A Segmentation-based Method to Extract Structural and Evolutionary Features for Protein Fold Recognition

Abdollah Dehzangi, Kuldip Paliwal, James Lyons, Alok Sharma, Abdul Sattar

Abstract—Protein fold recognition (PFR) is considered as an important step towards the protein structure prediction problem. Despite all the efforts that have been made so far, finding an accurate and fast computational approach to solve the PFR still remains a challenging problem for bioinformatics and computational biology. 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 for 7.4% over the best results reported in the literature. We also report 73.8% prediction accuracy for a data set consisting of proteins with less than 25% sequence similarity rates and 80.7% prediction accuracy for a data set with proteins belonging to 110 folds with less than 40% sequence similarity rates. We also investigate the relation between the number of folds and the number of features being used and show that the number of features should be increased to get better protein fold prediction results when the number of folds is relatively large.

Index Terms—Protein Fold Recognition, Feature Extraction, Structural-based Features, Evolutionary-based Features, Segmented distribution, Segmented Auto Covariance, Occurrence, Support Vector Machine (SVM)

A Protein consists of a sequence of monomers called amino acids that are connected to each other through peptide bonds. Proteins are considered as the most important biological micro-molecules and play a vital role in most of the biological interactions. Therefore, determining how it functions is an important task in biology and biomedical science. *Protein Fold Recognition (PFR)* is considered as an important step towards protein function prediction. 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. Despite all the efforts that have been made during the last two decades, finding an effective computational approach to solve this problem still remains crucial for computational biology and bioinformatics [1]. In the pattern recognition terminology, the PFR is defined as solving a multi-class classification task in which its performance is crucially relied on the features and classification techniques being used. Features that capture significant global and local discriminatory information and the classification techniques that perform consistently with these extracted features have been used in the literature [2]. A wide range of classification techniques such as, *Artificial Neural Networks (ANN)* [3], [4], [5], [6], [7], [8], Meta Classifiers [9], [10], [11], [12], [13], K-Nearest Neighbors [14], [15], [16], [17], [18] and *Support Vector Machines (SVM)* [19], [20], [21], [22], [23], [24], [25], [26], [27] have been used for the PFR. Among the classifiers employed to tackle the PFR, using *Support Vector Machine (SVM)* have attained the best results [26], [27], [28], [29], [30], [31], [32]. Similarly, a wide range of features have been extracted and used to tackle the PFR such as, Physicochemical-based features [19], [23], [33], [34], Sequence-based features [6], [14], [15], [32] Evolutionary-based features [18], [25], [28], [30], and Structural-based features [17], [18], [23], [35], [36], [37], [38]. Achieved results have shown that the most significant enhancement for the protein fold prediction accuracy has been achieved by relying on the feature extraction approaches rather than the classification techniques being used [4], [15], [19], [27], [28], [29], [39]. In most of the studies that addressed the PFR by feature extraction techniques, global discriminatory information has been represented using the composition of the amino acids feature group (the percentage of the occurrence of the amino acids along the protein sequence divided by the length of protein sequence [19], [22], [30]). However, it has been shown that this feature group is not able to adequately reveal global information as it is not able to

- *Abdollah Dehzangi is with Institute for Integrated and Intelligent Systems (IIIS), Griffith University, and National ICT Australia (NICTA), Brisbane, Australia. Email: a.dehzangi@griffith.edu.au*
- *Kuldip Paliwal is with School of Engineering, Griffith University, Brisbane, Australia. Email: k.paliwal@griffith.edu.au*
- *James Lyons is with School of Engineering, Griffith University, Brisbane, Australia. Email: j.lyons@griffith.edu.au*
- *Alok Sharma is with School of Engineering and Physics, University of the South Pacific, Fiji and Adjunct Associate Professor at the Institute for Integrated and Intelligent Systems (IIIS), Griffith University. Email: sharma_al@usp.ac.fj*
- *Abdul Sattar is with Institute for Integrated and Intelligent Systems, Griffith University, and National ICT Australia (NICTA), Brisbane, Australia. Email: a.sattar@griffith.edu.au*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2013.2296317, IEEE/ACM Transactions on Computational Biology and Bioinformatics

capture information regarding the length of the protein sequence [39], [40] which was shown as effective feature for the PFR [33], [41].

Compared to the methods adopted to extract global discriminatory information, a wider range of methods were used to extract local discriminatory information for the PFR [42] such as, pseudo amino acid composition [14], [15], [22], [23], [43], cross covariance [28], auto covariance [28], [42], bi-gram [29], [39], and tri-gram [3]. Despite the significant local discriminatory information provided using these approaches, most of these methods produce large number of features as well as large amount of redundant features [14], [29] which makes them computationally expensive for large protein data banks (e.g. cross covariance and tri-gram [3], [28]). At the same time, in all these methods the whole protein sequence as a single entity have been used to extract local information. In other words, they aimed to extract local information by exploring whole protein sequence as a global entity [17], [28]. Therefore, they could not appropriately explore local information embedded in protein sequence [33], [40]. As a results, the protein fold prediction accuracy remains limited especially when the sequence similarity rate is low [39].

In this study, we aim at enhancing protein fold prediction accuracy by addressing these limitations. We propose a segmentation-based feature extraction method 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. In this method, we divide the protein sequence into several segments and extract distribution and auto-covariance features from each segment. 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 the PFR. We investigate the impact and generality of our proposed methods on four data sets including a data set consisting of proteins with less than 25% sequence similarity and a data set consisting of proteins belonging to 110 different folds with less than 40% sequence similarity. By applying SVM to our extracted features we enhance the protein fold prediction accuracy by 7.4% better than the highest reported results found in the literature.

1 Data sets

In this study, four data sets namely TG, EDD, F92, and F110 are used to investigate the performance of our proposed methods. The TG data set introduced by [40] consists of 1612 proteins belonging to 30 folds with less than 25% sequence 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 the PFR when the sequence similarity is very low [33], [40], [44]. We also extract EDD (extended version of DD data set [19] 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% sequence similarities. The EDD data set extracted from an older version of SCOP has been widely used for the PFR [21], [28], [29]. This data set enables us to directly compare our results with previously reported results found in the literature.

