Journal of Computer Science

Original Research Paper

Protein Secondary Structure Prediction using Hybrid Recurrent Neural Networks

Romana Rahman Ema, Akhi Khatun, Md. Alam Hossain, Mostafijur Rahman Akhond, Nazmul Hossain and Md. Yasir Arafat

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

Article history
Received: 06-01-2022
Revised: 06-05-2022
Accepted: 18-06-2022

Corresponding Author:
Romana Rahman Ema
Department of Computer Science and Engineering, Jashore University of Science and Technology, Bangladesh
Email: rr.ema@just.edu.bd

Abstract: The most important and challenging problem in bioinformatics is protein secondary structure prediction. The molecules of all protein organisms have three-dimensional (primary, secondary, 3-D) structures which are completely recognized by the sequence of amino acids. Protein secondary structure attributes to the polypeptide backbone of the local configuration of proteins. Most generally, the second-level prediction is indicated such as: If there is an amino acid sequence of the protein, then predict that all amino acid has in the α-Helices (H), β-sheet (E), and other Random Coils (C). In this study, Hybrid Recurrent Neural Networks (HRNN) have been proposed for the prediction of protein secondary structure to improve the prediction performance. The purpose of the work is to predict the protein secondary structure and bring out a highly accurate solution that would be easily solved in computational biology. The proposed method can experimentally perform exceedingly better than other previous work and this study could be easily understandable by researchers for solving the protein structure prediction problems. The five techniques are used for this implementation. These are Gated Recurrent Unit (GRU), Long Short-Term Memory (LSTM), Bidirectional, Bidirectional Gated Recurrent Unit (BGRU), and Bidirectional Long Short-Term Memory (BLSTM) neural networks. Especially, the proposed two-dimensional recurrent neural network (2D-RNN) framework consisted of five models: 2D-GRU_RNN, 2D-LSTM_RNN, 2D-Bi_RNN, 2D-BiGRU_RNN, and 2D-BiLSTM_RNN. In this study, firstly the 2D recurrent neural network has been generated and combined the extracted features of protein sequence with Position-Specific Scoring Matrix (PSSM). After that, the model has been trained and tested with those datasets. Finally, the model has been evaluated for prediction. Besides, all prediction accuracy has been compared and improved with existing methods. These achievements are obtained 91% (BiGRU and BiLSTM), 92% (BiGRU), 89% (BiGRU and BiLSTM), 93% (BiGRU and BiLSTM), 88% (BiGRU), 86% (BiGRU), 91% (GRU), 87% (BiLSTM), 88% (BiGRU) and 93% (GRU and BiGRU) for predicting accuracy of Q3.

Keywords: Protein Secondary Structure Prediction (PSSP), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), Position-Specific Scoring Matrix (PSSM), Database of Secondary Structure Assessments (DSSP)

Introduction

Protein is a macromolecular polypeptide (Zhao *et al*., 2020). Secondary structures are the building blocks of the macromolecule structure. It formally refers to the pattern of hydrogen bonds in both the amino group and therefore the carboxyl oxygen atoms within the peptide backbone. There are two types of secondary structure. One is a regular secondary structure and it has three types which are α-Helices (H) and β-sheet (E), Coil (C) (Akkaladevi *et al*., 2005; Hobbes, 2019) and other is irregular secondary structure and it has many types such as Tight turns, Bulges, etc. According to DSSP, there are 8-class of secondary structures i.e., G (3_{10} Helix) H, I (π-Helix) B

(Isolated Bridge), E (Beta-strand), C, S(Bend), T (Tight turns) are converted into 3-class of secondary structure (Hu *et al*., 2019; Zhang *et al*., 2018; Yang *et al*., 2018; Hanson *et al*., 2019). {B, S, T, G, I, C} are converted into C, {H}> H, {E}>E. Methods of predicting protein secondary structures based on deep learning techniques are the most crucial problems in molecular biology. These different techniques are applied for the prediction of 3-state or 8-state secondary structures which are correctly predicted by Q3 or Q8 accuracy with PSSM. But Q3 accuracy (C, H, E) is the best for the PSSP (Zhang *et al*., 2018; Wang *et al*., 2017; 2016).

Over the early years, the researchers predicted protein secondary structure prediction using various techniques (Zhang *et al*., 2018; Wardah *et al*., 2018). Furthermore, the accuracy of the prediction has improved compared to the existing methods. Such strategies are based on all the information needed to determine the three-dimensional structure and the sequence of amino acids which are encoded. A recurrent neural network is a class of artificial neural networks. It is also called a form of feed-forward Neural network. It typically handles sequential data. A recurrent neural network has many types and techniques that sequentially organize data (Babaei *et al*., 2010). We used a two-dimensional recurrent neural network with long short-term memory cells and a Gated Recurrent Unit with a Bidirectional recurrent neural network for the prediction of the protein secondary structure and predicted accuracy using some kinds of datasets. Deep RNN is used in the model's hidden layers, which are referred to as Gated Recurrent Units (GRUs). GRU is a form of RNN that can model data in sequential order. GRU is faster and consumes less memory than LSTM, however, LSTM is more accurate when working with datasets with longer sequences. Gated recurrent unit and long-short term memory network which is eventually combined with a bidirectional network by a bi-LSTM and Bi-GRU layer. Ten types of datasets are used having all the features in each model and after selecting the models their performances have been evaluated and these performances have been added to the final model.

BiGRU uses deep RNN will functionally improve the performance of algorithms. This technique also increases the accuracy of PSSP than other single RNN techniques. It also shows better performance for the prediction on smaller and larger datasets. BiLSTM will also improve the performance of prediction accuracy and sequentially organize data (Hu *et al*., 2019). It also helps to extend the amount of information that is available on the network (Hu *et al*., 2019).

Related Work

Zhao *et al.* (2020) have proposed a new strategy based on generative confrontation and convolutional neural networks to predict protein secondary structures. They created a confrontation network for the extraction of the protein features and then combined the extracted features with the original position-specific scoring matrix data as input from the convolutional neural network to get predicted results.

Hu *et al*. (2019) introduced a Bi-LSTM-based ensemble algorithm to predict the secondary structure of proteins. They introduced the ensemble model. This technique included five Sub-Models (PSSM, HMM (Hidden Markov Model), PSSM-Count, Wordem, and PPS model). Bi_LSTM layer has been created for these models. Each model contained 2 Bi_LSTM layers and composed an ensemble model. Finally, the Bi_LSTM layer is joined with the sub-model. The ensemble model and sub-model trained concurrently and observed the performance of each model. This model achieved the highest 84% accuracy.

Zhang *et al.* (2018) presented a novel deep learning architecture based on a convolutional neural network, residual network, and bidirectional recurrent neural network to improve the prediction performance of protein secondary structure. This model applied RNN to verify the structural class of protein for low and high-dimensional data sets. The Stock well transformation is applied to improve the prediction performance of protein structural class.

