

Protein secondary structure prediction by a neural network architecture with simple positioning algorithm techniques

Romana Rahman Ema¹, Sharmin Sultana², Shakil Ahmed Shaj², Syed Md. Galib¹

¹Department of Computer Science and Engineering, Faculty of Engineering and Technology, Jashore University of Science and Technology, Jashore, Bangladesh

²Department of Computer Science and Engineering, Faculty of Science and Technology, North Western University, Khulna, Bangladesh

Article Info

Article history:

Received Aug 29, 2021

Revised Feb 11, 2022

Accepted Mar 3, 2022

Keywords:

Neural network architecture

Protein secondary structure

Q3 Prediction

SIMPA technique

ABSTRACT

Protein secondary structure is an immense achievement of bioinformatics. It's an amino acid residue in a polypeptide backbone. In this paper, an innovative method has been proposed for predicting protein secondary structures based on 3-state protein secondary structure by neural network architecture with simple positioning algorithm (SIMPA) technique. Q3 (3-state prediction of protein secondary structure) is a fundamental methodology for our approach. Initially, the prediction of the secondary structure of the protein using the Q3 prediction method has been done. For this, a model has been built from its primary structure. Then it will retrieve the percentage of amino acid sequences from the original sequence to the accuracy of the predicted sequence. Utilizing the SIMPA technique from the 3-state secondary structure predicted sequence, the percentage of dissimilar residues of the three types (α -helix, β -sheet and coil) of Q3 has been extracted. Then the verification of the Q3 predicted accuracy through the SIMPA technique was done. Finally using a new method of neural network, it is verified that the Q3 prediction method gives good results from the neural network approach.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Syed Md. Galib

Department of Computer Science and Engineering, Faculty of Engineering and Technology, Jashore

University of Science and Technology

Jashore-7408, Bangladesh

Email: galib.cse@just.edu.bd

1. INTRODUCTION

Protein secondary structure prediction is the 3-D form of amino acid sequence that based on hydrogen bonding patterns and some geometric constraints. Secondary prediction has focused on the kind of amino acid that a residue's backbone adopts for an individual sequence. In support of this work prospect, three types: α -helix (H), β -sheet (E), and coil (C) have been analyzed [1], [2]. Simple positioning algorithm (SIMPA) is a concept of nearest neighborhood strategy that its traits towards a transition. It is often used to indicate data points based on how its neighbors are classified [3]. Neural networks are complex structures made up of artificial neurons that can capture 1 input and 1 output layer [4]. Here one hidden layer is observed. The number of 10-100 neurons are present here. The input level will be 20 window size order. This will reduce the window size of order window size to the hidden level. Activation function in a neural network which explains how the loaded amount of the input is converted from a node to output at a layer of the network. Nearest neighbor rule is a test case in point of protein structure [5]. Protein secondary structure prediction is usually performed at the input level in the form of sequence profiles and in addition to sequential structure matches [6]. Secondary structure states are divided into three categories: α -helix, β sheet,

and coil. To classify secondary structures into three states, Sander created a define secondary structure of protein (DSSP) algorithm [7]. Artificial neural network (ANN) and recurrent neural networks (RNN) are the two foremost sorts of deep learning architectures. There are two skills needed to train a deep learning network and to adapt the network optimally [8], [9]. The secondary structure of protein can be predicted using the pattern recognition of hydrogen-bonded and geometrical features method [10]. The known sequence and known structure of α and β -hemoglobin are possible correlation between specific amino acid and location of helical and non-helical part [11].

In this paper, the residue of α -helix, β -sheet, and coil from Q3 prediction using the nearest neighbor calculation have been extracted. The prediction accuracy from Q3 is 85%. The SIMPA method was then used to input the predicted protein structure from Q3 prediction. The whole predicted result has not utilized as input to the SIMPA method for convenience's sake. The resulting protein structure was shortened to 7 characters, which is known as the window size. Similarity matrix has been used to compare the two outcomes. In this paper, the exact percentage of residue α -helix (H), β -sheet (E), and coil (C) has been calculated individually. Finally, the residue of result from the Q3 secondary protein prediction and the SIMPA approach both have the same percentage of residue. 100 percent accuracy in terms of residue comparison using the SIMPA approach has been acquired by our test. In the neural network method, the same dataset with window size 7 have been used. After applying the method, verifying the dedication data using SIMPA technique has been done. But training dataset does not verify the actual sequence of the amino acid with the predicted sequence obtained by the neural network method. So, the Q3 prediction method is better than the neural network approach.

The precision of the secondary structure of protein can be calculated in different ways. But the amount of α -helix, β -sheet and coil in this accuracy has not been shown in any method. And no predicted accuracy was shown with verification. Here, it is proposed to predict Q3 prediction and also verified it through SIMPA technique. Residue of α -helix, β -sheet and coil are also shown.

