Annals of Medical Research
DOI: 10.5455/annalsmedres.2019.10.658 2020;27(1):348-59

Original Article

A proposal of a hybrid model to predict the secondary protein structures based on amino acid sequences

Muhammed Kamil Turan[1], Hasan Bagci[2]

[1]Karabuk University, Faculty of Medice, Department of Medical Biology, Karabuk, Turkey
[2]19 Mayıs University, Faculty of Department Of Medical Biology, Samsun, Turkey

Abstract

Aim: Predicting the secondary structure of proteins based on amino acid sequences is one of the most significant issues in bioinformatics that requires clarification. A high accuracy in determining the secondary structure is a key to programmatically uncover 3D structure of proteins and for individual drug applications of programmable proteins. The success rates in predicting the secondary structures (Q3 score) were around 0.60 when relevant research was initiated and now the rates have reached to the limit of 0.80.

Material and Methods: In this study, the secondary structure was predicted through 3-state (Helix, Strand and Turn). Artificial neural networks and machine learning algorithms were used as a hybrid model and a framework was developed. The probability of the paired presence of amino acids in sequences was used in digitizing amino acid sequences. Calculations were completed separately for each secondary structural element and the cascade mean filter was used as a threshold method to clarify the differences. The generated matrices were used to digitize the protein sequences. Secondary structure was predicted through the Helix-Strand, Helix-Turn, Strand-Turn, and subsequently, a final decision as Helix, Strand and Turn was reached via machine learning models.

Results: It was determined that the success rates in the dual estimation of secondary structural elements were 0.797 for helix-strand, 0.848 for helix-turn and 0.829 for strand-turn. The average success rate for paired estimation of secondary structural elements was calculated as 0.824. In the proposed model, accuracy was calculated as 0.742 for Helix, 0.703 for Strand and 0.880 for Turn. Q3 score was obtained as 0.775.

Keywords: Protein secondary structure prediction; amino acid encoding; artificial neural networks; machine learning methodes

INTRODUCTION

Predicting the secondary structure from the protein sequence as Helix, Strand and Turn is one of the most significant problems such that bioinformatics science has pursued a solution for a long time (1). The problems, described as the Holy Grail, are also a key for 3D simulations of proteins. Holy Grail was also used for all problems that attempted to extract meaningful information from complex and raw biological data (2). Protein secondary structures were commonly characterized as 3-state. The 3-state addresses Helix, Strand and Turn structures. Helix structure was divided into three sub-groups as 310-helix, α-helix and π-helix and Strand structure was divided into two sub-groups as the isolated bridge and the extended sheet (3). A time-dependent examination of the problem related to predict the secondary structure of proteins reveals three periods. In the first period, the success rate (Q3 score) was 0.60 as a limit value and in the second period it was 0.70 as a threshold value. The third period could be defined as the recent timeframe, in which, success rates over 0.70 were achieved due to the application of deep learning algorithms to the research field (4). An example for the first basic algorithm, with success rate of 0.60, was provided by Chou and Fasman (5), whereas, a basic algorithm until a success rate of 0.70 was GOR (6) and the first algorithm that surpassed success rate of 0.70 was PHD (7). A success rate of 0.70 could be defined as the vertical limit. PSIPRED (8), which used a two-stage neural network structure to predict the secondary structure 2-state (Helix and Strand), with the position specific scoring matrix, developed by PSI-BLAST, achieved the level of 0.765. Further studies

Received: 22.10.2019 **Accepted:** 29.01.2020 **Available online:** 19.02.2020
Corresponding Author: Muhammed Kamil Turan, Karabuk University, Faculty of Medice, Department of Medical Biology, Karabuk, Turkey **E-mail:** kamilturan@karabuk.edu.tr

applied Deep Neural Network algorithms to the field and the success rate exceeded the 0.80 limit (4,9). These studies indicated a highest mean success value of 0.79. Based on the accessed literature and best of our knowledge, the highest success rate is 0.847 (9). Feature vectors generated through running multiple algorithms were reported in studies that employed methodologies such as Decision tree (DT) (10-12), Support vector machine (SVM), Bayesian approach, Gaussian Naive Bayes (GNB) (13,14) and Random forest (RF) (10–14).

The present study focuses on the prediction of the secondary structure based on amino acid sequences. Initially, the secondary structures were interpreted as pairs and their results presented a feature matrix for the prediction of the secondary structure. The generated feature vectors provided to predict the secondary structure of the desired amino acid sequence using Machine Learning (ML) algorithms. Instead of the highest average value used in the literature, the present study used the actual average values obtained via the networks, which were repeatedly trained in different training sets and tested with test sets that were not used during training.

MATERIAL and METHODS

Python programming language (Version 3.7.3) was chosen to develop the program that predicted the secondary structures of proteins. Python programming language was chosen for the program since it was an open source programming language, fast and easy to learn, suitable for scientific programming and data visualization. It was generally preferred in bioinformatics programs and supported object-oriented programming (15,16). The analyses were completed with the Protein Secondary Structure Prediction Framework (PsspF) developed by the authors of the manuscript in Python programming language.

The National Center for Biotechnology Information (NCBI) and The Universal Protein Resource Knowledge base (UniProtKB) websites were used respectively to access the amino acid sequences and the secondary structure element sequences of these sequences (17,18). Pandas, Numpy, SciPy, and StatsModels python frameworks were used for the local storage and processing of the data that was obtained from the official open source websites (19–22). MatPlotLib and SeaBorn open source python frameworks were used for the visualization of the obtained and processed data (23,24). The scikit-learn library was used to predict the secondary structural elements. Such specific framework was preferred since it included simple and effective tools for data mining, data analysis, ML and multi-layered artificial neural networks, it was accessible by everyone and had an interchangeable open source code for everyone, and due to its architecture based on NumPy, SciPy, and MatPlotLib (25,26).

The Basic Database

The data obtained from NCBI and UniProtKB databases were locally stored in a computer as a basic database. The basic database contained non-repeated records of the protein UniProtId (accession number), protein length, amino acid sequence of the specific protein, and the secondary structure element sequence of that protein. The basic database was titled as mainDB.

Secondary Structure Element Databases

The secondary structure element databases were created using the mainDB. The secondary structural elements consist of Helix, Strand and Turn in the proposed model. Therefore, mainDB consists of 3 database tables named as Helix, Strand and Turn, respectively, each for one secondary structural element. Tables of mainDB were composed of UniProtId belonging to proteins, secondary structure element type, segment start point, segment end point, segment amino acid sequence and segment length fields of the proteins. These tables were used by the PsspF, in order to provide the data required for the analyzes.

