Protein Fold Recognition Using an Overlapping Segmentation Approach and a Mixture of Feature Extraction Models

Abdollah Dehzangi[1,2], Kuldip Paliwal[1], Alok Sharma[1,3], James Lyons[1], Abdul Sattar[1,2]
{a.dehzangi, k.paliwal, a.sattar}@griffith.edu.au,
sharma_al@usp.ac.fj, and james.lyons@griffithuni.edu.au

[1] Institute for Integrated and Intelligent Systems (IIIS), Griffith University, Brisbane, Australia
[2] National ICT Australia (NICTA), Brisbane, Australia
[3] University of the South Pacific, Fiji

Abstract. Protein Fold Recognition (PFR) is considered as a critical step towards the protein structure prediction problem. PFR has also a profound impact on protein function determination and drug design. Despite all the enhancements achieved by using pattern recognition-based approaches in the protein fold recognition, it still remains unsolved and its prediction accuracy remains limited. In this study, we propose a new model based on the concept of mixture of physicochemical and evolutionary features. We then design and develop two novel overlapping segmented-based feature extraction methods. Our proposed methods capture more local and global discriminatory information than previously proposed approaches for this task. We investigate the impact of our novel approaches using the most promising attributes selected from a wide range of physicochemical-based attributes (117 attributes) which is also explored experimentally in this study. By using Support Vector Machine (SVM) our experimental results demonstrate a significant improvement (up to 5.7%) in the protein fold prediction accuracy compared to previously reported results found in the literature.

Keywords: Mixture of Feature Extraction Model, Overlapping Segmented distribution, Overlapping Segmented Auto Covariance, Support Vector Machine

1 Introduction

Prediction of the three dimensional structure (tertiary structure) of a protein from its amino acid sequence (primary structure) still remains as an unsolved issue for bioinformatics and biological science. *Protein Fold Recognition (PFR)* is considered as an important step towards protein structure prediction problem. PFR is defined as classifying a given protein to its appropriate fold (among finite number of folds). It also provides critical information about the functionality of proteins and how they are evolutionarily related to each other. Recent

advancement in the pattern recognition field stimulates enormous interest in this problem.

During the last two decades, a wide range of classifiers such as, Bayesian-based learners [1], *Artificial Neural Network (ANN)* [2], *Hidden Markov Model (HMM)* [3], Meta classifiers [4, 5], *Support Vector Machine (SVM)* [6–8] and ensemble methods [1, 9, 10] have been implemented and applied to this problem. Despite the crucial impact of the classification techniques used in solving this problem, the most important enhancements achieved were due to the attributes being selected and feature extraction methods being used [2, 6, 11–15]. Generally, features have been extracted to attack this problem can be categorized into three groups namely, *sequential* (extracted from the alphabetic sequence of the proteins (e.g. composition of the amino acids)), *physicochemical* (extracted based on different physical, chemical, and structural attributes of the amino acids and proteins (e.g. hydrophobicity)), and *evolutionary* (extracted from the scoring matrices generated based on evolutionary information (e.g. *Position Specific Scoring Matrix (PSSM)* [16])) feature groups.

The study of [8] and followup works explored the impact of physicochemical-based features in conjunction with sequential-based features for the PFR and attained promising results [17]. The main advantage of using physicochemical-based features is that these features do not rely on sequential similarities. Hence, they maintain their discriminatory information even when the sequential similarity rate is low. Furthermore, they are able to provide important information about the impact of physicochemical-based attributes on the folding process. However, they are unable to provide sufficient information to solve this problem individually. On the other hand, sequential-based features have the merit that they are able to provide critical information about the interaction of the amino acids in proteins based on the sequence similarity. However, they fail to maintain this information when the sequential similarity rate is low. Thus, relying solely on these two categories of features did not lead to better results.