In Q3 prediction method, the prediction accuracy is obtained by comparing the accuracy of the sequence of amino acid with its actual accuracy. SIMPA technique has to be given as input with window size 7 from the predicted sequence. Then the overall score of the predicted sequence has to be calculated. Through the conformation prediction of SIMPA technique, different residues of α -helix, β -sheet, and coil of protein secondary structure can be visible and using those residues the Q3 method can be verified. Here, the update of the weight matrix using neural network approach has been done and prediction of the secondary structure was measured with the help of conformation matrix. Then it is needed to find out the residue using SIMPA technique in the same way as it was done in Q3 method.

2. RELATED WORKS

Chou and Pasman [1] projected conformations of helical and sheet intended in favor of protein polypeptide backbones. Based on hydrogen bonding designs and three secondary structure states, those are characterized aptly. Qian and Sejnowski [4] stated that SIMPA can be established by nearest neighbor calculation in bioinformatics. It is a contrivance intended for predicting secondary protein structure. SIMPA contains a matching matrix that portrays scores used for a single amino acid substitute subsequent to another.

Yoo *et al.* [5] stated that the nearest neighbor rule is a test case in point that classified concurring in the direction of the classification of relative training examples commencing a recognized structures database. Magnan and Baldi [6] proposed a perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity in Bioinformatics. Wang *et al.* [7] anticipated by protein secondary structure that alludes towards the protein of native adaptation of the polypeptide backbone. Secondary structure states are divided into three categories: α -helix, β sheet, and coil. To classify secondary structures into three states, Sander created a define secondary structure of protein (DSSP) algorithm [7].

Holley and Karplus [8] proposed a neural networks system. Complex and recurrent neural networks are the two foremost sorts of deep learning architectures. Spencer *et al.* [9] proposed that, the forecast of the protein secondary structure of the position-specific matrix raised by the PSI explosion and the deep learning network is called DNSS. The paper further said that there are two skills needed to train a deep learning network and to adapt the network optimally. The secondary structure of protein can be predicted using the pattern recognition of hydrogen-bonded and geometrical features method [10]. The known sequence and known structure of α and β -hemoglobin are possible correlation between specific amino acid and location of helical and non-helical part [11].

Based on 3-state prediction and the SIMPA technique, a new method for predicting protein structure has been developed. An accuracy of 85 percent in Q3 prediction for sequences of 128 or fewer amino acids was attained and then the residue percentage using the SIMPA approach was verified.

3. THE PROPOSED METHOD

At first, the Q3 prediction method has been done. Then the SIMPA technique has been applied and lastly, the neural network approach has been used. Below are the three methods.

3.1. 3-state prediction

Q3 prediction algorithms are substantial for predicting protein secondary structure. The accuracy of the Q3 prediction method is 85% which is better than other secondary structure prediction methods. Exact protein structure and function prediction depend somewhat on the accuracy of 3-state protein secondary structure prediction. Q3 prediction works in the following steps: i) collect data from the Kaggle community in tabular, ii) build a model to predict the Q3 secondary structure of a protein from its primary structure, iii) then to use the LSTM layer composed of neural networks, and iv) obtain a Q3 accuracy higher than 85% for sequences with 128 or fewer amino acids. By the following steps, the data set lists chains of protein.PDB id and chain code show the sequence of amino acids and the secondary structure.

3.2. SIMPA technique

SIMPA described a structural classification of protein residue based on the nearest neighbor method's top of their 3-D configuration and sequence similarity [12], [13]. SIMPA method follows: i) fulfilled a conformation matrix, ii) differentiate the test sequence of residues, iii) the contrast of the residues of the three homologues must be compared, iv) calculate the test sequence's general score, and v) allocate a conformation prediction table for each residue.

First, the conformation matrix has been created. To help to construct this conformation matrix, we have used the distance values. We'll learn more about this in the upcoming section. Then the training set was utilized, and the residue was found by comparing the test sequences. Then the general score of the sequence of the test set was calculated. Finally, the estimation of the actual predicted result from conformation prediction was done.

3.3. Neural network approaches

Neural network allows one more approach to apprehend more sophisticated residue interactivity. This approach initially utilized to predict secondary structures and are also based on some of the most effective contemporary strategies in a neural network [14], [15]. Residues of 20 amino acid sequences are allowed in the neural network input layer. Neural network approach follows: i) Regarded small samples of amino acid sequence, ii) the network input size turns on the size of window [16]. input level is 20*window size, iii) create weight matrix, iv) training sequence divided into sub-sequence, v) create hidden layer and calculation the output and vi) put in a compressing reason for all bracing result to obtain the ending production.

At first the 7-window size of amino acid had taken as input dataset which was used in Q3 prediction method. Then a weight matrix was taken that can be updated. Then the training dataset was divided into 7 parts with three window sizes. Then the output by creating the hidden layer was calculated. Then the final output petition as a squashing outcome has been found for all activation output.

4. METHOD

This research has proposed three easy methods based on Q3, SIMPA techniques and neural network architecture intended for protein secondary structure prediction. Combining the Q3 and SIMPA methods, prediction of the secondary structure of the protein was measured. Neural network architecture has added with the two methods. In the Q3 method, the sequence of amino acids is predicted from the primary structure to the secondary structure. Predicting the protein secondary structure is one of the most important and challenging issues in bioinformatics. Neural network techniques have been applied to solve the problem and have achieved considerable success in this research field.

