Briefings in Bioinformatics, 2023, **24(1)**, 1–11

https://doi.org/10.1093/bib/bbac493

Problem Solving Protocol

Grain protein function prediction based on self-attention mechanism and bidirectional LSTM

Jing Liu, Xinghua Tang and Xiao Guan

Corresponding author. Xiao Guan, School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China. E-mail: gnxo@163.com

Abstract

With the development of genome sequencing technology, using computing technology to predict grain protein function has become one of the important tasks of bioinformatics. The protein data of four grains, soybean, maize, indica and japonica are selected in this experimental dataset. In this paper, a novel neural network algorithm Chemical-SA-BiLSTM is proposed for grain protein function prediction. The Chemical-SA-BiLSTM algorithm fuses the chemical properties of proteins on the basis of amino acid sequences, and combines the self-attention mechanism with the bidirectional Long Short-Term Memory network. The experimental results show that the Chemical-SA-BiLSTM algorithm is superior to other classical neural network algorithms, and can more accurately predict the protein function, which proves the effectiveness of the Chemical-SA-BiLSTM algorithm in the prediction of grain protein function. The source code of our method is available at https://github.com/HwaTong/Chemical-SA-BiLSTM.

Keywords: grain, protein function prediction, deep learning, self-attention, bidirectional long short-term memory, chemical property.

Introduction

Grain is essential for human survival, supplying more than half of the calories consumed by humans [1]. As a food additive, it plays an irreplaceable role in the food industry. At the same time, grain protein content reaches 8–12%, which is the main source of protein in human diet. Protein is considered a key factor determining nutritional quality [2]. Therefore, the study of grain protein is of great significance to the development of human daily life, food agriculture and food industry.

With the advancement of sequencing technology, the early method of protein function prediction through biological experiments consumes a lot of manpower, material resources, time and funds. It can no longer adapt to the increasing growth rate of grain protein sequence data [3]. Therefore, the computational method has become one of the mainstream methods for protein function prediction [4].

The early computational methods used BLAST, PSI-BLAST, FASTA and other software to search for similar sequences of each protein in the training set, and then assumed that similar sequences had similar functions [5], and migrated protein function annotations. With the development of artificial intelligence, many machine learning methods are widely used to predict protein function. SVM-Prot [6] utilized protein composition and transformation, distribution features and SVM algorithm for protein function prediction. ProMK [7] combined the KNN algorithm with five different methods of measuring distances between characteristic values to predict protein function on different datasets. Many other researchers used different machine learning methods to predict protein function and achieved good results, such as co-learning [8], Naive Bayes model [9, 10], Random Forest [11] and so on.

However, shallow protein function prediction methods are often difficult to mine deep (nonlinear) relationships between proteins and Gene Ontology (GO) functional terms. Compared with traditional machine learning methods, the deep learning method can learn from massive protein sequence data without feature engineering. As long as the amino acid sequence data are simply processed, it can be directly input into the neural network for learning. The deep learning method solves the problems that are difficult to be solved by traditional machine learning algorithms in the past such as high dimension, redundancy and high noise caused by massive protein sequence data. DeepGO [12] as one of first deep learning models used the convolutional neural network (CNN) algorithm for protein function prediction using different datasets. It was an algorithm for predicting protein functions from protein sequences and PPI networks. Based on the DeepGO algorithm, DeepGOPlus [13] was developed for predicting protein function from amino acid sequences alone which combined CNN model with similarity-based method BLAST. It combined neural network predictions with methods based on sequence similarity to capture interaction information. In ProtConv [14], the CNN algorithm was presented and trained for protein function prediction task. It converted the vector representation of the protein or peptide sequence into two-dimensional image with single channel which is fed into the CNN. In Deep_CNN_LSTM_GO [15], the CNN algorithm combined with

Jing Liu is an associate professor of the College of Information Engineering, Shanghai Maritime University. Her research interests include artificial intelligence algorithms, bioinformatics and machine learning.

Xinghua Tang is a master student at the College of Information Engineering, Shanghai Maritime University. Her research interests include machine learning and computational biology.

Xiao Guan is a Professor of the School of Health Science and Engineering, University of Shanghai for Science and Technology. His research interests include protein function prediction, quantitative structure–activity relationship and optimization of food processing.

Received: June 9, 2022. **Revised:** October 13, 2022. **Accepted:** October 18, 2022

the Long Short-Term Memory (LSTM) algorithm was proposed for predicting protein function. It can be trained on any standard CPU without the need for a dedicated GPU.

Although the models proposed above achieve relatively good prediction results in solving protein function prediction task, there are still some problems. On the one hand, the network structure cannot effectively capture the long-term dependency between the same protein sequence and cannot fully extract the amino acids sequence information. Long-term dependence refers to the long-distance dependence relationship between each amino acids in a protein sequence. By establishing this relationship, the overall information of the sequence can be better learned. On the other hand, it is difficult to effectively distinguish the valid information and invalid information of the protein sequence. It is difficult to capture the amino acid sequence that has a greater effect on the protein function. Valid information refers to protein sequence information that has a great impact on protein function. Correspondingly, invalid information refers to protein sequence information that has less impact on protein function.

Based on the thinking of the above problems, it is of great significance to develop a new prediction method to solve the problem of protein function prediction. Firstly, for the sequence information of amino acids that cannot be fully extracted, the bidirectional Long Short-Term Memory network (BiLSTM) [16] is used to extract the global and local feature information of proteins. At the same time, the sequence relationship between the feature information can be effectively preserved, so that the model can obtain better prediction effect. Secondly, in order to better utilize the sequence relationship between feature information and reflect the importance of different sequence positions, the self-attention mechanism [17, 18] is used in this experiment to make the model pay more attention to the important features in the sequence, thus enhancing the robustness and generality of the protein function prediction model. Additionally, Corral [19] used a variety of machine learning methods to explore the mapping relationship between protein key residues and functions, and found that modeling with chemical properties can achieve higher accuracy. Therefore, in this experiment, six chemical properties are added to the amino acid sequence after data processing as the input of the model to enrich the information of the input data. To a certain extent, the input data can provide more useful information for protein function prediction model.

In summary, grain protein is used as research object in this paper. The Chemical-SA-BiLSTM algorithm is proposed for the task of grain protein function prediction. The Chemical-SA-BiLSTM algorithm adds chemical properties to the original amino acid sequence as the input data of the model, and combines the self-attention mechanism and the BiLSTM algorithm. The experimental results show that, whether compared with the classical neural network algorithm (the CNN algorithm and the LSTM algorithm) and the combined version of the CNN algorithm and the BiLSTM algorithm (the CNN-BiLSTM algorithm), or compared with the classical neural network algorithm combining chemical properties (the Chemical-CNN algorithm), the Chemical-SA-BiLSTM algorithm proposed in this paper can achieve better prediction results in the grain protein function prediction. It is proved that the Chemical-SA-BiLSTM algorithm proposed in this paper can fully extract the amino acid sequence information and effectively use the sequence relationship between the feature information, which has high effectiveness and robustness.

