Indian Journal of Science and Technology, *Vol 8(16), 63605, July 2015*

ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

A Meta-heuristic Framework for Secondary Protein Structure Prediction using BAT-FLANN Optimization Algorithm

Kailash Shaw* and Debahuti Mishra

Computer Science & Engineering, Institute of Technical Education and Research, Siksha O Anusandhan University, Bhubaneswar - 751030, Odisha, India; kailash.shaw@gmail.com, mishradebahuti@gmail.com

Abstract

Background/Objectives: Proteins are the fundamental units of biology; the mechanism by which primary sequence of proteins is predicted into its secondary structure is not yet accurately achieved. **Methods/Statistical analysis**: In this paper, BAT inspired FLANN (Functional Link Artificial Neural Network) model for protein secondary structure prediction with low computation cost and accuracy has been proposed. The proposed model consists of three different phases; i) First, the primary sequence of amino acid is converted into dynamic matrix for different window sizes then this dynamic matrix is used to derive correlation matrix, ii) Second, FLANN is used to classify each sequence of correlation matrix with different learning parameters and random weights. BAT inspired optimization algorithm has been used to optimize the weight and learning parameters of BAT-FLANN, and (iii) finally, refinement of secondary structure result. **Results:** Experiments were conducted with real datasets of some primary sequence on RS126 and CB396 datasets. Proposed method has been compared with existing DSC, NNSSP, PHD, PREDATOR, ZPRED, MULPRED, SVM models and found to be more promising. **Conclusion/Application:** The proposed method achieves average Q3 accuracy 81.2% and 82.7% for CB396 and RS126 dataset respectively. Moreover the segment overlap (SOV) is 76.1% and 75.3% for CB396 and RS126 dataset respectively.

Keywords: BAT-FLANN, Classifier, Dynamic Matrix, Protein Data Bank (PDB), Secondary Structure

1. Introduction

The phenomena of forecasting the secondary state of a protein from primary structure are referred as Protein Secondary structure Prediction (PSP). To determine the biological function[1] of three dimensions folded structure of amino acid its secondary structure is of great importance. Knowing 3D structure of protein will help in medicine for design of drug and enzymes[2] that shows the importance of PSP in the field of bioinformatics and theoretical chemistry. Protein is built up of 20 different amino acids repeating to form a polymer chain with different characteristics. Basically any primary structure can be distinguished according to their Secondary Structure (SS). Usually they are classified into eight classes: H (α-helix), G (310-helix), I (p-helix), E (b-strand), B (b-bridge), T (turn), S (bend) and – (rest), out of which they are mostly classified as three classes: *helices*, *sheets* and other structural conformations such as *loops*, *turns* and *coils*. The spiral string formed by hydrogen bond between CO and NH are identified as Helix, whereas sheets are formed by stretched polypeptide backbone[3–5]. Most of the researcher state that X-ray crystallography and multi-dimensional magnetic resonance methods are the two most effective methodologies in identifying the 3D protein structure. They can achieve a high precision in the cost of long time period and the molecular weight is also confined to 20,000. In this paper, first a dynamic matrix of sliding window of different sizes has been generated. This matrix serves as the input data for training the FLANN model[6] and the weight of this model has been optimized using BAT algorithm. The paper addresses the

**Author for correspondence*

problem of protein secondary structure prediction[7–10] using a Meta heuristic approach by improving accuracy comparing with the existing model. Results show that, the prediction accuracy achieved is almost 83%.

Protein structure prediction was early made by Chou-Fasman and GOR[11,12] method 1960s and 1970s. There work based on *helix-coil* transition models and predicting *alpha helices*. In 1970s predictions on *beta sheets* were made that relies on statistics and probability parameters generated from known solved structures. The best accuracy predicted by this method is near about 60-65%[3–5]. Whereas with current technology: Neural Nets, SVM, and large databases of known protein structures can achieve upto 80%[13–16]. Probably the earliest attempts at using multiple sequence information for secondary structure were helpful in prediction of alpha sub unit[17–19]. However the use of multiple sequence data is popularised by the work done by Benner & Gerloff[20] for Secondary Structure Prediction (SSP) for the cAMP–dependent kinases. Here an effort is made to discover most conserved regions of a protein sequence buried in core or functionally important for secondary structure prediction. By the help of clustering Benner and Gerloff[20] demonstrated that the residue can be predicted with reasonable accuracy. Rost and Sander [21] had made a break-through by introducing the feed forward neural network strained by back-propagation[22] that replaces the use of human intervention for SSP. Our proposed framework is also uses neural networks but in simplified form that is Functional Link Artificial Neural Network (FLANN) that achieves high degree of prediction accuracy and can be easily tested and run on any common computer system.

The layout of the paper is as follows: background knowledge required for PSP and materials and method is discussed in section 2. Proposed PSP model is discussed in section 3 with the working procedure of BAT, section 4 deals with experimental evaluation and section 5 gives the conclusion and future direction of the proposed work.

