A Tri-Gram Based Feature Extraction Technique Using Linear Probabilities of Position Specific Scoring Matrix for Protein Fold Recognition

Kuldip K. Paliwal, *Member, IEEE*, Alok Sharma, *Member, IEEE*, James Lyons, and Abdollah Dehzangi*, *Member, IEEE*

***Abstract*—In biological sciences, the deciphering of a three dimensional structure of a protein sequence is considered to be an important and challenging task. The identification of protein folds from primary protein sequences is an intermediate step in discovering the three dimensional structure of a protein. This can be done by utilizing feature extraction technique to accurately extract all the relevant information followed by employing a suitable classifier to label an unknown protein. In the past, several feature extraction techniques have been developed but with limited recognition accuracy only. In this study, we have developed a feature extraction technique based on tri-grams computed directly from Position Specific Scoring Matrices. The effectiveness of the feature extraction technique has been shown on two benchmark datasets. The proposed technique exhibits up to 4.4% improvement in protein fold recognition accuracy compared to the state-of-the-art feature extraction techniques.**

***Index Terms*—Feature extraction technique, position specific scoring matrix (PSSM), protein fold recognition, support vector machine (SVM), tri-gram.**

I. Introduction

PROTEIN fold recognition is an important and challenging task in biological science, biomedicine, bioinformatics and drug design. The identification of a three dimensional structure of a protein sequence provides objective information about the characterization of a protein. This would assist in understanding protein heterogeneity, protein-protein interactions and protein-peptide interactions. Though it is possible to determine the structure of a protein by crystallography methods, it is usually a very slow and time consuming process. The copiousness of protein data in this era requires the advancements of computational ways to decipher protein structure in a reasonable amount of time.

Manuscript received May 06, 2013; revised September 27, 2013; accepted December 04, 2013. Date of current version February 27, 2014. *Asterisk indicates corresponding author.*

K. K. Paliwal is with the School of Engineering, Griffith University, Brisbane 4111, Australia (e-mail: k.paliwal@griffith.edu.au).

A. Sharma is with the School of Engineering and Physics, University of the South Pacific, Fiji, and also with the Institute for Integrated and Intelligent Systems (IIIS), Griffith University, Brisbane 4111, Australia (e-mail: sharma_al@usp.ac.fj).

J. Lyons is with the School of Engineering, Griffith University, Brisbane 4111, Australia (e-mail: james.lyons@griffithuni.edu.au).

*A. Dehzangi is with the Institute for Integrated and Intelligent Systems (IIIS), Griffith University, Brisbane 4111, Australia, and National ICT Australia (NICTA), Brisbane 4111, Australia (e-mail: a.dehzangi@griffith.edu.au).

Digital Object Identifier 10.1109/TNB.2013.2296050

The prime objective of protein fold recognition is to find the fold of a protein sequence. Assigning of protein fold to a protein sequence is a transitional stage in the recognition of three dimensional structure of a protein. The protein fold recognition broadly covers feature extraction task and classification task. For the former task, several feature extraction techniques have been developed. Dubchak *et al.* have proposed syntactical and physicochemical-based features for protein fold recognition [1]. They used amino acids' composition (AAC) as syntactical-based features and 5 following attributes of amino acids for deriving physicochemical-based features namely, hydrophobicity (H), predicted secondary structure based on normalized frequency of á-helix (X), polarity (P), polarizability (Z), and van der Waals volume (V). They used three descriptors (composition, transition and distribution) to compute the features. The AAC features comprise of 20 features and physicochemical-based features comprise of 105 features (21 features for each of the attributes used). The features proposed by [1] have been widely used in the field of protein fold recognition [2]–[11]. Apart from the above mentioned 5 attributes used by [1], features also extracted by incorporating other attributes of the amino acids; and if the number of features is large then top few can be selected [12], [13]. Some of the other attributes used are: solvent accessibility [14], flexibility [15], bulkiness [16], first and second order entropy [17], and size of the side chain of the amino acids [11]. These physicochemical attributes are usually selected in an arbitrary way and recently a systematic way of selecting physicochemical attributes was proposed by [18]. Taguchi and Gromiha [19] proposed features which are based on amino acids' occurrence; Shamim *et al.* [20] have extracted features from the structural information of amino acid residues and amino acid residue pairs; Ghanty and Pal, [21] proposed pairwise frequencies of amino acids separated by one residue (PF1) and pairwise frequencies of adjacent amino acid residues (PF2). There are 400 features each in PF1 and PF2. These pairwise frequency features (PF) are used as in the augmented form in the study conducted by [22], thereby, having 800 features. Thus, the feature vector of PF has 800 features. Chou [23] proposed pseudo-amino acid composition (A) based features to effectively represent protein sequence. Dong *et al.* [24] have shown autocross-covariance (ACC) transformation for protein fold recognition. Shen and Chou [25], Kurgan *et al.* [26] and Liu *et al.* [27] have shown autocorrelation features for protein sequence, and Dehzangi *et al.* [28] derived fea-

