Advancing the Accuracy of Protein Fold Recognition by Utilizing Profiles From Hidden Markov Models

James Lyons, Abdollah Dehzangi*, Rhys Heffernan, Yuedong Yang*, Yaoqi Zhou, Alok Sharma, and Kuldip Paliwal*

***Abstract*—Protein fold recognition is an important step towards solving protein function and tertiary structure prediction problems. Among a wide range of approaches proposed to solve this problem, pattern recognition based techniques have achieved the best results. The most effective pattern recognition-based techniques for solving this problem have been based on extracting evolutionary-based features. Most studies have relied on the *Position Specific Scoring Matrix (PSSM)* to extract these features. However it is known that profile-profile sequence alignment techniques can identify more remote homologs than sequence-profile approaches like PSIBLAST. In this study we use a profile-profile sequence alignment technique, namely HHblits, to extract HMM profiles. We will show that unlike previous studies, using the HMM profile to extract evolutionary information can significantly enhance the protein fold prediction accuracy. We develop a new pattern recognition based system called HMMFold which extracts HMM based evolutionary information and captures remote homology information better than previous studies. Using HMMFold we achieve up to 93.8% and 86.0% prediction accuracies when the sequential similarity rates are less than 40% and 25%, respectively. These results are up to 10% better than previously reported results for this task. Our results show significant enhancement especially for benchmarks with sequential similarity as low as 25% which highlights the effectiveness of HMMFold to address this problem and its superiority over previously proposed approaches found in the literature. The HMMFold is available online at: http://sparks-lab.org/pmwiki/download/index.php?Download=HMMFold.tar.bz2**

***Index Terms*—Evolutionary-based features, HMM profile, HMMFold, protein fold recognition, PSSM profile, support vector machine (SVM).**

I. Introduction

THE FUNCTION of a protein in biological interactions is dependent on its three-dimensional structure. *Protein Fold Recognition* (PFR) is an important step towards protein structure and function prediction. PFR is defined as assigning a given protein sequence to a fold that contains proteins with similar general three-dimensional configuration. It has been shown that proteins belonging to the same fold have similar function in biological interactions [1]. The number of folds is not infinite and it is estimated to be less than two thousand. Currently, according to the latest version of *Structural Classification of Protein (SCOP)* Data Bank, in total less than 1400 folds have been identified [2]. Hence, we can relate new proteins to known proteins based on the fold relationship. The general three-dimensional configuration of proteins facilitates the process of protein structure prediction for template-based techniques and also provides important information about the function of the proteins.

In the last two decades, a wide range of computational approaches have been proposed to address this problem. Among these approaches, the pattern recognition-based techniques have achieved promising results. PFR can be defined as solving a multiclass classification task where its performance relies on features as well as the classification technique being used [3]. The most important factor to enhance protein fold prediction accuracy using pattern recognition-based system is to extract highly discriminative features that effectively represent protein sequences. Early studies mainly focused on the alphabetic sequence of proteins for feature extraction [1]. They mainly extracted features based on how amino acids are distributed along the protein sequence (e.g., occurrence and composition of the amino acids, etc.). However, these features are not able to provide any information about the physicochemical or evolutionary information of the protein sequence. Later, new effective sequence based feature extraction techniques were proposed to extract more discriminatory information from the interaction of neighboring amino acids. Two of the most effective techniques are dipeptide and tripeptide features. Dipeptides and tripeptides represent the interaction of two and three neighboring amino acids along the protein sequence, respectively. These feature groups have been effectively used to extract local discriminatory information [4]–[10].

Later studies shifted their focus to extract features from the physicochemical properties of the amino acids [5], [10]–[12]. Despite providing better understanding about the interaction of the amino acids and their properties, they have not been able to enhance the protein fold prediction accuracy significantly better than using sequence-based features [13].

More recent studies have shifted their focus to evolutionary information for feature extraction and significantly enhanced the protein fold prediction accuracy compared to previous studies [14]–[16]. To extract these features, most authors have relied on *Position Specific Scoring Matrices (PSSM)* for feature extraction. The PSSM is a sequence profile that is extracted

Manuscript received March 06, 2015; revised April 20, 2015; accepted July 15, 2015. *Asterisks indicate corresponding authors.*

J. Lyons and R. Heffernan are with the School of Engineering, Griffith University, Brisbane, Australia (e-mail: j.lyons@griffith.edu.au; james.lyons@griffithuni.edu.au).

*A. Dehzangi is with the Institute for Integrated and Intelligent Systems (IIIS), Griffith University, Brisbane, Australia (e-mail: abdollah.dehzangi@griffith.edu.au).

*Y. Yang is with the Institute for Glycomics, Griffith University, Gold Coast, Australia (e-mail: yuedong.yang@griffith.edu.au).

Y. Zhou is with the Institute for Glycomics, Griffith University, Gold Coast, Australia (e-mail: yaoqi.zhou@griffith.edu.au).

A. Sharma is with the 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, Brisbane, Australia (e-mail: sharma_al@usp.ac.fj).

*K. Paliwal is with the School of Engineering, Griffith University, Brisbane, Australia (e-mail: k.paliwal@griffith.edu.au).

Digital Object Identifier 10.1109/TNB.2015.2457906

from PSIBLAST [17]. For a given protein sequence, PSI-BLAST searches a protein data bank, finds the highly similar protein sequences and extracts a profile that provides the substitution probabilities of each amino acid based on its position with other amino acids which is called the PSSM. In this way, they are able to provide important discriminatory information based on investigating highly similar and evolutionary related proteins. The PSSM has been shown to be an important source of information for feature extraction in many problems in protein science [18], [19]. In fact, to the best of our knowledge, the most promising results for the PFR have been achieved by using the PSSM for feature extraction [20]–[23]. For example, in the Yang *et al.* [21] and Paliwal *et al.* [24] that reported the best results for the PFR, PSSM was directly used to extract evolutionary information and also used indirectly to extract structural information (using predicted secondary structure from SPINE-X [25] or PSIPRED [26], respectively which in turn use PSSM for their predictions).

As explained in [17], [18], [27]–[29] studies, PSIBLAST finds highly similar protein sequences to build the PSSM sequence profile. Therefore, it is sensitive in detecting highly similar sequences. That is why using this method for problems that rely on finding very similar homologies is so effective [25]. However, it is unable to detect remote homology similarities effectively, while the PFR problem, especially when the sequential similarity rate is very low, is considered as a remote homology detection problem [30]–[32]. Therefore, using alternative techniques that are able to detect remote homologous sequences should be more effective than PSIBLAST. To investigate this hypothesis, instead of PSSM, we use another remote homology detection technique to extract evolutionary information.

In this study, we use HHblits to produce a HMM profile [28]. The HMM profile has been shown to be a more effective approach for remote homology detection compared to PSSM [18], [28], [29]. To the best of our knowledge, the evolutionary based profile extracted using a remote homology detection technique has not been explored for the PFR or similar studies (e.g., protein structural class or subcellular localization prediction problems). We develop the HMMFold method that extracts n-gram feature groups from the HMM-profile to solve the PFR. We then apply SVM to these features as the state-of-the-art classification technique for the PFR. We investigate the effectiveness of HMMFold on three of the most popular benchmarks used for the PFR namely DD, EDD, and TG benchmarks and achieve to 81%, 93%, and 86% prediction accuracies, respectively, which are up to 10% better than previously used techniques [22], [24]. We also, for the first time, achieve over 85% prediction accuracy for the PFR when the sequential similarity rate is below 25% (for the TG benchmark). Our promising results propose a new direction for extracting highly discriminatory features using evolutionary information utilizing remote homology detection techniques.

