The International Society of Data Scientists
Journal of Data Science and Artificial Intelligence
Vol. 5, No. 1. Online ISSN: 2831-4794

Protein secondary structure prediction based on LSTM neural network approach

Pham Uyen Mai Tram[1], Pham Uyen Mai Anh[1], Pham The Bao[2*]
[1]School of Biotechnology, International University, Vietnam National University HCM City, Vietnam
[2]Information Science Faculty, Sai Gon University, HCM City, Vietnam

*Corresponding author: ptbao@sgu.edu.vn
E-mail: pumtram1106@gmail.com, phamuyenmaianh2003@gmail.com

ABSTRACT

Proteins are crucial for maintaining cellular, organ, and tissue structure and function in the body, predicting protein structure and function from sequence remains challenging. Computational methods are essential for predicting protein properties. In Vietnam, protein technology and bioinformatics have gained attention, especially during the Covid-19 pandemic. This study presents a deep learning approach using a four-layer LSTM model and protein datasets to predict protein secondary structure. The proposed method incorporates RPCA to reduce dimensionality, eliminate errors and outliers, and enhance machine learning model effectiveness. The accuracy of the LSTM model is 88.73%, surpassing modern methods. While limitations and challenges exist, this research contributes to organizing knowledge and building an experimental program for protein function prediction. The proposed method provides a valuable tool for accurately predicting protein function based on secondary structure. Future studies using deep learning, projected protein sequences and structures hold promise for further advancements in protein function prediction.

Keywords: Four-layer LSTM model, RPCA, protein function prediction, deep learning models

1. INTRODUCTION

Proteins are essential macromolecules that play a foundational role in all living organisms and are involved in a wide range of cellular functions, including DNA replication, molecule transport, structural support, and catalyzing chemical reactions. However, processing and analyzing the vast amounts of biological and protein data is a challenging task that requires sophisticated algorithms to handle issues such as redundancy, memory requirements, and computational power consumption. Recent advancements in feature extraction techniques have enabled the prediction of protein structure and biological function.

Several methods have been developed for predicting the secondary structure of proteins, such as Support Vector Machines (SVM), artificial neural networks, Bidirectional Cyclic Neural Networks (BRNN), and probability graph models. Modern protein structure prediction (PSP) systems generally comprise four main components: an input module that generates input features from a protein sequence, including multiple sequence alignment (MSA) of homologous proteins; a "trunk" or artificial neural network that converts input features into spatial information encoding part of the 3D structure; an output module that converts spatial information into the original 3D structure; and a fine-tuning module that improves the original structure and generates all atomic coordinates.

One such method, proposed by Nur Aini and Zuraini Ali in 2020 [7], utilized RPCA (Robust Principal Component Analysis) and SVM to classify structural proteins. This method extracted required features from the input vector, reduced the data dimensions, and tested the SVM with a multiplicative RBF (Radial Basis Function) to achieve high-quality classification of structural proteins. However, this method did not exploit the

association in protein structure and could not remove exceptions or error data, limiting its performance compared to modern artificial neural network models.

In a more recent study, a synthetic model comprising five sub-models, each describing protein sequences according to individual amino acid characteristics, homologous proteins, and math coding, was developed. These sub-models were combined using a layer of Bi-directional LSTM (long short-term memory) to predict the structure of protein sequences from three protein datasets: data1199, CB513, and CASP203. This model achieved an accuracy of 84.3% in Q3 accuracy and 81.9% in the allowed decimal fraction (SOV) using 10-fold cross-validation [8].

Predicting the 3D structural patterns of a protein is a significant computational challenge, but understanding the function of a protein at the molecular level requires knowledge about its structure. In this study, we propose a method that uses RPCA for data preprocessing and extracts the main properties of protein sequences. To exploit the linking between molecules in protein sequences, a deep learning model containing four layers of LSTM is utilized to predict the secondary structure of proteins. This model serves as a foundation for making future predictions of the biological function of proteins by increasing the exploitation of the binding features between elements in protein chains.

2. METHODOLOGY

We propose a method using RPCA to preprocess data and extract key features of protein sequences, with RPCA providing fast and efficient key feature extraction. RPCA is an effective method for feature extraction because it has better extraction quality than PCA and performs better in training complex datasets with sample data that are flawed, redundant, and have high dimensions. PCA is a good and widely used method for extracting local object features because it emphasizes variation and reveals key patterns in the dataset so that the data is easily mined and visualized. However, PCA cannot detect and ignore outliers of the sample dataset, thus greatly affecting the system's efficiency [10, 11, 12].