1536-1241 © 2014 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

tures by considering more physicochemical properties. Sharma *et al.* [29] have derived bi-gram features using evolutionary information (PSSM). It is also shown that by fusion of features the recognition rates can be improved [30]–[33]. For the latter task case, several classifiers have been developed or used including linear discriminant analysis [34], [35], Bayesian classifiers [2], Bayesian decision rule [36], k-nearest neighbor [25], [37], Hidden Markov model [38], [39], artificial neural network [40], [41], support vector machine (SVM) [6], [20], [21], [42], [43], and ensemble classifiers [20], [33], [41], [44], [45]. Among these classifiers, SVM (or SVM-based for ensemble strategy) classifier exhibits quite promising results [21], [26], [27].

In order to decipher protein structure properly, the features extracted from a protein sequence should have relevant information for fold discrimination. This implies the necessity of carefully developing the feature extraction technique. Therefore, in this paper we focus on developing feature extraction technique to examine the recognition performance. Since SVM classifier provides high recognition accuracy, we use SVM classifier to compare the performance of our feature extraction technique with other feature extraction techniques. Our proposed feature extraction technique is based on the novel way of finding the neighborhood information of amino acids in a protein sequence via tri-grams.

Markowetz *et al.* [46] have shown the importance of using tri-gram features for protein fold recognition. The tri-gram features capture the neighborhood information of amino acids. Isik *et al.* [47], and Ghanty and Pal [21] also used tri-gram features, however, by reducing the dimensionality of the feature vectors. The performance in terms of recognition accuracy was not very promising for tri-gram features [21]. Since there are 20 amino acids of interest, there will be $20 \times 20 \times 20 = 8000$ combinations of the amino acid triplets (or tri-grams), giving an 8000 dimensional feature vector for a given protein sequence. If we use the frequency of each tri-gram occurring in the primary protein sequence for feature extraction, then this usually leads to a feature vector consisting of mostly zeros. Therefore, in this procedure, there is a high possibility of losing out vital information useful for protein fold recognition. This could be one of the main reasons of tri-gram features exhibiting low recognition performance [21].

It has been seen in the literature that by forming consensus sequence significantly improves the recognition performance of protein fold recognition. The consensus sequence is obtained by incorporating evolutionary information of amino acids from position specific scoring matrix (PSSM) [48]. The consensus sequence is derived from a protein sequence by replacing the amino acid of a primary protein sequence with the amino acid of the highest probability as dictated by PSSM. However, if tri-gram features are extracted from the consensus sequence instead, then the problem of having mostly zeros in a feature vector still remains.

Instead of computing tri-gram features either from the primary protein sequence or the consensus sequence, we compute in this paper tri-gram features directly from PSSM. This is done by accumulating the probabilities of each of the tri-gram using the probability information contained in PSSM. Since in this procedure we are utilizing linear probabilities to compute tri-gram features and all the combinations of tri-grams occur in PSSM, we avoid having zeros in the feature vector. Therefore, our procedure would retrieve more information useful for the protein fold recognition. Note that we can interpret our procedure of computing the tri-gram features from PSSM as the soft procedure, while the procedure used in earlier studies (where tri-gram features are computed by counting the occurrence of individual amino acid triplets from the protein sequence) can be considered as the hard procedure.

In experiment, we apply our procedure on two benchmarks namely Taguchi and Gromiha (TG), [19] dataset and extended Ding and Dubchak (EDD) [6] dataset. We performed k-fold cross-validation on the datasets and obtained very promising recognition performance. On TG dataset we get protein fold recognition accuracy of 72.5% and on EDD-dataset we get 86.2% using SVM classifier.

II. Dataset

In this study, two datasets TG and EDD are utilized. The TG dataset extracted by [19] consists of 1612 proteins belonging to 30 most populated folds from the SCOP 1.73. The sequence similarity of protein of TG datasets is no more than 25%. We extract the EDD dataset from the latest version of the SCOP 1.75 consisting of 3418 proteins belonging to 27 folds. These 27 folds were also being used in the original Ding and Dubchak dataset. In the EDD dataset the protein sequences have sequence similarity no more than 40%. For both the datasets, the major structural classes are α, β, α/β, and $\alpha + \beta$. The summary of TG and EDD datasets are given in Tables I and II.