To investigate the impact of our proposed methods for more complex (regarding to the number of folds containing in the data set) data sets consisting of larger number of folds, we have extracted two new data sets namely F92 and F110 from the SCOP 1.75. F92 data set consists of 6331 proteins belonging to 92 most populated folds in SCOP. This data set is generated by making sure that each fold at least consists of 25 proteins. The F110 contains 6723 proteins belonging to 110 folds in which each fold has more than 20 proteins. Having larger number of folds, these two data sets are able to provide more reliable and general results for the PFR. Furthermore, comparison of the results achieved for the EDD, F92, and F110 can provide important information regarding the impact of increasing the number of folds and consequently complexity of the problem on the PFR performance. These four data sets are available upon request.

2 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 [45] 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. Extracted features from PSSM have been widely used for the PFR and attained promising results [27], [28], [29].

We also use predicted secondary structure using SPINE-X which was recently proposed by [46] and attained better results (especially for the coded area) than PSIPRED on predicting protein secondary structure [47]. 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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2013.2296317, IEEE/ACM Transactions on Computational Biology and Bioinformatics

structure using SPINE-X provides significant structural information for the PFR similar to or even better than PSIPRED due to its better performance [17], [23], [30], [46].

During the last decade, the substitution score (extracted from the PSSM) and predicted secondary structure using PSIBLAST and SPINE-X (or PSIPRED before that) have been widely used in protein science (e.g. protein fold recognition, protein function prediction, protein structure prediction, protein subcellular localization) and extracted features from these sources attained promising results [23], [28], [30], [38], [42], [48]. As it is highlighted in [49], the most sensitive methods for fold recognition use sequence profiles to represent both the query and the data base proteins. The robustness and sensitivity of PSSM and SPINE-X for feature extraction have been addressed in [23], [46], [49]. In continuation, the global and local features extracted in this study will be explained in detail. The SPINE-M and PSSM for all four data sets used in this study are available upon request.

2.1 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 the secondary structure elements, respectively.

2.1.1 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_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 [21], [23], [50].

2.1.2 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).

2.2 Local Features

To extract these features, we use a segmentation method described below and extract distribution and auto covariance features from the individual segments. In this manner, we are able to provide more local information compared to the global features described earlier.

2.2.1 Segmented Distribution Features:

Here, we first apply a segmentation method to individual columns of PSSM and SPINE-M matrices, and represent each segment by a distribution feature. 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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2013.2296317, IEEE/ACM Transactions on Computational Biology and Bioinformatics

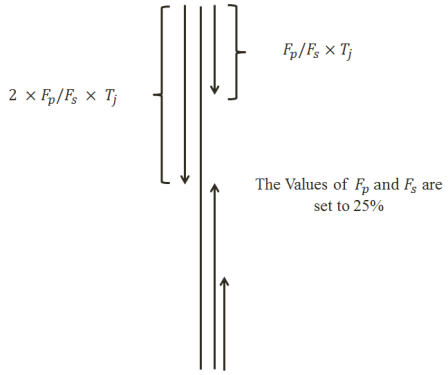


Fig. 1: The segmentation method adopted in this study. F_p and F_s are segmentation factors used for PSSM and SPINE-M, respectively.

first amino acid in the protein sequence), we compute the sum ($S_1 = \sum_{i=1}^{I_j^1} P_{ij}$) of the substitution probabilities corresponding to the j^{th} column until reaching to less than or equal to F_P (segmentation factor) of T_j, where I_j^1 is the number of amino acids used in computing the sum S_1. The amino acids from index 1 to I_j^1 are considered to define the first segment for the j^{th} column of the PSSM. We use the index of last amino acid of this segment (i.e., I_j^1) as the first segmented distribution feature. Next, we calculate the index of the last amino acid in the second segment (i.e., 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, and use it as second segmented distribution feature. In this study, F_P is set to 25% as other values of F_P (10% and 5%) attained similar performance.

Next, we calculate two more features (I_j^3, I_j^4) for the j^{th} column of PSSM by carrying out the segmentation in reverse order. Instead of starting from the first row of j^{th} column of PSSM as done previosuly to compute I_j^1 and I_j^2, we now 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). Thus, we calculate 4 segmented distribution features for each column in PSSM. This means that we will have a total of 4 $\times$ 20 = 80 features for 20 columns in PSSM to build segmented distribution feature group (called *PSSM_SD*). The configuration of the segmentation scheme adopted in this study is shown in Figure 1.

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. In segmentation-based distribution feature extraction technique, we extract the distribution factor in which explains how amino acids are distributed along protein sequence with respect to their substitution scores. Therefore, it returns the index which is equal to the number of amino acids for each segment while semi-occurrence returns the summation of substitution scores. The distribution factor has been shown as effective features which it's emphasized as well in this study [19].

2.2.2 Segmented Auto Covariance Features:

The concept of auto covariance has been widely used in the literature to capture local discriminatory information and has attained better results compared to bi-gram [29], [39] or tri-gram features [3]. Pseudo amino acid composition based features are good examples of these types of features [14], [17]. 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 [28]. 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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2013.2296317, IEEE/ACM Transactions on Computational Biology and Bioinformatics

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:

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

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:

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

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).

3 Support Vector Machine

In pattern recognition, SVM is considered as the-state-of-the-art classification technique. It was introduced by [51] 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:

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

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. Among a wide range of complex classification techniques used for the PFR [5], [16], [20], [52], [53], the best results reported in the literature was attained using this classifier [28], [29], [30]. In this study, the SVM classifier implemented in LIBSVM (C-SVC type and using one-versus-one approach to extend it for multi-class classification task) toolbox with *Radial Basis Function (RBF)* as its kernel function is used [54]. RBF kernel is adopted here due to its better performance than other kernels functions (e.g. polynomial kernel, linear kernel, and sigmoid [28]). 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. The grid search algorithm tries various pairs of γ and C values and selects the values with the best classification accuracy [54].

