Journal of Mathematics Research; Vol. 16, No. 4; August 2024
ISSN 1916-9795 E-ISSN 1916-9809
Published by Canadian Center of Science and Education

Protein Secondary Structure Prediction Using Convolutional Bidirectional GRU

Yumeng, Lu

Correspondence: School of Mathematics, Yunnan Normal University, Kunming, Yunnan 650500, China

Received: July 2, 2024 Accepted: July 31, 2024 Online Published: August 31, 2024

doi:10.5539/jmr.v16n4p11 URL: https://doi.org/10.5539/jmr.v16n4p11

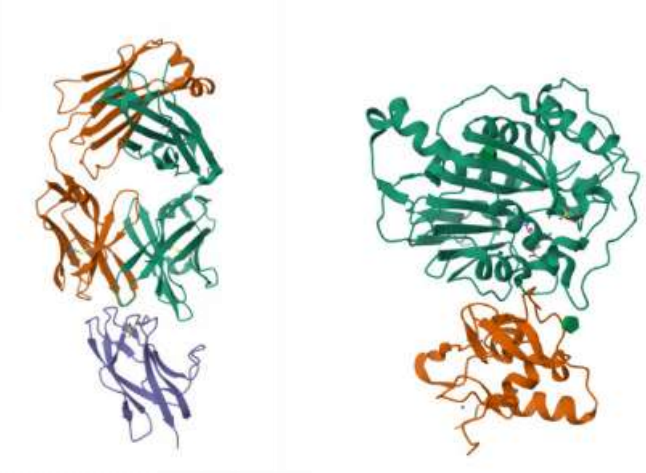
Abstract

In this paper, a protein secondary structure prediction method based on convolutional bidirectional GRU Model (CBi-GRU model) is adopted, which combines the advantages of sliding window in extracting local features of data. The use of CNN and Bi-GRU in the construction of the model improves the feature expression and data utilization, and improves the performance of the model. Protein data from FoxChase Institute were used, and high quality, complete and representative CullPDB dataset, CB513, CASP10 and CASP11 datasets were selected to train, test and validate the model. The results show that the proposed method achieves good prediction performance on CASP10 and CASP11 datasets, and the prediction accuracy of Q8 is 76.2% and 76.4%, respectively. Compared with RaptorX-SS, DeepCNF, CGAN-PSSP and other methods, the Q8 evaluation indicators are improved. Compared with the latest research data, our Q8 prediction accuracy is improved by 2% and 5.1%, which shows the effectiveness and superiority of the proposed model.

Keywords: sliding window, Convolutional bidirectional GRU, Protein Secondary structure prediction

1. Introduction

Deoxyribonucleic acid (DNA) contained in the gene encoding protein sequences, the sequence is present in the nucleus. DNA molecules through transcription and translation form protein sequences, this process is called protein formation. Any change in DNA coding can lead to changes in the genetic traits of the organism, and the results of the changes can be beneficial, detrimental, or neutral. In the process of protein to form any mutation could change in the protein folding, will have negative or positive effect on organisms. Proteins are part of the basis of life activity, is very important to maintain the normal function of the organism and health. In particular, proteins with similar sequences exist in organisms that bend into different structures based on their unique amino acid arrangement. As shown in Figure 1.



(a) (b)

(b)SARS - CoV - 2 the three-dimensional structure of membrane protein 6W4H, (a)Zika membrane protein 6DFI of three-dimensional structure.

Prediction of protein secondary structure is a key area of bioinformatics and molecular biology, to understand the protein three-dimensional structure and biological function provides the basis. Accurate protein secondary structure prediction not only helps to reveal how to perform the biology character (Whisstock & Lesk, 2001), and for the

optimization of the mechanism of disease research, drug design, as well as the evolution process of exploration has extremely important significance. In addition, it plays an indispensable role in protein engineering, prediction before structural biology experiments (Kabsch & Sander, 1983), and the construction of systems biology models. With the continuous progress of computational methods, secondary structure prediction has become an important bridge connecting genomic information with protein function research.

In protein secondary structure prediction, 8 class label system will protein secondary structure subdivided into eight different types. This segment provides than the traditional three or four classification (such as spiral Angle and no rules curly), folding, more detailed structure characteristics. An 8-level label (Smith & Johnson, 2021) system typically includes the following types: Helix (α-helix), π-helix, 310-helix, β-fold, β-turn, turn, random coil, others (such as bending, bridging, etc.). The function of this detailed classification is that it can more accurately describe the secondary structure characteristics of proteins, thus providing more abundant information for the study of 3D structure, function and dynamics of proteins. In addition, the eight-level tag system can enhance the predictive power of models in protein structure prediction, especially for proteins with complex structures or special conformational features. Through the use of 8 class labels, can be a deeper understanding of protein conformation of diversity, and the role of different secondary structure elements in protein function.