III. Classification Technique

In this paper we used support vector machine (SVM) [49] as a classifier. SVM is considered to be the state-of-the-art machine learning and pattern classification algorithm. It has been extensively applied in classification and regression tasks. SVM aims to find maximum margin hyperplane (MMH) to minimize classification error. In SVM a function called the kernel K is used to project the data from input space to a new feature space, and if this projection is non-linear it allows non-linear decision boundaries [50].

To find a decision boundary between two classes, SVM attempts to maximize the margin between the classes, and choose linear separations in a feature space. The classification of some known point in input space $\mathbf{x}_i$ is y_i which is defined to be either -1 or $+1$. If $\mathbf{x}'$ is a point in input space with unknown classification then

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

where y' is the predicted class of point $\mathbf{x}'$. The function K() is the kernel; n is the number of support vectors; α_i are adjustable weights and b is a bias. In this study, the complexity parameter (C) is set to be 1000. We use LibSVM for training and testing with the radial basis function (RBF) kernel [51]. The RBF kernel function can be given by $K(\mathbf{z}_i, \mathbf{z}_j) = \exp(-g * \|\mathbf{z}_i - \mathbf{z}_j\|\text{^}2)$,

TABLE I
SUMMARY OF TAGUCHI AND GROMIHA DATASET.

No.	Fold	No. of samples in each fold
	α	
1	Cytochrome C	25
2	DNA/RNA binding 3-helical bundle	103
3	Four helical up and down bundle	26
4	EF hand-like fold	25
5	SAM domain-like	26
6	α-α super helix	47
	β	
7	Immunoglobulin-like β-sandwich	173
8	Common fold of diphtheria toxin/transcription factors/cytochrome	28
9	Cupredoxin-like	30
10	Galactose-binding domain-like	25
11	Concanavalin A-like lectins/glucanases	26
12	SH3-like barrel	42
13	OB-fold	78
14	Double-stranded α-helix	34
15	Nucleoplasmin-like	42
	α/β	
16	TIM α/β-barrel	145
17	NAD(P)-binding Rossmann-fold domains	77
18	FAD/NAD(P)-binding domain	31
19	Flavodoxin-like	55
20	Adenine nucleotide a hydrolase-like	34
21	P-loop containing nucleoside triphosphate hydrolases	95
22	Thioredoxin fold	32
23	Ribonuclease H-like motif	49
24	S-adenosyl-L-methionine-dependent methyltransferases	34
25	α/β-Hydrolases	37
	α + β	
26	β-Grasp, ubiquitin-like	42
27	Cystatin-like	25
28	Ferredoxin-like	118
29	Knottins	80
30	Rubredoxin-like	28

TABLE II
SUMMARY OF EXTENDED DING AND DUBCHAK DATASET

No.	Fold	No. samples in each fold
	α	
1	Globin-like	41
2	Cytochrome*c*	35
3	DNA-binding 3-helical bundle	322
4	4-Helical up-and-down bundle	69
5	4-Helical cytokines	30
6	Alpha; EF-hand	59
	β	
7	Immunoglobulin-like β-sandwich	391
8	Cupredoxins	47
9	Viral coat and capsid proteins	60
10	ConA-like lectins/glucanases	57
11	SH3-like barrel	129
12	OB-fold	156
13	Trefoil	45
14	Trypsin-like serine proteases	45
15	Lipocalins	37
	α/β	
16	(TIM)-barrel	336
17	FAD (also NAD)-binding motif	73
18	Flavodoxin-like	130
19	NAD (P)-binding Rossmann-fold	195
20	P-loop containing nucleotide	239
21	Thioredoxin-like	111
22	Ribonuclease H-like motif	128
23	Hydrolases	83
24	Periplasmic binding protein-like	16
	α + β	
25	β-Grasp	121
26	Ferredoxin-like	339
27	Small inhibitors, toxins, lectins	124

where g is gamma parameter. The gamma and C parameters are optimized using LibSVM. The data is not normalized before processing to the SVM classifier.

IV. TRI-GRAM FEATURE EXTRACTION TECHNIQUE FOR PROTEIN FOLD RECOGNITION

As mentioned earlier that if tri-gram features are extracted from the primary protein sequence or the consensus sequence, the problem of having mostly zeros in a tri-gram feature vector still remains. Therefore, in this work, we do not use a protein sequence directly or a consensus sequence for computing tri-gram features. Further, instead of using hard decision rule for computing tri-gram features, we use soft decision rule for computing the features. We use PSSM linear probabilities of a given protein sequence to compute the probabilities of individual tri-grams to form a tri-gram probability matrix $\mathbf{T}$. The $\mathbf{T}$ matrix is a 3-dimensional matrix of size $20 \times 20 \times 20$. The elements of this matrix define a tri-gram feature vector of size 8000. To define the elements of matrix $\mathbf{T}$, let us denote $\boldsymbol{P}$ the matrix of PSSM linear probabilities for the given protein. The matrix $\boldsymbol{P}$ has L rows and 20 columns (where L is the length of the protein sequence). Let $P_{i,j}$ be its element at ith row and jth column which can be interpreted as the relative probability of jth amino acid at the ith location of the protein sequence $\sum_{j=1}^{20} P_{i,j} = 1$, for $i = 1, 2, \ldots, L$ Then the (m, n, r) element of $\mathbf{T}$ matrix can be computed as:

$$\mathrm{T}_{m,n,r} = \sum_{i=1}^{\mathrm{L}-2} \mathrm{P}_{i,m}\mathrm{P}_{i+1,n}\mathrm{P}_{i+2,r}. \tag{2}$$

These 8000 elements of $\mathbf{T}$ matrix define the tri-gram feature vector $\boldsymbol{f}$ that is used to represent the given protein for protein fold recognition task. Since in the computation of tri-gram feature vector $\boldsymbol{f}$ f all the information of PSSM probability has been used and there is very low or no sparsity in the feature vector $\boldsymbol{f}$ (i.e., it has very low or no zero components), intuitively $\boldsymbol{f}$ contains more information useful for protein fold recognition task than computing tri-gram directly from the protein sequence (or from a consensus sequence by PSSM).

In order to illustrate the drawback of the conventional tri-gram feature extraction method and to present the effectiveness of our proposed feature extraction technique, we use a simple toy example in this section. Let us assume that there be in total only 3 amino acids namely V, H and R that form any protein sequence. Let a protein sequence of interest be given as $VVHVR$ of length $L = 5$ and its PSSM be given as in Table III. Using the probability information in PSSM, we can find out the consensus sequence (where each amino acid is replaced by the one that has the highest probability in PSSM) for this protein as $VVHRH$. The tri-gram features computed from the original protein sequence $VVHRH$ and the consensus sequence $VVHRH$ are shown in Tables IV and V, respectively. It can be seen from Table IV and Table V that out of 27 features only 3 have the values as 1. The remaining combinations of features do not exist in the protein sequence as well as in the consensus sequence and

TABLE III
POSITION SPECIFIC SCORING MATRIX OF THE PROTEIN SEQUENCE

Amino acids	V	H	R
V	0.45	0.30	0.25
V	0.45	0.25	0.30
H	0.20	0.60	0.20
V	0.30	0.30	0.40
R	0.20	0.65	0.15

TABLE IV
TRI-GRAM FREQUENCIES FROM THE ORIGINAL PROTEIN SEQUENCE $VVHVR$

Tri-gram	Frequency	Tri-gram	Frequency	Tri-gram	Frequency
VVV	0	*HVV*	0	*RVV*	0
VVH	1	*HVH*	0	*RVH*	0
VVR	0	*HVR*	1	*RVR*	0
VHV	1	*HHV*	0	*RHV*	0
VHH	0	*HHH*	0	*RHH*	0
VHR	0	*HHR*	0	*RHR*	0
VRV	0	*HRV*	0	*RRV*	0
VRH	0	*HRH*	0	*RRH*	0
VRR	0	*HRR*	0	*RRR*	0

TABLE V
TRI-GRAM FREQUENCIES FROM THE CONSENSUS SEQUENCE $VVHRH$

Tri-gram	Frequency	Tri-gram	Frequency	Tri-gram	Frequency
VVV	0	*HVV*	0	*RVV*	0
VVH	1	*HVH*	0	*RVH*	0
VVR	0	*HVR*	0	*RVR*	0
VHV	0	*HHV*	0	*RHV*	0
VHH	0	*HHH*	0	*RHH*	0
VHR	1	*HHR*	0	*RHR*	0
VRV	0	*HRV*	0	*RRV*	0
VRH	0	*HRH*	1	*RRH*	0
VRR	0	*HRR*	0	*RRR*	0

