Journal of Theoretical Biology 393 (2016) 67–74

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Protein fold recognition using HMM–HMM alignment and dynamic programming

James Lyons [a], Kuldip K. Paliwal [a], Abdollah Dehzangi [b], Rhys Heffernan [a], Tatsuhiko Tsunoda [c,d], Alok Sharma [d,e,f,*]

[a] School of Engineering, Griffith University, Brisbane, QLD 4111, Australia
[b] University of Iowa, USA
[c] Department of Medical Science Mathematics, Medical Research Institute, Tokyo Medical and Dental University, 1-5-45 Yushima, Bunkyo-ku, Tokyo 113-8510, Japan
[d] Laboratory for Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences, 1-7-22 Suehiro-cho, Tsurumi, Yokohama 230-0045, Japan
[e] Institute of Integrated and Intelligent Systems, Griffith University, Brisbane, Australia
[f] School of Engineering and Physics, University of the South Pacific, Fiji

HIGHLIGHTS

- Performance of protein fold recognition has been improved.
- A new feature extraction method has been proposed.
- An improvement of around 2.7–11.6% has been observed.

ARTICLE INFO

Article history:
Received 26 October 2015
Received in revised form
17 December 2015
Accepted 18 December 2015
Available online 19 January 2016

Keywords:
Protein fold recognition
HMM–HMM alignment profile
Dynamic time warping
Classification

ABSTRACT

Detecting three dimensional structures of protein sequences is a challenging task in biological sciences. For this purpose, protein fold recognition has been utilized as an intermediate step which helps in classifying a novel protein sequence into one of its folds. The process of protein fold recognition encompasses feature extraction of protein sequences and feature identification through suitable classifiers. Several feature extractors are developed to retrieve useful information from protein sequences. These features are generally extracted by constituting protein's sequential, physicochemical and evolutionary properties. The performance in terms of recognition accuracy has also been gradually improved over the last decade. However, it is yet to reach a well reasonable and accepted level. In this work, we first applied HMM–HMM alignment of protein sequence from HHblits to extract profile HMM (PHMM) matrix. Then we computed the distance between respective PHMM matrices using kernalized dynamic programming. We have recorded significant improvement in fold recognition over the state-of-the-art feature extractors. The improvement of recognition accuracy is in the range of 2.7–11.6% when experimented on three benchmark datasets from Structural Classification of Proteins.

1. Introduction

Protein fold recognition (PFR) is considered as a transitional step towards protein tertiary structure identification. It has two main components: feature extraction of protein sequence and then feature identification using suitable classifiers. The purpose of feature extraction is to find informative features from primary protein sequence which can be effectively used in the classification stage. The target of PFR is to associate a fold to a novel protein sequence. The tertiary structure of protein helps in understanding protein interactions, protein heterogeneity and in drug design. Since the identification of tertiary structures through experimental methods (such as x-ray crystallography and nuclear magnetic resonance) is very time consuming, computational approaches attracted considerable attention over the years.

* Corresponding author at: Laboratory for Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences, 1-7-22 Suehiro-cho, Tsurumi, Yokohama 230-0045, Japan.
E-mail addresses: alok.sharma@griffith.edu.au, alok.fj@gmail.com, alokanand.sharma@riken.jp (A. Sharma).

http://dx.doi.org/10.1016/j.jtbi.2015.12.018
0022-5193/

Since feature extraction plays a pivotal role in PFR, we concentrate on developing a feature extraction scheme that would have a better performance. Features are generally extracted by encompassing sequential, physicochemical and evolutionary information (Ding and Dubchak, 2001; Chou 2001, 2005; Sharma et al., 2013b, 2015c; Lyons et al., 2014a, 2014b; Paliwal et al., 2014a, 2014b; Dong et al., 2009; Yang et al., 2011; Dehzangi et al., 2010a, 2010b, 2010c; Heffernan et al., 2015a, 2015b; Mamun and Sharma, 2014). Some authors have also used functional information as features (Zakeri et al., 2014). The use of functional information has shown to improve the performance. However, the functional information is usually extracted either by experimental methods or by previously known structural information. In this paper, we primarily use evolutionary information of protein sequence to extract features. Over the years, a wide range of feature extraction techniques are proposed (Nemethy and Scheraga, 1984; Nemethy et al., 1985, 1986; Rumsey et al., 1985; Pottle and Scheraga, 1989; Maggiora and Scheraga, 1992; Carlacci, 1991; Chou and Scheraga, 1982). A brief description of the work carried in the last decade has been given in the Related Work section.

Some popular methods of computing evolutionary information of protein is by using PSI-BLAST (Altschul et al., 1997). This technique generates profile of a protein sequence known as Position Specific Scoring Matrix (PSSM). For a protein sequence of length L would have PSSM of size $L \times 20$. This profile can be used as a basis to retrieve vital information regarding a protein sequence. It has shown in the literature that features extracted using PSSM has considerably improved the performance of PFR (Shen, 2009; Wu and Xiao, 2011a, 2011b; Lin et al., 2012; Lin and Fang, 2013; Chou and Shen, 2007; Dehzangi et al., 2015a, 2015b, 2015c; Saini et al., 2014, 2015a, 2015b; Sharma et al., 2015c). Recently a hidden Markov model (HMM) based protein sequence search tool HHblits (Remmert et al., 2011) has been proposed. This technique provides profile of HMM (PHMM) for a protein sequence, the size of which is $L \times 20$. It is experimented to be faster and more sensitive than PSI-BLAST tool.

The HMM profile has been shown to be a more effective approach for remote homology detection compared to PSSM (Yan et al., 2013). Most of methods utilized evolutionary information from PSSM, however, in this work, we employed profile of HMM. For this, we first utilize HHblits to obtain PHMM for a protein sequence. Then we incorporate PHMM to constitute distance matrix between two protein profiles. We use this distance matrix to find alignment path using dynamic programming. Since different proteins of varying lengths could share the same fold, we can extract meaningful features from the aligned homologous proteins. In other words, if the distance matrix between two proteins is low, they belong to the same fold otherwise they are not. Therefore, we can use training set to estimate the distance matrix which would help in recognizing de novo protein sequences.

2. Related work

Protein fold recognition research has two main components, feature extraction of protein sequence to retrieve useful information and classification of extracted features to associate a unique fold to a de novo protein sequence. Over the last several years, progress has been made to develop both these components. In this section, we briefly cover these aspects and describe some of the preceding works.

