Enhancing Protein Fold Prediction Accuracy Using Evolutionary and Structural Features

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

[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 an important step towards the protein structure prediction problem. It also provides crucial information about the functionality of the proteins. Despite all the efforts that have been made during the past two decades, finding an accurate and fast computational approach to solve PFR still remains a challenging problem for bioinformatics and computational biology. It has been shown that extracting features which contain significant local and global discriminatory information plays a key role in addressing this problem. In this study, we propose the concept of segmented-based feature extraction technique to provide local evolutionary information embedded in Position Specific Scoring Matrix (PSSM) and structural information embedded in the predicted secondary structure of proteins using SPINE-X. We also employ the concept of occurrence feature to extract global discriminatory information from PSSM and SPINE-X. By applying a Support Vector Machine (SVM) to our extracted features, we enhance the protein fold prediction accuracy to 7.4% over the best results reported in the literature.

Keywords: Protein Fold Recognition, Feature Extraction, Segmented distribution, Segmented Auto Covariance, Occurrence, Support Vector Machine (SVM)

1 Introduction

Protein Fold Recognition (PFR) is defined as assigning a given protein to a fold (among a finite number of folds) that represents its functionality as well as its major tertiary structure. Therefore, PFR is considered as an important step towards the protein structure prediction problem. Despite all the efforts that have been made so far to find an effective computational approach to solve this problem, it still remains an unsolved problem for computational biology. From the pattern recognition perspective, PFR is defined as solving a multi-class classification task. Therefore, extracting features that capture significant

global and local discriminatory information as well as the classification technique being used play the main roles in solving this problem. During the past two decades, a wide range of classification techniques have been used for PFR [1–9]. Among the classifiers employed to tackle this problem, *Support Vector Machine (SVM)* based classifiers have attained the best results [10, 11]. However, the most significant enhancement of PFR accuracy has been achieved by relying on the feature extraction approaches rather than the classification techniques being used [1, 9, 10, 12–14]. In most of the studies that addressed PFR by feature extraction techniques, global discriminatory information has been represented using the composition of the amino acids feature group (the occurrence of the amino acids along the protein sequence divided by the length of protein sequence [1, 8]). However, it has been shown that this feature group is not able to adequately reveal global information [15]. Furthermore, composition feature group is not able to capture information regarding the length of the protein sequence that was shown as an effective feature for PFR [13].

Compared to the methods adopted to extract global discriminatory information, a wider range of methods were used to extract local discriminatory information for PFR such as, pseudo amino acid composition [3, 8, 9], cross covariance [10], auto covariance [10], bi-gram [11, 14], and tri-gram [16]. Despite the significant local discriminatory information provided using these approaches, most of these methods produce large number of features which makes them computationally expensive for large protein data banks (e.g. cross covariance and tri-gram [10, 16]). At the same time, in all these methods the whole protein sequence as a single entity have been used to extract local information. In another words, they aimed to extract local information by exploring whole protein sequence as a global entity. Therefore, they could not appropriately explore local information embedded in protein sequence. Furthermore, despite all the efforts have been made to enhance the protein fold prediction accuracy so far, its prediction accuracy remains limited especially when the sequential similarity rate is low.

In this study, we aim at enhancing protein fold prediction accuracy by addressing these limitations. We propose segmented-base feature extraction to extract local evolutionary information embedded in *Position Specific Scoring Matrix (PSSM)* as well as structural information embedded in the predicted secondary structure using SPINE-X. We also employ the concept of an occurrence feature of the transformed protein sequence using evolutionary and structural information embedded in PSSM and SPINE-X to extract adequate global discriminatory information for PFR. By applying SVM to our extracted features we enhance the protein fold prediction accuracy to 7.4% better than the highest reported results found in the literature.

2 Data sets

In this study, two data sets namely TG and EDD are used to investigate the performance of our proposed methods. The TG data set introduced by [15] consists of 1612 proteins belonging to 30 folds with less than 25% sequential

similarities. TG is extracted from *Structural Classification of Proteins (SCOP)* 1.73 which has been previously used to investigate the performance of proposed methods for PFR when the sequential similarity is very low [13, 15, 17]. We also extract EDD (extended version of DD data set [1] which is extracted from SCOP 1.75). This data set consists of 3418 proteins belonging to 27 folds that was used originally in DD data set with less than 40% sequential similarities. The EDD data set extracted from an older version of SCOP has been widely used for PFR [5, 10, 11]. Using this data set enables us to directly compare our results with previously reported results found in the literature.

