Gene Reports 12 (2018) 261–275

Contents lists available at ScienceDirect

Gene Reports

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

Enhanced Artificial Neural Network for Protein Fold Recognition and Structural Class Prediction

P. Sudha, D. Ramyachitra*, P. Manikandan

Department of Computer Science, Bharathiar University, Coimbatore 641 046, India

ARTICLE INFO

Keywords:
Protein structure prediction
Protein Fold Recognition
Structural Class Prediction
Artificial Neural Network
Enhanced Artificial Neural Network

ABSTRACT

In Bioinformatics Protein Fold Recognition (PFR) and Structural Class Prediction (SCP) is a significant problem in predicting protein with a three dimensional structure. Extraction of valuable features of protein that consists of 20 amino acids to acquire more desirable classifiers is fundamental to this PFR and SCP. Feature extraction technique predominantly exploits Forward Consecutive Search Scheme (FCS) that supplements syntactical-based, evolutionary-based and physicochemical-based information. In this research work, a classifier known as Enhanced Artificial Neural Network (ANN) is employed as it is more efficient than Forward Consecutive Search scheme in order to improve the performance of PFR and SCP. The Enhanced ANN algorithm is an improved version of Artificial Neural Network when compared with various existing algorithms such as Support Vector Machine (SVM), ANN, K-Nearest Neighbor (KNN) and the Bayesian. The experiments are conducted on four datasets namely DD, EDD, TG and RDD. Ultimately, the statistical imputation of Enhanced ANN algorithm hypothesizes gives better results than other algorithms to improve the performance of PFR and SCP.

1. Introduction

Proteins are the components which play important roles in the activities of organisms. Protein's function depends on the interactions with other proteins and its folding. Mismatch protein folding usually leads to changing in properties of the protein, which causes some diseases (Hashemi et al., 2009). To acquire knowledge about the protein function, interactions and regulations the prediction of protein structural classes is extremely useful (Jian-Yi Yang et al., 2010). To Increase the prediction accuracy of secondary structure and also to reduce the testimony of hunting scope in three dimensional structure predictions, the mastery of the structural class is helpful (Mohammad and AliYaghoubi, 2016). The SCP has become one of the most important features for characterizing the overall folding type of a protein in protein research. The first definition of protein structural class was introduced by Levitt and Chothia in 1976 and the globular proteins are normally classified into four structural classes such as (i) the all-α class consists of only little amount of strands, (*ii*) *the* all-β class consists of only little amount of helices, (iii) the α/β class consists of helices and almost all parallel strands, and $\alpha + \beta$ class consists of helices and almost all anti-parallel strands (Levitt and Chothia, 1976). Basically, the structural class of protein prediction from 20 amino acids is a significant task in the field of molecular biology.

Proteins with unique length and similarities to be a part of the same fold having the identical significant protein secondary structure in the identical arrangement with the identical topology certainly they have a regular origin of evolutionary (Yang et al., 2011). PFR is used to model the proteins which have the similar fold as proteins of known structure, but do not have homologous proteins with known protein structure. PFR is the acquiring of three dimensional structure of the protein sequences independent from the sequence identities (Ding and Dubchak, 2001). PFR and SCP are prohibited as a transitional step for identifying the protein three dimensional structures. The PFR and SCP consist of two main concepts such as feature extraction techniques and classification techniques. The main goal of PFR and SCP is to allocate the novel protein sequence to a particular fold type and to a particular class type. Computational approaches considered more attention over the years due to the expense and the time involved in identifying the three dimensional structure of protein by using X-ray crystallography and Nuclear Magnetic Resonance (NMR) (Ibrahim and Abadeh, 2017).

Many feature extraction techniques have been developed for protein Structural Class Prediction such as syntactical and physicochemical

Abbreviations list: PFR, Protein Fold Recognition; SCP, Structural Class Prediction; ANN, Artificial Neural Network; SVM, Support Vector Machine; KNN, K-Nearest Neighbor; FCS, Forward Consecutive Search scheme; NMR, Nuclear Magnetic Resonance; PF, Pairwise Frequency; MD-SFS, Multi Dimensional- Successive Feature Selection; PNNE, Probabilistic Neural Network Ensemble; RNNs, Recurrent and Recursive Artificial Neural Networks

* Corresponding author.
E-mail address: ramyachitra@buc.edu.in (D. Ramyachitra).

https://doi.org/10.1016/j.genrep.2018.07.012
Received 25 January 2018; Received in revised form 19 June 2018; Accepted 15 July 2018
Available online 17 July 2018

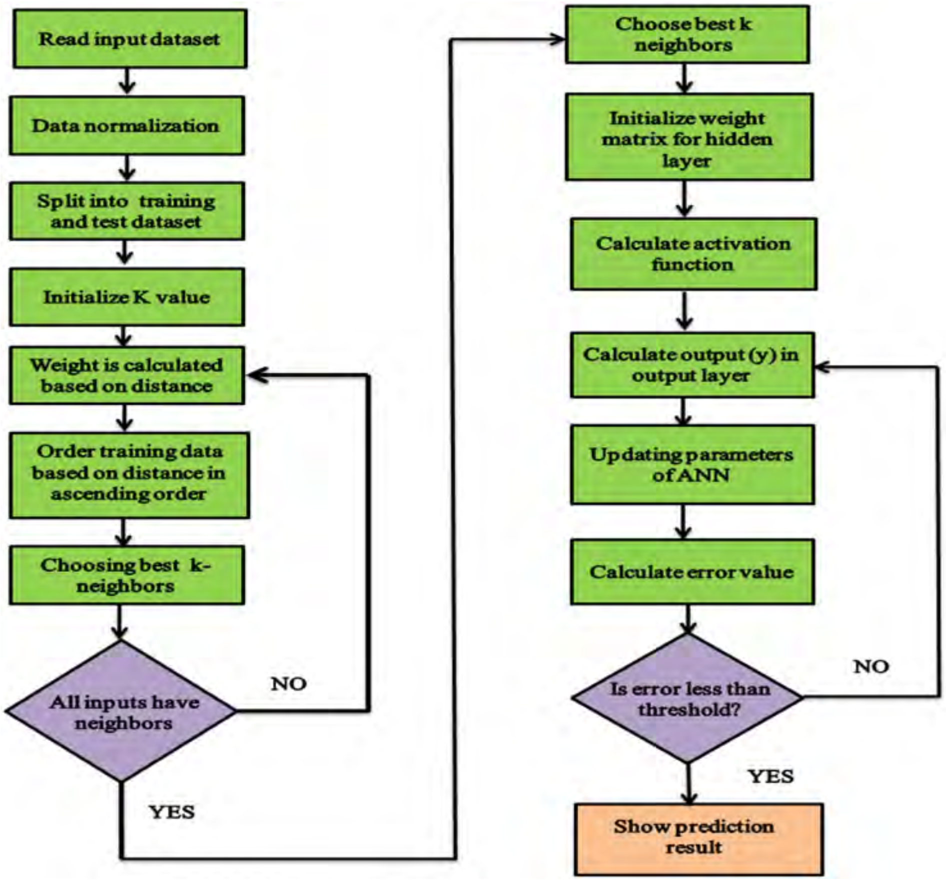

Read input dataset
Data normalization
Split into training and test dataset
Initialize K value
Weight is calculated based on distance
Order training data based on distance in ascending order
Choosing best k-neighbors
All inputs have neighbors
NO
YES
Choose best k neighbors
Initialize weight matrix for hidden layer
Calculate activation function
Calculate output (y) in output layer
Updating parameters of ANN
Calculate error value
Is error less than threshold?
NO
YES
Show prediction result

Fig. 1. Generalized flow chart of enhanced ANN.

based features (Dehzangi et al., 2013a, 2013b; Dubchak et al., 1997; Huang and Tian, 2006), Pairwise frequency (PF) carried out by (Yang et al., 2011), PF1 and PF2 (Ghanty and Pal, 2009), Bigram feature (Hayat et al., 2014a, 2014b; Sharma et al., 2013; Saini et al., 2014), Trigram (Lyons et al., 2016; Paliwal et al., 2014a), Separated dimmers (Saini et al., 2015), Pseudo-Amino Acid Composition (Chou, 2001), feature selection techniques such as syntactical, evolutionary and physicochemical-based features (Guyon and Elisseeff, 2003; Sharma et al., 2012, 2012b, 2013; Raicar et al., 2016; Cormen et al., 1990; Dehzangi and Phon-Amnuaisuk, 2011). Also several computational classifiers are used for protein Structural Class Prediction such as SVM (Hae-Jin et al., 2004), KNN (Shen and Chou, 2006; Ding and Zhang, 2013), ANN (Raicar et al., 2016), Bayesian classifiers (Chinnasamy et al., 2005), Hidden Markov Model (Bouchaffra and Tan, 2006), Ensemble classifiers (Dehzangi et al., 2009, 2010a, 2010b, Dehzangi and Karamizadeh, 2011; Shen and Chou, 2006; Yang et al., 2011), Hierarchical classification (Sharma et al., 2016) and Bayesian decision rule (Wang and Yuan, 2000) for both PFR and SCP. These techniques have many disadvantages such as poor performance when the dataset is large and sometimes may lead to over fitting and data loss or complexity, difficulties in debug and complex optimal design.

To overcome the drawbacks of the existing classification technique a new approach called Enhanced ANN have been developed. This approach focuses on improving the performance of PFR and SCP accuracies using physico-chemical properties of amino acids, which overcome the drawback of classification techniques. Hence our proposed algorithm finds the overlapping communities and works with weighted network. To evaluate the performance of the proposed algorithm with the existing techniques four benchmark datasets namely, DD (Murzin et al., 1995; Ding and Dubchak, 2001; Alok Sharma, 2013), EDD (Dong et al., 2009; Alok Sharma, 2013), TG (Taguchi and Gromiha, 2007; Alok Sharma, 2013) and RDD (Xia et al., 2017) are used.

In this paper, the input for our proposed algorithm have been derived from the feature extraction technique, namely Forward Consecutive Search scheme (FCS) that combines physico-chemical based feature by syntactical based or evolutionary based feature. Then the proposed algorithm, namely Enhanced Artificial Neural Network is compared with four other popular classification algorithms, namely SVM, KNN, ANN and Bayesian for same datasets. Finally, the results are compared using the performance metrics to measure the performance of the proposed algorithm that indicates that it performs very efficiently for both PFR and SCP with high accuracy when compared to the other algorithms. The remaining sections of the paper are organized as follows. Section 2 describes the related works of the existing techniques. The methodology has been explained in the Section 3. Section 4 shows the experimental results and discussion. The biological significance is given in Section 5. Finally the conclusion and future is given in Section 6.