In general, features are extracted by using structural, physicochemical and evolutionary information. Dubchak et al. (1997) have proposed syntactical and physicochemical-based features for PFR. They used amino acids' composition (AAC) as syntactical-based features and the 5 following attributes of amino acids for deriving physicochemical-based features namely, hydrophobicity (H), predicted secondary structure based on normalized frequency of α-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 Dubchak et al. (1997) have been widely used in the field of PFR (Chinnasamy et al., 2005; Krishnaraj and Reddy, 2008; Valavanis et al., 2010; Ding and Dubchak, 2001; Dehzangi et al., 2009; Kecman and Yang, 2009; Kavousi et al, 2011, Dehzangi and Amnuaisuk, 2011; Chmielnicki and Stapor, 2012; Dehzangi et al., 2013a, 2013b, 2013c). Apart from the above mentioned 5 attributes used by Dubchak et al. (1997), features have also been extracted by incorporating other attributes of amino acids. Some of the other attributes used are: solvent accessibility (Zhang et al., 2010), flexibility (Najmanovich et al., 2000), bulkiness (Huang and Tian, 2006), first and second order entropy (Zhang et al., 2008), size of the side chain of the amino acids (Dehzangi and Amnuaisuk, 2011). These physicochemical attributes are selected in an arbitrary way and recently a systematic way of selecting physicochemical attributes was proposed by Sharma et al. (2013a, 2012b). Ohlson et al. (2004) proposed a profile–profile alignment method to improve PFR. Taguchi and Gromiha (2007) proposed features which are based on amino acids' occurrence; Shamim et al. (2007) have extracted features from the structural information of amino acid residues and amino acid residue pairs; Ghanty and Pal (2009) 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 concatenated in the study conducted by Yang et al. (2011), thereby, having 800 features. If the dimensionality of a feature vector is very large then a few important features can be selected for further processing using feature selection or dimensionality reduction schemes (Sharma and Paliwal, 2006b, 2006c, 2007, 2008a, 2008b, 2008c, 2010a, 2010b, 2015a, 2015b, 2012c, 2012f, 2012g; Sharma et al., 2005, 2006a, 2011, 2012a, 2012d, 2012e, 2012h, 2014b, 2014c; Paliwal and Sharma, 2010, 2011, 2012). To avoid completely losing the sequence-order information, the pseudo amino acid composition (Chou, 2001, 2005) or Chou's PseAAC (Lin and Lapointe, 2013; Shen and Chou, 2008) was proposed to replace the simple amino acid composition (AAC) for representing the sample of a protein. Dong et al. (2009) have shown autocross-covariance (ACC) transformation for protein fold recognition. Shen and Chou (2006), Kurgan et al. (2008) and Liu et al. (2012) have shown autocorrelation features for protein sequence, and Dehzangi and Amnuaisuk (2011) derived features by considering more physicochemical properties. Sharma et al. (2013b) have derived bi-gram features using evolutionary information (PSSM). Paliwal et al. (2014a, 2014b) have proposed trigram, and amalgamation of evolutionary and structural based features (PSSM-SPINE-X) based features using evolutionary information. Sharma et al. (2014a) used intrinsically disordered regions for protein function estimation and Lyons et al. (2014a) use alignment method incorporating dynamic programming for feature extraction. For classification task, several classifiers have been developed or used including linear discriminant analysis (Klein, 1986), Bayesian classifiers (Chinnasamy et al., 2005), Bayesian decision rule (Wang and Yuan, 2000), k-nearest neighbor (Shen and Chou, 2006; Ding and Zhang, 2008), hidden Markov model (Bouchaffra and Tan, 2006; Deschavanne and Tuffery, 2009), artificial neural network (Chen et al., 2007, Ying et al., 2009), support vector machine (SVM) (Ding and Dubchak, 2001; Shamim et al., 2007; Ghanty and Pal, 2009) and ensemble classifiers (Dehzangi et al., 2009, 2010a, 2010b; Yang et al., 2011; Dehzangi and

Karamizadeh, 2011). Among these classifiers, SVM (or SVM-based for ensemble strategy) classifier exhibits quite promising results (Liu et al., 2012; Kurgan et al., 2008; Ghanty and Pal, 2009).

In order to improve performance of PFR, it is important to extract relevant and meaningful information from protein sequence. In this view, we focus on carefully developing the feature extraction method. It has been seen in the literature that since SVM classifier (Vapnik, 1995) provides high recognition accuracy, we use this classifier to compare the performance of our feature extraction method with other feature extraction methods. SVM often employs Radial Basis Function (RBF) kernel. The RBF kernel (along with other common SVM kernels such as the linear and polynomial kernel) requires fixed length feature vectors. This has motivated many previous works to try and extract fixed length representations of proteins so that they can then be efficiently compared. In this work we define a kernel designed to work with variable length data. This allows us to directly compare PHMM matrices, instead of first transforming the matrix into a fixed length vector prior to comparison.

As demonstrated by a series of recent publications (Chen et al., 2014; Xu et al., 2014; Lin et al., 2014; Jia et al., 2015a, 2015b) in compliance with Chou's 5-step rule (Chou, 2011), to establish a really useful sequence-based statistical predictor for a biological system, one should follow the following five guidelines: (a) construct or select a valid benchmark dataset to train and test the predictor; (b) formulate the biological sequence samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (c) introduce or develop a powerful algorithm (or engine) to operate the prediction; (d) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; (e) establish a user-friendly web-server for the predictor that is accessible to the public. Here, we are to describe how to deal with these steps one-by- one.

3. Dataset

Three datasets are used in this study to gauge the performance of the proposed feature extraction scheme compared with the other existing techniques. These datasets are DD-dataset (Ding and Dubchak, 2001), EDD-dataset (Dong et al., 2009) and TG-dataset (Taguchi and Gromiha, 2007). DD-dataset contains a training set of 311 proteins and a test set of 384 proteins. Any two proteins have 35–40% of sequence identity for aligned subsequence longer than 80 residues. The dataset has 27 Structural Classification of Proteins (SCOP) folds which represent all major structural classes: α, β, α/β, and $\alpha+\beta$ (Ding and Dubchak, 2001). The training and test sets are merged to perform 10-fold cross-validation to evaluate the performance.

The EDD-dataset consists of 3418 proteins with less than 40% sequential similarity belonging to the 27 folds that originally used in DD-dataset. We extracted the EDD-dataset from SCOP in similar manner as Dong et al. (2009) did, in order to study our proposed method using a larger number of samples.

The TG-dataset consists of 1612 protein sequences belonging to 30 different folding types of globular proteins from SCOP. The names of the number of protein sequences in each of 30 folds have been described in Taguchi and Gromiha (2007). The sequence similarity of protein of TG datasets is no more than 25%.

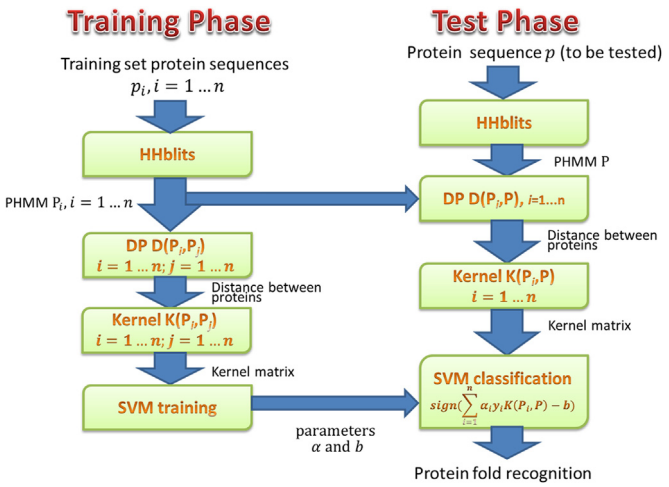

Fig. 1. A flow-diagram illustrating protein fold recognition using profile of HMM and dynamic programming through SVM classifier. In the training phase, protein sequences are processed through HHblits to get HMM profiles (PHMM). Then distance matrix between two PHMMs is evaluated in the DP step. This distance matrix is further sent to the kernelization procedure to obtain kernel matrix which is used in the SVM training to find α and b parameters. In the test phase, HMM profile P of a protein sequence p is obtained by HHblits which is used in the DP step to evaluate distance between P and protein sequences of the training set. After this step, kernel matrix is obtained which is used in the SVM classification stage to find a fold of this protein sequence p.