Figure 1 shows the Q3 method of protein secondary structure prediction. First, data from Kaggle dataset was collected. A model was built to predict structure from its primary structure and applied neural network architecture of LSTM layer. In data processing steps, first the evaluation of the model was done and after that training dataset was selected. Then training dataset converted the sequence to numerical format. After that training set stopped processing. Q3 accuracy compared the predicted sequence to actual sequence and got a predicted accuracy on the test dataset. SIMPA is another form of predicting secondary protein structure. It can be found in the nearest-neighbor calculation [17]. Q3 utilizes the sequence of predicted amino acids of the protein secondary structure prediction as data in the SIMPA technique.

Figure 2 shows the SIMPA technique flowchart. The first contained a similarity matrix for it and then compared the three homologues test sequences and residues. Then it calculated the overall score for test

sequence and construction of the conformation matrix was done. After that it allocated the scores in a conformation prediction in the SIMPA technique. Then the final residue result in the test sequence was found. Neighbor base classifiers use some or all of the training set patterns to classify the type of test. These classifiers are involved in finding out the similarities between the test and each pattern in the training set. It contains a similarity matrix that appears in Figure 3, which portrays the result replacement of one by another amino acid.

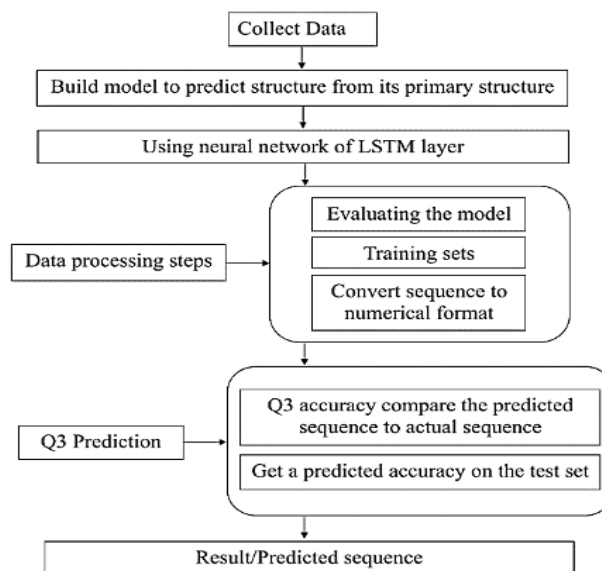


Figure 1. 3-state protein secondary structure (Q3 method)

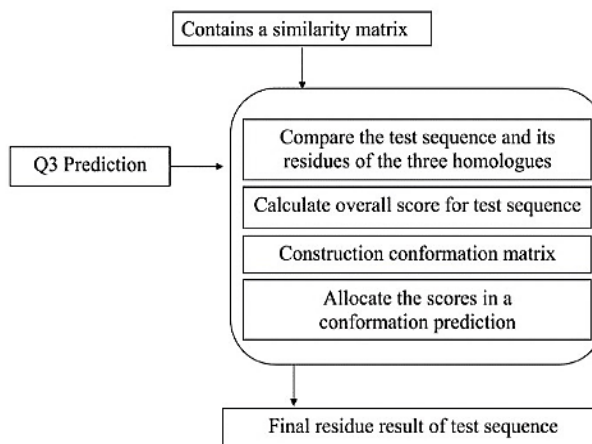
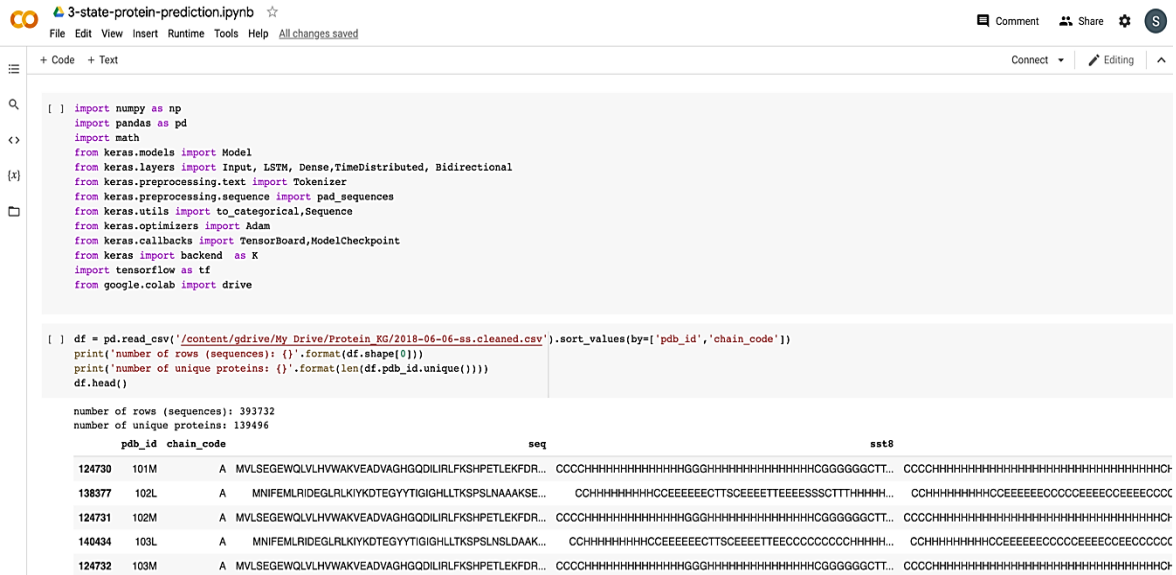