2. Related work

FCS is used to select physico-chemical attributes for PFR and SCP (Raicar et al., 2016). A novel mixture of physico-chemical and evolutionary based feature extraction methods that depend on the concepts of segmented distribution and density is developed (Dehzangi et al., 2014b). The Hidden Markov- Support Vector Machines (HM-SVMs) classifier is introduced to predict the residues that participate in a beta sheet with hydrogen bonds between adjacent sheets in structural class (Blaise Gassend et al., 2006). The feature extraction techniques named tri-grams, computed directly from Position Specific Scoring Matrices have a problem of time complexity due to its iterative process (Paliwal

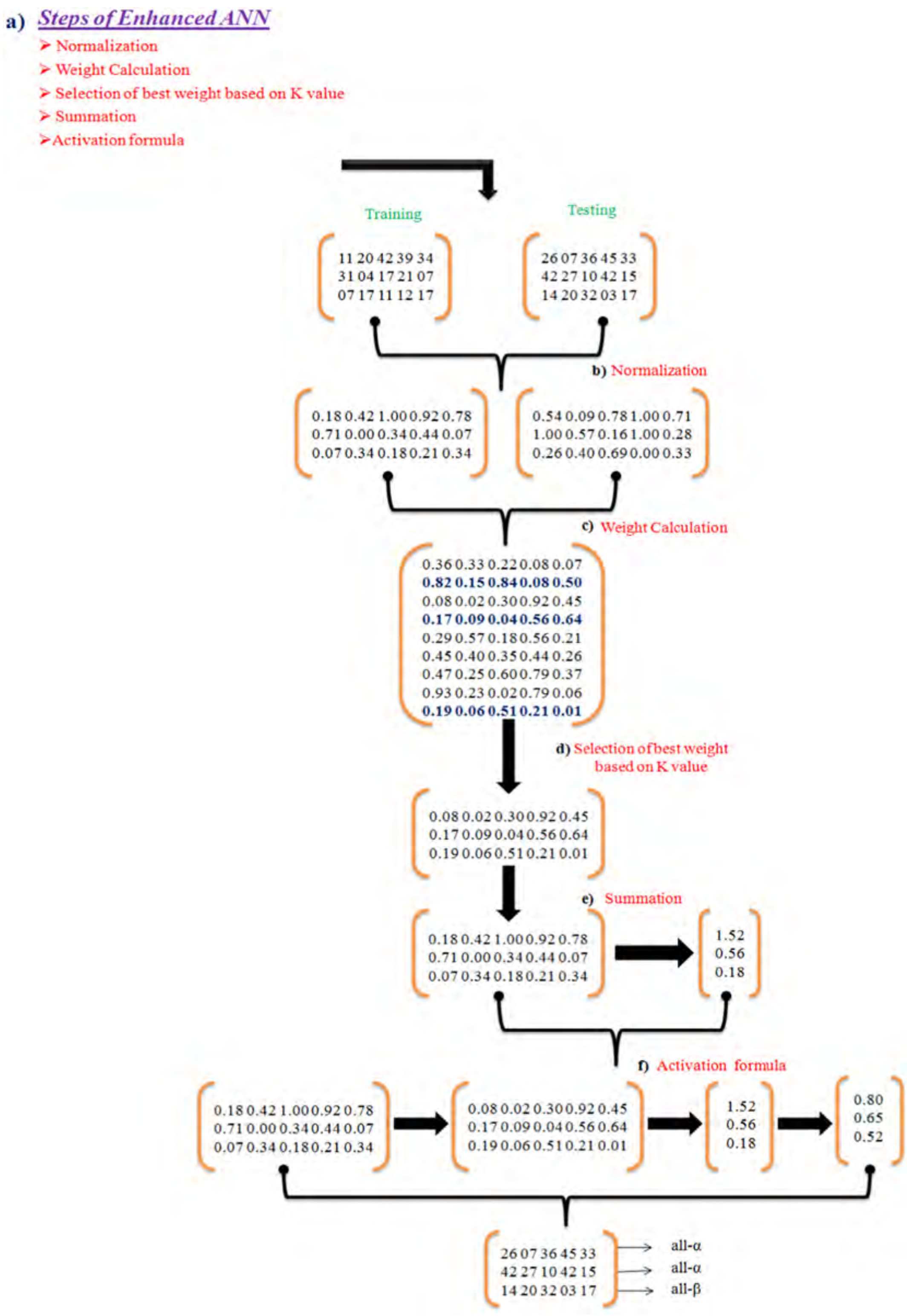
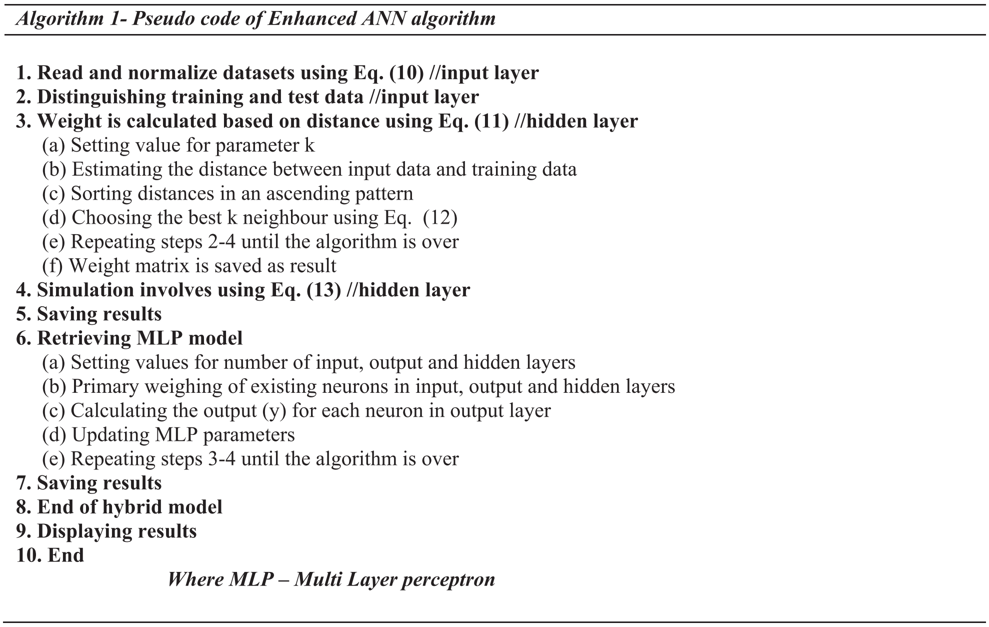


Fig. 2. Flow chart of enhanced ANN.

et al., 2014a). The Multi Dimensional-Successive Feature Selection (MD-SFS) method is generated (Alok Sharma, 2013) to select physico-chemical attributes of the amino acids for Protein Fold Recognition determination.

The bi-gram based feature extraction is calculated from Position Specific Scoring Matrices directly for Protein Fold Recognition and it has a problem of limited accuracy (Sharma et al., 2013). The feature selection method is recommended (Ghanty and Pal, 2009) to select few features named as trio AACs and trio potential for protein fold and structural class determination. SVM is projected as a base classifier (Dubchak et al., 2001) for multi-class Protein Fold Recognition based on two new methods named as unique one-against-others and the all-against-all (Wang and Liu, 2004) proposed kernel method-SVM classifier for protein fold and structural class recognition. This kernel model was quite sensitive and high complexity.

Protein Fold Recognition is solved by multi-class support vector



```
Algorithm 1- Pseudo code of Enhanced ANN algorithm

1. Read and normalize datasets using Eq. (10) //input layer
2. Distinguishing training and test data //input layer
3. Weight is calculated based on distance using Eq. (11) //hidden layer
    (a) Setting value for parameter k
    (b) Estimating the distance between input data and training data
    (c) Sorting distances in an ascending pattern
    (d) Choosing the best k neighbour using Eq.  (12)
    (e) Repeating steps 2-4 until the algorithm is over
    (f) Weight matrix is saved as result
4. Simulation involves using Eq. (13) //hidden layer
5. Saving results
6. Retrieving MLP model
    (a) Setting values for number of input, output and hidden layers
    (b) Primary weighing of existing neurons in input, output and hidden layers
    (c) Calculating the output (y) for each neuron in output layer
    (d) Updating MLP parameters
    (e) Repeating steps 3-4 until the algorithm is over
7. Saving results
8. End of hybrid model
9. Displaying results
10. End
                    Where MLP – Multi Layer perceptron
```

Fig. 3. Pseudo code for Enhanced ANN algorithm.

machine classifier (Minh et al., 2003) based on sequence-derived features and having the difficulty in optimal design. The SVM classifier applied to feature extraction method based on the physical and physicochemical properties of amino acids for PFR and SCP having a problem of over fitting (Dehzangi and Phon-Amnuaisuk, 2011). A Probabilistic Neural Network Ensemble (PNNE) model is developed by (Chen et al., 2007) for multi-class PFR problem. It is evaluated by two datasets containing 27 SCOP folds, but it would computationally intensify to train on very large datasets. The Kohonen's self–organizing neural network is recommended to predict the structural classes of proteins that have difficulties in debugging (Cai and Zhou, 2000).

In (Baldi and Pollastri, 2003) the new approach is introduced, namely Recurrent and Recursive Artificial Neural Networks (RNNs). The architectures for large-scale applications that are derived from the state-of-the-art predictors for protein structural features such as secondary structure (1D) still have an issue of high processing time. The pair of neural network-based algorithms are given to predict tertiary structural class and the secondary structure presented with non-homologous protein (Chandonia and Karplus, 1995). Protein Structural class is predicted by neural network classifier with amino acid composition and hydrophobic pattern frequency information as input to derive the Structural Class Prediction that has a poor performance for small datasets (Metfessel and Saurugger, 1993). An Optimized Evidence Theoretic K -Nearest Neighbor (OET-KNN) classifier is discovered by (Shen and Chou, 2005) which is combined through a weighted voting of features to give a final determination for classifying a query protein for recognition with poor performance when the datasets is large.

Adaptive Fuzzy r-Nearest Neighbor (AFK-NN) method is used to predict enzyme subfamily class without overlap that is not suitable for large datasets (Wen-Lin Huang et al., 2007). The new method, namely *Tree-Augmented Bayesian* Networks (TAN) based on the theory of learning Bayesian networks structure probabilities determines the significance of each feature for each class that helped in protein structure with data complexity (Chinnasamy et al., 2005). A novel method called Hierarchical Ensemble of Bayesian Classifiers (HensBC) is deployed to predict protein subcellular location used only for small datasets (Bulashevska and Eils, 2006). The Bayesian model was introduced (Li et al., 2014) based on the knob-socket model of protein packing in secondary structures that have a limitation of data loss. From the literature review, it is observed that many feature extraction and classification techniques can be used for the PFR and SCP problems.

Especially for SCP some of the feature extraction techniques are handled such as tri-grams (Tao et al., 2015), Primary and Secondary sequences (Nanni et al., 2014), Chaos Game Representation (CGR) (Zhang et al., 2016), overlapping segmented distribution and auto correlation-based (Dehzangi et al., 2013a, 2013b), Segmented distribution and segmented auto covariance (Dehzangi et al., 2014a), Hybrid feature spaces of Bayes of multi profile and probability of bi-gram (Hayat et al., 2014a, 2014b), Chou's pseudo amino acid composition and wavelet denoising (PseAAC-WD) (Yu et al., 2017), 11- dimensional feature vector (Liu and Jia, 2010), Evolutionary collocation based sequence representation (Chen et al., 2008) and classification algorithms such as, Ensemble classifiers (Nanni et al., 2014, Nanni, 2006) and SCPRED (Kurgan et al., 2008).