TABLE VI
TRI-GRAM PROBABILITY MATRIX $\mathbf{T}$

Tri-gram	Probability	Tri-gram	Probability	Tri-gram	Probability
VVV	$T_{1,1,1} = 0.0795$	*HVV*	$T_{2,1,1} = 0.0780$	*RVV*	$T_{3,1,1} = 0.0525$
VVH	$T_{1,1,2} = 0.1875$	*HVH*	$T_{2,1,2} = 0.2130$	*RVH*	$T_{3,1,2} = 0.1245$
VVR	$T_{1,1,3} = 0.0855$	*HVR*	$T_{2,1,3} = 0.0740$	*RVR*	$T_{3,1,3} = 0.0555$
VHV	$T_{1,2,1} = 0.1155$	*HHV*	$T_{2,2,1} = 0.0960$	*RHV*	$T_{3,2,1} = 0.0785$
VHH	$T_{1,2,2} = 0.1875$	*HHH*	$T_{2,2,2} = 0.2070$	*RHH*	$T_{3,2,2} = 0.1305$
VHR	$T_{1,2,3} = 0.1395$	*HHR*	$T_{2,2,3} = 0.1020$	*RHR*	$T_{3,2,3} = 0.0935$
VRV	$T_{1,3,1} = 0.0700$	*HRV*	$T_{2,3,1} = 0.0810$	*RRV*	$T_{3,3,1} = 0.0490$
VRH	$T_{1,3,2} = 0.1600$	*HRH*	$T_{2,3,2} = 0.2250$	*RRH*	$T_{3,3,2} = 0.1150$
VRR	$T_{1,3,3} = 0.0750$	*HRR*	$T_{2,3,3} = 0.0740$	*RRR*	$T_{3,3,3} = 0.0510$

therefore have the values as 0. On the other hand, the tri-gram features computed from (1) is shown in Table VI. This gives 27 dimensional tri-gram feature vector $\boldsymbol{f}$ which does not have the sparsity as had in Tables IV and V. Intuitively, the feature vector of Table VI has more information than the feature vector of Tables IV and V. This is demonstrated through experimentation, described in the next section.

V. EXPERIMENTAL RESULTS

Experiments are conducted on two datasets (TG and EDD) to demonstrate the effectiveness of the proposed feature extraction technique. The results related to TG dataset and EDD dataset are shown in Tables VII and VIII, respectively. For classification, the SVM classifier is employed and the classification performance is measured in terms of accuracy of the protein fold recognition, where the accuracy is defined as the percentage of correctly recognized proteins of the test set. In the experiments, the k-fold cross-validation [1] procedure is used to find the classification performance for different feature extraction techniques. The values of k are taken to be 5, 6, 7, 8, 9, and 10. For the SVM classifier, RBF kernel is used. The RBF kernel parameters, gamma and C, are optimized using LibSVM. The following feature sets are computed from the original protein sequences for the experiment: PF1, PF2 [21], PF [22], Occurrence (O) [19], AAC and $\mathrm{AAC} + \mathrm{HXPZV}$ [6]. We have used PSSM probabilities to find the consensus sequence for each of the original protein sequence in both of the datasets. This is done by replacing the amino acid of the original protein sequence by the amino acid having the highest probability in PSSM. We also use the above-mentioned feature extraction techniques (PF1, PF2, PF, O, AAC and $\mathrm{AAC} + \mathrm{HXPZV}$) on the consensus sequences to obtain the additional feature sets. In addition, ACC [24] and Bigram [29] have also been used for feature extraction. In the Tables VII and VIII the feature sets obtained from the consensus sequence are denoted as $\mathrm{PSSM} + FEAT$, where $FEAT$ is any feature extraction technique. For our tri-gram feature extraction technique, (1) has been employed to compute the features. Thus, there are 17 types of feature sets shown in Tables VII and VIII the first 8 are computed from the original protein sequences, the next 7 are extracted from the consensus sequences and the last two are extracted from the full PSSMs. These feature sets are evaluated in terms of classification performance using k-fold cross-validation procedure and the results are shown in these tables. The highest recognition accuracy of a particular k-fold cross-validation is indicated in bold face.

It can be seen from Tables VII and VIII that for the original protein sequences and consensus sequences, PF is giving better recognition accuracy then other feature extraction techniques. The tri-gram feature set does not perform satisfactorily when it is computed from the original protein sequences; however, its performance improves and becomes comparable to other feature sets when it is computed from the consensus sequences. Dong's feature set (ACC) exhibits quite promising results on both the datasets. Bigram feature set [29] is also giving quite promising results on both the datasets. For EDD dataset, the bigram features reached to 84.5% recognition accuracy. The tri-gram feature (of this paper) gives the best recognition performance for both the datasets. For TG dataset the proposed tri-gram feature is giving between 71.4% and 72.5% recognition accuracy. For EDD dataset the recognition accuracy is between 85.7% and 86.2%. The improvement in terms of recognition performance is quite promising for the proposed feature extraction technique.

[1]For statistical stability we performed 100 times k-fold cross-validation in this paper.