At present, the Protein Structure Prediction (Protein Structure Prediction, PSP) the experiment method of the nuclear magnetic resonance spectrum (NMR) (Morelli, Dolla, Czjzek, Palma, Blasco, Krippahl & Guerlesquin, 2000), X-ray crystallography and frozen electron microscopy (sem). 2000 nuclear magnetic resonance (NMR), provides the low resolution of Protein Structure, It is limited to a subset of proteins, and access to deeper information about protein structural changes is limited. X-ray crystallography provides high resolution 3 d shape of the protein, can be applied to small protein and protein, but it doesn't provide data on protein structure change. Frozen electron microscopy (Callaway, 2015) is a new experimental method, 2015 research and development of this technology is in use on the basis of the low resolution of frozen photos on improvement to speed up the research of protein structure. Although these methods have proved to be very effective, they are time-consuming and expensive. Sometimes, it is impossible to find the structural coordinates of proteins using experimental methods, but the efficiency of such high-throughput sequencing technology allows researchers to find a large number of protein sequences (Rost & Sander, 1996). Thus, the use of the known protein sequences and use experiments to determine the number of the gap between protein structure has expanded significantly. The limitations of using experimental techniques to determine protein structure, coupled with the availability of protein sequences, have prompted researchers to develop machine learning algorithms.

Recently, a number of deep learning based protein secondary structure prediction methods have achieved competitive results. Most of the early methods perform well in three-state prediction, but do not perform well in eight-state prediction due to increased complexity. To address this problem, many neural network-based eight-state prediction methods have been explored, including in 2011, Yang et al. (Yang, Wu, Ying & Sui, 2011) proposed an efficient method for protein secondary structure prediction based on two-layer support vector machine (SVM) and position specific scoring matrix (PSSMs). Wang et al. (Wang, Zhao, Peng & Xu, 2011) proposed for 8-class SS prediction using Conditional neural fields (CNF), a recently invented probabilistic graphical model called RaptorX-SS. Zhou and Troyanskaya (Zhou & Troyanskaya,2014) proposed a method based on supervised generation of random networks to predict protein secondary structures with deep hierarchical representation. In 2016, Li and Yu et al.(Li & Yu,2016)introduced the DCRNN model that utilizes convolutional neural networks with different kernel sizes to extract multi-scale local context features. Wang et al. (Wang, Peng, Ma & Xu, 2016) proposed DeepCNF, an integrated model of conditional neural field and shallow neural network, for protein secondary structure prediction. In 2018, Fang et al. (Fang, Shang & Xu, 2018) proposed the MUFOLD-S architecture, which can effectively handle local and global interactions between amino acids to make accurate predictions. Zhou et al. (Zhou, Wang, Zhao, Xu & Lu, 2018) introduced a protein prediction method using multi-scale CNN and the CNNH-PSS model of highway, which was used to transfer information from the current layer to the output of the next layer. Zhang et al. (Zhang, Li & L ü, 2018)used the comprehensive synergy of convolutional neural network, residual network and bidirectional recurrent neural network CRRNN prediction to improve the performance of protein secondary structure prediction. In 2020, GuoZhiye et al. (Guo, Hou & Cheng, 2021) used multiple advanced deep learning architectures (DNSS2) to predict protein secondary structure. Kumar Prince et al. (Kumar, Bankapur & Patil, 2020) proposed a prediction model consisting of 42-dimensional hybrid features combined by Convolutional Neural Network (CNN) and Bidirectional Recurrent Neural Network (BRNN). In 2021, XuYing et al. (Xu & Cheng, 2021) introduced a protein secondary structure prediction model based on multi-scale convolutional attention neural network using multi-channel and multi-scale parallel architecture. In 2022, Jin et al. (Jin, Guo, Jiang, Wu & Yao,2022)introduced the concept of adversarial game into the prediction of secondary structure, and proposed a prediction model based on conditional Generative Adversarial Network (GAN). In 2023, Kim et al. (Kim & Kwon, 2023) used AttSec to extract a self-attention map corresponding to pairwise features between amino acid embeddings and passed it to a convolutional block to capture local patterns for a

novel prediction model based on transformer architecture. In 2024, Abramson et al. (Abramson, Adler & Dunger, et al, 2024) published a paper in Nature in which they introduced AlphaFold 3.0, an AI system developed by Google DeepMind and Isomorphic Labs. It achieves high-precision prediction of the structure and interaction of biomolecules such as proteins, DNA and RNA. Through the upgraded version of Evoformer module and the new Diffusion Network, the accuracy of molecular interaction prediction is significantly improved by at least 50%. It even doubles the accuracy in predicting key interactions between proteins.