Figure 2. SIMPA technique

Figure 3 shows the similarity matrix used to calculate the overall score in the SIMPA technique. Neural network approaches are essentially an extension of the empirical approach with parameter fitting, although it is a sophisticated approach. They involve mathematical assessments of complex interrelationships within the system. Figure 4 shows the neural network method. Here, the short sequence of amino acids is taken as the dataset. Then created a weight matrix. After training, the dataset is divided into sub-sequences, and it gives the input layer. Finally, it creates a hidden layer and calculates the output.

We completed the entire project on Google utilizing tensor processing units (TPU) hardware acceleration. Python was the primary language we used to program with. In addition, we've utilized a number of other API and libraries. Figure 5 shows that, we have made extensive use of KERAS, especially its layers. And all of this is backed by TensorFlow.

test datasets are used in this paper. In this paper with data from Kaggle datasets [18] have been used. The chain is characterized by both a chain code and the protein ID embedded in the protein database (PDB) [19]. In addition to PDB IDs and chain codes, the Table 1 also shows the sequence of amino acids and secondary structures for a chain.



```

import numpy as np
import pandas as pd
import math
from keras.models import Model
from keras.layers import Input, LSTM, Dense, TimeDistributed, Bidirectional
from keras.preprocessing.text import Tokenizer
from keras.preprocessing.sequence import pad_sequences
from keras.utils import to_categorical, Sequence
from keras.optimizers import Adam
from keras.callbacks import TensorBoard, ModelCheckpoint
from keras import backend as K
import tensorflow as tf
from google.colab import drive

df = pd.read_csv('/content/gdrive/My Drive/Protein_KG/2018-06-06-ss.cleaned.csv').sort_values(by=['pdb_id', 'chain_code'])
print('number of rows (sequences): {}'.format(df.shape[0]))
print('number of unique proteins: {}'.format(len(df.pdb_id.unique())))
df.head()

```

			seq	sst8
124730	101M	A	MVLSEGEWQLVLIHWAKVEADVAGHGQDILIRLFKSPETLEKFRD...	CCCCHHHHHHHHHHHHGGGHHHHHHHHHHHHHHHHHHHHHHHHH...
138377	102L	A	MNIFEMLRIDEGLRLKIYKDTGGYTTIGIHLLTKSPSLNAAAKSE...	CCHHHHHHHHHHCCHEEEECTTSCEEEETEEESSCTTTHHHHH...
124731	102M	A	MVLSEGEWQLVLIHWAKVEADVAGHGQDILIRLFKSPETLEKFRD...	CCCCHHHHHHHHHHHHGGGHHHHHHHHHHHHHHHHHHHHHHHHH...
140434	103L	A	MNIFEMLRIDEGLRLKIYKDTGGYTTIGIHLLTKSPSLNDAAK...	CCHHHHHHHHHHCCHEEEECTTSCEEEETEECCCCCCHHHHH...
124732	103M	A	MVLSEGEWQLVLIHWAKVEADVAGHGQDILIRLFKSPETLEKFRD...	CCCCHHHHHHHHHHHHGGGHHHHHHHHHHHHHHHHHHHHHHHHH...

Figure 5. Setup environment

4.3. Model building after collection

A model was built to predict the 3-states secondary structure of a protein (Q3) from its primary structure. This structure section is hidden layer that throw out different amino acids [20]. The top three-state precision without relying on composition formats ranges from 82 to 84 percent, which was unthinkable just a few years ago. Deep learning algorithms are used to manage large datasets for computational protein design by predicting the potential of 20 amino acids in a single protein [21]. These upgrades are achieved from dynamically proportioned databases and constructions for tutelage of protein sequences, the work of data formatting of secondary structure and extra definitive deep learning procedures.

4.4. 3-state prediction dataset

Q3 method predicts the secondary structure of a protein. The data was collected on the protein structure from research collaboratory for structural bioinformatics (RCSB). It was made available to the Kaggle Community [18]. Dataset lists chains of protein row-wise. Both chain code and protein data bank (PDB) id identify chains. 73% of proteins have less than 3 chains. Furthermore, 10% have more than 4 chains [22]. Consistent inclusion of single observed frequency and pair of residues in the local order of 7. There are 139496 proteins listed in the dataset. Table 1 shows the protein data, which is taken from the Kaggle. By processing this data, extraction was done of the predicted structure in the Q3 prediction method.

