Leveraging Deep Learning Methodologies for Protein Secondary Structure Prediction

Sai Madhur Mallampalli
Graduate Student, Computer Science
School of Science, Engineering and Technology
Penn State Harrisburg, The Pennsylvania State University
Middletown, PA 17057, United States
sbm6433@psu.edu

Hyuntae Na
Associate Professor, Computer Science
School of Science, Engineering and Technology
Penn State Harrisburg, The Pennsylvania State University
Middletown, PA 17057, United States
hzn17@psu.edu

***Abstract*—Protein secondary structure prediction (PSSP) is critical for understanding protein interactions and aids in drug discovery. Despite being foundational, traditional statistical techniques can lack the accuracy and efficiency required for complex protein structures. This research investigates how advanced deep learning approaches, notably Bi-directional Long Short-Term Memory (Bi-LSTM) networks, might improve protein secondary structure prediction. The study seeks to considerably improve the predictive accuracy of secondary structure prediction from amino acid primary sequences by implementing these approaches on large protein datasets, such as those obtained from PISCES.**

***Index Terms*—Protein Secondary Structure Prediction, BiLSTM, Deep Learning, Data Preprocessing and Visualization**

I. Introduction

Proteins are essential components of our bodies and are typically classified into four distinct categories. Protein structures are composed of amino acids that are linked with peptide bonds. The primary structure refers to the way the amino acids are sequenced. The secondary structure is formed by the dihedral angles that surround peptide bonds. The tertiary structure refers to the protein folds. Quaternary structure is produced when folded polypeptide molecules bind to the functional proteins. In molecular biology, knowledge of protein secondary structure is essential for designing medications and modeling genetic processes. Alpha-helices, beta-sheets, and random coils are examples of secondary structures in proteins that are essential for determining a protein's three-dimensional conformation, which in turn dictates interactions and functionality. Originally, applying statistical techniques derived from empirical data, the Chou-Fasman [1] and GOR [2] algorithms have been crucial in predicting these structures. Using information theory and Bayesian statistics, the GOR technique [2] estimates the probability of secondary structures based on surrounding amino acid characteristics at each place in a protein sequence. However, these traditional methods often struggle to meet the high accuracy demands posed by increasing protein complexity and larger datasets. With the advent of deep learning, and the usage of Recurrent Neural Networks (RNNs) [3] and Convolutional Neural Networks (CNNs) [4] in the secondary structure prediction has shown significant advancements in numerous bioinformatics applications, including protein secondary structure prediction. This paper explores the applications of Bi-LSTM (Bi-directional Long Short-Term Memory) network, a kind of recurrent neural network that processes data points from both past and future states, enhancing learning dynamics and prediction accuracy.

The methodology leverages comprehensive datasets, culled through PISCES [5], to train and evaluate the Bi-LSTM model and the objective is to demonstrate that these models significantly outperform traditional statistical methods in predicting protein secondary structures, offering both higher accuracy and computational efficiency. These advancements have implications beyond academia, potentially accelerating the development of new therapeutics and enhancing our understanding of genetic functions.

II. Literature Review

The initial approaches to PSSP were based on statistical methods, leveraging the observation that certain amino acids were more likely to be found in certain kinds of secondary structures. The Chou-Fasman method as reviewed by Wolfgang Kabsch and Christian Sander [6] is a classic example, using propensities of amino acids to predict secondary structure with a 56% accuracy. The introduction of machine learning algorithms, including SVMs and early neural networks, marked a significant improvement as they leveraged sequence profiles and sometimes structural information from known protein structures, pushing accuracy rates to 73%-77% through SVM and neural networks by Ward JJ, McGuffin LJ, Buxton BF, Jones DT [7].

With the advent of deep learning, CNNs became popular for their ability to capture local patterns along the sequence. Zhou J and Olga G, [8] present a new approach for PSSP, using a supervised generative stochastic network (GSN), they incorporated a Markov chain that samples from a conditional distribution of outputs based on inputs by employing a position specific scoring matrices for capturing evolutionary information and binary vectors for encoding amino acid sequences. The convolutional architecture allows the model to process the high-dimensional space of protein sequences effectively, using both local and global information and it achieved an accuracy of 66.4%

Models like DeepCNF (Deep Convolutional Neural Fields) by Wang, S., Peng, J., Ma, J. et al. [9] further improved accuracy by integrating sequence features with conditional random fields to capture both complex sequence-structure relationships and interdependencies between adjacent secondary structure labels, achieving accuracies above 80% for Q3 and 72% for Q8.