3 Feature Extraction Method

In this study, we rely on PSSM and the predicted secondary structure using SPINE-X to extract evolutionary and structural information respectively. PSSM is calculated by applying PSIBLAST [18] to EDD and TG data sets (using NCBI's non redundant (NR) database with its cut off value (E) set to 0.001). PSSM consists of an $L \times 20$ matrix (L is the length of a protein and the columns of the matrices represent 20 amino acids). It provides the substitution probability of a given amino acid based on its position along a protein sequence.

We also use predicted secondary structure using SPINE-X which was recently proposed by [19] and attained better results (especially for the coded area) than PSIPRED on predicting protein secondary structure [20]. Given a protein sequence, it returns an $L \times 3$ matrix (which will be referred to as SPINE-M for the rest of this study) consisting of the normalized probability of contribution of a given amino acid based on its position along the protein sequence to build one of the three secondary structure elements namely, α-helix, β-strands, and coils. It also returns a transformed version of the protein sequence (also extracted from SPINE-M) in which each amino acid along the protein sequence is replaced with H (represents helix), E (represents strand), or C (represents coil) based on its tendency to incorporate in building one of these secondary structure elements. In this study, we will refer to this sequence as the structural consensus sequence. It is expected that predicted secondary structure using SPINE-X provides significant structural information for PFR similar to or even better than PSIPRED due to its better performance [19]. In continuation, the global and local features extracted in this study will be explained in detail.

3.1 Global Features

To extract global discriminatory information embedded in PSSM and SPINE-M we mainly relied on the concept of the occurrence feature. We extract evolutionary and structural consensus sequence-based occurrence from the transformed protein sequence using PSSM and SPINE-M respectively. We also extract semi-occurrence feature group directly from PSSM and SPINE-M which represents the summation of the substitution probability of the amino acids and normalized probability of secondary structure elements respectively.

Consensus Sequence-based Occurrence: In this method, we extract occurrence of the amino acids as well as occurrence of the secondary structure elements derived from the evolutionary-based and the structural-based consensus sequences respectively. To extract the occurrence feature group from the evolutionary consensus sequence, we first need to extract this sequence from PSSM. In the evolutionary consensus sequence, amino acids along the original protein sequence $(O_1, O_2, ..., O_L)$ are replaced with the corresponding amino acids with the maximum substitution probability $(C_1, C_2, ..., C_L)$. This is done in the following two steps. In the first step, for a given amino acid, the index of the amino acid with the highest substitution probability is calculated as follows:

$$I_i = \text{argmax}\{P_{ij} : 1 \leq j \leq 20\},\ 1 \leq i \leq L, \tag{1}$$

where P_{ij} is the substitution probability of the amino acid at location i with the j^{th} amino acid in PSSM. In the second step, we replace the amino acid at i^{th} location of original protein sequence by the I^{th} amino acid to form the consensus sequence. After calculating the evolutionary consensus sequence, we count the occurrence of each amino acid (for all the 20 amino acids) along this sequence and produce the occurrence feature from the evolutionary based consensus sequence which we call (AAO). Similarly, we calculate the occurrence of each *secondary structure elements (SSEO)* (for all three elements) in the structural consensus sequence and extract the corresponding feature group. The occurrence feature group is used in this study as the global descriptor of the proteins since it maintains the information regarding the length of protein sequence which is disregarded using composition feature group [2, 5].

Semi-Occurrence: In this method, we calculate semi-occurrence feature group from both PSSM and SPINE-M. It is called semi-occurrence because instead of using the protein sequence directly to calculate the occurrence of each amino acid, we calculate the summation of the substitution probability for each amino acid from the PSSM or normalized frequency of each secondary structure element from SPINE-M. The semi-occurrence derived from the PSSM ($PSSM\_AAO$) is calculated as follows:

$$\text{PSSM-AAO}_j = \sum_{i=1}^{L} P_{ij}, (j = 1, ..., 20). \tag{2}$$

In a similar manner, we calculate the semi-occurrence of the normalized frequency of the secondary structure elements from SPINE-M ($SPINE\_SSEO$) as follows:

$$\text{SPINE-SSEO}_j = \sum_{i=1}^{L} S_{ij}, (j = 1, 2, 3), \tag{3}$$

where S_{ij} is the normalized probability of the occurrence of the j^{th} secondary structure element for the i^{th} amino acid in the SPINE-M. These feature groups

are able to provide important global discriminatory information about the substitution probability of the amino acids as well as normalized frequency of secondary structure elements based on PSSM and SPINE-M. For the rest of this study, the combination of all these four global feature groups (AAO + SSEO + PSSM-AAO + SPINE-SSEO) will be referred as F_{global} (consisting of 46 features in total).

3.2 Local Features

To extract these features, we extract distribution and auto covariance features using segmentation method. In this manner, we are able to provide more local information compared to use of whole protein sequence as a global entity to extract these features.

Segmented Distribution: This method is specifically proposed to extract more local discriminatory information for PFR based on the amino acids' substitution probability with each other (extracted from PSSM) as well as their tendency to incorporate in one of the secondary structure elements (extracted from SPINE-M). For PSSM, for the j^{th} column, we first calculate the total sum of substitution probability $T_j = \sum_{i=1}^{L} P_{ij}$. Then, starting from the first row of PSSM (which corresponds to the first amino acid in the protein sequence) we sum the substitution probabilities corresponding to the j^{th} column until reaching to less than or equal to F_P (segmentation factor) of T_j ($S_1 = \sum_{i=1}^{I_j^1} P_{ij}$). I_j^1 is the number of amino acids such that the summation of their substitution probability is equal to S_1 and is the corresponding feature for this segment. We calculate I_j^2 by summing the substitution probability of amino acids (again, starting from the first row of PSSM) until reaching $2 \times F_P$ of T_j. Similarly, I_j^2 is the number of amino acids such that the summation of their substitution probability is equal to S_2 ($2 \times F_P$ of T_j) and is the corresponding feature for this segment. In this study F_P is set to 25% since it attained similar performance as adopting 10% and 5% for this parameter. In other words, dividing the protein sequence into four segments provide similar local discriminatory information in comparison with dividing it to 10 or 20.

We also calculate I_j^3, I_j^4 features for the j^{th} column of PSSM. Dissimilar to I_j^1 and I_j^2, we start from the last row of PSSM (corresponding to the last amino acids of the protein sequence). To calculate I_j^3, starting from the last row of PSSM, we sum the substitution probabilities of amino acids until reaching less than or equal to F_P of T_j. In the similar manner, we calculate I_j^4, summing substitution probability of amino acids (starting from the last row of PSSM) until reaching to $2 \times F_P$ of total sum (T_j). In this manner, we also cover whole protein sequence as well (50% of T_j is covered by starting from the first row and 50% of T_j is covered by starting from the last row). Therefore, for a given column in PSSM we calculate 4 segmented distribution features (which in total $4 \times 20 = 80$ features are extracted corresponding to 20 columns in PSSM) to build segmented distribution feature group (called *PSSM_SD*).

In a similar manner, we calculate the segmented distribution feature group of the normalized frequency of the secondary structure elements from SPINE-M (called *SPINE_SD*) using $F_S = 25\%$ (where F_s is used as the distribution factor for SPINE-M equivalent to F_P used for PSSM) and respectively extract $3 \times 4 = 12$ features in total for all three elements.

Segmented Auto Covariance: The concept of auto covariance have been widely used in the literature to capture local discriminatory information and has attained better results compared to similar methods used for this task such as bi-gram [14, 11] or tri-gram features [16]. Pseudo amino acid composition based features are good examples of these types of features [3, 21]. These features have been computed using the whole protein sequence as a single entity for feature extraction. Therefore, they could not adequately explore the local discriminatory information embedded in protein sequence [10]. In the present study, we extend the concept of segmented distribution features as described in the previous subsection to compute the auto covariance features. This provides more local evolutionary and structural information from PSSM and SPINE-M. First for PSSM, we segment the protein sequence using $F_P = 25\%$. Using a procedure similar to the one described in the previous subsection, for the j^{th} column in PSSM we divide the protein sequence into 4 segments (from first amino acid corresponding to first row of PSSM until reaching I_j^1; from first amino acid corresponding to first row of PSSM until reaching I_j^2; from last amino acid corresponding to the last row of PSSM until reaching I_j^3; and from last amino acid corresponding to the last row of PSSM until reaching I_j^4). we calculate auto covariance feature using K_P(distance factor used for PSSM for each segment) as follows:

$$\text{PSSM-seg}_{n,m,j} = \frac{1}{(I_j^n - m)} \sum_{i=1}^{I_j^n - m} (P_{i,j} - P_{ave,j}) \times (P_{(i+m),j} - P_{ave,j}),$$
$$(n = 1, 2, 3, 4 \ \& \ m = 1, ..., K_P \ \& \ j = 1, ..., 20), \quad (4)$$

where, $P_{ave,j}$ is the average substitution probability for the j^{th} column in PSSM. We also compute the global auto covariance coefficient (K_P features) as follows:

$$\text{PSSM-AC}_{m,j} = \frac{1}{(L - m)} \sum_{i=1}^{L-m} (P_{i,j} - P_{ave,j}) \times (P_{(i+m),j} - P_{ave,j}),$$
$$(m = 1, ..., K_P \ \& \ j = 1, ..., 20). \quad (5)$$

Thus, we extract a total of ($2K_P + 2K_P + K_P = 5K_P$) auto covariance features ($2K_P$ features for segments corresponding to I_j^1 and I_j^2, $2K_P$ features for segments corresponding to I_j^3 and I_j^4 and K_P features corresponding to global auto covariance) in this manner. Then by combining PSSM-AC and PSSM-seg (extracted for all 20 columns of PSSM) we build the corresponding feature group which is called PSSM-SAC ($20 \times (5 \times K_P)$) features in total).

This procedure is also repeated for SPINE-M in the same way (K_S is used as the distance factor for SPINE-M equivalent to K_P used for PSSM) for all three columns of SPINE-M and segmented auto covariance of normalized frequency of secondary structure elements are extracted as follows:

$$\text{SPINE-seg}_{n,m,j} = \frac{1}{(I^n_{max} - m)} \sum_{i=1}^{I^n_{max}-m} (S_{i,j} - S_{ave,j}) \times (S_{(i+m),j} - S_{ave,j}),$$
$$(n = 1, 2, 3, 4 \ \& \ m = 1, ..., K_S \ \& \ j = 1, 2, 3), \quad (6)$$

where, $S_{ave,j}$ is the average substitution probability for the j^{th} column in SPINE-M. Similarly, the global auto covariance is computed as follows:

$$\text{SPINE-AC}_{m,j} = \frac{1}{(L - m)} \sum_{i=1}^{L-m} (S_{i,j} - S_{ave,j}) \times (S_{(i+m),j} - S_{ave,j}),$$
$$(m = 1, ..., K_S \ \& \ j = 1, 2, 3). \quad (7)$$

The combination of SPINE-seg and SPINE-AC builds SPINE-SAC consisting of $3 \times (5K_S)$ features in total (extracted for all three columns of SPINE-M).

4 Support Vector Machine

In pattern recognition, SVM is considered as the-state-of-the-art classification technique. It was introduced by [22] aiming at finding the *Maximum Margin Hyper-plane (MMH)* based on the concept of support vector theory to minimize classification error. It transforms the input data to higher dimensionality using the kernel function to find support vectors. 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:

$$y' = sign\left(\sum_{i=1}^{n} a_i y_i K(\mathbf{x}_i, \mathbf{x}') + b \right), \quad (8)$$

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. The best results reported in the literature for PFR was attained using this classifier [10, 11, 4, 16]. In this study, the SVM classifier implemented in LIBSVM (C-SVC type) toolbox with *Radial Basis Function (RBF)* as its kernel function is used [23]. RBF kernel is adopted here due to its better performance than other kernels functions (e.g. polynomial kernel, linear kernel, and sigmoid [10]). In this study, the width parameter γ in addition to the cost parameter C of the SVM are optimized using grid search algorithm implemented in the LIBSVM package.