More recent studies shifted the focus to evolutionary-based features which have significantly enhanced the performance of the PFR [6, 12]. Relying on the PSSM, evolutionary-based features are able to provide important information about the dynamic substitution score of the amino acids with each other. However, similar to the sequential-based features, they do not provide any information about the impact of different physicochemical-based attributes on the folding process. Furthermore, they lose their discriminatory information dramatically when the sequential similarity rate is low.

In this study, we propose a novel approach to enhance the protein fold prediction accuracy and at the same time to provide more information about the impact of the physicochemical-based attributes on the folding process. In our proposed approach, first we transform the protein sequence using evolutionary-based information. Then, physicochemical-based features are extracted from the transformed sequence of the proteins using segmentation, density, distribution, and autocorrelation-based methods in an overlapping style. We explore our proposed feature extraction methods for 15 most promising attributes which are

selected from 117 experimentally explored physicochemical-based attributes. Finally, by applying SVM on the combinations of the extracted features, we enhance the protein fold prediction accuracy for 5.7% over previously reported results found in the literature.

2 Benchmarks

To evaluate the performance of our proposed method against previous studies found in the literature, the EDD (extended version of DD data set introduced by Ding and Dubchak [8]) and the TG (introduced by Taguchi and Gromiha [18]) benchmarks are used. In earlier studies, DD was considered as the most popular benchmark for the PFR. However, it is no longer used [12, 13] due to its inconsistency with the latest version of *Structural Classification of Proteins (SCOP)* [19]. Extracted from the latest version of the SCOP, the EDD has been widely used as a replacement for the original DD [6, 3, 11, 12]. In this study, we extract the EDD benchmark from the SCOP 1.75 consisting of 3418 proteins belonging to the 27 folds that was originally used in the DD with less than 40% sequential similarities. We also use the TG benchmark [18] consisting of 1612 proteins belonging to 30 folds with less than 25% sequential similarities.

3 Physicochemical-based Attributes

In this study, we investigate the impact of our proposed approaches using 15 physicochemical-based attributes. These 15 attributes have been selected from 117 physicochemical-based attributes (which are taken from the AAindex [20], the APDbase [21], and previous studies found in the literature [22]) in the following manner. For a given attribute, we extracted six feature groups based on the overlapped segmented distribution and overlapped segmented autocorrelation approaches which are the subjects of this study. Then we applied five classifiers namely, Adaboost.M1, Random Forest, Naive Bayes, *K-Nearest Neighbor (KNN)*, and SVM to each feature group separately. Therefore, 30 prediction accuracies were achieved for each physicochemical-based attribute for each benchmark (five classifiers applied to six feature groups separately ($5 \times 6 = 30$)). Considering this experiment for EDD and TG benchmarks, 60 prediction accuracies ($2 \times 30 = 60$) are achieved for each individual attribute ($60 \times 117 = 7020$ prediction accuracies in total for all 117 attributes) [4]. Then we compared these results for all 117 attributes and selected 15 attributes that attained the best results in average for all 60 prediction accuracies [5]. The feature selection process was conducted manually. This process was also explored in our previous studies for the PFR and protein structural class prediction problem [15, 23]. The

[4] The experimental results achieved in this step for all five classifiers for EDD and TG benchmarks are available upon request.

[5] Details about the attribute selection process as well as the list and references of all 117 physicochemical-based attributes are available upon request.

selected attributes are: (1) structure derived hydrophobicity value, (2) polarity , (3) average long range contact energy, (4) average medium range contact energy, (5) mean *Root Mean Square (RMS)* fluctuational displacement, (6) total non-bounded contact energy, (7) amino acids partition energy, (8) normalized frequency of alpha-helix, (9) normalized frequency of turns, (10) hydrophobicity scale derived from 3D data, (11) *High Performance Liquid Chromatography (HPLC)* retention coefficient to predict hydrophobicity and antigenicity, (12) average gain ratio of surrounding hydrophobicity, (13) mean fractional area loss, (14) flexibility, and (15) bulkiness. Note that to the best of our knowledge, most of the selected attributes (attributes number 3, 4, 5, 6, 7, 10, 11, 12, 13, and 14) have not been adequately (or not at all) explored for the PFR. However, in our conducted comprehensive experimental study, they have outperformed many popular attributes that have been widely used for PFR [2, 8, 9, 13].