Particularly for PFR part of feature extraction techniques pre-owned such as Hidden Markov Model by Dynamic Programming (PHMM-DP) (Lyons et al., 2016), k-Amino Acid Pair (k-AAP) (Paliwal et al., 2014b), Hydrophobicity and pair-wise amino acids (Pal and Chakraborty, 2003), amino acid alignment feature extraction method is computed based on Kernalized dynamic time warping (Lyons et al., 2014) and Classification techniques namely Sparse representation based classification (SRC) (Yan et al., 2017), Random forest (Dehzangi et al., 2010a, 2010b), Heterogeneous classifier (Dehzangi and Karamizadeh, 2011), Ensemble classifiers (Dehzangi et al., 2009), Single probabilistic multi-class multikernel machine (Damoulas and Girolami, 2008).

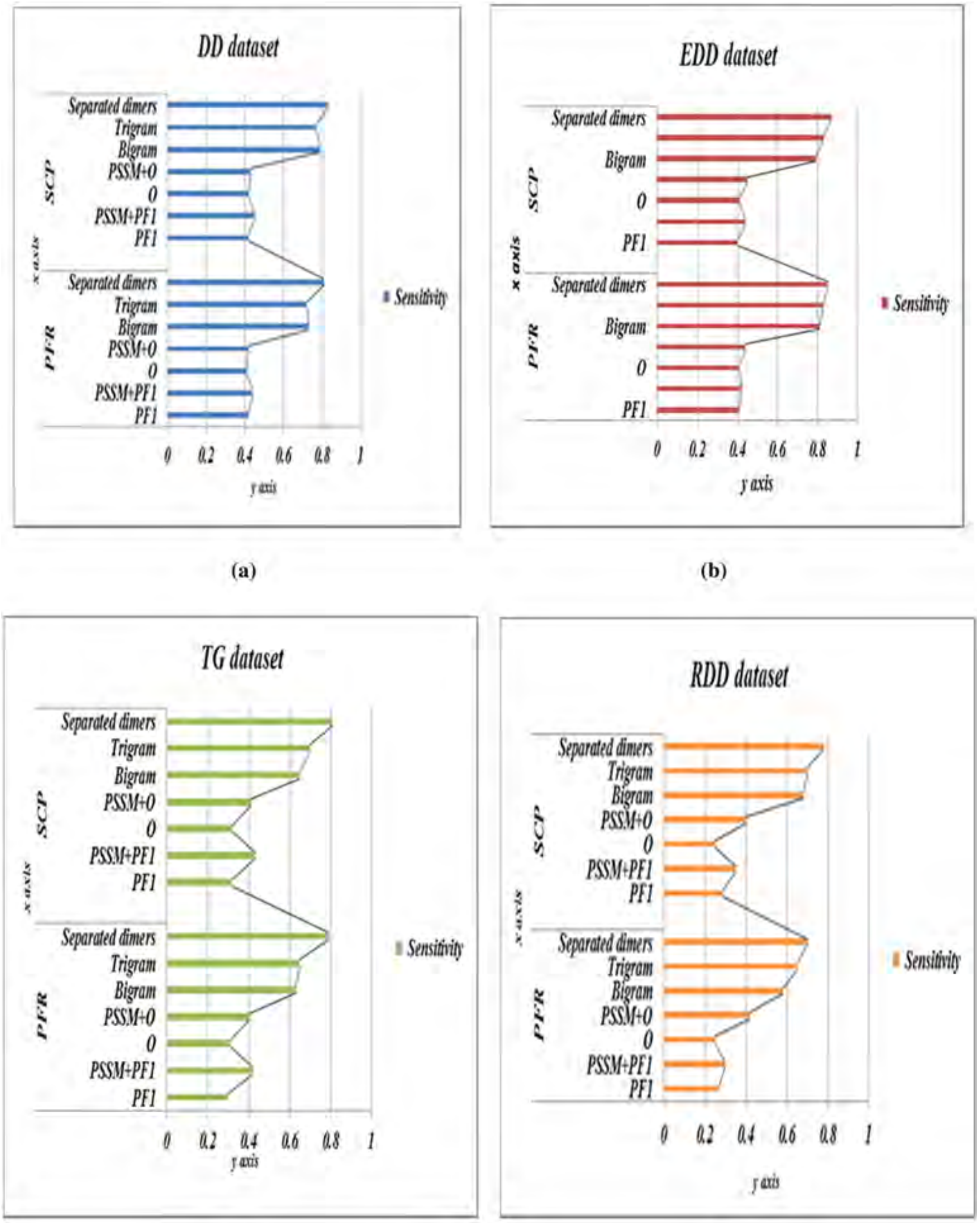
3. Methodology

In this research the existing algorithms such as SVM, ANN, KNN, Bayesian and the proposed approach, namely Enhanced Artificial Neural Network algorithm are compared to predict the PFR and SCP. From the results it is inferred that the Enhanced ANN algorithm provides better results than the existing algorithms.

3.1. Syntactical and evolutionary based features

Occurrence (**O**) is the frequency of amino acids in a protein sequence that produce 20 features

$$\mathrm{n}_i = (n_{i\,1}, n_{i\,2},..., n_{ij},..., n_{i\,20}) \qquad (1)$$



(a)

(b)

(c)

(d)

Fig. 4. Sensitivity of all feature sets for PFR and SCP on DD, EDD, TG and RDD datasets.

where n $_i$ constitutes the number of amino acids of each type in i th protein (Taguchi and Gromiha, 2007).

Pairwise frequency (PF1) is the frequency of pairs of amino acids separated by one residue in a protein sequence that produce 400 features (Ghanty and Pal, 2009)

$$PP1\,(\mathrm{a,b}) = \frac{1}{R}\sum_{x_i=a,x_j=b,j=i+2} e^{\frac{h(x_i)+h(x_j)}{M}},\ \mathrm{a,\ b} = 1,\ 2,\ \ldots,20 \tag{2}$$

where a and b represent two residues; h (a) is the hydrophobicity of residue a and M is a constant; the role of M is to scale the value of hydrophobicity so that numerical overflow is avoided; R is a normalizing constant computed by.

$$\mathrm{R} = \underset{S\,\in\,X_{Tr}}{\dot{max}}\left\{\underset{a,\ b}{max}\sum_{\mathbf{x_i=a,x_j=b,j=i+2}} e^{\frac{h(x_i)\,+\,h(x_j)}{M}}\right\} \tag{3}$$

where (X_{Tr}) is the entire training set.

A prefix of PSSM + before a (O, PF1) indicates that the feature was extracted from consensus protein sequences rather than the raw protein sequences with the amino acids in the PSSM consist of high probability (Raicar et al., 2016).

The Bigram feature is based on PSSM matrix that represents the transitional probabilities from one amino acid to another and also produces 400 features.

$$\mathrm{B}_{m,n} = \sum\nolimits_{i=1}^{L-1} \mathrm{p}_{i,m}\mathrm{p}_{\mathrm{i+1,n}},\ where\ 1 \le m \le 20\ and\ 1 \le n \le 20 \tag{4}$$

where B indicate the bigram occurrence matrix; P is the matrix of a given protein that represents PSSM with L rows and 20 columns; L is the length of the primary sequence and also an element at ith-row and jth-column is denoted by $p_{i,\,j}$ which is the relative probability of the primary protein sequences (Sharma et al., 2013). The feature vector for bigram is computed by

$$\mathrm{F} = [\mathrm{B}_{1,1}, \mathrm{B}_{1,2},\ldots, \mathrm{B}_{1,20}, \mathrm{B}_{2,1},\ldots, \mathrm{B}_{2,20},\ldots, \mathrm{B}_{20,1},\ldots, \mathrm{B}_{20,20}]^T \tag{5}$$

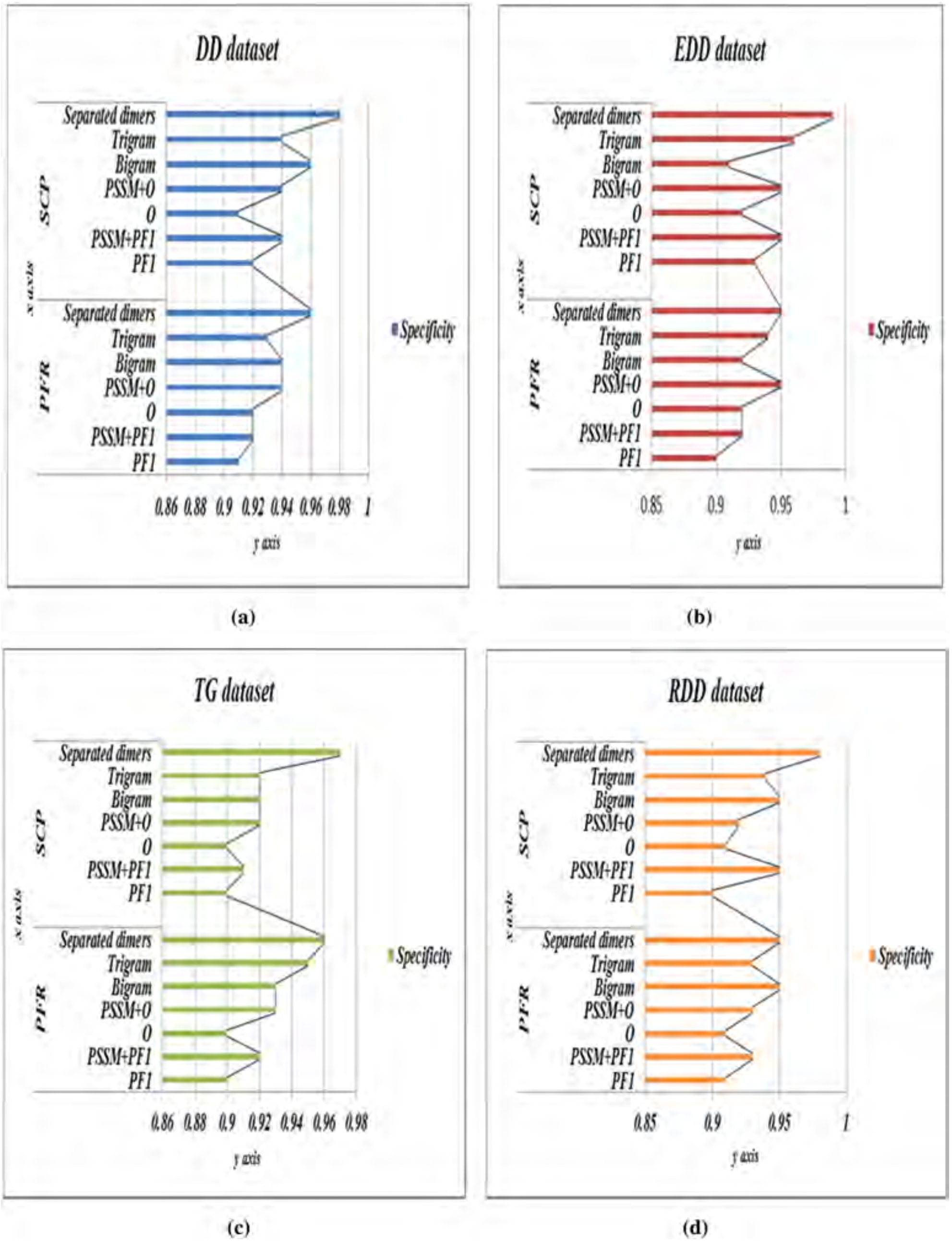


Fig. 5. Specificity of all feature sets for PFR and SCP on DD, EDD, TG and RDD datasets.

where F determines as bigram feature vector, B is the bigram occurrence matrix and T indicates the transpose of the vector.