2 Materials and Method

2.1 Dataset

Many researchers have opted for RS126 dataset used by Rost, & Sander 1993[21] and CB396 dataset by Cuff, & Barton[23] for creating an effective prediction tool. These datasets are extreme non-homologous in nature RS126 is one of the ideal dataset that helps the researcher for implementing different methodology for PSP. The dataset consist of 23,346 amino acids from 126 non-homologous amino acid sequences. The population of helix, beta strand and coil in terms of percentage is found to be 32%, 21% and 47% respectively. Online database such as PDB is used to download RS126. For better result analysis another dataset CB396 is used. It consists of 396 numbers of non-redundant proteins. Standard DSSP[24] labels is being used in this paper for the training samples. These eight structural classes as per DSSP labels are drop down to three using following methods:

- {H,I,G} → H(Helix), {E,B} → B(Beta Sheet), Rest{S,T,C} → C(Coil)
- {H,G}→ H(Helix), {E} → B(Beta Sheet), Rest{S,T,B,I,C} → C(Coil)
- {H}→ H(Helix), {E}→ B(Beta Sheet), Rest{G,S,T,B,I,C} → C(Coil)
- {H,G}→ H(Helix), {E,B}→ B(Beta Sheet), Rest{S,T,B,I,C}→ C(Coil)

2.2 Methods for Secondary Structure Prediction

There are many methods which are either probabilistic or derived from fusion of statistics and artificial intelligence. Some of them are: a) DSC[24] which applies GOR residue and amino acid position[26] combined with the information from multiple sequence alignment. Linear discrimination with filtering is applied to deduced weights that remove erroneous predictions, b) NNSSP[27] is a technique based upon environmental scoring scheme. It considers N and C terminal positions of *helices* and *strands* for prediction; c) PHD[21] is the ideal architecture consists of three level neural networks with a window of 13 amino acids. It is also known as a structure to structure network that improves prediction accuracy secondary structures. d) PREDATOR[28] embedded with SIM software[29] uses an internal pair for alignment in contrast to global multiple sequence alignment then algorithm is used to predict secondary structure segments and e) ZPRED[18] is based on the GOR method by Garnier J. *et al.*,[26] with an addition to extra calculated conserved value weights[30].

2.3 Assessment of Accuracy

For measurement of accuracy of the prediction we used two methods 1) Average[31] Q_3 and 2) segment overlap (*sov*)[32]. Q_3 measures the overall percentage of predicted residues in terms of (H, E and C), to observe as given in (1).

$$Q_3 = \sum_{(i=H,E,C)} \frac{predicted_i}{observed_i} * 100 \quad (1)$$

Similarly segment overlap computation is performed for each datasets. Segment overlap tries to capture segment prediction and their ignorance level varies from of 35% (random protein pairs) to an average 91% for homologous protein pairs. Segment overlap is calculated by (2)[32]

$$sov = 100 \times \frac{1}{N} \sum_{i\in H,E,C} \sum_{s} \frac{minov\left(s_{obs};s_{pred}\right) + \delta\left(s_{obs};s_{pred}\right)}{maxov\left(s_{obs};s_{pred}\right)} \times len(s) \quad (2)$$

Where, N denotes the total number of residues, *minov* and *maxov* are the minimum and maximum overlap in the extent of the segment. δ is the minimum variation with a ratio of 1.0 where there are only minor deviations at the ends of segments. Per-class accuracy criterion $Q_i^{\%obs}$ as given in (3) for $class_i$ is defined as the percentage of correctly classified residues in the $class_i$, to all residues observed in $class_i$[33]. Where, M_{ii} is the number of residues is observed in $class_i$ and classified as i, and obs_i is the total number of residues observed in $class_i$.

$$Q_i^{\%obs} = 100 * \frac{M_{ii}}{obs_i} \quad (3)$$

3. A Meta-heuristic Framework for Secondary Protein Structure Prediction

The proposed meta-heuristic framework for secondary protein structure prediction is divided into three phases as shown in Figure 1. All the phases are described below.

Phase 1: Protein primary sequence data is collected from PDB RS126 and is stored in FASTA format. Algorithm-I helps in converting primary sequence of protein in dynamic matrix $X = \{x_1, x_2, \ldots, x_n\}$, where, $|x_i| = k$, k is an any odd number also termed as window size of protein primary sequence that is used to feed into network for prediction of primary residue at $k/2$ position. For example, PDB ID: 2MHU has got primary sequence as: "MDPNCSCAAGDSCTCAGSCKCKECKCTSCK" of length l=30. For a window size of l=17 we get; a dynamic matrix of order $n \times m$. where, n and m can be computed using n=L-k+1 and m=k. Table 1 depicts first 17 columns are generated due to primary sequence and 18th column indicates the class to which $k/2$=9th residue (indicated in bold) of every $x_i \in X$ of primary sequence belongs to. This is nothing but the expected secondary structure which our network will be trained with.

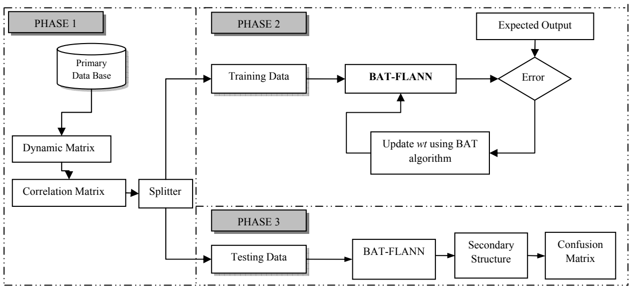


Figure 1. A meta-heuristic framework for secondary protein structure prediction.

Algorithm 1: Creation of Dynamic Matrix

```
function dm=create_dm (ps, k, ss)
where, ps=" MDPNCSCAAGDSCTCAGSCKCKE CKCTSCK", k=17, ss="--SS--SSSSS----TT----SS---GGG-"
dm=[];
[x y]=size(ps);
t=1;
mid=ceil(k/2);
for i=1:x-k+1
    dm(t,1:k)= ps(i:k+i-1)';
    dm(t, k+1)= ss(mid+i-1);
    t=t+1;
end
```

It can be noted that primary sequence is in alphabetical order and this paper proposes a novel way to find correlation among them and achieve correlation matrix. Correlation between the matrixes can be calculated using the following three steps:

Step 1. Finding the impact factor ζ_{xi} for each residue of $x_i \in X$, looking into the Figure 2; impact factor of i^{th} row can be calculated as:

$$\tau_{x_i} = \left(\frac{count\_occurance\_of\left(a_p\right)}{m} | \forall p = 1 : m \, in \, x_i \right) \quad (4)$$

