Molecular Therapy
Nucleic Acids
Original Article

ACP-DL: A Deep Learning Long Short-Term Memory Model to Predict Anticancer Peptides Using High-Efficiency Feature Representation

Hai-Cheng Yi,[1,2,3] Zhu-Hong You,[1,3] Xi Zhou,[1] Li Cheng,[1] Xiao Li,[1] Tong-Hai Jiang,[1] and Zhan-Heng Chen[1]

[1]The Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, Urumqi 830011, China; [2]University of Chinese Academy of Sciences, Beijing 100049, China

Cancer is a well-known killer of human beings, which has led to countless deaths and misery. Anticancer peptides open a promising perspective for cancer treatment, and they have various attractive advantages. Conventional wet experiments are expensive and inefficient for finding and identifying novel anticancer peptides. There is an urgent need to develop a novel computational method to predict novel anticancer peptides. In this study, we propose a deep learning long short-term memory (LSTM) neural network model, ACP-DL, to effectively predict novel anticancer peptides. More specifically, to fully exploit peptide sequence information, we developed an efficient feature representation approach by integrating binary profile feature and *k*-mer sparse matrix of the reduced amino acid alphabet. Then we implemented a deep LSTM model to automatically learn how to identify anticancer peptides and non-anticancer peptides. To our knowledge, this is the first time that the deep LSTM model has been applied to predict anticancer peptides. It was demonstrated by cross-validation experiments that the proposed ACP-DL remarkably outperformed other comparison methods with high accuracy and satisfied specificity on benchmark datasets. In addition, we also contributed two new anticancer peptides benchmark datasets, ACP740 and ACP240, in this work.

INTRODUCTION

Cancer is one of the most devastating killers of human beings, accounting for millions of deaths around the world each year.[1,2] Conventional physical and chemical methods, including targeted therapy, chemotherapy, and radiation therapy, remain the principle modes to treat cancer, which focus on killing the diseased cells, but normal cells are also adversely affected.[3,4] More obviously, these treatments are expensive and inefficient, which means there is an urgent need to develop novel efficient measures to solve this deadly disease.[5] The discovery of anticancer peptides (ACPs), a kind of short peptide generally with a length less than 50 amino acids and most of which are derived from antimicrobial peptides (AMPs), often cationic in nature, has led to the emergence of a novel alternative therapy to treat cancer.

ACPs open a promising perspective for cancer treatment, and they have various attractive advantages,[6,7] including high specificity, ease of synthesis and modification, low production cost, and so on.[8] ACPs could interact with the anionic cell membrane components of only cancer cells, and, for this reason, they can selectively kill cancer cells with almost no harmful effect on normal cells.[4,9] In addition, few ACPs, e.g., cell-penetrating peptides or peptide drugs, inhibit the cell cycle or any other functionality. Thus, they are safer than traditional broad-spectrum drugs, which have become the most competitive choice as therapeutics compared to small molecules and antibodies. In recent years, ACP therapeutics have been extensively explored and used to fight various tumor types across different phases of preclinical and clinical trials.[10–14] However, only a few of them can eventually be employed for clinical treatment. Furthermore, it's time-consuming, expensive, and lab-limited to identify potential new ACPs by experiment.

With the huge therapeutic importance of ACPs, there is an urgent need to develop highly efficient prediction techniques. Some notable research has been reported in the prediction of ACPs.[15] Tyagi et al.[16] developed a support vector machine (SVM) model using amino acid composition (AAC) and dipeptide composition as input features on experimentally confirmed anticancer peptides and random peptides derived from the Swiss-Prot database. Hajisharifi et al.[17] also reported an SVM model using Chou's[18,19] pseudo AAC (PseAAC) and the local alignment kernel-based method. Vijayakumar and Ptv[20] proposed that, between ACPs and non-ACPs, there was no significant difference in AAC observed. Also, they presented a novel encoding measure, which achieved better predictive performance than AAC-based features, considering both compositional information and centroidal, distributional measures of amino acids. Shortly afterward, based on the optimal g-gap dipeptide components, by exploring the correlation between long-range residues and sequence-order effects, Chen et al.[21] described iACP, which exhibited the best predictive

Received 9 March 2019; accepted 8 April 2019;
https://doi.org/10.1016/j.omtn.2019.04.025.

[3]These authors contributed equally to this work.

Correspondence: Zhu-Hong You, The Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, Urumqi 830011, China.
E-mail: zhuhongyou@ms.xjb.ac.cn

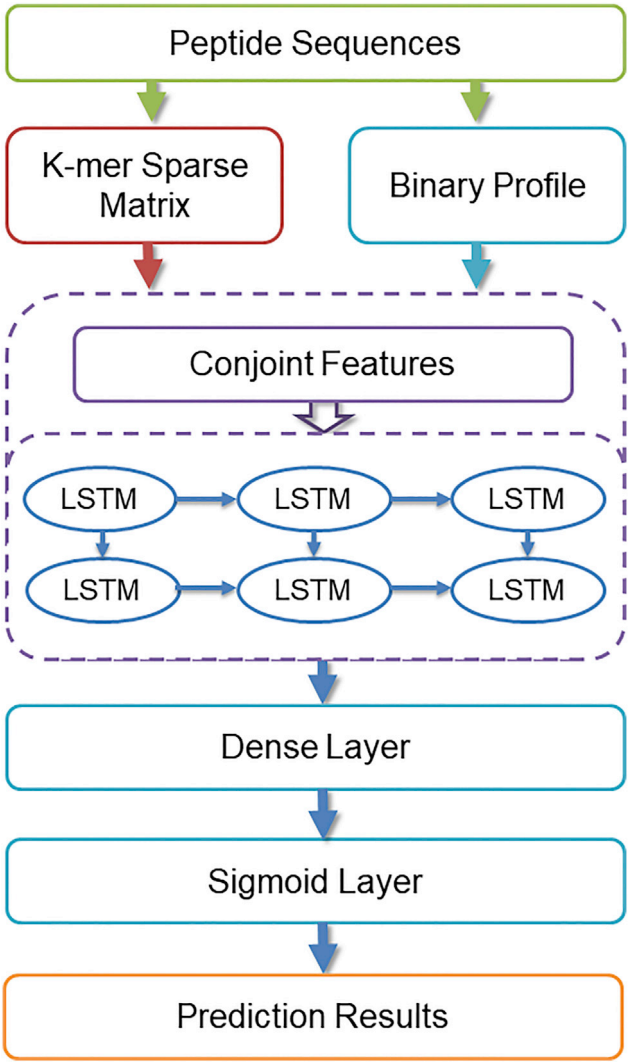

Figure 1. The Flowchat of Our ACP-DL Method

We used the *k*-mer sparse matrix and binary profile feature to represent peptide sequences, and the deep LSTM model is trained to predict anticancer peptides.

performance at that time. More recently, Wei et al.[22] developed a sequence-based predictor called ACPred-FL, which uses two-step feature selection and seven different feature representation methods.