This paper adopts a Model based on convolution two-way GRU helped (CBi-GRU helped Model) of protein secondary structure prediction methods, combined with statistical analysis of data processing and the sliding window depending on the local feature extraction, aims to improve the accuracy of protein secondary structure prediction. Use CullPDB protein server publicly available two data set for training and testing of the model, then using CASP10 and CASP11 validation data sets. Eight types of fine granularity were used as the evaluation criteria for secondary structure prediction, and the experimental results were compared with RaptorX-SS, DeepCNF, CGAN-PSSP and other methods. The experimental results show that the proposed method achieves good prediction performance on CASP10 and CASP11 datasets, with prediction accuracy of 76.2% and 76.4%, respectively. By comparing with RaptorX-SS, DeepCNF, CGAN-PSSP and other methods, the Q8 evaluation index has been improved. Compared with the latest research data, the prediction accuracy of Q8 has been improved by 2% and 5.1%, showing the effectiveness and superiority of the proposed model.

2. Data and Material

2.1 Dataset

Three datasets were used in this study:

The first to use CullPDB(Finn,Clements & Eddy ,2015) protein server two publicly available data sets, including CullPDB data has 5926 protein sequences, the CB513 data set has 513 protein sequences, we randomly divide the data set to training and testing the model. Then using CASP10 (Kryshtafovych, Barbato, Fidelis, Monastyrskyy, Schwede & Tramontano,2014) and CASP11(Moult, Fidelis, Kryshtafovych, Schwede & Tramontano,2014) according to the set for validation.

2.2 Input Features

In this study, we represent protein sequences as one-hot encoded forms of their amino acids combined with their position-specific scoring matrix (PSSM)(Altschul , Madden , Schäffer , Zhang , Miller & Lipman, 1997)as an input dataset. Standard amino acids are represented by letters from A to Y, and any nonstandard amino acids are represented by X. This allows us to use a sequence of 21 characters to encode the primary structure of any protein. To enable the sequence information to be effectively learned by the prediction model, we employ a transformation method that converts each amino acid into a 21 one-hot encoding vector containing only two states: 0 or 1, where 1 indicates the type of amino acid and 0 for the remaining positions. The sequence of N amino acids is then converted into an N×21 matrix.

PSSM is an important tool for revealing evolutionary information about biological sequences, capturing evolutionarily important information by recording the conservation patterns of sequences at specific positions. The main function of PSSM is to quantify the amino acid preference at each position in the sequence. These preferences reflect sequence characteristics that have been retained during natural selection and are often closely related to specific secondary structural elements of the protein (e.g., α-helix, β-fold, etc.). By analyzing these scores, key amino acid residues that may affect the structural stability and function of the protein can be identified. PSI-BLAST algorithm is used to generate PSSM. The process includes the following four steps: First, all sequences in the database that are similar to the target sequence are identified; Secondly, the amino acid frequency matrix of each position was constructed based on these sequences. Then, the location probability matrix was further developed. Finally, the final PSSM is generated to provide key information for protein structure prediction.

This paper adopts a specific approach to handle feature representations of protein sequences, ensuring that they can be adapted to our prediction model. First, we utilize Position-specific scoring moments (PSSMS), which have a dimension of N×21, where N represents the length of the sequence and 21 corresponds to the number of types of amino acids. In order to transform the PSSM score values into the [0, 1] interval, we apply the sigmoid function s(x) = 1/e^(-x) for normalization. Considering that most protein sequences do not exceed 700 amino acid residues, we decided to uniformly adjust the one-hot encoding and PSSM of the sequences to a size of 700×21. For sequences with more than 700 amino acids, a chunking strategy was adopted to split them into two parts with some overlapping regions between them. For those sequences less than 700 in length, we padded the length to 700 amino acids by adding zeros to the end.