II. Materials and Methods

A. Benchmarks

We will use three popular benchmarks that have been widely used to evaluate different techniques for protein fold recognition namely, DD, EDD, and TG [21], [33]–[35]. This enables us to directly compare our results with a wide range of studies as well as the state-of-the-art techniques used for the PFR in the literature.

The DD benchmark was extracted by Ding and Dubchak in 2001 [10] from the *Structural Classification of Proteins* (SCOP) version 1.63. This data set originally consisted of 311 proteins in its training set with less than 40% sequential similarity and 383 proteins with less than 35% sequential similarity in the test set. These proteins belong to 27 folds. To conduct a more reliable experiment and to produce more statistically significant prediction results, later studies combined these two sets and used the 10-fold cross validation evaluation method. Hence, the new benchmark consists of 694 samples.

We also use the *Extended Ding and Dubchak* (EDD) dataset. We extract this benchmark from the SCOP (version 1.75) which contains more proteins [36]. This benchmark consists of 3418 proteins with less than 40% sequential similarity belonging to the same 27 folds as used in DD. This benchmark is extracted to incorporate the latest changes to SCOP and to use new proteins that have been added to this protein data bank [37]. Recent studies have used this benchmark as the main benchmark to compare their prediction results for the PFR when the sequential similarity rate is below 40%.

The third benchmark that we used is called the TG benchmark and extracted by Taguchi and Gromiha from the SCOP version 1.73 [13]. This benchmark consists of 1612 proteins with less than 25% sequential similarity belonging to 30 folds. This benchmark is used as the main benchmark to investigate the performance of protein fold recognition when the sequential similarity rate is below 25%.

In addition to these three benchmarks, we use the Lindahl benchmark to compare our proposed method with template-based threading methods [38]. This benchmark consist of 976 proteins with the sequential similarity rate of less than 40%.

B. Feature Extraction Methods

In this study, we extract monogram, bigram, and trigram feature groups. These three feature groups have been shown to be effective features for capturing local discriminatory information from the evolutionary profile using the PSSM. These three feature groups have been previously extracted from the protein sequence as well and attained good results [4], [5], [7], [8], [15]. For protein sequence, monogram features have been widely referred as occurrence of the amino acids feature group while bigram and trigram referred to as dipeptides and tripeptides. However, as it was shown in [22]–[24], [39], extracting these feature groups directly from the PSSM significantly enhanced the protein fold prediction accuracy compared to extracting these features from the protein sequence. The best results reported for protein fold recognition has been achieved by extracting evolutionary information using trigrams from the PSSM as well as using a combination of bigrams using evolutionary and structural profiles extracted directly and indirectly from the PSSM (extracting structural information using SPINE-X which also uses the PSSM for its prediction) [22]–[24].

We extract monogram, bigram, and trigram features from the HMM profile. The HMM profile is calculated by applying HHblits with its cut off value (E) is set to 0.001 on our explored

benchmark (using the latest version of Uniprot20 protein data base which was updated in 2013) in four iterations [28], [40]. Given a protein sequence, the HMM profile produces an $L \times 30$ matrix where L is the length of protein sequence. The values output by HHblits in the HMM profile are converted to linear probabilities using the formula $p = 2^{-N/1000}$ where N is the probability number from the profile. The first 20 columns represent the substitution probability of the amino acids along its sequence, based on their position, with all 20 amino acids. The next 10 columns represent the probability of three states that are defined in HHblits to represents the changes in the sequences namely, *insertion (I)*, *deletion (D)*, and *match (M)*. Comparing two sequences to find the alignment based on the HMM-profiles, insertion refers to the case that an amino acid is appended in a specific position in the sequence, deletion refers to the case that an amino acid is removed from a specific position in the sequence and match refers to the case that an exact match is detected [30], [31], [41]. The first seven rows provided the probability of changes between these tree stages (M -> M, M -> I, M -> D, I -> I, I -> M, D -> D, D -> M). The other three columns refers to the number of each stage occurring in the alignment process [28], [30].

We also produce the PSSM to extract n-gram features and compare the effectiveness of extracting these features compared to use of the HMM profile. We extract PSSM using the PSI-BLAST tool on NCBI's non redundant (nr) database with a cutoff value (E) of 0.001 in three iterations [17]. The output of PSIBLAST produced two matrices: the linear substitution probabilities and the log odds. Only the linear probabilities are used in this work. To be able to directly compare the effectiveness of the PSSM and the HMM profile for detecting remote homologies for protein fold recognition, we will extract our features from the first 20 columns of the HMM-profile. This consists of substitution probabilities of the amino acids that correspond to the 20 columns of the PSSM. We will introduce the monogram, bigram, and trigram features in the following subsections in more detail.

1) Monogram Features: The monogram feature group extracted from the HMM is a global descriptor of the proteins as it does not provide any information about the local interactions of the amino acids along the proteins [11], [13]. Monogram feature group is extracted as the global summation of the substitution probabilities of a given amino acid with all the other amino acids along the protein sequence. We calculate this feature as the summation of the substitution probabilities for the first 20 columns of the HMM-profile. To compute monogram features from the HMM profile, the following formula is used:

$$M(i) = \sum_{m=1}^{L} h_{m,i} (i = 1, \ldots, 20), \tag{1}$$

where L is the length of the protein, and $h_{m,i}$ represents the i^{th} column of the m^{th} row of the linear probabilities from the HMM-profile [23]. In this manner, we extract monogram (M) feature group consisting of 20 features in total.

2) Bigram Features: The Bigram feature group provides information about the interaction of the neighboring amino acids [39]. To extract this feature group from the HMM profile, the following formula is used:

$$B(i,j) = \sum_{m=1}^{L-1} h_{m,i} h_{m+1,j}, \tag{2}$$

where $1 \leq i \leq 20, 1 \leq j \leq 20$, B is a 20×20 matrix of features. The 20×20 matrix B is flattened into a length 400 vector which forms the final feature for a given protein:

$$[B(1,1), B(1,2), \ldots, B(1,20), B(2,1), \ldots, B(19,20), B(20,1), \ldots, B(20,20)].$$

The bigram features can be interpreted as the likelihood of two consecutive amino acids appearing along the protein sequence. Using the evolutionary profile to extract bigram features has been shown to be an effective technique to reduce the number of redundant features and at the same time to extract important local evolutionary information for protein fold recognition [22], [24], [39].

3) Trigram Features: Despite increasing the number of features dramatically compared to the number of features extracted using the bigram technique (400 bigrams compared to 8000 trigram features), the trigrams is still an effective feature group for protein fold recognition [23]. Similar to the bigram, to extract the trigram feature group we consider the substitution probabilities of consecutive amino acids. However, instead of considering two neighboring amino acids as in the bigram, for the trigrams we calculate the expected occurrence of three consecutive amino acids. To compute the trigram features from the HMM profile, the following formula is used:

$$T(i,j,k) = \sum_{m=1}^{L-2} h_{m,i} h_{m+1,j} h_{m+2,k}, \tag{3}$$

where $1 \leq i \leq 20, 1 \leq j \leq 20, 1 \leq k \leq 20$, T is a $20 \times 20 \times 20$ block of features. The $20 \times 20 \times 20$ matrix T is flattened into a length 8000 vector which forms the final feature for a given protein:

$$[T(1,1,1), T(1,1,2), \ldots, T(1,1,20), T(1,2,1), \ldots, T(20,20,20)].$$

As mentioned earlier, the trigram feature group has been used successfully to represent local information from the protein profile (PSSM) and attained promising results for protein fold recognition [23]. In fact, despite their simplicity, all three n-gram techniques that are introduced here (monogram, bigram, and trigram techniques) have been successfully used to extract global and local discriminatory information from the protein profile (PSSM) and are considered state-of-the-art feature extraction techniques for protein fold recognition [7], [13], [15], [22], [24], [39], [42].