In methoding section, the PsspF methodology was described through its all processing steps, starting with the first data obtained to making the final prediction, in a sequential order. In short, the PsspF methodology was based on the sequential steps of creating the conditional probability matrices (CPM) for secondary structural elements, the use of multilayer artificial neural network models for paired predictions (Multilayer perceptron classifier, MLPC), the extraction of feature matrices from the results of paired estimation tools (Feature extraction, FE) and making the final prediction from the ML models of the extracted features (Machine learning layer, ML).

Creating the Conditional Probability Matrices (CPM)

Three conditional probability matrices existed in secondary structural elements. One for each secondary structural element was designed. The probability of the paired presence of amino acids at a given window size was calculated. The calculated values were recorded in the relevant field on the size 20 to 20 CPM matrixes. The sequence indexes were based on the alphabetical order of the amino acids. For the paired amino acid Alanine - Tyrosine, the first index in the CPM matrix would be 0 and the second index would be 19, and the CPM [0][19] would yield the value of the Alanine-Tyrosine pair. Equation -1 was used to calculate the probability of the paired presence of amino acids.

$$P_{AB} = \frac{F_A + F_B}{F_{AB}}$$

Equation-1

($P_{"AB"}$) is the probability of the paired presence of amino acids in a given window size (wl=20). Here, F_A represents the number of amino acids, F_B represents the number of B amino acids and $F_{"AB"}$ represents the number of AB amino acids in a given wl.

The created matrices were respectively named as CPM^{HELIX}, CPM^{STRAND} and CPM^{TURN}. Each CPM was normalized to the range of [0,1] using a threshold value calculated by the cascade mean filter method (27,28). Equation-2 was used for normalization.

$$CPM_{Normalized} = \begin{Bmatrix} CPM \leqslant CMF \Rightarrow 0 \\ CPM > CMP \Rightarrow 1 \end{Bmatrix} \cdot CPM$$

Equation-2

The equation used for normalization, where the conditional probabilities of the paired amino acids calculated by CPM for Helix, Strand or Turn secondary structural elements were symbolized by CMF and the threshold value was calculated by the cascade mean filter method.

Binary Prediction Networks

MLPC method was used to predict the paired combinations of the secondary structural elements. Three paired combinations were determined as the Helix-Strand, Helix-Turn and Strand-Turn. Strand-Helix, Turn-Helix and Turn-Strand networks were excluded since they were similar to the previous in terms of classification. The MLPC method was designed as layers of basic processing units called as perceptron. The layer types were designated as an input layer, one or more hidden layers and an output layer, respectively. The input layer was designed as a feature vector that would be applied to the MLPC for classification, and the output layer was designed as the classes that belonged to the problem. It was essential to digitize the amino acids in order to organize them as an input vector within an artificial neural network.

Digitization of the Amino Acid Sequences

Given that amino acids are essentially non-numerical data, they cannot be used as a direct input vector in methods such as machine learning or MLPC. Therefore, they initially need to be digitized. Orthogonal encoding is often employed for such purpose. In orthogonal encoding, each amino acid is expressed as a 20x1 vector. Each vector has 1, in only one amino acid-specific position and has 0 for all other positions (29). Other commonly used methods include two amino acid encodings scheme (30), codon coding scheme (17), amino acids physicochemical properties coding scheme (32). The present study employed normalized CPM matrices in order to digitize the amino acids. The input vector for paired prediction networks was obtained through the orthogonal digitization of amino acids within a given window size, using the CPM^{HELIX}, CPM^{STRAND} and CPM^{TURN} matrices. The sliding window method was used to digitize a segment in any protein. In this method, the first amino acid of the respective segment was aligned over the median amino acid of the window, and at each turn, the window was shifted by one unit, until it reached the last amino acid in the segment. Hence, a sampling matrix that consists of amino acids as much in number as the window size, was obtained based on the number of segment lengths. Numerical information for each line of the sampling matrix could be obtained through digitizing the two adjacent amino acids using CPM matrices. Digitizing a segment containing 20 amino acids with a window containing 12 amino acids provided an input matrix with the size of 20x11. In this study, numbers between 9 and 21 were chosen as a window size. For example, "For example, when window size is selected as 12, 11 paired amino acids are obtained for only this window. With each window is digitized using 3 matrices, an input vector containing 33 parameters are obtained. In this study, numbers between 9 and 21 were chosen as a window size. The final window size has been chosen 16 in this study.

Multilayer Perceptron Classifier (MLPC)

MLPC was used for Helix-strand, Helix-Turn and Strand-Turn paired prediction networks. Figure 1 presents the smallest processing unit of the MLPC schematically. Hyperbolic tangent function was used as an activation function.

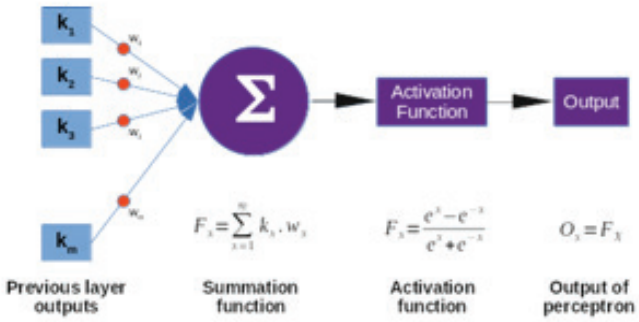


Figure 1. Perceptron schematic, the basic processing unit for the MLPC model

MLPC model topology was generated through using perceptrons in layers. MLPC model associated all perceptrons in a layer through linking them in one to one correspondence with the perceptrons in next and previous layer. Given that the input layer had no previous layer, it was only associated with the first hidden layer. Similarly, the output layer was only associated with last hidden layer. The network weights were designed to be one to one relationship. An example of the MLPC topology is presented in Figure 2.

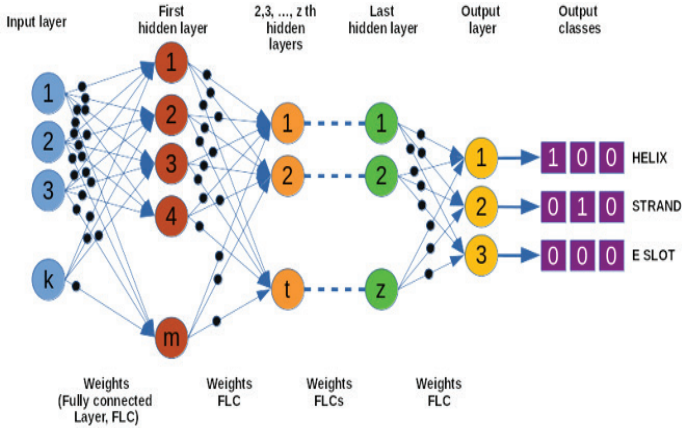


Figure 2. MLPC topology schematic based on the Helix – Strand paired prediction network.