Table 1. The example of one-hot encoding for input sequence

C	G	Q	C	Y	Q	I	A	C	A	...
0	0	0	0	0	0	0	1	0	1	...
1	0	0	1	0	0	0	0	1	0	...
0	0	0	0	0	0	0	0	0	0	...
...	...	...	...	...	...	...	...	...	...	...
0	0	0	0	1	0	0	0	0	0	...

Materials and Methods

Data representation

To use neural network for protein function prediction based on amino acid sequences, the first task is to find the best way to represent the input data so that the protein sequences can be recognized by the program. The current popular encoding methods include one-hot encoding, learned embeddings and BLOSUM62 embeddings. In general, the learned embedding method has a large number of model parameters. Compared with the learning embedding method, the one-hot coding method can not only reduce the number of model parameters, but also avoid the overfitting problem [13]. In addition, the BLOSUM62 embedding is one of the popular encoding methods. It represents each amino acid by its corresponding row in the BLOSUM62 matrix. Instead of treating all amino acids independently, the BLOSUM62 matrix keeps the evolutionary information about which pairs of amino acids are easily interchangeable during evolution [20]. A study showed that the one-hot encoding method achieved lower model error compared with BLOSUM62 embedding [21]. Therefore, the n-gram of sequence of amino acid codes is encoded by the one-hot encoding method in this study. This method maps each amino acid letter to a specific real number from 1 to 20. And then, each term of the n-gram is assigned a vector consisting of all zeros, except for a one at the position reserved for that term. For example, the real number corresponding to the letter D is 3, which means that the third position of its vector is assigned one, and the remaining positions are zero. It is worth noting that the lengths of protein sequences are mostly unequal and vary greatly. In order to unify the format of the input data and reduce the calculation time of the model, each protein sequence length is unified to 1002 in this experiment. Despite the restriction that the sequence length is 1002 and does not contain ambiguous amino acid codes, about 90% of protein sequences in UniProt satisfy these conditions [13]. In other words, protein sequences with the length greater than 1002 are filtered out. If the protein sequences with the initial length less than 1002, they are padded with zeros on the left until the sequence length is 1002. Finally, all protein sequences with ambiguous amino acid codes (B, J, O, U, X, Z) are deleted. The example of one-hot encoding for input sequence is shown in Table 1.

The protein function depends on the chemical properties of its amino acids. Adding chemical properties to the sequence data can make the input data more informative and useful in protein function prediction task to a certain extent. The six chemical properties of amino acid [22] added in this experiment are shown in Table 2. In general, each amino acid sequence is encoded as a 26-dimensional vector. The first 20 dimensions of the vector represent the original amino acid sequence, and the other six dimensions are six amino acid chemical properties newly added.

Table 2. Six amino acid chemical properties

Chemical properties	value
Charge	positive:1; negative:-1; neutral:0.1
Hydrophobicity	from -4.5 to 4.5
isPolar	yes:1; no:0
isAromatic	yes:1; no:0
hasHydroxyl	yes:1; no:0
hasSulfur	yes:1; no:0

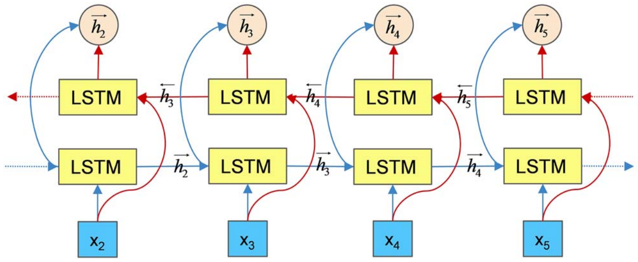

Figure 1. BiLSTM network architecture.

BiLSTM and self-attention mechanism

Hochreiter [23, 24] proposed the LSTM network, which contains three gates (input gate, forget gate and output gate) and a memory cell. The LSTM network can effectively retain the historical information of the input data and learn the dependency information of long sequences. However, in the LSTM network, the cell state at each time can only store the current time and the previous time information. Its single direction transmission mode makes it impossible for the algorithm to obtain information from the back to the front. For the purpose of solving this problem, it is a good choice to build the model with the BiLSTM algorithm [16]. The BiLSTM algorithm can capture important information of amino acid sequences bidirectionally, fully consider the contextual correlation information of the current amino acid sequences, and can learn protein sequences feature more deeply. The BiLSTM network architecture is shown in Figure 1.

Due to the long amino acid sequence, the BiLSTM model is unable to capture the most direct relationship between the feature vector and the result label. Adding self-attention mechanism [17, 18] to the model can solve this kind of problem. It can weigh the input features and measure the importance of each feature to the experimental object. The self-attention mechanism is widely used in the fields of text and image classification [25, 26], machine translation [18, 27] and bioinformatics [28, 29]. In this experimental model, the computational relationship within the self-attention mechanism is shown in Figure 2.

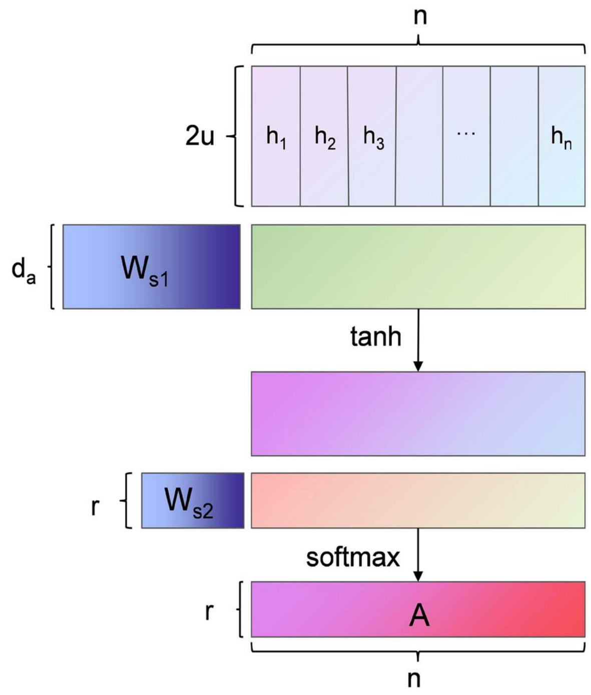

Figure 2. The computational relationship within the self-attention mechanism.

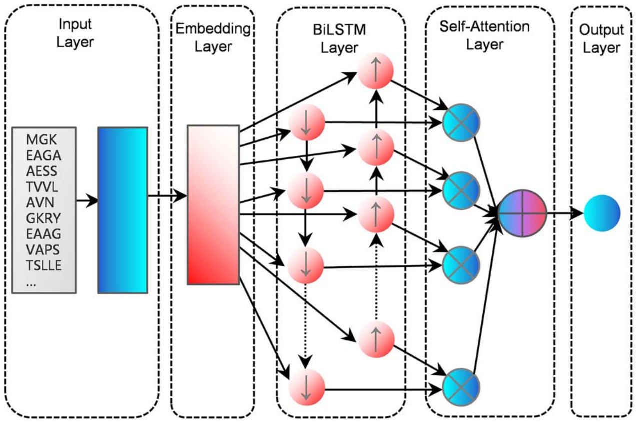

Figure 3. Chemical-SA-BiLSTM Architecture.