Note that moving to higher n-grams is not practical as they produce very large number of features (e.g., quadgrams produce 160 000 features for each protein) which is not feasible for large protein data banks or when the number of samples are limited

[11], [43]. As a result we limit ourselves to extracting trigram features.

C. Support Vector Machines

During the past two decades, a wide range of classification techniques, such as *artificial neural networks (ANN)* [4], [42], *Meta Classifiers* [44], [45], *K-nearest neighbor (KNN)* [8], [46], *support vector machine (SVM)* [20], [21], and *Ensemble Classifiers* [43], [47] have been implemented and used for protein fold recognition. However, the best results for this task is achieved by using SVM classifiers [21]–[24].

SVM is considered a state-of-the-art machine learning and pattern classification algorithm [48]. It has been extensively applied in classification and regression tasks. SVM aims to find a maximum margin hyperplane to minimize classification error. A function called the kernel $(K())$ is used to project the data from input space to a new feature space. By using a nonlinear projection, it allows for nonlinear decision boundaries [49].

The optimization problem solved by SVM is to find values of α_i that maximize L_D:

$$L_D = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j K(\mathbf{x_i}, \mathbf{x_j}) \tag{4}$$

subject to the conditions:

$$0 \leq \alpha_i \leq C, \qquad \text{for all } i$$
$$\text{and}$$
$$\sum_i \alpha_i y_i = 0.$$

To classify a new point $\mathrm{x}\prime$ with class $y\prime \in \{-1, 1\}$, the following equation is used:

$$y\prime = \text{sign}\left(\sum_{i=1}^{n} \alpha_i y_i K(\mathrm{x_i}, \mathrm{x}\prime) + b \right) \tag{5}$$

where $K()$ is the kernel function. Common kernel functions include linear, polynomial, and *radial basis function (RBF)*. In this paper, we use the RBF kernel defined as: $K(\mathrm{x_i}, \mathrm{x}\prime) = \exp\left(-\gamma(\mathrm{x_i} - \mathrm{x}\prime)^T(\mathrm{x_i} - \mathrm{x}\prime)\right)$. where γ is a tunable parameter.

We apply SVM to the DD, EDD and TG datasets as described in Section III-A. In all of our experiments we use the SVM classifier with the RBF kernel. We use the implementation of SVM from the libsvm library [50]. To provide general results and also to avoid any over training, we have used the same values for C and γ parameters as they have used in [23]. They found these parameters ($\gamma = 0.0038$ and $C = 1000$) using grid search which is also implemented in the libsvm package.

As the evaluation criterion, we use 10-fold cross validation to be able to directly compare our results with previous studies as it has been widely used in the literature [21], [24]. In 10-fold cross validation technique, the data set is randomly divided into 10 mutually exclusive subsets. Then in 10 different experiments, 9 subsets are combined and used as training set and the remaining subset is used as the test set. This process repeats until all 10 subsets are used exactly once as a testing subset. Then the results of all the samples are averaged to produce the final prediction performance. We conduct our experiments 10 times and report the average prediction accuracy (using 10 times 10-fold cross validation and averaging the results).

TABLE I
THE RESULTS FOR MONOGRAM, BIGRAM, AND TRIGRAM FEATURES EXTRACTED FROM THE ORIGINAL PROTEIN SEQUENCE AS WELL AS THE PSSM AND HMM PROFILES. WE PRODUCED THE RESULTS WHEN USING THE 20 COLUMNS OF THE HMM CORRESPONDING TO THE 20 COLUMNS IN PSSM AND WHEN USING ALL 30 COLUMNS OF THE HMM PROFILE (THE AVERAGE PREDICTION ACCURACY OF USING 10 TIMES 10-FOLD CROSS VALIDATION)

Feature Group	DD	TG	EDD
Sequence-Monogram	51.0	36.2	46.9
Sequence-Bigram	53.4	43.1	55.6
Sequence-Trigram	38.4	34.4	45.7
PSSM-Monogram	69.6	50.1	69.6
HMM-Monogram (20)	76.2	75.2	86.7
HMM-Monogram (30)	56.1	56.9	77.6
PSSM-Bigram	74.1	68.1	84.5
HMM-Bigram (20)	79.4	83.1	92.6
HMM-Bigram (30)	67.6	65.3	83.4
PSSM-Trigram	74.6	73.4	88.7
HMM-Trigram (20)	81.8	86.0	93.8
HMM-Trigram (30)	68.5	65.8	84.3

III. RESULTS AND DISCUSSION

A. Results

In the first step, we extract monogram, bigram, and trigram features from the original protein sequence, from the PSSM, and also from the HMM profiles. We extract features from the HMM using its first 20 columns corresponding to the 20 columns of the PSSM and all its 30 columns. The results achieved by applying an SVM classifier to these features for the DD, EDD, and TG benchmarks are shown in Table I. As shown in this table, the best results are achieved by applying SVM to the monogram, bigram, and trigram feature groups extracted from the HMM profile using its first 20 columns for feature extraction. Note that we have used HMM profile with all 30 columns but our results was not as good as using just the first 20 columns of the HMM profile corresponding to the 20 column of the PSSM. When we tuned the parameters for the SVM for HMM (30), our results was just slightly worst than using HMM (20). However, when we used the same parameters extracted for HMM (20), the results using HMM (30) was much worse. The main reason can be because the extra 10 columns of the HMM profile do not provide any additional discriminatory information to its first 20 columns for the protein fold recognition problem. While adding the next 10 columns, we increase the number of features dramatically (using 20 columns of HMM, we extract $30\times30\times30\times = 27\,000$ features in Trigram feature group instead of 8000 using first 20 columns). Therefore, we proceed our experiments using the first 20 columns of the HMM profile for feature extraction. For the rest of this study, we extract our features from the first 20 columns of the HMM profile. Note that to maintain the consistency of our experiments, we just report the results using HMM (30) with the same parameters for the SVM classifier (γ and C) used for HMM (20) in Table I.

