Secondary Protein Structure Prediction using CullPDB dataset

Hrishikesh Mahajan[1,☯], Pratik Kamble[1,☯], Smridhi Bhat[1,☯], Yash Shekhadar[1,☯],

1 Affiliation Dept of Computer and Information Sciences, University of Florida , Gainesville, Florida, USA

☯These authors contributed equally to this work.
* mahajanh@ufl.edu
* kamblep@ufl.edu
* bhatsmridhi@ufl.edu
* yash.shekhadar@ufl.edu

Abstract

Sequence of Amino acids joined by peptide bonds are called as Proteins. Many structures of proteins are possible due to different amino acids and rotation of chains in different directions. These structures of protein are responsible for various interactions leading to protein functions. Thus protein structure prediction has an significance in the fields of drug design,medicine and biotechnology. In this paper, we try to predict secondary protein structure prediction using ML techniques. An incremental approach using Recurrent Neural Networks and Bidirectional LSTM is proposed in this paper. CullPDB5926 is the data-set used for this project. This data-set is included in ICML 2014 Deep Supervised and Convolutional generative stochastic network for Protein Secondary Structure Prediction. [1].

Author summary

Protein secondary structure is the shape in which the local segments of the protein sequences are arranged in, this shape is taken as a result of the various chemical and physical laws. Developing an approach for protein secondary structure prediction is necessity of the time. With the advent of fast GPU's and recent developments in the field of Machine Learning and Deep Learning, the secondary structure prediction can be achieved with a decent accuracy. In this paper, Bidirectional Recurrent Neural Network are used to predict secondary structure of the protein using the publicly available cullpdb data set from the Pisces server. This project is made public available on github under MIT Licence at https://github.com/mahajanhrishikesh/ML-Genomics-Project

Introduction

Proteins play crucial role in the body. The structure of protein can be used to determine function of protein. Thus to know the structure of protein is really important. Many methods of protein prediction exists from the corresponding amino acid sequence. X-ray crystallography and Nuclear Magnetic Resonance (NMR) spectrometry are few techniques used for protein structure prediction. These techniques are time consuming

and costly. Now with the improvement of processing power Machine Learning methods can be used to predict protein accurately and quickly.

Protein structure prediction can be done on three levels namely secondary, tertiary and Quaternary. The secondary level of prediction consists of a plethora of methods such as Chou Fasman's Method [11], PSIPRED, LSTM [4] and BiDirectional LSTM [5]. We have implemented the Bi-directional LSTM method. Also in order to determine the performance of the model on a global scale (not just concentrating on local structures) we used Principle Component Correlation Analysis [14]. Additionaly, we have also surveyed two state of the art methods that perform the task of protein secondary structure prediction; MUFOLD-SS [12] and Deep-ACLSTM [13]. Datasets ususally used for protein secondary structure prediction are CB513 [18], Cull-PDB(Protein Data Bank) [2], versions of CASP(Critical Assessment of protein Structure Prediction) [19] like CASP10, CASP11

For the tertiary structure prediction, Comparative modeling [15] is based on the observation that evolutionary related sequences have identical structures. Due to this fact, it is possible to generate the 3D structure of a protein by building upon the known 3D structures. In this technique, a sequence of steps are executed iteratively. Firstly, a suitable protein template is found out. Then the target and the template sequences are aligned. This is followed by identification of structurally similar areas. Further, the structure of the target is modelled. Each of the step might introduce an error and it may snowball into a larger error depending on the stage that the error has occurred.

Another method is Fold Recognition, also known as 'Threading' [16]. This method takes into account the current sequence and compares it with the known types of folds. This is ideal for predicting the structure of proteins which do not have any evolutionary linked structures.The last method is Ab Initio Prediction [17]. In this method, the prediction uses only amino acid sequence. Template proteins are not used in this technique.

Our study focuses on experimenting various deep learning techniques to perform Protein Secondary Structure Prediction from amino acid sequence data. We aim to train the models and test their prediction accuracy on CB513, Cull-PDB datasets.

Materials and methods

Data

This data-set is included in ICML 2014 Deep Supervised and Convolutional generative stochastic network for Protein Secondary Structure Prediction paper [1]. The dataset can be accessed through the following link provided by Jian Zhou who is the author of the mentioned paper: Dataset [2] This project uses data-sets from "Deep supervised and convolutional generative stochastic network for protein secondary structure prediction at ICML 2014" [1]. Two datasets "cullpdb+profile_5926_filtered.npy.gz" and "cb513+profile_split1.npy.gz" are used.

The first one has been divided into training, validation and test set. This data-set is used to evaluate the later data-set CB513. CB513 contains sequences up to 700 amino acids.The data is represented in a NumPy matrix format. It can be shaped in the following way: (N proteins x 700 amino acids x 57 features).

As per the Fig 1, Both training and test data have maximum sequences with length of 100, while the sequence length ranges from 0 to 700. Fig 2 shows that the dataset contains all 20 amino acid residues with Leucine (L) been the most occurring. Fig 3 shows us the frequency of different secondary structures present in the dataset. The different structures present in the dataset are as follows:

1. Loop

Table 1. 57 Features in the dataset.

Data Rows	Features
$[0-22)$	amino acid residues, with the order of 'A', 'C', 'E', 'D', 'G', 'F', 'I', 'H', 'K', 'M', 'L', 'N', 'Q', 'P', 'S', 'R', 'T', 'W', 'V', 'Y', 'X', 'NoSeq'
$[22, 31)$	Secondary structure labels, with the sequence of 'L', 'B', 'E', 'G', 'I', 'H', 'S', 'T', 'NoSeq'
$[31, 33)$	N- and C- terminals
$[33, 35)$	relative and absolute solvent accessibility, used only for training. (absolute accessibility is thresholded at 15; relative accessibility is normalized by the largest accessibility value in a protein and thresholded at 0.15; original solvent accessibility is computed by DSSP)
$[35, 57)$	sequence profile. Note the order of amino acid residues is ACDEFGHIKLMNPQRSTVWXY and it is different from the order for amino acid residues

The last feature of both amino acid residues and secondary structure labels just mark end of the protein sequence. [22,31) and [33,35) are hidden during testing.