Table 1. Dynamic matrix for PDBID: 2MHU for window size $k = 17$

Sl. No	PDBID: 2MHU for window size $k = 17$																	Secondary structure
1	‘M’	‘D’	‘P’	‘N’	‘C’	‘S’	‘C’	‘A’	**‘A’**	‘G’	‘D’	‘S’	‘C’	‘T’	‘C’	‘A’	‘G’	**‘S’**
2	‘D’	‘P’	‘N’	‘C’	‘S’	‘C’	‘A’	‘A’	**‘G’**	‘D’	‘S’	‘C’	‘T’	‘C’	‘A’	‘G’	‘S’	**‘S’**
3	‘P’	‘N’	‘C’	‘S’	‘C’	‘A’	‘A’	‘G’	**‘D’**	‘S’	‘C’	‘T’	‘C’	‘A’	‘G’	‘S’	‘C’	**‘S’**
4	‘N’	‘C’	‘S’	‘C’	‘A’	‘A’	‘G’	‘D’	**‘S’**	‘C’	‘T’	‘C’	‘A’	‘G’	‘S’	‘C’	‘K’	**‘-’**
5	‘C’	‘S’	‘C’	‘A’	‘A’	‘G’	‘D’	‘S’	**‘C’**	‘T’	‘C’	‘A’	‘G’	‘S’	‘C’	‘K’	‘C’	**‘-’**
6	‘S’	‘C’	‘A’	‘A’	‘G’	‘D’	‘S’	‘C’	**‘T’**	‘C’	‘A’	‘G’	‘S’	‘C’	‘K’	‘C’	‘K’	**‘-’**
7	‘C’	‘A’	‘A’	‘G’	‘D’	‘S’	‘C’	‘T’	**‘C’**	‘A’	‘G’	‘S’	‘C’	‘K’	‘C’	‘K’	‘E’	**‘-’**
8	‘A’	‘A’	‘G’	‘D’	‘S’	‘C’	‘T’	‘C’	**‘A’**	‘G’	‘S’	‘C’	‘K’	‘C’	‘K’	‘E’	‘C’	**‘T’**
9	‘A’	‘G’	‘D’	‘S’	‘C’	‘T’	‘C’	‘A’	**‘G’**	‘S’	‘C’	‘K’	‘C’	‘K’	‘E’	‘C’	‘K’	**‘T’**
10	‘G’	‘D’	‘S’	‘C’	‘T’	‘C’	‘A’	‘G’	**‘S’**	‘C’	‘K’	‘C’	‘K’	‘E’	‘C’	‘K’	‘C’	**‘-’**
11	‘D’	‘S’	‘C’	‘T’	‘C’	‘A’	‘G’	‘S’	**‘C’**	‘K’	‘C’	‘K’	‘E’	‘C’	‘K’	‘C’	‘T’	**‘-’**
12	‘S’	‘C’	‘T’	‘C’	‘A’	‘G’	‘S’	‘C’	**‘K’**	‘C’	‘K’	‘E’	‘C’	‘K’	‘C’	‘T’	‘S’	**‘-’**
13	‘C’	‘T’	‘C’	‘A’	‘G’	‘S’	‘C’	‘K’	**‘C’**	‘K’	‘E’	‘C’	‘K’	‘C’	‘T’	‘S’	‘C’	**‘-’**
14	‘T’	‘C’	‘A’	‘G’	‘S’	‘C’	‘K’	‘C’	**‘K’**	‘E’	‘C’	‘K’	‘C’	‘T’	‘S’	‘C’	‘K’	**‘S’**

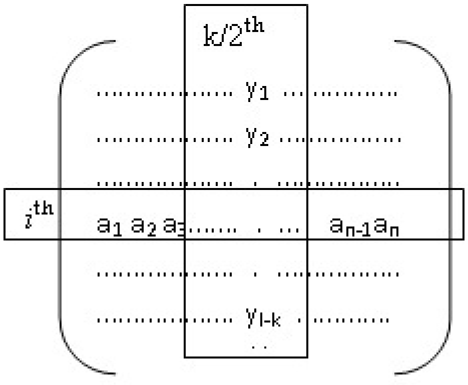


Figure 2. Dynamic matrix used for calculation of ζ_{xi}.

$$\tau_{y_{k/2j}} = \left(\frac{count\_occurance\_of\left(y_j\right)}{n} \middle| \forall j = 1:n \; in \; y_j \right) \tag{5}$$

Step 2: Calculating the mean $\overline{\tau}_{x_i}$ and $\overline{\tau}_{y_{k/2_j}}$

$$\overline{\tau}_{x_i} = \frac{1}{m} \sum_{i=1}^{m} x_i \tag{6}$$

$$\overline{\tau}_{y_{k/2_j}} = \frac{1}{m} \sum_{j=1}^{n} y_j \tag{7}$$

Step 3: Calculating the correlating value ψ_{x_i} for feature x_i as:

$$\psi_{x_i} = \left(\frac{\left(\tau_{x_i} - \overline{\tau}_{x_i}\right)}{\sum_{i=1}^{m} \tau_{x_i} - \overline{\tau}_{x_i} * \sum_{j=1}^{n} \tau_{y_{\frac{k}{2}j}} - \overline{\tau}_{y_{\frac{k}{2}j}}} \middle| \forall i = 1:m, x_i \in X \right) \tag{8}$$

By applying above three steps we get the correlation matrix of PDB: 2MHU for window size 17 as shown in Table 2.

Matrix of Table 2 has been passed through splitter algorithm that splits matrix into two different proportions, one for training and other for testing as shown in Figure 3 and demonstrated in algorithm 2.