Chemical-SA-BiLSTM

In order to enrich the information of the input data, six chemical properties are added to the amino acid sequence after data processing. The final amino acid sequence and GO annotations are jointly used as the input of the neural network model for model training. Sequence features are extracted based on BiLSTM network in this experiment. But the BiLSTM network needs to perform calculations in sequence of amino acid sequence. For the features that are far apart interdependent, it takes a certain amount of time and steps to obtain enough information accumulation to connect them. The farther apart they are, the less likely the BiLSTM network captures effective information. This means that when an amino acid may be related to its surrounding amino acids or farther amino acids, using the BiLSTM algorithm only considers the information before and after the protein sequence within a certain range and cannot solve the correlation problem between discontinuous amino acids. It is worth noting that a single amino acid or a few amino acids may have a great impact on protein function. In the calculation process, the self-attention mechanism can directly connect the correlation between any two features in the sequence through one calculation step, which greatly shortens the distance between long-distance dependent features. Therefore, the combination of the BiLSTM algorithm and the self-attention mechanism gives more attention to the amino acids that may have a great influence on protein function, so that the amino acid sequence has a greater contribution to the accurate prediction of protein function. The architecture of the Chemical-SA-BiLSTM algorithm is shown in Figure 3.

The input sequence of the BiLSTM layer is $S = \{x_1, x_2, \cdots, x_t\}$. The model sequentially inputs each item $x_1, x_2, \cdots, x_t$ of the input sequence into the BiLSTM network. And then the forward output $\overrightarrow{h_t}$ and reverse output $\overleftarrow{h_t}$ at each time are calculated in the forward and reverse directions, respectively. After that, the output vectors in both forward and reverse directions are added. The final output vector h_t can be obtained. Then, the feature vector $H = (h_1, h_2, \cdots, h_n)$ obtained by the BiLSTM model is input into the self-attention model to calculate the weight vector a. The expression of the weight vector a can be obtained by Equation (1). By multiplying the feature vector and the weight vector a, the final vector m of the self-attention layer can be obtained, and its calculation equation is shown in Equation (2).

$$a = softmax\left(W_{s2}\,\mathbf{tanh}\left(W_{s1}H^T\right)\right) \tag{1}$$

$$m = aH \tag{2}$$

In the above equations, W_{s1} and W_{s2} are parameter matrices. The dimension of the weight vector a is n. The dimension of the vector m is 2u.

Next, the vector m output from the self-attention layer is transferred to the fully connected layer. Subsequently, this vector is mapped to a specific number by the Dense operation. To prevent overfitting, the Dropout layer [30] is added to the model. Finally, the output is mapped in the range [0, 1] through the sigmoid function of the activation layer, thereby obtaining the prediction result of the protein function.

Datasets

All protein data in the UniProtKB-SwissProt database have been carefully verified by experienced protein chemists and molecular biologists through consulting literature and computer tools, which can provide high-quality amino acid sequences and GO function annotation of grain protein for this experiment. Gene Ontology has been widely recognized as the gold standard for protein function annotation [31, 32], which covers three different sub-ontologies according to different function categories: Biological Process (BP), Molecular Function (MF) and Cellular Component (CC) [33]. Each GO function annotation describes a unique biological concept. And a protein is annotated by several GO function annotations.

In this experiment, the protein data of soybean, maize, indica and japonica are downloaded from the UniProtKB-SwissProt database. Only GO terms with experimental evidence codes (EXP, IDA, IPI, IMP, IGI, IEP, TAS and IC) can be retained. All proteins not annotated by GO terms are deleted. According to the category of GO function annotation, separate datasets are constructed for the three sub-ontologies of MF, BP and CC. For each grain protein dataset, we propagate annotations using the GO hierarchy structure [33]. If a protein is annotated with a GO annotation, it will be annotated with all of its ancestral annotations. When propagating annotations, we consider the GO term problem of proteins with different sub-ontologies. For instance, if a protein P has GO annotations of BP and MF, the protein P is classified into both the BP dataset and the MF dataset. After this step, the number of annotated proteins for each GO class is calculated and all classes with 50 or more annotations are selected for our prediction model. After this step, the data processed above are randomly divided into training set (80%) and test set (20%). Then the training set is divided into training set (80%) and validation set (20%) randomly. The final training set is used to train the model, and the validation set is used to evaluate how well the model predicts and select the best model. The best model is then used to evaluate the performance by the test set. After the above processing, the number of protein sequences of the four grain protein datasets is shown in Table 3.

Table 3. The distribution of protein sequences samples in the grain protein dataset for each sub-ontology.

Datasets	Sub-ontology	Training samples	Test samples	Total
Soybean	BP	264	67	331
	MF	283	71	354
	CC	256	65	321
Maize	BP	534	134	668
	MF	608	153	761
	CC	612	153	765
Indica	BP	652	164	816
	MF	756	189	945
	CC	754	189	943
Japonica	BP	2588	647	3235
	MF	2858	715	3573
	CC	2796	700	3496

Table 4. CNN network structure

Layer	Output Shape
Conv+Relu (64 filters of size 9, dropout=0.2)	(994,64)
Max Pooling (size 3, stride 3)	(331,64)
Conv+Relu (64 filters of size 7, dropout=0.2)	(325,64)
Max Pooling (size 3, stride 3)	(108,64)
Conv+Relu (64 filters of size 7, dropout=0.2)	(102,64)
Max Pooling (size 3, stride 3)	(34,64)
K Max Pooling (K=10)	(10,64)
Flatten	(640)

Baseline comparison methods

The model in this paper is compared with common neural network learning models: CNN, LSTM and CNN-BiLSTM. In addition, the CNN model combined with chemical properties, called Chemical-CNN, is used as comparison algorithm. The network structure of the CNN baseline model used in this experiment refers to the design of Zuallaertet [34]. Taking the BP dataset of japonica as an example, the CNN network structure is shown in Table 4. Additionally, the workflow of the CNN-BiLSTM model, a model combining the CNN algorithm and the BiLSTM algorithm in this experiment, is shown in Figure 4.