Table 2. $cullpdb + profile_5926\_filtered.npy.gz$ **dataset division**

Data Rows	Division
$[0, 5435)$	training
$[5435, 5690)$	test
$[5690, 5926)$	validation

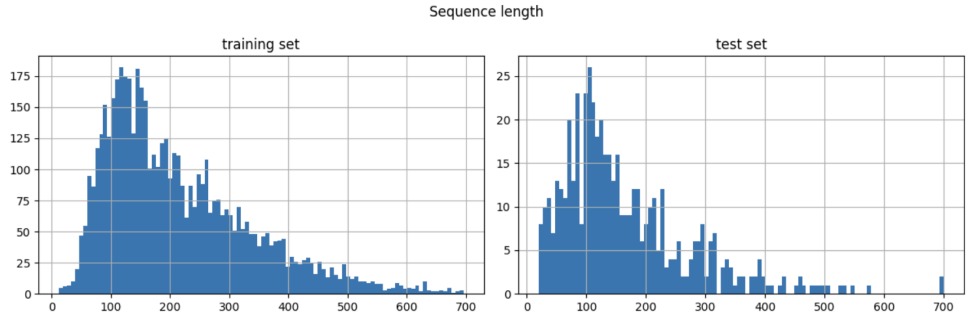

Fig 1. Sequence Length Frequency Graph in training and test set. In the training set it can be observed that most of the proteins are in the range of 100-200 while in the testing set the most of the proteins are in range 50-150. The maximum sequence length observed is 700.

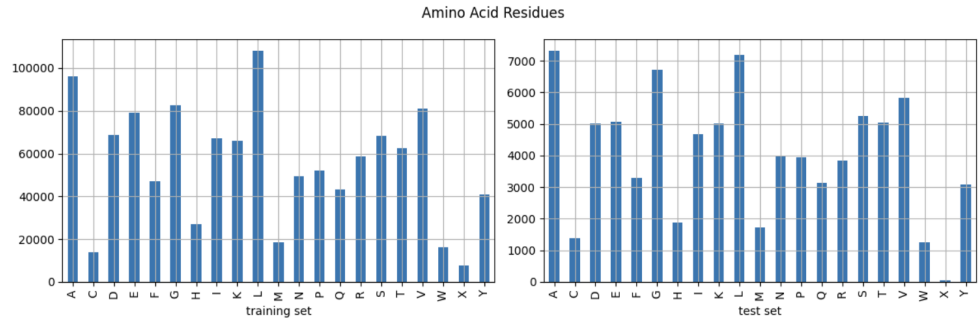

Fig 2. Amino acid Residue Frequency Graph. We observe that both training and testing data-sets contains all of the Amino Acid residues.However, the training data contains much more residues ranging up to 100000 while testing data range up to 7000.

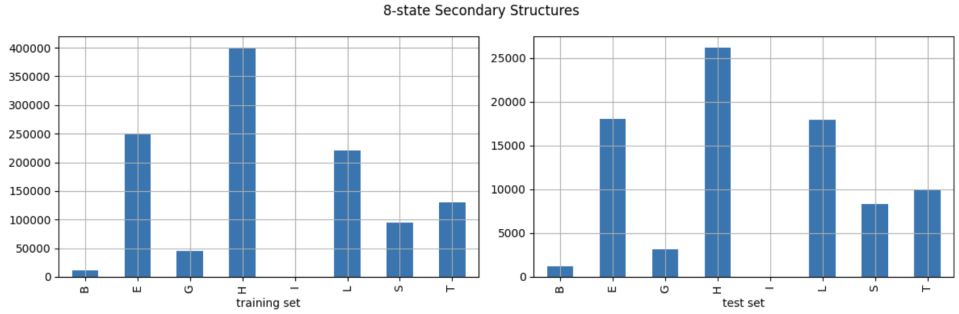

Fig 3. Frequency of 8 Secondary Structures present in the data-set. We observe that most of the protein structures are present in both data-sets expect Pi Helix. Helix is most occurring structure with about 38000 occurrences in the training set and about 26000 occurrences in. the test data.

2. Beta Bridge
3. Beta Strand
4. Helix
5. Pi Helix
6. Alpha Helix
7. Bend
8. Beta-Turn

Method

Primary Methods:
To solve our prediction problem, we first used normal LSTM (Long Short Term Memory Networks) over Recurrent Neural Networks to overcome the problem of vanishing gradients. Unlike traditional RNNs, they use feedback connections to retain the information about the previous data. The LSTM technique is known to perform well with sequential data, such as text recognition and speech recognition, so it is aligned with our data, which is either a . Our best model achieved an accuracy of 67%.For our second trial, we used 2 layer - LSTM to improve the performance of our model by detecting more complex features.

One of the major drawbacks of LSTM network was that it was unidirectional and hence it would not know any context from the future sequence of amino acids or profiles. To overcome this problem, we trained our data with bidirectional LSTM. Bidirectional LSTMs (BiLSTM) are recurrent neural networks used primarily on natural language processing. Unlike standard LSTM, the input flows in both directions, and it is capable of utilizing information from both sides (Past Future). It is also a powerful tool for modelling the sequential dependencies between words and phrases in both directions of the sequence. However, even with the BiLSTM the accuracy did not improve. This was later on rectified by making a change to the training data, we started to train the model with sequence profile data. With BiLSTM, we improved our model's performance by a wide margin and achieved 67% accuracy. Additionally, we were able to prevent over fitting our model using regularisation.

In order to survey a few methods similar to this field of secondary structure prediction of proteins we also look at two novel methods namely Deep-ACLSTM [**?**] and

MUFOLD-SS [**?**]. First, we take a look at MUFOLD-SS, this method concentrates on convolutional neural networks. We have recreated their code and run the same on our CULLPDB dataset. The trick this model uses is that we do parallel convolutions by moving different sized kernels over the same region and then merge them. This helps us greatly as sequential convolution would increase the number of parameters and also would not learn as much as the parallel convolution approach. Multiple such inception blocks **??** are strung along together in this approach. This model was able to achieve a Q8 accuracy of 70.63% on the public benchmark dataset of CB513.

Another interesting approach proposed in the paper Deep-ACLSTM was the stacking of convolutional neural networks above bi-directional LSTMs(Long Short Term Memory Cells). [13] The convolutional layers with different kernel sizes help in capturing multi-scale local contextual features. The features obtained from convolutions of different kernel sizes are concatenated and then further fed to a block of stacked bi-directional LSTMs. The long range dependencies in amino acid sequences and their impact on secondary structure of protein are taken care of by the bidirectional LSTMs layers. In this way, both local and global contextual features are incorporated in the model. This model was able to achieve a Q8 accuracy of 70.5% on the public benchmark dataset of CB513.