TABLE II
THE RESULTS (%) FOR NAIVE BAYES, BAYES NETWORK, KNN, SVM (WITH LINEAR KERNEL), AND RF (USING 10, 50, AND 100 BASE LEARNERS) FOR TRIGRAM FEATURE VECTOR EXTRACTED FROM THE PSSM, AND HMM PROFILES FOR DD, EDD, AND TG BENCHMARKS

	DD Benchmark		EDD Benchmark		TG Benchmark	
Classifier	PSSM	HMM	PSSM	HMM	PSSM	HMM
Naive Bayes	64.2	70.4	66.4	69.8	47.0	59.0
Bayes Network	46.3	55.0	47.0	67.0	34.2	53.2
KNN	61.3	75.1	79.2	88.9	53.7	75.8
SVM	73.1	80.6	88.4	93.6	73.5	84.5
RF (10)	52.1	69.9	55.0	82.8	37.2	66.9
RF (50)	62.6	76.3	66.0	87.6	44.7	75.0
RF (100)	63.3	76.7	67.2	88.5	45.9	75.7

TABLE III
THE RESULTS ACHIEVED FOR EACH FOLD FOR THE DD BENCHMARK USING SVM ON THE TRIGRAM FEATURE GROUP EXTRACTED FROM THE PSSM AND HMM PROFILE, RESPECTIVELY

No.	Fold	No. of samples	PSSM	HMM
	α			
1	Globin-like	19	100.0	100.0
2	Cytochrome C	16	100.0	100.0
3	DNA-binding 3-helical bundle	32	84.4	90.6
4	4-Helical up-and-down bundle	15	80.0	86.7
5	4-Helical cytokines	18	88.9	94.4
6	α EF-hand	15	73.3	86.7
	β			
7	Immunoglobulin-like β-sandwich	74	90.5	90.5
8	Cupredoxins	21	90.5	90.5
9	Viral coat and capsid proteins	29	82.8	86.2
10	ConA-like lectins/glucanases	13	61.5	76.9
11	SH3-like barrel	16	56.3	62.5
12	OB-fold	32	50.0	65.6
13	Trefoil	12	100.0	100.0
14	Trypsin-like serine proteases	13	61.5	61.5
15	Lipocalins	16	93.8	93.8
	α/β			
16	(TIM)-barrel	77	80.5	76.6
17	FAD (also NAD)-binding motif	23	78.3	91.3
18	Flavodoxin-like	24	50.0	62.5
19	NAD (P)-binding Rossmann-fold	40	62.5	87.5
20	P-loop containing nucleotide	22	54.6	63.6
21	Thioredoxin-like	17	82.4	94.1
22	Ribonuclease H-like motif	22	45.5	77.3
23	Hydrolases	18	77.8	88.9
24	Periplasmic binding protein-like	15	40.0	86.7
	α+β			
25	β-Grasp	15	40.0	53.3
26	Ferredoxin-like	40	60.0	55.0
27	Small inhibitors, toxins, lectins	40	95.1	97.6

We achieve 92.6%, and 93.8% prediction accuracies using bigram and trigram feature groups extracted from the HMM profile for the EDD benchmark which are 8.1% and 5.1% better than extracting these feature groups from the PSSM profiles, respectively [23], [39]. By achieving 93.8% prediction accuracy for this benchmark, we enhance the protein fold prediction ac-

TABLE IV
THE RESULTS ACHIEVED FOR EACH FOLD FOR THE EDD BENCHMARK USING SVM ON THE TRIGRAM FEATURE GROUP EXTRACTED FROM THE PSSM AND HMM PROFILE, RESPECTIVELY

No.	Fold	No. of samples	PSSM	HMM
	α			
1	Globin-like	41	92.7	97.6
2	Cytochrome C	35	91.4	97.1
3	DNA-binding 3-helical bundle	322	94.7	97.8
4	4-Helical up-and-down bundle	69	76.8	88.4
5	4-Helical cytokines	30	83.3	90.0
6	α EF-hand	59	84.8	88.1
	β			
7	Immunoglobulin-like β-sandwich	391	94.4	96.7
8	Cupredoxins	47	89.4	91.5
9	Viral coat and capsid proteins	60	71.7	83.3
10	ConA-like lectins/glucanases	57	87.7	87.7
11	SH3-like barrel	129	69.8	87.6
12	OB-fold	156	77.6	85.3
13	Trefoil	45	88.9	95.6
14	Trypsin-like serine proteases	45	86.7	91.1
15	Lipocalins	37	78.4	89.2
	α/β			
16	(TIM)-barrel	336	97.6	98.5
17	FAD (also NAD)-binding motif	73	90.4	95.9
18	Flavodoxin-like	130	80.0	92.3
19	NAD (P)-binding Rossmann-fold	195	95.4	98.5
20	P-loop containing nucleotide	239	93.3	96.7
21	Thioredoxin-like	111	85.6	91.9
22	Ribonuclease H-like motif	128	72.7	91.4
23	Hydrolases	83	98.8	100.0
24	Periplasmic binding protein-like	16	81.3	100.0
	α+β			
25	β-Grasp	121	81.0	87.6
26	Ferredoxin-like	339	84.4	91.7
27	Small inhibitors, toxins, lectins	124	100.0	100.0

curacy by 3.2% when compared to the best results reported for this benchmark [20]–[22], [24].

Our results for the DD and TG benchmarks are even more promising. Using bigram and trigram feature groups extracted from the HMM profile, we achieve 83.1% and 86.0% prediction accuracies for the DD and TG benchmarks, respectively [21], [23], [24]. To the best of our knowledge, it is the first time that the protein fold prediction accuracy has reached over 80% prediction accuracy for a benchmark with less than 25% sequential similarity. Achieving 86.0% prediction accuracy for the PFR when the sequential similarity is less than 25% is a great breakthrough in solving this problem. This highlights the effectiveness of using the HMM profile for feature extraction.

We also achieve 81.8% prediction accuracy for the DD benchmark using the trigram feature group which is 7.2% better than using trigram feature group extracted from the PSSM. It is also the first time that we can achieve to over 80% prediction accuracy (using 10-fold cross validation) for the DD benchmark since the introduction of this benchmark in 2001 by Ding and Dubchak [10].

To investigate the generality of the achieved enhancement with respect to the classifier being used, we have used several of the most popular classification techniques that have been widely used in the literature. We use these classifiers for HMM-Trigram and PSSM-Trigram feature groups and compared the results [43], [51]–[54]. We used *naive Bayes*, *Bayes network*, *K-nearest neighbor (KNN)*, SVM (using linear kernel), and *Random Forest (RF)* using three different numbers of base learners (10, 50, and 100). We used the implementation of these classifiers in WEKA machine learning toolbox [55]. Except for RF that is used with three different number of base learners, the default parameters of the WEKA are used for other classifiers. We used RF with three

TABLE V
THE RESULTS ACHIEVED FOR EACH FOLD FOR THE TG BENCHMARK USING SVM ON THE TRIGRAM FEATURE GROUP EXTRACTED FROM THE PSSM AND HMM PROFILE, RESPECTIVELY

No.	Fold	No. of samples	PSSM	HMM
	α			
1	Cytochrome C	25	84.0	96.0
2	DNA/RNA binding 3-helical bundle	103	86.4	92.2
3	Four helical up and down bundle	26	46.2	69.2
4	EF hand-like fold	25	48.0	76.0
5	SAM domain-like	26	42.3	73.1
6	α-α super helix	47	74.4	78.7
	β			
7	Immunoglobulin-like β-sandwich	173	89.0	89.0
8	Common fold of diphtheria toxin/transcription factors/cytochrome	28	28.6	53.6
9	Cupredoxin-like	30	86.7	93.3
10	Galactose-binding domain-like	25	52.0	72.0
11	Concanavalin A-like lectins/glucanases	26	84.6	84.6
12	SH3-like barrel	42	47.6	76.2
13	OB-fold	78	51.3	71.8
14	Double-stranded α-helix	34	76.5	88.2
15	Nucleoplasmin-like	42	59.5	61.9
	α/β			
16	TIM α/β-barrel	145	95.2	97.9
17	NAD(P)-binding Rossmann-fold domains	77	80.5	97.4
18	FAD/NAD(P)-binding domain	31	87.1	96.8
19	Flavodoxin-like	55	56.4	81.8
20	Adenine nucleotide a hydrolase-like	34	38.2	94.1
21	P-loop containing nucleoside triphosphate hydrolases	95	84.2	95.8
22	Thioredoxin fold	32	50.0	87.5
23	Ribonuclease H-like motif	49	32.7	81.6
24	S-adenosyl-L-methionine-dependent methyltransferases	34	82.4	94.1
25	α/β-Hydrolases	37	91.9	100.0
	α+β			
26	β-Grasp, ubiquitin-like	42	47.6	73.8
27	Cystatin-like	25	48.0	88.0
28	Ferredoxin-like	118	69.5	79.7
29	Knottins	80	100.0	98.8
30	Rubredoxin-like	28	57.1	82.1

different base learners to investigate the impact of the number of base learners on its prediction performance. Note that we used SVM using linear kernel to investigate the impact of using different kernel and compare its results using SVM with RBF kernel. Our achieved results are shown in Table II.