Then, to enhance the ability to exploit the linkages between elements in the protein sequence (this is time-series data), we designed a deep learning model containing 4 layers of LSTM networks to predict the secondary structure of proteins and as a foundation for future protein biological function prediction. LSTM neural network architecture is one of the effective solutions for processing secondary protein structure data; in which information from the same sample is linked and interacts with each other to form a complete protein structure. We focus on predicting eight types of secondary protein structures based on their sequence arrangement by constructing a one-dimensional LSTM neural network architecture [1, 2, 3-8].

The data were encoded as shown in Table 1 to normalize the DNA sequences of the dataset. The main processing steps of the study are as follows:

Step 1 - Data Preprocessing: Clean the data to remove unwanted data, missing values, rows and columns, duplicate values, data type conversion, etc. Separate the cleaned data into two sets - a training set and a test set.

Step 2 - Extract the features needed for training with RPCA. To preprocess the protein sequence data, we utilize RPCA to extract the key features. Initially, synthetic data with low ranking is generated, where N determines the size of the validation matrix, num_groups specifies the number of groups of values stored in each row, num_values_per_group determines the number of values in each group, *p_missing* is the number of outliers in the data, and *ds* is the list of elements in the group. Algorithm 1 is used to build up synthetic data that has low rank for protein sequence analysis. The algorithm takes as input the size of the validation matrix (N), the number of groups of values stored in each row (num_groups), the number of values in each group (num_values_per_group), the number of error data points (*p_missing*), and the list of elements in the group (*ds*). The output is a list of initialized protein sequences (D).

We developed algorithm 2 to remove some of the noise from the data. This study uses RPCA to estimate

attenuated data. RPCA estimates data by summing two matrices, L and S, where L is the low-rank component and S is the sparse component. In our context, matrix L contains the signal, while matrix S represents noise or the sparse component. However, in some cases, matrix S can be ignored and only matrix L can be used to evaluate the data after noise removal.

Algorithm 3 initializes the RPCA function to identify the main components of the sample data matrix M, which consists of protein sequences. The algorithm iteratively computes the low rank matrix L and the sparse data matrix S until convergence. The algorithm returns the low rank matrix L and the sparse data matrix S, which represent the main components and the noise of the protein sequence data, respectively.

The RPCA method is applied to the input matrix D' to obtain the low-rank matrix L and the sparse matrix S, which is shown in Algorithm 4. The RPCA function is initialized on the matrix D', and then the L and S values are computed via RPCA. The output low-rank matrix L retains the main components of the protein sequence dataset, while the sparse matrix S contains the noise and outlier values.

Finally, we obtain matrix L from the initial protein sequence dataset, which retains the important data dimensions and effectively reduces the dimensions from the data matrix D. This approach is particularly useful in the case of exceptional samples or incorrect inputs, allowing the input data for the LSTM deep learning model to become 'cleaner' and increasing the efficiency of model training. Once the main components of the model dataset have been identified, we use the LSTM deep learning model to study the collected model data and build a set of weights for the protein structure prediction model on the evaluation set.

Step 3 - Train the LSTM model with the data from step 2. The complete structure of the LSTM model will be used for training and testing on the sample dataset. (1) The first step is to pass the data as input to the deep learning model; (2) Then, the first layer includes two Conv-1D steps and resize, where Conv-1D is used to make the input data into a 1-dimensional input, then the second layer includes Conv-2D for 2-dimensional input and an array resize layer, then there is a Dropout layer to prevent the model from overfitting the sample data and the layers are fully connected with categorical-sparse, then there are the main layers of our model which include 4 interconnected LSTM layers, as shown in Figure 1.

Algorithm 1: Building up the synthetic data that has low ranking

Input: N, num_groups, num_values_per_group, $p_missing$, ds
Output: D: List of initialized protein sequences

1. For k from 0 to num_groups do:
2. $d = \text{numpy.ones}((N, \text{num_values_per_group})) * (k+1) * 10$
3. $ds.\text{append}(d)$
4. $D = \text{numpy.hstack}(ds)$

Algorithm 2: Noise Removal

Input: D: List of protein sequences
Output: D: List of protein sequences after removing noise.