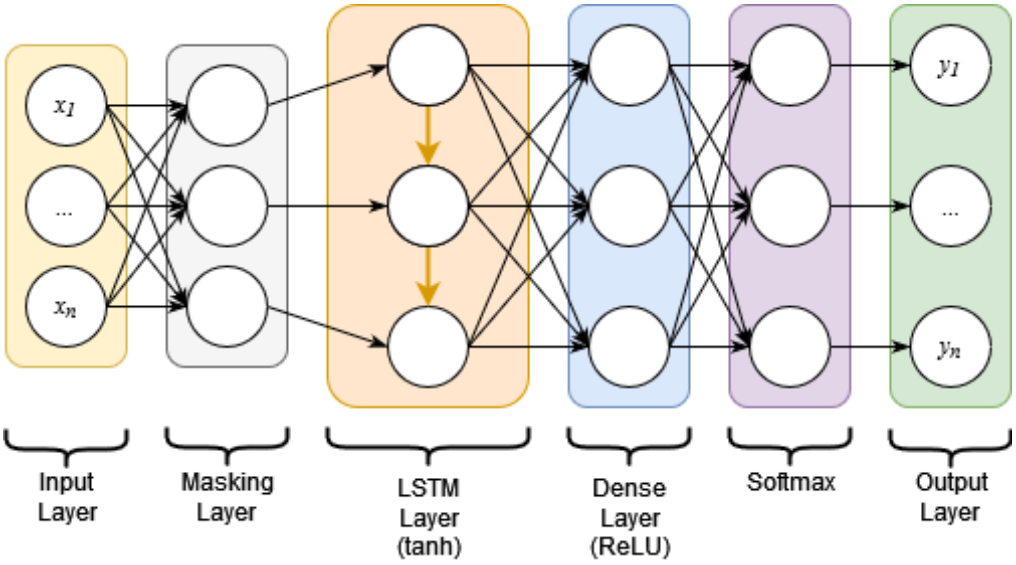

Fig 4. LSTM single layer architecture In this Recurrent Neural Network model, the amino residues are passed as the input which goes into a Masking Layer to make the input length equal which then goes undergoes into one LSTM layer. A fully connected dense layer with ReLU activation function followed by softmax activation is added to get the final output which is the protein structure among the 8 different structures in the dataset.

The following link can be used to access our project:
https://github.com/mahajanhrishikesh/ML-Genomics-Project

Secondary Methods:

To prove the structural similarity between true and predicted protein structure, we made use of Principle Component Analysis. A principal component analysis (PCA) examines underlying correlations among multiple variables (In this scenario it is the eight feature data of structure type which is predicted). This can be achieved by creating uncorrelated variables that maximize variance successively. It simplifies complex data while preserving patterns and trends by transforming the data into fewer

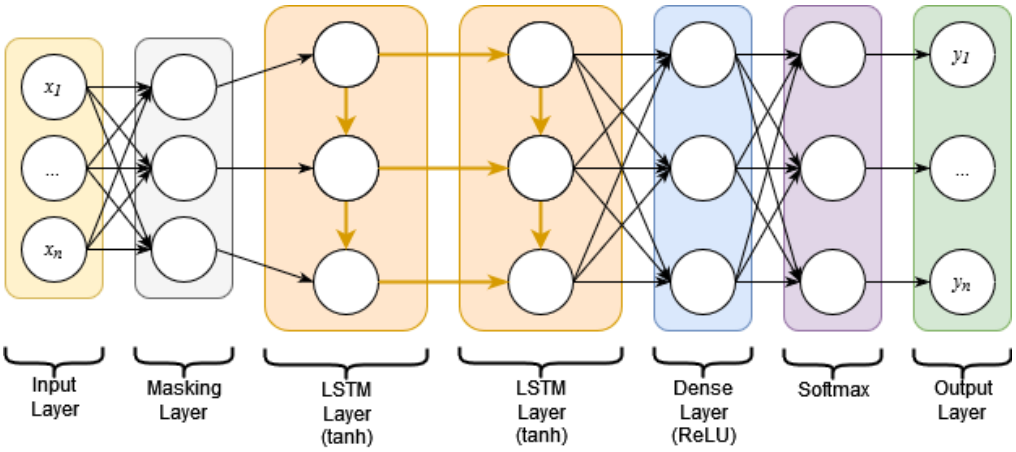

Fig 5. LSTM two layer architecture In this Recurrent Neural Network model, the amino residues are passed as the input which goes into a Masking Layer to make the input length equal which then goes undergoes into one LSTM layer. One more LSTM layer with tanh activation is added. A fully connected dense layer with ReLU activation function followed by softmax activation is added to get the final output which is the protein structure among the 8 different structures in the dataset.

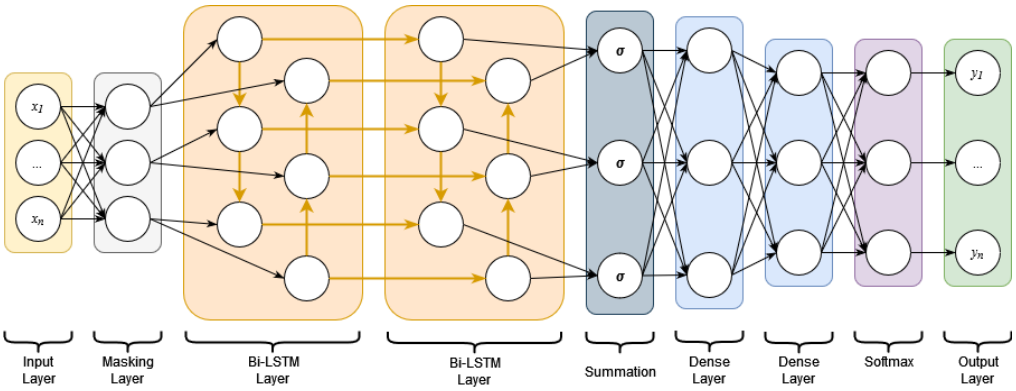

Fig 6. Bidirectional LSTM architecture A Bilateral LSTM layer is used in this model over Unidirectional LSTM. The amino residues are passed as the input which goes into a Masking Layer to make the input length equal which then goes undergoes into two Biderectional LSTM layers. A fully connected dense layer with ReLU activation function followed by softmax activation is added to get the final output which is the protein structure among the 8 different structures in the dataset.

dimensions, which act as summaries of features. After running the analysis and plotting the graph for PCA, we observed that the our model was performing good as the predicted protein structure data was aligned with experimented data. The utilisation of PCA also gave us the insight as in where our model was not performing well which we wouldn't have attained by metrics like Mean Squared Error. As seen in the figure 10 if the predicted secondary level protein structure aligns well with its respective ground truth then it is to be said that the model is working well.