4 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-SAC). The architecture of our proposed system is shown in Figure 2. To evaluate the performance of our proposed methods, 10-fold cross validation evaluation criterion is adopted in this study as it was mostly used for this task in the literature [19], [21], [29], [40]. We first investigate the impact of our proposed method for the 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.

4.1 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 [28] that using a distance factor larger than 10 to extract auto covariance feature group attains similar results with using 10 for the 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 3. We also calculate the SVM parameters on the 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 all four data sets used in this study (where C = 0.075 and $\gamma = 100$). We have also conduct the parameter tuning for the F92 data set in which attained to its best results with the same parameters extracted for the EDD. Note that the TG and F110 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 to 88.1%, 73.1%, 81.2%, and 80.4% prediction accuracies for the EDD, TG, F92, and F110 data sets respectively. However, it is not clear which one of K_p and K_s has the main impact on the achieved results. Furthermore, the impact of increasing K_p and K_s on the EDD and TG data sets are slightly different from F92 and F110. As it is shown in Figure 3, increasing K_p and K_s, from 1 to 10, the prediction accuracy almost monotonically increases for the EDD and TG data sets while for the F92 and F110, it monotonically increase until $K_p = 7$ and $K_s = 7$ and then it remains unchanged (and slightly drops). Having significantly different number of folds in the F92 and F110, it is expected that addressing the PFR using

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2013.2296317, IEEE/ACM Transactions on Computational Biology and Bioinformatics

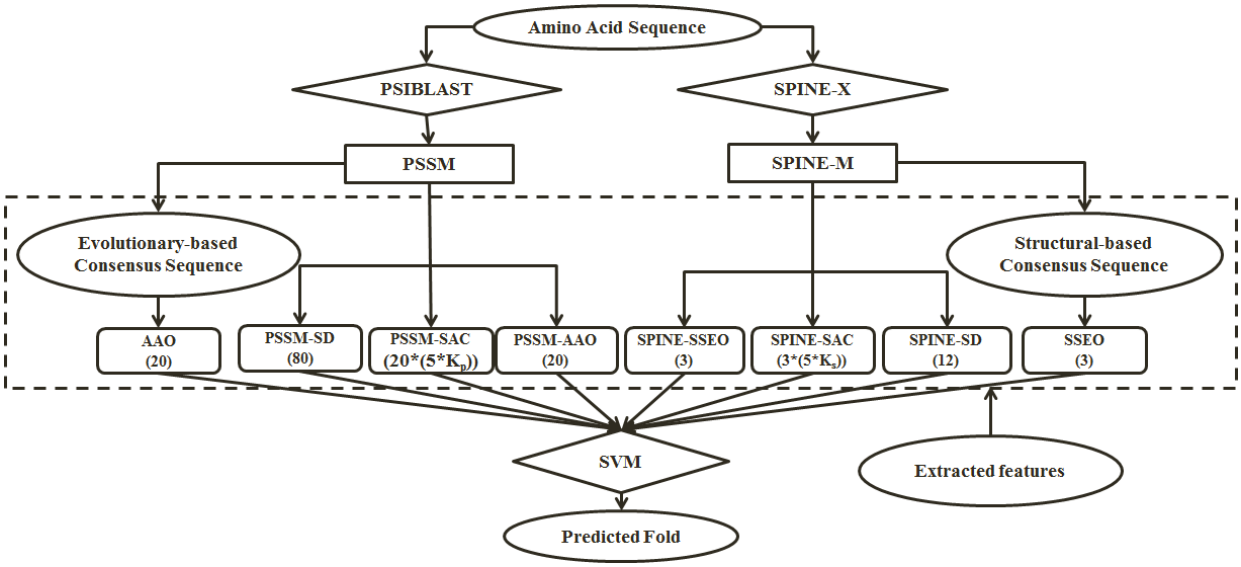


Fig. 2: The general architecture of our proposed feature extraction model. The number of features extracted in each feature group is shown in the brackets below the feature groups' names.

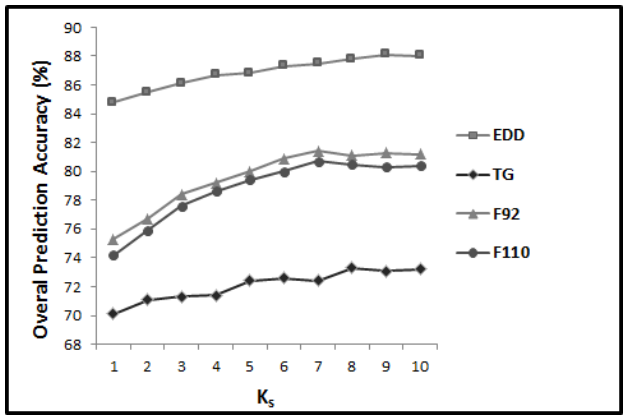


Fig. 3: The results achieved for EDD, TG, F93, and F110 data sets with respect to K_p and K_s which are monotonically increase from 1 to 10.

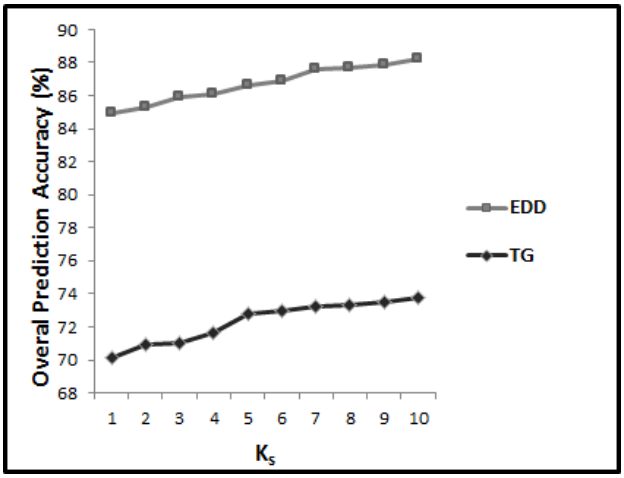
these two data sets would be more complex and slightly different from the EDD and TG.

