Development and performance evaluation of FLANN based model for protein structural class prediction

Bishnupriya Panda, Ambika Prasad Mishra, Babita Majhi and Minakhi Rout

Abstract— ***During last few decades' accurate prediction of protein structural class has been a challenging problem. Efficient and meaningful representation of protein molecule plays a significant role. In this paper Chou's pseudo amino acid composition along with amphiphillic correlation factor has been used to represent protein data. A simple functionally linked artificial neural network has been used for structural class prediction***.

Keywords— **AAC, AmPseAAC, Functional link artificial neural network (FLANN), Protein Domain, Structural Class**

I. Introduction

In molecular biology, sequence to structure prediction plays a significant role. Specifically in protein molecular biology, function of protein molecule is highly dependent on the structure of protein molecule. Protein structural class plays a significant role in determining protein folding. Protein database is growing every day. So a computationally efficient method is highly required for protein structural class prediction. Levitt and Chothia (1976) reported 10 structural class, 4 principal and 6 small classes in a dataset of 31 globular proteins. However biological community recognizes 4 principal classes depending on percentage of alpha helices and beta strand.

α class : Contains more than 45% alpha strand and less than 5% beta strands.

β class : Contains more than 45% beta strands and less than 5% alpha strand.

α+β class : Contains more than 30% alpha helices and more than 20 % beta strands and beta strands are anti- parallel.

α/β class : Contains more than 30% alpha helices and more than 20 % beta strands and beta strands are parallel.

Accurate determination of protein structural class is a two step process: Effective representation of protein sequence and then developing a prediction model. Many in-sillico structural class prediction algorithm and methods have been proposed earlier. Amino Acid Composition (AAC) is highly related to protein structural class [1]. Several classification methods such as distance classifier, principal component analysis [2], Bayesian classifier, fuzzy clustering [3], support vector machine [4] and multilayer artificial neural network [5] have been proposed in the literature. Though many promising results have been achieved, AAC of protein lacks sequence order and sequence length information. Sequence order and sequence length information also play a significant role in predicting protein structural class because amino acid composition do not differentiate between protein molecules of different sequence order and sequence length. In this paper along with pseudo amino acid composition, amphiphillic correlation factors of protein molecule [6] has been used to capture the sequence order information. Many authors have proposed neural network as a good candidate for classification of protein structural class. But how to choose the number of layers and number of neurons in each layer to enhance the classification accuracy is highly complex problem. To alleviate this problem in this paper we propose a low complexity single layer single neuron neural network known as functional link artificial neural network (FLANN) [7] for classification of protein structural class.

This paper is organized as follows: Section II describes the Amino Acid Composition of data and design of amphiphillic pseudo amino acid composition of data using Hydrophilicity and Hydrophobicity of amino acids. Working principle and discussion of functional expansion of all the three types are carried out in Section III. Section IV deals with the results obtained from the simulation study followed by discussions. Finally Section V presents the conclusion of the investigation.

Bishnupriya Panda, Ambika Prasad Mishra' and Minakhi Rout
Dept. of CSE, ITER, Siksha O Anusandhan University
Bhubaneswar, India
e-mail:- panda.bishnupriya@gmail.com/ ambikaprasad.mishra@gmail.com\ minakhi.rout@gmail.com

Babita Majhi
Dept. of CSIT, G. G. Vishwavidyalaya, Central University
Bilaspur, India
e-mail:- babita.majhi@gmail.com

II. Protein Data and Feature Extraction

A. *Amino acid composition (AAC) feature of protein*

Amino acid composition representation of protein molecule is a 20-dimensional feature vector in Euclidian space. The protein x in the composition space is defined as

$$P(x)=\left[p_1(x),p_2(x),.....p_{20}(x)\right] \tag{1}$$

where

$$P_k(x)=\frac{f_k(x)}{\sum_{i=1}^{20} f_i(x)}\, i,k=1,2.....20 \tag{2}$$

B. *Amphiphillic Pseudo amino acid (AmPseAAC) composition feature of protein*

Sequence order and Sequence length information of a protein must be retained. However protein sequence lengths vary widely which poses an additional problem. Chou [9] has proposed an effective way of representing protein character sequence by some of its physiochemical properties. Hydrophobicity and Hydrophilicity of protein molecule play important role in folding of protein molecule.