As it is shown in Table II, for all our employed classifiers and all three benchmarks (DD, EDD, and TG benchmark), using Trigram feature vector extracted from HMM profile achieved better results than extracting this feature vector from PSSM profile. It highlights the effectiveness of using HMM profile for feature extraction for PFR compared to PSSM profile regardless of the classification technique being used. As it is shown in this table, for the Random Forest classifier, increasing the number of base learners from 10 to 50 significantly increases the prediction performance. While the increase in prediction performance is not as significant when we increase the number of base learners from 50 to 100. It shows that increasing the number of base learners for the Random Forest further than 100, we can not ex-

pect significant prediction enhancement as it was discussed and supported in Dehzangi *et al.* [44]. The best results among the employed classifier is achieved by using SVM classifier using linear kernel. It shows the superiority of SVM classifier with respect to the features being used compared to the other employed classifiers. Despite promising results achieved using SVM with linear kernel, they are still lower than the results achieved by using SVM with RBF kernel which shows the preference of RBF kernel for this task.

We then produce results for each individual fold for the DD, EDD, and TG benchmarks using SVM to the trigram feature group extracted from these two profiles. It is done to provide more information about the impact of using the HMM profile compared to the use of PSSM to extract features. The results are shown in Tables III–V. In these tables, the name of the folds represented in each benchmark is shown in column two, the number of samples in each fold is represented in column three, and the results achieved for the trigram feature group extracted from the PSSM and HMM profiles are shown in columns four and five, respectively.

As it is shown in Table III, our results achieved using the HMM profile to extract trigram for most of the folds (except fold number 26 "Ferredoxin-like") is equal to or better than using the PSSM for feature extraction. The same pattern is repeated in Table IV with similar folds which highlight the consistency of our achieved results with respect to the targeted folds. As it is shown in Table V, our achieved enhancement with respect to the folds is repeated here again and even more significant for the TG benchmark. The results achieved using the HMM profile compared to the PSSM profile to extract trigram feature group shows the preference of HMM over PSSM for feature extraction for most of the folds (except fold number 29 "Knottins"). It is also shown that using the HMM profile we achieve over 50% prediction accuracy for all the folds investigated in the TG benchmark.

B. Comparison With Existing Methods

To be able to directly compare our results with the best results reported in the literature, we have reproduced those results for all the three benchmarks that we have used. This comparison is provided in Table VI. Note that to be able to directly compare our results with the results found in the literature, for DD benchmark, we divide this benchmark to train (311 samples) and test (383) samples, and conducted our experiments. For the EDD, and TG benchmark, we proceed with reporting our results using the average result of 10 times 10-fold cross validation experiments and reproduced previous results found in the literature for those experiments. The features tested include the following: PF1, PF2 (sequence based bigram features [42]), PF [8], O (Occurrence [13]), AAC (amino acid composition [10]), AAC + HXPVZ (composition plus physicochemical features from 5 attributes [10]). The previous 6 features are computed from the original protein sequence, the same features are also computed from the PSSM-based consensus sequence, these are prepended with CONS- before the feature abbreviation. Finally, ACCfold (Auto and cross covariance [20]), Bigram [39], PSSM-SPINE-S [3], Trigrams [23], DTW [22] and k-AAP [24] are used.

As it is shown in this table and was highlighted in the previous subsection, we significantly outperform previously reported results for the PFR. Using trigram features, we achieve 86.0%, 75.8%, and 93.8% prediction accuracies which are 11.0%, 4.3%, and 3.2% better than previously reported results for the TG, DD, and EDD benchmarks, respectively [3], [15], [21], [22], [24], [35].

TABLE VI
PFR ACCURACY FOR FEATURES ON THE DD, TG AND EDD DATASETS. ACCURACIES ARE COMPUTED FOR THE INDEPENDENT TEST SET FOR THE DD BENCHMARK AND 10×10-FOLD CROSS-VALIDATION FOR THE EDD, AND TG BENCHMARKS

References	Features	DD	TG	EDD
[42]	PF1	50.6	38.8	50.8
[42]	PF2	48.2	38.8	49.9
[42]	PF	53.4	43.1	55.6
[13]	O	51.0	36.2	46.9
[10]	AAC	45.1	32.0	40.9
[10]	AAC+HXPZV	47.2	36.3	40.9
[39]	CONS-PF1	64.6	52.7	75.2
[39]	CONS-PF2	64.7	51.1	74.9
[39]	CONS-PF	67.5	58.8	79.3
[39]	CONS-O	63.5	46.7	68.5
[39]	CONS-AAC	59.2	44.0	61.9
[3]	PSSM-SPINE-S	-	73.8	88.2
[20]	ACCfold (k=10)	70.1	66.4	85.9
[22]	DTW	-	74.0	90.2
[24]	k-AAP	-	77.0	90.6
[21]	TAXFOLD	71.5	-	90.0
[39]	PSSM-Monogram	-	50.1	69.6
[39]	PSSM-Bigram	-	68.1	84.5
[23]	PSSM-Trigram	-	73.4	88.7
This study	HMM-Monogram	66.2	75.2	86.7
This study	HMM-Bigram	74.2	83.1	92.6
This study	HMM-Trigram	**75.8**	**86.0**	**93.8**

We also compare HMMFold with the conventional template-based methods [56]–[62]. To do this, we use Lindahl benchmark and adapt 2-fold cross validation evaluation criterion as used in [21] and [62]. Similar to [21], we preprocessed in a way that each fold contains at least $N_{\min}$ number of samples. We conduct our experiments when $N_{\min} = 1$ and $N_{\min} = 5$ and compare our results with previous results found in the literature. We also make sure that samples from the same superfamily were placed in the same group for cross validation. Therefore, similar to [21], the training and testing samples came from different superfamilies. Note that we did not tune the SVM parameters for this benchmark. We used the same SVM parameters that we used for the other benchmarks. In addition to the template-based techniques, we compare our results with TAXFOLD [21] and ACCfold [20] (two taxonomy based techniques) that have been previously used for this benchmark and outperformed other template-based methods. The results achieved for this experiment are shown in Table VII. As it is shown in the table, we achieve 14.9% and 5.3% when $N_{\min} = 5$ and 2.9% and 0.3% when $N_{\min} = 1$ improvements over BoostThreader and SPARKS-X as the best results reported in the literature for this benchmark using a template-based technique. We also achieve 4.9% and