Although there exist two classes in paired prediction networks (only Helix and Strand for Helix -Strand), 3 output classes were used in the MLPC topology employed in the present study. Output layer patterns for paired prediction networks were presented in Table 1.

Table 1. Output layer patterns and empty slots for paired prediction networks

Binary prediction network and their classes		Output layer and their slots		
		Slot 0	Slot 1	Slot 2
Helix – Strand	Helix	1	0	0*
	Strand	0	1	0*
Helix – Turn	Helix	1	0*	0
	Turn	0	0*	1
Strand – Turn	Strand	0*	1	0
	Turn	0*	0	1

*:Empty slots

Paired prediction networks were designed with two and three hidden layers. Literature reported no mathematical formula that clearly stated the number of hidden layers that should be in MLPCs, number of perceptrons that should be in each hidden layer based on the parameters of the problem. Methods such as trial and error, rule of thumb, simple two-phase method and the sequential orthogonal approach were suggested in the literature in order to estimate the number of hidden layers and number of perceptrons (33). The present study adopted the trial and error method. Consequently, networks with two hidden layers and three hidden layers were designed. The perceptron numbers in each hidden layer were chosen between 1 and 20 with the increment as 1.

Performance Criteria

Each paired prediction network was trained with the first 80 percent of the random raw data generated through the mainDB and was tested with the last 20 percent that was not ever introduced to the network during training. Confusion matrix and performance criteria were utilized to determine which network had a better topological structure. The performance criteria employed in the present study were accuracy (ACC), sensitivity (SEN), specificity (SPE), F1 score (F1) and Mathew's correlation coefficient (MCC). The set of equations in Equation-3 were used to determine the performance criteria.

$$ACC = \frac{TP}{TP + FP + FN + TN}$$

$$SEN = \frac{TP}{TP + FN}$$

$$SPE = \frac{TN}{TN + FP}$$

$$MCC = \frac{TN.TP - FP.FN}{\sqrt{(TP + FP).(TP + FN).(TN + FP).(TN + FN)}}$$

$$F_1 = 2.\frac{SEN.SPE}{SEN + SPE}$$

Equation- 3. MLPC performance criteria

The performance measures generated from different topologies for each binary prediction network were calculated using both for the training and test sets. In order to decide which topology was better, results of the performance criteria on test set were sorted ascendingly according to MCC values. The topology with the highest MCC value was named as the best MLPC model. MCC is a performance criterion, defined between the [-1, + 1] interval and is used for ML. A MCC value of -1 indicates that classes were completely inversely predicted, +1 indicates that there was no error throughout the prediction, and 0 indicates that the results were completely in random occurrence. The F1-score is a weighted average, within [0,1] interval, and is used to interpret SEN and SPE values together (34,35). One for each binary prediction network, three best MLPC models were selected and used further in this study. 600 datasets selected randomly from database were trained and tested to state performance criteria of the designed networks in a more accurate way.

Feature Extraction for Machine Learning Algorithms

In order to predict a particular segment in a protein sequence, the results obtained via the digitizing and processing of the segment in binary prediction networks were accepted as the extracted features for machine learning models. Data from the entire segment is used in creating input parameters for machine learning. Input parameters were determined as the cascade mean filter values of empty slots for Helix – Strand, Helix – Turn and Strand – Turn networks, maximum and minimum values for empty slots, Helix – Strand numbers in the order of Helix – Strand network predictions, Helix – Turn numbers in the order of Helix – Turn network predictions, Strand – Turn numbers in the order of Strand – Turn network predictions and the length of protein segment, respectively. This is called the main feature vector and consists of a total of 12 parameters. Final prediction was made by performing rule-based classification via outputs of ML algorithms which accepts the 12 parameters as inputs. In this study, segment length of the protein sequence that was predicted was selected as minimum four.

Hybrid Final Decision Model Combined with Machine Learning

ML algorithms used in this phase of the study were determined as k-Nearest Neighbors (KNN), DT, RF, GNB, Ada boost classifier (ADABC) and Gaussian boost classifier (GBC). Each machine learning algorithm was run

on test sets obtained by fundamental parameters and one having the highest MCC value was used for final decision. Final decision model was obtained by running rules on outputs generated by KNN, DT, RF, GNB, ADABC and GBC algorithms. The rules were used as:

• KNN, DT, RF, GNB, ADABC and GBC algorithms were provided to predict as Helix, Strand or Turn by using fundamental parameters.

• After doing majority voting, if the result is Helix, the given amino acid sequence is directly classified as Helix.

• If the result is either Strand or Turn, new rules are developed by using GBC, KNN and ADABC results. The rules are:

- If GNB predicts as Turn and both KNN and ADABC predict as Helix, the given amino acid sequence is classified as helix.

- If GNB predicts as turn and both KNN and ADABC predict as Strand, the given amino acid sequence is classified as strand.

- In other situations, the given amino acid sequence is classified as turn.

- If GNB predicts as either Helix or Strand and KNN predicts as Strand, the given amino acid sequence is classified as Strand.

- In other situations, the given amino acid sequence is classified as Helix.

The diagram of the developed model was presented in Figure 3.