5 Results and Discussion

We construct the input feature vector to use with SVM consisting of our extracted feature (F_{global} + PSSM-SD + SPINE-SD + PSSM-SAC + SPINE-

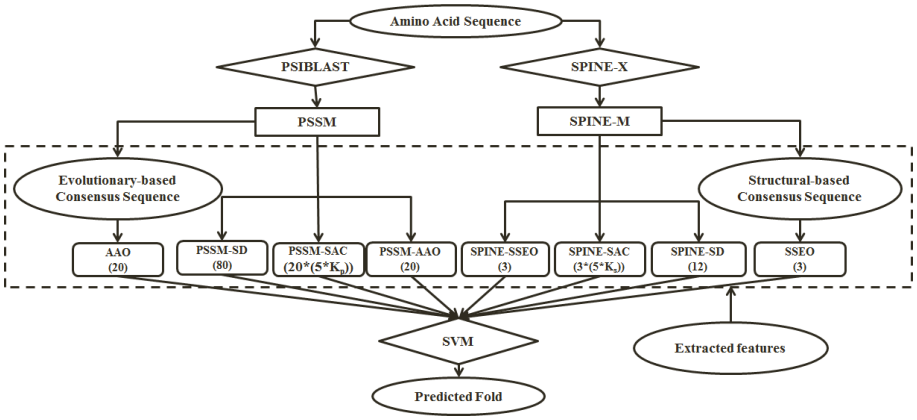


Fig. 1. The general architecture of our proposed feature extraction model. The number of features extracted in each feature group is shown in the brackets below the feature groups' names.

SAC). The architecture of our proposed system is shown in Figure 1. To evaluate the performance of our proposed methods, 10-fold cross validation evaluation criterion is adopted in this study as it was often used for this task in the literature [1, 5, 11, 15]. We first investigate the impact of our proposed method for PFR with respect to the K_p and K_s parameters in PSSM-SAC and SPINE-SAC respectively. Then we investigate the impact of each of the proposed feature groups in this study separately on the achieved prediction accuracy. Finally, we compare our achieved results with previously reported results for the PFR.

5.1 Investigating the Impact of K_p and K_s

As it was mentioned earlier, K_p and K_s values between 1 and 10 are investigated here (since it was shown in [10] that using a distance factor larger than 10 to extract auto covariance feature group attains similar results with using 10 for PFR). To do this, in 10 different experiments, we apply SVM to our proposed feature vector while K_p and K_s are monotonically increased from 1 to 10 ($K_p = 1$ and $K_s = 1$, $K_p = 2$ and $K_s = 2$, ... , $K_p = 10$ and $K_s = 10$). The results for this experiment is shown in Figure 2. We also calculate the SVM parameters on EDD data set (where $K_p = 10$ and $K_s = 10$) for our proposed feature vector using the grid search algorithm. Calculated parameters are used for the rest of this study (to avoid over tuning parameters) for both TG and EDD data sets (where C = 0.07 and $\gamma = 100$). Note that the TG data sets have not been used at all for parameter tuning.

As we can see, increasing the K_p and K_s, prediction accuracy almost monotonically increases as well. Using $K_p = 10$ and $K_s = 10$, we reach 88.1% and 73.1% prediction accuracies for EDD and TG data sets respectively. However, it is not clear which one of K_p and K_s has the main impact on the achieved

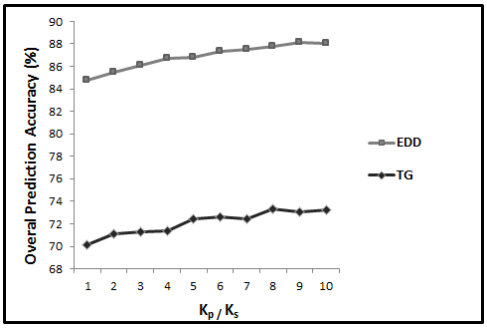
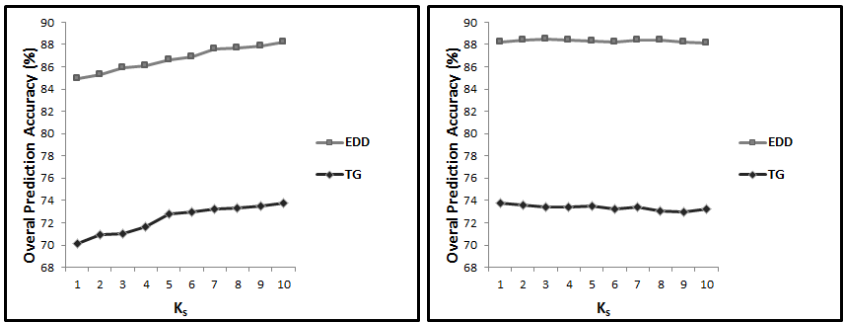