Suppose a protein molecule is represented by $P_1P_2P_3.......P_l$ where P_1 represents the residue at location 1 along with the sequence and so on. The sequence order effect along with a protein chain is approximated by a set of sequence order correlation factor which is defined as

$$\theta_\tau=\frac{1}{L-1}\sum_{1}^{L-\tau}\theta\left(P_i,P_\tau\right),\left(i=1,2,3....\lambda\right) \tag{3}$$

In (3) L and θ_τ denote length θ_τ order correlation factor. The correlation function $\theta\left(P_i,P_\tau\right)$ is calculated as

$$\theta\left(P_i,P_\tau\right)=H(P_i)*H(P_j) \tag{4}$$

In (3) θ_1 is first tier correlation factor and θ_2 is second tier correlation factor.

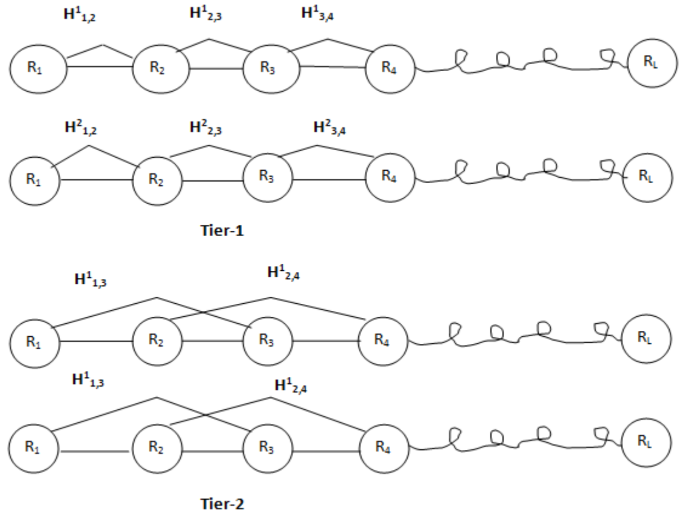


Fig.1 A schematic representation to show sequence order correlation factor using Hydrophobicity and Hydrophilicity values

Before substituting the Hydrophobicity and Hydrophilicity values in (3) they are subjected to some standard conversion. The objective of the conversion is to make the coded sequence zero mean over the 20 amino acids. The standard conversion is described by the following expression:

$$h^1(R_i)=\frac{h_0^1(R_i)-\sum_{k=1}^{20}h_0^1(R_k)/20}{\sqrt{\sum_{u=1}^{20}[h_0^1(R_u)-\sum_{k=1}^{20}h_0^1(R_k)/20]^2/20}}$$

$$h^2(R_i)=\frac{h_0^2(R_i)-\sum_{k=1}^{20}h_0^2(R_k)/20}{\sqrt{\sum_{u=1}^{20}[h_0^2(R_u)-\sum_{k=1}^{20}h_0^2(R_k)/20]^2/20}} \tag{5}$$

So a protein sample is represented as

$$P=\begin{bmatrix} P_1 \\ P_2 \\ . \\ . \\ P_{20} \\ P_{20+1} \\ . \\ . \\ P_{20+2\lambda} \end{bmatrix}$$

where

$$P_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{20} f_i + w\sum_{j=1}^{2\lambda}\theta_j} (1 \le u \le 20) \\ \dfrac{w\theta_{u-20}}{\sum_{i=1}^{20} f_i + w\sum_{j=1}^{2\lambda}\theta_j} (21 \le u \le 20+2\lambda) \end{cases}$$

and w=0.01

C. *Data Set*

The dataset constructed by Chou [10] contains 204 proteins where the dataset constructed by Zhou [11] contains 277 and 498 proteins. The number of protein domains in each class is listed in Table -I.