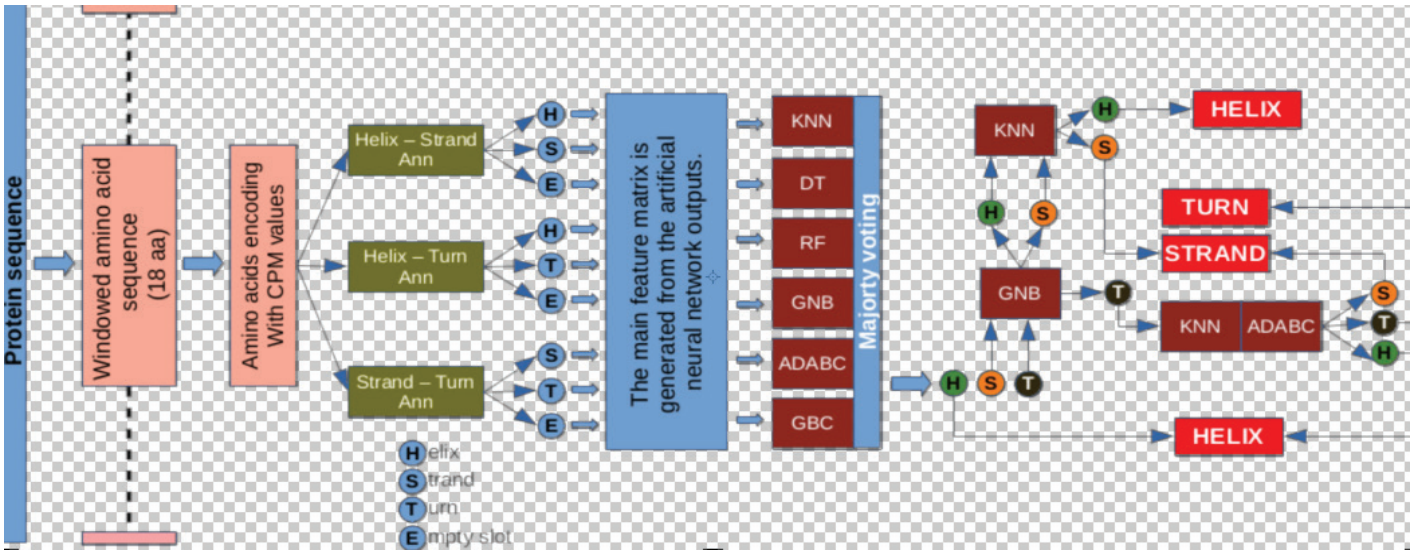


Figure 3. The diagram of the developed model (H: Helix, S: Strand, T: Turn,Ann: Artificial neural network, KNN: k-Nearest Neighbors, DT:Decision tree, RF:Random forest, GNB:Gaussian Navie Bayes, ADABC:Ada boost classifier, GBC:Gaussian boost classifier, H:Helix, S:Strand, T:Turn)

RESULTS

All findings of the present study were presented in a hierarchical order. The present section discusses the amino acid distributions and segment lengths in Helix, Strand and Turn databases, CPM matrices and the disposition of amino acids in secondary structures, paired prediction networks and the final decision results of the hybrid model created by ML algorithms.

Helix, Strand and Turn Secondary Structure Databases

Helix database contains 56,274 records. The protein segment lengths in the Helix database were tested based on suitability for normal distribution by Anderson Darling (AD) and the p value was calculated as 0.0. The minimum value as 3, the maximum value as 148 and the median value as 9 were calculated for the lengths of the Helix segments that were unsuitable for normal distribution. Strand database contains 63,657 records. The p value for the protein segment lengths in the Strand database was calculated as 0.0 and was unsuitable for normal distribution. The minimum value as 3, the maximum value as 48 and the median value as 5 were calculated for the lengths of the Strand segments. Turn database contains 15,420 records. The p values for the protein segment lengths in the Turn database was calculated as 0.0 and were unsuitable for normal distribution. The minimum value as 3, the maximum value as 12 and the median value as 3 were calculated for the lengths of the Turn segments. Results belonging to secondary structure element databases were presented in Table 2 and Figure 3 comparatively.

CPM values were used as digitization matrices. CPM values of all three secondary structural elements (Helix, Strand and Turn) were thresholder at 0.02086, 0.01753 and 0.01558 levels, respectively, based on the cascade mean filter values. The matrices generated due to thresholding were visualized and presented in Figure 4.

Table 2. Secondary structure element databases

Secondary structure elements	Records	The length of protein segments			
		p-value	Max	Min	Median
Helix	56274	0.0	148	3	9
Strand	63657	0.0	48	3	5
Turn	15420	0.0	12	3	3

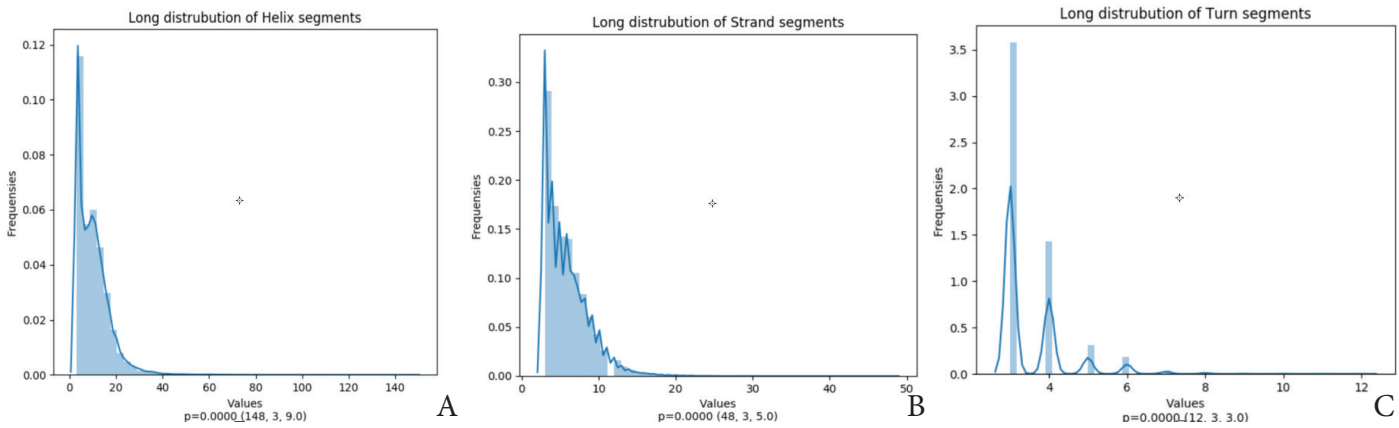


Figure 4. Segment lengths for secondary structural element databases

Helix, Strand and Turn secondary structure databases were used to calculate the distribution of amino acids based on the secondary structure elements. Given that the secondary structure elements were different in length and their amount in the database, selected randomized sets were generated within a certain window size in order to calculate the ratio of amino acids in secondary structure elements. 20 amino acid windows were preferred as the window size and 2000 sets of secondary structure for Helix and Strand and 600 sets for Turn were created randomly and their distributions were calculated through amino acid frequencies. It was observed that distribution of all amino acids in secondary structure were within normal distribution (AD p>0.05).

