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-$\alpha$ class consists of only little amount of strands, (*ii*) *the* all-$\beta$ class consists of only little amount of helices, (iii) the $\alpha/\beta$ 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

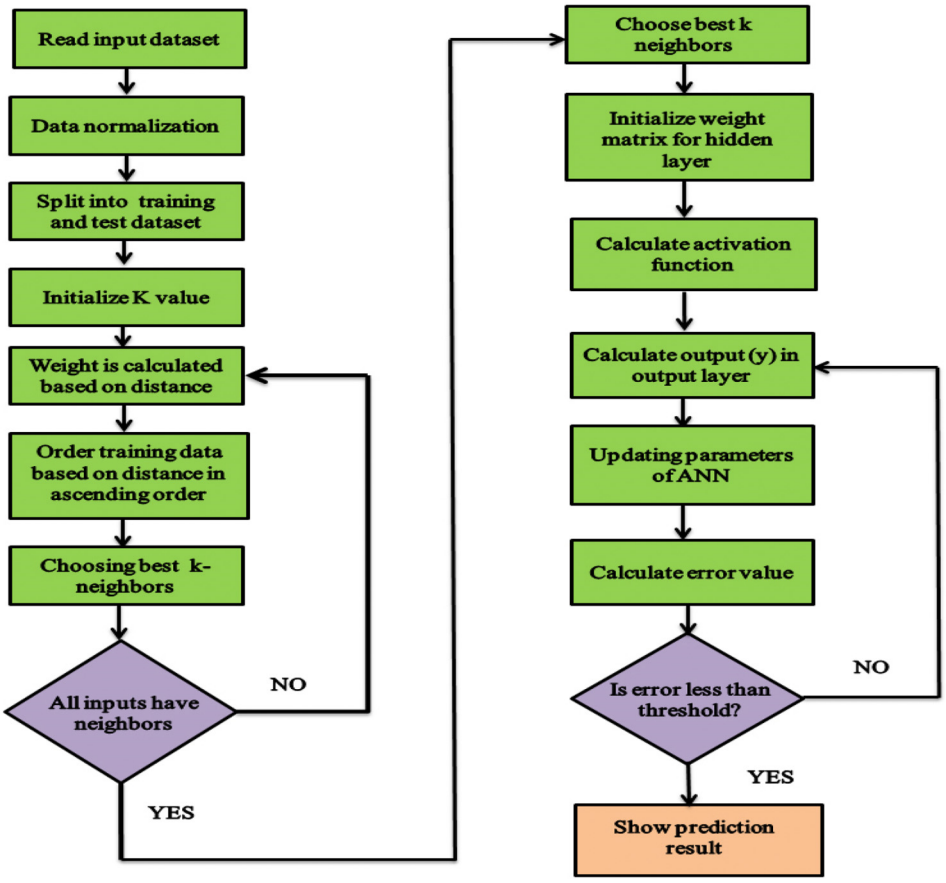


**Fig. 1.** Generalized flow chart of enhanced ANN.

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