Guo *et al*. (2018) introduced a hybrid deep learning framework, 2-dimensional Convolutional Bidirectional Recurrent Neural Networks (2C-BRNNs) to improve the predictability of 8-grade secondary structures. This model also extracted differential local interactions between amino acid residues by 2dimensional convolutional neural networks This hybrid framework comprised four models which 2DConvBGRUs, 2DCNN-BGRUs, 2DConv-BLSTM, and 2DCNN-BLSTM. This model performed better for the prediction of protein secondary structure than the benchmark models. This is also helpful for feature extraction.

Li and Yu (2016) proposed an EEDN (end-to-end deep network) that predicted protein secondary structures from integrated local and global contextual features. They presented a CCNN (Cascaded Convolutional Neural Networks) and RNN to predict the secondary structure. This model comprised four parts, one feature embedding layer, 2nd; multi-scale Convolutional Neural Network (CNN) layers, 3rd; three stacked Bidirectional Gated Recurrent Unit (Bi-GRU) layers, and 4th; two fully connected hidden layers. The embedded sequence features and the original profile features are fed into multiscale CNN layers with different kernel sizes to extract multiscale local relevant features and improve the prediction performance. This model was effective to achieve the performance of art.

Cheng *et al*. (2020) proposed a prediction method of protein secondary structure based on the CNN and LSTM model. CNN has two convolutional layers, one carpooling layer and the other ReLU activation layer in its architecture. They used the Soft Max classifier. It is fed with the feature maps extracted from the second

convolutional layer and the first probability output is obtained. There is a sequence layer and a last layer in the LSTM model. To get the second probability output, the feature is retrieved from the last layer and fed into a random forest classifier. To obtain the prediction model EN-CSLR in this study, the two probabilistic outputs are weighted and combined.

Recent studies show that the prediction of protein secondary structure is a vital issue. Accuracy results and time complexity issues of the prediction process of the existing methods using deep learning techniques are not satisfactory. So, the proposed Hybrid Recurrent Neural Networks (HRNN) are helpful for the prediction of protein secondary structure.

Protein

Proteins are called large and complex organic molecules that take part in a vital role in the body. Proteins are building blocks of amino acid sequence (Yang *et al*., 2018; Babaei *et al*., 2010; Dongardive and Abraham, 2017). They function mostly in cells and are essential for the formation, function, and control of body tissues and organs (Wardah *et al*., 2018).

Figure 1 shows the structure of the protein. Generally, protein structure has four types: Primary, secondary, tertiary, and quaternary structure (Protein Structure, 2019; OPS, 2022).

Primary Structure (PS)

The primitive stage of protein structure is called primary structure. It is a linear sequence of amino acids in a polypeptide chain (Wardah *et al*., 2018). The hormone insulin contains two polypeptide chains. Every chain has its own set of amino acids, grouped in a specific order. Each amino acid sequence can be linked to the next amino acid sequence by a peptide bond formed during the process of biosynthesis. Protein starts to form the amino-terminal (N) end and ends in the Carboxyl-terminal (C) end (Protein Structure, 2019; OPS, 2022).

Primary Structure = Sequence of Amino Acid.

Figure 2 shows the 3 letter code of the amino acid sequence. The order of a protein is found by the DNA of the gene which is encoded by the part of a protein or multi-subunit protein.

Secondary Structure (SS)

The second level of protein structure is called the secondary structure. Mostly common and available secondary structure is alpha-helices and beta-strand (Li and Yu, 2016). It is folded or pleated. It is formed into a polypeptide chain by hydrogen bonds between the carbonyl O group and amino hydrogen H. Besides, this contains random coils, bulges, turns, Beta-bends, etc., (Protein Structure, 2019; OPS, 2022).

Regular Secondary structure = Alpha -Helices (H), Beta- Strand (E), and Coil (C) (Guo *et al*., 2019).

Figure 3(a) shows the α helix structure and intermolecular hydrogen bonding. It has 3.6 amino acids per turn. Its inner-facing side chains are hydrophobic.

This Fig. 3(b) indicates the β-sheet of the secondary structure. Each 5 to 10 amino acid in each region forms beta-sheets.

Tertiary Structure (TS)

It refers to the 3-D structure (Babaei *et al*., 2010; Guo *et al*., 2018). It contains many forms of the polypeptide chain. It has R-group interactions. Besides, this structure has many properties following hydrophobic interactions, hydrogen bonding, ionic bonding and disulfide bridge, and dipole-dipole interactions. Hydrogen bonds can be formed by polar R-groups. Hydrophobic interactions and dipole-dipole interactions are very important for the three-dimensional structure.

Tertiary structure = fold helices and strands into domains.

Figure 4 shows the tertiary structure. This structure fold helices and strands into domains for the prediction of protein (Protein Structure, 2019; OPS, 2022).

Quaternary Structure (QS)

It gives a specific overall shape of a protein (Guo *et al*., 2018). It involves interactions and cross-links between different parts of the polypeptide chain. Some units stabilize QS (Wardah *et al*., 2018). Example:

- Hydrophobic and Hydrophilic interactions
- Salt bridges
- Hydrogen bonds
- Disulfide bonds (Protein Structure, 2019; OPS, 2022)

Quaternary Structure (Biological Units) = functional assemblies of chains (subunits).

Figure 5 points out the quaternary structure. It has two or more tertiary subunits. For example, two alpha chains and two beta chains are included in hemoglobin.

In this study, the secondary structure of protein function has been used for the prediction. There are many techniques for the prediction like the Machine Learning algorithm, Chou-Fasman method, Hidden Markov Model, etc., but hybrid recurrent neural network techniques have been used in this study. This technique helps to improve the prediction performance and accuracy of the existing methods. In this study, the proposed hybrid recurrent neural network consists of Gated Recurrent Unit (GRU), Long Short-Term Memory (LSTM), and Bidirectional RNN and their combined techniques are BiGRU and BiLSTM.

Gated Recurrent Unit (GRU)

A Gated recurrent unit is one kind of recurrent neural network. This network unit can be used to improve the memory capacity of a recurrent neural network as well as to facilitate the training of a model. The hidden unit can also be used to solve the invisible gradient problems in recurrent neural networks. This illustrates better performance for the sequence-based processing models and predictions on smaller and larger datasets (Wardah *et al*., 2018; Phi, 2018).

Figure 6 illustrates the architecture of GRU. The different gates of the Gated recurrent network are discussed below:

1. Update Gate: It indicates how much knowledge of the past has to be passed with the future. This is almost the same as the LSTM output gate (Li and Yu, 2016; Phi, 2018)
2. Reset Gate: This gate indicates the past knowledge of how should be reset. This is similar to the consolidation of input gate and forgets gate like LSTM (Li and Yu, 2016; Phi, 2018)
3. Current memory gate: This is included in the reset gate as the input modulation gate. It is a sub-portion of the input gate. It helps to present some non-linearity input and create the input zero-minute. It is helpful to minimize the impact of past data on the current data which processed in the future (Li and Yu, 2016; Phi, 2018)

Fully gated unit, initial value for t = 0, input vector = y_t and output vector ($g_0 = 0$):

$$u_t = \sigma_g\left(P_u y_t + V_u g_{t-1} + C_u\right) \quad (1)$$

[u_t: Update gate vector for the fully gated unit]:

$$q_t = \sigma_g\left(P_q y_t + V_q g_{t-1} + C_q\right) \quad (2)$$

[q_t: reset gate vector for the fully gated unit]:

$$\hat{g}_t = \varphi_g\left(P_g y_t + V_g\left(q_t \odot g_{t-1}\right) + C_g\right) \quad (3)$$

[$\hat{g}_t$: candidate activation vectot]:

$$g_t = \left(1-u_t\right)\odot g_{t-1} + u_t \odot \hat{g}_t \quad (4)$$

[g_t: Output vector for the fully gated unit].