4 Feature Extraction Method

In the continuation, we first use PSIBLAST for the EDD and TG benchmarks (using NCBI's non redundant (NR) database with three iterations and cut off E-value of 0.001) and extract the PSSM [12]. The PSSM consists of two $L \times 20$ matrices (where L is the length of a protein sequence) namely, PSSM_cons and PSSM_prob. PSSM_cons contains the log-odds while PSSM_prob contains the normalized probability of the substitution score of an amino acid with other amino acids depending on their positions along a protein sequence. Then four main sets of features are extracted (two sets from the transformed protein sequences and two sets directly from the PSSM). In continuation, each feature extraction approach will be explained in detail (overlapped segmented distribution, overlapped segmented autocorrelation, semi-composition, and evolutionary-based auto-covariance).

4.1 Physicochemical-based Feature Extraction:

In this study, a new mixture of physicochemical and evolutionary-based feature extraction method is proposed based on the concepts of overlapped segmented distribution and autocorrelation. The main idea of our proposed method is to extract physicochemical-based features from the transformed sequences (so called consensus sequence) using evolutionary-based information to get benefit of discriminatory information embedded in both of these groups of features, simultaneously. In our proposed method, we first extract the consensus sequence and then, two feature groups namely overlapped segmented distribution and overlapped segmented autocorrelation are extracted from it.

Consensus Sequence Extraction Procedure: An amino acid sequence when transformed using evolutionary-based information embedded in the PSSM is called a consensus sequence [12]. Previously, to extract this sequence the PSSM_cons

have been popularly used [12]. In this method, each amino acid based on its position along the protein sequence $(O_1, O_2, ..., O_L)$ is replaced with the amino acid that has the highest (maximum) substitution score according to the PSSM_cons $(C_1, C_2, ..., C_L)$. Consensus sequence was also effectively used to extract sequential-based features and attained promising results for the PFR [12]. However, it fails to address an important issue. For the case of unknown proteins the PSSM_cons does not provide any information and simply returns all equal substitution scores with the other amino acids (equal to -1).

To address this limitation, we use a modified method which relies on the PSSM_prob for feature extraction. In the PSSM_prob if a sequence similarity is found in NR, it returns a substitution probability score for even unknown amino acids. Using PSSM_prob dramatically reduces the number of unknown amino acids in the consensus sequence while previous approaches. Using PSSM_prob, we have successfully replaced over 360 unknown amino acids (out of 362 unknown amino acids) for the EDD benchmark while for the TG benchmark, we have successfully replaced all of the unknown amino acids.

Overlapped Segmented Distribution (OSD): Global density of an specific attribute is considered as a popular feature for the PFR. However, it does not properly explore the local discriminatory information available in the sequence [22]. To address this issue, the distribution of different segments of a specific density is extracted and added to this feature. In this method, we first replace the amino acids in the consensus sequence $(C_1, C_2, ..., C_L)$ with the values assigned to each of them based on a given physicochemical-based (e.g. hydrophobicity) attribute $(S_1, S_2, ..., S_L)$. Then we calculate the global density $T_{gd} = \frac{\sum_{i=1}^{L} S_i}{L}$. Next, beginning from each side of the sequence, the given attribute summation again is calculated until reaching to the first $K_s\%$ of the T_{gd} as follows:

$$I_k = (T_{gd} \times L \times K_s)/100. \quad (1)$$

Finally, the number of summed amino acids divided by the length of the protein is returned as the distribution of the first $K_s\%$ of global density. For example, if the summation of the hydrophobicity of m amino acids is equal to I_k, then the output for $K_s\%$ distribution factor is m/L. In this study, K is set to 5 based on the experimental study conducted by the authors due to similar performance of using $K_s = 5$ compared to the larger distribution factors (10 or 25) and trade of between the number extracted features and achieved prediction accuracy. This process is repeated until reaching to $K_s = 75$ (5%, 10%, 15%, ..., 75%) of the global density from each side (Figure 1).