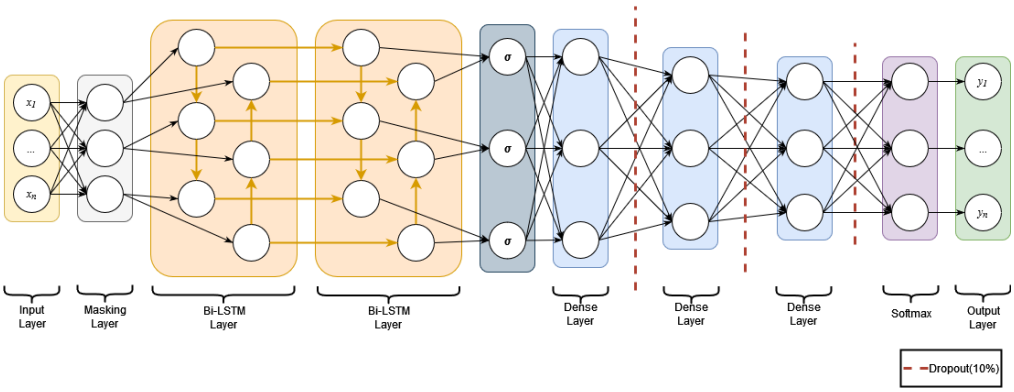

Fig 7. Bidirectional LSTM architecture with dropout Regularisation is added to the previous model as seen in figure 6. A dropout of 0.1 is added, which drops 10 % of nodes randomly during training in each epoch. This handles over-fitting of the training data thus making the model more reliable.

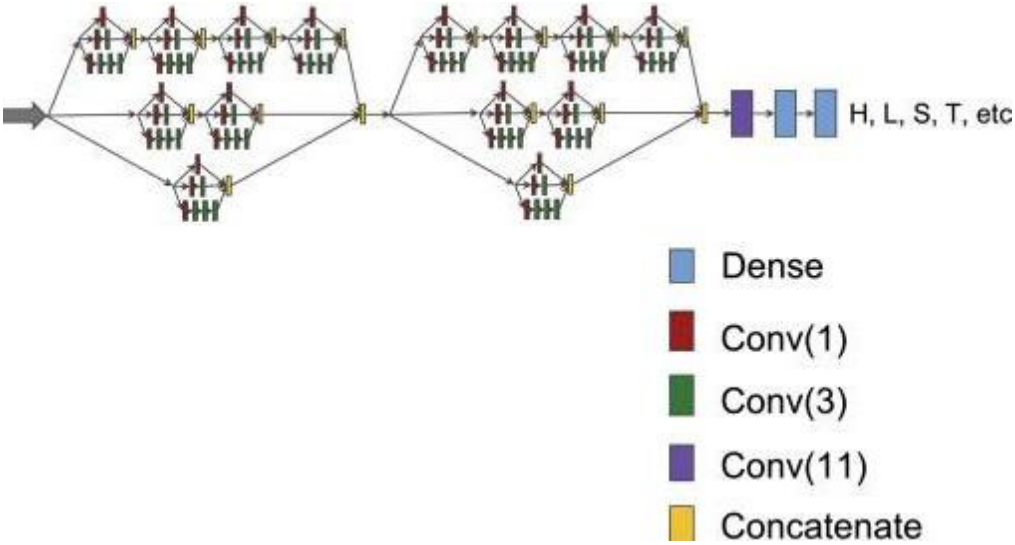

Fig 8. MuFold-SS Architecture [12] Neural Network architecture of the MuFold-SS model. This model uses the conception of using Incpetion Module inside an inception Module. An Inception module basically contains convolutional layers inside convolutional layers. An inception module consists of three parallel convolution operations concatenated at the end of the block. Conv(1), Conv(3), Conv(11) refers to convolutional layers with filter size of 1, 3, 11 respectively. Local features obtained from these parallel convolutions are concatenated. The inception part is followed by a Conv(11) layer, two fully connected layers to predict the secondary structure os the protein.

Results

We have iteratively improved the deep learning based architectures that we have used. As seen in the previous section, 5 incremental steps were tried with improving accuracy each time. We used LSTM single layer as our first model. As seen in Fig 11, this model has the lowest accuracy among the successor. The accuracy graph during every epoch is shown in the Fig 19. The accuracy obtained is around 44%. For our second model , we chose LSTM with two layers. As seen in Fig 12, this model has the accuracy slightly

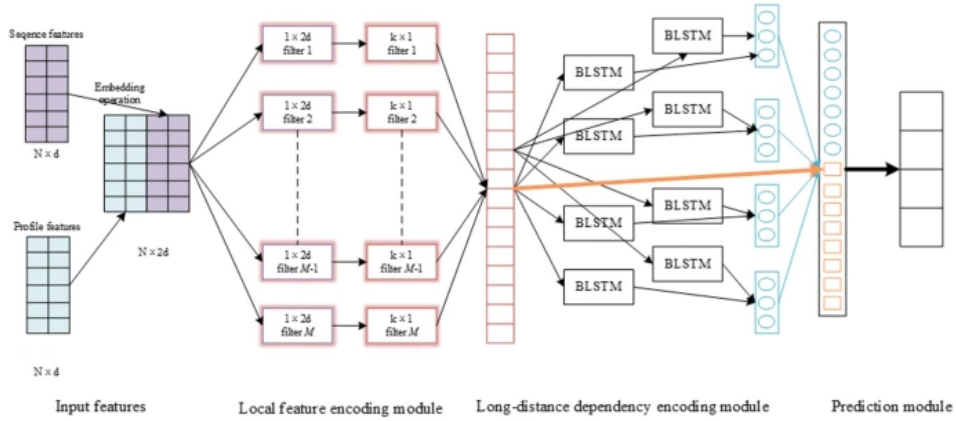

Fig 9. Deep ACLSTM Architecture [13] Architecture diagram of the Deep ACLSTM(Deep Asymmetric Convolutional Long Short Term Memory) model. The model uses two inputs; main(sequence data) and auxiliary(sequence profile). After the embedding, the result is concatenated with auxiliary input and passed to multi-scale CNN layer. The output from CNN is fed to the stacked bi-directional LSTM layers. The final two fully-connected layers perform the task of classification of the structure.