Finally, the input features received by the prediction model are a 700×42 matrix. In this matrix, the first 21 columns represent the one-hot encoded form of the residue sequence, while the subsequent 21 columns correspond to the PSSM

of each amino acid residue, together constituting a comprehensive feature set that provides rich sequence information and evolutionary context for our model. This method not only ensures the consistent processing of sequences with different lengths, but also enhances the ability of the model to capture the characteristics of the sequence.

2.3 Feature Extraction

After we understand the inputs and outputs, we can see that they are: An 8-state label arrangement of amino acids in a protein is predicted (a tensor of shape [700, 9] or [700, 2]) based on the linear sequence of amino acids in a protein (a tensor of shape [700, 22]), where 700 is the length of the alignment of amino acid sequences for each protein in the dataset. Because the actual length of the amino acid sequence of each protein is not consistent, directly using 700 amino acid sequences as input features to train the model may lead to poor learning ability of fine-grained protein sequence features.

Therefore, we propose a feature construction method based on sliding window (Fengping, Xiaowei & Xiang, 2022). By selecting window size w and sliding frequency s, we can convert each protein from a length of 700 amino acid sequences to ⌈(700-w)/s⌉+1 groups of length w amino acid sequences for finer-grained and smaller-scale feature learning. Thus, the accuracy of the prediction model is improved. It should be noted that when using this method sequence model, in addition to the init function and forward function, we also need to implement a predict function when defining our deep neural network. Given a feature of a protein of the shape [1, 700, 22], the predict function partitions the 700 amino acid sequences into ⌈700/w⌉ sets of length w amino acid sequences, and then concatenates the results of each set. Some amino acids may be predicted multiple times. At this time, the value of the mode or the first prediction is taken, and the secondary prediction structure of a protein is obtained after splicing.

We employed a sliding window to extract features and labels from the original dataset, creating input-output pairs for training and testing. Firstly, the default window size was 17 time steps, and the feature sequence was extracted from the dataset in order and stored in column X, while the label sequence was correspondingly extracted from the dataset and stored in the list Y. After that, these lists are reshaped into arrays of specified dimensions, and finally a function is used to split the dataset into training and test sets for subsequent training and evaluation of machine learning models.

3. Method Building

In this paper, a protein secondary structure prediction method based on convolutional bidirectional GRU Model (CBi-GRU model) is proposed, which combines the advantages of sliding window in local feature extraction to make the model better optimized. In the following, the process of building the model will be introduced, and the results of the test will be illustrated with tables in the experimental section.

3.1 CNN

Convolutional Neural Networks (CNNs)(Zhao&Zhang,2020) is a kind of deep learning model that can automatically extract features from raw data without manually designing feature extractors, which is crucial for recognizing complex features in protein sequences. Parameter sharing reduces the complexity and computation of the model, while end-to-end learning simplifies the mapping process from amino acid sequences directly to secondary structure categories. In addition, the multi-layer structure of CNN is able to capture feature representations ranging from simple to complex, mimicking the working principle of biological visual cortex. With the improvement of computing power, CNN can efficiently process large-scale protein data. Bayesian optimization and other techniques are used to further optimize the network structure and parameters to improve the prediction performance.

A convolution operation usually involves a convolution kernel (or filter) and an input signal. A new output signal is generated by sliding the convolution kernel over the input signal and computing a weighted sum of local regions.

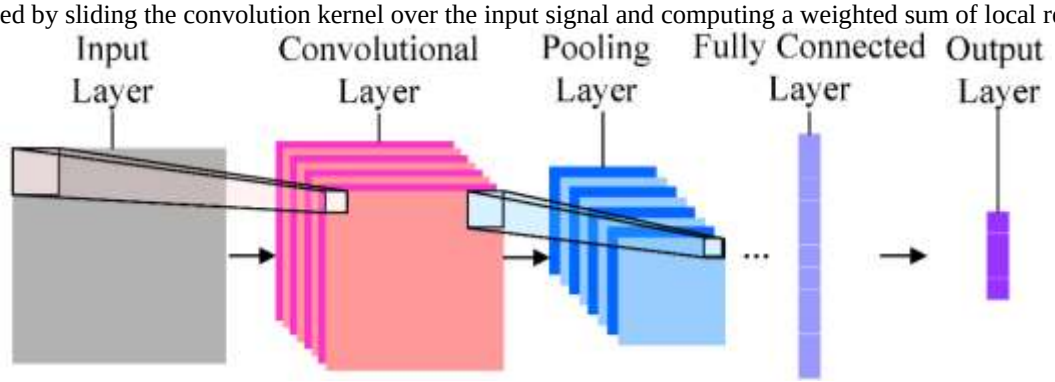


Figure 2. Structure diagram of CNN

3.2 Bi-GRU

Bi-directional Gated Recurrent Unit (Bi-GRU)(Abdelkader, Njimbouom, Oh & Kim,2023).A deep learning neural