4. Kernelized dynamic programming based on profile HMM for feature extraction

This section indulges on the proposed feature extraction scheme for PFR. To provide an overview of the scheme, a flow diagram has been illustrated in Fig. 1. Features from protein sequences are extracted and processed through SVM classifier to predict folds. In the training phase, the parameters of SVM classifier are estimated and in the test phase a novel protein sequence is associated to one of its folds. In order to compute the necessary features, the first step is to obtain PHMM profiles from HHblits. The size of PHMM profile of a L length protein sequence is $L \times 20$ (actually, it produces $L \times 30$ matrix where 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 depict 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). Then in the dynamic programming (DP) step, distance matrix is computed by comparing all row vectors in two PHMM matrices. This distance matrix is then used to compute the overall distance between the two proteins. This distance is further used in kernelization stage to compute kernel distance measure which is further used to train the parameters of SVM classifier. If the parameters are evaluated then a fold can be predicted for a protein sequence in the test phase. For brevity, we refer this method as PHMM-DP.

In order to explain the scheme in detail, let us assume profile of HMM of two protein sequences be A and B of sizes $L_A \times 20$ and $L_B \times 20$. Let a_i (for $i = 1, 2, \ldots, L_A$) and b_j (for $j = 1, 2, \ldots, L_B$) be row vectors of A and B, respectively. The cosine distance between a_i and b_j can be depicted as

$$d(a_i, b_j) = 1 - \frac{a_i b_j^{\mathrm{T}}}{||a||\ ||b||}, \tag{1}$$

where $||a|| = a_i a_i^{\mathrm{T}}$ and $||b|| = b_j b_j^{\mathrm{T}}$. The consine distance has been used by previous investigators (Lyons et al., 2014a, 2014b; Cai, 2006; Chou, 1993). Computing distances d for L_A and L_B rows

would give a distance matrix D of size $L_A \times L_B$. After obtaining distance matrix D, we can apply dynamic programming to obtain minimum cost path. This procedure will help to find cumulative distance matrix C which defines the total cost between (a_1, b_1) and (a_i, b_j). If proteins are similar then the cost will be low. The cumulative distance matrix can be obtained as

$$C_{i,j} = \min(C_{i-1,j}, C_{i,j-1}, C_{i-1,j-1}) + D_{i,j} \quad (2)$$

where $C_{i,j}$ is an empty set for $i \leq 0$ and/or $j \leq 0$ and $D_{i,j} = d(a_i, b_j)$.

Computing $C_{i,j}$ for all i and j would give distance between two PHMM matrices A and B, denoted here as $C_{dtw}(A, B)$. This distance is a measure of similarity between the two aligned proteins. Let a kernel matrix between A and B be $K(A, B)$, having γ as a kernel parameter (selected by cross-validation on the training set). Then kernel function K can be represented in C_{dtw} as $K(A, B) = \exp(-C_{dtw}(A, B)^2/\gamma^2)$. By computing distance K between all the pairs of proteins, we obtain a kernel matrix of size $n \times n$ where n is the number of training samples. This matrix is then processed via the SVM classifier for estimating model for classification.

5. Support vector machine as a classifier

SVM (Vapnik, 1995) is considered to be the state-of-the-art machine learning and pattern classification algorithm. It has been widely applied in classification and regression tasks. SVM aims to find maximum margin hyper-plane (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 (Bishop, 2006). This function K is usually considered as RBF kernel, polynomial kernel or linear kernel. These kernels require fixed length feature vectors. Since the protein sequences are of varying lengths, we can't use these kernels. However, in this work we have defined a kernel function that can cater for this varying length (of proteins) problem without limiting the proteins to a fixed length vector. This would provide SVM more relevant and useful information for protein fold recognition.

In order to find a decision boundary between two folds, SVM attempts to maximize the margin between the folds, 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

$$y' = \text{sign}\left(\sum_{i=1}^{n} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}') + \beta\right) \quad (3)$$

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 β is a bias. We use libsvm (Chang and Lin, 2011) for training and test with our kernel function.

6. Results and discussions

In this section we show the effectiveness of the proposed scheme on three benchmark datasets: DD, EDD and TG. To find the PFR accuracy we have used SVM classifier from libsvm (Chang and Lin, 2011). The accuracy is a percentage of correctly recognized proteins to all the proteins in the test set. The SVM parameters (gamma and C) are optimized using grid search. We adopted 10-fold cross-validation in this study as done by many investigators with SVM as the prediction engine. We performed 50 times 10-fold cross-validation in this paper to get statistical stable values.[1] The cross-validation process was done in the following manner:

Table 1
Recognition accuracy by 10-fold cross validation procedure for various feature extraction techniques using SVM classifier on DD, EDD and TG datasets.

Feature sets	DD	EDD	TG
PF1 (Ghanty and Pal, 2009)	50.6	50.8	38.8
PF2 (Ghanty and Pal, 2009)	48.2	49.9	38.8
PF (Yang et al., 2011)	53.4	55.6	43.1
O (Taguchi and Gromiha, 2007)	51.0	46.9	36.3
AAC (Ding and Dubchak, 2001)	45.1	40.9	32.0
AAC+HXPZV$^+$ (Ding and Dubchak, 2001)	47.2	40.9	36.3
ACC (Dong et al., 2009)	68.0	85.9	66.4
consensus+PF1	64.6	75.2	52.7
consensus +PF2	64.7	74.9	51.1
consensus +PF	67.5	79.3	58.8
consensus +O	63.5	68.5	46.7
consensus +AAC	59.2	61.9	44.0
consensus +AAC+HXPZV	58.2	67.9	46.6
SSP (this paper)	42.2	58.3	50.2
PSSM+PHMM (this paper)	77.7	92.3	82.0
Mono-gram (Sharma et al., 2013b)	69.6	76.9	58.8
Bi-gram (Sharma et al., 2013b)	74.1	84.5	68.1
Tri-gram (Paliwal et al., 2014a, 2014b)	73.4	86.2	72.5
Alignment method (Lyons et al., 2014a)	74.7	90.2	74.0
PHMM-DP (this paper)	**82.7**	**92.9**	**85.6**

Step 1: Given training data, partition it randomly into n roughly equal segments.
Step 2: Hold out one segment as validation data and the remaining $n-1$ segments as learning data from the training data.
Step 3: Use the learning data for finding the model parameters.
Step 4: Use validation data to compute classification accuracy. Store the obtained classification accuracy.
Step 5: Repeat steps 1–4 n times.
Step 6: Evaluate average classification accuracy over n repetitions.