Therefore, we investigate the impact of K_p and K_s for two cases. We first investigate the impact of these two parameters on the EDD and TG data sets and then for the F92, and F110 data sets, separately. To investigate the effectiveness of K_p and K_s on the EDD data set (then explore the impact on the TG data set), two different experiments are conducted. 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 4.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 $K_s = 10$ (since the best results attained by adjusting $K_s = 10$) and in 10 different experiments, increase the value of K_p from 1 to 10. As we can see in Figure 4.b, the performance does not change by increasing the K_p. As it is shown in the Figures 4.a and 4.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 $K_S = 10$). Note that this number of features is dramatically lower than the number of features used in [28] and [29] to enhance the PFR accuracy.

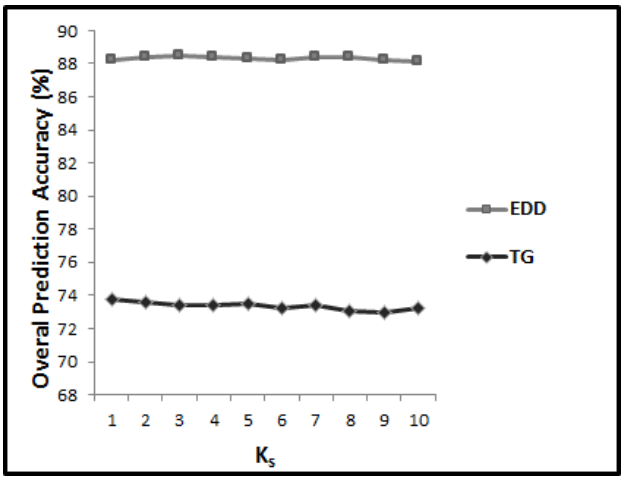
In the similar manner, we explore the impact of K_p and K_s on the F92 in two different experiments. First, we set the value of $K_p = 1$ and in 10 different experiments, increase the value of K_s from 1 to 10 (Figure 5.a). As we can see, increasing K_s monotonically increases the prediction accuracy until $K_s = 7$ and then the prediction accuracy remains almost similar (slightly drops). Therefore, we set $K_s = 7$. In a different experiment, we set the value of $K_s = 7$ and in 10 different experiments, increase the value of K_p from 1 to 10. As we can see in Figure 5.b, different to the EDD and TG data sets, the prediction accuracy for the F92 by increasing the K_p until $K_p = 7$ increases and then the (increasing K_p/K_s from 7 to 10) prediction performance remains unchanged. As it is shown in Figures 5.a and 5.b, similar results are achieved for the F110 data set as well.

As it was discussed earlier, the number of folds investigated in the F92 and F110 are significantly higher than the number of folds explored in the EDD and TG data sets (over three times). The difference between the number of features (from 388 to 943) used for the EDD and TG data sets compared to the F92 and F110 data sets to provide more effective features indicates that by increasing the complexity of the problem, the amount of discriminatory information for classification task needs to be increased as well. Therefore, the number of folds explored in the employed data set can be considered as a parameter that impact on the segmentation factors to extract segmentation based features from PSSM and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2013.2296317, IEEE/ACM Transactions on Computational Biology and Bioinformatics

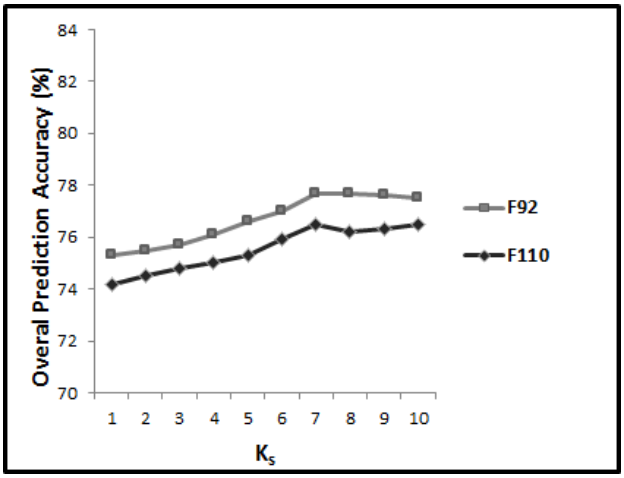


(a) The impact of increasing K_s from 1 to 10 while $K_p = 1$ for EDD and TG data sets.

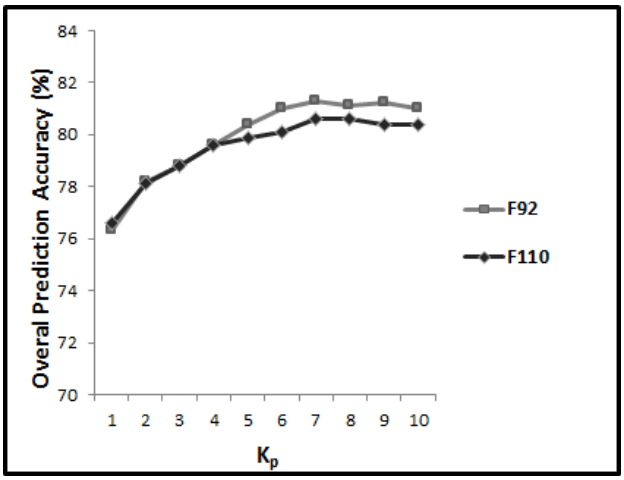


(b) The impact of increasing K_p from 1 to 10 while $K_s = 10$ for EDD and TG data sets.

Fig. 4: The impact of adjusting K_p and K_s for EDD and TG data sets



(a) The impact of increasing K_s from 1 to 10 while $K_p = 1$ for F92 and F110 data sets.