1. # Get the size of matrix data build from the first step

   $n1, n2 = D.\text{shape}$

2. # Randomly take 1 of the 2 dimensions, then assign of the data in the D matrix to 0

   $S = \text{numpy.random.rand}(n1, n2)$

$D[S < k] = 0$

Algorithm 3: Initializing the RPCA function to find the main component

Input: M: Sample Data Matrix
Output: L: Low Rank Matrix; S: Sparse data matrix

1. **Initialization:** $S_0 = Y_0 = 0$, $\mu > 0$
2. **While** not converged **do:**

   Calculate $L_{k+1} = \mathcal{D}_{\mu}(M - S_k - \mu^{-1}Y_k)$

   Calculate $S_{k+1} = \mathcal{S}_{\lambda\mu}(M - L_{k+1} + \mu^{-1}Y_k)$

   Calculate $Y_{k+1} = Y_k + \mu(M - L_{k+1} - S_{k+1})$

   end while

3. **return** L, S

Algorithm 4: Applying RPCA to calculate the decreasing data

Input: D': List of protein sequences
Output: L: Low Rank Matrix; S: Sparse data matrix

1. # Initialize the RPCA function on the matrix D

   $rpca = \text{R_PCA}(D)$

2. # Calculating L, S values via RPCA

   $L, S = rpca.\text{fit}(\text{max_iter} = 10000, \text{iter_print} = 100)$

3. RESULTS AND DISCUSSION

3.1 Empirical datasets

To evaluate the accuracy of protein secondary structure prediction methods, a benchmark dataset of 500 diverse protein sequences can be employed, with none of them sharing sequence similarity to any known 3D structures. Numerous prediction models have been developed based on a set of 126 protein sequences used by Rost and Sanders [13] or on a more recent collection of 396 non-redundant protein fragments curated by Cuff JA and Barton GJ [14] using proteins from the RS126 protein sequence. It is worth noting that the CB513 dataset has also been extensively studied and employed as a benchmark in various scholarly publications.

Training and evaluation datasets

In this study, we used the CullPdB and CB513 datasets, both of which contain secondary structure sequence and protein sequence information [15]. To enhance the effectiveness of training the LSTM model for predicting protein secondary structure, we employed RPCA to preprocess the data. The protein data from these datasets are stored as a matrix of N (protein) $\times$ k (feature), so we needed to reshape the size to N (protein) $\times$ 700 (amino acids) $\times$ 57 (features). The dataset has 57 features, with the following segment meanings:

- $[0,22]$: amino acid radicals, in the order 'A', 'C', 'E', 'D', 'G', 'F', 'I', 'H', 'K', 'M', 'L', 'N', 'Q', 'P', 'S', 'R', 'T', 'W', 'V', 'Y', 'X', 'NoSeq'.
- $[22,31]$: substructure labels, with sequences being 'L', 'B', 'E', 'G', 'I', 'H', 'S', 'T', 'NoSeq'.
- $[31,33]$: N- and C-terminus indicators.
- $[33,35]$: evidence of both relative and absolute solvent accessibility; exclusively used in training (absolute accessibility is thresholded at 15; relative accessibility is normalized by the largest accessibility value in a protein and is thresholded at 0.15; solvent accessibility is initially calculated by DSSP).
- $[35,57]$: protein sequence annotations. Note that the order of amino acid residues is ACDEFGHIKLMNPQRSTVWXY, which is different from the order for amino acid residues.

The CullPdB_6133 training dataset has been divided into three parts:

- $[0,5600)$: used for training
- $[5605,5877)$: used for validation
- $[5877,6133)$: used for testing

All proteins in this dataset can be used for testing on the CB513 dataset.

Figure 1: LSTM model architecture.

Test Datasets

The CB513 dataset consists of 513 files with the extension *.all. These files contain information about protein sequences with no sequence similarity between protein sequences, divided into two sets:

- 413 training protein sequences
- 100 experimental protein sequences.

The data is formatted as paired character strings, where the primary amino acid sequence is directly aligned with its corresponding secondary structure labels. A representative sample is shown below:

```
MVLSEGEWQLVLHVWAKVEADVAGHGQDILI...
CCCCHHHHHHHHHHHHHHGGGHHHHHHHHHH...
MNIFEMLRIDEGLRLKIYKDTEGYYTIGIGH...
CCHHHHHHHHHCCEEEEEECTTSCEEEETTE...
```

To facilitate computational processing, these alphabetical sequences must be converted into numerical formats. Tables 1 and 2 detail the encoding criteria and representation rules used to translate the raw amino acid characters and structural classes within the CB513 dataset into machine-readable inputs.