According to the cognition of the short length of ACPs, it's difficult to exploit the efficient features of many mature feature representation methods, which are widely used on protein sequences.[23] With the rapid growth of the number of ACPs that has been identified experimentally, by machine learning, and by bioinformatics research,[24–40] the computational prediction methods of ACPs still need further development.

In this study, we proposed a deep learning long short-term memory (LSTM) neural network model to predict anticancer peptides, which we named ACP-DL. The efficient features exploited from peptides

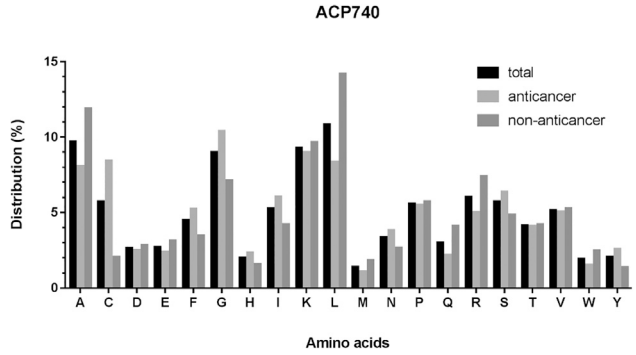

Figure 2. Comparison of Amino Acid Composition of Anticancer, Non-anticancer, and Total Peptides in Dataset ACP740

sequences are fed as input to train the LSTM model. More specifically, peptide sequences are transformed by *k*-mer sparse matrix of the reduced amino acid alphabet,[41,42] which is a 2D matrix, and retained almost complete sequence order and amino acid component details. Meanwhile, peptide sequence are also converted by a binary profile feature,[43] which can be regarded as one-hot encoding of categorical variables and has been suggested to be an efficient feature extraction technique.[16,22] Finally, these features are fed into our LSTM model to predict new anticancer peptides.

To further evaluate the performance of our model, we evaluated the ACP-DL on two novel benchmark datasets. We also compared the purposed ACP-DL with existing state-of-the-art machine-learning models, e.g., SVM,[44,45] Random Forest (RF),[46]and Naive Bayes (NB).[47] The 5-fold cross-validation experimental results showed that our method is suitable for the anticancer prediction mission with notable prediction performance. The workflow of ACP-DL is show in Figure 1.

RESULTS AND DISCUSSION

Above all, we compared the different distributions of amino acids in anticancer peptides, non-anticancer peptides, and all peptides in datasets ACP740 and ACP240. The results for ACP740 are shown in Figure 2, the composition of all 20 amino acids in these peptides were counted and compared. Certain residues, including Cys, Phe, Gly, His, Ile, Asn, Ser, and Tyr, were found to be abundant in anticancer peptides compared to non-anticancer peptides, while Glu, Leu, Met, Gln, Arg, and Trp were abundant in non-anticancer peptides compared to anticancer peptides. Similarly, as shown in Figure 3, in dataset ACP240, the Phe, His, Ile, and Lys were abundant in anticancer peptides. Since terminal residues play essential roles in biological functions of peptides.

Evaluation of ACP-DL's Capability to Predict Anticancer Peptides

First, we executed our model ACP-DL on the ACP740 and ACP240 datasets to evaluate its ability of predicting anticancer peptides. The 5-fold cross-validation details are offered in Tables 1 and 2.

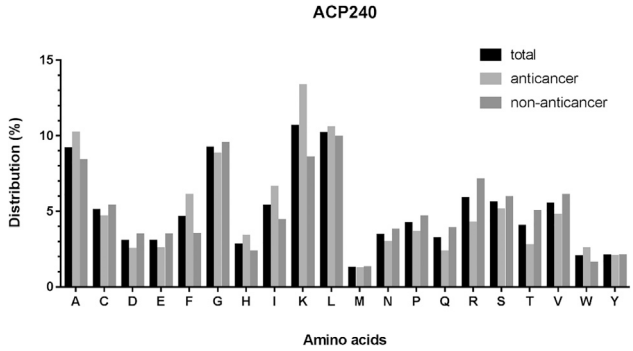

Figure 3. Comparison of Amino Acid Composition of Anticancer, Non-anticancer, and Total Peptides in Dataset ACP240

The average accuracy of 5-fold cross-validation on ACP740 was 81.48% with 3.12% SD, the average sensitivity (Sens) was 82.61% with 3.36% SD, the average specificity (Spec) was 80.59% with 4.01% SD, the mean precision (Prec) was 82.41% with 3.81% SD, and the Matthews correlation coefficient (MCC) was 63.05% with 6.23% SD. ACP-DL showed an outstanding capability to identify anticancer peptides, performed an area under the receiver operating characteristic (ROC) curve (AUC) of 0.894, as shown in Figure 4A, and has achieved the best performance on the ACP740 dataset among all comparison methods.