Here, σ_g is the sigmoid function, φ_g is the hyperbolic tangent, and P, V, and C is the parameter metrices and vector (Zhang *et al*., 2018; Wang *et al*., 2017; Li and Yu, 2016; Panda and Majhi, 2021; Phi, 2018)

Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) is one kind of artificial Recurrent Neural Network (RNN) architecture. It is used in deep learning architectures (Cheng *et al*., 2020). Unlike standard feedforward neural networks, LSTM has feedback connections. It processes both single and entire sequences of data points. LSTM cells can note both long and short-range interactions by applying constant error flow (Heffernan *et al*., 2018). This allows the input of the entire protein sequence, regardless of sequence segmentation It can be used to predict the performance problems, classification problems, processing problems, protein homology detection, and also sequential problems. We can use an element-wise multiplication vector ($\odot$) for the first calculation (Sønderby and Winther, 2014; Phi, 2018).

Figure 7 indicates a cell state, an input gate, an output gate and a forget gate which are the general components of a typical LSTM unit.

Hidden state and new inputs-the input at a current timestep and the hidden state from a prior timestep are combined before being passed through various gates.

Forget the gate-this gate determines which information should be forgotten. The sigmoid function ranges between 0 and 1 and determines which values in the cell state should be deleted, recalled, or partially remembered. (Multiplied by some value between 0 and 1).

Input gate-This gate facilitates the identification of critical components that must be introduced to the cell state. The cell state candidate gets to multiply the output of the input gate, with just the information the input gate deems important being included in the cell state.

Update cell state-The output of the forget gate gets to multiply the prior cell state gets. For the instance of input gate*cell state candidate, we get new information for the latest cell states.

Update hidden state-It is the last part. The most recent cell state is multiplied by the outcomes of the output gate using the tanh activation function.

The equations of the LSTM cell with forget gate are:

$$f_t = \sigma_g\left(P_f y_t = V_f h_{t-1} + C_f\right) \quad (5)$$

$$j_t = \sigma_g\left(P_j y_t = V_j h_{t-1} + C_j\right) \quad (6)$$

$$o_t = \sigma_g\left(P_o\, y_t = V_o\, h_{t-1} + C_o\right) \quad (7)$$

$$\bar{d}_t = \sigma_g\left(P_o\, y_t = V_o\, h_{t-1} + C_o\right) \quad (8)$$

$$d_t = f_t \odot d_{t-1} + j_t \odot \bar{d}_t \quad (9)$$

$$h_t = o_t \odot \sigma_h\left(d_t\right)[12][19][19][27] \quad (10)$$

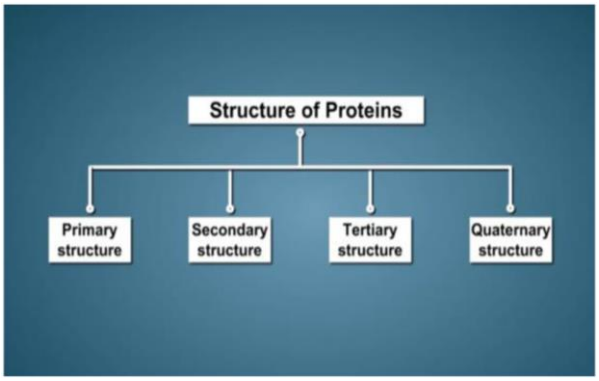


Fig. 1: Structure of protein

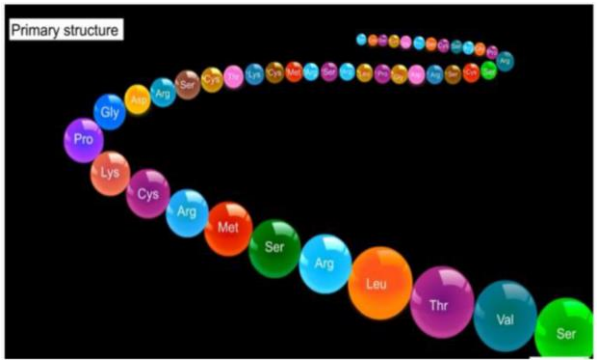
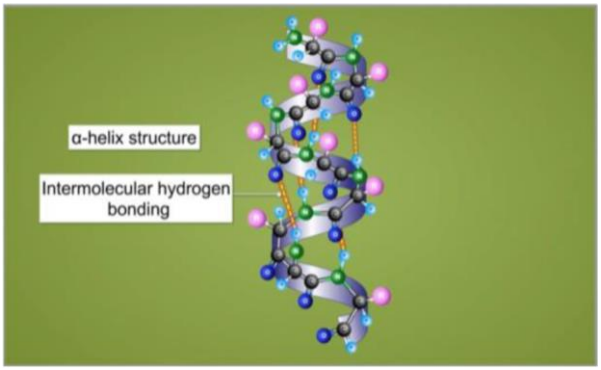
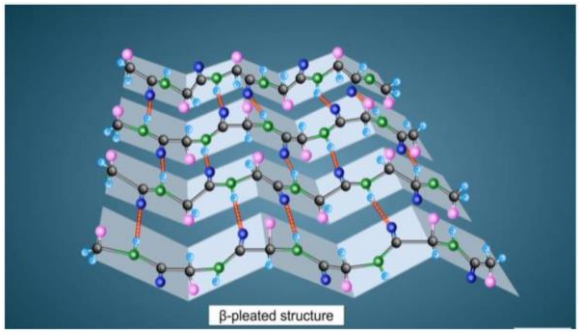


Fig. 2: Primary structure



(a)



(b)

Fig. 3: (a) Alpha-Helix structure and intermolecular hydrogen bonding of protein secondary structure; (b): Beta-sheet of protein secondary structure

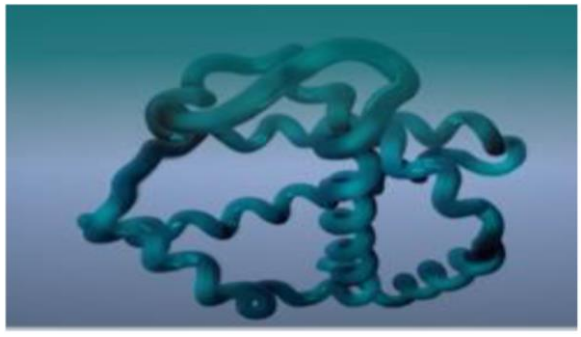


Fig. 4: Tertiary structure of protein

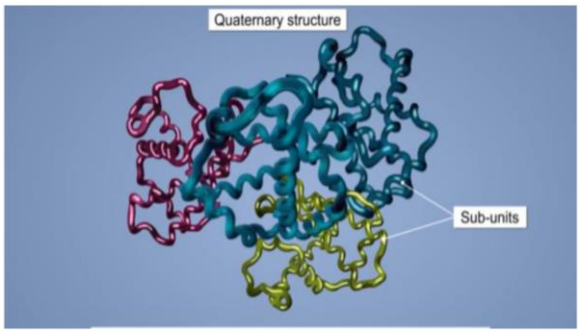


Fig. 5: Quaternary structure of protein

Bidirectional Recurrent Neural Networks (BRNN)

