 2008 IEEE International Conference on Bioinformatics and Biomedicine, pp. 393–396 (2008)
4. Dehzangi, A., Phon-Amnuaisuk, S., Dehzangi, O.: Using random forest for protein fold prediction problem: An empirical study. Journal of Information Science and Engineering 26(6), 1941–1956 (2010)
5. Ding, C., Dubchak, I.: Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics 17, 349–358 (2001)
6. Kurgan, L.A., Zhang, T., Zhang, H., Shen, S., Ruan, J.: Secondary structure-based assignment of the protein structural classes. Amino Acids 35, 551–564 (2008)
7. Gromiha, M.M., Oobatake, M., Sarai, A.: Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins. Biophysical Chemistry 82, 51–67 (1999)
8. Dehzangi, A., Phon-Amnuaisuk, S.: Fold prediction problem: The application of new physical and physicochemical- based features. Protein and Peptide Letters 18(2), 174–185 (2011)
9. Chen, K., Kurgan, L.A.: Pfres: protein fold classification by using evolutionary information and predicted secondary structure. Bioinformatics 23(21), 2843–2850 (2007)
10. Shen, H.B., Chou, K.C.: Predicting protein fold pattern with functional domain and sequential evolution information. Journal of Theoretical Biology 256(3), 441–446 (2009)
11. Yang, J.Y., Chen, X.: Improving taxonomy-based protein fold recognition by using global and local features. Protein 79(7), 2053–2064 (2011)
12. Shen, H.B., Chou, K.C.: Ensemble classifier for protein fold pattern recognition. Bioinformatics 22, 1717–1722 (2006)
13. Taguchi, Y.H., Gromiha, M.M.: Application of amino acid occurrence for discriminating different folding types of globular proteins. BMC Bioinformatics 8(1), 404 (2007)
14. Mathura, V.S., Kolippakkam, D.: Apdbase: Amino acid physico-chemical properties database. Bioinformation 12(1), 2–4 (2005)
15. Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. In: International Conference on Machine Learning, pp. 148–156 (1996)
16. Witten, I., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)
17. Breiman, L.: Random forests. Machine Learning 45, 5–32 (2001)