Algorithm 2: Splitter

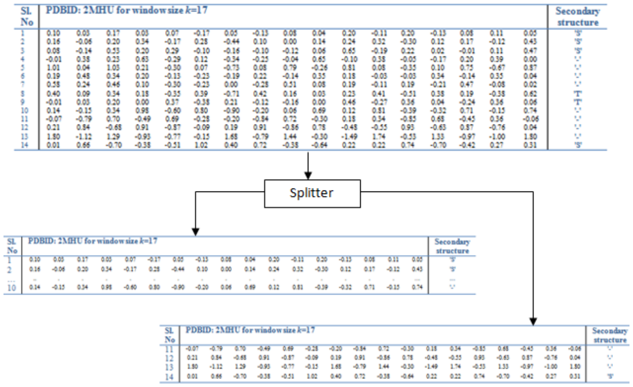
```
function [train_mat test_mat]=splitter(dm,per)
where, dm is the dynamic matrix created by algorithm 1and per is the training percentage
[x y]=size(dm);
train=ceil(x×per/100);
test=x-train;
train_mat=dm(1:train,:);
test_mat=dm(train+1:x,:);
```

Phase 2: Training of BAT-FLANN Model

When, BAT sends signal with pulse rate (sound wave of frequency) 20 kHz to 200 kHz as shown in Figure 4. This signal deflects back after striking the object to BAT

Table 2. Correlation matrix of PDB: 2MHU for window size $k = 17$

Sl. No	PDBID: 2MHU for window size $k = 17$																	Secondary structure
1	0.10	0.03	0.17	0.03	0.07	-0.17	0.05	-0.13	0.08	0.04	0.20	-0.11	0.20	-0.13	0.08	0.11	0.05	‘S’
2	0.16	-0.06	0.20	0.34	-0.17	0.28	-0.44	0.10	0.00	0.14	0.24	0.32	-0.30	0.12	0.17	-0.12	0.43	‘S’
3	0.08	-0.14	0.53	0.20	0.29	-0.10	-0.16	-0.10	-0.12	0.06	0.65	-0.19	0.22	0.02	-0.01	0.11	0.47	‘S’
4	-0.01	0.38	0.23	0.63	-0.29	0.12	-0.34	-0.25	-0.04	0.65	-0.10	0.38	-0.05	-0.17	0.20	0.39	0.00	‘-’
5	1.01	0.04	1.03	0.21	-0.30	0.07	-0.73	0.08	0.79	-0.26	0.81	0.08	-0.35	0.10	0.75	-0.67	0.87	‘-’
6	0.19	0.48	0.34	0.20	-0.13	-0.23	-0.19	0.22	-0.14	0.35	0.18	-0.03	-0.03	0.34	-0.14	0.35	0.04	‘-’
7	0.58	0.24	0.46	0.10	-0.30	-0.23	0.00	-0.28	0.51	0.08	0.19	-0.11	0.19	-0.21	0.47	-0.08	0.02	‘-’
8	0.40	0.09	0.34	0.18	-0.35	0.39	-0.71	0.42	0.16	0.03	0.23	0.41	-0.51	0.38	0.19	-0.38	0.62	‘T’
9	-0.01	0.03	0.20	0.00	0.37	-0.38	0.21	-0.12	-0.16	0.00	0.46	-0.27	0.36	0.04	-0.24	0.36	0.06	‘T’
10	0.14	-0.15	0.34	0.98	-0.60	0.80	-0.90	-0.20	0.06	0.69	0.12	0.81	-0.39	-0.32	0.71	-0.15	0.74	‘-’
11	-0.07	-0.79	0.70	-0.49	0.69	-0.28	-0.20	-0.84	0.72	-0.30	0.18	0.34	-0.85	0.68	-0.45	0.36	-0.06	‘-’
12	0.21	0.84	-0.68	0.91	-0.87	-0.09	0.19	0.91	-0.86	0.78	-0.48	-0.55	0.93	-0.63	0.87	-0.76	0.04	‘-’
13	1.80	-1.12	1.29	-0.93	-0.77	-0.15	1.68	-0.79	1.44	-0.30	-1.49	1.74	-0.53	1.33	-0.97	-1.00	1.80	‘-’
14	0.01	0.66	-0.70	-0.38	-0.51	1.02	0.40	0.72	-0.38	-0.64	0.22	0.22	0.74	-0.70	-0.42	0.27	0.31	‘S’