Ahmadi et al. [10] focused on enhancing the accuracy of predicting protein secondary structures using evolutionary optimized neural network (EONN) and evolutionary optimized support vector machine (EOSVM) models, incorporating algorithms like genetic algorithm (GA), differential evolution (DE), and particle swarm optimization (PSO). They encoded the input features and integrated evolutionary algorithms coupled with a sliding window size to handle the sequence predictions by achieving accuracy upto 78.83%. Lyu et al. [3] proposed a reductive model that combines multilayer perceptrons with bidirectional gated recurrent units and hidden markov model profiles to capture information from CB513 dataset for predictions resulting in an increase in accuracy by achieving Q3 accuracy of 83.32% and Q8 accuracy of 70.51%. Tamzid et al. [11] delve into the usage of Graph Neural Networks for PSSP with an intense focus on graph construction, node embeddings, and the use of SVM for classification through the usage of orthogonal encoding and padding the amino acid sequences on both ends to accommodate the window size and achieved an accuracy of 76.89%. Hu et al. [12] incorporated five types of protein features including physio-chemical properties, position specific substitution matrices (PSSM) scores, PSSM count, hidden Markov model (HMM) sequence profiles, and word embedding for amino acid encoding and proposed an ensemble algorithm on BiLSTM to combine multiple sub models with each sub model solving a processing subproblem for each feature with an accuracy of 84.3% . [13] Yang et. al, approaches have achieved accuracy (about 84%) that is close to the theoretical limit (88%) in the Deep Learning domain, and similarly as given by Dewi Pramudi et al. [14], AlphaFold has disrupted the way protein structure prediction by achieving accuracies upto 90% and pushing the envelope higher.

III. Methods

The methodology employed in this project involved extensive data pre-processing and exploratory data analysis on datasets culled through PISCES [15] [16] and CB513 [17] before integrating it with a Bidirectional LSTM Model for the secondary structure prediction from the primary sequence of the amino acids.

A. Data Preprocessing and Visualization

Multiple datasets containing protein structures with varied sequence identity and resolution cutoffs were preprocessed to gauge the distribution of data and understand the scope and nuances of the subproblems that define the secondary structure predictions. These datasets were imported from CSV files sought through Github repositories [15] [16] that used PISCES server to cull data suitable for predictive analysis. The datasets from 2022 contained protein structures filtered at 25% and 30% sequence identity and 2.0 and 2.5 Angstrom resolution cutoffs. The inclusion of an archival dataset from 2018 aided in the comparison of the distribution of data. A curated version of a mid-2022 dataset where duplicate entries based on primary sequence identity were removed, retaining only the first occurrence to enhance the training mechanism. The features included features like pdb, primary sequence(seq), and secondary sequence in sst8 and sst3 classifications, length of the sequences named len_x, nonstdaa which checks if the amino acid is standard or redundant, and irrelevant features like rfac, freerfac and resol as they deal with 3D diffraction. Hence, feature extraction was performed by selecting seq and sst3 as they are the only two features that are required for the classification task.

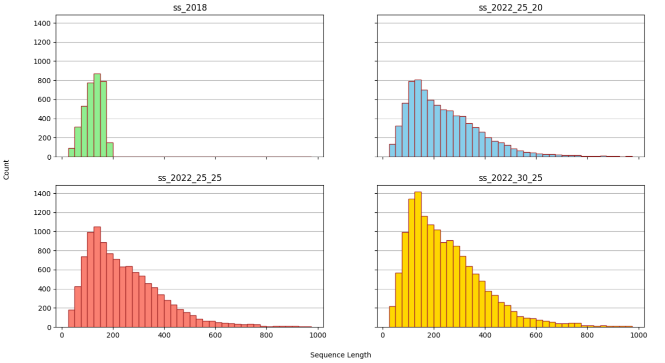


Fig. 1. Sequence distribution across datasets based on length.