III. Functional Linked Artificial Neural Network (FLANN)

The Functional Link Artificial Neural Network (FLANN) has been developed as an alternative architecture to the well known Multilayer Perceptron (MLP) network with application to both function approximation and pattern recognition. The FLANN proposed by Pao [12] is a single layer artificial neural network structure, a nonlinear network with simple operations and provides comparable performance as that of multilayer artificial neural network. The weights of the FLANN are updated using simple LMS algorithm as given in (7).

$$w(k+1) = w(k) + 2*\mu*e(k)*\phi \qquad (7)$$

TABLE I. DETAILS OF PROTEIN DATASETS

Dataset	All α	All β	α+β	α/β
204 domain	52	61	46	45
277 domain	70	61	81	65
498 domain	107	126	136	129

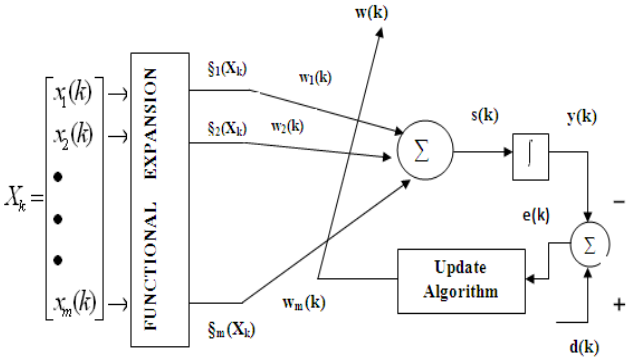


Fig.2 A simple functional link artificial neural network architecture

Fig. 2 shows an adaptive FLANN architecture with one neuron and nonlinear inputs. The nonlinearity in the input is introduced by trigonometric expansions of input values. After nonlinear mapping of the input features simple linear combiner is used to obtain the output which is then passed through a nonlinear function. According to Covers theorem, a complex pattern classification problem cast in a high-dimensional space is more likely to be linearly separable than in a low dimensional space. The functional expansion block makes use of a functional model comprising of a subset of orthogonal sine and cosine basic functions and the original pattern along with its outer products. For an input pattern consisting of (x_1, x_2) can be expanded using trigonometric functions as

$$\phi(x) = [x_1, \sin\Pi(x_1), \cos\Pi(x_1), x_2, \sin\Pi(x_2), \cos\Pi(x_2), ...] \qquad (8)$$

The intermediate output s(k) is calculated as

$$s(k) = \phi(k).W(k) \qquad (9)$$

The final output, y(k) is given as

$$y(k) = \tanh(s(k)) \qquad (10)$$

The output of the FLANN is compared with the target value to give the error value

$$e(k) = d(k) - y(k) \qquad (11)$$

IV. Simulation Study and Discussion

The flow graph of the proposed model can be visualized as:

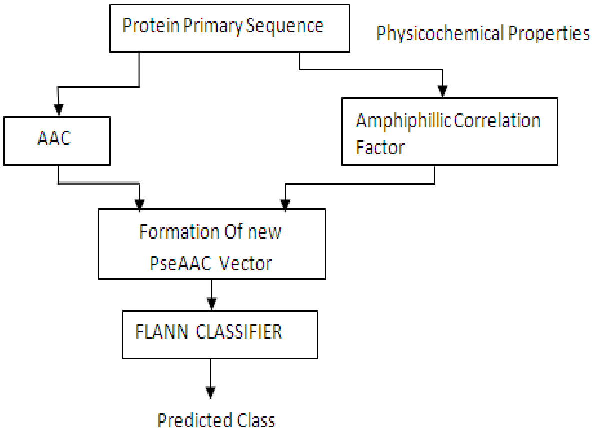

Protein Primary Sequence
Physicochemical Properties
AAC
Amphiphillic Correlation Factor
Formation Of new PseAAC Vector
FLANN CLASSIFIER
Predicted Class

Fig.3 The flow graph of the proposed FLANN based classification scheme

Initially for each of the standard protein dataset the Amino acid composition (AAC) and Amphiphillic Pseudo amino acid (AmPseAAC) composition features are extracted using the formulae as described in Section 2. Then each of the feature pattern is expanded to five terms using the trigonometric expansion. The nonlinearly mapped input pattern is weighted, added together and passed through the activation function, tanh ( ) to give the final value, y(k). The output of FLANN is compared with the target value to produce the error. Then the simple LMS based algorithm is used to update the weights of the classifier and process continues until the error square is equal to zero. The value of the error square corresponding to each iteration is stored and plotted in Figs. 4-6 to show the convergence characteristics for 204, 277 and 498 domain protein dataset respectively.