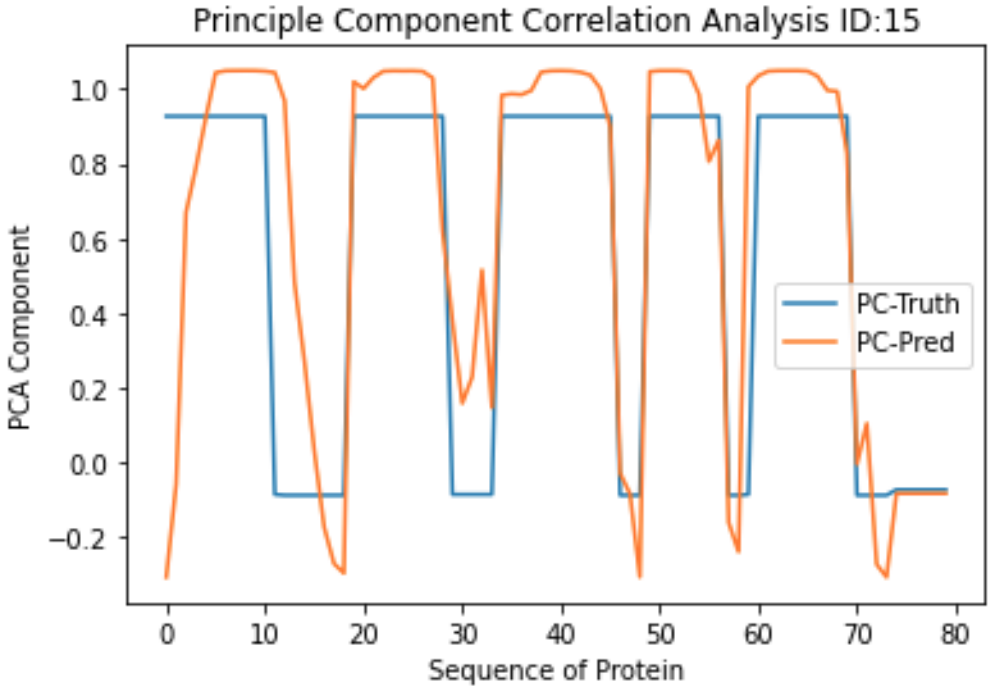

Fig 10. Principal Component Correlation Analysis for a protein. Randomly a protein is selected from the test data-set, in this case 15th Protein to find the co-relation. PCA Analysis on actual secondary data which is the row [22,32) as seen in table 2 is performed. Then PCA is performed on the predicted output from the model. The graph shows the actual data in blue and predicted output in orange.

more than the predecessor. The accuracy graph during every epoch is shown in the Fig 19. The accuracy obtained is around 45%. To further improve the accuracy, we

used Bi-directional LSTM. As we are predicting structure of protein here, a sequential model was not sufficient as the accuracy of last two models depicted. Here we tried using an Bidirectional Model, which gave us significantly good accuracy. The reason being structure cannot be defined by a sequential data. As seen in Fig 13, this model's better accuracy can be seen . The accuracy graph during every epoch is shown in the Fig 19. The accuracy obtained is around 54%. We utilized Bidirectional LSTM with Regularisation for our fourth model. In this model, dropout is added to the previous model. Dropout randomly selects neurons and ignore them during training. Making the model accurate to varied data. As per Fig 14, the accuracy is been affected due to the dropout . The accuracy graph during every epoch is shown in the Fig 19. The accuracy obtained is around 47.5%. To further improve our model, we used Bidirectional LSTM with Sequence Profile data. The same Bidirectional LSTM model is trained using Sequence Profiles, which is present in the rows [35,57) of the CullPDB 5926 dataset. As per Fig 15, the accuracy has improved significantly. The accuracy graph during every epoch is shown in the Fig 19. The accuracy obtained is around 65%. As per figure 19, our best model has the ROC Curve.Summarizing, moving from sequential LSTM model to Bidirectional LSTM model was the most improvement for the model. The accuracy was improved from 44% to 65%, a staggering 21% of accuracy improvement.
The MUFOLD-SS delivers a Q8 accuracy of 70.63% on CB513. This was the best method mentioned in this work as it has the highest accuracy. The hierarchically stacked convolution blocks work together along with regularization techniques such as Dropout help in making the model robust and can generalize efficiently.
The Deep-ACLSTM with 300 LSTM nodes performs protein secondary structure prediction with a Q8 accuracy of 70.5%, 75% and 73% on CB513, CASP10 and CASP11 datasets respectively. The state of the art accuracy of Deep-ACLSTM is justified by its technique of incorporating both the local and global contextual features and thus capturing short and long range dependencies between the amino acids and its impact on the protein's secondary structure.

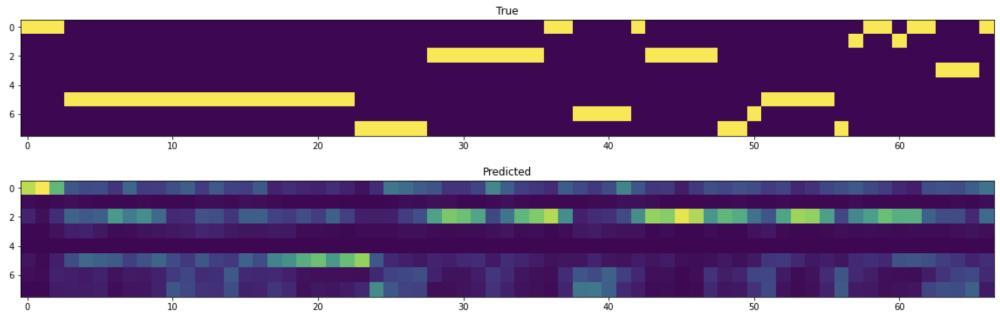

Fig 11. LSTM Single Layer Secondary Structure Prediction Y-axis shows the eight secondary types and X axis is the sequence of first protein in the dataset.This model gave an accuracy of about 44% as seen in figure 19

Discussion

In this paper we tried out multiple methods for prediction of the secondary structure of proteins. In addition to creating models ourselves we have reviewed two more state of the art methods [] and []. Using our methods of using the bidirectional LSTM we have attained an accuracy of 67.89% over our test dataset. Our journey of began with the simple LSTM model for the secondary structure prediction. This however was not enough as the unidirectional LSTM does not provide insight from the data that is to