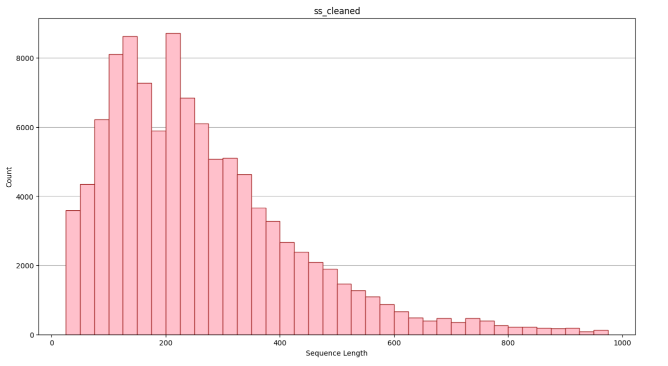
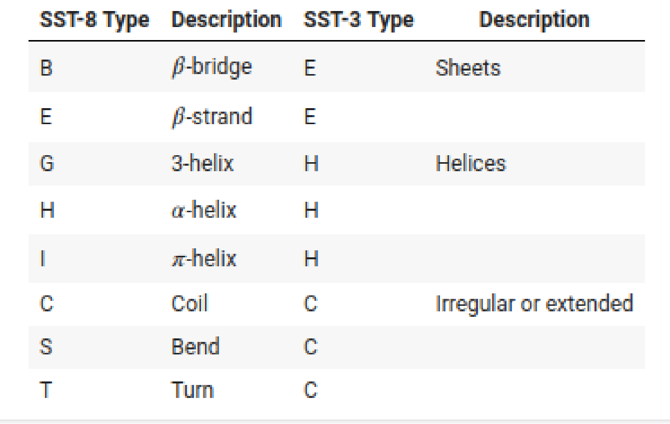


Fig. 2. Sequence distribution in curated dataset based on length.

The above Figures 1 and 2 show the distribution of primary sequences on all the datasets and it is clear that majority of the sequences are between the length 40 - 300 across all data sets.

1) Q3 vs Q8: Figure 3 illustrates how the secondary structure is divided into two types. SST-8, which categorizes into eight specific types, and SST-3, which groups these into three broader categories. In SST-8, the types include: β-bridge (B), β-strand (E), 3-helix (G), α-helix (H), π-helix (I), Coil (C), Bend (S), and Turn (T). The SST-3 system simplifies these



SST-8 Type	Description	SST-3 Type	Description
B	β-bridge	E	Sheets
E	β-strand	E	
G	3-helix	H	Helices
H	α-helix	H	
I	π-helix	H	
C	Coil	C	Irregular or extended
S	Bend	C	
T	Turn	C	

Fig. 3. Secondary Structure Classification

into Sheets (E), encompassing both β-strands and β-bridges; Helices (H), which include α-helices, π-helices, and 3-helices; and Irregular or extended structures (C) such as coils, bends, and turns.

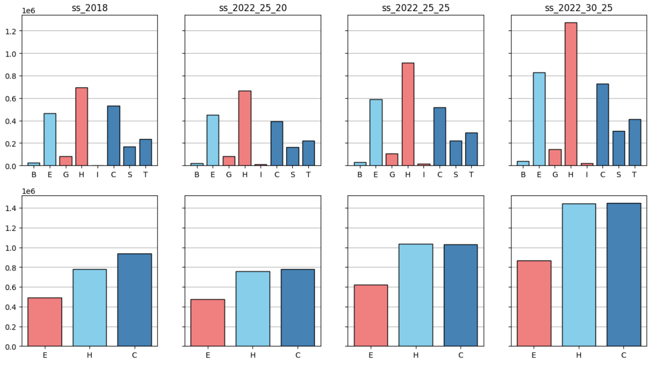


Fig. 4. Secondary Structure distribution among the datasets for testing

It was observed that Figures 4 and 5 show the distribution of secondary structure labels on all the datasets and it is evident that the frequency of C and H is higher than E.