Evaluation metrics

In this paper, the accuracy, precision, recall and F1 values are used as the model evaluation metrics [35] to measure the prediction effect of the algorithm. The definitions of these evaluation metrics are shown in Equation (3), Equation (4), Equation (5) and Equation (6), respectively. TP, TN, FP and FN denote true positives, true negatives, false positives and false negatives, respectively. In general, when the precision is high, the recall is usually low. When the recall is high, the precision is slightly lower. Therefore, the F1 score is also used as the model evaluation metric in this experiment. The F1 score is the harmonic mean of the precision and the recall, which can synthetically consider the precision and the recall, and comprehensively evaluate the algorithm performance. The higher the F1 score, the better the algorithm performance and the

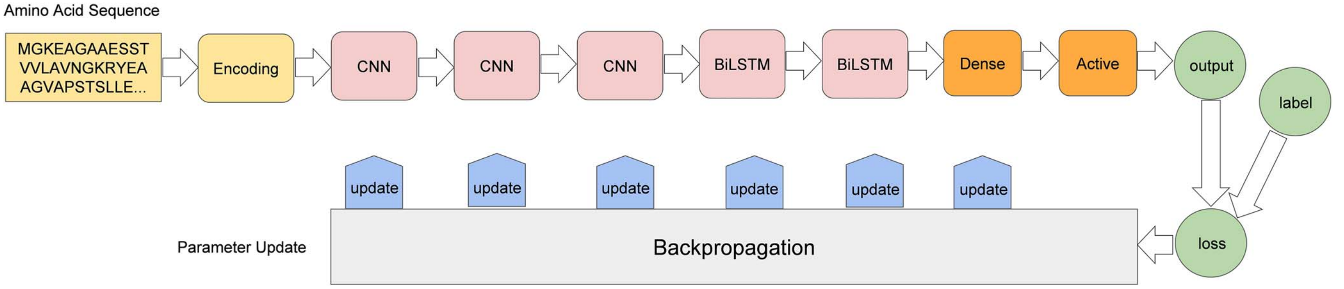

Figure 4. CNN-BiLSTM model workflow.

stronger the algorithm stability.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$F1 = \frac{2 \times Precison \times Recall}{Precison + Recall} \tag{6}$$

Experimental results and discussion

The performance comparison and analysis of the Chemical-SA-BiLSTM algorithm with other algorithms

In order to verify the performance superiority of the Chemical-SA-BiLSTM algorithm proposed in this paper in the grain protein function prediction model, the soybean protein, the maize protein, the indica protein and the japonica protein are selected as experimental objects. The Chemical-SA-BiLSTM algorithm is compared with the classical neural network algorithms (the CNN algorithm, the LSTM algorithm, the CNN-BiLSTM algorithm and the transformer algorithm). The experiments of each algorithm are run for 10 times, and all the prediction results of the 10 experiments are averaged. The function prediction results of the final four grain proteins are shown in Table 5. For each evaluation metric, the larger the evaluation value, the better the algorithm performance.

The accuracy, precision, recall and F1 score of the CNN, LSTM, CNN-BiLSTM and Chemical-SA-BiLSTM algorithm for soybean, maize, indica and japonica proteins are shown in Table 5. In terms of accuracy, the Chemical-SA-BiLSTM algorithm is superior to the CNN, LSTM, CNN-BiLSTM and transformer algorithms on the four grain protein datasets. On the BP dataset of indica protein, compared with the CNN algorithm, the accuracy of the Chemical-SA-BiLSTM algorithm is improved by 16.694889%, and its value is as high as 90.550564%. The accuracy of the LSTM algorithm is also higher than that of the CNN algorithm and the CNN-BiLSTM algorithm. For the values of precision, the results of the Chemical-SA-BiLSTM algorithm are higher than those of the CNN algorithm and the CNN-BiLSTM algorithm on the four grain protein datasets. The precision of the Chemical-SA-BiLSTM algorithm is generally higher than that of the LSTM algorithm. Only on the BP and CC datasets of maize protein and CC dataset of japonica protein, the precision of the Chemical-SA-BiLSTM algorithm is slightly lower than that of the LSTM algorithm. But on these datasets, the recall of the Chemical-SA-BiLSTM algorithm is higher than that of the LSTM algorithm. In order to comprehensively consider the precision value and recall value, the F1 score is used as an index to evaluate the performance of the algorithm in this experiment. In terms of F1 score, the LSTM algorithm is superior to the CNN and CNN-BiLSTM algorithms on the four grain protein datasets. The F1 score of the Chemical-SA-BiLSTM algorithm is higher than the other four algorithms, and its value on the BP dataset of soybean protein is as high as 86.806706%. On the BP dataset of indica protein, compared with the CNN algorithm, the F1 score of the Chemical-SA-BiLSTM algorithm is improved by 17.166506%. In general, the Chemical-SA-BiLSTM algorithm proposed in this paper is superior to the classical deep learning algorithms (the CNN algorithm, the LSTM algorithm the CNN-BiLSTM algorithm and the transformer algorithm) in terms of accuracy and F1 score, and can effectively predict the function of grain proteins.

It can also be seen from Table 5 that both the F1 score and accuracy of soybean with fewer samples are higher than those of maize with more samples. The F1 score and accuracy of the BP and CC datasets of soybean are even higher than those of japonica with thousands of samples. However, the F1 score and accuracy of the MF dataset of soybean are lower than those of japonica. Probably because the length of protein sequence selected for the experiment is 1002, hundreds of protein sequences with the length of 1002 are not a small amount of training for the model. When a certain amount of data is reached, the difference in algorithm performance may not be necessarily related to the amount of data. It does not mean that the larger the amount of data, the better the algorithm performance.

In addition, the 10-fold cross validation experiments are performed for all baseline models (CNN, LSTM, CNN-BiLSTM and transformer) and the Chemical-SA-BiLSTM model proposed in this paper on the soybean protein dataset to benchmark and compare the effectiveness of these models. All experiments are performed 10 times and the average value of the 10 experiments is taken as the final experimental results. The experimental results are presented in the form of mean ± standard deviation, as shown in Table 6. As can be seen in Table 6, the Chemical-SA-BiLSTM algorithm is higher than the CNN, LSTM, CNN-BILSTM and transformer algorithms in terms of accuracy and F1 score. On the standard deviation values of accuracy, precision, recall and F1 score, the standard deviation value of the Chemical-SA-BiLSTM algorithm is lower than those of the other four algorithms. It is worth noting that the standard deviations of the Chemical-SA-BiLSTM algorithm are all less than 1%. It further proves the superiority of Chemical-SA-BiLSTM algorithm in the prediction of grain protein function.