A bidirectional neural network is a special kind of recurrent neural network. This network connects two hidden layers of opposite directions to the same output (Guo *et al*., 2018; BRNN, 2022). The output layer or model will be able to get information from forwarding and backward states (BRNN, 2022). This recurrent neural network can be used to predict the protein secondary structure and performance problems. Besides, we can use this network for speech, hand-writing recognition, and translation. Two bidirectional RNN with GRU and LSTM are stacked to increase the prediction performance (Guo *et al*., 2018).

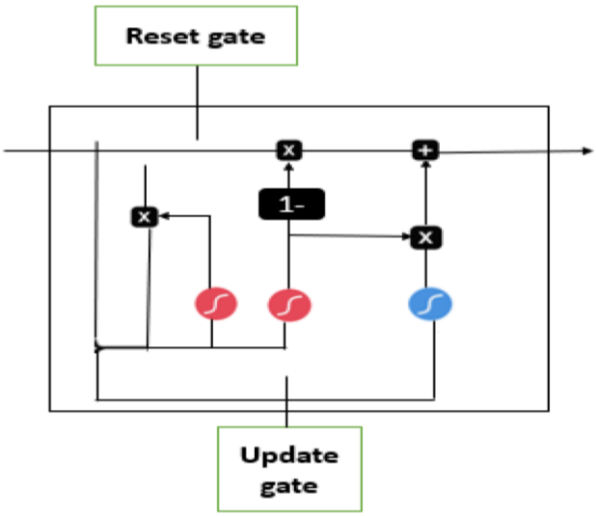


Fig. 6: Architecture of GRU network (Li and Yu, 2016; RS126Data, 2022)

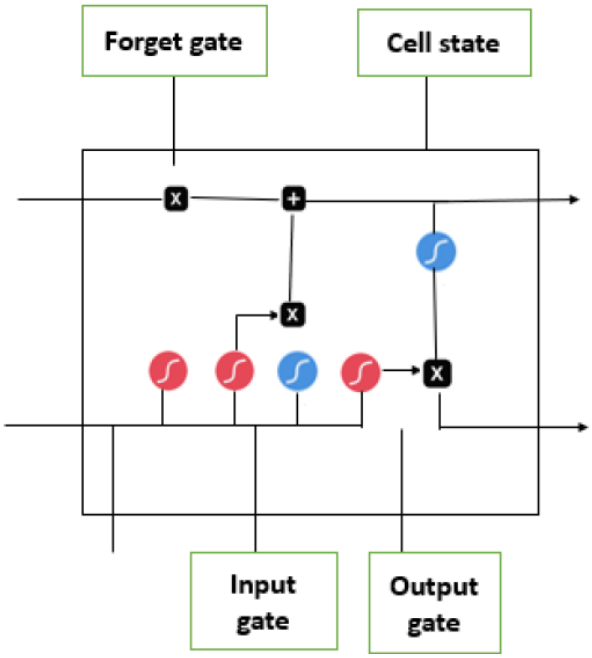


Fig. 7: Architecture of LSTM network (Guo *et al*., 2018; RS126Data, 2022)

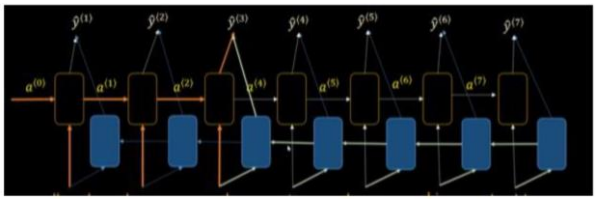


Fig. 8: Architecture of bidirectional RNN

Figure 8 indicates the bidirectional RNN. The BRNN splits the neurons of a regular RNN. It includes two directions. One is a forward state which points to a positive time. The other is backward states which point to the negative time (BRNN, 2022). The output of the bidirectional-recurrent neural network model, yt is determined by:

$$M\left(y_t \mid \{x_i\} i \neq t\right) = \sigma\left(p_y^f g_t^f + p_y^c g_t^c + C_y\right) \quad (11)$$

where:

$$g_t^f = tang\left(P_g^f g_{t-1}^f + P_x^f x_t + C_g^f\right) \quad (12)$$

$$g_t^c = tang\left(P_g^c g_{t-1}^c + P_x^c x_t + C_g^c\right) \quad (13)$$

Bidirectional Gated Recurrent Unit (BiGRU)

A bidirectional gated recurrent unit is a sequence processing model that consists of two GRU: One taking the input in a forward's direction and the other in a backward direction (Guo *et al*., 2018). It is a bidirectional recurrent neural network. It has an input and forgets the gate (Li and Yu, 2016). This bidirectional network increases the performance of PSSP (Li and Yu, 2016; Kumar *et al*., 2020; BGRU, 2016).

Bidirectional Long Short-Term Memory (BiLSTM)

A bidirectional LSTM is a bidirectional recurrent neural network. It is called a sequence processing model based on two LSTMs: One taking the input in a forward direction and the other in a backward direction (Guo *et al*., 2018). Bi-LSTMs also improve the performance of algorithms like LSTM Network. Bi-LSTMs essentially improve the number of information accessible to the network. This technique increases the accuracy of PSSP than other simple RNN techniques (Hanson *et al*., 2019; Kumar *et al*., 2020).

Proposed Methods

In this study, a prediction method has been proposed for the hybrid recurrent neural networks for the protein secondary structure. The recurrent neural network-based algorithms have been implemented. The amino acid sequences have been utilized for a better prediction of secondary structure with the help of PSSM.

Figure 9 shows at first the dataset has been loaded, after that, it will remove the unnecessary raw data. Further, the dataset has been preprocessed. After processing the dataset, a popular sequence comparison tool i.e., PSSM has been used. Again, a 2D layer of a recurrent neural network model has been created. Before creating the 2D layer, 5 types of RNN have been used and combined these types within 2D-RNN. At the end, the

model of 2D-GRU_RNN, 2D-LSTM_RNN, 2D-Bidirectional_RNN, 2D-BiGRU_RNN and 2D- BiLSTM RNN have been built. The datasets have been trained and tested with these models. Finally, the accuracy of these models has been calculated.

Secondary Structure Datasets

I RS126 dataset
II CB513
III PDB
IV CASP12
V CASP10
VI 2018-06-06-pdb-intersect-pisces.csv
VII 2018-06-06-pdb-ss.cleaned.csv
VIII Validation_secondary_structure_valid.csv
IX Training_secondary_structure_train.csv
X TS115 dataset

CB513 Dataset

This is an essential and suitable dataset for the improvement of algorithms and the prediction of secondary structure. In this study, we used this type of dataset that assist to increase the prediction performance of protein secondary structure (Wang *et al*., 2017; Zhou *et al*., 2018).