Table 1. Protein dataset

PDB id	Chain code	seq	SS8	SS3	Has non-std aa
2BP3	T	LRGSLPTFRSSLFVLWVR	CCCCSSCCEEEEEEE	CCCCCCEEEEEEE	FALSE
2GPV	H	TNMGLEAIIRKALMGK	CCCCHHHHHHHHHH	CCCCHHHHHHHH	FALSE
2JO5	B	AAAAAAIKAIAAIKAG	CCTTTTHHHHHHHH	CCCCCHHHHHHH	TRUE
2N07	X	GHCSDFRNFYDHPHPEIC	CCTTSHHHHHHHCH	CCCCCHHHHHHC	FALSE
2R35	A	RGIVEQCCTSICSLYQL	CCHHHHHHSSCCC	CCHHHHHHCCCC	FALSE
2ZVW	N	GRKRRQTSMTDFYHS	CCCCCBCCGGGTC	CCCCCECCHHHC	FALSE
3OGK	Q	RRASLHRFLEKRKDRV	CCTTHHHHHHHHHC	CCCCHHHHHHHH	FALSE
3U50	L	ARTKQTARKSTGGKAP	CCCCCCCCSCCTT	CCCCCCCCCCCC	FALSE
4H25	F	QHIRCNPKRIGPSKVA	CBCCCCCSCSCC	CECCCCCCCC	TRUE
5DOW	B	KRNKALKKIRKLQKRG	CCCHHHHHHHHHH	CCCHHHHHHHHH	TRUE

5. RESULT AND DISCUSSION

A neural network architecture was presented that utilizes the unified integration of predictions by a 3-state prediction and complex neural network to refine the predictive effectiveness of protein secondary structure. Our suggested neural network has achieved 85% accuracy on the Kaggle dataset for 3-state prediction. Then the SIMPA approach was used to double-check the result, making the total prediction accuracy more definite and error-free. Figure 6 shows the amino acid sequence predicted structure. Using the Q3 method, this predicted structure was measured. The figure also shows the actual structure of the sequence along with the predicted structure.

```

test sequence 1 of 4:
original sequence:
MHHHHHHHMESSDISAMQPVNPKPFLKGLVNHVRVGVKLFNSTEYRGLVSTDNYFNLQLNEA
predicted structure:
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCEEEEECCCCCEEEEEEEEECCCCCEEEEEEE
actual structure:
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCEEEEECCCCCEEEEEEEEECCCCCEEEEEEE
=====
test sequence 2 of 4:
original sequence:
MIQNHKINMTPEICKASRALVNLTKELALMAGIATPTIADFERGARKPHGNLRSIIIAFENKGL
predicted structure:
CCCCCCCCCHHHHHHHHHHCCCHHHHHHHHCCCCCHHHHHHHCCCCCCCCCHHHHHHHH
actual structure:
CCCCCCCCCHHHHHHHHHHCCCHHHHHHHHCCCHHHHHHHHCCCCCCCCCHHHHHHHHHH
test sequence 3 of 4:
original sequence:
PVSPKKKENALLRYLLDKDDT
predicted structure:
CCCCCCCCCHHHHHHHHHCCCCC
actual structure:
CCCCCCCCCHHHHHHHHHCCCCC
=====
test sequence 4 of 4:
original sequence:
ADLEDNMETLNDNLKVEKADNAAQVEKALEKMLAAAADALKATPPKLEDKSPDPEMHDFRHGFAL
predicted structure:
CCHHHHHHHHHHHHHHHHHHCCCHHHHHHHHHHHHHHHHHHHCCCCCCCCCCCCCCCCCHHHHHHH
actual structure:
CCHHHHHHHHHHHHHHHHHHCCCHHHHHHHHHHHHHHHHHHHCCCCCHHHCCCCCCHHHHHHHHHH

```

Figure 6. Amino acid sequence prediction

5.1. 3-State prediction or Q3 method

In this model, the dataset applied into four amino acid sequences. By predicting this sequence, better accuracy has been achieved than the actual sequences and it gives the output of residue α -helix (H), β -sheet (E), and coil (C) individually. Once preprocessing the input sequence, it was conceded into an LSTM layer that returns a 100-dim sequence. Then it goes throughout a dense layer and finally, a SoftMax activation function is applied in order to predict the probability of each of the 3 states at every time step. Improvement have come from an increasingly large database of protein sequence & training structure [23]. Q3 accuracy compares the predicted sequence to the real Q3 structure element by part and returns the ratio of matches. Comparable accuracy was got on the test set.