3.2 Experimental environment

In consideration of the large size of the dataset and the complexity of the LSTM model, potential out-of-memory (OOM) errors were mitigated by implementing an experimental program using the Colab-Pro environment. This program offers high RAM capacity and a GPU with significant memory to support the training requirements of the model. Figure 2 illustrates the specifications of the utilized hardware and software components.

Figure 2: Specifications of experimented GPU.

Table 1: Representation criteria for the CB513 dataset.

Amin acids	Amino acids	Alphabet	Code
Alanine	Alanine	A	01
Cysteine	Cysteine	C	02
Aspartate	Asparaginska kiselina	D	03
Glutamate	Glutaminska kiselina	E	04
Phenylalanine	Fenilanin	F	05
Glycine	Glicin	G	06
Histidine	Histidin	H	07
Isolecine	Izoleucin	I	08
Lysine	Lizin	K	09
Leucine	Leucin	L	10
Methionine	Metionin	M	11
Asparagine	Asparagin	N	12
Proline	Proline	P	13
Glutamine	Glutamin	Q	14
Arginine	Arginin	R	15
Serine	Serin	S	16
Threonine	Treonin	T	17
Valine	Valin	V	18
Tryptophan	Triptofan	W	19
Threonine (Tyrosine)	Tirozin	Y	20

Table 2: Code for Secondary Protein Structure in CB513.

Class of Protein Structure	Secondary Protein Structure	Code
H, G	Helix	01
E	Strand	02
B, I, S, T, C, L	Others	03

3.3 Results

3.3.1 Reliability

The LSTM model was optimized using the Adam optimization method, a stochastic gradient descent approach based on adaptive moment estimation of first- and second-order moments, for fifty epochs on the entire dataset. The model's performance was evaluated using accuracy metrics, where a higher accuracy value indicates better performance.

To test the trained model on the test dataset, the CB513 dataset was used. The model achieved an error value of 0.31 and an accuracy value of 88.73%, which is an excellent performance for an LSTM model. The trained model was then used to predict the secondary structure from protein sequences.

3.3.2 Comparison

Zhang B et al. utilized the TR12148 dataset, which merges the TR5534 and TR6614 canonical datasets and contains 12,148 polypeptide chains, in their 2018 study [15]. They trained several deep learning models using this dataset, such as Conditional SCGSN, DCRNN, and CNN, and evaluated the performance of their models on the CB513 dataset. Table 3 shows the results of their experiments, with their CRRNN model achieving a maximum accuracy of 86% for predicting the substructure value "H.".

In 2021, Alexander Rives et al. [9] used five publicly available datasets, including CullPDB53, CB513, CASP1054, CASP1155, and CAMEO, to evaluate the performance of their deep learning models. DeepCNF, which achieved an accuracy rate of 72%, is one of their notable models [16].

Compared to the prior research, the proposed model achieved an accuracy of 88.73% on the CB513 dataset, which is superior. However, both previous works employed deep learning models, and their best performance on the CB513 dataset was approximately 86%. Table 3 illustrate our results with other approaches.

Table 3: Accuracy comparison with some other models on CB513 dataset.

Method	**Accuracy (%)**
CRRNN	86.0
DeepCNF	72.0
Transformer	74.3
Recommend method	**88.73**

4. DISCUSSION

Through our research and experimentation, we have found that our proposed method has contributed to the systematization of relevant knowledge and the development of an experimental program to solve the problem. Our method has several notable advantages, including the use of RPCA technique to reduce the dimensionality of protein data, thereby eliminating error and outlier data samples due to data construction and collection. This has greatly improved the effectiveness of machine learning models in predicting protein structure, which is a crucial aspect of predicting protein function.

Our model with four layers of LSTM works efficiently on various parameters, such as learning rate, input port bias, and output port bias. We have found that a high learning rate drives the output gates asymptotically towards zero, thereby limiting the negative consequences of model training. Additionally, the accuracy of our proposed method in predicting protein structure is higher than that of many modern methods.

However, we acknowledge that our experimental program and proposed method still have some limitations and challenges. For instance, the LSTM model is prone to overloading, and dropping "learning" is more challenging to implement in LSTM network architectures. Additionally, the LSTM is sensitive to different random weight initializations. These challenges provide avenues for future research to address and improve upon our proposed method.