2) Data Preparation for Training: We subset protein sequence data to prepare distinct training, validation, and testing sets, essential for the robust evaluation of predictive models. Sequences were filtered to include only those within a specified length range 40 - 300, enhancing the focus on sequences of relevant sizes for structural prediction and removing outliers. The data underwent further partitioning to create non-overlapping training and validation sets, with additional steps taken to remove internal duplicates within the validation set to maintain data integrity and prevent overfitting. The primary sequences were transformed into k-mers similar to [18] which is an incorporation of n-gram methodology that is utilized in natural language processing tasks to transform sequences to subsequences of specific length k. This technique was followed by the capturing of local sequence features, tokenizing them and encoding them as integers with substantial padding to

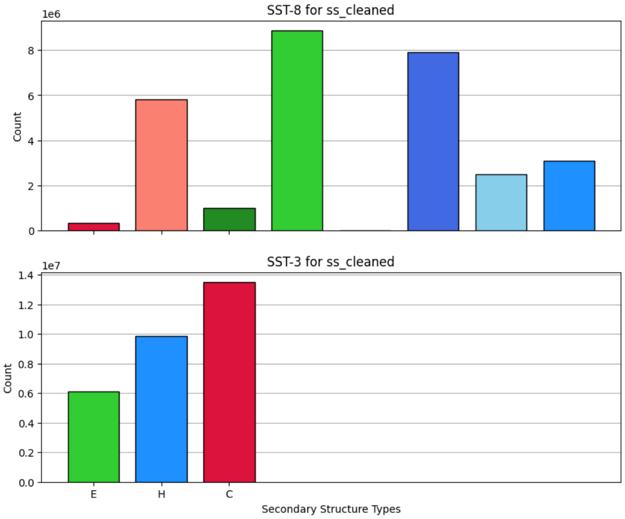
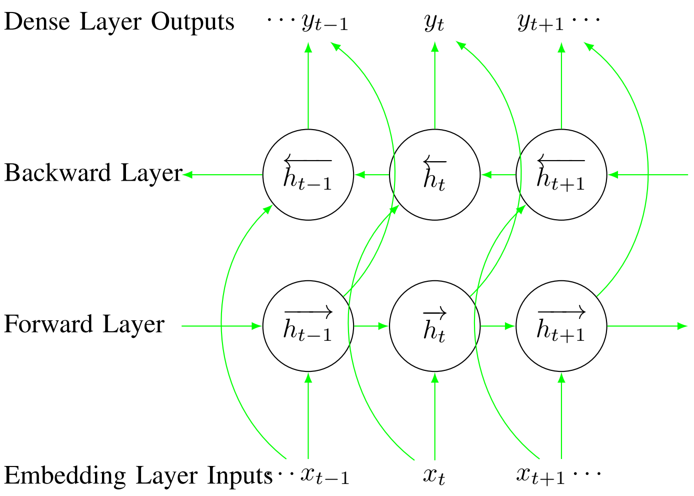


Fig. 5. Secondary Structure distribution in the curated dataset for training and validation

the specified length to ensure consistency and suitability of data before being fed into the model. Subsequently, the target secondary structures were encoded using the character-level tokenizer that maps each unique structural character to a unique integer and then padded to the same fixed length as the sequence data and then perform one hot encoding on vectors for preparation.

B. Model - Training and Testing

A Bi-LSTM model was built to implement the prediction and decoding of the secondary structure from the one-hot encoded primary sequence. Bi-LSTM was employed to construct the model as LSTM captures the long-range dependencies within the sequences as the understanding of the entire sequence is crucial in this task as the interaction between the surrounding amino acids can influence the secondary structure. The bidirectionality allowed the model to capture the context from both ends of the sequence providing an inherent well-rounded process that does not overlook integral aspects during the training, validation and testing.



The model receives the encoded input at each time step and the embedding layer transforms the discrete tokens into continuous vectors that represent the embedded tokens. The Bidirectional layer processes the embedded inputs in both forward and backward directions, generating a sequence of hidden states for each direction at each time step. The time-distributed dense layer applies a dense neural network to each time step independently, converting the LSTM output at each time step into a more refined representation that maps directly to the desired output space through a softmax activation function and a dropout rate of 0.1 % to prevent overfitting during training. Once the model was built, it was trained based on varied lengths of the sequences. The model was trained on the curated dataset with no duplication and no overlap with the test sets. To decrease the categorical cross-entropy loss between the actual and predicted values, the accuracy was computed for the encoded characters in the input and output sequences and ignoring padding classes combined with an RMSprop optimizer [19].