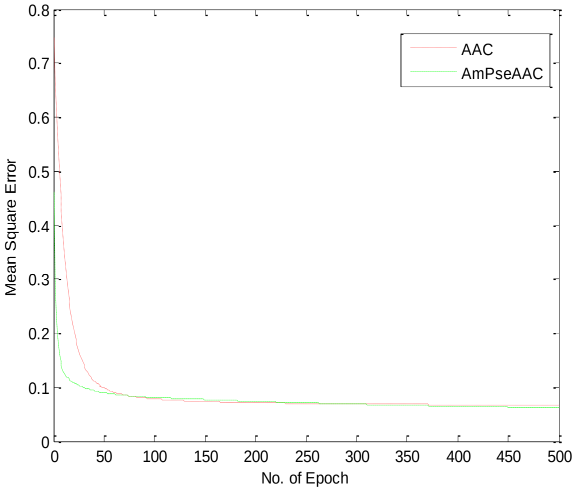


Fig. 4 Comparison of convergence characteristics of FLANN based classifier for the 204 protein domain using AAC and AmPseAAC features

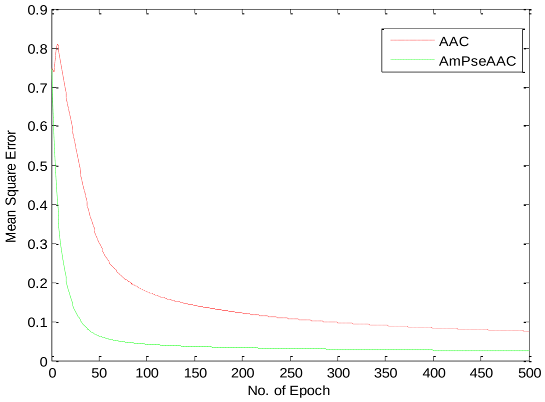


Fig. 5 Comparison of convergence characteristics of FLANN based classifier for the 277 protein domain using AAC and AmPseAAC features

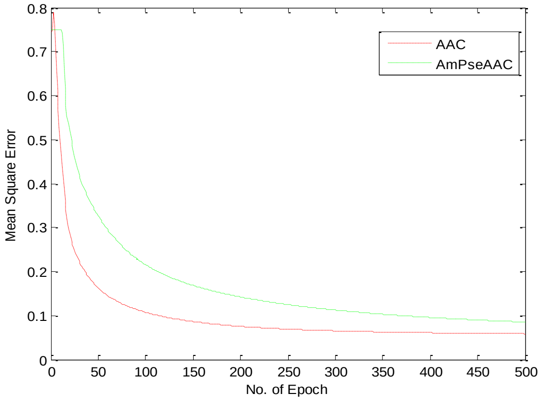


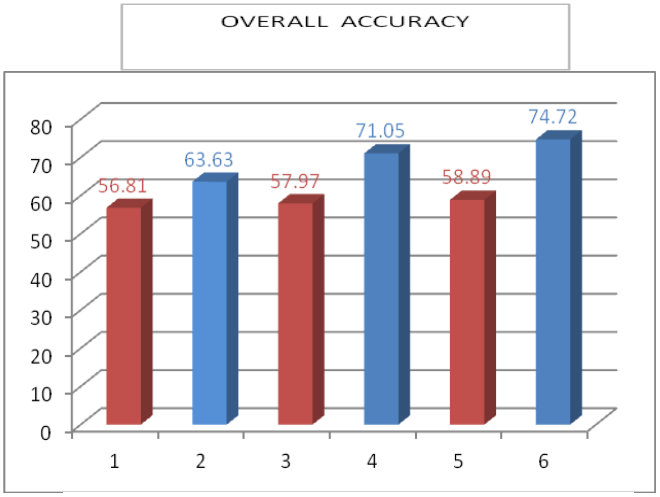
Fig. 6 Comparison of convergence characteristics of FLANN based classifier for the 498 protein domain using AAC and AmPseAAC features