The mean accuracy of 5-fold cross-validation on ACP240 was 85.42%, the average Sens was 84.62%, the average Spec was 89.94%, the mean Prec was 80.28%, and the MCC was 71.44%; and, the AUC of ACP-DL was 0.906, as shown in Figure 4C. In general, the performance of the deep learning model will become better with the increase in the scale of data, and the model that can achieve good results on smaller datasets will also achieve good results on huger data. Actually, the data scale of anticancer peptides is not very large, so we didn't implement a neural network model with very complex architecture; and, to a certain extent, the 5-fold cross-validation is not conducive to the neural network model, because it further reduces the amount of training data. It is noteworthy that, although the dataset ACP240 was smaller than ACP740, our model ACP-DL still performed very well. The experimental results of rigorous cross-validation on benchmark dataset ACP740 and dataset ACP240 confirmed that our model has a good capability to predict anticancer peptides.

Comparison with Three Widely Used Machine-Learning Models

To evaluate the ability of the purposed method, we further compared ACP-DL with other widely used machine-learning models on the same benchmark datasets, including ACP740 and ACP240. Here we have selected the SVM, RF, and NB models, and we built them using the same cross-validation datasets. The implementation of these three machine-learning models comes from Scikit-learn,[48] and they were tested with default parameters. Since these methods were evaluated using the same evaluation criteria, the comparison model and deep learning model ACP-DL results are shown in Table 3 and Figures 4 and 5. ACP-DL obtained the most significant performance among the contrasted methods.

Table 1. The 5-Fold Cross-Validation Details in the ACP740 Dataset

Fold Set	Acc (%)	Sens (%)	Spec (%)	Prec (%)	MCC (%)
1	79.73	81.94	77.63	81.94	59.58
2	83.11	85.71	80.00	86.30	66.39
3	81.08	79.75	84.00	78.08	62.22
4	85.81	86.49	85.33	86.30	71.63
5	77.70	79.17	76.00	79.45	55.47
Average	81.48 ± 3.12	82.61 ± 3.36	80.59 ± 4.01	82.41 ± 3.81	63.05 ± 6.23

Table 3 shows the details of the comparison. In the ACP740 dataset, our method ACP-DL significantly outperformed other methods with an accuracy of 81.48%, a Sens of 82.61%, a Spec of 80.59%, a Prec of 82.41%, an MCC of 63.05%, and an AUC of 0.894. ACP-DL increased the accuracy by over 5%, the MCC by over 10%, and the AUC by more than 5%, respectively. In the dataset ACP240, ACP-DL also performed remarkably with an accuracy of 85.42%, a Sens of 84.62%, a Spec of 89.94%, a Prec of 80.28%, an MCC of 71.44%, and an AUC of 0.906. ACP-DL improved the accuracy by over 8%, the Spec by over 10%, the MCC by over 14%, and the AUC by more than 5%, respectively. Obviously, the deep learning model shows its power, and our model is suitable for anticancer peptide identification and prediction. ACP-DL is a competitive model used to predict anticancer peptides and accelerate related research. The comparison experiment results proved our assumption.

Conclusions

In this study, we proposed a deep learning LSTM model to predict potential anticancer peptides using high-efficiency feature representation. More specifically, we developed an efficient feature representation approach by integrating binary profile feature and *k*-mer sparse matrix of reduced amino acid alphabet feature to fully exploit peptide sequence information. Then we implemented a deep LSTM model to automatically learn how to identify anticancer peptides and non-anticancer peptides. To the best of our knowledge, this is the first time that the deep LSTM model has been applied to predict anticancer peptides.

Meanwhile, to evaluate the capability of the proposed method, we further compared ACP-DL with widely used machine-learning models in the same benchmark datasets, including ACP740 and ACP240; experimental results on the 5-fold cross-validation show that the proposed method achieved outstanding performance compared with existing methods, on benchmark dataset ACP740 with 81.48% accuracy at the AUC of 0.894 and on dataset ACP240 with an accuracy of 85.42% at the Spec of 89.94 and the AUC of 0.906, respectively. The improvement is mainly from the deep

Table 2. The 5-Fold Cross-Validation Details in the ACP240 Dataset

Fold Set	Acc (%)	Sens (%)	Spec (%)	Prec (%)	MCC (%)
1	93.75	89.66	99.99	86.36	87.99
2	81.25	77.42	92.31	68.18	63.02
3	87.50	88.46	88.46	86.36	74.83
4	83.33	90.91	76.92	90.91	67.83
5	81.25	76.67	92.00	69.57	63.53
Average	85.42	84.62	89.94	80.28	71.44

LSTM model's model parameter optimization and effective feature representation from original peptide sequences. In addition, we have contributed two novel anticancer peptide benchmark datasets, ACP740 and ACP240, in this work.

It is anticipated that ACP-DL will become a very useful high-throughput and cost-effective tool, being widely used in anticancer peptide prediction as well as cancer research. Further, as demonstrated in a series of recent publications in developing new prediction methods,[49–51] user-friendly and publicly accessible web servers will significantly enhance their impacts. It is our wish to be able to provide in the future a web server for the prediction method presented in this paper.

MATERIALS AND METHODS

In this study, we proposed a novel deep learning LSTM model to predict anticancer peptides, named ACP-DL, using high-efficiency features provided by *k*-mer sparse matrix and the binary profile feature. Furthermore, we evaluated ACP-DL's predictive performance of anticancer peptides in benchmark datasets ACP740 and ACP240. Moreover, we compared ACP-DL with three widely used machine-learning models in the same datasets, including SVM,[44] RF,[46] and NB,[47] to prove the robustness and effectiveness of the proposed method. Eventually, we made a summary, analysis, and discussion of the ACP-DL.

Construction of Datasets

We constructed two novel benchmark datasets in this work for ACP identification, named ACP740 and ACP240. As previous studies suggested, the new datasets comprised both positive and negative datasets, while positive samples were experimentally validated ACPs and AMPs without anticancer function were collected as negative samples.