PDB and other Datasets

This dataset is used to predict the 2-D and 3-D protein structures. It is the largest dataset. It is also used to fold the protein structure and organize the classifying data. We used PDB datasets and other types of datasets. For all of the datasets, the accuracy of 's dataset is better than the other types of datasets (SPS, 2022). RS126 dataset is the oldest and largest dataset for the protein secondary structure prediction. Rost and Electric Sander created the scheme. It is one of the most effective datasets to predict the supermolecule structure. It is also applied to bioinformatics research. It can carry 23,347 residues with an average supermolecule sequence length of 185. There are 3 % α-helices (H), 1% β-sheet as well as 47% random coil in RS126 datasets (RS126Data, 2022).

GRU Recurrent Neural Network Algorithm

GRU (D, P, S, C)
Input: D-train data, P- test data, S- sample size, C-number of sequences

Output: Prediction results

1. At the first step, the network needs to initialize the size of S and C
2. For j= 1 to n do
3. To calculate update gate u_t; initially, t = 0, input vector = y_t and output vector, $g_0 = 0$, using the e:

$$u_t = \sigma_g(P_u y_t + V_u g_{t-1} + C_u)$$

4. To calculate reset gate q_t, this model needs a sigmoid function and parameter matrix, using the Eq. 2:

$$q_t = \sigma_g(P_q y_t + V_q g_{t-1} + C_q)$$

5. Finally, calculate the output vector, a new memory use reset gate q_t to store the previous information. Here, the activation function ($\hat{g}_t$) is Needed to hold the information using Eq. 3 and 4:

$$\hat{g_t} = \varphi_g(P_g y_t + V_g(q_t \odot g_{t-1}) + C_g)$$

$$g_t = (1 - u_t) \odot g_{t-1} + u_t \odot \hat{g_t}$$

LSTM Recurrent Neural Network Algorithm

Input: y (y_1, ….) Where $y_t \varepsilon \mathbb{R}^n$
Parameter: P_f, V_f, C_f, P_j, V_j, C_j, P_o, V_o, C_o, P_d, V_d, C_d
Output: h
Seq = number of data instances

1. For t = 1 to seq n,
Initially, values $d_0 = 0$ and $h_0 = 0$,
2. To calculate forget gate f_t, update gate j_t and cell input gate $\bar{d}_t$, using these Eq. 5, 6 and 8:

$$f_t = \sigma_g\left(P_f y_t = V_f h_{t-1} + C_f\right)$$
$$j_t = \sigma_g\left(P_j y_t = V_j h_{t-1} + C_j\right)$$
$$\bar{d}_t = \sigma_d\left(P_d y_t = V_d h_{t-1} + C_d\right)$$

3. To update cell state d_t, using element- wise multiplication, using Eq. 9:

$$d_t = f_t \odot d_{t-1} + j_t \odot \bar{d}_t$$

4. After calculating internal cell state, then calculate update gate by calculating element-wise multiplication using activation function using Eq. 7:

$$o_t = \sigma_g\left(P_o y_t + V_o h_{t-1} + C_o\right)$$

End for

5. In the last step, calculate the output vector, also known as the hidden state vector, using Eq. 10:

$$h_t = o_t \odot \sigma_h(d_t)$$

Output: h (h_1 ,……, n) here $h_t \varepsilon \mathbb{R}^n$

Bidirectional Recurrent Neural Network Algorithm

Input layers = I_l
Output layers = O_l
Hidden layers = H_l Seq = number of data set instances.
Forward Hidden layer's activation function = $g_t f$

Backward Hidden la er's activation function = g_t^c

Forward Pass

1. For i = 1 to H_l
2. For j = 1 to Seq
3. To Calculate the forward pass for g_t^f, using the Eq. 12:

$$g_t^f = tang\left(P_g^f g_{t-1}^f + P_x^f x_t + C_g^f\right)$$

4. End for
5. For j = Seq to 1
6. To calculate the backward pass for g_t^c, using Eq. 13:

$$g_t^c = tang\left(P_g^c g_{t-1}^c + P_x^c x_t + C_g^c\right)$$

7. End for
8. End for
9. For i = 1 to O_l
10. To Calculate the forward pass for the output layer using the previously-stored activation function, using Eq. 11:

$$M\left(y_t \mid \{x_i\} i \neq t\right) = \sigma\left(p_y^f g_t^f + p_y^c g_t^c + C_y\right)$$

11. End for

Backward pass

12. For I = O_l to 1
13. To calculate the backward pass for the output layer using the previously-stored activation function, using Eq. 11:

$$M\left(y_t \mid \{x_i\} i \neq t\right) = \sigma\left(p_y^f g_t^f + p_y^c g_t^c + C_y\right)$$