TABLE VII
THE RESULTS (%) ACHIEVED BY USING HMMFOLD COMPARED TO OTHER TEMPLATE-BASED METHODS ON LINDAHL BENCHMARK (USING 2-FOLD CROSS VALIDATION AND A SETUP AS IT WAS USED IN [21]. NOTE THAT THE RESULTS FOR THE PREVIOUS STUDIES ARE TAKEN FROM [21] AND [62]

	Family			Superfamily		Fold	
Methods	$N_{min}=1$	$N_{min}=3$	$N_{min}=4$	$N_{min}=1$	$N_{min}=5$	$N_{min}=1$	$N_{min}=5$
No. of Sequences	555	97	47	434	225	321	177
No. of Categories	176	13	5	86	23	38	8
SPARKS-X	84.1	-	-	59.0	76.3	45.2	67.0
BoostThreader	86.5	-	-	66.1	76.4	42.6	57.4
DescFold-II	81.1	-	-	60.6	-	32.4	-
DescFold-I	80.7	-	-	57.8	-	24.9	-
FOLDpro	85.0	-	-	55.5	70.0	26.5	48.3
SP5	82.4	-	-	59.8	-	37.9	58.7
SPARKS	81.6	-	-	52.5	69.1	24.3	47.7
HHpred	82.9	-	-	58.8	-	25.2	-
Fugue	82.2	-	-	41.9	-	12.2	-
RAPTOR	86.6	-	-	56.3	-	38.2	-
ACCFold	53.9	79.6	95.7	23.1	78.3	29.9	51.9
TAXFOLD	68.6	90.7	100.0	39.3	84.5	40.6	67.7
HMMFold (this study)	74.0	80.5	80.9	43.6	77.5	**45.5**	**72.3**

4.6% better results than TAXFOLD as the best taxonomy technique used for this benchmark when $N_{\min}=1$ and $N_{\min}=5$, respectively. These results show the preference of HMMFold compared to other template and taxonomy based techniques for PFR.

We also conduct the paired T-test to study the statistical significance of our achieved enhancement compared to previous results (on DD, EDD, and TG benchmarks). The achieved P-value for the paired T-test ($\mathrm{P}=0.0002$) supports the statistical significance of our reported improvements over the previously reported results found in the literature for protein fold recognition.

Similar to [20] and [21], we have conducted our experiments to predict proteins in family and superfamily level as well. As it is shown in Table VII, our results are lower than other temple based methods. The main reason is that for these two problems, there are just a few samples which prevent our method to train properly for these tasks. Since we extract 8000 features using Trigram feature group, there is a need to have more sample to be able to properly train our method. Furthermore, we optimized our parameters for the fold recognition task and used the same parameters for family and superfamily prediction as fold recognition is the main focus of this problem.

C. Discussion

Our results show that features based on the HMM profiles significantly outperform the same features computed from the PSSM. We claim that even from the beginning, the HMM profiles should have been used to extract evolutionary information for the protein fold recognition problem instead of using the PSSM profiles. As it was explained in the introduction, the HMM profiles are designed to extract remote homology information while the PSSM profile is more sensitive in finding highly significant alignments [18], [28], [32]. In Yan *et al.* [18], it was shown that PSIBLAST has lower *false positive rate* (FPR) and *false negative rate* (FNR) for finding highly similar samples than HHsearch which shows the sensitivity of its sequence alignment compared to the HMM profile. HHsearch shows better performance in finding similar alignments as it identifies the remote homology relations well. Because protein fold is defined based on the relationship between proteins with similar general configurations and secondary structure shapes, finding remote homology can provide more discriminatory information [18], [30], [32].

Since the introduction of dynamic evolutionary information, using the PSSM profile to extract this information have been successfully used for the PFR and its similar problems such as protein structural class prediction, protein function prediction, etc. [19], [63]–[67]. Our results show the superiority of the HMM profile for protein fold recognition and support the findings of [27]–[29], [31], [32] studies. We believe that using the HMM profile is potentially able to enhance the prediction performance for similar problems as well. In fact, our observations in this study as well as findings in Solding *et al.* [19] support the idea that extracting evolutionary information from a profile that is better at detecting remote homology information can enhance our ability to solve many problems in protein science. Therefore, HMMFold as a HMM profile based technique inherits this advantage which is its ability to use the remote homology information for PFR. Note that HMMFold can be improved by proposing new techniques to reduce the number of features to speed up the classification task and be more suitable for large protein data banks.

We hypothesize that not just for protein fold recognition, but also for similar problems (e.g., protein structural class and protein subcellular localization prediction problems) using HMM profiles could significantly enhance the prediction performance and is capable of revealing more discriminatory information [28]. Extracting this information and finding remote homology

relationships more accurately is also made possible the increases in the number of sequences deposited in the protein data banks such as NCBI and UniProt.

IV. Conclusion

In this study we have proposed the HMMFold technique. This technique is based on extracting evolutionary information from HMM profiles. We used monogram, bigram, and trigram features which are extracted from the HMM profile. We then applied SVM to classify these features. We have shown that by using SVM to the trigram feature group we are able to significantly enhance protein fold prediction accuracy compared to previous results found in the literature. We achieved 86.0%, 75.8%, and 93.8% prediction accuracies which are 11.0%, 4.3%, and 3.2% better than the previously reported results for the TG, DD, and EDD benchmarks, respectively.

We have, for the first time, achieved over 80% prediction accuracy (86%) for protein fold recognition when the sequential similarity rate is less than 25% which is 11% better than previously reported results. This breakthrough can be considered as a promising achievement in addressing protein fold recognition. Our results also highlight the effectiveness of the HMM profiles compared to the PSSM profiles for protein fold recognition.

Considering our results as well as the findings in [28], [29], and [18], we showed that using the HMM profile is more effective than the PSSM profile for the protein fold recognition problem and should be considered the main resource to extract sequence profiles and evolutionary information for this task.

For our future works, we aim to investigate the impact of using the HMM profiles for other similar studies that depend on detecting remote relationships between protein structures such as protein structural classes, protein subcellular localization, protein function prediction, protein domain prediction, and investigate the performance of the HMMFold for these problems. For public use, HMMFold is freely available at: http://sparks-lab.org/pmwiki/download/index.php?Download=HMMFold.tar.bz2 We have also provided options for users to extract n-gram features and download the source code for the HMMFold from this web page.

References

[1] C. Chothia and A. V. Finkelstein, "The classification and origins of protein folding patterns," *Annu. Rev. Biochem.*, vol. 59, no. 1, pp. 1007–1035, 1990.

[2] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, "Scop: A structural classification of proteins database for the investigation of sequences and structures," *J. Mol. Biol.*, vol. 247, no. 4, pp. 536–540, 1995.

[3] A. Dehzangi, J. Lyons, A. Sharma, K. K. Paliwal, and A. Sattar, "A segmentation-based method to extract structural and evolutionary features for protein fold recognition," *IEEE/ACM Trans. Comput. Biol. Bioinformat.*, pp. 1–11, 2014.

[4] 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," *IEEE Trans. NanoBiosci.*, vol. 6, no. 2, pp. 186–196, 2007.

[5] N. R. Pal and D. Chakraborty, "Some new features for protein fold prediction," in *Proc. 2003 Joint Int. Conf. Artif. Neural Netw. Neural Inf. Process., Ser. ICANN/ICONIP'03*, 2003, pp. 1176–1183.

[6] L. Nanni, S. Brahnam, and A. Lumini, "High performance set of pseaac and sequence based descriptors for protein classification," *J. Theoretical Biol.*, vol. 266, no. 1, pp. 1–10, 2010.