The positive anticancer peptide samples can be represented as P^+, and the negative non-anticancer peptides can be represented as P^-. So, the whole dataset can be represented as P.

$$P = P^+ \cup P^- \quad \text{(Equation 1)}$$

Moreover, there is no overlap between the positive and negative datasets.

$$\varnothing = P^+ \cap P^- \quad \text{(Equation 2)}$$

Dataset ACP740

We selected 388 samples as the initial positive data on the basis of Chen et al.'s[21] and Wei et al.'s[24] studies, of which 138 were from Chen et al.'s work and 250 were from Wei et al.'s work. Correspondingly, the initial negative data were 456 samples, of which 206 were from Chen et al.'s work and 250 were from Wei et al.'s work, respectively. To avoid the bias of dataset, the widely used tool CD-HIT[52] was further used to remove those peptides sequences with a similarity of more than 90%. As a result, we finally obtained a dataset containing 740 samples, of which 376 were positive samples and 364 were negative samples.

Dataset ACP240

As the same procedure, to validate the generalization ability of the predictive model, we further constructed an additional dataset, named ACP240, which initially included 129 experimentally validated anticancer peptide samples as the positive dataset and 111 AMPs without anticancer functions as the negative dataset, respectively.

Moreover, those sequences with a similarity of more than 90% were removed using the popular tool CD-HIT.[52] The similarity setting was consistent with previous studies.[21,22] The CD-HIT is available at http://weizhong-lab.ucsd.edu/cdhit-web-server. There was no overlap between dataset ACP740 and dataset ACP240, and these two datasets are both non-redundant datasets. The two benchmark datasets are publicly available at https://github.com/haichengyi/ACP-DL.

Representation of the Peptide Sequences

A peptide sequence can be represented as follows:

$$P = p_1 p_2 p_3 p_{\ldots} p_l, \quad \text{(Equation 3)}$$

where p_1 represents the first residue in the peptide P, p_2 denotes the second residue in the peptide P, and so on; l represents the length of P. Note that the residue p_i is an element of the standard amino acid alphabet to train a machine-learning model; the first step is to convert diverse-length peptides into fixed-length feature vectors. In this study, we exploited two feature representation methods, as described below.

Binary Profile Feature (BPF)

As mentioned above, there are 20 different amino acids in the standard amino acid alphabet (A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, and Y). Each amino acid type is encoded with the following feature vector composed of 0/1. More specifically, the first amino acid type A in the alphabet is encoded as f(A) = (1,0,…,0), the second amino acid type C is encoded as f(C) = (0,1,…,0), and so on. Subsequently, for a given peptide sequence P, its N terminus with the length of k amino acids was encoded as the following feature vector:

$$\mathrm{BPF(k)} = \left[f(p_1), f(p_2), \ldots f(p_k)\right], \quad \text{(Equation 4)}$$

where k represents the length of the N terminus of the peptide P.[22] Thus, the dimension of BPF(P) is 1×20. Experiments show that the best results can be achieved when k is set to 7. So, one given peptide sequence is encoded to a 1×140 feature vector by binary profile.

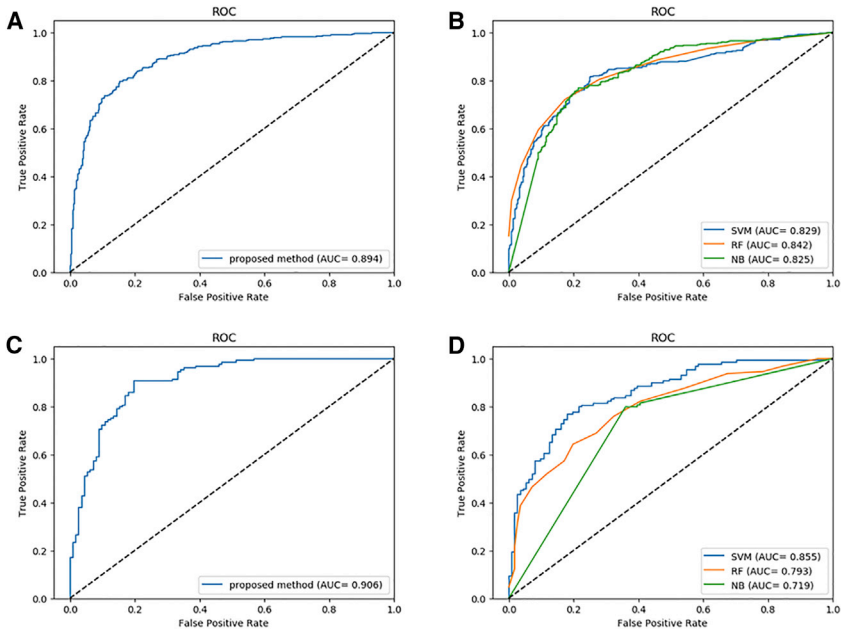

Figure 4. Performance of the Proposed Model ACP-DL and Comparison Model on Datasets ACP740 and ACP240

(A) The performance of the proposed model ACP-DL in dataset ACP740. (B) The performance of the comparison models in dataset ACP740, including SVM, RF, and NB. (C) The performance of the proposed model ACP-DL in dataset ACP240. (D) The performance of the comparison models in dataset ACP240, including SVM, RF, and NB.

K-mer Sparse Matrix

We also encoded the peptide sequence by using the *k*-mer sparse matrix previously proposed.[41] In detail, its *k*-1 consecutive nucleotides and *k* consecutive nucleotides are regarded as a unit. 3-mer of peptides is composed of 3 amino acids.[53] First the 20 amino acids were reduced into 7 groups based on their dipole moments and side chain volume: Ala, Gly, and Val; Ile, Leu, Phe, and Pro; Tyr, Met, Thr, and Ser; His, Asn, Gln, and Tpr; Arg and Lys; Asp and Glu; and Cys.[16,54,55] So, the peptide sequence was reduced into a 7-letter alphabet. Then we scanned each peptide sequence from left to right, stepping one amino acid at a time, which is considered the characteristics of each amino acid.

Suppose an above-mentioned peptide sequence length is L, there would be 7^k different possible *k*-mer and an $L-k+1$ step appearing in the RNA sequence.

One peptide sequence is transformed into a $7^k \times (L-k+1)$ *k*-mer sparse matrix M. Initialization of all elements is 0. When $m_j m_{j+1} m_{j+2}$ are just equal to the i_{th} *k*-mer among 7^k different *k*-mer, set the element $a_{ij} = 1$. The rest can be handled in the same way. Thus, an input peptide sequence is converted into a $7^k \times (L-k+1)$ matrix M.