The value of λ is chosen as 10 for finding the correlation factor up to 10th tier to preserve sequence order information. It is observed that the α+β class is more difficult to predict as it contains more variability of helices. The classification accuracy of the model is compared using both AAC and AmPseAAC in all the three data sets. Table-II shows the comparison result of prediction accuracy for 204, 277 and 498 protein domain datasets with AAC and AmPseAAC features. The same is also represented in histogram in Fig. 7. It is evident from Table 2 that AmPseAAC feature representation of the protein data gives more accurate classification result over AAC .

UACEE International Journal of Artificial Intelligence and Neural Networks – IJAINN
Volume 3 : Issue 3 [ISSN 2250 – 3749]
Publication Date : 09 September 2013

TABLE –II COMPARISON OF CLASSIFICATION ACCURACY OF THREE DIFFERENT PROTEIN DOMAIN DATASETS WITH AAC AND AMPSEAAC FEATURES

Data Set	**Features**	**Classification Accuracy**				
		All α	**All β**	**All α+β**	**All α/β**	**Overall Accuracy**
204	AAC	53.8	45	46.6	100	56.81
	AmPseAAC	90	17.6	30	68	63.63
277	AAC	22.7	71.5	74.07	76.9	57.97
	AmPseAAC	100	52.09	66.7	72.3	71.05
498	AAC	75	60	50	62.9	58.89
	AmPseAAC	63	92	52.9	88.5	74.72



204 Domain	277 Domain	498 Domain
1:AAC 2:AmPseAAC	3:AAC 4:AmPseAAC	5:AAC 6:AmPseAAC

Fig. 7 Comparison of overall classification accuracy of three protein datasets using AAC and AmPseAAC features

V. Conclusion

In this paper the low complexity neural network FLANN has been used as classifier. The accuracy of the model is enhanced as AmPseAAC representation of the proteins is supplied to the model. Therefore it is suggested that if more number of features are added then the classification accuracy can further be enhanced.

References

[1] K. C Chou, "A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space," Proteins, vol. 21 (4), pp. 319–344, 1995.

[2] Q. S Du, Z.Q Jiang, W.Z He, D.P Li, and K.C Chou, "Amino acid principal component analysis (AAPCA) and its applications in protein structural class prediction," J. Bimol.Struct.Dynam, vol. 23, pp. 635–640, 2006.

[3] Y.S Ding, T.L Zhang, and K.C Chou, "Prediction of protein structure classes with pseudo amino acid composition and fuzzy support vector machines network," Protein Peptide Lett., Vol. 14, pp. 811–815, 2007.

[4] Y.D. CAI, X.J. Liu., and X. Xu, G.P. Zhou, "Support vector machines for predicting protein structural class," BMC Bioinformatics 2, 3, 2001.

[5] Y. CAI, and G. Zhou, "Prediction of protein structural classes by neural network," Biochimie, vol. 82 (8), pp. 783–785, 2000.

[6] Sitanshu Sekhar Sahu, and Ganapati Panda. "A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction," Computational Biology and Chemistry, vol. 34, pp. 320–327, 2010.

[7] Ritanjali Majhi, G. Panda, and G. Sahoo. "Development and performance evaluation of FLANN based model for forecasting of stock markets," Expert Systems with Applications, vol. 36, pp. 6800-6808, 2009.

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

[9] K.C Chou,"A key driving force in determination of protein structural classes", Biochem. Biophys, Res. Commun, Vol. 264, pp. 216–224, 1999.

[10] G.P Zhou,"An intriguing controversy over protein structural class prediction," J. Protein Chem, 17 (8), pp. 729–738, 1998.

[11] Y.H Pao, "Adaptive pattern recognition & neural networks," Reading, MA: Addison-Wesley, 1989.