Sl. No	PDBID: 2MHU for window size k=17																	Secondary structure
1	0.10	0.03	0.17	0.03	0.07	-0.17	0.05	-0.13	0.08	0.04	0.20	-0.11	0.20	-0.13	0.08	0.11	0.05	'S'
2	0.16	-0.06	0.20	0.34	-0.17	0.28	-0.44	0.10	0.00	0.14	0.24	0.32	-0.30	0.12	0.17	-0.12	0.43	'S'
3	0.08	-0.14	0.53	0.20	0.29	-0.10	-0.16	-0.10	-0.12	0.06	0.65	-0.19	0.22	0.02	-0.01	0.11	0.47	'S'
4	-0.01	0.38	0.23	0.63	-0.29	0.12	-0.34	-0.25	-0.04	0.65	-0.10	0.38	-0.05	-0.17	0.20	0.39	0.00	'-'
5	1.01	0.04	1.03	0.21	-0.30	0.07	-0.73	0.08	0.79	-0.26	0.81	0.08	-0.35	0.10	0.75	-0.67	0.87	'-'
6	0.19	0.48	0.34	0.20	-0.13	-0.23	-0.19	0.22	-0.14	0.35	0.18	-0.03	-0.03	0.34	-0.14	0.35	0.04	'-'
7	0.58	0.24	0.46	0.10	-0.30	-0.23	0.00	-0.28	0.51	0.08	0.19	-0.11	0.19	-0.21	0.47	-0.08	0.02	'-'
8	0.40	0.09	0.34	0.18	-0.35	0.39	-0.71	0.42	0.16	0.03	0.23	0.41	-0.51	0.38	0.19	-0.38	0.62	'T'
9	-0.01	0.03	0.20	0.00	0.37	-0.38	0.21	-0.12	-0.16	0.00	0.46	-0.27	0.36	0.04	-0.24	0.36	0.06	'T'
10	0.14	-0.15	0.34	0.98	-0.60	0.80	-0.90	-0.20	0.06	0.69	0.12	0.81	-0.39	-0.32	0.71	-0.15	0.74	'-'
11	-0.07	-0.79	0.70	-0.49	0.69	-0.28	-0.20	-0.84	0.72	-0.30	0.18	0.34	-0.85	0.68	-0.45	0.36	-0.06	'-'
12	0.21	0.84	-0.68	0.91	-0.87	-0.09	0.19	0.91	-0.86	0.78	-0.48	-0.55	0.93	-0.63	0.87	-0.76	0.04	'-'
13	1.80	-1.12	1.29	-0.93	-0.77	-0.15	1.68	-0.79	1.44	-0.30	-1.49	1.74	-0.53	1.33	-0.97	-1.00	1.80	'-'
14	0.01	0.66	-0.70	-0.38	-0.51	1.02	0.40	0.72	-0.38	-0.64	0.22	0.22	0.74	-0.70	-0.42	0.27	0.31	'S'

Sl. No	PDBID: 2MHU for window size k=17																	Secondary structure
1	0.10	0.03	0.17	0.03	0.07	-0.17	0.05	-0.13	0.08	0.04	0.20	-0.11	0.20	-0.13	0.08	0.11	0.05	'S'
2	0.16	-0.06	0.20	0.34	-0.17	0.28	-0.44	0.10	0.00	0.14	0.24	0.32	-0.30	0.12	0.17	-0.12	0.43	'S'
...	..	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	...	...
10	0.14	-0.15	0.34	0.98	-0.60	0.80	-0.90	-0.20	0.06	0.69	0.12	0.81	-0.39	-0.32	0.71	-0.15	0.74	'-'

Sl. No	PDBID: 2MHU for window size k=17																	Secondary structure
11	-0.07	-0.79	0.70	-0.49	0.69	-0.28	-0.20	-0.84	0.72	-0.30	0.18	0.34	-0.85	0.68	-0.45	0.36	-0.06	'-'
12	0.21	0.84	-0.68	0.91	-0.87	-0.09	0.19	0.91	-0.86	0.78	-0.48	-0.55	0.93	-0.63	0.87	-0.76	0.04	'-'
13	1.80	-1.12	1.29	-0.93	-0.77	-0.15	1.68	-0.79	1.44	-0.30	-1.49	1.74	-0.53	1.33	-0.97	-1.00	1.80	'-'
14	0.01	0.66	-0.70	-0.38	-0.51	1.02	0.40	0.72	-0.38	-0.64	0.22	0.22	0.74	-0.70	-0.42	0.27	0.31	'S'

Figure 3. Splitter used for dividing data into training set and testing set.

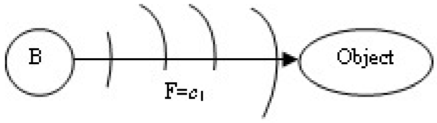


Figure 4. BAT sends sound signal with frequency c_1.

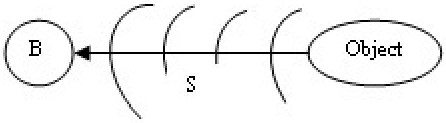


Figure 5. Echo signal use to calculate the distance S.

as echo signal (Figure 5) has been used to calculate the distance S[34,35]. The minimum distance from BAT to any object is the destination of the BAT. BAT flies towards the minimum distance object. BAT reduces its pulse rate when it reaches nearer the object. BAT continues to do so till the distance becomes S=0. Traditional FLANN model have set of input neurons and one output neuron. These neurons are connected with some random weight *wt*. Training data $X = \{x_1, x_2, \ldots, x_r\}$, of order $r \times m$ can be mapped into higher dimensional space ϑ of order $r\times$ $(m \times 3)$ by functional expansion using trigonometric function (9).

$$\vartheta = \left[\left(x_1, sin\pi_{x_1}, cos\pi_{x_1}\right), \left(x_2, sin\pi_{x_2}, cos\pi_{x_2}\right), \ldots, \left(x_r, sin\pi_{x_r}, cos\pi_{x_r}\right)\right] \tag{9}$$

In this paper we embedded the concept of BAT by further increasing the high dimension dataset ϑ by φ as given in equation (10).

$$\varphi = \left[\left(\vartheta_1, f_1\right), \left(\vartheta_2, f_2\right), \ldots, \left(\vartheta_r, f_r\right)\right] \tag{10}$$

Where, f_K is the pulse rate of k^{th} BAT and can be given as (11) where, c_1 is the pulse rate used to control the frequency f_K of BAT B_K. The value of c_1 is auto adjusted in each iteration. Initially, c_1 is set to 0.2.

$$f_k = c_1 * \frac{\sum_{i=1}^{m}\left(x_{ki}\right)}{m} \tag{11}$$

As given in the Figure 6, training data φ is input to the FlANN, weight *wt* of FLANN is randomly taken in between [0.5 to –0.5].