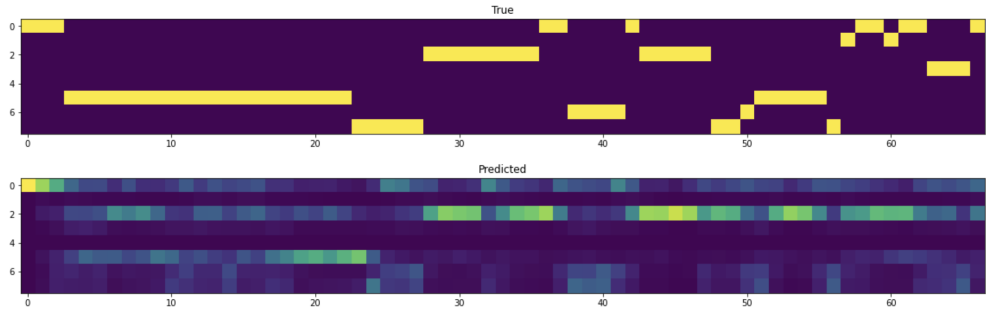

Fig 12. LSTM two layers Secondary Structure Prediction Y-axis shows the eight secondary types and X axis is the sequence of first protein in the dataset.This model gave an accuracy of about 45% as seen in figure 19

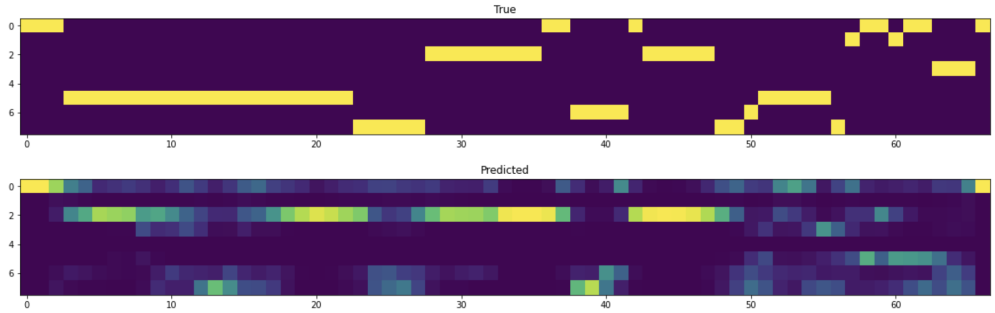

Fig 13. Bi-directional LSTM Secondary Structure Prediction Y-axis shows the eight secondary types and X axis is the sequence of first protein in the dataset.This model gave an accuracy of about 54% as seen in figure 19

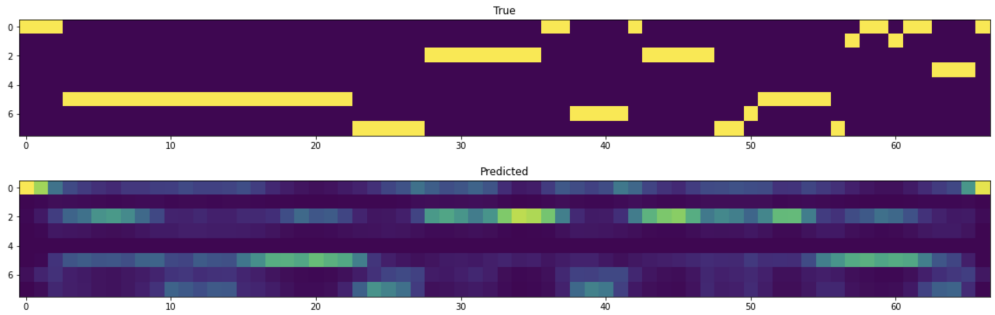

Fig 14. Bidirectional LSTM with Regularisation (Dropout) Secondary Structure Prediction Y-axis shows the eight secondary types and X axis is the sequence of first protein in the dataset.This model gave an accuracy of about 47.5% as seen in figure 19

follow which is a critical piece of information in a task such as this. Moreover the outputs from this model were mostly non definitive in nature and none of the sections in the proteins were clearly matching the expected values.

The novelty of our approach lies in picking the correct amount of model complexity along with the correct data being selected. The crux of predicting the secondary structure is having the future values considered as well. Hence there needs to be a way to propagate the future values or atleast a window of the values back to the nodes. Considering the limited window view of CNN we considered using the BiDirectional LSTMs. After adding the bidirectional LSTM the accuracy improved considerably (Up

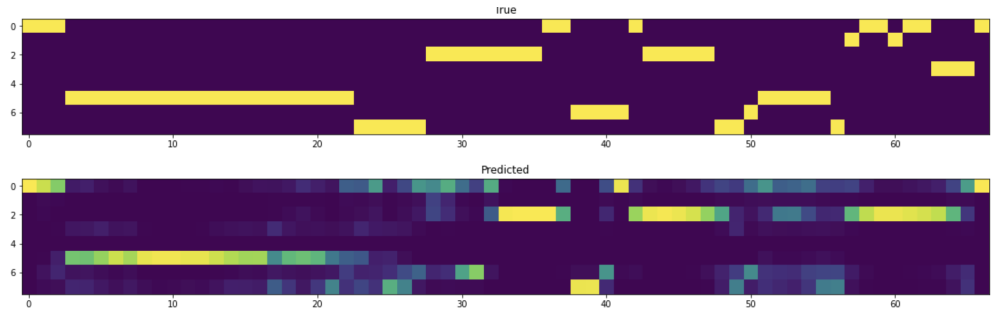

Fig 15. Bidirectional LSTM with Sequence Profile data Secondary Structure Prediction Y-axis shows the eight secondary types and X axis is the sequence of first protein in the dataset. This model gave an accuracy of about 65% as seen in figure 19

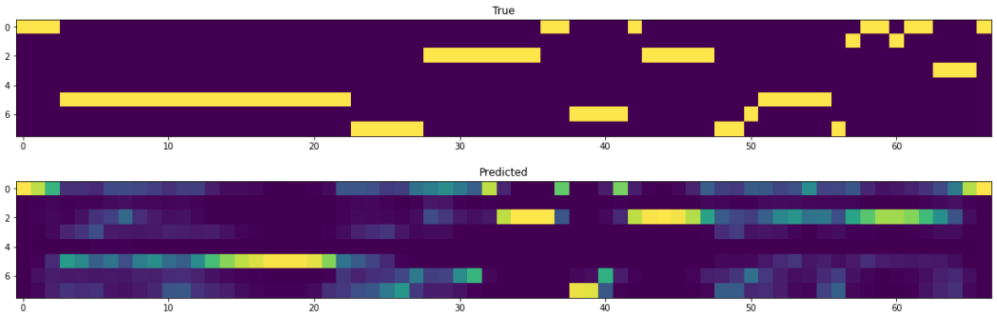

Fig 16. Sequence + Profile Data Y-axis shows the eight secondary types and X axis is the sequence of first protein in the dataset. This model gave an accuracy of about 68% as seen in figure 19