network architecture for processing sequential data, which combines two GRU layers, forward and backward, to capture the past and future information at each time point in the sequence for a more comprehensive understanding of time series data. Then, the gating mechanism of update gate and reset gate is used to control the information flow, and the reset gate(Kim,Lee,Moon,Kim,Kim,Jin, Shin, Lee,Jang, Hu & Park,2023)is beneficial to capture the short-term dependencies in the sequence. The update gate helps to replenish long-term dependencies in the sequence. It effectively solves the problem of gradient disappearance and gradient explosion, and can stack multiple layers to form a deep model to learn more complex features. Such models are suitable for sequence-to-sequence mapping tasks and can be trained end-to-end. The forward and backward processing of Bi-GRU can be carried out in parallel, which accelerates the training process. Meanwhile, it also has good generalization ability and a wide range of applications.

Due to the unique bidirectional information processing ability of Bi-directional Gated Recurrent Unit (Bi-GRU), the model can simultaneously capture the past and future context information of each position in the sequence, and more comprehensively understand the complex patterns of amino acid sequences. Its gating mechanism can effectively process sequence data to improve the accuracy of prediction, and due to its end-to-end learning feature, it can directly map from amino acid sequence to secondary structure categories, simplifying the prediction process. In addition, Bi-GRU has good generalization ability and computational efficiency, can adapt to different protein families and types, as well as combine with experimental data to provide more
comprehensive protein structure information.

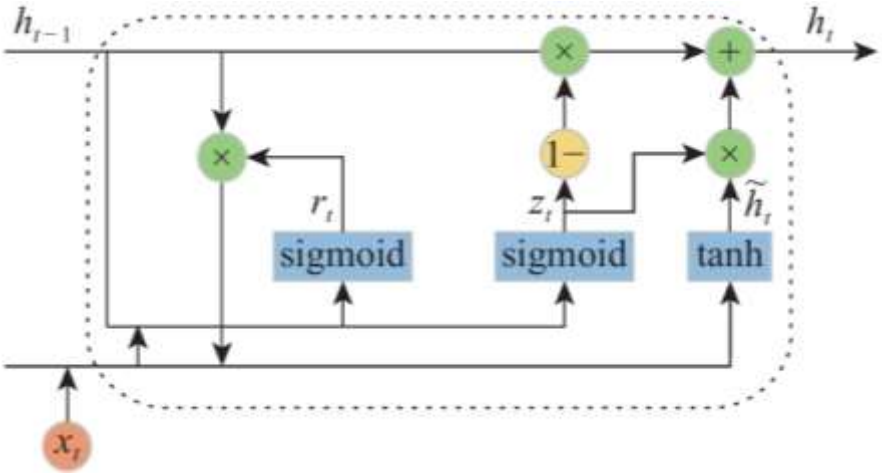


Figure 4. Structure diagram of Bi-GRU

3.3 CBi- GRU

In this paper, CNN and Bi-directional Gated Recurrent Unit (Bi-GRU) are integrated to predict protein secondary structure, which can capture local features and long-distance sequence information at the same time and improve the prediction accuracy. The translation invariance and automatic feature learning properties of convolutional networks, as well as the sensitivity of Bi-GRU to sequence data, make the model not only insensitive to position changes in protein sequences, but also effectively deal with long-term dependencies in sequence data.

The advantage of this ensemble approach is its end-to-end learning capability, which can be directly mapped from amino acid sequences to secondary structure categories, simplifying the prediction process. In addition, the parallel processing ability of the model also significantly improves the computational efficiency. By adjusting the network structure and parameters, the model shows strong adaptability and flexibility, which enhances the accuracy and reliability of prediction. The model structure is shown in fig 4:

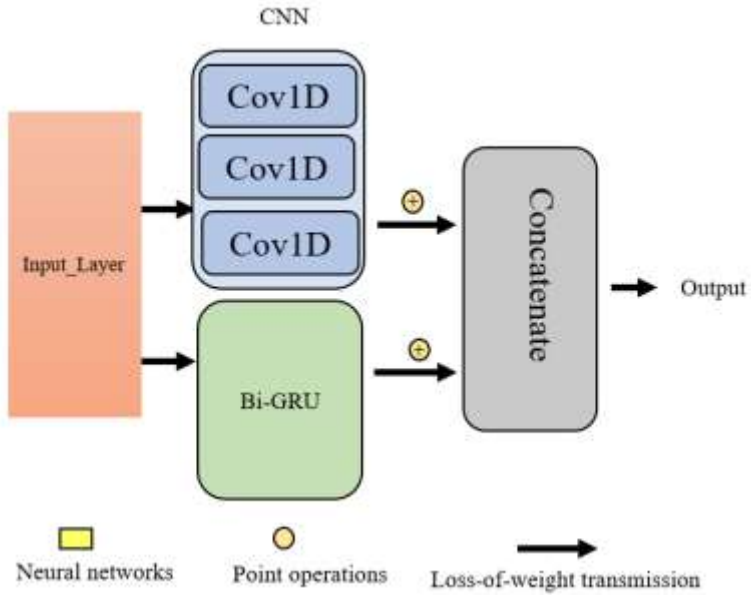


Figure 4. Structure diagram of CBi-GRU

The model first receives data with a 3D shape through the input layer. Three one-dimensional convolutional layers with different kernel sizes (3, 7, 11) are used to extract the local features of the sequence, and the sequence length is kept constant by padding. The outputs of these convolutional layers are merged to capture features at different scales. At the same time, a Bi-directional Gated recurrent Unit (Bi-GRU) layer processes the input sequence to capture long-distance dependencies. Finally, a Softmax activated fully connected output layer predicts the secondary structure class for each amino acid residue. The whole model structure integrates the local perception ability of the convolutional layer and the sequence processing ability of the Bi-GRU layer, aiming to accurately predict the secondary structure of proteins.

4. Experimental Results and Discussion

4.1 Evaluation Metrics

In protein structural biology, the Q8 scoring system covers eight different protein secondary structure categories, including helices (α-helices, 3-10 helices, π-helices), folds (extended strands, β-bridges), turns, bends, and irregular structures, providing a comprehensive measure of prediction accuracy. Compared with the tertiary structure labels, the use of 8-level labels can better capture the complexity and diversity of protein structure and improve the description ability of the prediction model for protein structure.

4.2 Experimental Design

In this paper, a convolutional neural network with bidirectional GRU is constructed, and the Q8 accuracy is 62.2% and 70.4% when using CullPDB and CB513 datasets for training and testing. Compared with the previously constructed single-layer Bi-LSTM double-layer Bi-LSTM neural network model with attention mechanism, the prediction results are improved.

Table 1. The performance of the model is evaluated by performing the above four model cases on the Cullpdb and CB513 datasets, using Q8 as the evaluation metric

Model	Accuracy of Cullpdb Q8 (%)	Accuracy of CB513 Q8 (%)
Single-layer BiLSTM	52. 1	65. 9
Double layer BiLSTM	50. 2	68.7
Bi-GRUCNN	62. 2	70. 4

4.3 Model Testing and Comparison

The CASP10 and CASP11 datasets were used to evaluate the performance of the CBi-GRU Model, and Q8 was used as the evaluation metric. Experimental results show that the proposed method achieves good prediction performance on CASP10 and CASP11 datasets, and the prediction accuracy is 75.2%and 76.4%respectively.

Table 2. Compared with RaptorX-SS, DeepCNF, CGAN-PSSP and other methods, the prediction accuracy of Q8 is improved by 2% and 5.1%respectively, which shows the effectiveness and superiority of the proposed model

Method	CASP10(%)	CASP11(%)
RaptorX-SS	64. 8	65. 1
DeepCNF	71. 8	72. 3
DCRNN	73. 9	71. 2
MUFOLD-SS	74. 2	71. 6
CRRNN	73. 8	71. 6
FIDCNN-SS	74. 9	71. 3
MCNN-PSSP	74. 9	71. 5
CGAN-PSSP	74. 6	71. 3
Our(model)	**76.2**	**76. 4**

5. Conclude

In this paper, a protein secondary structure prediction method based on convolutional bidirectional GRU Model (CBi-GRU model) is proposed, which combines the advantages of data statistical analysis processing and sliding window in local feature extraction, aiming to improve the accuracy of protein secondary structure prediction.

Through sliding Windows, the interaction information of adjacent amino acid residues in the protein sequence is further captured. Then the CBi-GRU Model is used to model the sequences within the window and predict the protein secondary structure. Its end-to-end learning capability, which can directly map from amino acid sequences to secondary structure categories, simplifies the prediction process. By adjusting the network structure and parameters, the model shows strong adaptability and flexibility, which can improve the accuracy of protein secondary structure prediction.