The distribution index is calculated from both sides of the proteins due to the fact that there is no rear or front for proteins. Furthermore, an approach of using one side calculation produces accumulative distribution in the other side. We also use the overlapping approach to explore distribution of the amino acids better with consideration of an specific attribute. 75% overlapping factor is selected experimentally based on the trade off between the number of features added and the discriminatory information provided. Therefore, using $K_s = 5$ distribution

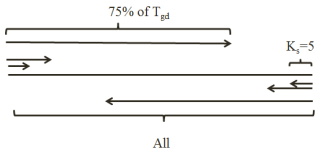

Fig. 1. Segmented distribution-based feature extraction method.

and 75% overlapping factors (in addition to the global density feature in each group), 31 features are extracted ($75/5 = 15$ features from each side).

Overlapped Segmented Autocorrelation (OSA): Similar to the density, autocorrelation-based features have been widely used for the PFR and attained promising results [17, 2]. However, even the most sophisticated approaches failed to provide adequate local discriminatory information (e.g. pseudo amino acid composition [9]). Therefore, segmented-base approach is used in this study. In the proposed approach, we segment the protein sequence using a segmented distribution approach explained in previous subsection and then calculate the autocorrelation in each segment accumulatively (in this case, K_s is set to 10, distance factor (F) is set to 10 and overlapping factor is set to 70%). The autocorrelation in each segment is equal to:

$$Seg\text{-}Auto_{i,a} = \frac{1}{(L(a/100) - i)} \sum_{j=m}^{n} S_j S_{j+i}\ , \quad (i = 1, ..., F\ \&\ a = 10, .., 70), \tag{2}$$

where L is the length of sequence, a is the segmentation factor, m and n are respectively the begin and the end of a segment, and S_j is the value of an attribute (normalized) for each amino acid. We also add the global autocorrelation (where F set to 10) which is calculated as follows:

$$Global\text{-}Auto_{i,a} = \frac{1}{L} \sum_{j=1}^{L-i} S_j S_{j+i} \quad (i = 1, ..., F). \tag{3}$$

Therefore, based on each attribute the autocorrelation of the 10%, 20%, 30%, ... , 70% from each end (14 segments in total) are accumulatively calculated (*Seg-Auto + Global-Auto = OSA*). *F=10* is adopted in this study because it was showed in [6] as the most effective distance factor for the PFR. The overlapping and the segmentation factors are also adjusted based on the experimental study conducted by the authors. In results, a feature group based on this approach is extracted consisting of 150 features *(70 + 70 + 10)*.

4.2 Evolutionary-based Feature Extraction

We also extract two sequenced-based feature groups namely, semi-composition and evolutionary-based auto-covariance directly from the PSSM.

Semi-composition (Semi-AAC): This feature group is extracted to provide more information about the occurrence of each amino acid along a protein sequence. However, instead of being extracted from the original protein sequence, we directly extract that from the PSSM. In this feature group, the composition of each amino acid is equal to the summation of its substitution scores divided by the length of the protein which is calculated as follows:

$$Semi\text{-}AAC_i = \frac{1}{L}\sum_{i=1}^{L} P_{ij}, (j = 1, ..., 20), \tag{4}$$

where P_{ij} is the substitution score for the amino acid at position i with the *j-th* amino acid in the PSSM. It was shown in [24] that Semi-AAC is able to provide more discriminatory information compared to the conventional composition feature group.