Trigram is based on PSSM matrix that represents the transitional probabilities with triplets from one amino acids to another and also produces 8000 features (Paliwal et al., 2014a, 2014b).

Separated dimers consist of amino acid dimers with probabilistic expressions that have spatial separations from k = 1,2, …,K, where K, denoted as the upper bound of k that produces 400 × k features

$$F(k) = [F_{1,1}(k), F_{1,2}(k),..., F_{1,20}(k), F_{2,1}(k),...F_{20,20}(k)] \tag{6}$$

where k depicts the spatial distance between the dimers; P determines the PSSM matrix representation of a protein sequence and also it consists of L (where L is the length of the protein sequence) rows with 20 columns. The mth amino acid ($1 \le m \le 20$) to nth amino acid ($1 \le n \le 20$) of probabilistic expression can be computed using Eq. (7) (Saini et al., 2015)

$$F_{m,n}(k) = \sum_{i=0}^{L-k} P_{i,m}P_{\,i+k,n} \tag{7}$$

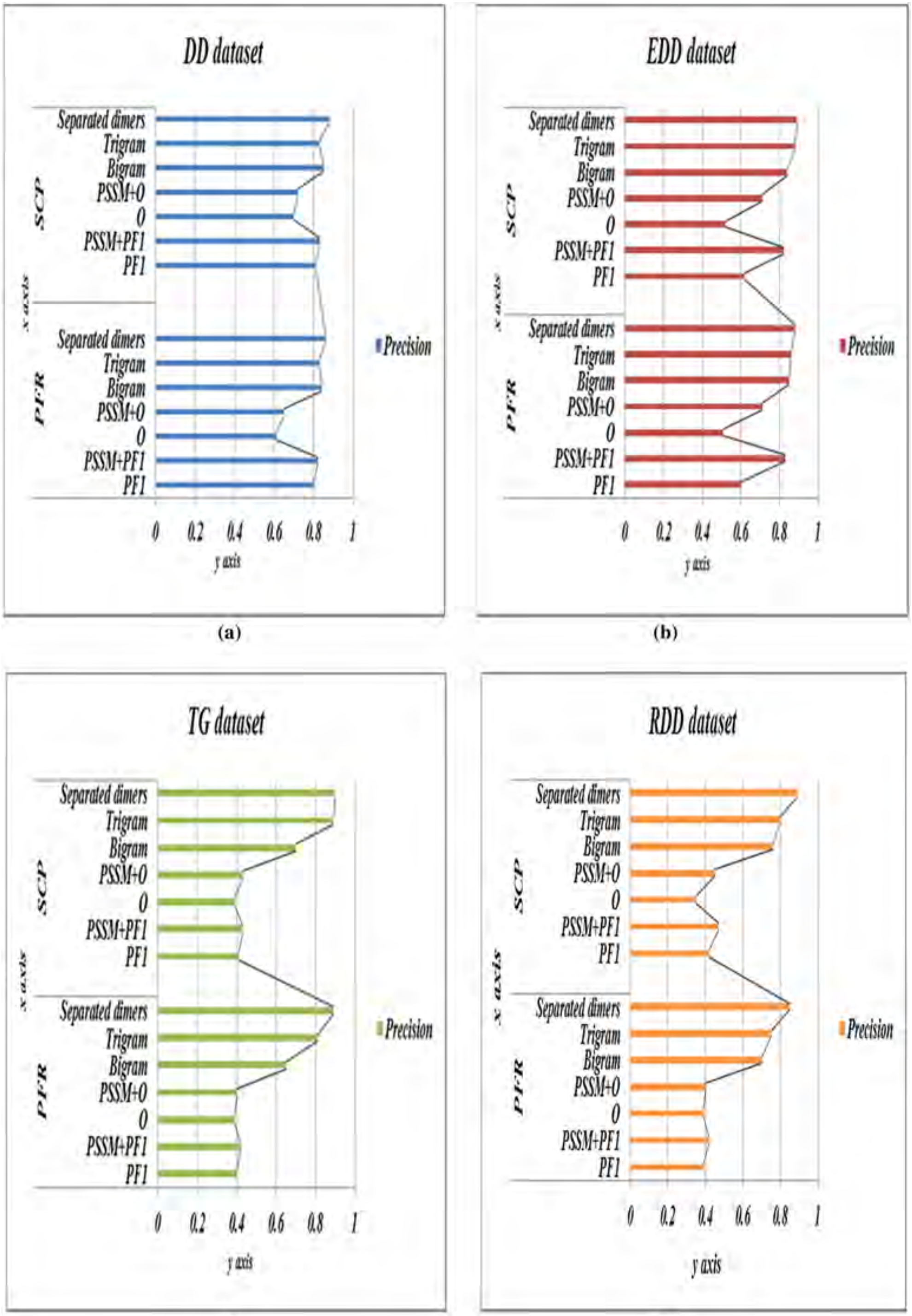
where F(k) computes as the feature sets for probabilistic occurrence of amino acid dimers with different values of k.

3.2. Physico-chemical based features

In order to improve the performance of PFR and SCP the physico-chemical based feature are used to preserve more judicious information from the amino acids of protein sequences and is examined by

$$A_i = \frac{1}{\mathrm{L}} \sum_{m=1}^{L-\mathrm{i}} (f_m - \mu)(f_{m+\mathrm{i}} - \mu) \tag{8}$$

where i consists of the values 1 to 20 that is fabricated for each protein sequences with 20 autocorrelation features, L is the length of



(a) (b)

(c) (d)

Fig. 6. Precision of all feature sets for PFR and SCP on DD, EDD, TG and RDD datasets.

probabilistic residues, f_m determines the probabilistic residue value of protein sequence with mth amino acid and the average value of L is depicted by μ (Raicar et al., 2016).

3.3. Forward consecutive search scheme

The FCS (Forward Consecutive Search Scheme) is a scheme that is basically used to combine physico-chemical based features with the existing syntactical and evolutionary based features to improve the performance of prediction of structural class as well as protein fold. This can be derived from Eq. (1). The main aim of FCS is relatively simple and its implementation is, therefore, straight forward for solving feature extraction problem (Raicar et al., 2016).

$$\Omega_S = \mathrm{arg}max_{j=1,2,r}\, \alpha(\{\mathrm{Feature}, \mathrm{SF}_{\Omega1}, \mathrm{SF}_{\Omega2}, \ldots, \mathrm{SF}_j\}),\ \mathrm{j} \neq \Omega_1,\ \Omega_{2,\ldots,}\Omega_{S-1} \tag{9}$$

where Ω_S is a physico-chemical attribute, selected at a given level S;S means total number of levels, j determines maximum number of physico-chemical attributes; α is a baseline accuracy; Feature consists of any one of the syntactical based and evolutionary based features; $SF_{\Omega1}$& $SF_{\Omega2}$ are the previous level successive features and SF_j means successive features.

3.4. The proposed approach: Enhanced Artificial Neural Network

The proposed algorithm is designed to predict the 27 protein fold type and protein structural class such as all-α, all-β, α/β and α + β from extracted features using physico-chemical properties based on artificial neurons and nearest neighbor. This is performed by Normalization, Euclidean distance, nearest neighbor and Sigmoid activation function. The generalized flow chart of Enhanced ANN is given in Fig. 1.

3.4.1. Normalization

$$\text{Norm} = \text{e} - E_{min}/E_{max} - E_{min} \tag{10}$$

where e determines data; E_{min} indicates the minimum values in each column; E_{max} determines the maximum values in each column. Fig. 2(b) illustrates the normalization step.

3.4.2. Euclidean distance

$$\text{d} = \sqrt{\sum_{i=1}^{v} (P_{1i} - P_{2i})^2} \tag{11}$$

where d indicates distance; v determines the length of amino acids in protein sequences; P_{1i} and P_{2i} are the two points (i.e. Euclidean vectors). The weight calculation of protein sequence depicted in Fig. 2(c).

3.4.3. Nearest neighbor

$$u_i\,(\text{r}) = \frac{\sum_{j=1}^{\text{K}} u_{ij}\left(1/\|d\,(r, r_j)^2\|\right.}{\sum_{j=1}^{\text{K}}\left(1/\|d\,(r, r_j)^2\|\right.} \tag{12}$$

where u_i determines the membership values in the i^{th} class; r is the current residue that is assigned to membership values and r_j indicates the neighbor residue of r; d means the distance between the residues; K is the count of nearest neighbor determined in Fig. 2(d).

Instead of picking all the weight with low and high, Eq. (12) is used to pick only minimum nearest neighbor based on k value that provides the results in an efficient manner when compared with existing classifier. The flow chart of Enhanced ANN with its input values is given in Fig. 2.

3.4.4. Sigmoidal activation function

$$\text{sig}\,(\text{x}) = 1/(1 + e^{-Z}) \tag{13}$$

where e^{-z} is an exponential function that is mainly used to compute activation function with small range of real number such as (−1 to +1). Fig. 2(e) and Fig. 2(f) constitutes the Summation process and Activation function. The pseudo code of Enhanced ANN is shown in Fig. 3.

Algorithm 1. Pseudo code of Enhanced ANN algorithm.

```
1. Read and normalize datasets using Eq. (10) //input layer
2. Distinguishing training and test data //input layer
3. Weight is calculated based on distance using Eq. (11) //hidden layer
    (a) Setting value for parameter k
    (b) Estimating the distance between input data and training data
    (c) Sorting distances in an ascending pattern
    (d) Choosing the best k neighbour using Eq.  (12)
    (e) Repeating steps 2-4 until the algorithm is over
    (f) Weight matrix is saved as result
4. Simulation involves using Eq. (13) //hidden layer
5. Saving results
6. Retrieving MLP model
    (a) Setting values for number of input, output and hidden layers
    (b) Primary weighing of existing neurons in input, output and hidden layers
    (c) Calculating the output (y) for each neuron in output layer
    (d) Updating MLP parameters
    (e) Repeating steps 3-4 until the algorithm is over
7. Saving results
8. End of hybrid model
9. Displaying results
10. End
                    Where MLP – Multi Layer perceptron
```

3.5. Datasets

In this research work, the proposed algorithm is inspected with the well-known benchmark datasets for analyzing the execution of the algorithm based on the strength. The benchmark datasets used in this research work are DD, EDD, TG and RDD. The DD dataset consists of 311 protein sequences presented in DD training datasets with similarity of < 35% along with 384 protein sequences presented in testing datasets with similarity lower than 40% (Ding and Dubchak, 2001; Murzin et al., 1995). The EDD dataset consists of 2082 protein sequences in training and 1336 protein sequences in testing. Totally it consists of 3418 protein sequences with similarity of < 40% (Dong et al., 2009). The TG dataset consists of 1010 protein sequences in training as well as 602 protein sequences in testing dataset. Totally it consists of 1612 protein sequences with similarity < 25%, which belongs to 30 fold types that represents all the major structural class. The training sets are mostly used to find out physico-chemical attributes (Taguchi and Gromiha, 2007). The RDD dataset consists of 311 protein sequences in training datasets and 380 protein sequences in testing datasets with similarity lower than 37% (Xia et al., 2017).