The performance of the proposed scheme has been gauged by comparing several other existing methods. The results are depicted in Table 1. In this experiment, the following feature sets are considered: PF1, PF2 (Ghanty and Pal, 2009), PF (Yang et al., 2011), Occurrence (O) (Taguchi and Gromiha, 2007), AAC, AAC+HXPZV (Ding and Dubchak, 2001), ACC (Dong et al., 2009), mono-gram and bi-gram (Sharma et al., 2013b), trigram (Paliwal et al., 2014a, 2014b) and amino acid alignment method (Lyons et al., 2014a). We have also updated the protein sequences to get the consensus sequence by using their corresponding PSSMs; i.e., each amino acid of a protein sequence is replaced by the amino acid that has the highest probability in PSSM. After this updating procedure, we have used the same feature extraction techniques (PF1, PF2, PF, O, AAC and AAC+HXPZV) again to obtain the recognition performance. In Table 1, we have placed the results for PSSM updated protein sequences (or the consensus sequence) in rows of consensus+*FET*, where *FET* is any feature extraction technique. We have also experimented using predicted secondary structure (SSP) and by combining profiles of PSSM and HMM (PSSM+PHMM). The

[1] In statistical prediction, the following three cross-validation methods are often used to examine a predictor for its effectiveness in practical application: independent dataset test, subsampling or k-fold crossover test, and jackknife test (Chou and Zhang, 1995). However, of the three test methods, the jackknife test is deemed the least arbitrary that can always yield a unique result for a given benchmark dataset as elaborated in Chou and Shen (2010) and demonstrated by Eqs.28-30 in Chou (2011). Accordingly, the jackknife test has been increasingly used and widely recognized by investigators to examine the quality of various predictors (see, e.g., Esmaeili et al., 2010, Chen et al., 2012, Hajisharifi et al., 2014, Chou et al., 2012). However, to reduce the computational time, we adopted the 10-fold cross-validation in this study as done by many investigators with SVM as the prediction engine.

highest recognition accuracy of a particular 10-fold cross-validation is mentioned in bold face.

The highest accuracies on DD, EDD and TG datasets are 82.7%, 92.9% and 85.6%, respectively. On DD dataset, we achieve an improvement of 8% compared to previous work of Lyons et al. (2014a, 2014b). Furthermore, on EDD we achieve 2.7% compared to the previous work (Lyons et al., 2014a, 2014b) and for TG we obtain 11.6% improvement when compared with Lyons et al. (2014a, 2014b). It should be noted that this improvement is achievement by using only sequential or evolutionary information of proteins. The performance can be improved further by incorporating other types of features for e.g. by using functional and physicochemical information. In general, the protein fold prediction accuracy by alignment method is around 2.7–11.6% higher than other methods.

Furthermore, we carried out paired *t*-test (David and Gunnink, 1997) on our achieved results compared to the highest reported results of the literature to study statistical significance of the prediction enhancement. The associated probability value for the paired *t*-test computed to be $p = 0.05$ which confirms the statistical significance of our reported enhancement in this study compared to the state-of-the-art results found in the literature for PFR. Note if $p \leq 0.05$, then the obtained results are considered to be significant and therefore it rejects the null hypothesis that the improvement is made purely by chance.

Next, in order to provide more statistical significance of our achieved results, we have carried out experiments to report analysis on precision, sensitivity and specificity. Further information regarding these three evaluation criteria can be found in Kurgan and Homaeian (2006), and Dehzangi et al. (2014a, 2014b). Sensitivity measures the ratio of correctly classified samples to the whole number of test samples for each class which are classified as correct samples and calculated as follows:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}} \times 100, \tag{4}$$

while TP represents true positive and FN represents false negative samples. Precision represents, how relevant the number of TP is to the whole number of positive prediction and is calculated as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \times 100. \tag{5}$$

where FP denotes false positive. Specificity, as other evaluation criterion used in this study measures the ratio of correctly rejected samples to the whole number of rejected test samples and is calculated as follows:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN}+\text{FP}} \times 100, \tag{6}$$

where TN denotes true negative. The sensitivity, specificity and precision are computed for each class and then average over all the classes are computed and reported in Figs. 2–4. There are other performance metrics used in the literature (see, e.g. Liu et al., 2015d, 2015a, 2015b, 2015c; Jia et al., 2015a, 2015b; Chen et al., 2015; Xiao et al., 2015). The set of metrics is valid only for the single-label systems. For the multi-label systems whose existence has become more frequent in system biology (Wu and Xiao, 2011a, 2011b, 2012) and system medicine (Xiao et al., 2013), a completely different set of metrics as defined in (Chou, 2013) is needed. In this paper, we have utilized accuracy measure to address this issue as performed in the literature (Sharma et al., 2013a, 2013b; Lyons et al., 2014a; Dehzangi et al., 2015a, 2015b, 2015c).

Fig. 2 shows the analysis on DD dataset, Fig. 3 depicts on EDD dataset and Fig. 4 on TG dataset. It can be observed from Figs. 2–4 that specificity is high for all the feature sets. However, precision

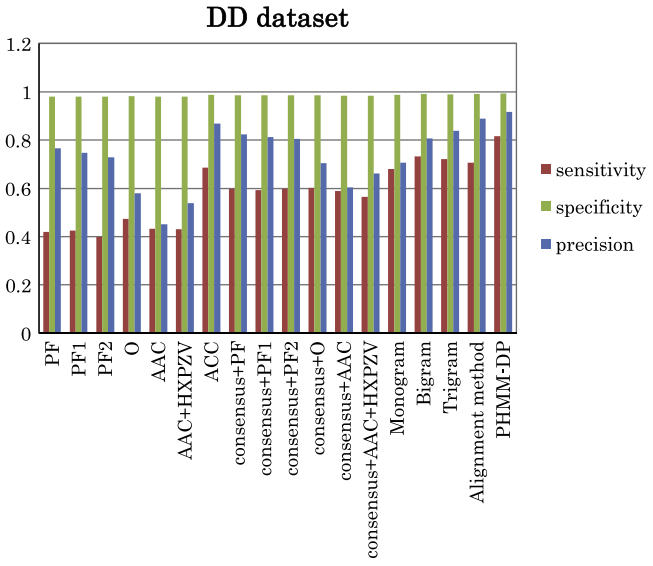

Fig. 2. Precision, sensitivity and specificity of all feature sets on DD dataset.

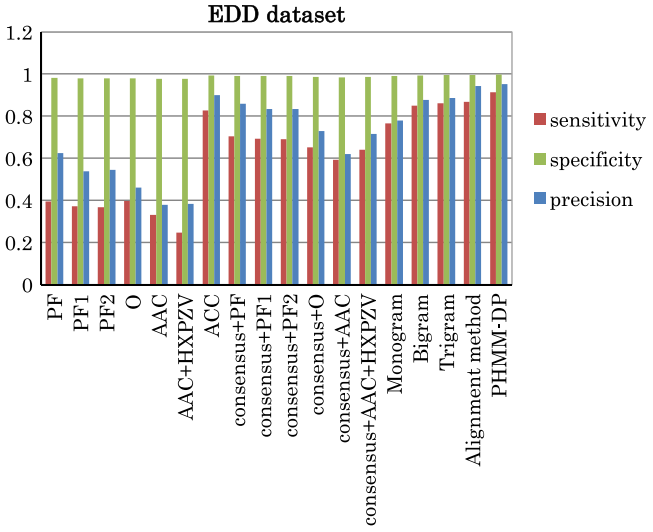

Fig. 3. Precision, sensitivity and specificity of all feature sets on EDD dataset.