Evolutionary-based auto covariance (PSSM-AC): This feature group provides crucial information about the local interaction of the amino acids from the PSSM and attained promising results for the PFR [6, 24]. In the PSSM-AC the auto covariance of the substitution score of each amino acid with another amino acids with the distance factor of 10 (the distance factor is set to 10 as the most effective value as the distance factor investigated in [6]) is calculated (from the PSSM_cons). The PSSM-AC can be calculated as follows:

$$PSSM\text{-}AC_{j,f} = \frac{1}{(L-f)}\sum_{i=1}^{L-f}(P_{i,j} - P_{ave,j})(P_{i+f,j} - P_{ave,j}), \quad (j = 1, ...20 \ \ \& \ \ f = 1, ..., 10), \tag{5}$$

where $P_{ave,j}$ is the average of substitution score for the *j-th* column of PSSM. Therefore, $20 \times F$ features calculated in this feature group ($20 \times 10 = 200$).

5 Support Vector Machine (SVM)

SVM introduced by [25] aims at finding the *Maximal Marginal Hyperplane (MMH)* based on the concept of the support vector theory to minimize the error. The classification of some known points in input space $\mathbf{x}_i$ is y_i which is defined to be either -1 or +1. If x' is a point in input space with unknown classification then:

$$y' = sign\left(\sum_{i=1}^{n} a_i y_i K(\mathbf{x}_i, \mathbf{x}') + b\right), \tag{6}$$

where y' is the predicted class of point $\mathbf{x}'$. The function *K()* is the kernel function; n is the number of support vectors and a_i are adjustable weights and b is the bias. This classifier is considered as the state-of-the-art classification techniques in the pattern recognition and attained the best results for the PFR [6, 12, 11]. Therefore, we will only use SVM to investigate the effectiveness of our proposed methods here rather than the five classifiers that used earlier in Section 3 for the

feature selection process. In this study, three different SVM-based classifiers are used to reproduce previous results as well as evaluating our proposed approaches. We use the SVM classifier implemented in the SVMLIB toolbox using *Radial Base Function (RBF)* as its kernel [26] (using grid algorithm implemented in SVMLIb to optimize its parameters (width (γ) and regularization (C) parameters)). We also use SVM using *Sequential Minimal Optimization (SMO)* as a *polynomial kernel* which its polynomial degree is set to one (which is called linear kernel) and three (implemented in WEKA with using its default parameters [27]).

6 Results and Discussion

In the first step, the performance of the modified consensus sequence extraction method is explored by extracting occurrence (occurrence of each amino acid in a protein sequence (20 features)) and composition (percentage of the occurrence of each amino acid along a protein sequence (20 features)) feature groups. We extract these feature groups from the original sequence, the consensus sequence extracted using conventional approach and the modified consensus sequence extraction method used in this study, and applied SVM (with linear kernel). In this study, 10-fold cross validation is used as the evaluation method as it has been mainly used for this purpose in the literature [6, 8].

Table 1. Comparison of the achieved results (%) using SVM (linear kernel) to evaluate the proposed consensus sequence extraction method compared to use of original sequence as well as previously used methods for the EDD and the TG benchmarks.

Methods	Composition		Occurrence	
	EDD	TG	EDD	TG
Original Sequence	32.4	31.6	41.2	33.6
Current consensus sequence extraction method	42.2	34.7	48.2	38.6
Proposed Method in this study	**44.4**	**36.3**	**48.9**	**38.8**

As shown in Table 1, the modified consensus sequence extraction method used in this study enhances the PFR performance considering composition and occurrence of the amino acids feature groups. Next, we extract features introduced in the previous stage and combine them to build the input feature vector to feed the employed SVM classifier. The input feature vector is built by combining Semi-AAC (20 features), segmented distribution (31 features), segmented autocorrelation (150 features), and PSSM-AC (200 features) feature groups in addition to the length of protein sequence feature (as used in [2, 1]). Therefore, for each attribute, a feature vector consists of 402 features is created and named *Comb_ph1* to *Comb_ph15*. The overall architecture of our proposed method is shown in Figure 2.