In this study, the value of *k* is set to 3 to process the peptide sequence. The *k*-mer sparse matrix M can be defined as follows:

$$M = \left(a_{ij}\right)_{7^k} \times (L-k+1) \quad \text{(Equation 5)}$$

$$a_{ij} = \begin{cases} 1, & if\ m_j m_{j+1} m_{j+2} = k-mer(i) \\ 0, & else \end{cases}. \quad \text{(Equation 6)}$$

The *k*-mer sparse matrix M is a low-rank matrix, which almost retained all the raw information, including sequence frequency, position, and order hidden information. Then, singular value decomposition (SVD)[56] is used to reduce one two-dimensional matrix M into a 1×343 feature vector.

Finally, we conjoined two different feature representation methods' output, each peptide sequence gain 1×483 conjoined feature vector. Meanwhile, the whole dataset was transformed as a 2D matrix here. The feature matrix was reshaped into a 3D tensor for training the LSTM model, while the feature matrix without being formally reshaped was used to train the comparison model.

Deep LSTM Model Architecture

LSTM is an improvement of a recurrent neural network (RNN), which is mainly used in the natural language processing (NLP) and speech recognition field.[57–59] Different from a traditional neural network, an RNN can take advantage of sequence information. Theoretically, it can utilize the information of arbitrary length sequence; but, because of the problem of vanishing gradient in the network structure, it can only retrospectively utilize the information on time steps that are close to it in practical applications. To solve this problem, LSTM was presented with specially designed network architecture, which can learn long-term dependency information naturally. A general architecture of LSTM is composed of an input gate, a forget gate, an update gate, and a memory block. The improvement of LSTM is mainly from incorporating a memory cell that accepts the network to learn when to forget previous hidden states and when to update hidden states, according to the input information through time. It uses dedicated storage units to store information. To our knowledge, the deep LSTM model was first applied to predict novel anticancer peptides in this work.

LSTM selectively passes information through a gate unit, which mainly is by means of a sigmoid neural layer and a dot multiplication operation. Each element of the sigmoid layer output (a vector) is a real number between 0 and 1, representing the weight (or percentage) that the corresponding information passes through. For example, 0 means no information is allowed, and 1 means let all information pass.

Table 3. Actual Performance of Comparison Models and ACP-DL in the Same Dataset

Dataset	Model	Acc (%)	Sens (%)	Spec (%)	Prec (%)	MCC (%)	AUC
ACP740	SVM	64.59	62.43	90.68[a]	37.57	33.57	0.829
	RF	76.35	75.10	80.34	72.27	53.06	0.842
	NB	69.73	84.70[a]	49.21	90.94[a]	43.98	0.825
	ACP-DL	81.48[a]	82.61	80.59	82.41	63.05[a]	0.894[a]
ACP240	SVM	77.50	85.89[a]	70.68	85.65[a]	57.31	0.855
	RF	72.08	73.53	76.09	67.63	44.38	0.793
	NB	72.50	72.26	79.94	63.95	45.44	0.719
	ACP-DL	85.42[a]	84.62	89.94[a]	80.28	71.44[a]	0.906[a]

[a]This measure of performance is the best among the compared methods.

Forget Gate

In the information flow processing of LSTM, the first step is to decide what information will discord from the cell state. This decision is accomplished by a way known as forget gate. Forget gate reads h_{t-1} and x_t, then outputs a value between 0 and 1 for each digit in cell state C_{t-1}; 1 means reserved absolutely and 0 means discard completely.

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \quad \text{(Equation 7)}$$

Here, the h_{t-1} represents the output of the previous cell, x_t represents the current cell input, and σ means Sigmoid function.

Input Gate

The next step is to decide how much new information will be added to the cell state. To do this, there are two steps: first, a Sigmoid layer called the input gate layer determines which information needs to be updated; and then, a tanh layer generates a vector, which is the

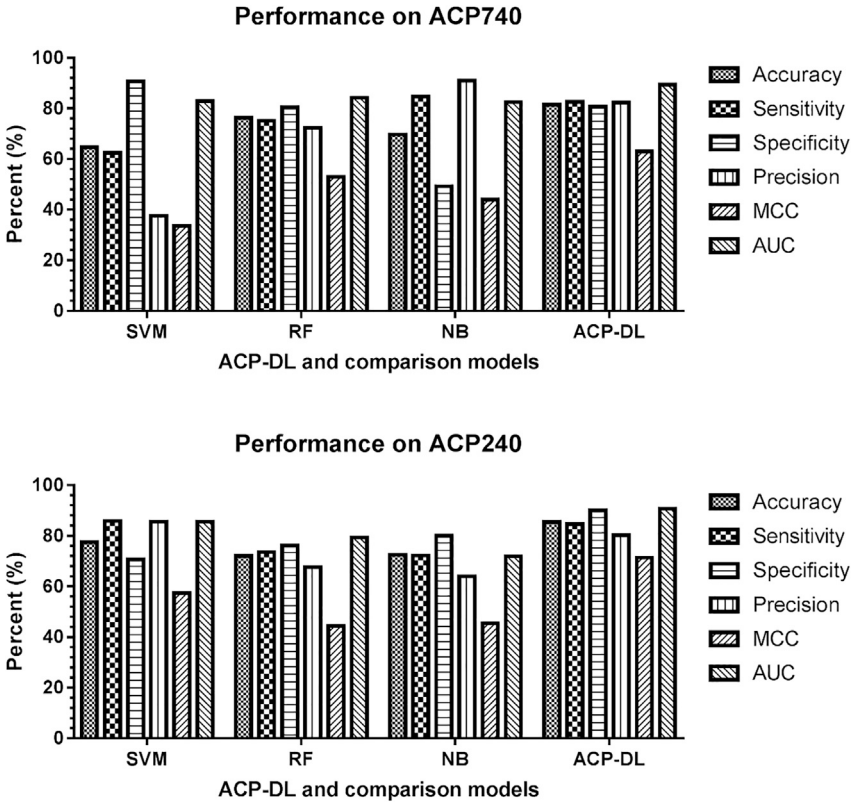

Figure 5. Comparison of SVM, Random Forest, Naive Bayes, and ACP-DL in Benchmark Datasets ACP740 and ACP240

alternate content $\tilde{C}_t$ to update. We combined the two parts to update the state of cell.

$$i_t = \sigma(W_i \bullet [h_{t-1}, x_t] + b_i) \quad \text{(Equation 8)}$$

$$\tilde{C}_t = \tanh(W_c \bullet [h_{t-1}, x_t] + b_c) \quad \text{(Equation 9)}$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad \text{(Equation 10)}$$

We multiply the old state with f_t and discard the information we need to discard. Then we add $i_t * \tilde{C}_t$. This is the new candidate value, which is changed according to the degree of each state we decide to update.