The model was trained for 10 epochs achieving a training accuracy of 0.8719 and a validation accuracy of 0.8553. The testing was performed mimicking the approach in training by initially generating the k-mers from the primary sequences from the test sets, tokenizing the k-mers by encoding the kmer strings into sequences of integers where each integer represents a token and padding sequences with zeros to maintain uniformity in length. The target secondary structures were tokenized into kmers, padded to match the length of the input sequences and then the encoded target integer sequences were converted into a binary matrix to be employed in categorical cross-entropy loss calculation during model evaluation. The one-hot encoded predictions were transformed back to the original secondary sequences through a reverse mapping function that uses a reverse index dictionary that interprets the model's predictions and compares them to actual sequences and the Q3 accuracy was calculated.

IV. Experiments and Results

The model was run on different test sets that were described in the data visualization with the Q3 accuracy ranging from 0.65 to 0.72 for the three-state predictions due to the varied limits for each model and the k-mer size as given in I. The model performs better when the max length is at 400 as it ensures that the outliers are not taken into consideration and a k-mer size of 3 is chosen.

TABLE I
Performance Evaluation on Test Sets

Datasets	*Q3 Accuracy*	*No of Instances*
ss_test	0.65	6968
ss_test_25_25	0.68	4923
ss_test_25_20	0.72	3774
ss_test_2018	0.69	4058
CB513	0.71	517

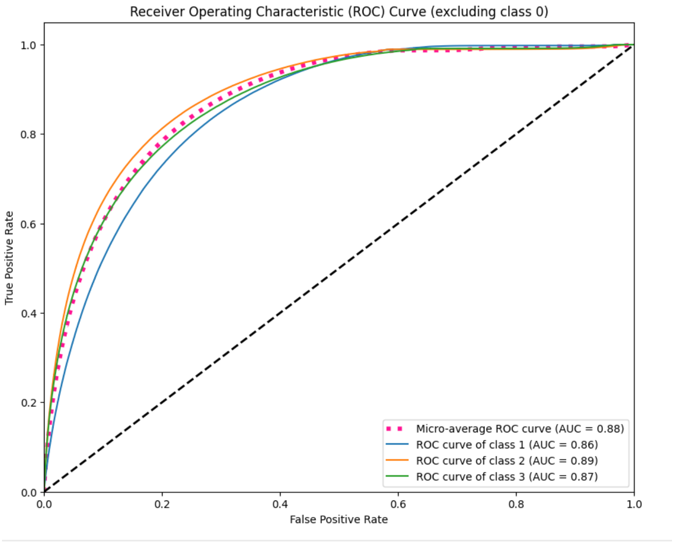


Fig. 6. AUC-ROC for the model

As shown in 6, the AUC is at an average of 0.88 which indicates a good metric for the model.

V. Conclusions and Future Scope

The aim of the paper was to explore and understand the dynamics between primary sequence and the three-state secondary structure of protens and demonstrate the efficacy of Bi-directional Long Short-Term Memory (Bi-LSTM) networks in modelling protein secondary structure prediction (PSSP).Using huge datasets combined with data preprocessing and visualization approaches led to a significant increase in accuracy for the BiLSTM model over conventional statistical methods. The model's capacity to examine the underlying intricacies of the protein sequences by utilizing data points at both ends of the k-mer window guaranteed that the complicated relationships are not overlooked. Usage of non-zero classes for training and testing ensured that padded characters do not cloud the model. The future scope in this research domain could include expanding the methodology to eight-state prediction using embedding techniques like orthogonal encoding and interweaving mechanisms like Attention and Transformer with various deep learning techniques like a fusion of BiLSTMs with GNNs that will substantially increase the accuracy of PSSP and eventually aid in 3D modelling of protein structures.

Acknowledgment

This document is prepared by referencing Zobel's book [20] and wikibook [21].