(b) The impact of increasing K_p from 1 to 10 while $K_s = 10$ for F92 and F110 data sets.

Fig. 5: The impact of adjusting K_p and K_s for F92 and F110 data sets

SPINE-M. For this case, having up to 110 folds we are able to provide effective discriminative information by adjusting K_P and K_S both to 7. Note that, this number of features is still much lower than the number of features used in [28] and [29] to reveal effective discriminatory information. Therefore, for the rest of this study, K_p and K_s are set to 1 and 10 respectively for the EDD and TG data sets as representatives of the data sets with limited number of folds (less than 30 folds) while they are both set to 7 for the F92 and F110 as representatives of the data sets with significantly higher number of folds and consequently more complex cases (up to 110 folds).

4.2 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 for the EDD and TG data sets (in which $K_p = 1$ and $K_s = 10$) are shown in Table 1 and the results for the F92 and F110 data sets (in which $K_p = 7$ and $K_s = 7$) are shown in Table 2. As we can see, for both of the Tables 1 and 2, all the feature groups used to reveal global and local discriminatory information are effectively contribute to the achieved protein fold prediction enhancement. It shows that the protein fold prediction enhancement reported here is dependent on the all of the feature groups proposed in this study.

4.3 Comparison with the Existing Methods

We compare 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 for the EDD and TG data sets and both are set to 7 for the F92 and F110) which will be referred as PSSM-SPINE-S (388 and 943 features in total for the EDD/TG and F92/F110 data sets) with the best results reported in the literature. The results are shown in the Tables 3 and 4. As we can see in Table 2, we report up to 73.8% and 88.2% prediction

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2013.2296317, IEEE/ACM Transactions on Computational Biology and Bioinformatics

TABLE 1: The impact of proposed feature groups proposed in this study (using SVM classifier) to enhance protein fold prediction accuracy (in %) for the EDD, TG, F92, and F110 data sets. For EDD and TG data sets and in PSSM-SAC and SPINE-SAC feature vectors, the values of K_p and K_s are set to 1 and 10 respectively while for the F92 and F110 feature vectors, these two values (K_p and K_s) are both set to 7.

Combination of features	EDD	TG	F92	F110
F_{global}	74.7	58.7	64.8	64.1
F_{global} + PSSM-SD	79.4	62.6	72.6	72.2
F_{global} + SPINE-SD	79.1	63.6	69.1	68.0
F_{global} + PSSM-SD + SPINE-SD	82.3	66.7	74.1	73.1
F_{global} + PSSM-SAC	80.1	64.0	77.6	77.1
F_{global} + SPINE-SAC	84.1	68.2	72.9	72.4
F_{global} + PSSM-SAC + SPINE-SAC	86.1	71.8	79.9	79.2
F_{global} + PSSM-SD + SPINE-SD + PSSM-SAC	87.5	72.6	79.1	78.7
F_{global} + PSSM-SD + SPINE-SD + SPINE-SAC	87.1	72.8	77.3	76.8
PSSM-SD + SPINE-SD + PSSM-SAC + SPINE-SAC	85.9	71.1	80.3	79.9
F_{global} + PSSM-SD + SPINE-SD + PSSM-SAC + SPINE-SAC	88.2	73.8	81.4	80.7

TABLE 2: Comparison of the results reported for the EDD, TG, F92, and F110 data sets (in %). Note that column named No. is referring to the number of features. For EDD and TG data sets and in PSSM-SAC and SPINE-SAC feature vectors, the values of K_p and K_s are set to 1 and 10 respectively which for the F92 and F110 data sets, these two values (K_p and K_s) are both set to 7.

Ref.	Features	No.	Method	EDD	TG	F92	F110
[40]	AAO (from original protein sequence)	20	LDA	46.9	36.3	32.7	31.3
[40]	AAC (from original protein sequence)	20	LDA	40.9	32.0	30.2	29.6
[19]	Physicochemical Features + AAC	125	SVM	50.1	39.5	39.1	38.4
[33]	Physicochemical Features + AAC	220	ANN(RBF)	52.8	41.9	43.4	41.9
[44]	Threading	-	Naive Bayes	70.3	55.3	56.3	54.8
[3]	PF (bi-gram)	400	SVM	75.2	52.7	60.2	59.5
[3]	TF (Tri-gram)	8000	SVM	71.0	49.4	63.2	62.1
[29]	Combination of bi-gram features	2400	SVM	69.9	55.0	69.9	55.0
[21]	PSIPRED and PSSM features	242	SVM	77.5	60.1	70.5	68.8
[28]	ACCfold-AC	200	SVM	80.1	58.8	68.2	68.0
[28]	ACCfold-ACC	4000	SVM	85.9	66.4	78.2	77.3
This study	PSSM-SPINE-S	388	SVM	88.2	73.8	81.4	80.7

accuracies for the 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 [28] for the TG and EDD data sets (66.4% and 85.9% prediction accuracies) respectively. The enhancement achieved compared to other similar approaches to reveal more local information such as bi-gram [29] and tri-gram [3] is much more significant (over 11% for the EDD and TG data sets). The higher enhancement achieved for the TG data set compared to [28] shows that our method is more effective when the sequence similarity rate is very low (up to 25%). It is also important to highlight that we outperformed [28] using 388 features (for the EDD and TG data sets) compared to 4000 features used in that study.

Similarly, as shown in Table 3, for the first time, we report over 80% prediction accuracy for a data set that has over 100 folds. We report 81.4% and 80.7% prediction accuracies for the F92 and F110 data sets respectively. These results are up to 3.2% and 3.4% better than the highest reported results for these two data sets that are achieved by reproducing the results reported in [28] for the F92 and F110 data sets (78.2% and 77.3% prediction accuracies) respectively. It is also important to highlight that we outperformed [28] using 943 features compared to 4000 features used in that study. Despite having larger number of features (943 features) compared to the number of features used for the EDD and TG data sets (388 features), it is still dramatically lower than the number of features used in the state-of-the-art methods used for the PFR. We have also conducted pairwise t-test to investigate the statistical significance of our reported results. the probability value calculated for the pairwise t-test ($p = 0.001$) emphasizes the statistical significance of our reported results and the enhancement achieved in this study. In result, our proposed methodology is able to significantly enhance the protein fold prediction accuracy compared to the previous studies found in the literature and at the same time reduce the number of features used for this task significantly. In other words, we are able to provide more local and global information from PSSM and SPINE-X for the PFR compared to previously proposed approaches found in the literature.