The proposed method achieves good prediction performance on CASP10 and CASP11 datasets, and the prediction accuracy is 76.2% and 76.4%, respectively. Compared with RaptorX-SS, DeepCNF, CGAN-PSSP and other methods, our Q8 prediction accuracy is improved by 2%and 5.1%respectively, demonstrating the effectiveness and superiority of our proposed model.

Acknowledgments

The authors would like to express their sincere gratitude to Professor Liu Haihong for his invaluable guidance and support throughout the research process. His insights and expertise have significantly contributed to the development of this study.

Special thanks go to the Scientific Research Fund project of Education Department of Yunnan Province Foundation for providing the financial support that made this research possible.

We are grateful to the anonymous reviewers for their constructive feedback, which has helped to improve the quality of this paper.

Finally, we extend our heartfelt thanks to our families and friends for their understanding and continuous support during the course of this research.

Authors contributions

In this paper, we propose a method for protein secondary structure prediction based on the Bi-GRU Model, which can extract sequence features at a smaller scale and more fine-grained, and capture long-distance dependencies. The ability to capture long distance dependencies in a sequence can be significantly improved. By adding the output of earlier layers in the network directly to the output of subsequent layers, it helps to solve the problem of gradient disappearance or gradient explosion in deep networks, thereby promoting the efficient flow of gradients in the network.

The feature engineering method of sliding window is used to extract more abundant local sequence information by sliding a fixed size window on the protein sequence.

The Q8 accuracy was used to evaluate the performance of the secondary structure prediction model, and the prediction accuracy was 76.2\% and 76.4\% respectively on the CASP10 and CASP11 data sets. Compared with RaptorX-SS, DeepCNF, CGAN-PSSP and other methods, our Q8 prediction accuracy is improved by 2\% and 5.1\%, demonstrating the effectiveness and superiority of our proposed model.

Funding

This work was supported by the Scientific Research Fund of the Education Department of Yunnan Province [project number 2024Y167].

Competing interests

No competing interest is declared.

Informed consent

Obtained.

Ethics approval

The Publication Ethics Committee of the Canadian Center of Science and Education. The journals policies adhere to the Core Practices established by the Committee on Publication Ethics (COPE).

Provenance and peer review

Not commissioned; externally double-blind peer reviewed.

Data availability statement

The above datasets are publicly available and can be accessed from the relevant websites. CullPDB and CB513 are available at \url{http://www.princeton.edu/~jzthree/datasets/ICML2014/}. CASP10 and CASP11 can be downloaded from \href{http://predictioncenter.org/}{Prediction Center}. The model was trained and tested by randomly partitioning the CullPDB and CB513 datasets. Then CASP10 and CASP11 datasets were used for validation.

Data sharing statement

No additional data are available.

Open access

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/).

References

Whisstock, J. C., & Lesk, A. M. (2003). Prediction of protein function from protein sequence and structure. *Q Review of Biophysics*, *36*(3), 307–340. https://doi.org/10.1017/S0033583503003818

Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, *22*(12), 2577-2637. https://doi.org/10.1002/bip.3602212027

Smith, J. A., & Johnson, E. O. (2021). The application of an 8-class labeling system in protein secondary structure prediction. *Journal of Bioinformatics and Systems Biology*, *18*(4), 523-532.

Morelli, X., Dolla, A., Czjzek, M., Palma, P. N., Blasco, F., Krippahl, L., Moura, J., & Guerlesquin, F. (2000). Heteronuclear NMR and soft docking: an experimental approach for a structural model of the cytochrome c553-ferredoxin complex. *Biochemistry*, *39*(10), 2530-7. https://doi.org/10.1021/bi9926337

Callaway, E. (2015). The revolution will not be crystallized: A new method sweeps through structural biology. *Nature News*, *525*(7568), 172. https://doi.org/10.1038/525172a

Rost, B., & Sander, C. (1996). Bridging the protein sequence-structure gap by structure predictions. *Annual Review of Biophysics and Biomolecular Structure*, *25*, 113-136. https://doi.org/10.1146/annurev.bi.25.110196.005551

Wang, Z., Zhao, F., Peng, J., & Xu, J. (2011). Protein 8-class secondary structure prediction using conditional neural fields. *Proteomics*, *11*(19), 3786–3792. https://doi.org/10.1002/pmic.201100421

Zhou, J., & Troyanskaya, O. (2014). Deep supervised and convolutional generative stochastic network for protein secondary structure prediction. *In Proceedings of the International Conference on Machine Learning*, pp. 745-753.