3.6. Performance measures

This research focuses on the performance measures such as precision, sensitivity and specificity in order to produce various statistical consequence in order to successfully bring out results.

3.6.1. Specificity

Specificity measures the negative proportion that is identified correctly to the whole number of test samples that are rejected and is premeditated by using Eq. (14).

$$\text{Specificity} = \frac{\text{T N}}{\text{TN + FP}} \tag{14}$$

where TN depicts true negative and FP denotes false positive samples.

3.6.2. Sensitivity

Sensitivity measures the positive proportion that is correctly

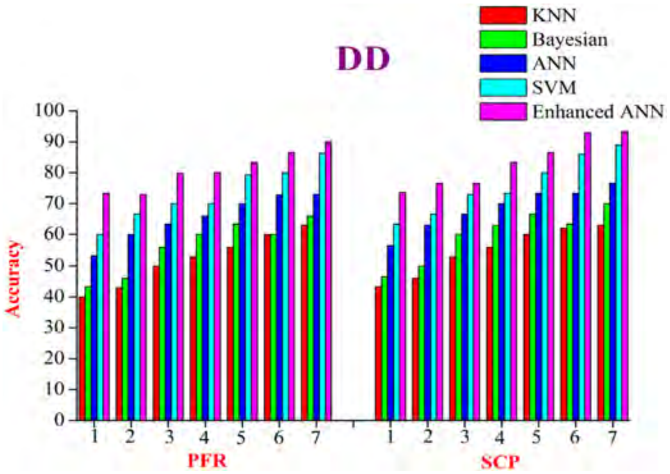


Fig. 7. Comparison of prediction accuracy for the existing and proposed algorithms on DD dataset.

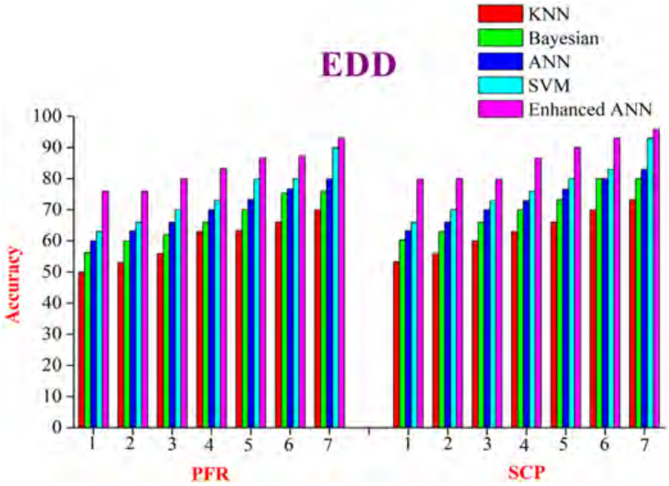


Fig. 8. Comparison of prediction accuracy for the existing and proposed algorithms on EDD dataset.

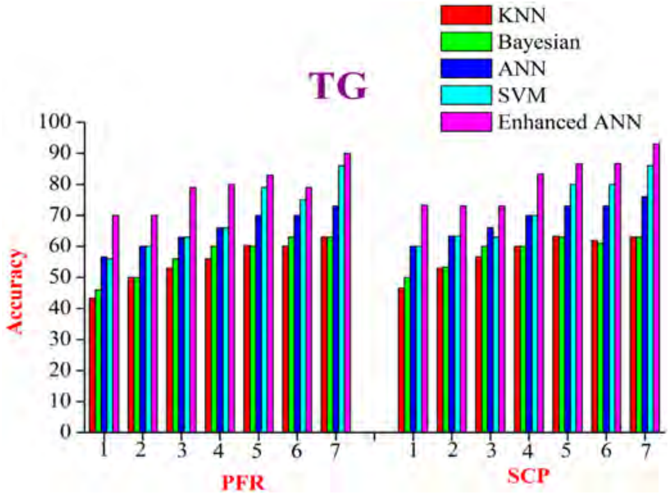


Fig. 9. Comparison of prediction accuracy for the existing and proposed algorithms on TG dataset.

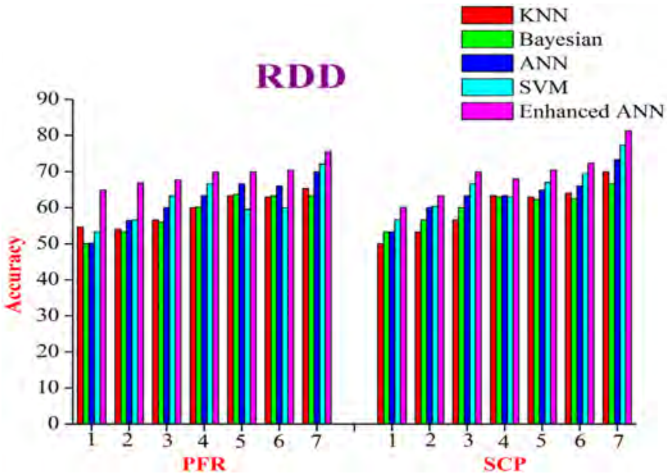


Fig. 10. Comparison of prediction accuracy for the existing and proposed algorithms on RDD dataset.

identified among the test samples which is correctly classified and deliberated by using Eq. (15).

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{15}$$

where TP represents true positive and FN denotes false negative samples.

3.6.3. Precision

Precision refers to how related number of TP is identified to the whole number of positive predictions and is denoted by using Eq. (16).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{16}$$

where TP determines true positive and FP constitutes false positive samples. The specificity, sensitivity and precision are evaluated for both PFR and SCP with seven features such as O, PF1, PSSM + O, PSSM + PF1, Bigram, Trigram and Separated dimers by using an Enhanced ANN algorithm.

Fig. 4(a), (b), (c) and (d) depicts the performance of Sensitivity for DD, EDD,TG and RDD datasets for proposed algorithm.

Fig.5(a), (b), (c) and (d) shows the performance of Sensitivity for DD, EDD,TG and RDD datasets for proposed algorithm.

Fig. 6 (a), (b), (c) and (d) shows the performance of Sensitivity for DD, EDD, TG and RDD dataset for proposed algorithms. However, when compared with all the datasets Sensitivity and precision are quite varied for both PFR and SCP.

4. Implementation and discussion

In addition, the public presentation of the enhanced algorithm has been gauged by comparison with many classification techniques, namely SVM, ANN, KNN and Bayesian classifier. From Figs. 7–10 is evidence that classifications of syntactical, evolutionary based features and physico-chemical based features have improved the performance of PFR and SCP that displayed better prediction accuracy when compared with both existing and proposed algorithms. From these results, it is inferred that the proposed Enhanced ANN classification algorithm has given better results than the other existing algorithms such as SVM, ANN, KNN and Bayesian for the syntactical and evolutionary based features such as PF1, PSSM + PF1, O, PSSM + O, Bigram, Trigram and Separated dimmers. Furthermore, the classification test was conducted

Table 1
Comparison of accuracy prediction for the existing and the proposed algorithm for the DD dataset.

	Features	KNN	Bayesian	ANN	SVM	Enhanced ANN
		(Nanni, 2006)	(Chinnasamy et al., 2005)	(Cai and Zhou, 2000)	(Chen et al., 2006)	(Proposed)
PFR	PF1 (Ghanty and Pal, 2009)	40.0	43.3	53.3	60.0	73.3
	PSSM + PF1 (Raicar et al., 2016)	43.0	46.0	60.0	66.6	73.0
	O (Taguchi and Gromiha, 2007)	50.0	56.0	63.3	70.0	79.9
	PSSM + O (Raicar et al., 2016)	53.0	60.0	66.0	70.0	80.0
	Bigram (Sharma et al., 2013)	56.0	63.3	70.0	79.3	83.3
	Trigram (Lyons et al., 2016)	60.0	60.0	72.8	80.0	86.6
	Separated dimers (Saini et al., 2015)	63.0	66.0	73.0	86.3	90.0
SCP	PF1 (Raicar et al., 2016)	43.3	46.6	56.6	63.3	73.6
	PSSM + PF1 (Lyons et al., 2016)	46.0	50.0	63.0	66.6	76.6
	O (Raicar et al., 2016)	53.0	60.0	66.6	73.0	76.6
	PSSM + O (Lyons et al., 2016)	56.0	63.0	70.0	73.3	83.3
	Bigram (Saini et al., 2014)	60.0	66.6	73.3	80.0	86.6
	Trigram (Paliwal et al., 2014a)	62.0	63.3	73.3	86.0	93.0
	Separated dimers (Saini et al., 2015)	63.0	70.0	76.6	89.0	93.3

Table 2
Comparison of accuracy prediction for the existing and the proposed algorithm for the EDD dataset.

	Features	KNN	Bayesian	ANN	SVM	Enhanced ANN
		(Nanni, 2006)	(Chinnasamy et al., 2005)	(Cai and Zhou, 2000)	(Chen et al., 2006)	(Proposed)
PFR	PF1 (Ghanty and Pal, 2009)	50.0	56.3	60.0	63.0	76.0
	PSSM + PF1 (Raicar et al., 2016)	53.0	60.0	63.3	66.0	76.0
	O (Taguchi and Gromiha, 2007)	56.0	62.0	66.0	70.0	80.0
	PSSM + O (Raicar et al., 2016)	63.0	66.0	70.0	73.0	83.3
	Bigram (Sharma et al., 2013)	63.3	70.0	73.3	79.9	86.6
	Trigram (Lyons et al., 2016)	66.0	75.4	76.7	80.0	87.3
	Separated dimers (Saini et al., 2015)	70.0	76.0	80.0	90.0	93.0
SCP	PF1 (Raicar et al., 2016)	53.3	60.3	63.3	66.0	79.9
	PSSM + PF1 (Lyons et al., 2016)	56.0	63.0	66.0	70.0	80.0
	O (Raicar et al., 2016)	60.0	66.0	70.0	73.0	79.9
	PSSM + O (Lyons et al., 2016)	63.0	70.0	73.0	76.0	86.6
	Bigram (Saini et al., 2014)	66.0	73.3	76.6	80.0	90.0
	Trigram (Paliwal et al., 2014a)	70.0	80.0	80.0	83.0	93.0
	Separated dimers (Saini et al., 2015)	73.3	80.0	83.0	93.0	96.0

Table 3
Comparison of accuracy prediction for the existing and the proposed algorithm for the TG dataset.