TABLE VII
RECOGNITION ACCURACY (IN PERCENTAGE) BY k-FOLD CROSS VALIDATION PROCEDURE FOR VARIOUS FEATURE EXTRACTION TECHNIQUES USING SVM CLASSIFIER ON TAGUCHI AND GROMIHA (TG) DATASET

Feature sets	$k=5$	$k=6$	$k=7$	$k=8$	$k=9$	$k=10$
PF1	38.1	38.4	38.6	38.7	38.8	38.8
PF2	38.0	38.4	38.5	38.6	38.7	38.8
PF	42.3	42.6	42.7	43.0	43.0	43.1
O	35.8	36.1	36.2	36.1	36.3	36.3
AAC	31.5	31.5	31.7	31.8	31.9	32.0
AAC+HXPZV	35.7	36.0	36.1	36.2	36.3	36.3
Trigram	34.3	34.6	34.9	34.9	35.1	35.1
ACC	64.9	65.4	65.9	66.2	66.4	66.4
PSSM+PF1	51.1	51.5	52.0	52.3	52.4	52.7
PSSM+PF2	50.2	50.4	50.7	50.8	51.0	51.1
PSSM+PF	57.2	57.8	58.0	58.3	58.5	58.8
PSSM+O	46.0	46.3	46.5	46.5	46.7	46.7
PSSM+AAC	43.2	43.5	43.6	43.8	43.8	44.0
PSSM+AAC+HXPZV	45.6	45.9	46.0	46.2	46.3	46.6
PSSM+Tri-gram	47.9	48.5	48.8	49.0	49.2	49.4
Bigram	67.1	67.5	67.6	67.8	68.1	68.1
Tri-gram (this paper)	**71.4**	**71.7**	**72.3**	**73.3**	**72.4**	**72.5**

TABLE VIII
RECOGNITION ACCURACY (IN PERCENTAGE) BY k-FOLD CROSS VALIDATION PROCEDURE FOR VARIOUS FEATURE EXTRACTION TECHNIQUES USING SVM CLASSIFIER ON EXTENDED DING AND DUBCHAK (EDD) DATASET

Feature sets	$k=5$	$k=6$	$k=7$	$k=8$	$k=9$	$k=10$
PF1	50.2	50.5	50.5	50.7	50.8	50.8
PF2	49.3	49.5	49.7	49.8	49.8	49.9
PF	54.7	55.0	55.2	55.4	55.5	55.6
O	46.4	46.6	46.6	46.7	46.7	46.9
AAC	40.3	40.6	40.7	40.7	40.9	40.9
AAC+HXPZV	40.2	40.4	40.6	40.7	40.9	40.9
Trigram	47.5	47.7	48.0	48.1	48.2	48.3
ACC	84.9	85.2	85.4	85.6	85.8	85.9
PSSM+PF1	74.1	74.5	74.7	75.0	75.1	75.2
PSSM+PF2	73.7	74.1	74.5	74.6	74.7	74.9
PSSM+PF	78.2	78.6	78.8	79.0	79.1	79.3
PSSM+O	67.6	68.0	68.1	68.3	68.3	68.5
PSSM+AAC	60.9	61.3	61.5	61.6	61.7	61.9
PSSM+AAC+HXPZV	66.7	67.2	67.4	67.7	67.8	67.9
PSSM+Tri-gram	69.8	70.2	70.5	70.8	70.9	71.0
Bigram	83.6	84.0	84.1	84.3	84.3	84.5
Tri-gram (this paper)	**85.7**	**85.9**	**86.0**	**86.1**	**86.2**	**86.2**

VI. CONCLUSION

The tri-gram feature extraction technique for protein fold recognition is proposed in this paper. The proposed technique utilizes PSSM linear probabilities to compute the features. This feature extraction technique is studied on two benchmark datasets and its performance is compared with that of the other existing feature extraction techniques. The results reported in terms of recognition performance show the effectiveness of the proposed technique. It is noted that the proposed technique exhibits up to $0.3 \sim 4.4\%$ improvement in protein fold recognition accuracy with respect to the other feature extraction techniques.

REFERENCES

[1] I. Dubchak, I. Muchnik, and S. K. Kim, "Protein folding class predictor for SCOP: approach based on global descriptors," in *Proc. 5th Int. Conf. Intell. Syst. Mol. Biol.*, 1997, pp. 104–107.

[2] A. Chinnasamy, W. K. Sung, and A. Mittal, "Protein structure and fold prediction using tree-augmented naive Bayesian classifier," *J. Bioinformat. Comp. Biol.*, vol. 3, no. 4, pp. 803–819, 2005.

[3] 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.

[4] Y. Krishnaraj and C. K. Reddy, "Boosting methods for protein fold recognition: an empirical comparison," in *Proc. IEEE Int. Conf. Bioinformat. Biomed.*, 2008, pp. 393–396.

[5] I. K. Valavanis, G. M. Spyrou, and K. S. Nikita, "A comparative study of multi-classification methods for protein fold recognition," *Int. J. Comput. Intell. Bioinformat. Syst. Biol.*, vol. 1, no. 3, pp. 332–346, 2010.

[6] C. Ding and I. Dubchak, "Multi-class protein fold recognition using support vector machines and neural networks," *Bioinformatics*, vol. 17, no. 4, pp. 349–358, 2001.