Working of BAT-FLANN can be explained using following steps:

Step 1: Calculation of distance

Distance *S* of the object *z* (here object is the target secondary structure code in terms of 'H', 'E', and 'C') from BAT B_K is calculated by multiplying φ_k with *wt* for each object *z*.

$$S_{object_z} = \varphi_k \text{ x } wt \tag{12}$$

Output of neuron can be calculated using tansigmoid function as given in (13)

$$S_{op_z} = \frac{1}{\left(1+\exp^{-S_{object_z}}\right)} \tag{13}$$

Step 2: Updating position of BAT

Position of BAT with successive iteration tries to reach nearer to object ['H' 'E' 'C']. With each iteration, error is calculated by (14) which helps to update position of BAT using (15).

$$E_k = S_{op_z} - 1 \tag{14}$$

$$P_K = P_K + E_K \tag{15}$$

When BAT starts flying it is assumed that the position is initialize to zero. Its position keeps on changing when it reaches nearer to the object.

Step 3: Updation of frequency f_k and weight wt

As the BAT reaches nearer to its object the frequency decreases. This can be achieved by controlling the value of c_1 of (11) by (16), c_2 is a constant treated as the BAT learning parameter and set to 0.0011 and μ is momentum initialized to 0.5.

$$c_1 = f_k + c_2 * E_k^2 * P_k \tag{16}$$

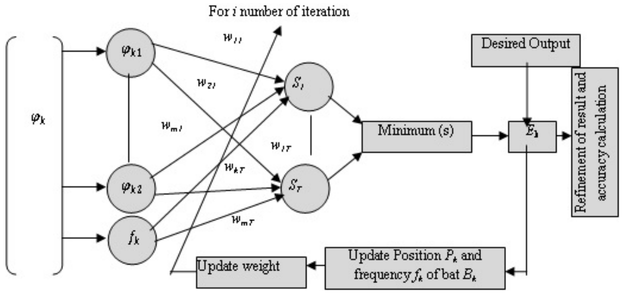


Figure 6. Working procedure of BAT-FLANN model.

$$wt = wt + 2 * \mu * E_k \tag{17}$$

Phase 3: Testing Data is used to Predict Secondary Structure by the BAT-FLANN

This secondary structure is compared with original secondary structure and accordingly confusion matrix is created. As per the DSSP: H(helix) ={G (3_{10} – helix), H (α- helix), I}, B(strands)={E (β-strand), B (β-bridge)}, C(coil, T, S)

{H, I, G} → H (Helix), {E,B} → E(Beta Sheet), Rest{S,T,C} → C(Coil)

According to the DSSP interpretation Table 3 has been formed. Where, first row is the primary sequence whose expected secondary structure is given in second row. DSSP interpretation secondary structure is given at third row. BAT-FLANN output is given in fourth row. DSSP interpretation and predicted output is compared and confusion matrix is created as shown in Table 4.

Accuracy Q_3 can be calculated by using (1)

$$Q_3 = \frac{9}{10} \times 100 = 90\% \tag{18}$$

3.1 Refinement of Result

Refinement of the result is done according to [37] which state that "For each predicted result with less than three consecutive H, it is sure that such prediction result contains some wrong prediction results. For better performance the pattern HXH (X is either E or C) is first converted to HHH.

Table 3. Accuracy comparison

Primary Sequence	A	G	D	S	C	T	C	A	G	S
Secondary Sequence	S	S	S	–	–	–	–	T	T	–
CASP interpretation[34]	C	C	C	–	–	–	–	C	C	–
Predicted O/p	C	C	C	–	–	–	H	C	C	–
Accuracy	1	1	1	1	1	1	0	1	1	1

Table 4. Confusion matrix

	H	E	C	–
H	0	0	0	0
E	0	0	0	0
C	0	0	5	0
–	1	0	0	4

In the patterns EHX/XHE, EHHX/XHHE, EHHHX/XHHHE, all H are converted to E. In the patterns CHC, CHHC and CHHHC, all H are converted to C".

4. Experimental Evaluation

The proposed model has been implemented and tested on MATLAB 10, on Pentium 4 with 2GB RAM. The prediction of protein structure is conducted using machine learning algorithm BAT-FLANN. Datasets RS126[21]and CB396[23] without any preprocessing step is being used for training and testing of BAT-FLANN algorithm followed by evaluation. Figure 7 and Figure 8 shows the Q3 accuracy for two datasets having PDB id: 1gdj with total residue length 153 and 1ppt having residue length 37 belongs to family of RS126. Those two figures clearly reveals that our proposed method surpass the accuracy predicted by other methods. Similarly, Figure 9 and Figure 10 shows the Q3 accuracy for two datasets having PDB id: *3rnt* with total residue length 103 and *1acx* having residue length 107 belonging to family of CB396. The RS126 and CB396 contain 126 and 396 different datasets respectively; therefore it is not possible to show the graph comparison for every data. We have randomly used two datasets out of RS126 and CB396 respectively for display. Q3 accuracy and segment overlap accuracy are evaluated and there average is tabulated in Table 5 and Table 6. This experiment is carried out using six different window sizes (segment length *k*): length 11, 13, 15, 17, 19 and 21. Proposed method is compared with earlier method whose results in the form of Q3 and SOV are shown in Table 7 for both datasets respectively. Earlier methods Q3 and SOV are taken directly from there paper. Looking to Table 5 and 6 it is clear that the length 17 is found to be optimal for local protein structure for both the datasets.