Output Gate

Ultimately, we need to determine what output is. This output will be based on our cell state, but it is also a filtered version. First, we run a sigmoid layer to determine which part of the cell state will be exported. Then, we process the cell state through a tanh function (to get a value between −1 and 1) and multiply it with the output of the Sigmoid gate, and eventually we just output the portion of the output we determine.

$$o_t = \sigma(W_o \bullet [h_{t-1}, x_t] + b_o) \quad \text{(Equation 11)}$$

$$h_t = o_t * \tanh(C_t) \quad \text{(Equation 12)}$$

In this experiment, considering the limited scale of anticancer peptide samples, we set the parameter of LSTM input layer to 128, and the output of LSTM layers was fed into a dense layer (a fully connected neural network layer) as input to obtain a final prediction result. We also used a sigmoid function as an activation function in the proposed model. The mathematical behaviors of a sigmoid function can be demonstrated as follows:

$$\sigma = \text{sigmoid}(x) = \frac{1}{(1 + e^{-x})}. \quad \text{(Equation 13)}$$

Between them, the dropout layer was applied to reduce over-fitting and enhance the neural network model robustness, and the parameter *dropout* was set to 0.25. Moreover, a loss function can measure the performance of machine-learning models. We selected to use log loss function (binary cross-entropy) corresponding to sigmoid function as loss function, which can be defined as:

$$\text{logloss}(t, p) = -((1 - p) \times \log(1 - p) + t \times \log(p)), \quad \text{(Equation 14)}$$

where p and t represent the prediction output of model and true target value, respectively. Finally, the Adam[60] optimizer was used to update the weights of network iteratively, which is popular in the deep learning field and combined the advantage of root-mean-square propagation (RMSProp) and adaptive gradient (AdaGrad) algorithm.

The implementation of the deep learning model is based on the Keras framework, which is capable of running on top of TensorFlow, Theano, or CNTK and is supported on both GPUs and CPUs. It was developed with a focus on enabling fast experimentation.[61]

Performance Evaluation Criteria

In this study, we proposed a novel deep learning LSTM model, ACP-DL, using an efficiency feature to predict potential anticancer peptides. We used 5-fold cross-validation to evaluate the performance of ACP-DL and comparison models. In each validation, all data randomly divide into five equal parts: the 4-fold set data are taken as training data, and the remaining 1-fold data are taken as test data. To guarantee the unbiased comparison, it was confirmed that there was no overlap between training data and test data. The final validation result was the average of 5-fold with SDs. We followed the widely used evaluation criteria,[62,63] including accuracy (Acc), Sens or recall, Spec, Prec, and MCC, defined as follows:

$$\text{Acc} = \frac{TN + TP}{TN + TP + FN + FP} \quad \text{(Equation 15)}$$

$$\text{Sens} = \frac{TP}{TP + FN} \quad \text{(Equation 16)}$$

$$\text{Spec} = \frac{TN}{TN + FP} \quad \text{(Equation 17)}$$

$$\text{Prec} = \frac{TP}{TP + FP} \quad \text{(Equation 18)}$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}, \quad \text{(Equation 19)}$$

where TN indicates the true negative number, TP denotes the true positive number, FN represents the false negative number, and FP stands for the false positive number. Certainly, the ROC curve and the AUC were also adopted to evaluate the performance.

AUTHOR CONTRIBUTIONS

H.-C.Y. and Z.-H.Y. conceived the algorithm, carried out analyses, prepared the datasets, carried out experiments, and wrote the manuscript. Other authors designed, performed, and analyzed experiments and wrote the manuscript. All authors read and approved the final manuscript.

CONFLICTS OF INTEREST

The authors declare no competing interests.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under grants 61572506 and 61722212 and in part by the Pioneer Hundred Talents Program of Chinese Academy of Sciences.