[7] 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 J.*, vol. 44, no. 3, pp. 887–901, 2013.

[8] T. Yang, V. Kecman, L. Cao, C. Zhang, and J. Z. Huang, "Margin-based Ensemble classifier for protein fold recognition," *Expert Syst. With Appl.*, vol. 38, pp. 12348–12355, 2011.

[9] I. Dubchak, I. B. Muchnik, and S. H. Kim, "Protein folding class predictor for SCOP: Approach based on global descriptors," in *Proc. 5th Int. Conf. Intell. Syst. Mol. Biol. (ISMB)*, 1997, pp. 104–107.

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

[11] A. Dehzangi and S. Phon-Amnuaisuk, "Fold prediction problem: The application of new physical and physicochemical- based features," *Protein Peptide Lett.*, vol. 18, no. 2, pp. 174–185, 2011.

[12] 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 Bioinformat.*, vol. 14, no. 233, p. 11, 2013.

[13] Y. H. Taguchi and M. M. Gromiha, "Application of amino acid occurrence for discriminating different folding types of globular proteins," *BMC Bioinformat.*, vol. 8, no. 1, p. 404, 2007.

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

[15] 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.

[16] 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.

[17] 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 Res.*, vol. 17, pp. 3389–3402, 1997.

[18] R. Yan, D. Xu, J. Yang, S. Walker, and Y. Zhang, "A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction," *Sci. Rep.*, vol. 3, 2013.

[19] J. Solding and M. Remmert, "Protein sequence comparison and fold recognition: Progress and good-practice benchmarking," *Curr. Opin. Struct. Biol.*, vol. 21, pp. 404–411, 2008.

[20] 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.

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

[22] J. Lyons, N. Biswas, A. Sharma, A. Dehzangi, and K. K. Paliwal, "Protein fold recognition by alignment of amino acid residues using kernelized dynamic time warping," *J. Theoretical Biol.*, vol. 354, no. 7, pp. 137–145, 2014.

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

[24] K. K. Paliwal, A. Sharma, J. Lyons, and A. Dehzangi, "Improving protein fold recognition using the amalgamation of evolutionary-based and structural-based information," *BMC Bioinformat.*, vol. 15, no. Suppl 16, p. S12, 2014.

[25] 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," *J. Comput. Chem.*, vol. 33, no. 3, pp. 259–267, 2012.

[26] D. T. Jones, "Protein secondary structure prediction based on position-specific Scoring matrices," *J. Mol. Biol.*, vol. 292, no. 2, pp. 195–202, 1999.

[27] R. C. Edgar and K. Sjolander, "A comparison of scoring functions for protein sequence profile alignment," *Bioinformatics*, vol. 20, no. 8, pp. 1301–1308, 2004.

[28] M. Remmert, A. Biegert, A. Hauser, and J. J. Soding, "Hhblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment," *Nature Methods*, vol. 9, no. 2, pp. 173–175, 2012.

[29] D. B. Kuchibhatla, W. A. Sherman, B. Y. W. Chung, S. Cook, G. Schneider, B. Eisenhaber, and D. G. Karlin, "Powerful sequence similarity search methods and in-depth manual analyses can identify remote homologs in many apparently orphan viral proteins," *J. Virol.*, vol. 88, no. 1, pp. 10–20, 2014.
[30] J. Soding, "Protein homology detection by HMM-HMM comparison," *Bioinformatics*, vol. 21, no. 7, pp. 951–960, 2005.
[31] J. Soding, M. Remmert, A. Biegert, and A. N. Lupas, "Hhsenser: exhaustive transitive profile search using HMM-HMM comparison," *Nucleic Acids Res.*, vol. 34, no. Suppl 2, pp. W374–W378, 2006.
[32] Q. Xu and R. L. Dunbrack, "Assignment of protein sequences to existing domain and family classification systems: Pfam and the pdb," *Bioinformatics*, vol. 28, no. 21, pp. 2763–2772, 2012.
[33] A. Dehzangi, K. K. Paliwal, J. Lyons, A. Sharma, and A. Sattar, "Enhancing protein fold prediction accuracy using evolutionary and structural features," in *Proc. 8th IAPR Int. Conf. Pattern Recog. Bioinformat., Ser. PRIB*, 2013, pp. 196–207.
[34] 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," *Comput. Biol. Chem.*, vol. 35, no. 1, pp. 1–9, 2011.
[35] 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," *Math. Biosci.*, vol. 240, no. 2, pp. 148–160, 2012.
[36] N. K. Fox, S. E. Brenner, and J. M. Chandonia, "Scope: Structural classification of proteins Extended, integrating SCOP and astral data and classification of new structures," *Nucleic Acids Res.*, vol. 42, no. D1, pp. D304–D309, 2014.
[37] A. Dehzangi and A. Sattar, "Ensemble of diversely trained support vector machines for protein fold recognition," in *Proc. 5th Asian Conf. Intell. Inf. Database Syst., Ser. ACIIDS05*, 2013, pp. 335–344.
[38] E. Lindahl and A. Elofsson, "Identification of related proteins on family, superfamily and fold level," *J. Mol. Biol.*, vol. 295, no. 3, pp. 613–625, 2000.
[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," *J. Theoretical Biol.*, vol. 320, pp. 41–46, 2013.
[40] A. Bairoch, R. Apweiler, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, A. D. Natale, C. O. N. Redaschi, and L. S. L. Yeh, "The universal protein resource (Uniprot)," *Nucleic Acids Res.*, vol. 33, no. Suppl 1, pp. D154–D159, 2005.
[41] J. Soding, M. Remmert, and B. Biegert, "Hhrep: de novo protein repeat detection and the origin of tim barrels," *Nucleic Acids Res.*, vol. 34, no. Suppl 2, pp. W137–W142, 2006.
[42] P. Ghanty and N. R. Pal, "Prediction of protein folds: Extraction of new features, dimensionality reduction, and fusion of heterogeneous classifiers," *IEEE Trans. NanoBiosci.*, vol. 8, no. 1, pp. 100–110, 2009.
[43] A. Dehzangi and S. Karamizadeh, "Solving protein fold prediction problem using fusion of heterogeneous classifiers," *Inf., Int. Interdisciplinary J.*, vol. 14, no. 11, pp. 3611–3622, 2011.
[44] A. Dehzangi, S. Phon-Amnuaisuk, and O. Dehzangi, "Using Random Forest for protein fold prediction problem: An empirical study," *J. Inf. Sci. Eng.*, vol. 26, no. 6, pp. 1941–1956, 2010.
[45] A. Dehzangi, S. Phon-Amnuaisuk, M. Manafi, and S. Safa, "Using rotation Forest for protein fold prediction problem: An empirical study," in *Proc. 8th Eur. Conf. (EvoBIO 2010), Lecture Notes in Computer Science*, vol. 6023, Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics, pp. 217–227.
[46] H. B. Shen and K. C. Chou, "Ensemble classifier for protein fold pattern recognition," *Bioinformatics*, vol. 22, pp. 1717–1722, 2006.
[47] 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.
[48] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.
[49] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2006.
[50] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, pp. 27:1–27:27, 2011.
[51] M. K. Bardsiri and M. Eftekhari, "Comparing ensemble learning methods based on decision tree classifiers for protein fold recognition," *Int. J. Data Mining Bioinformat.*, vol. 9, no. 1, pp. 89–105, 2014.
[52] A. Dehzangi, S. Phon-Amnuaisuk, and O. Dehzangi, "Enhancing protein fold prediction accuracy by using Ensemble of different classifiers," *Australian J. Intell. Inf. Process. Syst.*, vol. 26, no. 4, pp. 32–40, 2010.
[53] A. Dehzangi, K. K. Paliwal, A. Sharma, J. Lyons, and A. Sattar, "Protein fold recognition using an overlapping segmentation approach and a mixture of feature extraction models," in *AI 2013: Advances in Artificial Intelligence*. New York: Springer, 2013, pp. 32–43.
[54] A. Dehzangi, A. Sharma, J. Lyons, K. K. Paliwal, and A. Sattar, "A mixture of physicochemical and evolutionary-based feature extraction approaches for protein fold recognition," *Int. J. Data Mining Bioinformat.*, vol. 11, no. 1, pp. 115–138, 2015.
[55] I. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. San Francisco, CA, USA: Morgan Kaufmann, 2005.
[56] J. Xu, M. Li, D. Kim, and Y. Xu, "Raptor: optimal protein threading by linear programming," *J. Bioinformat. Comput. Biol.*, vol. 1, no. 1, pp. 95–117, 2003.
[57] J. Soding, A. Biegert, and A. N. Lupas, "The hhpred interactive server for protein homology detection and structure prediction," *Nucleic Acids Res.*, vol. 33, no. Suppl 2, pp. W244–W248, 2005.
[58] J. Cheng and P. Baldi, "A machine learning information retrieval approach to protein fold recognition," *Bioinformatics*, vol. 22, no. 12, pp. 1456–1463, 2006.
[59] W. Zhang, S. Liu, and Y. Zhou, "Sp5: improving protein fold recognition by using torsion angle profiles and profile-based gap penalty model," *PLoS One*, vol. 3, no. 6, p. E2325, 2008.
[60] R. X. Yan, J. N. Si, C. Wang, and Z. Zhang, "Descfold: a web server for protein fold recognition," *BMC Bioinformat.*, vol. 10, no. 1, p. 416, 2009.
[61] J. Peng and J. Xu, "Boosting protein threading accuracy," in *Research in Computational Molecular Biology*. New York: Springer, 2009, pp. 31–45.
[62] Y. Yang, E. Faraggi, H. Zhao, and Y. Zhou, "Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates," *Bioinformatics*, vol. 27, no. 15, pp. 2076–2082, 2011.
[63] K. C. Chou and H. B. Shen, "Review: recent advances in developing web-servers for predicting protein attributes," *Natural Sci.*, vol. 1, no. 02, p. 63, 2009.
[64] M. Mizianty and L. A. Kurgan, "Modular prediction of protein structural classes from sequences of twilight-zone identity with predicting sequences," *BMC Bioinformat.*, vol. 10, no. 1, p. 414, 2009.
[65] A. Dehzangi, K. K. Paliwal, J. Lyons, A. Sharma, and A. Sattar, "Proposing a highly accurate protein structural class predictor using segmentation-based features," *BMC Genomics*, vol. 15, no. Suppl 1, p. S2, 2014.
[66] A. Dehzangi, R. Heffernan, A. Sharma, J. Lyons, K. K. Paliwal, and A. Sattar, "Gram-positive and gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou's general PseAAC," *J. Theoretical Biol.*, vol. 7, no. 364, pp. 284–294, 2015.
[67] C. Huang and J. Yuan, "Using Radial Basis function on the general form of Chou's pseudo amino acid composition and PSSM to predict subcellular locations of proteins with both single and multiple sites," *Biosystems*, vol. 113, no. 1, pp. 50–57, 2013.