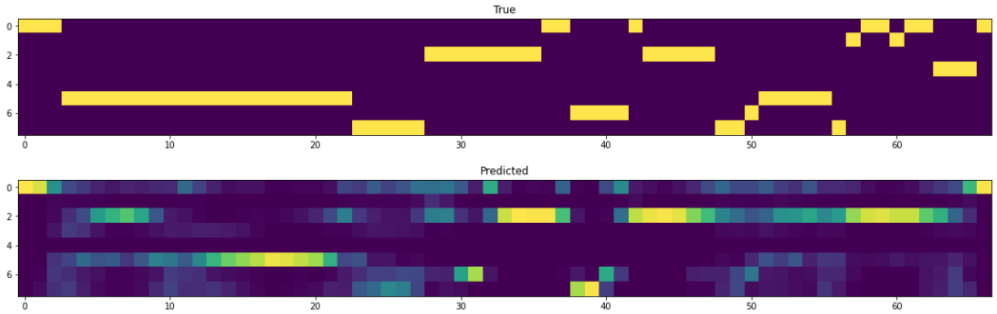

Fig 17. MuFold-SS Prediction Y-axis shows the eight secondary types and X axis is the sequence of first protein in the dataset.This model gave an accuracy of about 70% as per the paper Inception-Inside-Inception Networks for Protein Secondary Structure Prediction [12]

to 54%). Up until now we had been using only the sequence data from our dataset. If we switched to training with the profile data instead the accuracy bumps up to 65%. This is observed as the sequence profiles contain more information than the regular sequence data. Furthermore, using sequence data along with the profile data yields an accuracy of 67.89%. The combination of profile as well as sequence data to train a bidirectional network for protein secondary structure prediction is our novel addition to this problem.

In order to survey other methods that do the same thing, we explored two other methods that go by the names MUFOLD-SS and DeepACLSTM. These methods make use of CNN (70.63% Accuracy) and CNN+BLSTM (70.50% Accuracy) respectively. An

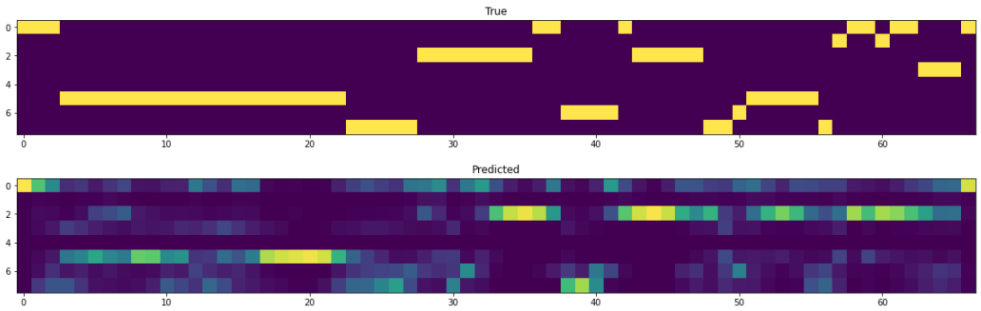

Fig 18. DeepACLSTM Prediction Y-axis shows the eight secondary types and X axis is the sequence of first protein in the dataset.This model gave an accuracy of about 70% as per the paper DeepACLSTM: deep asymmetric convolutional long short-term memory neural models for protein secondary structure prediction. [13]

interesting point to note is that the addition of the CNN likely gives the model an added edge. This is because the CNNs work on capturing the local patterns and the LSTMs take into account the long range patterns.

It is observed that these methods consume a lot of resources and are very compute intensive however, our Bidirectional LSTM is able to provide somewhat similar results albeit a bit less accurate but at a lesser computation cost. Perhaps, a better tuning of the Bidirectional LSTM only model along with a better exposure to the data will make the model perform as optimally as these two other methods.

Conclusion

We have satisfactorily solved the scientific goal that we stated early on which was to do secondary protein structure prediction using only amino acid data by experimenting with multiple methods as well as adding our own contribution to the problem. For solving the problem of protein structure prediction we used multiple LSTM based methods and settled upon the approach of using bidirectional LSTMs along with the sequence data as well as the profile data. Our approach fetched us an accuracy of 67.89% which is comparatively better in terms of computation expense than the other compute intensive methods available such as MUFOLD-SS and DeepACLSTM. However the other two methods do deliver a slightly greater accuracy than our model. Increasing the accuracy of a purely LSTM based approach is also possible by fine tuning the parameters and exposing the model to more varied data.

Supporting information

Bidirectional LTSM Bidirectional LTSM connects two hidden layers of opposite directions to the same output. Using this model, the output layer can receive information from past states as well as future states.

Inception CNN An inception model lets the CNN use mutiple types of the filter size inspite of a single filter size in a single image block and then concatenate and pass it onto the next layer.