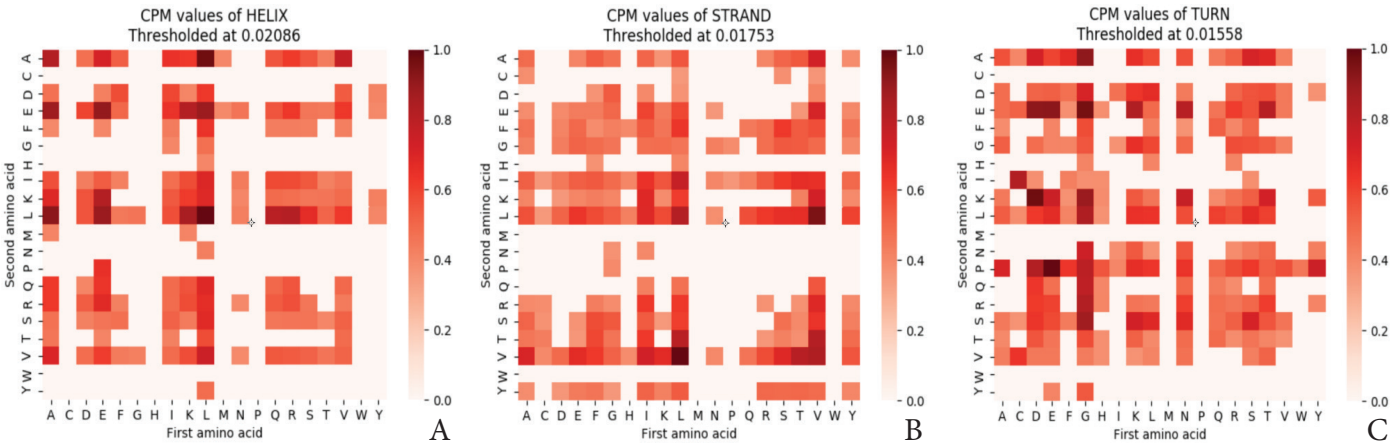


Figure 5. Heat map graphics for the thresholder CPM values. The graphic was plotted for Helix, Strand and Turn denoted as (A), (B) and (C), respectively.

It was determined that the amino acids with a higher probability of presence in the Helix secondary structure were A, E, K, L, M, Q and R. The amino acids with a higher probability of presence in the Strand secondary structure were C, F, I, T, V, W and Y. Within Turn secondary structure elements C, D, G, H, K, N, P, S and T were found to be the amino acids with higher probability of presence. The comparative graph was presented in Figure 6. All secondary amino acids, except C, yielded p <0.05 result in the comparative Student t-test and were found to be

significantly different in terms of their distribution. Amino acid C presented no significant difference in terms of probability of presence in Strand and Turn secondary structural elements (p = 0.635) and a significant difference in terms of probability of presence in the Helix and Strand secondary structural elements (p = 0.0).

Binary Prediction Network Results

The highest MCC values for all models developed for paired networks occurred in topologies with three hidden layers. Each network was tested for different window sizes and the highest MCC values were determined on the set with a window size of 16. The topology with the highest MCC value for the Helix-Strand paired prediction network included 3 hidden layers, with 4, 3 and 6 neurons, respectively, in each hidden layer (MCC = 0.496 ± 0.026). The topology with the highest MCC value for the Helix -Turn paired prediction network included 3 hidden layers, with 2, 12 and 2 neurons, respectively (MCC=0.696± 0.022). Finally, the topology with the highest MCC value for the Strand-Turn paired prediction network included 3 hidden layers and the number of neurons in each hidden layer was 6, 5 and 4, respectively (MCC = 0.666 ± 0.022). Performance criteria for these topologies were presented in Table 3.

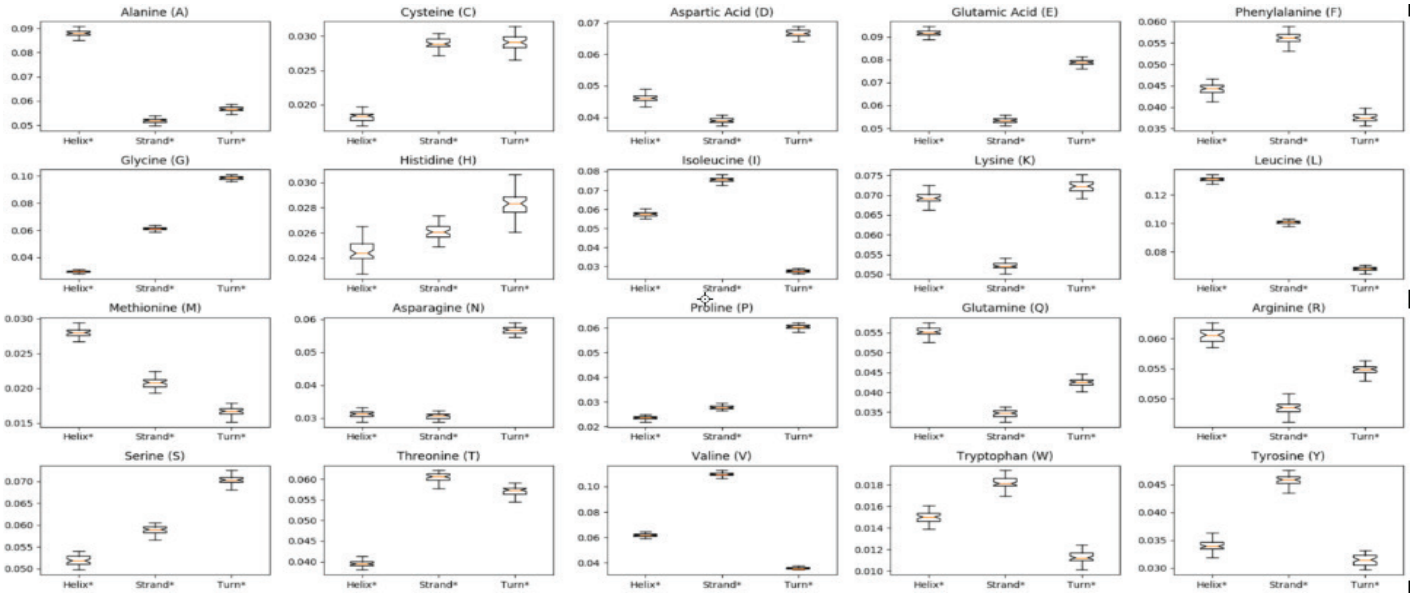


Figure 6. The probability of presence of the amino acids in secondary structural elements. The graph was provided using 2000 sets of 20 amino acids for the Helix and Strand structures and 600 sets of 20 amino acids for the Turn structure. Those, within normal distribution based on AD test, were indicated with *, and their p was calculated as p>0.05. IUPAC one-letter codes of the amino acids were provided in brackets, in the title of the graphs

Table 3. Performance criteria for paired prediction networks (The first number in the topology column represents the number of nodes in the first hidden layer, the second number refers to the nodes in the second hidden layer, and the third indicates the nodes in the third hidden layer. ACC: Accuracy, SEN: Sensitivity, SPE: Specificity, F1: F1-score, MCC: Mathews correlation coefficient. Equal values were presented only once in the table, in order to avoid data repetition.)