James Lyons received a B.Eng. degree with honors and a B.I.T. degree from Griffith University, Brisbane, Australia, in 2007. He is now pursuing a Ph.D. degree in robust automatic speech and speaker recognition at Griffith University. His research interests include automatic speech and speaker recognition, bioinformatics, protein fold, and structural class prediction problems and pattern recognition.

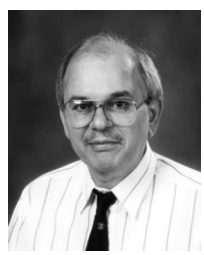
Abdollah Dehzangi (M'14) received the B.Sc. degree in computer engineering-hardware from Shiraz University, Iran, in 2007, the M.S. degree in bioinformatics from Multi Media University (MMU), Cyberjaya, Malaysia, in 2011, and the Ph.D. degree in bioinformatics at Griffith University, Brisbane, Australia, in 2015. He worked as a researcher in National ICT Australia (NICTA) from 2011 to 2014. He joined the Institute for Integrated and Intelligent Systems (IIIS) as research assistant in 2014. His research interests include bioinformatics, protein secondary, fold and structural class prediction problems, protein local and subcellular localization prediction problems, data mining, statistical learning theory, and pattern recognition. He reviewed several articles and is in the editorial board of several journals.

Rhys Heffernan received a B.Eng. degree with honors from Griffith University, Brisbane, Australia, in 2012. He is now pursuing a Ph.D. degree in machine learning with applications in bioinformatics, from Griffith University. His research interests include machine learning and pattern recognition, deep learning, and bioinformatics.

Yuedong Yang received the B.S. degree in biology and Ph.D. degree in bioinformatics from the University of Science and Technology of China in 2000 and 2006, respectively. Afterwards, he started his postdoctoral fellowship in the Schools of Informatics and Medicine, Indiana University, Bloomington, IN, USA, and was later promoted to Research Assistant Professor in 2011. In 2013, he moved to Griffith University, Australia, as a Research Fellow. His research interests include machine learning, protein structure prediction, protein function prediction, virtual ligand screening, and classification of human genetic variations. He has published 29 articles in high impact journals, such as *PNAS*, *Genome Biology*, *Nucleic Acid Research*, and *Bioinformatics*. According to Google Scholar, he has a H-index of 14 and more than 660 citations. Scopus database shows more than 85% citations were contributed by other scientists.

Yaoqi Zhou received his B.Sc. degree in chemical physics from University of Science and Technology of China in 1984 and a Ph.D. degree in chemical physics from State University of New York at Stony Brook, NY, USA, in 1990. After his postdoctoral studies at North Carolina State University (1994–1995) and Harvard University (1995–2000), he was an Assistant Professor and later Associate Professor at Department of Physiology and Biophysics, State University of New York at Buffalo, NY, USA, from 2000 to 2006, and a full Professor at Schools of Medicine and Informatics, Indiana University, from 2006 to 2013. He joined the Institute for Glycomics and School of Information and Communication Technology at Griffith University, Australia, as a Professor of Computational Biology in 2013. His research interests include development of computational algorithms and bioinformatics techniques for protein/RNA structure prediction, protein/RNA function prediction and design, computer-assisted ligand and target discovery, and discrimination of disease-causing and neutral genetic variations.

Alok Sharma (M') received the B.Tech. degree from the University of the South Pacific (USP), Suva, Fiji, in 2000 and the M.Eng. degree, with an academic excellence award, and the Ph.D. 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 on the editorial board of several journals.



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, NJ, USA, AT&T Shannon Laboratories, Florham Park, NJ, USA, and Advanced Telecommunication Research Laboratories, Kyoto, Japan. Since July 1993, he has been a professor at Griffith University, Brisbane, Australia, in the School of Microelectronic 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 the Acoustical Society of India. He has served the IEEE Signal Processing Society's 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).