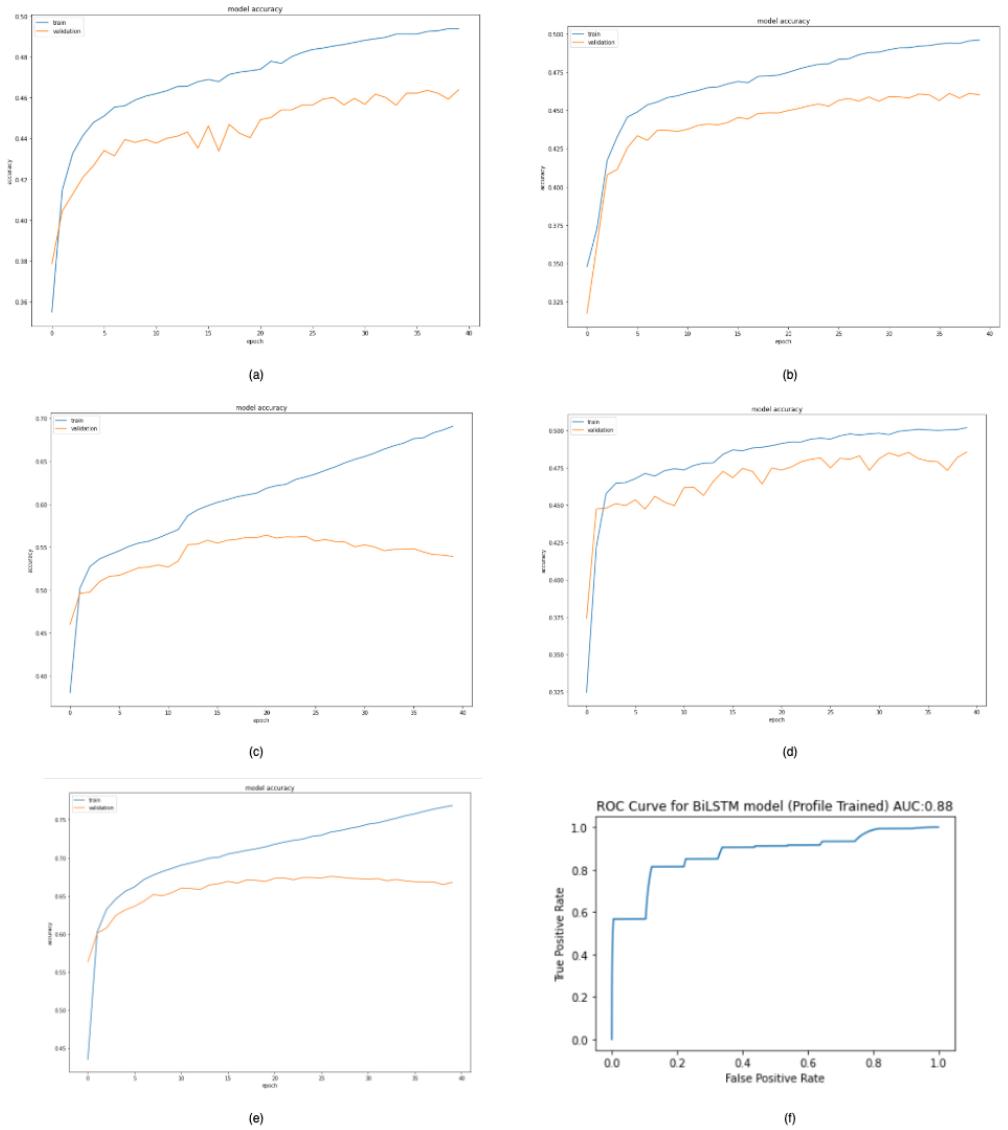

Fig 19. Accuracy Plot a) The graph shows accuracy of single layer LSTM model. Being the lowest, this model has accuracy of 44%. b) The graph shows accuracy of two layer LSTM model. This model has accuracy of 45%, slightly better than previous model. c) The graph shows accuracy of Bidirectional LSTM model. This model has accuracy of 54%, a good rise over the LSTM model. d) The graph shows accuracy of Bidirectional LSTM Model with Regularisation. This model has accuracy of 47.5%. e) The graph shows accuracy of Bidirectional LSTM Model with Sequence Profile Data. This model gave the highest accuracy of 54%. f) The graph shows accuracy of ROC Curve of the Bidirectional LSTM Model with Sequence Profile Data.

Acknowledgments

We would like to thank Prof. Dr. Kiley Graim for the constant support, encouragement and insightful suggestions during the course of the paper.

References

1. Deep supervised and convolutional generative stochastic network for protein secondary structure prediction. https://dl.acm.org/doi/10.5555/3044805.3044890.
2. ICML2014 data-set CullPDB. https://www.princeton.edu/ jzthree/datasets/ICML2014/.
3. Zhou X, Chou J and TC Wong Protein structure similarity from principle component correlation analysis. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1386710/
4. Hochreiter, Sepp & Schmidhuber, Jurgen. Long Short-term Memory. Neural computation. https://www.researchgate.net/publication/13853244_Long_Short-term_Memory
5. M. Schuster, K.K. Paliwal. Bidirectional recurrent neural networks. https://ieeexplore.ieee.org/document/650093
6. Jian Zhou and Olga G. Troyanskaya (2014) - "Deep Supervised and Convolutional Generative Stochastic Network for Protein Secondary Structure Prediction" - https://arxiv.org/pdf/1403.1347.pdf.
7. Sheng Wang et al. (2016) - "Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields" - https://arxiv.org/pdf/1512.00843.pdf.
8. Li, Zhen; Yu, Yizhou, Protein Secondary Structure Prediction Using Cascaded Convolutional and Recurrent Neural Networks, 2016.
9. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, Attention Is All You Need https://arxiv.org/abs/1706.03762
10. Qian Jiang, Xin Jin, Shin-JyeLee, Shaowen Yao Protein secondary structure prediction: A survey of the state of the art. https://www.sciencedirect.com/science/article/pii/S1093326317304217
11. Hang Chen, Fei Gu, Zhengge Huang Improved Chou-Fasman method for protein secondary structure prediction. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-7-S4-S14
12. Chao Fang, Yi Shang, and Dong Xu1, MUFOLD-SS: New Deep Inception-Inside-Inception Networks for Protein Secondary Structure Prediction. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6120586/
13. Yanbu Guo, Weihua Li, Bingyi Wang, Huiqing Liu , Dongming Zhou , DeepACLSTM: deep asymmetric convolutional long short-term memory neural models for protein secondary structure prediction. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-2940-0
14. Gia G. Maisuradze, Adam Liwo, and Harold A. Scheraga. Principal component analysis for protein folding dynamics. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2652707/

15. Narayanan Eswar, Ben Webb,Marc A. Marti-Renom, M.S. Madhusudhan, David Eramian, Min-yi Shen, Ursula Pieper, and Andrej Sali. Comparative Protein Structure Modeling Using Modeller.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4186674/

16. B Rost 1, R Schneider, C Sander Protein fold recognition by prediction-based threading. https://pubmed.ncbi.nlm.nih.gov/9237912/

17. Jooyoung Lee, Peter L. Freddolino and Yang Zhang Ab Initio Protein Structure Prediction. https://zhanggroup.org//papers/2017_3.pdf

18. Artificial Intelligence in Prediction of Secondary Protein Structure Using CB513 Database. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3041573/

19. Protein Structure Prediction Center CASP.
https://predictioncenter.org/index.cgi