Table 5. Grain protein function prediction results

Datasets	Sub-ontology	Algorithm	Accuracy(%)	Precision(%)	Recall(%)	F1(%)
Soybean	BP	CNN	81.901	79.435	82.203	80.777
		LSTM	84.879	76.137	92.340	83.299
		CNN-BiLSTM	82.701	79.457	84.335	81.804
		transformer	83.927	77.098	89.436	82.668
		Chemical-SA-BiLSTM	**87.820**	85.008	89.068	**86.807**
	MF	CNN	84.357	73.340	89.567	80.546
		LSTM	86.384	72.644	93.679	81.819
		CNN-BiLSTM	85.228	76.138	88.716	81.759
		transformer	86.255	79.750	86.686	82.885
		Chemical-SA-BiLSTM	**87.097**	82.990	85.491	**84.101**
	CC	CNN	80.370	78.235	80.899	79.486
		LSTM	83.181	82.755	82.623	82.660
		CNN-BiLSTM	80.760	79.797	80.638	80.197
		transformer	84.208	89.177	77.430	82.790
		Chemical-SA-BiLSTM	**84.835**	83.979	84.557	**84.182**
Maize	BP	CNN	77.881	73.025	62.072	67.089
		LSTM	79.523	75.102	67.108	70.863
		CNN-BiLSTM	77.927	69.936	67.120	68.488
		transformer	78.484	71.621	64.924	67.926
		Chemical-SA-BiLSTM	**80.421**	74.801	69.763	**72.091**
	MF	CNN	82.567	64.687	73.956	69.001
		LSTM	83.747	70.761	71.080	70.900
		CNN-BiLSTM	83.736	68.977	69.539	69.144
		transformer	84.425	74.580	67.913	70.620
		Chemical-SA-BiLSTM	**86.281**	74.898	75.043	**74.939**
	CC	CNN	78.143	79.064	61.307	68.872
		LSTM	82.424	79.624	70.760	74.920
		CNN-BiLSTM	78.830	78.080	63.111	69.799
		transformer	80.498	84.671	61.801	71.383
		Chemical-SA-BiLSTM	**83.940**	79.421	75.423	**77.151**
Indica	BP	CNN	73.856	50.237	88.228	64.020
		LSTM	89.140	81.669	77.071	79.294
		CNN-BiLSTM	87.973	78.994	74.139	76.461
		transformer	88.574	82.770	71.911	76.611
		Chemical-SA-BiLSTM	**90.551**	83.113	79.443	**81.186**
	MF	CNN	84.454	73.636	51.654	60.673
		LSTM	86.190	66.831	61.324	63.911
		CNN-BiLSTM	84.796	63.966	63.482	63.660
		transformer	86.053	71.717	55.493	61.959
		Chemical-SA-BiLSTM	**87.274**	75.071	63.248	**68.608**
	CC	CNN	81.506	78.081	77.655	77.800
		LSTM	82.186	80.101	77.618	78.825
		CNN-BiLSTM	81.691	78.400	77.655	78.006
		transformer	83.154	85.246	74.326	79.334
		Chemical-SA-BiLSTM	**85.299**	81.409	83.919	**82.434**
Japonica	BP	CNN	78.798	71.1667	58.817	64.276
		LSTM	80.766	71.490	63.920	67.489
		CNN-BiLSTM	79.374	65.345	66.320	65.817
		transformer	80.186	71.903	60.083	65.362
		Chemical-SA-BiLSTM	**81.367**	73.773	64.354	**68.715**
	MF	CNN	83.362	75.888	85.992	80.306
		LSTM	87.215	82.618	88.216	85.309
		CNN-BiLSTM	84.603	78.035	85.939	81.685
		transformer	85.720	80.523	86.480	83.375
		Chemical-SA-BiLSTM	**87.910**	84.229	88.207	**86.105**
	CC	CNN	78.649	75.274	78.009	76.524
		LSTM	81.614	81.079	80.268	80.656
		CNN-BiLSTM	79.214	74.139	82.478	77.935
		transformer	79.209	80.544	75.759	78.003
		Chemical-SA-BiLSTM	**83.728**	80.792	84.984	**82.689**

Table 6. Results of cross validation experiment for soybean protein function prediction

Sub-ontology	Algorithm	Accuracy(%)	Precision(%)	Recall(%)	F1(%)
BP	CNN	83.053 ± 1.319	78.898 ± 1.542	85.672 ± 1.587	82.128 ± 1.003
	LSTM	86.049 ± 0.709	81.747 ± 1.929	87.844 ± 1.271	84.672 ± 1.295
	CNN-BiLSTM	83.438 ± 1.078	80.472 ± 1.075	84.518 ± 1.296	82.442 ± 1.051
	transformer	84.292 ± 0.754	80.432 ± 1.157	91.653 ± 0.947	85.671 ± 0.794
	Chemical-SA-BiLSTM	**89.039 ± 0.411**	85.650 ± 0.937	93.003 ± 0.793	**89.170 ± 0.568**
MF	CNN	85.018 ± 1.663	74.724 ± 1.733	89.036 ± 2.482	81.224 ± 1.336
	LSTM	85.943 ± 1.502	78.939 ± 2.002	88.822 ± 1.204	83.576 ± 1.357
	CNN-BiLSTM	85.684 ± 1.370	77.654 ± 2.339	89.347 ± 2.748	83.038 ± 1.417
	transformer	86.959 ± 1.377	80.859 ± 1.119	87.173 ± 1.509	83.886 ± 0.900
	Chemical-SA-BiLSTM	**89.039 ± 0.985**	83.433 ± 0.657	89.045 ± 0.882	**86.145 ± 0.621**
CC	CNN	81.233 ± 0.704	79.845 ± 1.683	80.291 ± 1.458	80.058 ± 1.312
	LSTM	83.455 ± 1.030	85.812 ± 2.076	82.093 ± 1.798	83.874 ± 0.768
	CNN-BiLSTM	81.786 ± 1.113	80.710 ± 1.379	82.072 ± 0.937	81.372 ± 0.548
	transformer	83.719 ± 1.189	88.306 ± 1.925	79.905 ± 1.930	83.877 ± 1.472
	Chemical-SA-BiLSTM	**85.643 ± 0.701**	86.062 ± 0.850	84.611 ± 0.522	**85.327 ± 0.418**

Table 7. Helicobacter pylori protein function prediction results

Sub-ontology	Algorithm	Accuracy(%)	Precision(%)	Recall(%)	F1(%)
BP	CNN	80.028	62.821	97.233	76.255
	LSTM	81.752	64.644	98.860	78.147
	CNN-BiLSTM	81.353	63.384	99.324	77.378
	Chemical-SA-BiLSTM	**84.791**	75.028	94.556	**83.556**
MF	CNN	74.631	53.415	89.724	66.581
	LSTM	75.120	51.339	98.326	67.454
	CNN-BiLSTM	75.196	50.902	99.223	67.284
	Chemical-SA-BiLSTM	**78.351**	64.069	87.108	**73.398**
CC	CNN	77.310	60.643	81.929	69.649
	LSTM	81.966	65.299	95.898	77.693
	CNN-BiLSTM	81.358	64.691	94.075	76.645
	Chemical-SA-BiLSTM	**95.397**	92.155	98.061	**94.885**

In order to verify the performance of the Chemical-SA-BiLSTM algorithm on other protein datasets, helicobacter pylori protein is selected as experimental object for protein function prediction. Similarly, the experimental results of the Chemical-SA-BiLSTM algorithm are compared with the CNN, LSTM and CNN-BiLSTM algorithms. The experiments are run for 10 times and the average of the 10 experimental results is taken. The experimental results are shown in Table 7. It can be seen from the experimental results that, compared with the CNN, LSTM and CNN-BiLSTM algorithms, the Chemical-SA-BiLSTM algorithm has the highest accuracy, F1 score and precision. This proves the feasibility of applying the Chemical-SA-BiLSTM algorithm to the function prediction of other proteins, and also lays a foundation for applying the Chemical-SA-BiLSTM algorithm to generic protein function prediction in the future.

The comparison and analysis of chemical properties effect

In this study, in order to further verify the influence of chemical properties on the grain protein function prediction model performance, the soybean protein and the maize protein are selected as the experimental objects. And the CNN algorithm and the SA-BiLSTM algorithm are compared with their algorithms integrating chemical properties. The grain protein function prediction results are shown in Table 8.