[7] A. Dehzangi, S. P. Amnuaisuk, K. H. Ng, and E. Mohandesi, "Protein fold prediction problem using ensemble of classifiers," in *Proc. 16th Int. Conf. Neural Inf. Process. II*, 2009, pp. 503–511.

[8] K. Kavousi, B. Moshiri, M. Sadeghi, B. N. Araabi, and A. A. Moosavi-Movahedi, "A protein fold classier formed by fusing different modes of pseudo amino acid composition via PSSM," *Comput. Biol. Chem.*, vol. 35, no. 1, pp. 1–9, 2011.

[9] V. Kecman and T. Yang, "Protein fold recognition with adaptive local hyper plane algorithm," in *Proc IEEE Symp. Comput. Intell. Bioinformat. Comput. Biol. (CIBCB '09)*, pp. 75–78.

[10] W. Chmielnicki and K. Stapor, "A hybrid discriminative-generative approach to protein fold recognition," *Neurocomputing*, vol. 75, pp. 194–198, 2012.

[11] A. Dehzangi and S. P. Amnuaisuk, "Fold prediction problem: the application of new physical and physicochemical-based features," *Protein Peptide Lett.*, vol. 18, pp. 174–185, 2011.

[12] A. Sharma, S. Imoto, and S. Miyano, "A top-r feature selection algorithm for microarray gene expression data," *IEEE/ACM Trans. Comput. Biol. Bioinformat.*, vol. 9, no. 3, pp. 754–764, 2012.