	Features	KNN	Bayesian	ANN	SVM	Enhanced ANN
		(Nanni, 2006)	(Chinnasamy et al., 2005)	(Cai and Zhou, 2000)	(Chen et al., 2006)	(Proposed)
PFR	PF1 (Ghanty and Pal, 2009)	43.3	46.0	56.6	56.0	70.0
	PSSM + PF1 (Raicar et al., 2016)	50.0	50.0	60.0	60.0	70.0
	O (Taguchi and Gromiha, 2007)	53.0	56.0	63.0	63.0	79.0
	PSSM + O (Raicar et al., 2016)	56.0	60.0	66.0	66.0	80.0
	Bigram (Sharma et al., 2013)	60.3	60.0	70.0	79.0	83.0
	Trigram (Lyons et al., 2016)	60.1	63.0	70.0	75.0	79.0
	Separated dimers (Saini et al., 2015)	63.0	63.0	73.0	86.0	90.0
SCP	PF1 (Raicar et al., 2016)	46.6	50.0	60.0	60.0	73.3
	PSSM + PF1 (Lyons et al., 2016)	53.0	53.3	63.3	63.3	73.0
	O (Raicar et al., 2016)	56.6	60.0	66.0	63.0	73.0
	PSSM + O (Lyons et al., 2016)	60.0	60.0	70.0	70.0	83.3
	Bigram (Saini et al., 2014)	63.3	63.0	73.0	80.0	86.6
	Trigram (Paliwal et al., 2014a)	62.0	61.0	73.0	80.0	86.7
	Separated dimers (Saini et al., 2015)	63.0	63.0	76.0	86.0	93.0

for both PFR and SCP by using various datasets such as DD, EDD, TG and RDD.

For the prediction accuracy shown in Table 1 and Fig. 7, it is determined that the Enhanced ANN algorithm performed 73.3%, 73.0%, 79.9%, 80.0%,83.3%, 86.6% and 90.0% better for PFR and also performed 73.6%,76.6%,76.6%,83.3%,86.6%,93.0% and 93.3% better for SCP for PF1,PSSM + PF1,O,PSSM + O,Bigram and separated dimers on DD dataset compared to other existing algorithms. From the Table 2 and Fig. 8, it is hypothesized that the Enhanced ANN algorithm performed 76.0%, 76.0%, 80.0%, 83.3%, 86.6%, 87.3% and 93.0% better for PFR and also performed 79.9%, 80.0%, 79.9%, 86.6%, 90.0%, 93.0% and 96.0% better for SCP for PF1, PSSM + PF1, O, PSSM + O, Bigram and separated dimers on EDD dataset compared to other existing algorithms.

From Table 3 and Fig. 9, it is concluded that the Enhanced ANN algorithm performed 70.0%, 70.0%, 79.0%, 80.0%, 83.0%, 79.0% and 90.0% better for PFR and also performs 73.3%, 73.0%, 73.0%, 83.3%, 86.6%, 86.7% and 93.0% better for SCP for PF1, PSSM + PF1, O, PSSM + O, Bigram and separated dimers on TG dataset compared to other existing algorithms. Table 4 and Fig. 10 shows that the Enhanced ANN algorithm performed 64.9%, 66.9%, 67.7%, 69.9%, 70.0%, 70.5% and 75.4% better for PFR and also performs 60%, 63.3%, 70.0%.

Table 4
Comparison of accuracy prediction for the existing and the proposed algorithm for the RDD dataset.

	Features	KNN (Nanni, 2006)	Bayesian (Chinnasamy et al., 2005)	ANN (Cai and Zhou, 2000)	SVM (Chen et al., 2006)	Enhanced ANN (Proposed)
PFR	PF1 (Ghanty and Pal, 2009)	54.6	50.0	50.2	53.3	64.9
	PSSM + PF1 (Raicar et al., 2016)	54.0	53.3	56.5	56.6	66.9
	O (Taguchi and Gromiha, 2007)	56.6	56.0	60.0	63.3	67.7
	PSSM + O (Raicar et al., 2016)	60.0	60.2	63.3	66.6	69.9
	Bigram (Sharma et al., 2013)	63.3	63.7	66.6	59.6	70.0
	Trigram (Lyons et al., 2016)	63.0	63.3	66.0	60.0	70.5
	Separated dimers (Saini et al., 2015)	65.3	63.3	70.0	72.1	75.4
SCP	PF1 (Raicar et al., 2016)	50.0	53.3	53.3	56.6	60.0
	PSSM + PF1 (Lyons et al., 2016)	53.3	56.6	60.0	60.4	63.3
	O (Raicar et al., 2016)	56.6	60.0	63.3	66.6	70.0
	PSSM + O (Lyons et al., 2016)	63.3	63.0	63.3	63.0	68.0
	Bigram (Saini et al., 2014)	63.0	62.3	64.9	67.0	70.5
	Trigram (Paliwal et al., 2014a)	64.1	62.5	66.0	69.4	72.3
	Separated dimers (Saini et al., 2015)	70.0	66.6	73.3	77.3	81.3

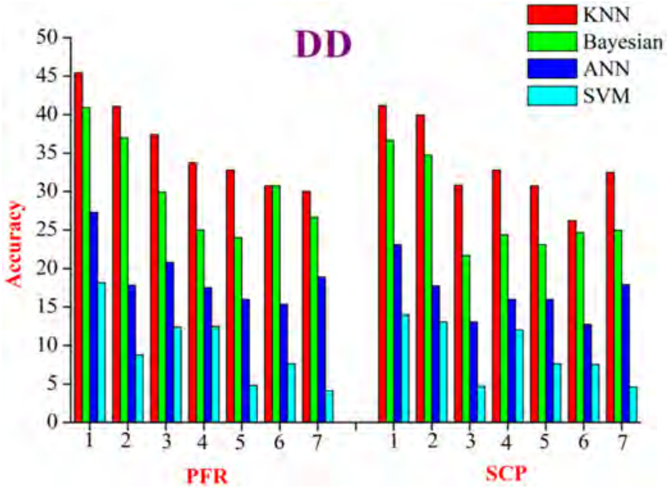


Fig. 11. Statistical significance of proposed and existing algorithms for DD dataset.

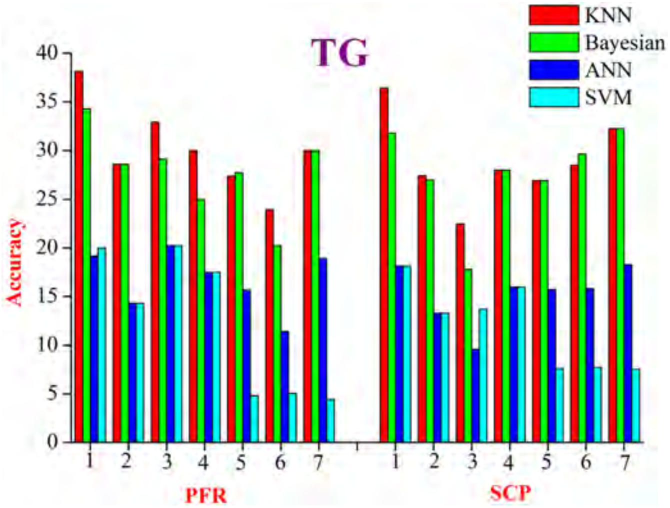


Fig. 13. Statistical significance of proposed and existing algorithms for TG dataset.

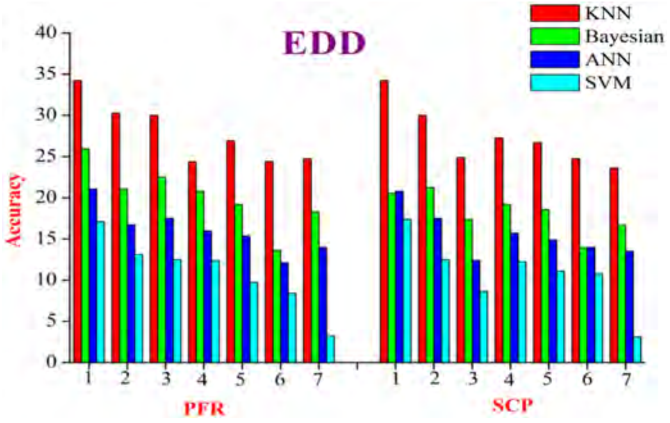


Fig. 12. Statistical significance of proposed and existing algorithms for EDD dataset.

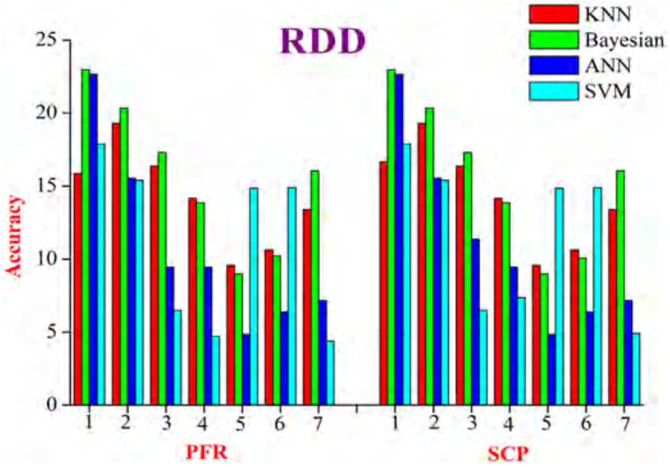


Fig. 14. Statistical significance of proposed and existing algorithms for RDD dataset.

Table 5
Statistical significance of proposed and existing algorithms for DD dataset.

	Features	KNN	Bayesian	ANN	SVM
PFR	PF1	45.43	40.93	27.29	18.15
	PSSM + PF1	41.1	36.99	17.81	8.77
	O	37.43	29.92	20.78	12.4
	PSSM + O	33.75	25.0	17.5	12.5
	Bigram	32.78	24.01	15.97	4.81
	Trigram	30.72	30.72	15.36	7.63
	Separated dimers	30.0	26.67	18.89	4.12
SCP	PF1	41.17	36.69	23.1	14
	PSSM + PF1	39.95	34.73	17.76	13.06
	O	30.81	21.68	13.06	4.7
	PSSM + O	32.78	24.37	15.97	12.01
	Bigram	30.72	23.1	15.97	7.63
	Trigram	26.2	24.65	12.74	7.53
	Separated dimers	32.48	24.98	17.9	4.61

Table 6
Statistical significance of proposed and existing algorithms for EDD dataset.