As can be seen from Table 8, on the soybean and maize protein datasets, the accuracy and F1 score of the Chemical-CNN algorithm are higher than those of the CNN algorithm. Moreover, the accuracy and F1 score of the Chemical-SA-BiLSTM algorithm are also higher than the CNN algorithm, the Chemical-CNN algorithm and the SA-BiLSTM algorithm. The reason is that the chemical properties of amino acids are related to protein function. This indicates that adding the chemical properties of amino acids to protein sequences can enrich the input information, which is beneficial to improve the performance of grain protein function prediction model. In addition, compared with the CNN algorithm and the SA-BiLSTM algorithm without amino acid chemical properties, the Chemical-CNN algorithm and the Chemical-SA-BiLSTM algorithm only slightly improved the prediction effect of grain protein function. The probable reason is that the chemical properties used in this experiment can be easily inferred from amino acid sequences, and neural network models can implicitly learn these properties automatically. In other words, future research can try to use other more complex chemical or physical properties, which may have a greater impact on the performance of protein function prediction models. In conclusion, compared with the algorithms without amino acid chemical properties, the algorithms with amino acid chemical properties are better in predicting the function of grain protein.

Cross-species protein prediction

In order to understand the prediction effect of one grain protein function prediction model on another grain protein, the MF dataset of japonica protein and the Chemical-SA-BiLSTM

Table 8. The comparison of chemical properties effect for grain protein function prediction

Datasets	Sub-ontology	Algorithm	Accuracy(%)	Precision(%)	Recall(%)	F1(%)
Soybean	BP	CNN	81.901	79.435	82.203	80.777
		Chemical-CNN	82.672	78.082	85.268	81.513
		SA-BiLSTM	85.013	78.656	90.047	83.906
		Chemical-SA-BiLSTM	**87.820**	85.008	89.068	**86.807**
	MF	CNN	84.357	73.340	89.567	80.546
		Chemical-CNN	84.375	74.022	88.531	80.577
		SA-BiLSTM	86.861	83.862	84.203	83.937
		Chemical-SA-BiLSTM	**87.097**	82.990	85.491	**84.101**
	CC	CNN	80.371	78.235	80.899	79.486
		Chemical-CNN	80.391	79.676	79.464	79.564
		SA-BiLSTM	84.195	85.390	81.832	83.187
		Chemical-SA-BiLSTM	**84.835**	83.979	84.557	**84.182**
Maize	BP	CNN	77.881	73.025	62.072	67.089
		Chemical-CNN	77.909	68.547	67.394	67.941
		SA-BiLSTM	80.134	75.747	67.859	71.630
		Chemical-SA-BiLSTM	**80.421**	74.801	69.763	**72.091**
	MF	CNN	82.567	64.687	73.956	69.001
		Chemical-CNN	82.730	63.018	76.800	69.113
		SA-BiLSTM	86.062	76.667	73.029	74.719
		Chemical-SA-BiLSTM	**86.281**	74.898	75.043	**74.939**
	CC	CNN	78.143	79.064	61.307	68.872
		Chemical-CNN	78.345	71.953	67.170	69.455
		SA-BiLSTM	83.463	79.695	73.480	76.407
		Chemical-SA-BiLSTM	**83.940**	79.421	75.423	**77.151**

Table 9. Cross-species protein prediction results

Datasets	Accuracy(%)	Precision(%)	Recall(%)	F1(%)
Soybean	85.472	84.766	80.127	82.381
Maize	76.954	89.330	63.249	74.060
Indica	80.889	75.128	60.929	67.287

Table 10. Examples of indica protein and japonica protein function prediction results

Datasets	Sub-ontology	Protein	Real function	Predicted function
Indica	BP	Q01N44 (FAAH_ORYSI)	GO:0016042 GO:0006629	GO:0016042 GO:0006629 GO:0006807
	MF	E0ZS48 (UREA_ORYSI)	GO:0009039 GO:0016787 GO:0046872 GO:0016810	GO:0009039 GO:0016787 GO:0046872
	CC	FAAH_ORYSI (Q01N44)	GO:0005783 GO:0005886 GO:0016020 GO:0005789	GO:0005783 GO:0005886 GO:0016020 GO:0005789 GO:0009536
Japonica	BP	Q9DE67 (LUM_COTJA)	GO:0030199 GO:0007601 GO:0032914 GO:0045944	GO:0030199 GO:0007601 GO:0032914
	MF	Q7XFK2 (BGA14_ORYSJ)	GO:0030246 GO:0016787 GO:0016798	GO:0030246 GO:0016787 GO:0016798 GO:0003824
	CC	Q7XWK5 (SAG39_ORYSJ)	GO:0005615 GO:0005764 GO:0010282 GO:0005773	GO:0005615 GO:0005764 GO:0010282 GO:0005773 GO:0005634

algorithm are used for model training, and then the trained model is used to predict the function of the MF datasets of soybean, maize and indica proteins. The prediction results are shown in Table 9. Combined with the results of the Chemical-SA-BiLSTM algorithm in Table 5, the F1 score obtained using the japonica model is relatively close to the F1 score obtained using its own dataset model. Compared with the CNN, LSTM and CNN-BiLSTM algorithm results in Table 5, the F1 score obtained using the japonica model in Table 9 is higher. In addition, compared with the results of the four algorithms in Table 5, the accuracy obtained using the japonica model is relatively low. It may be that the prediction results are not stable due to the use of models trained on other datasets to make predictions. Nevertheless, the F1 score and accuracy obtained using the japonica model for the prediction of soybean, maize and indica proteins are relatively high. This finding could be of great help in predicting protein function with insufficient data.

Comparison and analysis of predicted function and real function

In order to further study the causes of the abnormal function prediction results of some proteins, the Chemical-SA-BilSTM algorithm is used to predict the grain protein function in this section. From the experimental results of the three sub-ontologies of indica and japonica protein, BP, MF and CC, a piece of protein data with an incorrect prediction is selected. The predicted protein function in this experiment is compared with the real function in the Swiss-Prot database. The comparison results of protein functions selected in this section are shown in Table 10.

First of all, the actual effect of indica protein function prediction is analyzed in this part. In the Swiss-Prot database, the Q01N44 (FAAH_ORYSI) protein is confirmed to have two protein functions: GO:0016042 and GO:0006629. The GO:0016042 function is represented as lipid catabolic process, specifically referring to the chemical reactions and pathways resulting in the breakdown of lipids. The GO:0006629 function stands for lipid metabolic process, which is defined as the chemical reactions and pathways involving lipids. Unexpectedly, in the protein function prediction results of this experiment, in addition to the complete prediction of the GO:0016042 function and the GO:0006629 function of the Q01N44 (FAAH_ORYSI) protein, the GO:0006807 function is also predicted. The GO:0006807 function is expressed as nitrogen compound metabolic process. This metabolic process specifically refers to the chemical reactions and pathways involving organic or inorganic compounds that contain nitrogen. The reason why the GO:0006807 function is predicted may be that it is related to the lipid metabolic process (GO:0006629) and the nitrogen compound metabolic process (GO:0006807). Additionally, it is also mentioned in several literatures [36, 37] that the GO:0006807 function is an indispensable function in the biological processes of indica protein.