REFERENCES

1. Siegel, R.L., Miller, K.D., and Jemal, A. (2018). Cancer statistics, 2018. CA Cancer J. Clin. *68*, 7–30.
2. Ferlay, J., Shin, H.R., Bray, F., Forman, D., Mathers, C., and Parkin, D.M. (2010). Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. Int. J. Cancer *127*, 2893–2917.
3. Holohan, C., Van Schaeybroeck, S., Longley, D.B., and Johnston, P.G. (2013). Cancer drug resistance: an evolving paradigm. Nat. Rev. Cancer *13*, 714–726.
4. Hoskin, D.W., and Ramamoorthy, A. (2008). Studies on Anticancer Activities of Antimicrobial Peptides. Biochim. Biophys. Acta *1778*, 357–375.
5. Tyagi, A., Tuknait, A., Anand, P., Gupta, S., Sharma, M., Mathur, D., Joshi, A., Singh, S., Gautam, A., and Raghava, G.P. (2015). CancerPPD: a database of anticancer peptides and proteins. Nucleic Acids Res. *43*, D837–D843.
6. Gaspar, D., Veiga, A.S., and Castanho, M.A.R.B. (2013). From antimicrobial to anticancer peptides. A review. Front. Microbiol. *4*, 294.
7. Huang, Y., Feng, Q., Yan, Q., Hao, X., and Chen, Y. (2015). Alpha-helical cationic anticancer peptides: a promising candidate for novel anticancer drugs. Mini Rev. Med. Chem. *15*, 73–81.
8. Otvos, L., Jr. (2008). Peptide-based drug design: here and now. Methods Mol. Biol. *494*, 1–8.
9. Mader, J.S., and Hoskin, D.W. (2006). Cationic antimicrobial peptides as novel cytotoxic agents for cancer treatment. Expert Opin. Investig. Drugs *15*, 933–946.
10. Hariharan, S., Gustafson, D., Holden, S., McConkey, D., Davis, D., Morrow, M., Basche, M., Gore, L., Zang, C., O'Bryant, C.L., et al. (2007). Assessment of the biological and pharmacological effects of the alpha nu beta3 and alpha nu beta5 integrin receptor antagonist, cilengitide (EMD 121974), in patients with advanced solid tumors. Ann. Oncol. *18*, 1400–1407.
11. Gregorc, V., De Braud, F.G., De Pas, T.M., Scalamogna, R., Citterio, G., Milani, A., Boselli, S., Catania, C., Donadoni, G., Rossoni, G., et al. (2011). Phase I Study of NGR-hTNF, a Selective Vascular Targeting Agent, in Combination with Cisplatin in Refractory Solid Tumors. Clin. Cancer Res. *17*, 1964–1972.
12. Barras, D., and Widmann, C. (2011). Promises of apoptosis-inducing peptides in cancer therapeutics. Curr. Pharm. Biotechnol. *12*, 1153–1165.
13. Boohaker, R.J., Lee, M.W., Vishnubhotla, P., Perez, J.M., and Khaled, A.R. (2012). The use of therapeutic peptides to target and to kill cancer cells. Curr. Med. Chem. *19*, 3794–3804.
14. Thundimadathil, J. (2012). Cancer Treatment Using Peptides: Current Therapies and Future Prospects. J. Amino Acids *2012*, 967347.
15. Su, R., Liu, X., Wei, L., and Zou, Q. (2019). Deep-Resp-Forest: A deep forest model to predict anti-cancer drug response. Methods. Published online February 14, 2019. https://doi.org/10.1016/j.ymeth.2019.02.009.
16. Tyagi, A., Kapoor, P., Kumar, R., Chaudhary, K., Gautam, A., and Raghava, G.P. (2013). In silico models for designing and discovering novel anticancer peptides. Sci. Rep. *3*, 2984.
17. Hajisharifi, Z., Piryaiee, M., Mohammad Beigi, M., Behbahani, M., and Mohabatkar, H. (2014). Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test. J. Theor. Biol. *341*, 34–40.
18. Chou, K.C. (2005). Using Amphiphilic Pseudo Amino Acid Composition to Predict Enzyme Subfamily Classes (Oxford University Press).
19. Shen, H.B., and Chou, K.C. (2007). Using ensemble classifier to identify membrane protein types. Amino Acids *32*, 483–488.
20. Vijayakumar, S., and Ptv, L. (2015). ACPP: A Web Server for Prediction and Design of Anti-cancer Peptides. Int. J. Pept. Res. Ther. *21*, 99–106.
21. Chen, W., Ding, H., Feng, P., Lin, H., and Chou, K.C. (2016). iACP: a sequence-based tool for identifying anticancer peptides. Oncotarget *7*, 16895–16909.
22. Wei, L., Zhou, C., Chen, H., Song, J., and Su, R. (2018). ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides. Bioinformatics *34*, 4007–4016.
23. Du, P., and Wang, L. (2014). Predicting human protein subcellular locations by the ensemble of multiple predictors via protein-protein interaction network with edge clustering coefficients. PLoS ONE *9*, e86879.
24. Wei, L., Ding, Y., Su, R., Tang, J., and Zou, Q. (2018). Prediction of human protein subcellular localization using deep learning. J. Parallel Distrib. Comput. *117*, 212–217.
25. Zou, Q., Xing, P., Wei, L., and Liu, B. (2019). Gene2vec: gene subsequence embedding for prediction of mammalian N^6-methyladenosine sites from mRNA. RNA *25*, 205–218.
26. Zhang, J., Ju, Y., Lu, H., Xuan, P., and Zou, Q. (2016). Accurate identification of cancerlectins through hybrid machine learning technology. Int. J. Genomics *2016*, 7604641.
27. Chen, W., Feng, P., Yang, H., Ding, H., Lin, H., and Chou, K.-C. (2018). iRNA-3typeA: identifying three types of modification at RNA's adenosine sites. Mol. Ther. Nucleic Acids *11*, 468–474.
28. Chen, W., Tang, H., Ye, J., Lin, H., and Chou, K.-C. (2016). iRNA-PseU: Identifying RNA pseudouridine sites. Mol. Ther. Nucleic Acids *5*, e332.
29. Jiao, Y., and Du, P. (2016). Performance measures in evaluating machine learning based bioinformatics predictors for classifications. Quant. Biol. *4*, 320–330.
30. Du, P., Tian, Y., and Yan, Y. (2012). Subcellular localization prediction for human internal and organelle membrane proteins with projected gene ontology scores. J. Theor. Biol. *313*, 61–67.
31. Wang, L., You, Z.-H., Chen, X., Li, Y.-M., Dong, Y.-N., Li, L.-P., and Zheng, K. (2019). LMTRDA: Using logistic model tree to predict MiRNA-disease associations by fusing multi-source information of sequences and similarities. PLoS Comput. Biol. *15*, e1006865.
32. Wang, Y., You, Z., Li, L., Cheng, L., Zhou, X., Zhang, L., Li, X., and Jiang, T. (2018). Predicting Protein Interactions Using a Deep Learning Method-Stacked Sparse Autoencoder Combined with a Probabilistic Classification Vector Machine. Complexity *2018*, 4216813.
33. Wang, Y.-B., You, Z.-H., Li, L.-P., Huang, Y.-A., and Yi, H.-C. (2017). Detection of interactions between proteins by using legendre moments descriptor to extract discriminatory information embedded in pssm. Molecules *22*, 1366.
34. Wang, L., You, Z.-H., Huang, D.S., and Zhou, F. (2018). Combining High Speed ELM Learning with a Deep Convolutional Neural Network Feature Encoding for Predicting Protein-RNA Interactions. IEEE/ACM Trans. Comput. Biol. Bioinformatics. Published online October 5, 2018. https://doi.org/10.1109/TCBB.2018.2874267.
35. You, Z.-H., Lei, Y.-K., Gui, J., Huang, D.-S., and Zhou, X. (2010). Using manifold embedding for assessing and predicting protein interactions from high-throughput experimental data. Bioinformatics *26*, 2744–2751.
36. Li, S., You, Z.-H., Guo, H., Luo, X., and Zhao, Z.-Q. (2016). Inverse-free extreme learning machine with optimal information updating. IEEE Trans. Cybern. *46*, 1229–1241.
37. You, Z.-H., Huang, Z.-A., Zhu, Z., Yan, G.-Y., Li, Z.-W., Wen, Z., and Chen, X. (2017). PBMDA: A novel and effective path-based computational model for miRNA-disease association prediction. PLoS Comput. Biol. *13*, e1005455.
38. You, Z.-H., Yin, Z., Han, K., Huang, D.-S., and Zhou, X. (2010). A semi-supervised learning approach to predict synthetic genetic interactions by combining functional and topological properties of functional gene network. BMC Bioinformatics *11*, 343.
39. Zhu, L., You, Z.-H., and Huang, D.-S. (2013). Increasing the reliability of protein–protein interaction networks via non-convex semantic embedding. Neurocomputing *121*, 99–107.
40. Chen, Z.-H., You, Z.-H., Li, L.-P., Wang, Y.-B., Wong, L., and Yi, H.-C. (2019). Prediction of Self-Interacting Proteins from Protein Sequence Information Based on Random Projection Model and Fast Fourier Transform. Int. J. Mol. Sci. *20*, 930.
41. You, Z.H., Zhou, M., Luo, X., and Li, S. (2016). Highly Efficient Framework for Predicting Interactions Between Proteins. IEEE Trans. Cybern. *47*, 731–743.
42. Yi, H.-C., You, Z.-H., Huang, D.-S., Li, X., Jiang, T.-H., and Li, L.-P. (2018). A Deep Learning Framework for Robust and Accurate Prediction of ncRNA-Protein Interactions Using Evolutionary Information. Mol. Ther. Nucleic Acids *11*, 337–344.
43. Gautam, A., Chaudhary, K., Kumar, R., Sharma, A., Kapoor, P., Tyagi, A., and Raghava, G.P.; Open source drug discovery consortium (2013). In silico approaches for designing highly effective cell penetrating peptides. J. Transl. Med. *11*, 74.
44. Vapnik, V.N. (1998). Statistical Learning Theory, First Edition (Wiley).