5 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 the PFR compared to previously used methods for this task. We have used four data sets namely, the EDD, TG, F92, and F110 have been used to investigate the impact of our proposed feature extraction methods. These data sets enabled us to investigate the impact of our proposed methods when the sequence similarity rate was very low (TG), when larger number of folds (and consequently more complex case) was used (F92, and F110), and to directly compare our results with the state-of-the-art methods found in the literature (EDD). By applying SVM to the combination of our extracted features we significantly enhanced protein fold prediction accuracy compared to previously reported results in the literature.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2013.2296317, IEEE/ACM Transactions on Computational Biology and Bioinformatics

For the EDD and TG data sets, we achieved up to 73.8% and 88.2% prediction accuracies, up to 7.4% and 2.3% better than previous results found in the literature, respectively [28]. These enhancements were achieved by using less than 1/10 of the number of features used previously in [28].

By investigating the PFR using the F92 and F110 data sets, we showed that by increasing the number of folds the complexity of the problem is increasing and therefore, more discriminatory information is required to tackle this problem. We also showed by using segmentation based feature extraction technique, we are able to tackle this problem as well and enhance the protein fold prediction accuracy. For the first time, we reported over 80% prediction accuracy for a data set containing proteins belonging to over 100 folds. We achieved to 81.4% and 80.7% prediction accuracies for the F92, and F110 data sets respectively, up to 3.2% and 3.4% better than previous studies found in the literature [28] using less than 1/4 of the number of features that they have used. In other words, we were able to extract more potential local and global discriminatory information for the PFR compared to previously proposed methods found in the literature using fewer features.

ACKNOWLEDGEMENTS

NICTA is funded by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Centre of Excellence Program.

REFERENCES

[1] A. Chinnasamy, W. K. Sung, and A. Mittal, "Protein structure and fold prediction using tree augmented naive bayesian classifier," *Bioinformatics and Computational Biology*, vol. 3, pp. 803–819, 2005.

[2] S. Y. M. Shi, P. N. Suganthan, and D. Kalyanmoy, "Multiclass protein fold recognition using multiobjective evolutionary algorithms," in *Computational Intelligence in Bioinformatics and Computational Biology, 2004. CIBCB '04. Proceedings of the 2004 IEEE Symposium on*, 2004, pp. 61–66.

[3] P. Ghanty and N. R. Pal, "Prediction of protein folds: Extraction of new features, dimensionality reduction, and fusion of heterogeneous classifiers," *NanoBioscience, IEEE Transactions on*, vol. 8, no. 1, pp. 100–110, 2009.

[4] C. D. Huang, C. T. Lin, and N. R. Pal, "Hierarchical learning architecture with automatic feature selection for multiclass protein fold classification," *NanoBioscience, IEEE Transactions on*, vol. 2, no. 4, pp. 221–232, 2003.

[5] Y. Chen, X. Zhang, M. Q. Yang, and J. Y. Yang, "Ensemble of probabilistic neural networks for protein fold recognition," in *Bioinformatics and Bioengineering, 2007. BIBE 2007. Proceedings of the 7th IEEE International Conference on*, 2007, pp. 66–70.

[6] K. L. Lin, C. Y. Lin, C. D. Huang, H. M. Chang, C. Y. Yang, C. T. Lin, C. Y. Tang, and D. F. Hsu, "Feature selection and combination criteria for improving accuracy in protein structure prediction," *NanoBioscience, IEEE Transactions on*, vol. 6, no. 2, pp. 186–196, 2007.

[7] A. Dehzangi, S. Phon-Amnuaisuk, and O. Dehzangi, "Enhancing protein fold prediction accuracy by using ensemble of different classifiers," *Australian Journal of Intelligent Information Processing Systems*, vol. 26, no. 4, pp. 32–40, 2010.

[8] S. Jahandideh, P. Abdolmaleki, M. Jahandideh, and S. H. S. Hayatshahi, "Novel hybrid method for the evaluation of parameters contributing in determination of protein structural classes," *Journal of Theoretical Biology*, vol. 244, no. 2, pp. 275–281, 2007.

[9] P. Jain and J. Hirst, "Automatic structure classification of small proteins using random forest," *BMC Bioinformatics*, vol. 11, no. 1, p. 364, 2010.

[10] Y. D. Cai, K. Feng, W. Lu, and K. Chou, "Using logitboost classifier to predict protein structural classes," *Theoretical Biollogy*, vol. 238, pp. 172–176, 2006.

[11] A. Dehzangi and S. Karamizadeh, "Solving protein fold prediction problem using fusion of heterogeneous classifiers," *INFORMATION, An International Interdisciplinary Journal*, vol. 14, no. 11, pp. 3611–3622, 2011.

[12] A. Dehzangi, S. Phon-Amnuaisuk, M. Manafi, and S. Safa, "Using rotation forest for protein fold prediction problem: An empirical study," in *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*, 2010, pp. 217–227.

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

[14] H. B. Shen and K. C. Chou, "Ensemble classifier for protein fold pattern recognition," *Bioinformatics*, vol. 22, pp. 1717–1722, 2006.

[15] T. Yang, V. Kecman, L. Cao, C. Zhang, and J. Z. Huang, "Margin-based ensemble classifier for protein fold recognition," *Expert Systems with Applications*, vol. 38, pp. 12 348–12 355, 2011.

[16] L. Nanni, "Ensemble of classifiers for protein fold recognition," *Neurocomputing*, vol. 69, no. 7-9, pp. 850–853, 2006.