We then apply the SVM classifier to our extracted features. We also duplicate the study of Dong and his co-workers [6] which to the best of our knowledge attained the best results to tackle PFR. Furthermore, the 49D feature group

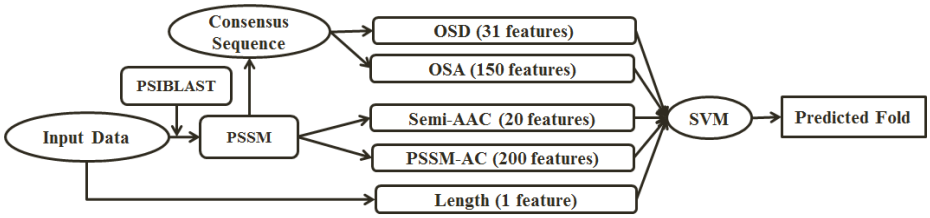

Fig. 2. The overall architecture of the proposed approach. The number of features extracted in each feature group is shown in the brackets.

extracted by [22] is also extracted from both of the employed benchmarks and is added to the extracted sequential-based features (Semi-AAC + PSSM-AC + length (221 features)). This feature vector consists of global density of 49 different physicochemical-based attributes (49 dimensional feature group) that has been extracted to provide sufficient physicochemical-based information for the PFR [18]. In this part, we aim at comparing the impact of using a wide range of features with exploring the impact of a single attribute considering our proposed feature extraction method. Note that using SVM classifier with linear kernel attains similar results to the other two version of SVM classifier investigated in this study (SVM classifier using SMO kernel function with $p = 3$, and SVM classifier using RBF kernel function) which emphasizes on the effectiveness of the employed features rather than the kernel function used for SVM. The best results for the EDD and the TG benchmarks compared to the state-of-the-art results found in the literature are shown in Table 2.

As it is shown in Table 2, we achieve up to 82.9% and 64.6% prediction accuracies for the EDD and the TG benchmarks which are 4.8% and 5.7% better than the best results reported in the literature for the employed benchmarks respectively. Considering the small enhancement achieved in previous studies (using DD benchmark), having over 4% enhancement is considered as a significant number for the PFR [1, 4, 14]. We also achieve to over 82% and 63% prediction accuracies respectively for the EDD and the TG benchmarks using extracted features from the attributes that have not been adequately explored (attribute 14), or (to the best of our knowledge) have not been explored at all for the PFR (attributes number 1, 5, 7, and 13). Also, the significant enhancement achieved for all of the explored attributes (over 80% and 61% prediction accuracies respectively for the EDD and the TG benchmarks) emphasizes on the importance of the proposed feature extraction methods in this study. We also achieve to 41.0% and 22.7% better prediction accuracies for the EDD and the TG benchmarks respectively compared the best results achieved without using evolutionary information for feature extraction (relying solely on the original protein sequence to extract physicochemical-based features [13]). It also emphasizes on the impact of our mixture of physicochemical-based and evolutionary-based feature extraction method to enhance the protein fold prediction accuracy.

Table 2. The best results (in percentage) achieved in this study compared to the best results found in the literature for the EDD and the TG benchmarks.

