ABSTRACT

Protein secondary structure plays a crucial role in determining cellular functions and make significant contributions to biomedicine, including drug development and vaccine design. However, experimental methods to identify these structures, such as X-ray crystallography and NMR spectroscopy, are costly and time-consuming, which slows progress in protein-related research and discoveries. Thus, computational approaches for protein secondary structure prediction have become a highly sought solution. This study aims to develop a model for predicting protein secondary structure into 8 classes scheme (Q8) using the Bidirectional Long Short-Term Memory (Bi-LSTM) architecture, which can capture sequential information from both directions, thereby enhancing prediction accuracy. The model is implemented using the CullPDB dataset for training and validation, and the CB513 dataset for testing. Key hyperparameters, such as optimizer, learning rate, dropout rate, and the number of LSTM layers, are fine-tuned to achieve optimal performance. The model's evaluation metrics include Q8 scheme accuracy and a confusion matrix, along with further analysis of precision, recall, and F1-score for each secondary structure class. The results show that the best hyperparameter combination—Adam optimizer, a learning rate of 1×10^{-3}, a dropout rate of 0.1, and two LSTM layers—achieved a validation accuracy of 91.33% and a test accuracy of 67.28% on the CB513 dataset. In conclusion, the Bi-LSTM model demonstrates strong performance in protein secondary structure prediction, though challenges remain in handling classes with smaller amounts of data.

Keywords : Protein secondary structure prediction, Bi-LSTM, Hyperparameter Tuning, Q8 Scheme