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

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

[19] C. Ding and I. Dubchak, "Multi-class protein fold recognition using support vector machines and neural networks," *Bioinformatics*, vol. 17, pp. 349–358, 2001.

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

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

[22] K. Kavousi, M. Sadeghi, B. Moshiri, B. N. Araabi, and A. A. Moosavi-Movahedi, "Evidence theoretic protein fold classification based on the concept of hyperfold," *Mathematical Biosciences*, vol. 240, no. 2, pp. 148–160, 2012.

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

[24] H. Zhang, X. Hu, and Q. Li, "The recognition of 27-class protein folds: approached by increment of diversity based on multi-characteristic parameters," *Protein and Peptide Letters*, vol. 16, no. 9, pp. 1112–1119, 2009.

[25] A. Dehzangi, K. K. Paliwal, J. Lyons, A. Sharma, and A. Sattar, "Exploring potential discriminatory information embedded in pssm to enhance protein structural class prediction accuracy," in *Proceeding of the Eighth IAPR International Conference on Pattern Recognition in Bioinformatics*, ser. PRIB, 2013, pp. 208–219.

[26] Z. Zimek, F. Buchwald, E. Frank, and S. Kramer, "A study of hierarchical and flat classification of proteins," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 7, pp. 563–571, 2010.

[27] A. Dehzangi, K. K. Paliwal, J. Lyons, A. Sharma, and A. Sattar, "Enhancing protein fold prediction accuracy using evolutionary and structural features," in *Proceeding of the Eighth IAPR International Conference on Pattern Recognition in Bioinformatics*, ser. PRIB, 2013, pp. 196–207.

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

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

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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2013.2296317, IEEE/ACM Transactions on Computational Biology and Bioinformatics

[31] W. Chmielnicki and K. Stapor, "A hybrid discriminative-generative approach to protein fold recognition," *Neurocomputing*, vol. 75, no. 1, pp. 194–198, 2012.
[32] Y. Ying, K. Huang, and C. Campbell, "Enhanced protein fold recognition through a novel data integration approach," *BMC Bioinformatics*, vol. 10, no. 1, p. 267, 2009.
[33] A. Dehzangi and S. Phon-Amnuaisuk, "Fold prediction problem: The application of new physical and physicochemical- based features," *Protein and Peptide Letters*, vol. 18, no. 2, pp. 174–185, 2011.
[34] A. Sharma, K. K. Paliwal, A. Dehzangi, J. Lyons, S. Imoto, and S. Miyano, "A strategy to select suitable physicochemical attributes of amino acids for protein fold recognition," *BMC Bioinformatics*, vol. 14, no. 233, p. 11, 2013.
[35] K. Chen, W. Stach, L. Homaeian, and L. Kurgan, "ifc2: an integrated web-server for improved prediction of protein structural class, fold type, and secondary structure content," *Amino Acids*, vol. 40, pp. 963–973, 2011.
[36] M. Mizianty and L. A. Kurgan, "Modular prediction of protein structural classes from sequences of twilight-zone identity with predicting sequences," *BMC Bioinformatics*, vol. 10, no. 1, p. 414, 2009.
[37] L. A. Kurgan, K. J. Cios, and K. Chen, "Scpred: Accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences," *BMC Bioinformatics*, vol. 9, p. 226, 2008.
[38] S. Zhang, S. Ding, and T. Wang, "High-accuracy prediction of protein structural class for low-similarity sequences based on predicted secondary structure," *Biochimie*, vol. 93, no. 4, pp. 710–714, 2011.
[39] A. Sharma, J. Lyons, A. Dehzangi, and K. K. Paliwal, "A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition," *Journal of Theoretical Biology*, vol. 320, no. 0, pp. 41–46, 2013.
[40] Y. H. Taguchi and M. M. Gromiha, "Application of amino acid occurrence for discriminating different folding types of globular proteins," *BMC Bioinformatics*, vol. 8, no. 1, 2007.
[41] A. Dehzangi and A. Sattar, "Protein fold recognition using segmentation-based feature extraction model," in *Proceedings of the 5th Asian Conference on Intelligent Information and Database Systems*, ser. ACIIDS05. Springer-Verlag, 2013, pp. 345–354.
[42] L. Nanni, A. Lumini, and S. Brahnam, "An empirical study on the matrix-based protein representations and their combination with sequence-based approaches," *Amino Acid Journal*, vol. 44, no. 3, pp. 887–901, 2013.
[43] L. Nanni, S. Brahnam, and A. Lumini, "High performance set of pseaac and sequence based descriptors for protein classification," *Journal of Theoretical Biology*, vol. 266, no. 1, pp. 1–10, 2010.
[44] M. M. Gromiha, "Multiple contact network is a key determinant to protein folding rates," *Journal of Chemical Information and Modeling*, vol. 49, no. 4, pp. 1130–1135, 2009.
[45] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. H. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped blast and psi-blast: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 17, pp. 3389–3402, 1997.
[46] E. Faraggi, T. Zhang, Y. Yang, L. Kurgan, and Y. Zhou, "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*, vol. 33, no. 3, pp. 259–267, 2012.
[47] D. T. Jones, "Protein secondary structure prediction based on position-specific scoring matrices," *Journal of Molecular Biology*, vol. 292, no. 2, pp. 195–202, 1999.
[48] A. A. Adl, A. Nowzari-Dalini, B. Xue, V. Uversky, and X. Qian, "Accurate prediction of protein structural classes using functional domains and predicted secondary structure sequences," *Journal of Biomolecular Structure and Dynamics*, vol. 29, no. 6, pp. 1127–1137, 2012.
[49] J. Solding and M. Remmert, "Protein sequence comparison and fold recognition: Progress and good-practice benchmarking," *Current Opinion in Structural Biology*, vol. 21, pp. 404–411, 2008.
[50] G. Bologna and R. D. Appel, "A comparison study on protein fold recognition," in *Neural Information Processing, 2002. ICONIP '02. Proceedings of the 9th International Conference on*, 2002, pp. 2492–2496.
[51] V. N. Vapnik, *The nature of statistical learning theory*. Springer-Verlag New York, Inc, 1995.
[52] C. Lampros, C. Papaloukas, T. P. Exarchos, Y. Goletsis, and D. I. Fotiadis, "Sequence-based protein structure prediction using a reduced state-space hidden markov model," *Computers in Biology and Medicine*, vol. 37, no. 9, pp. 1211–1224, 2007.
[53] H. B. Hashemi, A. Shakery, and M. P. Naeini, "Protein fold pattern recognition using bayesian ensemble of rbf neural networks," in *Soft Computing and Pattern Recognition, 2009. SOCPAR '09. International Conference of*, 2009, pp. 436–441.
[54] C. C. Chang and C. J. Lin, "Libsvm: a library for support vector machines," 2001.