	Features	KNN	Bayesian	ANN	SVM
PFR	PF1	34.22	25.93	21.06	17.11
	PSSM + PF1	30.27	21.06	16.72	13.16
	O	30.0	22.5	17.5	12.5
	PSSM + O	24.37	20.77	15.97	12.37
	Bigram	26.91	19.17	15.36	9.74
	Trigram	24.4	13.64	12.15	8.4
	Separated dimers	24.74	18.28	13.98	3.23
SCP	PF1	34.22	20.54	20.78	17.4
	PSSM + PF1	30.0	21.25	17.5	12.5
	O	24.91	17.4	12.4	8.64
	PSSM + O	27.26	19.17	15.71	12.25
	Bigram	26.67	18.56	14.89	11.12
	Trigram	24.74	13.98	13.98	10.76
	Separated dimers	23.65	16.67	13.55	3.13

Table 7
Statistical significance of proposed and existing algorithms for TG dataset.

	Features	KNN	Bayesian	ANN	SVM
PFR	PF1	38.15	34.29	19.15	20.0
	PSSM + PF1	28.58	28.58	14.29	14.29
	O	32.92	29.12	20.26	20.26
	PSSM + O	30.0	25.0	17.5	17.50
	Bigram	27.35	27.72	15.67	4.82
	Trigram	23.93	20.26	11.4	5.07
	Separated dimers	30.0	30.0	18.89	4.45
SCP	PF1	36.43	31.79	18.15	18.15
	PSSM + PF1	27.40	26.99	13.29	13.29
	O	22.47	17.81	9.59	13.7
	PSSM + O	27.98	27.98	15.97	15.97
	Bigram	26.91	26.91	15.71	7.63
	Trigram	28.49	29.65	15.81	7.73
	Separated dimers	32.26	32.26	18.28	7.53

68.0%, 70.5%, 72.3% and 81.3% better for SCP for PF1, PSSM + PF1, O, PSSM + O, Bigram and separated dimers on TG dataset compared to other existing algorithms.

The X-axis in Figs. 7–10 and Figs. 11–14 are represented as follows, Where,
1-PF1.
2-PSSM + PF1.
3-O
4-PSSM + O.
5-Bigram.
6-Trigram.
7-Separated dimers.

Table 8
Statistical significance of proposed and existing algorithms for RDD dataset.

	Features	KNN	Bayesian	ANN	SVM
PFR	PF1	15.88	22.96	22.66	17.88
	PSSM + PF1	19.29	20.33	15.55	15.4
	O	16.4	17.29	9.45	6.5
	PSSM + O	14.17	13.88	9.45	4.73
	Bigram	9.58	9.0	4.86	14.86
	Trigram	10.64	10.22	6.39	14.9
	Separated dimers	13.4	16.05	7.17	4.4
SCP	PF1	16.67	22.96	22.66	17.88
	PSSM + PF1	19.29	20.33	15.55	15.4
	O	16.4	17.29	11.38	6.5
	PSSM + O	14.17	13.88	9.45	7.36
	Bigram	9.58	9.0	4.86	14.86
	Trigram	10.64	10.08	6.39	14.9
	Separated dimers	13.4	16.05	7.17	4.93

4.1. Statistical analysis

For the statistical metric given in Table 5 and Fig. 11, it is inferred that for both PFR and SCP the Enhanced ANN algorithm performed 30.0% and 32.48% better than KNN, 26.67% and 24.98% better than Bayesian, 18.89% and 17.9% better than ANN and also 4.12% and 4.61% better than SVM for the separated dimers feature on DD dataset. For the dataset EDD, for both PFR and SCP the Table 6 and Fig. 12 deduced that Enhanced ANN algorithm performed 24.74% and 23.65 better than KNN, 18.28% and 16.67% better than Bayesian, 13.98% and 13.55% better than ANN and also 3.23% and 3.13% better than SVM for the separated dimers feature. For the dataset TG, Table 7 and Fig. 13 it is concluded for both PFR and SCP that the Enhanced ANN algorithm performed 30.0% and 32.26% better than KNN, 30.0% and 32.26% better than Bayesian, 18.89% and 18.28% better than ANN and also 4.45% and 7.53% better than SVM for separated dimers feature. Finally, Table 8 and Fig. 14 shows the statistical improvements for both PFR and SCP that the Enhanced ANN algorithm performed 13.4% and 13.4% better than KNN, 16.05% and 16.05% better than Bayesian, 7.17% and 7.17% better than ANN and also 4.4% and 4.93% better than SVM for a separated dimers feature on the RDD dataset.

5. Biological significance

In computational biology, predicting the protein function from the primary protein is a very ***decisive*** task. The homology of sequence with secondary structure is helpful to predict protein function. The protein sequences which consist of secondary structure that can be used to identify several features of protein function like active site residues, cellular location, interactions with ligands and other proteins (Tiwari and Srivastava, 2014).

The Biological significance is conducted to analyze the performance of the proposed Enhanced ANN. The testing is carried out by using the Genome Motif tool. For testing three protein sequences from each structural class were taken from all-α, all-β, α/β and α + β. Totally 12 protein sequences are tested of which 1,2 and 3 protein sequences from all-α, 4,5 and 6 protein sequences from all-β, 7,8 and 9 protein sequences from α/β and finally 10,11 and 12 protein sequences from α + β. The result of biological significance is shown in Table 9. The protein sequences 1, 2,3 have a similar *NCBI-CDD function ID* such as 271265, 225223, 271289, 307346, 271305, 271308, 271306, 271304 and 271269 for all-*α class*. The protein sequences 4,5,6 have a similar NCBI-CDD function ID such as 143185, 319275, 143181, 311561, 319330, 319287, 143307, 214650, 214653, 214652, 143183, 319326, 316449 and 197706 for all-*β*. The protein sequences 7, 8, 9 have a similar NCBI-CDD function ID such as 307128 for α/β and finally the protein sequences 10,11,12 have a similar NCBI-CDD function ID such as 235193.

Table 9
Biological significance of proposed algorithm.

Protein function	Structural class	Protein sequence	NCBI-CDD function ID	Function definition
1.	All-α	Slfeqlggqaavqavtaqfyaniqadatvatffngidmpnqtnktaaflcaalggpnawtgrnlkevhanmgvsnaqfttvighlrsaltgagvaaalveqtvavaetvrgdvvtv	271265	TrHb1_N
2.		Stlyeklggttavdlavdkfyervlqddrikhffadvdmakqrahqkafltyafggtdkydgrymreahkelvenhglngehfdavaedllatlkemgvpedliaevaavagapahkrdvinq	225223	YjbI
3.		Gllsrlrkrepisiydkiggheaievvvedfyvrvladdqlsaffsgtnmsrlkgkqveffaaalggpepytgapmkqvhqgrgitmhhfslvaghladaltaagvpsetiteilgviaplavdvts	271289	TrHb
			307346	Bac_globin
			271305	TrHb2_Bs-trHb like_O
			271308	TrHb2_O_2
			271306	TrHb2_PhHbO-like_O
			271304	TrHb2_Mt-trHbO-
			271269	like_O
				TrHb3_P
4.	All-β	divmtqaapsvpvtpgesvsiscrssksllhsngntylywflqrpgqspqlliyrmsnlasgvpdrfsgsgsgtaftlrisrveaedvgvyyclqhleypftfgagtklelk	143185	IgV_L_lambda
5.		qavvtqesalttspgetvtltcrsstgavttsnyanwvqekpdhlftgliggtnnrapgvparfsgslignkaaltitgaqtedeaiyfcalwysnhlvfgggtkltvlg	143181	IgV_L_kappa
6.		Evqlqqpgaelvkpgasvklsckasgytftnywinwvkqrpgqglewigniypgssythynekfknkatltvdtssstaymqlssltsddsavyycanklgwfpywgqgtlvtvsa	319275	IgV
			311561	V-set
			319330	IgV_TCR_delta
			319287	IgV_TCR_alpha
			143307	IgV_TCR_beta
			214650	IGv
			214653	IG_like
			2148652	IG
			143183	IgV_TCR_gamma
			319326	IgV_CD8_beta
			316449	Ig_3
			197706	IGc2
7.	α/β	vlsegewqlvlhvwakveadvaghgqdilirlfkshpetlekfdrfkhlkteaemkased	271266	Mb_like
		lkkhgvtvltalgailkkkghheaelkplaqshatkhkipikylefiseaiihvlhsrhp	306539	Globin
8.		gdfgadaqgamnkalelfrkdiaakykelgy	271287	GbX
		slsaaeadlagkswapvfanknangldflvalfekfpdsanffadfkgksvadikaspkl	271299	CeGLB25_like
		rdvssriftrlnefvnnaanagkmsamlsqfakehvgfgvgsaqfenvrsmfpgfvasva	271275	Cygb
		appagadaawtklfgliidalkaaga		
9.		galtesqaalvkssweefnanipkhthrffilvleiapaakdlfsflkgtsevpqnnpel		
		qahagkvfklvyeaaiqlevtgvvvtdatlknlgsvhvskgvadahfpvvkeailktike		
		vvgakwseelnsawtiaydelaivikkemddaa		
10.	α + β	yvetrelqylypegeemvfmdletyeqfavprsrvvgaeffkegmtalgdmyegqpikvtppt	235193	PRK03999
11.		Ddpllaqlkqqlhsqtpraegvvkatekgfgflevdaqksyfipppqmkkvmhgdriiavihsekeresaepeelvepf		
12.		eentvdfqlefvfnevqdpdlderdptstyldydakpnliklrvwqgsnvhtdfakldlsdddwerlkaleqplqgriaecrqsttkkgywemlrfrndksngnhisvvekilvsikdgvkekeviewcpkisrawkkrendrrq		

6. Conclusion

In Structural Bioinformatics the prediction of three dimensional structure of protein without using these PFR and SCP becomes a very difficult task. Sometimes not correctly folded or structured proteins produce many diseases in living organisms. Predicting the protein structure is mainly used to avoid the diseases that arise in the living cells. Several methods have been introduced to overcome this problem but still some issues persist.

In this research work, to improve the performance of PFR and SCP the existing feature extraction techniques such as syntactical-based information and evolutionary-based information are not sufficient. In addition, here we are extracting the features from protein sequence by combining existing techniques with physico-chemical based information using FCS. To classify these extracted features efficiently the Enhanced Artificial Neural Network Algorithm has been introduced. The real protein sequences with unique length are used to test the enhanced algorithm. The results are compared with four existing algorithms such as DD, EDD, TG and RDD. The Enhanced Artificial Neural Network provides higher accuracy than others.

In future, classification can be done with the more syntactical and evolutionary features and a new feature extraction method can be introduced to supplement existing feature techniques efficiently. Divergent objectives may be advanced to find better solutions for PFR and SCP.

Acknowledgement

The authors thank the Department of Science and Technology (DST), New Delhi (DST/INSPIRE Fellowship/2015/IF150093) for the financial support under INSPIRE Fellowship for this research work.

References

Baldi, Pierre, Pollastri, Gianluca, 2003. The principled design of large-scale recursive neural network architectures–DAG-RNNs and the protein structure prediction problem. J. Mach. Learn. Res. 4, 575–602.

Bouchaffra, D., Tan, J., 2006. Protein fold recognition using a structural hidden Markov model. In: Proceedings of the 18th International Conference on Pattern Recognition. 3. pp. 186–189.

Bulashevska, A., Eils, R., 2006. Predicting protein subcellular locations using hierarchical ensemble of Bayesian classifiers based on Markov chains. BMC Bioinf. 7 (1), 298.