Fig. 2. The results achieved for TG and EDD data sets with respect to K_p and K_s which are monotonically increase from 1 to 10.

results. To investigate the effectiveness of K_p and K_s, two different experiments are conducted on the EDD data set. First, we set the value of $K_p = 1$ and in 10 different experiments, increase the value of K_s from 1 to 10 (Figure 3.a). As we can see, increasing K_s monotonically increases the prediction accuracy and setting $K_s = 10$ attain the best result for this task. In a different experiment, we set the value of $K_s = 10$ and in 10 different experiments, increase the value of K_p from 1 to 10. As we can see in Figure 3.b, the performance does not change by increasing the K_p. As it is shown in Figure 3.a and 3.b, similar results are achieved for the TG data set. In other words, using segmented auto covariance approach, we are able to reveal more local discriminatory information from PSSM and SPINE-M based on the concept of auto covariance compared to previous studies ($K_P = 1$ and $K_S = 10$). It is dramatically lower than the number of features used in [10] and [11] to reveal this information. Therefore, for the rest of this study K_p and K_s are set to 1 and 10 respectively.



(a) The impact of increasing K_s from 1 to 10 while $K_p = 1$ for EDD and TG data sets

(b) The impact of increasing K_p from 1 to 10 while $K_s = 10$ for EDD and TG data sets

Fig. 3. Investigating the effective values for K_s and K_p in our proposed feature extraction method.

5.2 Determining the Effect of the Proposed Feature Groups on the Protein Fold Prediction Accuracy

In continuation, we investigate the effectiveness of each of the feature groups used in this study separately to our reported protein fold prediction accuracy. The results are shown in Table 1. As we can see, all the feature groups used to reveal global and local discriminatory information are effectively contribute to the achieved protein fold prediction enhancement.

Table 1. The impact of proposed feature groups proposed in this study (using SVM classifier) to enhance protein structural class prediction accuracy (in %). For PSSM-SAC and SPINE-SAC, the values of K_p and K_s are respectively set to 1 and 10.

Combination of features	EDD	TG
F_{global}	74.7	58.7
F_{global} + PSSM-SD	79.4	62.6
F_{global} + SPINE-SD	79.1	63.6
F_{global} + PSSM-SD + SPINE-SD	82.3	66.7
F_{global} + PSSM-SAC	80.1	64.0
F_{global} + SPINE-SAC	84.1	68.2
F_{global} + PSSM-SAC + SPINE-SAC	86.1	71.8
F_{global} + PSSM-SD + SPINE-SD + PSSM-SAC	87.5	72.6
F_{global} + PSSM-SD + SPINE-SD + SPINE-SAC	87.1	72.8
PSSM-SD + SPINE-SD + PSSM-SAC + SPINE-SAC	85.9	71.1
F_{global} + PSSM-SD + SPINE-SD + PSSM-SAC + SPINE-SAC	88.2	73.8

5.3 Comparison with the Existing Methods

We compared the results achieved by applying SVM to the combination of features proposed in this study (F_{global}, PSSM-SAC, PSSM-SD, SPINE-SAC, SPINE-SD where K_p and K_s are set to 1 and 10 respectively) which will be referred as PSSM-SPINE-S (388 features in total) with the best results reported in the literature. The results are shown in Table 2. As we can see, we report up to 73.8% and 88.2% prediction accuracies for TG and EDD data sets respectively. These results are up to 7.4% and 2.3% better than the highest reported results for these two data sets that are achieved by reproducing the results reported in [10] for TG and EDD data sets respectively. The enhancement achieved compared to other similar approaches to reveal more local information such as bi-gram [11] and tri-gram [16] is much more significant (over 11% for EDD and TG data sets). The higher enhancement achieved for TG data set compared to [10] shows that our method is more effective when the sequential similarity rate is very low (up to 25%). It is also important to highlight that we outperformed [10] using 388 features compared to 4000 features used in that study. Therefore, our proposed methodology is able to significantly enhance protein fold prediction accuracy compared to the state-of-the-art methods found in the literature and at the same time reduce the number of features used for this task dramatically. In other words, we are able to provide more local and global information from PSSM and SPINE-X for PFR compared to previously proposed approaches found in the literature.