From the Figure 7, comparable accuracy on the test set has been seen. The accuracy of Q3 prediction with a neural network reaches 85% Figure 7 shows the Q3 predicted accuracy. That is extracted from the actual structure and predicted structure of the amino acid sequence. The result obtained in the Q3 method focused on not getting a good result in the neural network.

```
mdl.evaluate(x_test,y_test)
186/186 [=====] - 8s 41ms/step - loss: 0.2455 - q3_acc: 0.8549
[0.24547389149665833, 0.8549241423606873]
```

Figure 7. Final evaluated Q3 accuracy

5.2. SIMPA technique

In terms of precision, SIMPA is the most accurate. Since it effectively traits compliance to a residue in the most recent trial on the basis of nearest neighbor residues with recognized conformations in homologues, it is a shape of nearest neighbor system. Our method is based on the idea of using a sort order window of certain size. Here, work was done with 7 window sizes from the amino acid sequence used in the Q3 prediction. The conformation matrix was developed, the expanse values from Figure 3 were utilized. The similarity between MIQNHK and MHHHHHH=2+(-1)+0+0+2+0+0=3. We have calculated the other two homologues for PVSPKKK=3+2+2+0+2+0+0=9 and ADLEDNM=2+2+2+2+0+3+2=13.

Figure 8 shows the residue of protein secondary structure for the SIMPA technique [24]. Here it is seen that the residue is absent in α -helix (H) and β -sheet (E) and 100% present in coil (C). By this, it can be said that all the coils exist in the predicted structure, which is the same as the actual structure of the sequence. So, the result obtained in the Q3 method is verified with the SIMPA technique.

The C adaptation is present in all three homologues for their first residue. As a result, three scores are embedded under the C column, indicating that this residue is supported. This process is continued for all the residues. If we predict the complete amino acid sequence, then calculations are made for residue 1 to 7 window size, then for 2 to 8 with the prediction made from the previous, then for 3 to 9, and so on.

```
result_table = print_result ('CCCCCC', 'CCCCCC', 'CCCCCC', [3, 9, 13])
print (tabulate (result_table [1:], result_table [0], tablefmt="grid"))
```

	H	E	C
Residue 0	(0%)	(0%)	3+9+13(100.0%)
Residue 1	(0%)	(0%)	3+9+13(100.0%)
Residue 2	(0%)	(0%)	3+9+13(100.0%)
Residue 3	(0%)	(0%)	3+9+13(100.0%)
Residue 4	(0%)	(0%)	3+9+13(100.0%)
Residue 5	(0%)	(0%)	3+9+13(100.0%)
Residue 6	(0%)	(0%)	3+9+13(100.0%)

Figure 8. Residue table for protein secondary structure

5.2.1. Neural network approach

In terms of precision SIMPA: the same dataset has been used in the Q3 prediction method with window size is 7. "MHHHHHH", "MIQNHK", "PVSPKKK", "ADLEDNM" are small representative of amino acid sequence dataset. The window size of three is taken for the sequence "MHHHHHH", arrangement of size three will be '-MH', 'MHH', 'HHH', 'HHH', 'HHH', 'HHH', 'HH-'. The weight of a sample matrix is estimated.

$$\begin{aligned} W_{11} &= 1 & W_{12} &= 0 & W_{13} &= -1 \\ W_{21} &= -1 & W_{22} &= 1 & W_{23} &= 0 \\ W_{31} &= 0 & W_{32} &= -1 & W_{33} &= 1 \end{aligned}$$

Input represented the conformation matrix sequence [25]. Calculate the output for sub-sequence ‘-MHH’.

Figure 9 shows the reduced matrix that is used to solve the elimination. It reduced the row echelon. Here reduced matrix is used for calculating the output. The activation function defines the output of input or set of inputs.

$$\begin{aligned} \text{Activation output} &= 0*W_{11}+0*W_{12}+1*W_{13}+0*W_{21}+1*W_{22}+0*W_{23}+1*W_{31}+0*W_{32}+0*W_{33} \\ &= 1*0+0*0+(-1)*1+(-1)*0+1*1+0*0+0*1+(-1)*0+0*0 \\ &= 0 \\ \text{Final output} &= 1/(1+e^{-\text{activation}}) \\ &= 1/2 \\ &= 0.5 \end{aligned}$$

We learned from Agarwal and Rizvi [26],

```

If result>0.5 later
    Presume α-helix mean by "H"
Else if result>0 & result<=0.5
    Presume β-sheet mean by "E"
Else
    Presume coil mean by "C"
    
```

In the similar way we can compute the values of windows ‘MHH’, ‘HHH’, ‘HHH’, ‘HHH’, ‘HHH’, ‘HH-’ is 0.12,0.5,0.5,0.5,0.5,0.73. Whose prediction structure is “EEEEEEH”. Which does not match the actual sequence of the amino acid sequence. Similarly, we get from the rest of the sequences:

```

“MIQNHK”---→”EEEEHEE”
“PVSPKKK”--→”EEEEHEH”
“ADLEDNM”--→”EEHEEHH”
    
```

These predicted sequences are obtained using the SIMPA method. Table 2 shows the residue for the neural network approach. Here the residue of the same sequence in the neural network method has been found. But the coil is missing even though the α-helix (H) and β-sheet (E) is present in that table. Only coil is present in the actual structure of the amino acid sequence. The structure obtained from this residue table does not match the actual structure. Thus, the neural network method cannot be verified with the SIMPA technique. From this, it can be said that the Q3 (85%) prediction method gives better results than the neural network method.