Li, Z., & Yu, Y. (2016). Protein secondary structure prediction using cascaded convolutional and recurrent neural networks. *In Proceedings of the 25th International Joint Conference on Artificial Intelligence*, pp. 2560–2567.

Wang, S., Peng, J., Ma, J., & Xu, J. (2016). Protein secondary structure prediction using deep convolutional neural fields. *Scientific Reports*, *6*(1), 1-11. https://doi.org/10.1038/srep23173

Fang, C., Shang, Y., & Xu, D. (2018). MUFOLD-SS: New deep inception-inside-inception networks for protein secondary structure prediction. *Proteins: Structure, Function, and Bioinformatics*, *86*(5), 592–598. https://doi.org/10.1002/prot.25488

Zhou, J. Y., Wang, H., Zhao, Z., Xu, R., & Lu, Q. (2018). CNNH_PSS: Protein 8-class secondary structure prediction by convolutional neural network with highway. *BMC Bioinformatics*, *19*(S4), 60. https://doi.org/10.1186/s12859-018-2108-4

Zhang, B., Li, J., & Lü, Q. (2018). Prediction of 8-state protein secondary structures by a novel deep learning architecture. *BMC Bioinformatics*, *19*(1), 293. https://doi.org/10.1186/s12859-018-2306-7

Guo, Z., Hou, J., & Cheng, J. (2021). DNSS2: Improved ab initio protein secondary structure prediction using advanced deep learning architectures. *Proteins: Structure, Function, and Bioinformatics*, *89*(2), 207-217. https://doi.org/10.1002/prot.26107

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

Xu, Y., & Cheng, J. (2021). Secondary structure prediction of protein based on multi-scale convolutional attention neural networks. *Mathematical Biosciences and Engineering*, *18*(4), 3404-3422. https://doi.org/10.3934/mbe.2020408

Jin, X., Guo, L., Jiang, Q., Wu, N., & Yao, S. (2022). Prediction of protein secondary structure based on an improved channel attention and multiscale convolution module. *Frontiers in Bioengineering and Biotechnology*, *10*, 901018. https://doi.org/10.3389/fbioe.2022.901018

Kim, Y., & Kwon, J. (2023). AttSec: Protein secondary structure prediction by capturing local patterns from attention map. *BMC Bioinformatics*, *24*, 183. https://doi.org/10.1186/s12859-023-05141-9

Abramson, J., Adler, J., & Dunger, J., et al. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*.

Finn, R. D., Clements, J., & Eddy, S. R. (2015).HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Research*, *43*, W30-W38. https://doi.org/10.1093/nar/gkv380

Kryshtafovych, A., Barbato, A., Fidelis, K., Monastyrskyy, B., Schwede, T., &Tramontano, A. (2014). Assessment of the assessment: Evaluation of the model quality estimates in CASP10. *Proteins: Structure, Function, and Bioinformatics*, *82*, 112–126. https://doi.org/10.1002/prot.24433

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., & Tramontano, A. (2014). Critical assessment of methods of protein structure prediction (CASP)—round x. *Proteins: Structure, Function, and Bioinformatics*, *82*, 1–6. https://doi.org/10.1002/prot.24432

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402. https://doi.org/10.1093/nar/25.17.3389

Fengping, A. N., Xiaowei, L. I., & Xiang, C. A. O. (2022). Medical Image Classification Algorithm Based on Weight Initialization-Sliding Window CNN. *Journal of Frontiers of Computer Science and Technology*, *16*(8), 1885-1897. https://doi.org/10.11896/JFCS.2022.h2239

Zhao, Y. W., & Zhang, H. (2020). Protein Secondary Structure Prediction Based on Generative Adversarial Networks and Convolutional Neural Networks. *Journal of Computational Biology*, *10*(4), 49-56. https://doi.org/10.1089/cmb.2020.0168

Abdelkader, G. A., Njimbouom, S. N., Oh, T., & Kim, J. (2023). ResBiGAAT: Residual Bi-GRU with attention for protein-ligand binding affinity prediction. *Computational biology and chemistry*, *107*, 107969 .

Kim, H., Lee, J., Moon, S., Kim, S., Kim, T., Jin, S. W., … Park, J. R. (2023). Visual field prediction using a deep bidirectional gated recurrent unit network model. *Scientific reports*, *13*(1), 11154. https://doi.org/10.1038/s41598-023-37360-1