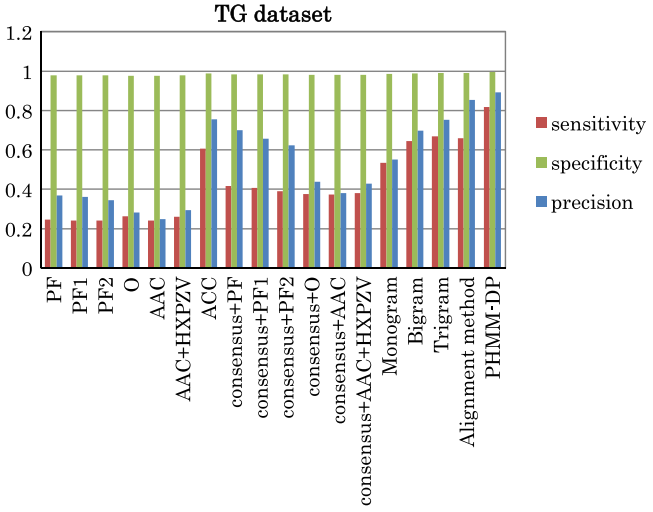

Fig. 4. Precision, sensitivity and specificity of all feature sets on TG dataset.

and sensitivity varies. For all the datasets, precision and sensitivity are quite promising for PHMM-DP method.

Since user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful models, simulated methods, or predictors (Chou and Shen, 2009; Lin and Lapointe, 2013; Chen et al., 2013; Lin et al., 2014; Guo et al., 2014; Liu et al., 2015), we shall make efforts in our future work to provide a web-server for the method presented in this paper to enhance the impact of the work (Chou, 2015).

7. Conclusion

In this study, we have presented a scheme of feature extraction for protein fold recognition problem. In this proposed scheme, we first extracted profile from HMM (we called as PHMM) of a protein sequence and then applied kernelized dynamic programming to find distance between two proteins. If the distance between proteins is low then there is a high probability that these two proteins belong to the same fold. By applying this phenomenon on three benchmark datasets we achieved reasonable performance. For DD, EDD and TG, we reported 82.7%, 92.9% and 85.6% prediction accuracies, respectively. The improvement is between 2.7% and 11.6%. It should be noted that this improvement is achieved by using sequential and/or evolutionary information of proteins only. It is possible to improve the performance further, if functional or physicochemical information is used. Some other statistical analyses have also been carried out like precision, specificity, sensitivity and paired *t*-test to check the significance of the results.

It is possible that by incorporating other features (e.g. features from physicochemical attributes), one might improve the performance of the presently available feature extraction models. However, the current extracted physicochemical-based features are not so informative to enhance the results. Therefore, in our further work we would explore the possibilities of utilizing physicochemical-based features with the evolutionary-based features to improve the classification accuracy of protein fold recognition (including other related problems like protein subcellular localization and secondary structure prediction).

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J.H., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped blast and psi-blast: a new generation of protein database search programs. Nucl. Acids Res. 17, 3389–3402.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer Science, NY.

Bouchaffra, D., Tan, J., 2006. Protein fold recognition using a structural Hidden Markov Model. In: Proceedings of the 18th International Conference on Pattern Recognition, pp. 186–189.

Cai, Y.D., 2006. Prediction of protease types in a hybridization space. Biochem. Biophys. Res. Commun. 339, 1015–1020.

Carlacci, L., 1991. Energetic approach to the folding of alpha/beta barrels. Proteins: Struct. Funct. Genet. 9, 280–295.

Chang, C.-C., Lin, C.-J., 2011. LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. 2 (3), pp. 27:1–27:27⟨http://www.csie.ntu.edu.tw/~cjlin/libsvm⟩.

Chen, K., Zhang, X., Yang, M.Q., Yang, J.Y., 2007. Ensemble of probabilistic neural networks for protein fold recognition. In: Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering (BIBE), pp. 66–70.

Chen, W., Lin, H., Feng, P.M., Ding, C., Zuo, Y.C., et al., 2012. iNuc-PhysChem: a sequence-based predictor for identifying nucleosomes via physicochemical properties. PLoS One 7, e47843.

Chen, W., Feng, P.M., Deng, E.Z., 2014. iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. Anal. Biochem. 462, 76–83.

Chen, W., Feng, P.M., Lin, H., 2013. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucl. Acids Res. 41, e68.

Chen, W., Feng, P., Ding, H., 2015. iRNA-Methyl: Identifying N6-methyladenosine sites using pseudo nucleotide composition. Anal. Biochem. 490, 26–33 (also, Data in Brief, 5, 2015, pp. 376–378).

Chinnasamy, A., Sung, W.K., Mittal, A., 2005. Protein structure and fold prediction using tree-augmented naive Bayesian classifier. J. Bioinf. Comp. Biol. 3 (4), 803–819.

Chmielnicki, W., Stapor, K., 2012. A hybrid discriminative-generative approach to protein fold recognition. Neurocomputing 75, 194–198.

Chou, K.C., 2001. Prediction of protein cellular attributes using pseudo amino acid composition. Proteins 43, 246–255 (erratum: 2001, vol. 44, 60).

Chou, K.C., Shen, H.B., 2010. Cell-PLoc: a package of web servers for predicting subcellular localization of proteins in various organisms (updated version: Cell-PLoc 2.0: an improved package of web-servers for predicting subcellular localization of proteins in various organisms. Nat. Sci. 2, 1090–1103. http://dx.doi.org/10.4236/ns.2010.210136 (Nature Protocols, 3, 2008, pp. 153–162).

Chou, K.C., Shen, H.B., 2009. Review: recent advances in developing web-servers for predicting protein attributes. Nat. Sci. 2, 63–92.

Chou, K.C., 2005. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21, 10–19.

Chou, K.C., 2011. Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review). J. Theor. Biol. 273, 236–247.

Chou, K.C., Wu, Z.C., Xiao, X., 2012. iLoc-Hum: using accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. Mol. Biosyst. 8, 629–641.

Chou, K.C., Shen, H.B., 2007. MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. Biochem. Biophys. Res. Commun. 360, 339–345.

Chou, K.C., Scheraga, H.A., 1982. Origin of the right-handed twist of beta-sheets of poly-L-valine chains. Proc. Natl. Acad. Sci. USA 79, 7047–7051.

Chou, K.C., 2015. Impacts of bioinformatics to medicinal chemistry. Med. Chem. 11, 218–234.

Chou, K.C., 2013. Some remarks on predicting multi-label attributes in molecular biosystems. Mol. Biosyst. 9, 1092–1100.

Chou, J.J., 1993. A formulation for correlating properties of peptides and its application to predicting human immunodeficiency virus protease-cleavable sites in proteins. Biopolymers 33, 1405–1414.

Chou, K.C., Zhang, C.T., 1995. Review: prediction of protein structural classes. Crit. Rev. Biochem. Mol. Biol. 30, 275–349.

David, H.A., Gunnink, J.L., 1997. The paired *t* test under artificial pairing. Am. Statistician 51 (1), 9–12.

Dehzangi, A.S., Phon-Amnuaisuk, O. Dehzangi, O., 2010a. Using random forest for protein fold prediction problem: an empirical study. J. Inf. Sci. Eng. 26 (6), 1941–1956.

Dehzangi, A., Amnuaisuk, S.P., Dehzangi, O., 2010b. Enhancing protein fold prediction accuracy by using ensemble of different classifiers. Aust. J. Intelli. Inf. Process. Syst. 26 (4), 32–40.

Dehzangi, A., Amnuaisuk, S.P., Manafi, M., Safa, S., 2010c. Using rotation forest for protein fold prediction problem: An empirical study. In: Proceedings of the 8th European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics, EvoBIO.

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

Dehzangi, A., Amnuaisuk, S.P., Ng, K.H., Mohandesi, E., 2009. Protein fold prediction problem using ensemble of classifiers. In: Proceedings of the 16th International Conference on Neural Information Processing, Part II, pp. 503–511.