	-	M	H
H			X
M		X	
-	X		

Figure 9. Reduced matrix

Table 2. Residue for neural network approach

	H	E	C
Residue 1		3+9+13(100%)	
Residue 2		3+9+13(100%)	
Residue 3	13(33%)	3+9(66%)	
Residue 4		3+9+13(100%)	
Residue 5	9(33%)	3+13(66%)	
Residue 6	13(33%)	3+9(66%)	
Residue 7	3+9+13(100%)		

6. CONCLUSION

In this paper the proposed attempt to improve prediction included two predictive models based on 3-state protein secondary structure prediction and the SIMPA technique with the neural network are described. Here, the Kaggle dataset have been used to predict the protein secondary structure. The neural network architecture connects the 3-state protein prediction and SIMPA technique to enhance the model and showed that the Q3 prediction method is much better than the neural network approach. The Q3 prediction method accuracy is 85%. The model can also predict erstwhile sequences and is not inadequate to bioinformatics nuisance. Q3 method can model complex sequence-structure relationships by neural network architecture and exploit unified secondary structure labels. 3-state prediction and the SIMPA technique are even better than the neural network method and straightforward. The exploratory result showed that the Q3 and SIMPA technique's overall performance was better than neural network method. Because α -helix (H) and β -sheet (E) are present in the predicted sequence obtained by the neural network approach, but coil (C) are present in the actual sequence of the input dataset. So here, the neural network approach seems not perfect to predict the correct estimated sequence.

The prediction has been completed predominantly for identified protein structures that are already available. Since our result is promising, we can expand it further. It is conceivable to predict the secondary structures for the unidentified protein structure based on this approach. With the avail of some other neural network architecture e.g. feed forward, reductive deep learning plus some further refined dataset e.g., CASP11 CASP12, this presage methodology can avail the current state-of-the-art bioinformatics research fields.




REFERENCES

- [1] P. Y. Chou and G. D. Pisman, "Conformational parameters for amino acids in helical, β -sheet, and random coil regions calculated from proteins," *Biochemistry*, vol. 13, no. 2, pp. 211–222, 1974, doi: 10.1021/bi00699a001.
- [2] G. Pollastri, D. Przybylski, B. Rost, and P. Baldi, "Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles," *Proteins: Structure, Function and Genetics*, vol. 47, no. 2, pp. 228–235, 2002, doi: 10.1002/prot.10082.
- [3] T. M. Yi and E. S. Lander, "Protein secondary structure prediction using nearest-neighbor methods," *Journal of Molecular Biology*, vol. 232, no. 4, pp. 1117–1129, 1993, doi: 10.1006/jmbi.1993.1464.
- [4] N. Qian and T. J. Sejnowski, "Predicting the secondary structure of globular proteins using neural network models," *Journal of Molecular Biology*, vol. 202, no. 4, pp. 865–884, 1988, doi: 10.1016/0022-2836(88)90564-5.
- [5] P. Yoo, B. Zhou, and A. Zomaya, "Machine learning techniques for protein secondary structure prediction: An overview and evaluation," *Current Bioinformatics*, vol. 3, no. 2, pp. 74–86, 2008, doi: 10.2174/157489308784340676.
- [6] C. N. Magnan and P. Baldi, "SSpro/ACCpro 5: Almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity," *Bioinformatics*, vol. 30, no. 18, pp. 2592–2597, 2014, doi: 10.1093/bioinformatics/btu352.
- [7] S. Wang, J. Peng, J. Ma, and J. Xu, "Protein secondary structure prediction using deep convolutional neural fields," *Scientific Reports*, vol. 6, no. 1, pp. 1–11, 2016, doi: 10.1038/srep18962.
- [8] I. H. Holley and M. Karplus, "Protein secondary structure prediction with a neural network," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 86, no. 1, pp. 152–156, 1989, doi: 10.1073/pnas.86.1.152.
- [9] M. Spencer, J. Eickholt, and J. Cheng, "A deep learning network approach to ab protein secondary structure prediction," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 12, no. 1, pp. 103–112, 2015, doi: 10.1109/TCBB.2014.2343960.
- [10] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, no. 12, pp. 2577–2637, 1983, doi: 10.1002/bip.360221211.
- [11] A. V Guzzo, "The influence of amino acid sequence on protein structure," *Biophysical Journal*, vol. 5, no. 6, pp. 809–822, 1965, doi: 10.1016/S0006-3495(65)86753-4.
- [12] M. Ankerst, M. Ankerst, G. Kastenmüller, H.-P. Kriegel, and T. Seidl, "Nearest neighbor classification in 3D protein databases," *Proceedings International Conference on Intelligent Systems for Molecular Biology*, pp. 34–43, 1999.
- [13] J. M. Levin, B. Robson, and J. Garnier, "An algorithm for secondary structure determination in proteins based on sequence similarity," *FEBS Letters*, vol. 205, no. 2, pp. 303–308, 1986, doi: 10.1016/0014-5793(86)80917-6.
- [14] D. G. Kneller, F. E. Cohen, and R. Langridge, "Improvements in protein secondary structure prediction by an enhanced neural network," *Journal of Molecular Biology*, vol. 214, no. 1, pp. 171–182, 1990, doi: 10.1016/0022-2836(90)90154-E.
- [15] S. A. Malekpour, S. Naghizadeh, H. Pezeshk, M. Sadeghi, and C. Eslahchi, "Protein secondary structure prediction using three neural networks and a segmental semi Markov model," *Mathematical Biosciences*, vol. 217, no. 2, pp. 145–150, 2009, doi: 10.1016/j.mbs.2008.11.001.
- [16] J. Peng, L. Bo, and J. Xu, "Conditional neural fields," *Advances in Neural Information Processing Systems (NIPS)*, pp. 1419–1427, 2009.
- [17] A. A. Salamov and V. V Solovyev, "Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments," *Journal of Molecular Biology*, vol. 247, no. 1, pp. 11–15, 1995, doi: 10.1006/jmbi.1994.0116.
- [18] A. Dom, "Protein Secondary Structure | Kaggle," 06 July, 2018, [Online], <https://www.kaggle.com/datasets/alfrandom/protein-secondary-structure>. [Accessed May. 13, 2020].
- [19] G. Wang and R. L. Dunbrack, "PISCES: recent improvements to a PDB sequence culling server," *Nucleic Acids Research*, vol. 33, no. Web Server, pp. W94–W98, Jul. 2005, doi: 10.1093/nar/gki402.
- [20] K. Asai, S. Hayamizu, and K. Handa, "Prediction of protein secondary structure by the hidden Markov model," *Bioinformatics*, vol. 9, no. 2, pp. 141–146, 1993, doi: 10.1093/bioinformatics/9.2.141.
- [21] J. Wang, H. Cao, J. Z. H. Zhang, and Y. Qi, "Computational protein design with deep learning neural networks," *Scientific Reports*, vol. 8, no. 1, Dec. 2018, doi: 10.1038/s41598-018-24760-x.
- [22] J. Garnier, J. F. Gibrat, and B. Robson, "GOR method for predicting protein secondary structure from amino acid sequence,"