14. End for
15. For i= 1 to H_l
16. For j=1 to Seq
17. To calculate the backward pass for $g^f{}_t$, using Eq. 13:

$$g_t^c = tang\left(P_g^c g_{t-1}^c + P_x^c x_t + C_g^c\right)$$

18. End for
19. For j = Seq to 1
20. Calculating the forward pass for g_t^c, using Eq. 12:

$$g_t^f = tang\left(P_g^f g_{t-1}^f + P_x^f x_t + C_g^f\right)$$

21. End for
22. End for (Heffernan *et al*., 2018; BRNN, 2022)

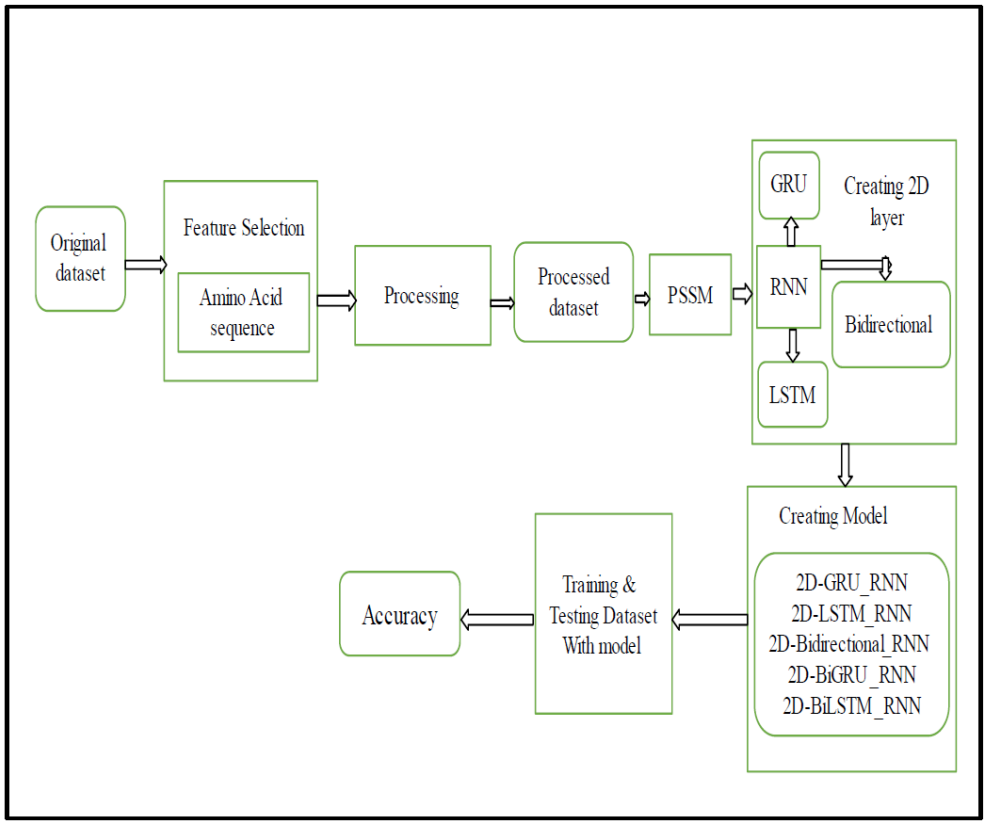


Fig. 9: Proposed flow-diagram of protein secondary structure prediction

Bidirectional GRU Algorithm

Input layers = I_l
Output layers = O_l
Hidden layers = H_l Seq = number of data set instances.
Forward Hidden layer's activation function = ${}^{h}{}_{t}f$
Backward Hidden layer's activation function = $h_t{}^{b}$
data ← read CSV formatted data
data ← import CSV data
data ← read and processing CSV formatted data.
Bi-GRU creates a layer with GRU and creates a BiGRU_RNN model.
For j = 1 to Seq
Forward pass
Calculate the forward pass for $h_t{}^{f}$ and $h_t{}^{b}$
End for
For j = Seq to 1
Backward pass
Calculate the backward pass for $h_t{}^{f}$ and $h_t{}^{b}$
End for
End for (Wardah *et al*., 2018; Kumar *et al*., 2020; BGRU, 2016)

Bidirectional LSTM Algorithm

Input layers = I_l
Output layers = O_l
Hidden layers = H_l Seq= number of data set instances.
Forward Hidden layer's activation function = ${}^{h}{}_{t}f$
Backward Hidden layer's activation function = $h_t{}^{b}$
data read CSV formatted data
data import CSV data.
data read and processing CSV formatted data.
Bi-LSTM create layer with LSTM and create a BiLSTM_RNN model
For j=1 to Seq
Forward pass
Calculate the forward pass for $h_t{}^{f}$ and $h_t{}^{b}$
End for
For j = Seq to 1 Backward pass
Calculate the backward pass for $h_t{}^{f}$ and $h_t{}^{b}$
End for
End for (Kumar *et al*., 2020; BGRU, 2016)

Performance Analysis and Results

If the better quality PSSP dataset is available then the accuracy will be better. In this study, ten types of datasets have been used. These models are trained very fast. For the training and testing datasets, batch size =128, verbose = 1, validation_split = 0.5, epochs = 20 have been used. The completion of the prediction process has taken time at 125 ms/step. As a result, these models achieved the highest accuracy of existing methods.

Performance Indices and PSSM

Accuracy: Accuracy is used to measure the performance prediction. It is needed to calculate the accurate result. Q3 Accuracy and PSSM (taking the input datasets) have been used to calculate the accurate prediction of secondary structure. It is calculated by:

$$Q_3 = \frac{N_C + N_H + N_E}{N} \times 100\% \text{ [1]} \quad (14)$$

Where:
N_C = The number of accurately predicted protein structural classes of C
N_H = The number of accurately predicted protein structural classes of H
N_E = The number of accurately predicted protein structural classes of E
N = The total number of amino acids:

$$Q_j = \frac{N_j}{N} \text{ where } j \in \{C, H, E\} \text{ [1]} \quad (15)$$

Q_j = Represents the total number of amino acid residues. Which are denoted in the state *j*

Figure 10 points out the test sequence were predicted and the actual sequence has been evaluated from the original sequence.

Table 1 and 2 shows the Q3 accuracy with Q_C, Q_H, and Q_E of the tested datasets based on BiGRU, GRU, and BiLSTM. It can be seen that we are shown some (Q_C, Q_H, Q_E) based Q3 accuracy on the proposed methods.

Figure 11 shows the performance among 10 types of datasets that are sequentially organized for computing the accuracy. These 5 techniques have been evaluated within those datasets. In this figure, we have been enabled to show a better performance. It is called a hybrid recurrent neural network model by using GRU, LSTM, Bidirectional, BiGRU, and BiLSTM techniques of recurrent neural network. Here, it is seen that the performance result of BiGRU and BiLSTM are higher than the single GRU, LSTM and Bidirectional RNN. From this figure, we got almost 93% accuracy from the PDB and validation secondary datasets because of larger datasets. This hybrid technique has increased the performance of protein secondary structure prediction and will be enabled to handle the sequential and protein fold data.

Fig. 12 shows that the comparison between the proposed model and the other CNN model is based on the test datasets. Here, the accuracy based on the proposed model is better than the existing methods. The proposed hybrid recurrent neural network model has improved closer to accurate prediction performance The test datasets CASP10, CASP12, CB513, and

PDB25 are obtained at 91, 92,89 and 93% which are 4, 5, 1, and 4% higher than the CNN model (Zhao *et al*., 2020). It can be seen that their dataset's 3-state accuracy results are not satisfactory. So, the proposed model is helpful to increase closer to accurate prediction performance and sequence alignment.

```
Test sequence 1 of 2:

Original sequence:
MHHHHHHMSESSDISAMQPVNPKPFLKGLVNHRVGVKLKFNSTEYRGTLVSTDNYFNLQLNEAEEFVAGVSHGTLGEIFIRCNNVLYIRELPN

Predicted structure:
CCCCCCCCCCCCCCCCCCCCCCCCCHCCCCCEEEEEECCCCCEEEEEEEEECCCCCEEEEEEEEECCCEEEEECCEEECHHHEEEEEECCC

Actual structure:
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCEEEEEECCCCEEEEEEEEECCCCCEEEEEEEEECCEEEEECCEEECHHHEEEEEECCC

========================================================================

Test sequence 2 of 2:

Original sequence:
MIQNHIKNMTPEICKASRALVNLTQKELALMAGIATPTIADFERGARKPHGNNLRSIIIAFENKGLDFVEEGGEIIGIFIRRKNVRAEESIDLSGHGSH

Predicted structure:
CCCCCHCCCCHHHHHHHHHHCCCCHHHHHHHCCCCCCHHCHCCCCCCCCCHHHHHHHHHHHCCCEEEEECCCCCCEEECHHCCCCCCCCCCCCCCC

Actual structure:
CCCCCCCCCCHHHHHHHHHHCCCCHHHHHHHCCCHHHHHHHHCCCCCCCHHHHHHHHHHHHCCEEEEEECCEEEEEECCCCCCCCCCCCCCCCCC

========================================================================
```