[13] A. Sharma, S. Imoto, S. Miyano, and V. Sharma, "Null space based feature selection method for gene expression data," *Int. J. Mach. Learn. Cybern.*, vol. 3, no. 4, pp. 269–276, 2012b.
[14] H. Zhang, T. Zhang, J. Gao, J. Ruan, S. Shen, and L. A. Kurgan, "Determination of protein folding kinetic types using sequence and predicted secondary structure and solvent accessibility," *Amino Acids*, pp. 1–13, 2010.
[15] R. Najmanovich, J. Kuttner, V. Sobolev, and M. Edelman, "Side-chain flexibility in proteins upon ligand binding," *Proteins: Struct., Funct., Bioinformat.*, vol. 39, no. 3, pp. 261–268, 2000.
[16] J. T. Huang and J. Tian, "Amino acid sequence predicts folding rate for middle-size two-state proteins," *Proteins: Struct., Funct., Bioinformat.*, vol. 63, no. 3, pp. 551–554, 2006.
[17] T. L. Zhang, Y. S. Ding, and K. C. Chou, "Prediction protein structural classes with pseudo amino acid composition: approximate entropy and hydrophobicity pattern," *Theoretical Biol.*, vol. 250, pp. 186–193, 2008.
[18] 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, pp. 1–11, 2013.
[19] Y. H. Taguchi and M. M. Gromiha, "Application of amino acid occurrence for discriminating different folding types of globular proteins," *BMC Bioinformat.*, vol. 8, p. 404, 2007.
[20] 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.
[21] 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. 2, no. 4, pp. 100–110, 2009.
[22] T. Yang, V. Kecman, L. Cao, C. Zhang, and J. Z. Huang, "Margin-based ensemble classifier for protein fold recognition," *Expert Syst. Appl.*, vol. 38, pp. 12348–12355, 2011.
[23] K. C. Chou, "Prediction of protein cellular attributes using pseudo amino acid composition," *Proteins*, vol. 43, pp. 246–255, 2001, *erratum: 2001, vol. 44, 60*.
[24] Q. Dong, S. Zhou, and J. Guan, "A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation," *Bioinformatics*, vol. 25, no. 20, pp. 2655–2662, 2009.
[25] H. B. Shen and K. C. Chou, "Ensemble classier for protein fold pattern recognition," *Bioinformatics*, vol. 22, pp. 1717–1722, 2006.
[26] L. A. Kurgan, T. Zhang, H. Zhang, S. Shen, and J. Ruan, "Secondary structure-based assignment of the protein structural classes," *Amino Acids*, vol. 35, pp. 551–564, 2008.
[27] T. Liu, X. Geng, X. Zheng, R. Li, and J. Wang, "Accurate prediction of protein structural class using auto covariance transformation of PSI-BLAST profiles," *Amino Acids*, vol. 42, pp. 2243–2249, 2012.
[28] A. Dehzangi, K. K. Paliwal, A. Sharma, O. Dehzangi, and A. Sattar, "A combination of feature extraction methods with an ensemble of different classifiers for protein structural class prediction problem," *IEEE Trans. Comput. Biol. Bioinformat.*, 2013, to be published.
[29] 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, no. 7, pp. 41–46, 2013.
[30] K. C. Chou and H. B. Shen, "Foldrate: A web-server for predicting protein folding rates from primary sequence," *Open Bioinformat. J.*, vol. 3, no. 2, pp. 31–50, 2009.
[31] L. Nanni, J. Y. Shi, S. Brahnam, and A. Luminim, "Protein classification using texture descriptors extracted from the protein backbone image," *J. Theoretical Biol.*, vol. 264, no. 3, pp. 1024–1032, 2010.
[32] L. Nanni, A. Lumini, and A. Brahnam, "An empirical study on the matrix-based protein representations and their combination with sequence-based approaches," *Amino Acids*, vol. 44, no. 3, pp. 887–901, 2013.
[33] H. B. Shen, J. N. Song, and K. C. Chou, "Prediction of protein folding rates from primary sequence by fusing multiple sequential features," *J. Biomed. Sci. Eng.*, vol. 2, pp. 136–143, 2009.
[34] A. Sharma and K. K. Paliwal, "Cancer classification by gradient LDA technique using microarray gene expression data," *Data Knowl. Eng.*, vol. 66, no. 2, pp. 338–347, 2008.
[35] P. Klein, "Prediction of protein structural class by discriminant analysis," *BiochimBiopjysActa*, vol. 874, pp. 205–215, 1986.
[36] Z. Z. Wang and Z. Yuan, "How good is prediction of protein-structural class by the component-coupled method?," *Proteins*, vol. 38, pp. 165–175, 2000.
[37] Y. S. Ding and T. L. Zhang, "Using Chou's pseudo amino acid composition to predict subcellular localization of apoptosis proteins: an approach with immune genetic algorithm-based ensemble classifier," *Pattern Recog. Lett.*, vol. 29, pp. 1887–1892, 2008.
[38] D. Bouchaffra and J. Tan, "Protein fold recognition using a structural hidden Markov model," in *Proc. 18th Int. Conf. Pattern Recog.*, 2006, pp. 186–189.
[39] P. Deschavanne and P. Tuffery, "Enhanced protein fold recognition using a structural alphabet," *Proteins: Struct., Funct., Bioinformat.*, vol. 76, pp. 129–137, 2009.
[40] K. Chen, X. Zhang, M. Q. Yang, and J. Y. Yang, "Ensemble of probabilistic neural networks for protein fold recognition," in *Proc. 7th IEEE Int. Conf. Bioinformat. and Bioengineering (BIBE)*, 2007, pp. 66–70.
[41] Y. Ying, K. Huang, and C. Campbell, "Enhanced protein fold recognition through a novel data integration approach," *BMC Bioinformat.*, vol. 10, no. 1, 267, 2009.
[42] A. Dehzangi, K. K. Paliwal, J. Lyons, A. Sharma, and A. Sattar, "Enhancing protein fold prediction accuracy using evolutionary and structural features," in *Eighth IAPR Int. Conf. Pattern Recognition in Bioinformat. (PRIB), LNBI*, 2013, vol. 7986, pp. 196–207.
[43] A. Dehzangi, K. K. Paliwal, J. Lyons, A. Sharma, and A. Sattar, "A segmentation-based method to extract structural and evolutionary features for protein fold recognition," *IEEE Trans. Comput. Biol. Bioinformat.*, 2013, to be published.
[44] A. Dehzangi, S. P. Amnuaisuk, and O. Dehzangi, "Enhancing protein fold prediction accuracy by using ensemble of different classifiers," *Australian J. Intell. Inform. Process. Syst.*, vol. 26, no. 4, pp. 32–40, 2010.
[45] A. Dehzangi and S. Karamizadeh, "Solving protein fold prediction problem using fusion of heterogeneous classifiers," *Inform. Int. Interdisciplinary J.*, vol. 14, no. 11, pp. 3611–3622, 2011.
[46] F. Markowetz, L. Edler, and M. Vingron, "Support vector machines for protein fold class prediction," *Biometrical J.*, vol. 45, no. 3, pp. 377–389, 2003.
[47] Z. Isik, B. Yanikoglu, and U. Sezeman, "Protein structural class determination using support vector machines," in *Proc. 19th Int. Symp. Comput. Inform. Sci. (ISCIS'04)*, 2004, pp. 82–89.
[48] 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.
[49] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.
[50] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer Science, 2006.
[51] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, 2011.

Kuldip K. 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, 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 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 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).

Alok Sharma (M'02) 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 in the editorial board of several journals.

James Lyons received a B.Eng. degree with Honors and a BIT 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 (GS'13) received the B.Sc. degree in computer engineering-hardware from Shiraz University, Iran, in 2007 and the M.S. degree in 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.