45. Chang, C.-C., and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. ACM Trans. Intell. Syst. Technol. *2*, 1–27.

46. Breiman, L. (2001). Random Forest. Mach. Learn. *45*, 5–32.

47. Zhang, H. (2004). The optimality of naive Bayes. In Proceedings of the 17th International Florida Artificial Intelligence Research Society Conference, V. Barr and Z. Markov, eds. (American Association for Artificial Intelligence), pp. 562–567.

48. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

49. Le, N.-Q.-K., Ho, Q.-T., and Ou, Y.-Y. (2017). Incorporating deep learning with convolutional neural networks and position specific scoring matrices for identifying electron transport proteins. J. Comput. Chem. *38*, 2000–2006.

50. Le, N.-Q.-K., Ho, Q.-T., and Ou, Y.-Y. (2018). Classifying the molecular functions of Rab GTPases in membrane trafficking using deep convolutional neural networks. Anal. Biochem. *555*, 33–41.

51. Le, N.Q.K., Yapp, E.K.Y., Ho, Q.-T., Nagasundaram, N., Ou, Y.-Y., and Yeh, H.-Y. (2019). iEnhancer-5Step: Identifying enhancers using hidden information of DNA sequences via Chou's 5-step rule and word embedding. Anal. Biochem. *571*, 53–61.

52. Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics *22*, 1658–1659.

53. Muppirala, U.K., Honavar, V.G., and Dobbs, D. (2011). Predicting RNA-protein interactions using only sequence information. BMC Bioinformatics *12*, 489.

54. Suresh, V., Liu, L., Adjeroh, D., and Zhou, X. (2015). RPI-Pred: predicting ncRNA-protein interaction using sequence and structural information. Nucleic Acids Res. *43*, 1370–1379.

55. Pan, X., Fan, Y.X., Yan, J., and Shen, H.B. (2016). IPMiner: hidden ncRNA-protein interaction sequential pattern mining with stacked autoencoder for accurate computational prediction. BMC Genomics *17*, 582.

56. Kolda, T.G., and O'Leary, D.P. (1998). A Semidiscrete Matrix Decomposition for Latent Semantic Indexing in Information Retrieval. ACM Trans. Inf. Syst. *16*, 322–346.

57. Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. Neural Comput. *9*, 1735–1780.

58. Gers, F.A., Schmidhuber, J., and Cummins, F. (2000). Learning to forget: continual prediction with LSTM. Neural Comput. *12*, 2451–2471.

59. Sundermeyer, M., Schlüter, R., and Ney, H. (2012). LSTM Neural Networks for Language Modeling (Interspeech), pp. 601–608.

60. Kingma, D.P., and Ba, J. (2014). Adam: A Method for Stochastic Optimization. arXiv, arXiv:1412.6980, https://arxiv.org/abs/1412.6980.

61. Chollet, F. (2018). Keras: The Python Deep Learning Library (Astrophysics Source Code Library).

62. Le, N.-Q.-K., and Ou, Y.-Y. (2016). Prediction of FAD binding sites in electron transport proteins according to efficient radial basis function networks and significant amino acid pairs. BMC Bioinformatics *17*, 298.

63. Le, N.-Q.-K., and Ou, Y.-Y. (2016). Incorporating efficient radial basis function networks and significant amino acid pairs for predicting GTP binding sites in transport proteins. BMC Bioinformatics *17* (*Suppl 19*), 501.