Binary Prediction Network	Topology	ACC	SEN	SPE	F1	MCC
Helix	4, 3, 6	0.797±0.018*	0.797±0.018*	0.696±0.021*	0.743±0.014*	0.496±0.026*
Strand			0.702±0.021*	0.794±0.018*	0.745±0.013*	0.498±0.026*
Helix	2, 12, 2	0.848**	0.870**	0.826±0.017*	0.847±0.011*	0.696±0.022*
Turn			0.826±0.017*	0.870**		
Strand	6, 5, 4	0.829±0.012*	0.751±0.019*	0.906±0.013*	0.821±0.012*	0.666±0.022*
Turn			0.906±0.013*	0.751±0.019*		

*: p>0.05 was accepted for AD normality test and mean ± std was provided.
**: p>0.05 was accepted for AD normality test and median value was provided.

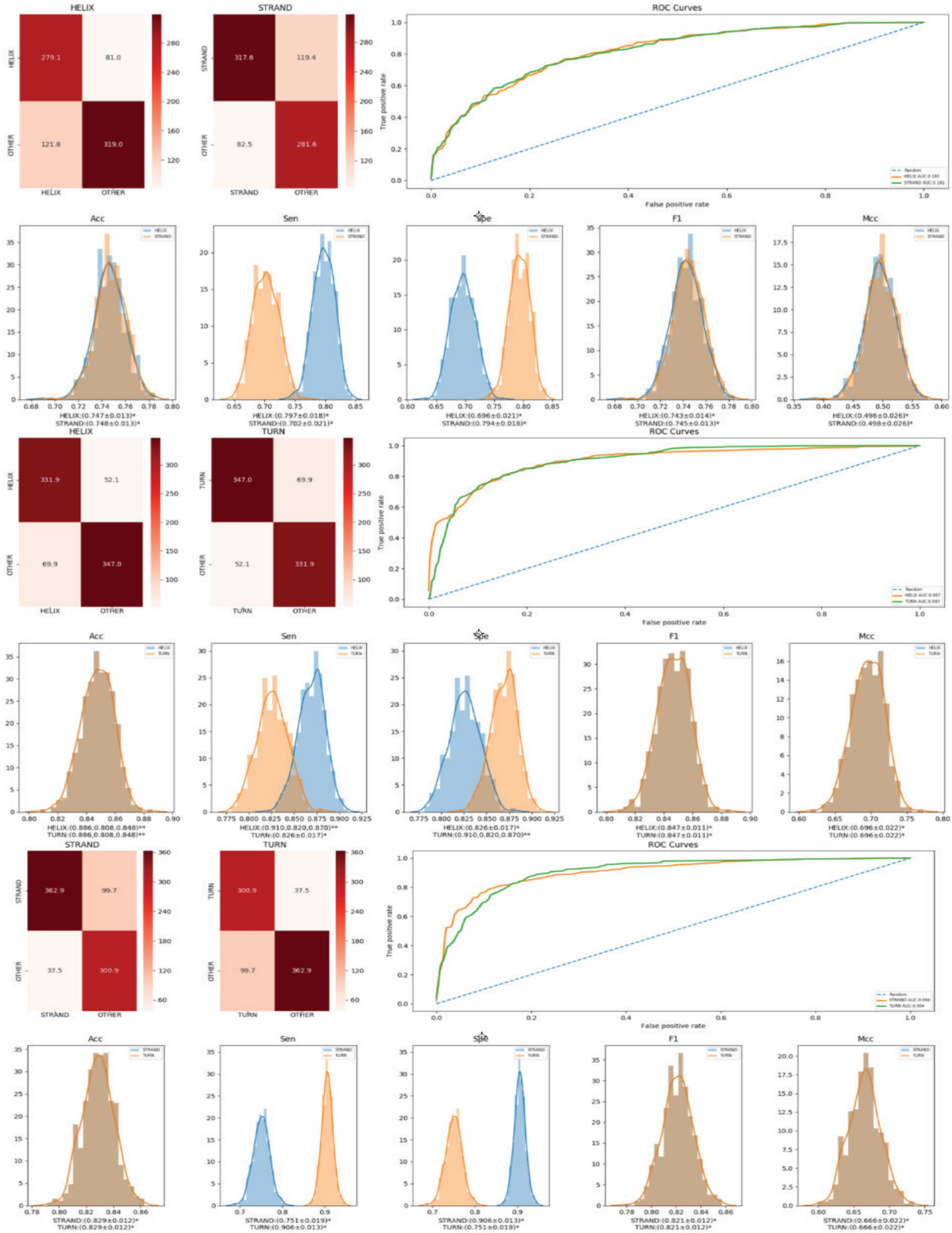


Figure 7. Performance graphs for paired prediction networks

The confusion matrix, performance criteria distribution graphs and receiver operating characteristic curves (ROC) for the best topologies of all three paired prediction networks were presented in Figure -7A, 7B and 7C.

Final Decision Results Combined with Machine Learning

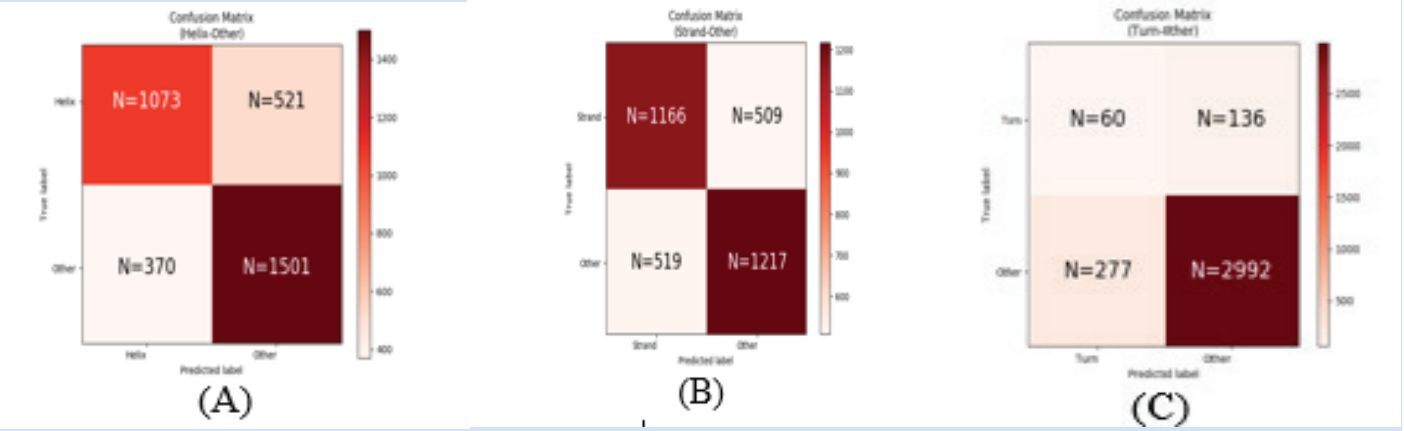
The feature vector obtained from the raw database that contains 600 proteins selected as completely randomized was separated into training and test sets. The first eighty percent data were assigned as the training set and the remaining twenty percent data were designated as the test set. Subsequently, ML algorithms were tried to predict secondary structures of protein segments whose lengths were four or more than four different from the obtained feature vector. As it was foreseen, each three secondary structures were not predicted with high MCC values by only one ML algorithm at once. The results were presented in Table 4. Therefore, final decision hybrid model presented in Figure 4 was used.