Cai, Y.D., Zhou, G.P., 2000. Prediction of protein structural classes by neural network. Biochimie 82 (8), 783–785.

Chandonia, John-marc, Karplus, Martin, 1995. Neural networks for secondary structure and structural class predictions. Open Struct. Biol. J. 1, 1–6.

Chen, C., Tian, Y.X., Zou, X.Y., Cai, P.X., Mo, J.Y., 2006. Using pseudo-amino acid composition and support vector machine to predict protein structural class. J. Theor. Biol. 243 (3), 444–448.

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). I. pp. 66–70.

Chen, K.E., Kurgan, L.A., Ruan, J., 2008. Prediction of protein structural class using novel evolutionary collocation-based sequence representation. J. Comput. Chem. 29 (10), 1596–1604.

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

Chou, K.C., 2001. Prediction of protein cellular attributes using pseudo amino acid composition. Proteins 43, 246–255.

Cormen, T.H., Leiserson, C.E., Rivest, R.L., 1990. Introduction to Algorithms.

Damoulas, T., Girolami, M.A., 2008. Probabilistic multi-class multi-kernel learning: on protein fold recognition and remote homology detection. Bioinformatics 24 (10), 1264–1270.

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

Dehzangi, A., Phon-Amnuaisuk, S., 2011. Fold prediction problem: the application of new physical and physicochemical-based features. Protein Pept. Lett. 18 (2), 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, pp. 503–511.

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

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

Dehzangi, A., Paliwal, K., Lyons, J., Sharma, A., Sattar, A., 2013a. Enhancing protein fold prediction accuracy using evolutionary and structural features. Pattern Recognit. Bioinform. 196–207.

Dehzangi, A., Paliwal, K., Sharma, A., Dehzangi, O., Sattar, A., 2013b. A combination of feature extraction methods with an ensemble of different classifiers for protein structural class prediction problem. IEEE/ACM Trans. Comput. Biol. Bioinform. 10 (3), 564–575.

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

Dehzangi, A., Sharma, A., Lyons, J., Paliwal, K.K., Sattar, A., 2014b. A mixture of physicochemical and evolutionary–based feature extraction approaches for protein fold recognition. Int. J. Data Min. Bioinf. 11 (1), 115–138.

Ding, C.H.Q., 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., 2013. Using Chou's pseudo amino acid composition to predict subcellular localization of apoptosis proteins. BMC Bioinf. 14 (233), 9–11.

Dong, Q., Zhou, S., Guan, G., 2009. A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation. Bioinformatics 25 (20), 2655–2662.

Dubchak, I., Muchnik, I.B., Kim, S.H., 1997. Protein folding class predictor for SCOP: approach based on global descriptors. InIsmb 104–107.

Gassend, Blaise, Charles O'Donnell, W., Thies, William, Lee, Andrew, van Dijk, Marten, Devadas, Srinivas, 2006. Predicting secondary structure of all-helical proteins using hidden Markov support vector machines. Copyright Springer-Verlag Berlin Heidelberg, pp. 93–104.

Ghanty, P., Pal, N.R., 2009. Prediction of protein folds: extraction of new features, dimensionality reduction, and fusion of heterogeneous classifiers. IEEE Trans. NanoBiosci. 8, 100–110.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182.

Hae-Jin, Hu, Pan, Yi, Harrison, Robert, Tai, Phang C., 2004. Improved protein secondary structure prediction using support vector machine with a new encoding scheme and an advanced tertiary classifier. IEEE Trans. NanoBiosci. 3 (4), 265.

Hashemi, H.B., Shakery, A., Naeini, M.P., 2009. Protein fold pattern recognition using Bayesian ensemble of RBF neural networks. In: International Conference of Soft Computing and Pattern Recognition. IEEE.

Hayat, Maqsood, et al., 2014a. Prediction of protein structure classes using hybrid space of multi-profile Bayes and bi-gram probability feature spaces. J. Theor. Biol. 346, 8–15.

Hayat, M., Tahir, M., Khan, S.A., 2014b. Prediction of protein structure classes using hybrid space of multi-profile Bayes and bi-gram probability feature spaces. J. Theor. Biol. 346, 8–15.

Huang, J.T., Tian, J., 2006. Amino acid sequence predicts folding rate for middle size two state proteins. Proteins Struct. Funct. Bioinform. 63, 551–554.

Huang, Wen-Lin, Chena, Hung-Ming, Hwang, Shiow-Fen, Hob, Shinn-Ying, 2007. Accurate prediction of enzyme subfamily class using an adaptive fuzzy *k*-nearest neighbor method. Biosystems 90, 405–413.

Ibrahim, Wisam, Abadeh, Mohammad Saniee, 2017. Extracting features from protein sequences to improve deep extreme learning machine for protein fold recognition. J. Theor. Biol. 421, 1–15.

Kurgan, L., Cios, K., Chen, K., 2008. SCPRED: accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences. BMC Bioinf. 9 (1), 226.

Levitt, M., Chothia, C., 1976. Structural patterns in globular proteins. Nature 261, 552–558.

Li, Q., Dahl, D.B., Vannucci, M., Joo, H., Tsai, J.W., 2014. Bayesian model of protein primary sequence for secondary structure prediction. PLoS One 9 (10), e109832.

Liu, T., Jia, C., 2010. A high-accuracy protein structural class prediction algorithm using predicted secondary structural information. J. Theor. Biol. 267 (3), 272–275.

Lyons, J., Biswas, N., Sharma, A., Dehzangi, A., Paliwal, K.K., 2014. Protein fold recognition by alignment of amino acid residues using kernelized dynamic time warping. J. Theor. Biol. 354, 137–145.

Lyons, J., Paliwal, K.K., Dehzangi, A., Heffernan, R., Tsunoda, T., Sharma, A., 2016. Protein fold recognition using HMM–HMM alignment and dynamic programming. J. Theor. Biol. 67–74.

Metfessel, B., Saurugger, P.N., 1993. Cross-Validation of Protein Structural Class Prediction Using Statistical Clustering and Neural Networks.

Minh, N., Nguyen Jagath, C., Rajapakse, 2003. Multi-class support vector machines for protein secondary structure prediction. Genome Inf. 14, 218–227.

Mohammad, H.Olyaee, AliYaghoubi, MahdiYaghoobi, 2016. Predicting protein structural classes based on complex networks and recurrence analysis. J. Theor. Biol. 404, 375–382.

Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C., 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247 (4), 536–540.

Nanni, L., 2006. Ensemble of classifiers for protein fold recognition. Neurocomputing 69 (7), 850–853.

Nanni, L., Brahnam, S., Lumini, A., 2014. Prediction of protein structure classes by incorporating different protein descriptors into general Chou's pseudo amino acid composition. J. Theor. Biol. 360, 109–116.

Pal, N.R., Chakraborty, D., 2003. Some new features for protein fold prediction. In: Artificial Neural Networks and Neural Information Processing—ICANN/ICONIP 2003. Springer, Berlin, Heidelberg, pp. 1176–1183.

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), 44–50.

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 Bioinf. 15 (Suppl16), S12.

Raicar, Gaurav, Saini, Harsh, Dehzangi, Abdollah, SunilLal, AlokSharma, 2016. Improving protein fold recognition and structural class prediction accuracies using physicochemical properties of amino acids. J. Theor. Biol. 402, 117–128.

Saini, H., Raicar, G., Sharma, A., Lal, S., Dehzangi, A., Rajeshkannan, A., Paliwal, K.K., 2014. Protein structural class prediction via k-separated bigrams using position specific scoring matrix. J. Adv. Comput. Intell. Intell. Inform. 8 (4).

Saini, H., Raicar, G., Sharma, A., Lal, S., Dehzangi, A., Lyons, J., Miyano, S., 2015. 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, 291–298.

Sharma, Alok, 2013. A strategy to select suitable physicochemical attributes of amino acids for protein fold recognition. BMC Bioinf. 14, 233.

Sharma, A., Imoto, S., Miyano, S., 2012a. A top-r feature selection algorithm for microarray gene expression data. IEEE/ACM Trans. Comput. Biol. Bioinform. 9 (3), 754–764.

Sharma, A., Imoto, S., Miyano, S., Sharma, V., 2012b. Null space based feature se- lection method for gene expression data. Int. J. Mach. Learn. Cybern. 3 (4), 269–276.

Sharma, A., Lyons, J., Dehzangi, A., Paliwal, K.K., 2013. A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition. J. Theor. Biol. 320, 41–46.

Sharma, A., Boroevich, K., Shigemizu, D., Kamatani, Y., Kubo, M., Tsunoda, T., 2016. Hierarchical maximum likelihood clustering approach. IEEE Trans. Biomed. Eng. https://doi.org/10.1109/TBME.2016.2542212,2016.

Shen, H., Chou, K.C., 2005. Using Optimized Evidence-Theoretic K-Nearest Neighbor Classifier and Pseudo-Amino Acid Composition to Predict Membrane Protein Types.

Shen, H.B., Chou, K.C., 2006. Ensemble classier for protein fold pattern recognition. Bioinformatics 22, 1717–1722.

Taguchi, Y.-h., Gromiha, M.M., 2007. Application of amino acid occurrence for discriminating different folding types of globular proteins. BMC Bioinf. 8 (1), 404.

Tao, P., Liu, T., Li, X., Chen, L., 2015. Prediction of protein structural class using tri-gram probabilities of position-specific scoring matrix and recursive feature elimination. Amino Acids 47 (3), 461–468.

Tiwari, Arvind Kumar, Srivastava, Rajeev, 2014. A Survey of Computational Intelligence Techniques in Protein Function Prediction. Int. J. Proteomics 2014, 1–22 845479.

Wang, Long-Hui, Liu, Juan, 2004. Predicting protein secondary structure by a support vector machine based on a new coding scheme. Genome Inform. 15 (2), 181–190.

Wang, Z.Z., Yuan, Z., 2000. How good is prediction of protein-structural class by the component-coupled method? Proteins 38, 165–175.

Xia, J., Peng, Z., Qi, D., Mu, H., Yang, J., 2017. An ensemble approach to protein fold classification by integration of template-based assignment and support vector machine classifier. Bioinformatics 33, 863–887.

Yan, K., Xu, Y., Fang, X., Zheng, C., Liu, B., 2017. Protein fold recognition based on sparse representation based classification. Artif. Intell. Med. 79, 1–8.

Yang, Jian-Yi, et al., 2010. Prediction of protein structural classes for low-homology sequences based on predicted secondary structure. BMC Bioinf. 11, S1–S9.

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.

Yu, B., Lou, L., Li, S., Zhang, Y., Qiu, W., Wu, X., Tian, B., 2017. Prediction of protein structural class for low-similarity sequences using Chou's pseudo amino acid composition and wavelet denoising. J. Mol. Graph. Model. 76, 260–273.

Zhang, L., Kong, L., Han, X., Lv, J., 2016. Structural class prediction of protein using novel feature extraction method from chaos game representation of predicted secondary structure. J. Theor. Biol. 400, 1–10.