References

[1] Chou, Peter Y. and Fasman, Gerald D. Prediction of protein conformation, Biochemistry Vol 13, https://doi.org/10.1021/bi00699a002

[2] Jean Garnier, Jean-François Gibrat, Barry Robson, [32] GOR method for predicting protein secondary structure from amino acid sequence, Methods in Enzymology, Academic Press, Volume 266, 1996,ISSN 0076-6879,ISBN 9780121821678, https://doi.org/10.1016/S0076-6879(96)66034-0

[3] Lyu Z, Wang Z, Luo F, Shuai J, Huang Y. Protein Secondary Structure Prediction With a Reductive Deep Learning Method. Front Bioeng Biotechnol. 2021 Jun 15;9:687426. doi: 10.3389/fbioe.2021.687426. PMID: 34211967; PMCID: PMC8240957.

[4] Jinyong Cheng, Yihui Liu, Yuming Ma, Protein secondary structure prediction based on integration of CNN and LSTM model,Journal of Visual Communication and Image Representation, Volume 71, 2020, 102844, ISSN 1047-3203, https://doi.org/10.1016/j.jvcir.2020.102844.

[5] https://dunbrack.fccc.edu/pisces/

[6] Wolfgang Kabsch and Christian Sander *How good are predictions of protein secondary structure?*, Volume 155, number 2, May 1983 https://febs.onlinelibrary.wiley.com/doi/epdf/10.1016/0014-5793%2882%2980597-8

[7] Ward JJ, McGuffin LJ, Buxton BF, Jones DT. Secondary structure prediction with support vector machines. Bioinformatics. 2003 Sep 1;19(13):1650-5. doi: 10.1093/bioinformatics/btg223. PMID: 12967961 https://pubmed.ncbi.nlm.nih.gov/12967961/

[8] Zhou J, Olga G, Deep Supervised and Convolutional Generative Stochastic Network for Protein Secondary Structure Prediction. Proceedings of Machine Learning Research, 2014 https://arxiv.org/pdf/1403.1347

[9] Wang, S., Peng, J., Ma, J. et al. Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. Sci Rep 6, 18962 (2016). https://doi.org/10.1038/srep18962

[10] Ahmadi Toussi, Cyrus & Haddadnia, Javad. (2019). Improving Protein secondary structure prediction; the evolutionary optimized classification algorithms. Structural Chemistry. 30. 10.1007/s11224-018-1271-5.

[11] T. H. Nahid, F. A. Jui and P. C. Shill, "Protein Secondary Structure Prediction using Graph Neural Network," 2021 5th International Conference on Electrical Information and Communication Technology (EICT), Khulna, Bangladesh, 2021, pp. 1-6, doi: 10.1109/EICT54103.2021.9733590.

[12] Hu, Hailong & Li, Zhong & Elofsson, Arne & Xie, Shangxin. (2019). A Bi-LSTM Based Ensemble Algorithm for Prediction of Protein Secondary Structure. Applied Sciences. 9. 3538. 10.3390/app9173538.

[13] Yang Y, Gao J, Wang J, Heffernan R, Hanson J, Paliwal K, Zhou Y. Sixty-five years of the long march in protein secondary structure prediction: the final stretch? Brief Bioinform. 2018 May 1;19(3):482-494. doi: 10.1093/bib/bbw129. PMID: 28040746; PMCID: PMC5952956.

[14] Dewi Pramudi Ismi, Reza Pulungan, Afiahayati, Deep learning for protein secondary structure prediction: Pre and post-AlphaFold, Computational and Structural Biotechnology Journal, Volume 20, 2022, Pages 6271-6286, ISSN 2001-0370, https://doi.org/10.1016/j.csbj.2022.11.012.

[15] https://github.com/zyxue/pdb-secondary-structure

[16] https://github.com/KirkDCO/pdb-secondary-structure-2022

[17] Avdagic Z, Purisevic E, Omanovic S, Coralic Z. Artificial Intelligence in Prediction of Secondary Protein Structure Using CB513 Database. Summit Transl Bioinform. 2009 Mar 1;2009:1-5. PMID: 21347158; PMCID: PMC3041573.

[18] https://www.kaggle.com/code/helmehelmuto/secondary-structure-prediction-with-keras/notebook

[19] https://medium.com/analytics-vidhya/a-complete-guide-to-adam-and-rmsprop-optimize

[20] Justin Zobel, *Writing for Computer Science*, 3rd ed., Springer, 2014.

[21] https://en.wikibooks.org/wiki/LaTeX

[22] https://en.wikibooks.org/wiki/LaTeX/Algorithms