Table 2. Comparison of the results reported EDD and TG data sets (in %). Note that column named No. is referring to the number of features.

Ref.	Features	No.	Method	EDD	TG
[15]	AAO (from original protein sequence)	20	LDA	46.9	36.3
[15]	AAC (from original protein sequence)	20	LDA	40.9	32.0
[1]	Physicochemical Features + AAC	125	SVM	50.1	39.5
[13]	Physicochemical Features + AAC	220	ANN(RBF)	52.8	41.9
[17]	Threading	-	Naive Bayes	70.3	55.3
[16]	PF (bi-gram)	400	SVM	75.2	52.7
[16]	TF (Tri-gram)	8000	SVM	71.0	49.4
[11]	Combination of bi-gram features	2400	SVM	69.9	55.0
[5]	PSIPRED and PSSM features	242	SVM	77.5	60.1
[10]	ACCfold-AC	200	SVM	80.1	58.8
[10]	ACCfold-ACC	4000	SVM	85.9	66.4
This study	PSSM-SPINE-S	388	SVM	88.2	73.8

6 Conclusion

In this study, we have proposed two novel segmentation based feature extraction techniques to reveal more local discriminatory information embedded in PSSM and SPINE-X. We also employed the concept of occurrence feature group and extend it to provide more global discriminatory information from PSSM and SPINE-X for PFR compared to previously used methods for this task. Then by applying SVM to the combination of our features extracted we significantly enhanced protein fold prediction accuracy compared to previously reported results in the literature. We achieved up to 73.8% and 88.2% prediction accuracies, up to 7.4% and 2.3% better than the highest results reported for TG and EDD data sets respectively [10]. These enhancements were achieved by using less than 1/10 of features used previously in [10]. In other words, we were able to extract more potential local and global discriminatory information for PFR compared to previously proposed methods found in the literature using fewer features.

References

1. Ding, C., Dubchak, I.: Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics **17** (2001) 349–358
2. Chen, K., Kurgan, L.A.: Pfres: protein fold classification by using evolutionary information and predicted secondary structure. Bioinformatics **23**(21) (2007) 2843–2850
3. Shen, H.B., Chou, K.C.: Ensemble classifier for protein fold pattern recognition. Bioinformatics **22** (2006) 1717–1722
4. Damoulas, T., Girolami, M.: Probabilistic multi-class multi-kernel learning: On protein fold recognition and remote homology detection. Bioinformatics **24**(10) (2008) 1264–1270
5. Deschavanne, P., Tuffery, P.: Enhanced protein fold recognition using a structural alphabet. Proteins: Structure, Function, and Bioinformatics **76**(1) (2009) 129–137
6. 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) (2010) 1941–1956

7. 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
8. 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
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. 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
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. Chen, K., Stach, W., Homaeian, L., Kurgan, L.: ifc2: an integrated web-server for improved prediction of protein structural class, fold type, and secondary structure content. Amino Acids **40** (2011) 963–973
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. Taguchi, Y.H., Gromiha, M.M.: Application of amino acid occurrence for discriminating different folding types of globular proteins. BMC Bioinformatics **8**(1) (2007)
16. 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
17. Gromiha, M.M.: Multiple contact network is a key determinant to protein folding rates. Journal of Chemical Information and Modeling **49**(4) (2009) 1130–1135
18. 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
19. Faraggi, E., Zhang, T., Yang, Y., Kurgan, L., Zhou, Y.: Spine x: Improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. Journal of Computational Chemistry **33**(3) (2012) 259–267
20. Jones, D.T.: Protein secondary structure prediction based on position-specific scoring matrices. Journal of Molecular Biology **292**(2) (1999) 195–202
21. Shen, H.B., Chou, K.C.: Predicting protein fold pattern with functional domain and sequential evolution information. Journal of Theoretical Biology **256**(3) (2009) 441–446
22. Vapnik, V.N.: The nature of statistical learning theory. Springer-Verlag New York, Inc (1995)
23. Chang, C.C., Lin, C.J.: Libsvm: a library for support vector machines. (2001)