The E0ZS48 (UREA_ORYSI) protein is found by manual annotation to have GO:0009039, GO:0016787, GO:0016810 and GO:0046872 protein functions. Among them, the GO:0009039 function represents urease activity. The GO:0016787 function indicates hydrolase activity, which can catalyze the hydrolysis of various bonds. The GO:0016810 function stands for hydrolase activity that catalyzes the hydrolysis of any carbon–nitrogen bond C-N, except peptide bonds. GO:0046872 means that the protein has the function of binding to a metal ion. Unfortunately, the experimental results show that the GO:0016810 function cannot be successfully predicted, which may be because the GO:0016810 function and the GO:0016787 function have similar functions. They both catalyze the hydrolysis of certain bonds. Therefore, it is difficult for experiments to make completely accurate function prediction on this problem.

Similarly, the FAAH_ORYSI (Q01N44) protein has four protein functions: GO:0005783, GO:0005886, GO:0016020 and GO:0005789. They are denoted as endoplasmic reticulum, plasma membrane, membrane, and endoplasmic reticulum membrane, respectively. In the prediction process, in addition to the complete prediction of the above four protein functions, the GO:0009536 function is additionally predicted. The GO:0009536 function is represented as plastid. The plastid is a member of a family of organelles find in the cytoplasm of plants and some protists that are membrane membrane-bounded DNA. Moreover, organelles are mainly composed of mitochondria, endoplasmic reticulum, centrosomes and so on. So, the GO:0009536 function is predicted in the experiment probably because the plastid is related to the membrane and endoplasmic reticulum. Notably, Xu et al. [38] clearly indicated that the GO:0009536 function is the core function of multiple proteins in indica.

Subsequently, examine the actual situation of japonica protein function prediction. The Q9DE67(LUM_COTJA) protein is shown to have GO:0030199, GO:0007601, GO:0045944 and GO:0032914 functions in the Swiss-Prot database. But the GO:0045944 function is not found in the prediction process. The GO:0045944 function stands for positive regulation of transcription by RNA polymerase II. The GO:0032914 function indicates positive regulation of transforming growth factor beta1 production. It may be that GO:0045944 and GO:0032914 have similar functions, and both have a positive regulatory effect on a certain protein component, which makes it difficult to correctly predict the GO:0045944 function in experiments.

In the MF subset of japonica dataset, the manual annotation functions for the Q7XFK2 (BGA14_ORYSJ) protein are GO:0030246, GO:0016787, GO:0016798 and GO:0003824. Compared with the function annotations displayed in the Swiss-Prot database, the GO:0003824 function is additionally found in this experiment for the prediction of the Q7XFK2 (BGA14_ORYSJ) protein, which is defined as the catalysis of a biochemical reaction at physiological temperatures. GO:0016787 and GO:0016798 both have the function of catalyzing the hydrolysis of certain bonds, which are similar to the GO:0003824 function. In addition, it was confirmed by multiple literatures [39–41] that most genes in japonica have GO:0003824 function.

Finally, for the Q7XWK5 (SAG39_ORYSJ) protein, it is confirmed to have GO:0005615, GO:0005764, GO:0010282 and GO:0005773 functions in the Swiss-Prot database. The GO:0005615 is defined as extracellular space. And the GO:0005764 is represented as lysosome. The GO:0010282 and GO:0005773 functions both represent some sort of vacuole organelle. Surprisingly, in the results of predicting the Q7XWK5 (SAG39_ORYSJ) protein function, it is shown that it also has the GO:0005634 function. The GO:0005634 function represents a membrane-bounded organelle of eukaryotic cells in which chromosomes are housed and replicated. It is represented as certain sort of organelle, like the GO:0010282 and GO:0005773 functions. Notably, the researchers showed that the GO:0005634 function predominated in the main categories of cellular component of japonica [42, 43].

In summary, the Chemical-SA-BiLSTM algorithm may not necessarily accurately predict the complete function of grain protein with very similar functions. But most grain protein functions can be predicted accurately. According to Table 5 and Table 10, the prediction effect of grain protein by the Chemical-SA-BiLSTM algorithm is generally good. What is more, in addition to completely predicting the protein function annotation, the experiment may also predict the GO function annotations that are not shown in the Swiss-Prot database. This provides a new direction for subsequent experimental research. In the future, researchers can try to verify whether the protein really contains these GO function annotations through biological method.

Conclusion

The study of grain protein is one of the research hotspots in biology in recent years. However, there is still a lack of research on the prediction of grain protein function. In this paper, the combination of the BiLSTM algorithm and the self-attention mechanism is applied to grain protein function prediction for the first time. At the same time, considering that the function of protein depends on the chemical properties of its amino acids, chemical properties are added to the amino acid sequence in this experiment. Finally, an accurate protein function prediction model, Chemical-SA-BiLSTM, is proposed. In order to evaluate the prediction effect of the model, four grain protein datasets, soybean, maize, indica and japonica, are used for experiments. The experimental results show that the performance of the Chemical-SA-BiLSTM algorithm is better than other classical neural network algorithms. The Chemical-SA-BiLSTM algorithm can achieve better prediction effect of grain protein function. With the improvement of prediction effect, we will extend the state-of-the-art Chemical-SA-BiLSTM algorithm to different protein function prediction tasks. Meanwhile, a new research direction for protein functional annotation is provided in this paper. That is, the possible protein function annotations can be determined by computational method. The biological experimental method can be used to determine whether the above function annotations exist.

Key Points

- Adding chemical properties to amino acid sequences can enrich protein information.
- The combination of BiLSTM algorithm and self-attention mechanism performs better than other classical neural network algorithms.
- Computational methods can be used to predict possible protein functional annotations, and then biological experiments can be used to determine whether such functional annotations exist.

Data availability

The data and source code for analyses in this manuscript are available at: https://github.com/HwaTong/Chemical-SA-BiLSTM.

Acknowledgements

Project supported by National Natural Science Foundation of China (32172247), Local University Capacity Building Project of Shanghai Science and Technology Commission(20060502100), the Major Project of Inner Mongolia Science and Technology Department(2021SZD0017).