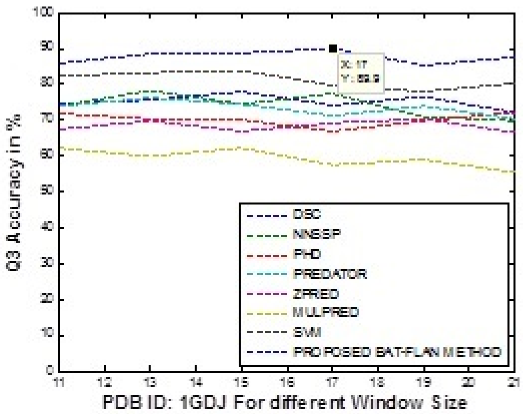


Figure 7. Q3 accuracy graph for *1gdj* data.

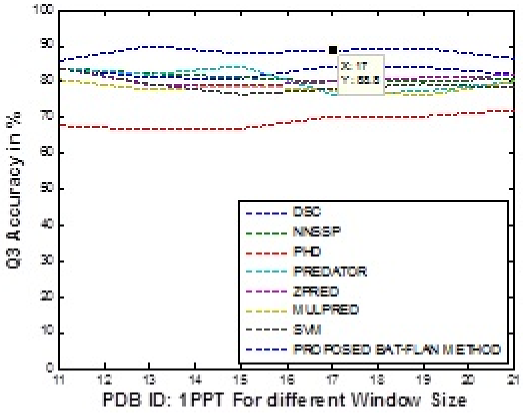


Figure 8. Q3 accuracy graph for *1ppt* data.

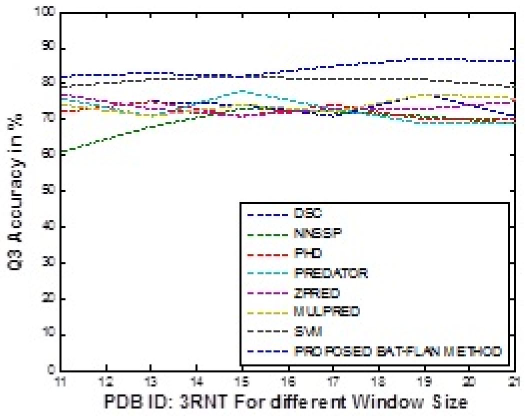


Figure 9. Q3 accuracy graph for *3rnt* data.

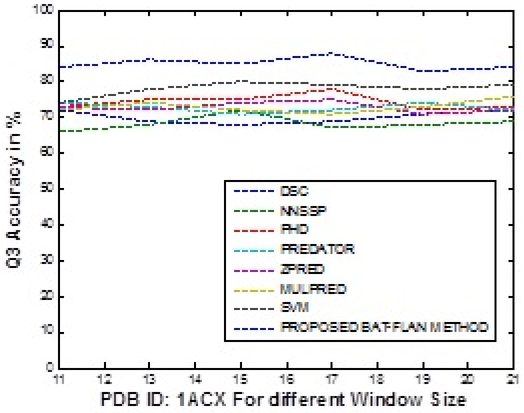


Figure 10. Q3 accuracy graph for *1acx* data.

Table 5. Accuracy per class for protein structure based on secondary structure state for RS126

WINDOW SIZE K	BAT-FLANN			
	$Q_H^{\%}$	$Q_E^{\%}$	$Q_C^{\%}$	Q3
11	0.8	0.82	0.87	0.830
13	0.86	0.81	0.79	0.820
15	0.87	0.79	0.83	0.830
17	**0.87**	**0.84**	**0.88**	**0.863**
19	0.88	0.82	0.76	0.820
21	0.83	0.81	0.76	0.800
Average				0.827

Table 6. Accuracy per class for protein structure based on secondary structure state for CB396

WINDOW SIZE K	BAT-FLANN			
	$Q_H^{\%}$	$Q_E^{\%}$	$Q_C^{\%}$	Q3
11	0.83	0.81	0.74	0.793
13	0.87	0.83	0.75	0.817
15	0.89	0.79	0.77	0.817
17	**0.84**	**0.8**	**0.85**	**0.830**
19	0.89	0.88	0.68	0.817
21	0.87	0.75	0.78	0.800
Average				0.812

Table 7. Q3 and Segment Overlap Results for the Set of RS126, and CB396 Proteins

Methods	**RS126**	**protein set**	**CB396**	**protein**
	Q3	SOV[33]	Q3	SOV[33]
PHD[33]	73.5	73.5	71.9	75.3
DSC[25]	71.1	71.6	68.4	72.0
PREDATOR[26]	70.3	69.9	68.6	69.8
NNSSP[27]	72.7	70.6	71.4	71.3
CONSENSUS[23]	74.8	74.5	72.9	75.4
Zpred[18]	66.7	–	64.8	–
2-StageMSVMs[38]	78.0	72.6	76.3	73.2
Proposed Method	**82.7**	**75.3 (SOV99)**[32]	**81.2**	**76.1 (SOV99)**[32]

5. Conclusion and Discussion

In this paper, an attempt had been made to map Protein Secondary Structure Prediction (PSSP) problem as a classification problem and used proposed BAT-FLANN algorithm for solving it. Proposed method is compared with DSC, NNSSP, PHD, PREDATOR, ZPRED, MULPRED and SVM protein classification techniques. It can be observed that proposed method achieves maximum accuracy 83% for CB396 at window size 17 and 86% for RS126 at window size 17. Whereas rest of the other methods on and average reaches up to 75% of accuracy. Proposed method has been tested with different percentage of training data and got the similar result. One of the major problems in PSSP is that, the data cannot be used directly for classification as it is in character format. For which we used a new and efficient technique that convert these categorical data into numerical forms. Proposed encoding scheme discussed in section Phase 1 of proposed model truly focuses on the impact of every residue over the prediction and accordingly transformed into correlating matrix. Table 5-7 shows the direct output of our classifier after post-processing technique.

6. References

1. Wang G, Zhao Y, Wang D. A protein secondary structure prediction framework based on the Extreme Learning Machine. Neurocomputing. 2008; 72(1-3):262–8.