Table 4. Final decision performance criteria for the Helix and others classification (KNN: k-Nearest neighbor, RF: Random forest, ADABC: Ada boost classifier, GBC: Gaussian boost classifier, GNB: Gaussian-Navie Bayes, SSE: Secondary structure element, ACC: Accuracy (SEN: Sensitivity, SPE: Specificity, MCC: Mathews correlation coefficient, Other: Merged Strand and Turn structures. Equal values were presented only once in the table, in order to avoid data repetition.)

Algorithm	SSE	ACC	SEN	SPE	F1	MCC
	Helix	0.681±0.007*	0.814±0.013*	0.561±0.016*	0.664±0.009*	0.386±0.012*
KNN	Strand	0.723±0.006*	0.613±0.015*	0.820±0.014*	0.701±0.008*	0.444±0.013*
	Turn	0.940±0.003*	0.036±0.003*	1.00**	0.069**	0.104**
	Helix	0.616±0.008*	0.591±0.014*	0.639±0.014*	0.614±0.008*	0.230±0.015*
DT	Strand	0.902±0.005*	0.201±0.031*	0.960**	0.330±0.043*	0.142±0.028*
	Turn	0.649±0.008*	0.621±0.014*	0.667±0.013*	0.643±0.008*	0.289±0.015*
	Helix	0.678±0.006*	0.740±0.013*	0.622±0.013*	0.676±0.007*	0.364±0.012*
RF	Strand	0.931±0.003*	0.087±0.020*	0.983±0.003*	0.160±0.033*	0.116±0.027*
	Turn	0.713±0.008*	0.653±0.013*	0.766±0.012*	0.705±0.006*	0.423±0.012*
	Helix	0.720±0.006*	0.496±0.012*	0.917±0.007*	0.644±0.010*	0.461±0.011*
GNB	Strand	0.621±0.007*	0.539±0.011*	0.695±0.011*	0.607±0.007*	0.237±0.013*
	Turn	0.736±0.007*	0.877±0.022*	0.727±0.008*	0.795±0.008*	0.308±0.012*
	Helix	0.934±0.004*	0.120**	0.998**	0.214**	0.072±0.033*
ADABC	Strand	0.737**	0.701**	0.801**	0.730**	0.470**
	Turn	0.694***	0.780±0.028*	0.665**	0.693**	0.398**
	Helix	0.680±0.007*	0.779±0.013*	0.591±0.006*	0.697**	0.374±0.013*
GBC	Strand	0.933±0.003*	0.110**	0.995**	0.103±0.029*	0.082±0.028*
	Turn	0.741**	0.632±0.014*	0.802±0.012*	0.707±0.007*	0.442±0.012*

*: p>0.05 was accepted for AD normality test and mean±std was provided.
**: p>0.05 was accepted for AnD normality test and median value was provided.

Table 5. Final results of the proposed algorithm for Helix, Strand and Turn classes (ACC: Accuracy, SEN: Sensitivity, SPE: Specificity, F1: F1-score, MCC: Mathew correlation coefficient, Other: Strand and Turn secondary structures)



Acc	0.742	Acc	0.703	Acc	0.880
Spe	0.742	Spe	0.714	Spe	0.956
Sen	0.743	Sen	0.692	Sen	0.178
F1	0.706	F1	0.694	F1	0.225
Mcc	0.480	Mcc	0.406	Mcc	0.172

Overall Accuracy (Q3) = 0.775

Confusion matrix evaluated at the result of the final decision hybrid model was shown in Table 5 for each secondary structure element

DISCUSSION

In the present study, amino acid sequences of proteins and their known secondary structures were retrieved from NCBI, UniProt web sites and were locally stored in a database. CPM values were used to digitize amino acids and make them utilizable by artificial neural networks. Secondary structures of amino acids digitized via CPM values were first predicted in pairs. The feature vector was developed based on the prediction results of these paired networks and this feature vector was used in training and testing of ML algorithms, which were used for final prediction. Given such characteristic of a hybrid model, PsspF generally surpassed the vertical limit of 0.70.

Digitization of amino acids is highly essential for the prediction of secondary structures. The main motivation for using CPM values in the present study was to prevent overgrowth of the input vector size. All solution networks that were developed had a narrowing structure in terms of topology. The most significant drawback is overfitting (29). Therefore, the present study employed CPM values to keep the input vector length short and to optimal network performance. Furthermore, cascade mean filter value and thresholding of CPM matrices provided the difference between the secondary structures to become more distinct.

The results of the paired prediction networks used to predict the secondary structure were re-run through an artificial neural networks model and the results were used for final decision. The reason for such process was due to the idea that a better outcome than the results of the three paired prediction networks might not be reached.

Rather than offering a single best ACC score in performance measures, using the results generated through repeated training and testing by different training and testing sets was found to be very useful in evaluating the performance of the networks. Thus, it became possible to obtain more detailed information about the success of the network. Furthermore, the use of MCC values to find the best topology was determined as a highly effective approach for evaluating both the SEN and SPE together (35).

