PROTEIN SECONDARY STRUCTURE PREDICTION USING DEEP LEARNING TECHNIQUES

डीप लर्निंग तकनीक के उपयोग से प्रोटीन की द्वितीयक संरचना का अनुमान

Thesis submitted in partial fulfillment
For the Award of Degree of
DOCTOR OF PHILOSOPHY

by

ASHISH KUMAR SHARMA
आशीष कुमार शर्मा

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY
(BANARAS HINDU UNIVERSITY) VARANASI – 221 005

Roll No: 17071511 **February 2021**

Chapter 7: Conclusion and Future Work

The conclusions of work of this thesis and suggestions for future research are presented in this chapter.

7.1 Conclusions

Protein secondary structure prediction is a sub problem of tertiary structure prediction, and secondary structure is utilized in the several protein folding prediction algorithms. Similarly, protein secondary structure information is used in a several field of bioinformatics, including proteome and gene annotation. The focus of this thesis is to develop some efficient algorithms for protein secondary structure prediction. Protein secondary structure prediction has been an active area of research for several decade and many applications developed using computational methods. The focus of this thesis is to develop some methods for secondary structure prediction from primary sequences.

This thesis contributes towards deep learning and subsequence based representation of protein primary sequences for protein secondary structure prediction. The protein secondary structure prediction improved by extracting the character sequence-based information of primary sequences. Each amino acid residue in a primary sequence is represented as a dense embedding vector. The primary sequences are represented as a summation of residues dense vectors. The recurrent neural network is used to map the complex non-local contextual information residues to predict their secondary

Structure. So I combined the dense embedding vector of residue with a recurrent neural network to predict secondary Structure.

The chapter-wise conclusion of this thesis is also being summarized as follows:

Chapter 1 discussed the motivation, background and problem description for the presented work including thesis scope/objectives, and contributions.

Chapter 2 discussed the theoretical background for Protein Secondary Structure Prediction. This chapter introduced an overview of different variants of feature extraction techniques, and similarities measures. Also, a literature survey of prominent Protein Secondary Structure Prediction approaches.

Chapter 3 demonstrated a simple and effective approach for retrieving contextual information from primary sequences using dense representation. The protein primary sequences are represented as linear combination of twenty amino acid residues. Each amino acid residue represented as dense embedding vector to capture the local contextual information. The whole primary sequence is combination of character embedding vectors. Further recurrent neural network are used to capture the non local contextual information. Further the complex contextual information for secondary structure prediction are given to Softmax layer for predicting the secondary structure classes. The performance of the proposed architecture in terms of accuracy for protein secondary structure prediction is better in comparison to the existing state-of-the-art single sequence-based methods for the chosen three publically available datasets.

Chapter 4 presented a deep learning framework for secondary structure prediction using character n-gram embedding. For preprocessing of primary sequences, n-gram based tokenization performed over sequences to capture frequent combination of amino acid residues. The bag of n-gram tokens are integer encoded based on their occurrence in dataset. Further character n-gram words are represented using dense embedding vector. These dense character n-gram words are summed up to represent the primary sequences. This work was evaluated on publicly available dataset and found advantageous in terms of performance for secondary structure prediction as compared to state-of-art methods.

Chapter 5 Protein primary sequences are segmented as variable length character n-gram words. The n-gram words are selected and integer encoded based on frequency in dataset. The most frequent n-gram words represented as dense embedding vector. The primary sequences are represented as combination of dense embedding vectors of residues. These dense vectors are given input to stacked bidirectional long short term memory network, which produced the non-local contextual information. The combination of variable length segmentation and long short term memory network better predict the protein secondary structure form their primary sequences.

Chapter 6 resolved the issues of protein secondary structure prediction as sequence to sequence translation. The primary sequences are one language and secondary structure element sequences as another language. The amino acid residues are represented as one hot encoding. Two long short term memory network used, one for encoder layer, second for decoder layer. The one hot encoding of primary sequences are input to encoder which keeps the state vectors and discard the output. These state vectors are input to decoder layer which uses the softmax function for secondary structure prediction. The proposed sequence-to-sequence model outperforms single sequence-based methods namely SPIDER3- Single[4] and PSIpred-Single[50].

7.2 Limitations of the proposed works in the thesis:

- Some of the publically available datasets are small in size and due to this performance of the methods are also limited. The performance of the proposed deep learning based approaches can further be improved by using the large datasets.
- The limit imposed on secondary structure prediction is the somewhat arbitrary definition of three states of secondary structures. The secondary structures helices and sheets do not have clear boundaries for classification which affects the prediction accuracy of the proposed methods in some cases.
- The protein sequences are of varying length ranges from 10 to 10K characters. So to cover representation for all sequences is not feasible, and word embedding representation for unknown n-gram words will have no idea how to interpret it as a vector.
- The architecture proposed for variable-length character n-gram based embedding contains a large corpus of n-gram words. In work presented in the thesis, select n-gram words on the basis of frequency and provide better results. However, if optimization based methods are investigated and used it may result in better accuracy to select the n-gram words from a vast corpus.
- The proposed Sequence-To-Sequence modeling architecture for protein secondary structure prediction has a limitation in terms of summarizing a long sequence with a context vector having small dimension.

7.3 Future Work

The research work presented in this thesis can help further into different directions. The scope for future works is as follows:

- Despite the advances in protein secondary structure prediction, there is still significant gap towards contextual information capture for secondary structure prediction. To proceed towards capturing the local as well as non-local contextual information better sequence representation methods can be used[10].

- Reinforcement learning techniques can be applied for protein secondary structure prediction as the numbers of primary sequences are large to further improve the prediction performance.
- The character-based feature extraction methods [101-103] can be explored for finding the contextual information of primary sequences to predict their structure and function.
- The protein secondary structure prediction depends on the local as well as global interactions of amino acids. To further improve the secondary structure prediction, we consider approaches utilizing methods that better capture both local and global interaction [104].
- Predicting 3D structures from the primary sequences may be challenging area that can be explored [105]. Also the precision of secondary structure datasets needs to be improved to avoid significant errors.