Fig. 10: The test sequence (Original, Predicted and Actual sequence)

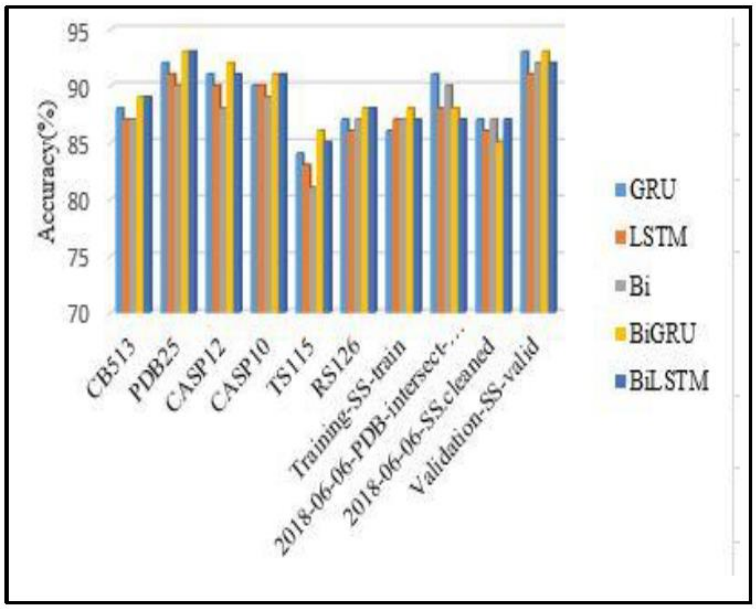


Fig. 11: Performance measurement and comparison of ten types of datasets using LSTM, GRU, Bidirectional, BiGRU, and BiLSTM techniques

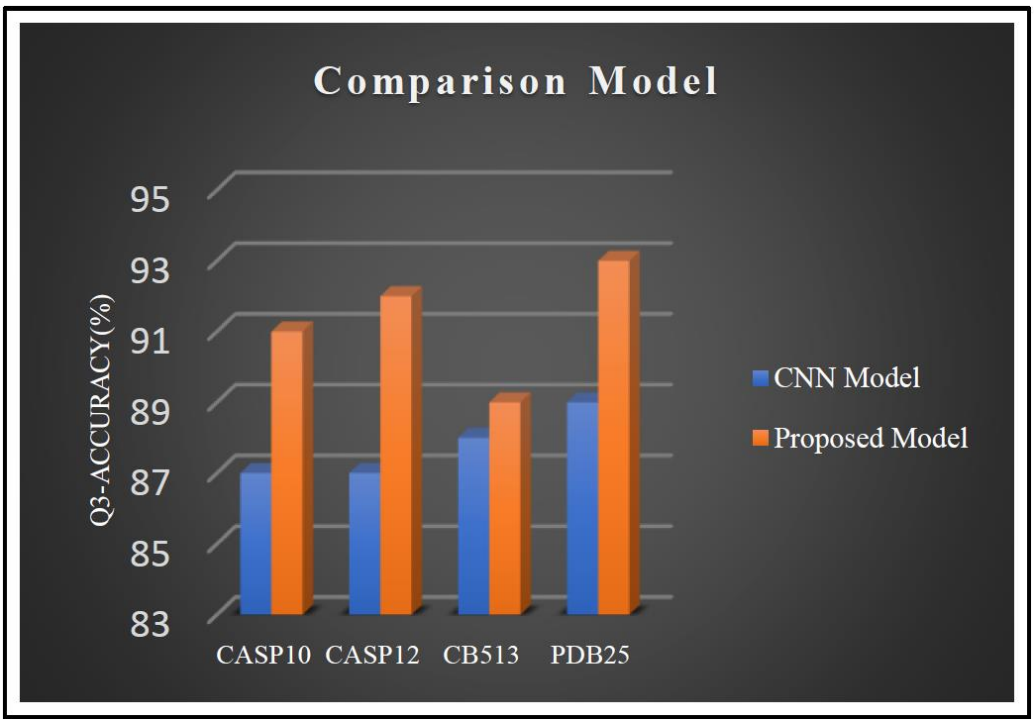


Fig. 12: The Q3-accuracy comparison between the proposed models (which are implemented on datasets CASP10, CASP12, CB513, and PDB25) and the CNN model (Zhao *et al*., 2020)

Table 1: Q_3 Accuracy(**Q_C**, **Q_H**, **Q_E**) of the tested methods based on datasets

Dataset	Q_3 Accuracy (%)	Q_C (%)	Q_H (%)	Q_E (%)	Method
CASP 10	91	91	87	84	BiLSTM
CASP 12	92	90	89	88	BiGRU
CB5 13	89	85	81	80	BiGRU
PDB 25	93	88	90	83	BiLSTM
TS 115	86	83	81	82	BiGRU
RS 126	88	84	82	83	BiGRU
Training-ss-train	88	82	85	81	BiGRU
2018-06-06-pdb-intersect-pisces	91	83	87	84	GRU
2018-06-06-ss. cleaned	87	81	83	84	BiLSTM
Validation-SS-valid	93	85	83	87	BiGRU

Table 2: Performance comparison and Q3 accuracy prediction based on proposed methods

	Method (Q3 Accuracy %)				
Dataset	GRU	LSTM	Bi	BiGRU	BiLSTM
CB 513	88	87	87	89	89
PDB 25	92	91	90	93	93
CASP 12	91	90	88	92	91
CASP 10	90	90	89	91	91
TS 115	84	83	81	86	85
RS 126	87	86	87	88	88
Training-SS-train	86	87	87	88	87
2018-06-06-pdb-intersect-pisces	91	88	90	88	87
2018-06-06-SS. Cleaned	87	86	87	85	87
Validation-SS-valid	93	91	92	93	92

Conclusion

The main conclusions of the experimental work should be presented. In this study, a hybrid recurrent neural network has been applied to improve the overall prediction performance of protein secondary structure. This technique includes five types of RNN based on GRU, Bidirectional LSTM, BiGRU, and BiLSTM. These techniques are applied to extract the features of protein structural class and sequence alignment. In this study, 2D RNN has been used for a better prediction performance. The BiGRU and BiLSTM also help to improve the 3-state accuracy of predictions compared to the other strategies in the recurrent neural network model. This hybrid recurrent neural network model provides a significant first step towards predicting the third-dimensional structure as well as providing information about protein activity, relationships, and functions. Protein folds based on amino acid sequences can revolutionize the design of accurately predicted drugs and explain the causes of new and old diseases. Having a protein structure provides a broader level of understanding of how a protein works, allowing us to make assumptions about how it can affect, control, or alter it. For example, knowing the structure of a protein allows you to design site-directed mutations to change functions. We are enabled to improve the prediction performance and achieved the highest 93% accuracy for the prediction performance than the existing works. As a neighborhood of future scope of labor, we'll propose the three-dimensional protein structure prediction using deep learning techniques. Also, we can explore more advanced techniques like the unsupervised or semi-supervised learning techniques, and other machine learning techniques, combined with the convolutional and recurrent neural network models, multilayer perceptron, inductive learning, and lots more.