Study	Attributes (No. of features)	Method	EDD	TG
[18]	AAO original sequence (20)	LDA	46.9	36.3
[18]	AAC original sequence (20)	LDA	40.9	32.0
[13]	Physicochemical(125)	Adaboost.M1	47.2	39.1
[8]	Physicochemical(125)	SVM	50.1	39.5
[13]	Physicochemical(220)	SVM(SMO)	52.8	41.9
[22]	Threading	Naive Bayes	70.3	55.3
[2]	Bi-gram (400)	SVM	75.2	52.7
[2]	Tri-gram (8000)	SVM	71.0	49.4
[11]	Combination of bi-gram features (2400)	SVM	69.9	55.0
[3]	PSIPRED and PSSM-based features (242)	SVM	77.5	57.1
[6]	ACCFold-AAC(200)	SVM(RBF)	76.2	56.4
[6]	ACCFold-AC(4000)	SVM(RBF)	**78.1**	**58.9**
This study	Comb_ph1 (402)	SVM(SMO)	82.3	63.3
This study	Comb_ph5 (402)	SVM(SMO)	82.8	**64.6**
This study	Comb_ph7 (402)	SVM(SMO)	**82.9**	64.0
This study	Comb_ph13 (402)	SVM(SMO)	82.5	63.7
This study	Comb_ph14 (402)	SVM(SMO)	82.4	63.8
This study	Original sequence (49+221)	SVM(SMO)	44.7	35.7
This study	Consensus sequence (49+221)	SVM(SMO)	59.7	45.9

We also achieve up to 23.2% and 18.7% better prediction performance for the EDD and TG benchmarks respectively compared to use of 49D (which is extracted from the consensus sequence and combined with the Semi-AAC, PSSM-AC and the length of the amino acid sequence (221 features)). In other word, by extracting features based on a single attribute using our proposed feature extraction method, we significantly enhance the PFR performance compared to use of a wide range of physicochemical-based attributes using global density as the main feature. These results emphasize on the effectiveness of the overlapped segmented-based feature extraction method to explore more discriminatory information. It is important to highlight that these results are achieved using 402 attributes, which is 10 times less than the number of attributes that was used in the ACCFold-AC model (4000 features). Besides enhancing the protein fold prediction accuracy, by proposing a mixture of physicochemical and evolutionary-based information, we introduce a new direction to obtain benefit from discriminatory power of these two groups of features simultaneously. Furthermore, by exploring physicochemical based features, the proposed method is able to provide crucial information about the impact of these attributes on the PFR. Note that our proposed features in this study (overlapped segmented-based distribution and overlapped segmented-based autocorrelation) have been investigated for the protein structural class prediction problem (in a different experiment) and obtained promising results as well [23] which highlights the generality of these approaches for similar studies.

7 Conclusion

In this study, we proposed a model to enhance the protein fold prediction accuracy as well as providing better understanding about the impact of the

physicochemical-based attributes on the PFR in the following five steps. In the first step, a modified consensus sequence extraction method was proposed. It addressed the issue of unknown proteins using evolutionary-based information. Proposed method also improved the protein fold prediction accuracy over the previous methods that extracted consensus sequence. In the second step, a comprehensive study on a wide range of physicochemical-based attributes (117 attributes) were conducted and 15 most promising attributes were selected. The selected attributes outperformed other attributes based on the density, distribution, and autocorrelation feature extraction methods. This comprehensive experimental study provided important information about the performance of these 117 physicochemical-based attributes on the PFR. In the third step, we proposed two novel feature extraction methods based on the concepts of segmented distribution and autocorrelation to provide more local and global discriminatory information for the PFR. In the next step, effective sequentially-based features that were directly extracted from the PSSM were combined with the proposed physicochemical-based features. In the final step, by using the SVM classifier (with linear kernel) to our extracted features, we achieved 82.9% and 64.6% prediction accuracies for the EDD and the TG benchmarks respectively which are 4.8% and 5.7% over the previously reported results found in the literature.