2. Petit-Zeman S. Treating protein folding diseases. Nature. 2009. Available from: http://www.nature.com/horizon/proteinfolding/background/treating.html
3. Lim VI. Algorithms for prediction of helices and structural regions in globular proteins. J Mol Biol. 1974; 88(4):873–94.
4. Yuksektepe FU, Yilmaz O, Turkay M. Prediction of secondary structures of proteins using a two-stage method. Computers and Chemical Engineering. 2008; 32(1-2):78–88.
5. Rose GD. Prediction of chain turns in globular proteins on a hydrophobic basis. Nature. 1978; 272:586–91.
6. Kumar R, Saithij MSB, Vaddadi S, Anoop SVKK. An intelligent functional link artificial neural network for channel equalization. Proceedings of International Conference on Signal Processing Robotics and Automation; 2009. p. 240–5.
7. Sleator RD. Prediction of protein functions. Functional Genomics. Springer; 2012. p. 15–24.
8. Bettella F, Rasinski D, Knapp EW. Protein secondary structure prediction with sparrow. Journal of Chemical Information and Modeling. 2012; 52(2):545–56.
9. Joo H, Chavan AG, Phan J, Day R, Tsai J. An amino acid packing code for a-helical structure and protein design. Journal of Molecular Biology. 2012; 419(3-4):234–54.
10. Joo H, Tsai J. An amino acid code for b-sheet packing structure. Proteins Structure Function and Bioinformatics. 2014; 82(9):2128–40.
11. Chou PY, Fasman GD. Conformational parameters for amino acids in helical, –sheet, and random coil regions calculated from proteins. Biochemistry. 1974; 13(2):211–22.
12. Chou P, Fasman G. Prediction of the secondary structure of proteins from their amino acid sequence. Adv Enzymol. 1978; 47:45–148.
13. Dor O, Zhou Y. Achieving 80% ten fold cross-validated accuracy for secondary structure prediction by large scale training. Proteins. 2006; 66(4):838–45.
14. Sarhan AM. Cancer classification based on micro array gene expression data using DCT and ANN. Proceedings of International Conference on General of Theoretical and Applied Information Technology; 2009. p. 208–16.
15. Gharehchopogh FS, Khaze SR, Maleki I. A New Approach in Bloggers Classification with Hybrid of K-Nearest Neighbor and Artificial Neural Network Algorithms. Indian Journal of Science and Technology. 2015 Feb; 8(3):237–46.
16. Das K, Ray J, Mishra D. Gene Selection Using Information Theory and Statistical Approach. Indian Journal of Science and Technology. 2015 Apr; 8(8):695–701.
17. Niermann T, Kirschner K, Crawford IP. Prediction of tertiary structure of the alpha-subunit of tryptophan synthase. Biol Chem Hoppe-Seyler. 1987; 368:1087–8.
18. Zvelebil MJJM, Barton GJ, Taylor WR, Sternberg MJE. Prediction of protein secondary structure and active sites using the alignment of homologous sequences. J Mol Biol. 1987; 195:957–61.
19. Mahendran R, Jenifer FJ, Palanimuthu M, Subasri S. Sequence and structural analysis of FtsZ homologs and comparison of bacterial FTsZ with eukaryotic tubulins. Indian Journal of Science and Technology. 2011; 4(2):141–6.
20. Benner SA, Gerloff D. Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure: a prediction of the structure of the catalytic domain of protein kinases. Advan Enzyme Reg. 1990; 31:121–81.
21. Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol. 1993; 232:584–99.
22. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 1986; 323:533–6.
23. Cuff JA, Barton GJ. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. Proteins Struct Funct Genet. 1999; 34:508–19.
24. Kabsch W, Sander C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983; 22(12):2577–637.
25. King RD, Sternberg MJE. Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. Protein Science. 1996; 5(11):2298–310.
26. Garnier J, Osguthorpe DJ, Robson B. Analysis and implications of simple methods for predicting the secondary structure of globular proteins. J Mol Biol. 1978; 120(1):97–120.
27. Salamov AA, Solovyev VV. Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. J Mol Biol. 1995; 247(1):11–5.
28. Frishman D, Argos P. Seventy five percent accuracy in protein secondary structure prediction. Proteins. 1997; 27:329–35.
29. Huang X, Miller WA. A time efficient, linear-space local similarity algorithm. Adv Appl Math. 1991; 12:337–57.
30. Taylor WR. Classification of amino acid conservation. J Theor Biol. 1986; 119:205–18.
31. Schulz GE, Schirmer RH. Principles of Proteins Structure. New York: Springer-Verlag; 1979. p. 1–314.
32. Zemla A, Venclovas C, Fidelis K, Rost B. A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. Proteins-Structure Function and Genetics. 1999; 34(2):220–3.
33. Rost BR, Sander C, Schneider R. Rede finding the goals of protein secondary structure prediction. J Mol Biol. 1994; 235:13–26.

34. Herman H, Gudra T. New Approach in BATs' Sonar Parameterization and Modelling. Physics Procedia. 2010; 3(1):217–24.
35. Altringham JD. BATs Biology and Behaviour. Oxford University Press; 2000.
36. Moult J, Hubbard T, Bryant SH, Fidelis K, Pedersen JT. Critical assessment of methods of protein structure prediction (CASP): round II. Proteins Struct Funct Genet. 1997; S1:2–6.
37. Wang G. A protein secondary structure prediction framework based on the Extreme Learning Machine. Neuro Computing. 2008; 72(1-3):262–8.
38. Nguyen MN, Rajapakse JC. Two-Stage Multi-Class Support Vector Machines to Protein Secondary Structure Prediction. Int J Data Min Bioinform. 2007; 1(3):248–69.