References

1. Reeves T, Thomas G, Ramsay G. Save and grow in practice: maize, rice, wheat. *A guide to sustainable cereal production (FAO UN, 2016)* 2016.
2. Raubenheimer D, Simpson SJ. Nutritional ecology and human health. *Annu Rev Nutr* 2016;**36**:603–26.
3. Saeidnia S, Manayi A, Abdollahi M. From in vitro experiments to in vivo and clinical studies; pros and cons. *Curr Drug Discov Technol* 2015;**12**(4):218–24.
4. Jiang Y, Oron TR, Clark WT, *et al.* An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biol* 2016;**17**(1):1–19.
5. Gillis J, Pavlidis P. Characterizing the state of the art in the computational assignment of gene function: lessons from the first critical assessment of functional annotation (cafa). *BMC bioinformatics* 2013;**14**(3):1–12.
6. Cai CZ, Han LY, Ji ZL, *et al.* Svm-prot: web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Res* 2003;**31**(13):3692–7.
7. Guoxian Y, Rangwala H, Domeniconi C, *et al.* Predicting protein function using multiple kernels. *IEEE/ACM Trans Comput Biol Bioinform* 2014;**12**(1):219–33.
8. Nam J-W, Shin K-R, Han J, *et al.* Human microrna prediction through a probabilistic co-learning model of sequence and structure. *Nucleic Acids Res* 2005;**33**(11):3570–81.
9. Nguyen CD, Gardiner KJ, Nguyen D, *et al.* Prediction of protein functions from protein interaction networks: a naïve bayes approach. In: Ho Tu-Bao, Zhou Zhi-Hua (eds). *Pacific Rim International Conference on Artificial Intelligence*. Berlin: Springer, 2008, 788–98.
10. Yousef M, Jung S, Showe LC, *et al.* Learning from positive examples when the negative class is undetermined-microrna gene identification. *Algorithms for molecular biology* 2008;**3**(1):1–9.
11. Chen X-W, Liu M. Prediction of protein–protein interactions using random decision forest framework. *Bioinformatics* 2005;**21**(24):4394–400.
12. Kulmanov M, Khan MA, Hoehndorf R. Deepgo: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics* 2018;**34**(4):660–8.
13. Kulmanov M, Hoehndorf R. Deepgoplus: improved protein function prediction from sequence. *Bioinformatics* 2020;**36**(2):422–9.
14. Sara ST, Hasan MM, Ahmad A, *et al.* Convolutional neural networks with image representation of amino acid sequences for protein function prediction. *Comput Biol Chem* 2021;**92**:107494.
15. Elhaj-Abdou MEM, El-Dib H, El-Helw A, *et al.* Deep_cnn_lstm_go: Protein function prediction from amino-acid sequences. *Comput Biol Chem* 2021;**95**:107584.
16. Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Netw* 2005;**18**(5–6):602–10.
17. Cheng J, Dong L, Lapata M. Long short-term memory-networks for machine reading. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 2016, 551–61.
18. Vaswani A, Shazeer N, Parmar N, *et al.* Attention is all you need. *Advances in neural information processing systems* 2017;**30**:5998–6008.
19. Corral-Corral R, Beltrán JA, Brizuela CA, *et al.* Systematic identification of machine-learning models aimed to classify critical residues for protein function from protein structure. *Molecules* 2017;**22**(10):1673.
20. Wen B, Zeng W-F, Liao Y, *et al.* Deep learning in proteomics. *Proteomics* 2020;**20**(21–22):1900335.
21. Hein A, Cole C, Valafar H. An investigation in optimal encoding of protein primary sequence for structure prediction by artificial neural networks. In: Hamid R. Arabnia, Hayaru Shouno, Quoc-Nam Tran, Leonidas Deligiannidis, Fernando G. Tinetti (eds). *Advances in Computer Vision and Computational Biology*. Cham: Springer, 2021, 685–99.
22. Szalkai B, Grolmusz V. Near perfect protein multi-label classification with deep neural networks. *Methods* 2018;**132**:50–6.
23. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* 1997;**9**(8):1735–80.
24. Hochreiter S. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 1998; **6**(02):107–16.
25. Mnih V, Heess N, Graves A, *et al.* Recurrent models of visual attention. *Advances in neural information processing systems* 2014;**27**:2204–12.
26. Yang Z, Yang D, Dyer C, *et al.* Hierarchical attention networks for document classification. In: Kevin Knight (ed). *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, San Diego, CA, USA: Association for Computational Linguistics; 2016, 1480–9.
27. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. *Proceedings of International Conference on Learning Representations*. 2015; arXiv:1409.0473.
28. Verga P, Strubell E, McCallum A. Simultaneously self-attending to all mentions for full-abstract biological relation extraction. *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2018, 872–84.
29. Wang D, Zeng S, Chunhui X, *et al.* Musitedeep: a deep-learning framework for general and kinase-specific phosphorylation site prediction. *Bioinformatics* 2017;**33**(24):3909–16.

30. Srivastava N, Hinton G, Krizhevsky A, *et al.* Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* 2014;**15**(1):1929–58.
31. Gene Ontology Consortium. The gene ontology in 2010: extensions and refinements. *Nucleic Acids Res* 2010;**38**(suppl_1): D331–5.
32. Huntley RP, Sawford T, Martin MJ, *et al.* Understanding how and why the gene ontology and its annotations evolve: the go within uniprot. *GigaScience* 2014;**3**(1):2047–17X.
33. Ashburner M, Ball CA, Blake JA, *et al.* Gene ontology: tool for the unification of biology. *Nat Genet* 2000;**25**(1):25–9.
34. Zuallaert J, Pan X, Saeys Y, *et al.* Investigating the biological relevance in trained embedding representations of protein sequences. In: *Workshop on Computational Biology at the 36th International Conference on Machine Learning (ICML 2019)*. Workshop on Computational Biology at ICML2019, Long Beach, CA, USA; 2019.
35. Jinbo X, Wang S. Analysis of distance-based protein structure prediction by deep learning in casp13. *Proteins: Structure, Function, and Bioinformatics* 2019;**87**(12):1069–81.
36. Wang Y, Zhao M, Zhang Q, *et al.* Genomic distribution and possible functional roles of putative g-quadruplex motifs in two subspecies of oryza sativa. *Comput Biol Chem* 2015;**56**:122–30.
37. Kumar V, Jain P, Venkadesan S, *et al.* Understanding rice-magnaporthe oryzae interaction in resistant and susceptible cultivars of rice under panicle blast infection using a time-course transcriptome analysis. *Genes* 2021;**12**(2):301.
38. Qun X, Yuan X, Wang S, *et al.* The genetic diversity and structure of indica rice in china as detected by single nucleotide polymorphism analysis. *BMC Genet* 2016;**17**(1):1–8.
39. Silveira RD, Abreu FRM, Mamidi S, *et al.* Expression of drought tolerance genes in tropical upland rice cultivars (oryza sativa). *Embrapa Milho e Sorgo-Artigo em periódico indexado (ALICE)* 2015; **14**:8181–200.
40. da Maia, Cadore PRB, Benitez LC, *et al.* Transcriptome profiling of rice seedlings under cold stress. *Funct Plant Biol* 2016;**44**(4): 419–29.
41. Zhang X, Fan X, Wang Y, *et al.* Exploring core response mechanisms to multiple environmental stressors via a genome-wide study in the brown alga saccharina japonica (laminariales, phaeophyceae). *J Phycol* 2021;**57**(1):345–54.
42. Azameti MK, Dauda WP, Panzade KP, *et al.* Identification and characterization of genes responsive to drought and heat stress in rice (oryza sativa l.). *Vegetos* 2021;**34**(2): 309–17.
43. Kim D-M, Kang J-W, Shim K-C, *et al.* Characterization of genes associated with salt tolerance using transcriptome analysis and quantitative trait loci mapping in rice. *Plant Breeding and Biotechnology* 2021;**9**(4):318–30.