References

1. Dehzangi, A., Phon-Amnuaisuk, S., Dehzangi, O.: Enhancing protein fold prediction accuracy by using ensemble of different classifiers. Australian Journal of Intelligent Information Processing Systems **26**(4) (2010) 32–40
2. Ghanty, P., Pal, N.R.: Prediction of protein folds: Extraction of new features, dimensionality reduction, and fusion of heterogeneous classifiers. NanoBioscience, IEEE Transactions on **8**(1) (2009) 100–110
3. Deschavanne, P., Tuffery, P.: Enhanced protein fold recognition using a structural alphabet. Proteins: Structure, Function, and Bioinformatics **76**(1) (2009) 129–137
4. Dehzangi, A., Phon-Amnuaisuk, S., Manafi, M., Safa, S.: Using rotation forest for protein fold prediction problem: An empirical study. In: Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics. (2010) 217–227
5. Dehzangi, A., Karamizadeh, S.: Solving protein fold prediction problem using fusion of heterogeneous classifiers. INFORMATION, An International Interdisciplinary Journal **14**(11) (2011) 3611–3622
6. Dong, Q., Zhou, S., Guan, G.: A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation. Bioinformatics **25**(20) (2009) 2655–2662
7. Chmielnicki, W., Stapor, K.: A hybrid discriminative-generative approach to protein fold recognition. Neurocomputing **75**(1) (2012) 194–198
8. Ding, C., Dubchak, I.: Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics **17** (2001) 349–358
9. Yang, T., Kecman, V., Cao, L., Zhang, C., Huang, J.Z.: Margin-based ensemble classifier for protein fold recognition. Expert Systems with Applications **38** (2011) 12348–12355

10. Kavousi, K., Sadeghi, M., Moshiri, B., Araabi, B.N., Moosavi-Movahedi, A.A.: Evidence theoretic protein fold classification based on the concept of hyperfold. Mathematical Biosciences **240**(2) (2012) 148–160
11. Shamim, M.T.A., Anwaruddin, M., Nagarajaram, H.A.: Support vector machine-based classification of protein folds using the structural properties of amino acid residues and amino acid residue pairs. Bioinformatics **23**(24) (2007) 3320–3327
12. Yang, J.Y., Chen, X.: Improving taxonomy-based protein fold recognition by using global and local features. Proteins: Structure, Function, and Bioinformatics **79**(7) (2011) 2053–2064
13. Dehzangi, A., Phon-Amnuaisuk, S.: Fold prediction problem: The application of new physical and physicochemical- based features. Protein and Peptide Letters **18**(2) (2011) 174–185
14. Sharma, A., Lyons, J., Dehzangi, A., Paliwal, K.K.: A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition. Journal of Theoretical Biology **320**(0) (2013) 41–46
15. Dehzangi, A., Sattar, A.: Protein fold recognition using segmentation-based feature extraction model. In: Proceedings of the 5th Asian Conference on Intelligent Information and Database Systems. ACIIDS, Springer-Verlag (2013) 345–354
16. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J.H., Zhang, Z., Miller, W., Lipman, D.J.: Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Research **17** (1997) 3389–3402
17. Shen, H.B., Chou, K.C.: Ensemble classifier for protein fold pattern recognition. Bioinformatics **22** (2006) 1717–1722
18. Taguchi, Y.H., Gromiha, M.M.: Application of amino acid occurrence for discriminating different folding types of globular proteins. BMC Bioinformatics **8**(1) (2007) 404
19. Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C.: Scop: A structural classification of proteins database for the investigation of sequences and structures. Journal of Molecular Biology **247**(4) (1995) 536–540
20. Kawashima, S., Pokarowska, P.P.M., Kolinski, A., Katayama, T., Kanehisa, M.: Aaindex: Amino acid index database, progress report. Neucleic Acids **36** (2008) D202–D205
21. Mathura, V.S., Kolippakkam, D.: Apdbase: Amino acid physico-chemical properties database. Bioinformation **12**(1) (2005) 2–4
22. Gromiha, M.M.: A statistical model for predicting protein folding rates from amino acid sequence with structural class information. Journal of Chemical Information and Modeling **45**(2) (2005) 494–501
23. Dehzangi, A., Paliwal, K.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 Transaction on Computational Biology and Bioinformatics (TCBB) **In Press** (2013)
24. 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 **42** (2012) 2243–2249
25. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer-Verlag (1999)
26. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2** (2011) 27:1–27:27
27. Witten, I., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. 2 edn. Morgan Kaufmann, San Francisco (2005)