Dehzangi, A., Karamizadeh, 2011. Solving protein fold prediction problem using fusion of heterogeneous classifiers. Inf. Int. Interdiscip. J. 14 (11), 3611–3622.

Dehzangi, A., Paliwal, K.K., Sharma, A., Dehzangi, O., Sattar, A., 2013a. A combination of feature extraction methods with an ensemble of different classifiers for protein structural class prediction problem. IEEE/ACM Trans. Comput. Biol. Bioinf. 10 (3), v564–v575.

Dehzangi, A., Paliwal, K.K., Lyons, J., Sharma, A., Sattar, A., 2013b. Exploring potential discriminatory information embedded in pssm to enhance protein structural class prediction accuracy. In: Proceeding of the Pattern Recognition in Bioinformatics, PRIB 2013, LNBI 7986, pp. 208–19.

Dehzangi, A., Paliwal, K.K., Lyons, J., Sharma, A., Sattar, A., 2013c. Enhancing protein fold prediction accuracy using evolutionary and structural features. In: Proceeding of the Pattern Recognition in Bioinformatics, PRIB 2013, LNBI 7986, pp. 196–207.

Dehzangi, A., Paliwal, K.K., Lyons, J., Sharma, A., Sattar, A., 2014a. Proposing a highly accurate protein structural class predictor using segmentation-based features. BMC Genom. 15 (Suppl 1), S2.

Dehzangi, A., Lyons, J., Sharma, A., Paliwal, K.K., Sattar, A., 2014b. A segmentation-based method to extract structural and evolutionary features for protein fold recognition. IEEE/ACM Trans. Comput. Biol. Bioinf. PP (99), 1–11.

Dehzangi, A., Sharma, A., Lyons, J., Paliwal, K.K., Sattar, A., 2015a. A mixture of physicochemical and evolutionary-based feature extraction approached for protein fold recognition. Int. J. Data Mining Bioinf. 11 (1), 115–138.

Dehzangi, A., Sohrabi, S., Lyons, J., Sharma, A., Paliwal, K.K., Sattar, A., 2015b. Gram-positive and gram-negative subcellular localization using rotation forest and physicochemical-based features. BMC Bioinf. 16 (Suppl 4:S1), 1–8.

Dehzangi, A., Heffernan, R., Sharma, A., Lyons, J., Paliwal, K.K., Sattar, A., 2015c. Gram-positive and gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou's general PseAAC. J. Theor. Biol. 364, 284–294.

Deschavanne, P., Tuffery, P., 2009. Enhanced protein fold recognition using a structural alphabet. Proteins: Struct. Funct. Bioinform. 76, 129–137.