Acknowledgment

We would like to express our gratitude and thanks to JUST, CSE laboratory, and the Faculty of Engineering and Technology for the funding and for those who are involved directly or indirectly in this research.

Author's Contributions

Romana Rahman Ema: The research idea was conceived and led by the author Romana. Overall supervision, revision, guidance, coding, analysis, comparison, and drafting of the manuscript were also done by her.

Akhi Khatun: Coding, analysis, comparison, and drafting of the manuscript were done.

Md. Alam Hossain: Overall supervision, revision, and guidance of the paper were done.

Mostafijur Rahman Akhond: Coding, analysis, comparison, and drafting of the manuscript.

Nazmul Hossain: Analysis, comparison, and drafting of the manuscript.

Md. Yasir Arafat: Analysis, comparison, and drafting of the manuscript.

Ethics

Informed Consent

We didn't use anyone's data or information in any way. We used publicly available data covered by the Database Contents License (DbCL) V1.0. Following that, the authors gave their unrestricted permission to use their work and information. The final manuscript was read and approved by all authors.

References

Akkaladevi, S., Katangur, A. K., Belkasim, S., & Pan, Y. (2005, October). Protein Secondary Structure Prediction using decision fusion of Genetic Algorithm and Simulated Annealing Algorithm. In 2005 International Conference on Neural Networks and Brain (1, pp. 467-472). IEEE.

Babaei, S., Geranmayeh, A., & Seyyedsalehi, S. A. (2010). Protein secondary structure prediction using modular reciprocal bidirectional recurrent neural networks. Computer Methods and Programs in Biomedicine, 100(3), 237-247. doi.org/10.1016/j.cmpb.2010.04.005

BGRU. (2016). Bi-Papers with code. https://paperswithcode.com/method/bigru

BRNN. (2022). En. Wikipedia.org. https://en.wikipedia.org/wiki/Bidirectional_recurrent_neural_networks

Cheng, J., Liu, Y., & Ma, Y. (2020). Protein secondary structure prediction is based on the integration of CNN and the LSTM model. Journal of Visual Communication and Image Representation, 71, 102844. doi.org/10.1016/j.jvcir.2020.102844

Dongardive, J., & Abraham, S. (2017). Reaching optimized parameter set: Protein secondary structure prediction using neural network. Neural Computing and Applications, 28(8), 1947-1974. doi.org/10.1007/s00521-015-2150-2

Guo, Y., Wang, B., Li, W., & Yang, B. (2018). Protein secondary structure prediction is improved by recurrent neural networks integrated with two-dimensional convolutional neural networks. Journal of Bioinformatics and Computational Biology, 16(05), 1850021. doi.org/10.1142/S021972001850021X

Guo, Y., Li, W., Wang, B., Liu, H., & Zhou, D. (2019). Deep aclstm: Deep asymmetric convolutional long short-term memory neural models for protein secondary structure prediction. BMC Bioinformatics, 20(1), 1-12. doi.org/10.1186/s12859-019-2940-0

Hanson, J., Paliwal, K., Litfin, T., Yang, Y., & Zhou, Y. (2019). Improving prediction of protein secondary structure, backbone angles, solvent accessibility and contact numbers by using predicted contact maps and an ensemble of recurrent and residual convolutional neural networks. Bioinformatics, 35(14), 2403-2410. doi.org/10.1093/bioinformatics/bty1006

Heffernan, R., Paliwal, K., Lyons, J., Singh, J., Yang, Y., & Zhou, Y. (2018). Single-sequence-based prediction of protein secondary structures and solvent accessibility by deep whole-sequence learning. Journal of Computational Chemistry, 39(26), 2210-2216. doi.org/10.1002/jcc.25534

Hobbes, T. (2019). Chapter 19. Elem. Law, pp, 99-106, doi.org/10.4324/9780429030772-19

Phi, M. (2018). Illustrated Guide to LSTM's and GRU's: A step by step explanation.

Hu, H., Li, Z., Elofsson, A., & Xie, S. (2019). A Bi-LSTM-based ensemble algorithm for prediction of protein secondary structure. Applied Sciences, 9(17), 3538. doi.org/10.3390/app9173538

Kumar, P., Bankapur, S., & Patil, N. (2020). An enhanced protein secondary structure prediction using a deep learning framework on hybrid profile-based features. Applied Soft Computing, 86, 105926. doi.org/10.1016/j.asoc.2019.105926

Li, Z., & Yu, Y. (2016). Protein secondary structure prediction using cascaded convolutional and recurrent neural networks. arXiv preprint arXiv:1604.07176.

OPS. (2022). Orders of protein structure: Primary, secondary, tertiary and quaternary. Alpha helix and beta-pleated sheet.

Panda, B., & Majhi, B. (2021). A novel improved prediction of protein structural class using a deep recurrent neural network. Evolutionary Intelligence, 14(2), 253-260. doi.org/10.1007/s12065-018-0171-3

Protein Structure. (2019). En. ikipedia.org. https://en.wikipedia.org/wiki/Protein_structure

RS126Data. (2022). For protein secondary structure prediction. https://www.kaggle.com/tamzidhasan/rs126data

Sønderby, S. K., & Winther, O. (2014). Protein secondary structure prediction with long short-term memory networks. arXiv preprint arXiv:1412.7828.

SPS. (2022). Sequence and metadata for various protein structures. https://www.kaggle.com/shahir/protein-data-set

Wang, S., Peng, J., Ma, J., & Xu, J. (2016). Protein secondary structure prediction using deep convolutional neural fields. Scientific Reports, 6(1), 1-11. doi.org/10.1038/srep18962

Wang, Y., Mao, H., & Yi, Z. (2017). Protein secondary structure prediction by using the deep learning method. Knowledge-Based Systems, 118, 115-123. doi.org/10.1016/j.knosys.2016.11.015

Wardah, W., Khan, M. G. M., Sharma, A., & Rashid, M. A. (2018). Protein secondary structure prediction using neural networks and deep learning: Review," Computer Biol. Chem., pp, 1-8. doi.org/10.1016/j.compbiolchem.2019.107093

Yang, Y., Gao, J., Wang, J., Heffernan, R., Hanson, J., Paliwal, K., & Zhou, Y. (2018). Sixty-five years of the long march in protein secondary structure prediction: The final stretch? Briefings in Bioinformatics, 19(3), 482-494.

Zhang, B., Li, J., & Lü, Q. (2018). Prediction of 8-state protein secondary structures by a novel deep learning architecture. BMC Bioinformatics, 19(1), 1-13. doi.org/10.1186/s12859-018-2280-5

Zhao, Y., Zhang, H., & Liu, Y. (2020). Protein secondary structure prediction based on generative confrontation and convolutional neural network. IEEE Access, 8, 199171-199178. doi.org/10.1109/ACCESS.2020.3035208

Zhou, J., Wang, H., Zhao, Z., Xu, R., & Lu, Q. (2018). CNNH_PSS: Protein 8-class secondary structure prediction by a convolutional neural network with the highway. BMC Bioinformatics, 19(4), 99-109. doi.org/10.1186/s12859-018-2067-8