Abdollah Dehzangi received the B.Sc. degree in Computer Engineering-Hardware from Shiraz University, Iran in 2007 and Master degree, in the area of bioinformatics from Multi Media University (MMU), Cyberjaya, Malaysia, in 2011. Since 2011, He is pursuing the Ph.D. degree in Bioinformatics at Griffith University Brisbane, Australia. He is also a researcher in National ICT Australia (NICTA). His research interests include Bioinformatics, protein fold and structural class prediction problems, data mining, statistical learning theory, and pattern recognition. He is a member of IEEE.

Kuldip Paliwal received the B.S. degree from Agra University, Agra, India, in 1969, the M.S. degree from Aligarh Muslim University, Aligarh, India, in 1971 and the Ph.D. degree from Bombay University, Bombay, India, in 1978. He has been carrying out research in the area of speech processing since 1972. He has worked at a number of organizations including Tata Institute of Fundamental Research, Bombay, India, Norwegian Institute of Technology, Trondheim, Norway, University of Keele, U.K., AT & T Bell Laboratories, Murray Hill, New Jersey, U.S.A., AT & T Shannon Laboratories, Florham Park, New Jersey, U.S.A., and Advanced Telecommunication Research Laboratories, Kyoto, Japan. Since July 1993, he has been a professor at Griffith University, Brisbane, Australia, in the School of Micro electronic Engineering. His current research interests include speech recognition, speech coding, speaker recognition, speech enhancement, face recognition, image coding, bioinformatics, protein fold and structural class prediction problems, pattern recognition and artificial neural networks. He has published more than 300 papers in these research areas. Dr. Paliwal is a Fellow of Acoustical Society of India. He has served the IEEE Signal Processing Societys Neural Networks Technical Committee as a founding member from 1991 to 1995 and the Speech Processing Technical Committee from 1999 to 2003. He was an Associate Editor of the IEEE Transactions on Speech and Audio Processing during the periods 1994–1997 and 2003-2004. He also served as Associate Editor of the IEEE Signal Processing Letters from 1997 to 2000. He was the editor–in–chief of Speech Communication Journal from 2005 to 2011. He was the General Co-Chair of the Tenth IEEE Workshop on Neural Networks for Signal Processing (NNSP2000).

James Lyones received a BEng degree with Honors and a BIT from Griffith University Brisbane, Australia in 2007. He is now pursuing a PhD degree in robust automatic speech and speaker recognition at Griffith University Brisbane, Australia. His research interests include Automatic Speech and Speaker recognition, Bioinformatics, protein fold and structural class prediction problems and pattern recognition.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2013.2296317, IEEE/ACM Transactions on Computational Biology and Bioinformatics

Alok Sharma Alok Sharma received the BTech degree from the University of the South Pacific (USP), Suva, Fiji, in 2000 and the MEng degree, with an academic excellence award, and the PhD degree in the area of pattern recognition from Griffith University, Brisbane, Australia, in 2001 and 2006, respectively. He was with the University of Tokyo, Japan (2010-2012) as a research fellow. He is an A/Prof. at the USP and an Adjunct A/Prof. at the Institute for Integrated and Intelligent Systems (IIIS), Griffith University. He participated in various projects carried out in conjunction with Motorola (Sydney), Auslog Pty., Ltd. (Brisbane), CRC Micro Technology (Brisbane), the French Embassy (Suva) and JSPS (Japan). His research interests include pattern recognition, computer security, human cancer classification and protein fold and structural class prediction problems. He reviewed several articles and is in the editorial board of several journals. He is a member of IEEE.

Abdul Sattar holds a BSc (Physics, Chemistry and Mathematics) and an MSc (Physics) from the University of Rajasthan, India, in 1977, and 1979 an MPhil in Computer and Systems Sciences from the Jawaharlal Nehru University, India, and an MMath in Computer Science from the University of Waterloo, Canada, and a PhD in Computer Science (with specialization in Artificial Intelligence) from the University of Alberta, Canada, in 1990. He is the founding Director of the Institute for Integrated and Intelligent Systems and a Professor of Computer Science and Artificial Intelligence at Griffith University. He is also a Research Leader at National ICT Australia (NICTA) Queensland Research Lab (QRL), where he has held the positions of QRL Education Director (2006-08) and Leader of the Smart Applications For Emergencies (SAFE) project (2005-08), and is currently leading the QRL node of NICTA's largest project, Advanced Technologies for Optimization and Modelling In Constraints (ATOMIC). He has been an academic staff member at Griffith University since February 1992 as a lecturer (1992-95), senior lecturer (1996-99), and professor (2000-present) within the School of Information and Communication Technology. Prior to his career at Griffith University, he was a lecturer in Physics in Rajasthan, India (1980-82), and a research scholar at Jawaharlal Nehru University, India (1982-85), the University of Waterloo, Canada (1985-87), and the University of Alberta, Canada (1987-1991). His research interests include knowledge representation and reasoning, constraint satisfaction, intelligent scheduling, rational agents, propositional satisfiability, temporal reasoning, temporal databases, and bioinformatics.