Ding, C., Dubchak, I., 2001. Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics 17 (4), 349–358.
Ding, Y.S., Zhang, T.L., 2008. Using Chou's pseudo amino acid composition to predict subcellular localization of apoptosis proteins: an approach with immune genetic algorithm-based ensemble classifier. Patt. Recog. Lett. 29, 1887–1892.
Dong, Q., Zhou, S., Guan, J., 2009. A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation. Bioinformatics 25 (20), 2655–2662.
Dubchak, I., Muchnik, I., Kim, S.K., 1997. Protein folding class predictor for SCOP: approach based on global descriptors. In: Proceedings of the 5th International Conference on Intelligent Systems for Molecular Biology, pp. 104–107.
Esmaeili, M., Mohabatkar, H., Mohsenzadeh, S., 2010. Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. J. Theor. Biol. 263, 203–209.
Guo, S.H., Deng, E.Z., Xu, L.Q., Ding, H., 2014. iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. Bioinformatics 30, 1522–1529.
Ghanty, P., Pal, N.R., 2009. Prediction of protein folds: extraction of new features, dimensionality reduction, and fusion of heterogeneous classifiers. IEEE Trans. Nano Biosci. 8, 100–110.
Hajisharifi, Z., Piryaiee, M., Mohammad Beigi, M., Behbahani, M., Mohabatkar, H., 2014. Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test. J. Theor. Biol. 341, 34–40.
Heffernan, R., Paliwal, K.K., Lyons, J., Dehzangi, A., Sharma, A., Wang, J., Sattar, A., Yang, Y., Zhou, Y., 2015a. Improving Prediction of Secondary Structure, Local Backbone Angles, and Solvent Accessible Surface Area of Proteins by Iterative Deep Learning, Scientific Reports. vol. 5. Nature Publishing Group, UK, pp. 1–11, Article number: 11476.
Heffernan, R., Dehzangi, A., Lyons, J., Paliwal, K.K., Sharma, A., Wang, J., Sattar, A., Zhou, Y., Yang, Y., 2015b. Highly accurate sequence-based prediction of half-sphere exposures of amino acid residues in proteins. Bioinformatics . http://dx.doi.org/10.1093/bioinformatics/btv665.
Huang, J.T., Tian, J., 2006. Amino acid sequence predicts folding rate for middle-size two-state proteins. Proteins: Struct. Funct. Bioinf. 63 (3), 551–554.
Jia, J., Liu, Z., Xiao, X., 2015a. iPPI-Esml: an ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC. J. Theor. Biol. 377, 47–56.
Jia, J., Liu, Z., Xiao, X., Liu, B., 2015b. Identification of protein-protein binding sites by incorporating the physicochemical properties and stationary wavelet transforms into pseudo amino acid composition (iPPBS-PseAAC). J. Biomol. Struct. Dyn. . http://dx.doi.org/10.1080/07391102.2015.1095116
Kavousi, K., Moshiri, B., Sadeghi, M., Araabi, B.N., Moosavi-Movahedi, A.A., 2011. A protein fold classier formed by fusing different modes of pseudo amino acid composition via PSSM. Comput. Biol. Chem. 35 (1), 1–9.
Kecman, V., Yang, T., 2009. Protein fold recognition with adaptive local hyper plane Algorithm. In: Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB'09, pp. 75–78.
Klein, P., 1986. Prediction of protein structural class by discriminant analysis. BiochimBiophysActa 874, 205–215.
Krishnaraj, Y., Reddy, C.K., 2008. Boosting methods for protein fold recognition: an empirical comparison. In: Proceedings of the IEEE Int. Conf. on Bioinfor. and Biomed. pp. 393–396.
Kurgan, L.A., Homaeian, L., 2006. Prediction of structural classes for protein sequences and domains-impact of prediction algorithms, sequence representation and homology, and test procedures on accuracy. Pattern Recognit. 39, 2323–2343.
Kurgan, L.A., Zhang, T., Zhang, H., Shen, S., Ruan, J., 2008. Secondary structure-based assignment of the protein structural classes. Amino Acids 35, 551–564.
Lin, S.X., Lapointe, J., 2013. Theoretical and experimental biology in one. J. Biomed. Sci. Eng. 6, 435–442.
Lin, W.Z., Fang, J.A., Xiao, X., 2012. Predicting secretory proteins of malaria parasite by incorporating sequence evolution information into pseudo amino acid composition via grey system model. PLoS One 7, e49040.
Lin, W.Z., Fang, J.A., 2013. iLoc-Animal: a multi-label learning classifier for predicting subcellular localization of animal proteins. Mol. Biosyst. 9, 634–644.
Lin, H., Deng, E.Z., Ding, H., Chen, W., 2014a. iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. Nucl. Acids Res. 42, 12961–12972.
Liu, Z., Xiao, X., Qiu, W.R., 2015d. iDNA-Methyl: identifying DNA methylation sites via pseudo trinucleotide composition. Anal. Biochem. 474, 69–77 (also, Data in Brief, 4, 2015, pp. 87–89).
Liu, B., Fang, L., Liu, F., Wang, X., 2015a. Identification of real microRNA precursors with a pseudo structure status composition approach. PLoS One 10, e0121501.
Liu, B., Fang, L., Long, R., 2015b. iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. Bioinformatics . http://dx.doi.org/10.1093/bioinformatics/btv1604.
Liu, B., Fang, L., Wang, S., Wang, X., 2015c. Identification of microRNA precursor with the degenerate K-tuple or Kmer strategy. J. Theor. Biol. 385, 153–159.
Liu, T., Geng, X., Zheng, X., Li, R., Wang, J., 2012. Accurate prediction of protein structural class using AutoCovariance transformation of PSI-BLAST profiles. Amino Acids 42, 2243–2249.
Lyons, J., Biswas, N., Sharma, A., Dehzangi, A., Paliwal, K.K., 2014a. Protein fold recognition by alignment of amino acid residues using kernelized dynamic time warping. J. Theor. Biol. 354, 137–145.
Lyons, S., Dehzangi, A., Heffernan, R., Sharma, A., Paliwal, K.K., Sattar, A., Zhou, Y., Yang, Y., 2014b. Predicting backbone Cα angles and dihedrals from protein sequences by stacked sparse auto-encoder deep neural network. J. Comput. Chem. 35 (28), 2040–2046.
Maggiora, G.M., Scheraga, H.A., 1992. The role of loop-helix interactions in stabilizing four-helix bundle proteins. Proc. Natl. Acad. Sci. USA 89, 7315–7319.
Mamun, K.A., Sharma, A., 2014. Importance of computational intelligent in proteomics. J. Adv. Comput. Intell. Intell. Inform. 18 (4), 469–473.
Najmanovich, R., Kuttner, J., Sobolev, V., Edelman, M., 2000. Side-chain flexibility in proteins upon ligand binding. Proteins: Struct. Funct. Bioinform. 39 (3), 261–268.
Nemethy, G., Scheraga, H.A., 1984. Energetic approach to packing of a-helices: 2. General treatment of nonequivalent and nonregular helices. J. Am. Chem. Soc. 106, 3161–3170.
Nemethy, G., Pottle, M.S., Scheraga, H.A., 1985. The folding of the twisted beta-sheet in bovine pancreatic trypsin inhibitor. Biochemistry 24, 7948–7953.
Nemethy, G., Rumsey, S., Tuttle, R.W., Scheraga, H.A., 1986. Interactions between two beta-sheets: energetics of beta/beta packing in proteins. J. Mol. Biol. 188, 641–649.
Ohlson, T., Wallner, B., Elofsson, A., 2004. Profile-profile methods provide improved fold-recognition: a study of different profile-profile alignment methods. Proteins: Struct. Funct. Bioinform. 57, 188–197.
Paliwal, K.K., Sharma, A., Lyons, J., Dehzangi, A., 2014a. A tri-gram based feature extraction technique using linear probabilities of position specific scoring matrix for protein fold recognition. IEEE Trans. Nanobiosci. 13 (1).
Paliwal, K.K., Sharma, A., Lyons, J., Dehzangi, A., 2014b. Improving protein fold recognition using the amalgamation of evolutionary-based and structural-based information. BMC Bioinform. 15 (Suppl16), S12.
Paliwal, K.K., Sharma, A., 2011. Approximate LDA technique for dimensionality reduction in the small sample size case. J. Pattern Recognit. Res. 6 (2), 298–306.
Paliwal, K.K., Sharma, A., 2012. Improved pseudoinverse linear discriminant analysis method for dimensionality reduction. Int. J. Pattern Recognit. Artif. Intell. 26 (1), pp. 1250002-1–1250002-9.
Paliwal, K.K., Sharma, A., 2010. Improved direct LDA and its application to DNA gene microarray data. Pattern Recognit. Lett. 31 (16), 2489–2492.
Pottle, M., Scheraga, H.A., 1989. Energy of stabilization of the right-handed beta-alpha-beta crossover in proteins. J. Mol. Biol. 205, 241–249.
Remmert, M., Biegert, A., Hauser, A., Söding, J., 2011. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat. Methods 9 (2), 173–175. http://dx.doi.org/10.1038/nmeth.1818.
Rumsey, S., Tuttle, R.W., Scheraga, H.A., 1985. Interactions between an alpha-helix and a beta-sheet: energetics of alpha/beta packing in proteins. J. Mol. Biol. 186, 591–609.
Saini, H., Raicar, G., Sharma, A., Lal, S., Dehzangi, A., Rajeshkannan, A., Lyons, J., Biswas, N., Paliwal, K.K., 2014. Protein structural class prediction via k-separated bigrams using position specific scoring matrix. J. Adv. Comput. Intell. Intell. Inform. 18 (4), 474–479.
Saini, H., Raicar, G., Sharma, A., Lal, S., Dehzangi, A., Lyons, J., Paliwal, K.K., Imoto, S., Miyano, S., 2015a. Probabilistic expression of spatially varied amino acid dimers into general form of Chou's pseudo amino acid composition for protein fold recognition. J. Theor. Biol. 380 (7), 291–298.
Saini, H., Raicar, G., Dehzangi, A., Lal, S., Sharma, A., 2015b. Subcellular localization for gram positive and gram negative bacterial proteins using linear interpolation smoothing model. J. Theor. Biol. 386, 25–33.
Sharma, A., Paliwal, K.K., 2015a. A deterministic approach to regularized linear discriminant analysis. Neurocomputing 151, 207–214.
Sharma, A., Paliwal, K.K., 2015b. Linear discriminant analysis for the small sample size problem: an overview. Int. J. Mach. Learn. Cybern. 6 (3), 443–454.
Sharma, A., Dehzangi, A., Lyons, J., Imoto, S., Miyano, S., Nakai, K., Patil, A., 2014a. Evaluation of sequence features from intrinsically disordered regions for the estimation of protein function. PLoS One 9 (2), e89890.
Sharma, A., Paliwal, K.K., Imoto, S., Miyano, S., 2014b. A feature selection method using improved regularized linear discriminant analysis. Mach. Vis. Appl. 25 (3), 775–786.
Sharma, A., Paliwal, K.K., Imoto, S., Miyano, S., Sharma, V., Ananthanarayanan, R., 2014c. A feature selection method using fixed-point algorithm for DNA microarray gene expression data. Int. J. Knowl. Intell. Eng. Syst. 18 (1), 55–59.
Sharma, A., Paliwal, K.K., Dehzangi, A., Lyons, J., Imoto, S., Miyano, S., 2013a. A strategy to select suitable physicochemical attributes of amino acids for protein fold recognition. BMC Bioinform. 14, 233.
Sharma, A., Lyons, J., Dehzangi, A., Paliwal, K.K., 2013b. A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition. J. Theor. Biol. 320 (7), 41–46.
Sharma, A., Imoto, S., Miyano, S., Sharma, V., 2012a. Null space based feature selection method for gene expression data. Int. J. Mach. Learn. Cybern. 3 (4), 269–276. http://dx.doi.org/10.1007/s13042-011-0061-9.
Sharma, A., Imoto, S., Miyano, S., 2012b. A top-r feature selection algorithm for microarray gene expression data. IEEE/ACM Trans. Comput. Biol. Bioinform. 9 (3), 754–764.
Sharma, A., Paliwal, K.K., 2012c. A two-stage linear discriminant analysis for face-recognition. Pattern Recognit. Lett. 33 (9), 1157–1162.
Sharma, A., Imoto, S., Miyano, S., 2012d. A filter based feature selection algorithm using null space of covariance matrix for DNA microarray gene expression data. Curr. Bioinform. 7 (3), 289–294.