- Methods in Enzymology*, vol. 266, pp. 540–553, 1996, doi: 10.1016/s0076-6879(96)66034-0.
- [23] Y. Yang *et al.*, “Sixty-five years of the long march in protein secondary structure prediction: The final stretch?,” *Briefings in Bioinformatics*, vol. 19, no. 3, pp. 482–494, 2018, doi: 10.1093/bib/bbw129.
- [24] P. Y. Chou and G. D. Fasman, “Prediction of protein conformation,” *Biochemistry*, vol. 13, no. 2, pp. 222–245, 1974, doi: 10.1021/bi00699a002.
- [25] D. T. Jones, “Protein secondary structure prediction based on position-specific scoring matrices,” *Journal of Molecular Biology*, vol. 292, no. 2, pp. 195–202, 1999, doi: 10.1006/jmbi.1999.3091.
- [26] P. Agarwal and S. A. Rizvi, “A technique based on neural network for predicting the secondary structure of proteins,” in *International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007)*, 2007, pp. 382–386, doi: 10.1109/ICCIMA.2007.26.

BIOGRAPHIES OF AUTHORS






Romana Rahman Ema    is currently working as a Lecturer at Jashore University of Science and Technology (JUST), Bangladesh. Before joining JUST, she was working as a Senior Lecturer in the department of Computer Science and Engineering at North Western University, Khulna, Bangladesh. Romana Rahman Ema received B.Sc. (Engg.) in Computer Science and Engineering from Jashore University of Science and Technology and M.Sc. (Engg) in Computer Science and Engineering from the same university. She can be contacted at email: rr.ema@just.edu.bd.






Sharmin Sultana    is a student in the department of Computer Science and Engineering, North Western University, Khulna. Her area of interest is data analytics, deep learning, machine learning, artificial intelligence and networking. She can be contacted at email: sailasharmin37@gmail.com.



Shakil Ahmed Shaj    graduated from North Western University in Khulna with a bachelor's degree in computer science engineering in 2021. He has over three technical publications published in scientific journals and international conferences to his credit. He is now working as a Software Engineer. His research interests include machine learning, deep learning, data science, and artificial intelligence. He can be contacted at email: shakilahmedshaj@gmail.com.



Syed Md. Galib    is a Professor in the department of Computer Science and Engineering (CSE) at Jashore University of Science and Technology (JUST), Bangladesh. Before joining at JUST, he was working as an Associate Professor in the department of Computer Science and Information Technology at Patuakhali Science and Technology University, Bangladesh. Dr. Galib finished his PhD from Swinburne University of Technology, Australia in 2015. Before that, he completed his Master of Science in Computer Engineering (specialization in Artificial Intelligence) from Dalarna University, Sweden in 2008. He obtained his Bachelor of Science in Computer Science and Engineering from Khulna University, Bangladesh in 2005. He can be contacted at email: galib.cse@just.edu.bd.