5. CONCLUSION

This article presents a method for predicting the secondary structure of protein sequences using an LSTM model and protein datasets. The secondary structure of a protein is important for its formation, and it can be used to predict the tertiary structure, which in turn is necessary to predict the protein's biological function. Secondary

structure prediction is comparatively accurate and simpler to solve than the three-dimensional structure prediction. With the most reliable techniques, the accuracy of assigning a strand, helix, or loop to a particular residue can reach 80%. The proposed LSTM model has an accuracy of 88.73% on the CB513 test dataset.

In the future, students and researchers in bioinformatics will continue to use deep learning models and projected protein sequences and structures to predict protein function. Protein function prediction methods are used by bioinformatics researchers to assign biological or biochemical functions to proteins based on genomic sequence data. These proteins are often poorly understood or unknown.

Declaration of Competing Interest
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

The data availability statement
The datasets used in this study are publicly available benchmark datasets, including data1199, CB513, and CASP203. These datasets can be accessed from their respective publicly available repositories. No new datasets were generated during this study.

The ethics statement
This study did not involve human participants, human tissues, animals, or any identifiable personal data. Therefore, ethical approval and informed consent were not required.

The conflict-of-interest statement
The authors declare that they have no competing interests. There are no financial or non-financial conflicts that could have influenced the outcomes of this research.

REFERENCES

[1] M. Torrisi, G. Pollastri, and Q. Le, "Deep learning methods in protein structure prediction," *Computational and Structural Biotechnology Journal*, 2020.

[2] V. Gligorijević, P. D. Renfrew, and T. Kosciolek, "Structure-based protein function prediction using graph convolutional networks," *Nature Communications*, 2021.

[3] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers: Original Research on Biomolecules*, 1983.

[4] M. AlQuraishi, "Machine learning in protein structure prediction," *Current Opinion in Chemical Biology*, 2021.

[5] P. Y. Chou and G. D. Fasman, "Prediction of protein conformation," *Biochemistry*, vol. 13, no. 2, 1974.

[6] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, Sep. 1997.

[7] N. A. Zakaria, Z. Ali Shah, and S. Kasim, "Protein Structure Prediction Using Robust Principal Component Analysis and Support Vector Machine," *International Journal of Data Science*, 2020.

[8] H. Hu, Z. Li, A. Elofsson, and S. Xie, "A Bi-LSTM Based Ensemble Algorithm for Prediction of Protein Secondary Structure," *MDPI*, 2019.

[9] A. Rives *et al.*, "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences," *Proceedings of the National Academy of Sciences*, vol. 118, no. 15, p. e2016239118, 2021.

[10] D. Sitani, A. Giorgetti, M. Alfonso-Prieto, and P. Carloni, "Robust principal component analysis-based prediction of protein-protein interaction hot spots," *Proteins: Structure, Function, and Bioinformatics*, vol. 89, no. 6, pp. 639–647, Feb. 2021.

[11] M. Clementi, S. Clementi, G. Cruciani, M. Pastor, A. M. Davis, and D. R. Flower, "Robust multivariate statistics and the prediction of protein secondary structure content," *Protein Engineering Design and Selection*, vol. 10, no. 7, pp. 747–749, Jul. 1997.

[12] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM (JACM)*, vol. 58, no. 3, pp. 1–37, 2011.

[13] B. Rost and C. Sander, "Prediction of Protein Secondary Structure at Better than 70% Accuracy," *Journal of Molecular Biology*, vol. 232, no. 2, pp. 584–599, 1993.

[14] J. A. Cuff and G. J. Barton, "Evaluation and improvement of multiple sequence methods for protein secondary structure prediction," *Proteins: Structure, Function, and Genetics*, vol. 34, no. 4, pp. 508–519, 1999.

[15] B. Zhang, J. Li, and Q. Lü, "Prediction of 8-state protein secondary structures by a novel deep learning architecture," *BMC Bioinformatics*, vol. 19, no. 1, 2018.

[16] S. Wang, S. Weng, J. Ma, and Q. Tang, "DeepCNF-D: Predicting Protein Order/Disorder Regions by Weighted Deep Convolutional Neural Fields," *International Journal of Molecular Sciences*, vol. 16, no. 8, pp. 17315–17330, 2015.