Sharma, A., Imoto, S., Miyano, S., 2012e. A between-class overlapping filter-based method for transcriptome data analysis. J. Bioinform. Comput. Biol. 10 (5), pp. 1250010-1–1250010-20.
Sharma, A., Paliwal, K.K., 2012f. A gene selection algorithm using Bayesian classification approach. Am. J. Appl. Sci. 9 (1), 127–131.
Sharma, A., Paliwal, K.K., 2012g. A new perspective to null linear discriminant analysis method and its fast implementation using random matrix multiplication with scatter matrices. Pattern Recognit. 45 (6), 2205–2213.
Sharma, A., Paliwal, K.K., Imoto, S., Miyano, S., 2012h. Principal component analysis using QR decomposition. Int. J. Mach. Learn. Cybern. . http://dx.doi.org/10.1007/s13042-012-0131-7
Sharma, A., Koh, C.H., Imoto, S., Miyano, S., 2011. Strategy of finding optimal number of features on gene expression data. Electron. Lett. 47 (8), 480–482.
Sharma, A., Paliwal, K.K., 2010a. Regularisation of eigenfeatures by extrapolation of scatter-matrix in face-recognition problem. Electron. Lett., IEE 46 (10), 682–683.
Sharma, A., Paliwal, K.K., 2010b. An improved nearest centroid classifier with shrunken distance measure for null LDA method on cancer classification problem. Electron. Lett., IEE 46 (18), 1251–1252.
Sharma, A., Paliwal, K.K., 2008a. A gradient linear discriminant analysis for small sample sized problem. Neural Process. Lett. 27 (1), 17–24.
Sharma, A., Paliwal, K.K., 2008b. Cancer classification by gradient LDA technique using microarray gene expression data. Data Knowl. Eng. 66 (2), 338–347.
Sharma, A., Paliwal, K.K., 2008c. Rotational linear discriminant analysis for dimensionality reduction. IEEE Trans. Knowl. Data Eng. 20 (10), 1336–1347.
Sharma, A., Paliwal, K.K., 2007. Fast principal component analysis using fixed-point algorithm. Pattern Recognit. Lett. 28 (10), 1151–1155.
Sharma, A., Paliwal, K.K., Onwubolu, G.C., 2006a. Class-dependent PCA, LDA and MDC: a combined classifier for pattern classification. Pattern Recognit. 39 (7), 1215–1229.
Sharma, A., Paliwal, K.K., 2006b. Subspace independent component analysis using Vector Kurtosis. Pattern Recognit. 39 (11), 2223–2226.
Sharma, A., Paliwal, K.K., 2006c. Rotational linear discriminant analysis using bayes rule for dimensionality reduction. J. Comput. Sci. 2 (9), 754–757.
Sharma, A., Paliwal, K.K., Onwubolu, G.C., 2005. Pattern classification: an improvement using VQ and PCA based techniques. Am. J. Appl. Sci. 2 (10), 1445–1455.
Sharma, R., Dehzangi, A., Lyons, J., Paliwal, K.K., Tsunoda, T., Sharma, A., 2015c. Predict Gram-positive and Gram-negative subcellular localization via incorporating evolutionary information and physicochemical features into Chou's general PseAAC. IEEE Trans. NanoBiosci. . http://dx.doi.org/10.1109/TNB.2015.2500186
Shamim, M.T.A., Anwaruddin, M., Nagarajaram, H.A., 2007. Support vector machine-based classification of protein folds using the structural properties of amino acid residues and amino acid residue pairs. Bioinformatics 23 (24), 3320–3327.
Shen, H.B., Chou, K.C., 2006. Ensemble classier for protein fold pattern recognition. Bioinformatics 22, 1717–1722.
Shen, H.B., Chou, K.C., 2008. PseAAC: a flexible web-server for generating various kinds of protein pseudo amino acid composition. Anal. Biochem. 373, 386–388.
Shen, H.B., 2009. Predicting protein fold pattern with functional domain and sequential evolution information. J. Theor. Biol. 256, 441–446.
Taguchi, Y.-h, Gromiha, M.M., 2007. Application of amino acid occurrence for discriminating different folding types of globular proteins. BMC Bioinform. 8, 404.
Valavanis, I.K., Spyrou, G.M., Nikita, K.S., 2010. A comparative study of multi-classification methods for protein fold recognition. Int. J. Comput. Intell. Bioinform. Syst. Biol. 1 (3), 332–346.
Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. Springer-Verlag, New York.
Wang, Z.Z., Yuan, Z., 2000. How good is prediction of protein-structural class by the component-coupled method? Proteins 38, 165–175.
Wu, Z.C., Xiao, X., 2011a. iLoc-Euk: a multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins. PLoS One 6, e18258.
Wu, Z.C., Xiao, X., 2011b. iLoc-Plant: a multi-label classifier for predicting the subcellular localization of plant proteins with both single and multiple sites. Mol. Biosyst. 7, 3287–3297.
Wu, Z.C., Xiao, X., 2012. iLoc-Hum: Using accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. Mol. Biosyst. 8, 629–641.
Xiao, X., Min, J.L., Lin, W.Z., Liu, Z., 2015. iDrug-Target: predicting the interactions between drug compounds and target proteins in cellular networking via the benchmark dataset optimization approach. J. Biomol. Struct. Dyn. 33, 2221–2233.
Xiao, X., Wang, P., Lin, W.Z., 2013. iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types. Anal. Biochem. 436, 168–177.
Xu, Y., Wen, X., Wen, L.S., Wu, L.Y., 2014. iNitro-Tyr: prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition. PLoS One 9, e105018.
Yan, R., Xu, D., Yang, J., Walker, S., Zhang, Y., 2013. A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction. Sci. Rep. 3.
Yang, T., Kecman, V., Cao, L., Zhang, C., Huang, J.Z., 2011. Margin-based ensemble classifier for protein fold recognition. Expert Syst. Appl. 38, 12348–12355.
Ying, Y., Huang, K., Campbell, C., 2009. Enhanced protein fold recognition through a novel data integration approach. BMC Bioinform. 10 (1), 267.
Zakeri, P., Jeuris, B., Vandebril, R., Moreau, Y., 2014. Protein fold recognition using geometric kernel data fusion. Bioinformatics . http://dx.doi.org/10.1093/bioinformatics/btu118.
Zhang, H., Zhang, T., Gao, J., Ruan, J., Shen, S., Kurgan, L.A., 2010. Determination of protein folding kinetic types using sequence and predicted secondary structure and solvent accessibility. Amino Acids, 1–13.
Zhang, T.L., Ding, Y.S., Chou, K.C., 2008. Prediction protein structural classes with pseudo amino acid composition: approximate entropy and hydrophobicity pattern. Theor. Biol. 250, 186–193.