The main idea behind using 3 output neurons for paired prediction networks was due to the fact that the training set was provided through using only the main population belonging to those networks for training the paired networks. Once a digitized protein segment was sent to a completed and ready-to-use network, it is probability of belonging to one of the Helix, Turn, or Strand main populations is determined, yet it cannot be clearly determined to which the segment belongs to. Given that the selected segment belongs either to Helix or the Turn main populations, the overall attitude of the Helix – Turn paired prediction network will tend to produce 1,0, and 0

output for Helix and 0,0,1 output for Turn. Hence, in cases that a segment belongs to the main population of Strand, it would be possible for the empty slot to have a non-zero value. Such situation provided flexibility to the model and the advantage to determine the outputs of paired prediction networks in single significance.

In the present study, the average final decision ACC values for the 3-state were established above 0.70 threshold and the CPM values were found to be successful in the digitization of amino acids.

Although a higher overall average of the paired prediction networks, compared to the final decision average success scores, supports the use of binary prediction networks directly as a prediction tool, it is as well essential to emphasize the need for ML algorithms in order to clearly determine the populations that the amino acids belong to

CONCLUSION

In this study, the average final decision ACC values for 3 structures are 0.775. Since it is higher than 0.70 that is vertical limit, CPM values are accepted as successful in the phase of digitization of amino acids. Similarly, the used rules for final decision hybrid model are accepted as successful. While a high specificity value (0.956) was calculated for Turn segment, the fact that sensitivity value was low was due to quite high TN value 2992.

Competing interests: The authors declare that they have no competing interest.
Financial Disclosure: There are no financial supports.
Ethical approval: Consent of Ethics is not required.

Muhammed Kamil Turan ORCID: 0000-0002-1086-9514
Hasan Bagci ORCID: 0000-0002-6216-9835

REFERENCES

1. Narloch PH, Parpinelli RS. The Protein Structure Prediction Problem Approached by a Cascade Differential Evolution Algorithm Using ROSETTA. Brazilian Conference on Intelligent Systems (BRACIS) 2017;294-9.
2. Weng JT-Y, Wu L-C, Chang W-C et al. Novel Bioinformatics Approaches for Analysis of High-Throughput Biological Data. BioMed Res Int 2014;1-3.
3. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577-637.
4. Yang Y, Gao J, Wang J, et al. Sixty-five years of the long march in protein secondary structure prediction: the final stretch? Brief Bioinform 2016;19:482-94.
5. Chou PY, Fasman GD. Empirical Predictions of Protein Conformation. Annu Rev Biochem 1978;47:251-76.
6. Garnier J, Osguthorpe DJ, Robson B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins J Mol Biol 1978;120:97-120.
7. Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol. 1993;232:584-99.
8. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195-202.
9. Jiang Q, Jin X, Lee S-J, et al. Protein secondary structure prediction: A survey of the state of the art. J Mol Graph Model 2017;76:379-402.
10. Selbig J, Mevissen T, Lengauer T. Decision tree-based formation of consensus protein secondary structure prediction. Bioinformatics. 1999;15:1039-46.
11. He J, Hu H-J, Harrison R, et al. Rule Generation for Protein Secondary Structure Prediction With Support Vector Machines and Decision Tree. IEEE Trans Nanobioscience 2006;5:46-53.
12. Yendralwar AA, Waghmare SL, Biyani RM, et al. Bayesian Approach to Prediction of Protein Secondary Structure 2014;5:3375-5.
13. Chawla N, Moore Jr, Bowyer KW, et al. Bagging-like effects for decision trees and neural nets in protein secondary structure prediction. Proceedings of the 1st International Conference on Data Mining in Bioinformatics. Springer, Verlag, 2001;50-9.
14. Lou W, Wang X, Chen F, et al. Sequence Based Prediction of DNA-Binding Proteins Based on Hybrid Feature Selection Using Random Forest and Gaussian Naïve Bayes. PLoS ONE 2014;9:e86703.
15. Python, Python.org. https://www.python.org/ access date 2019.
16. Guzzi PH. Computing Languages for Bioinformatics: Python. Encyclopedia of Bioinformatics and Computational Biology 2019;1:195-8.
17. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/ access date 2019.
18. The Universal Protein Resource Knowledge base. https://www.uniprot.org/ access date 2019.
19. McKinney W. pandas: a foundational Python library for data analysis and statistics. Python for High Performance and Scientific Computing. 2011;14:1-9.
20. van der Walt S, Colbert SC, Varoquaux G. The NumPy Array: A Structure for Efficient Numerical Computation. Comput Sci Eng. 2011;13:22-30.
21. SciPy.org. https://www.scipy.org/ access date 2019.
22. StatsModels: Statistics in Python, statsmodels 0.9.0 documentation. https://www.statsmodels .org/stable/index.html access date 2019.
23. Hunter JD. Matplotlib: A 2D Graphics Environment. Comput Sci Eng 2007;9:90–5.
24. Mwaskom/Seaborn: V0.8.1. https://zenodo.org/record/883859 access date 2019.
25. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res 2011;12:2825-30.
26. scikit-learn: machine learning in Python, scikit-learn 0.21.2 documentation. https://scikit-learn.org/stable/index.html access date 2019.
27. Sehirli E, Turan MK, Demiral E. A randomized automated thresholding method to identify comet objects on comet assay images. Proceedings of the 3rd International Conference on Communication and Information Processing, 2017; 464-7.

28. Turan MK, Yücer E, Sehirli E, et al. Estimation of population number via light activitieson night-time satellite images. ISPRS - Int Arch Photogramm Remote Sens Spat Inf Sci. 2017;103-5.
29. Lin K, May ACW, Taylor WR. Amino Acid Encoding Schemes from Protein Structure Alignments: Multi-dimensional Vectors to Describe Residue Types. J Theor Biol. 2002;216:361-5.
30. Swanson R. A, Vecctor representation for amino acid sequences. Bull Math Biol. 1984;64:623-39.
31. Zamani M, Kremer SC. Amino acid encoding schemes for machine learning methods. IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW), Atlanta, GA, 327-33.
32. Jing X, Dong Q, Hong D, et al. Amino acid encoding methods for protein sequences: a comprehensive review and assessment. IEEE/ACM Trans Comput Biol Bioinform 2018;1-14.
33. Panchal G, Ganatra A, Kosta YP, et al. Behaviour Analysis of Multilayer Perceptronswith Multiple Hidden Neurons and Hidden Layers. Int J Comput Theory Eng. 2011;332-7.
34. Jurman G, Riccadonna S, Furlanello C. A Comparison of MCC and CEN Error Measures in Multi-Class Prediction. Biondi-Zoccai G, editor. PLoS ONE. 2012;7:41882.
35. Raschka S. An Overview of General Performance Metrics of Binary Classifier Systems. arXiv preprint arXiv:1410.5330 2014;1-5.