2022 21st International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

Identifying anti-tumor heat shock proteins based on evolutionary information using deep learning method

Yi Fu[1,2], Ji Zhao[1,2], Juan Mei[1,2], Yi Ding[1,2]
1. School of Internet of Things Engineering, Wuxi City College of Vocational Technology, Wuxi, China
2. Wuxi Research Center for Environmental Science & Engineering, Wuxi, China
Email:fuyi@wxcu.edu.cn

***Abstract*—Heat shock proteins (HSPs) belong to stress proteins. The functions of HSPs are mainly reflected in three aspects: molecular chaperones, regulation of apoptosis and immune responses. Recent studies have shown that there is a certain correlation between HSPs and tumor cell. HSPs are participated in the invasion, proliferation and metastasis of tumor cells. Therefore, developing an accurate model for identification anti-tumor HSPs is a key step to understand molecular functions of HSPs and human tumor diseases. In this study, we propose using deep learning methods to identify anti-tumor HSPs. To seek out the optimal model for the dataset, several hyper-parameters are optimized according to the results of 10-fold cross-validation. Finally, the performance of the proposed model is further determined through an independent dataset. The experimental results indicated that the proposed model could classify anti-tumor HSPs with accuracy (ACC) of 93.76%, sensitivity (SN) of 92.80%, specificity (SP) of 93.33%, and Matthew's correlation coefficient (MCC) of 86.39% on the 10-fold cross-validation. Compared with other deep learing methods, using convolutional neural network (CNN) can achieve a significant improvement for identifying of anti-tumor HSPs.**

Keywords-HSPs; anti-tumor function; position specific scoring matrix; convolutional neural network

I. INTRODUCTION

Heat shock proteins (HSPs) belong to stress-induced proteins. HSPs can inhibit or reverse the denaturation of cellular proteins under extreme conditions. They are widely found in prokaryotic and eukaryotic cells. Because of their physiological and protective effects in cells, HSPs are also known as molecular chaperones [1]. In addition to being a molecular chaperone, HSPs also have the function of regulating cell apoptosis and participating in body immunity. Recent studies have found that HSPs are overexpressed in various tumor cells and participated in the invasion, proliferation, differentiation and metastasis of tumor cells. In clinic, HSPs can be used as biomarkers for cancer diagnosis, to track disease development or response to treatment, and also as therapeutic targets for cancer treatment [2].

According to molecular weight, HSPs are generally categorized into six major families: HSP27, HSP40, HSP60, HSP70, HSP 90 and HSP110 [3]. Because HSPs play a significant role in regulating cellular function and response during tumor development and metastasis, many researchers have studied the relationship between heat shock proteins and tumors. Regimbeau et al. described the function of HSPs from the aspects of anticancer therapeutics targets, disease monitoring biomarkers and cancer vaccines [1]. Albakova et al. outlined the relationship between HSPs and cancer and proposed the model for identification of HSPs in cancer [2]. Wu et al. elucidated the function of HSPs in pharmacology and cancer biology [3]. Zheng et al. studied the tumor immune effect of heat shock protein gp96 [4]. Elmallah et al. discussed the function of Hsp70 in cancer, especially focusing on the extracellular membrane-bound Hsp70 [5].

In the past decade, many computational methods have been developed for predicting HSPs. They commonly focused on two aspects, one is to identify heat shock proteins or not, the other is to classify HSP sequences into one of the six HSP families. In 2013, Feng et al. first identified the heat shock protein families using support vector machine and jackknife test [6]. Subsequently, Ahmad et al. studied the classification and prediction of heat shock protein family based on a variety of machine learning algorithms and multiple features [7]. In 2020, Jing et al. classified heat shock protein family using SVM algorithm and combined features [8]. Recently, Min et al. proposed a novel deep learning algorithm DeepHSP to classify both non-HSPs and six HSP families simultaneously [9]. With the successful application of deep neural network in many fields, researchers began to use it in biomedicine, such as protein secondary structure prediction, cancer prediction and drug design, and so on. These studies have achieved good performances.

HSPs play an important regulatory role in the development, diagnosis, and treatment of tumor. To our knowledge, there are no reports on the identification of anti-tumor heat shock proteins. Based on the above reasons, it prompts us to establish an accurate model to identify anti-tumor HSPs. In this study, we proposed a deep neural network model based on evolutionary information feature extraction method to classify anti-tumor HSPs. The experimental results indicated that our model could identify anti-tumor HSPs with ACC value of 93.76%, SN value of 92.80%, SP value of 93.33%, and MCC value of 86.39% on the 10-fold cross-validation dataset. Compared with other prediction methods, using CNN deep learning network can achieve a significant improvement for identifying of anti-tumor HSPs. The accurate identification of anti-tumor HSPs could assist people in comprehending the role of anti-tumor HSPs in tumor related diseases and designing drugs for promoting treatment.

2473-3636/22/$31.00 ©2022 IEEE
DOI 10.1109/DCABES57229.2022.00038

II. MATERIALS AND METHODS

A. Dataset

In this study, all the sequences come from UniProt database (https://www.uniprot.org). The proposed problem is to classify the HSPs with anti-tumor function and those without anti-tumor function. Therefore, we use the HSPs dataset with anti-tumor function as positive data and a dataset of HSPs without anti-tumor function as negative data. The obtained dataset was screened from the following two aspects: (1) the amino acid sequences with less than 50 residues were removed; (2) non-standard letters containing "B", "J", "O", "X" or "Z" were deleted. CD-HIT was applied to decrease the sequence homology, here sequence similarity was set to 50%. After this process, the dataset includes 290 HSPs with anti-tumor function and 502 HSPs without anti-tumor function. Dataset can be formulated as follows:

$$Set = Set^{+} \bigcup Set^{-} \quad (1)$$

Set is constituted with Set^{+} and Set^{-}. Set^{+} indicates the positive dataset (the 290 anti-tumor HSPs), Set^{-} indicates negative dataset (the 502 non-anti-tumor HSPs).

The dataset was randomly divided into the cross-validation dataset and the independent dataset. There were 200 anti-tumor HSPs and 300 non-anti-tumor HSPs in the cross-validation dataset. Remaining protein sequences were used as the independent dataset, there were 90 anti-tumor HSPs and 202 non-anti-tumor HSPs.

B. Deep neural network structure

Deep learning is considered a powerful method, which can automatically learn features in the neural network model. In the model, abstract high-level features are composed of low-level features, which are more suitable for discovering the distributed feature of data. In this study, we construct a deep learning framework based on convolutional neural network to accurately predict the anti-tumor HSPs. CNN model has been proved to be effective in many fields, such as in the computer vision domain, natural language processing systems and biomedical field [10], and so on.

In our model, CNN is composed of an input layer, multiple hidden layers and an output layer. It takes the numerical vector transformed by PSSM algorithm as input to the network. Our model includes two convolution layers, the number of filters in each layer are set to 64 and 128 respectively. The size of convolution kernel is set to 3. Padding is set as same to keep the dimension. And activation function in the convolution layer adopts a rectified linear unit (ReLU) for the normalized results. In order to prevent over-fitting, we add a dropout layer to the model to enhance the robustness of the neural network model.

$$Conv(R)_{jf} = \mathrm{Re}LU\left(\sum_{s=0}^{S-1}\sum_{n=0}^{N-1} W_{sn}^{f} R_{j+s,n}\right) \quad (2)$$

$$\mathrm{Re}LU(z) = \max(0, z) \quad (3)$$

Also note, we have added another three fully connected layers, which have 128, 64 and 32 neurons, respectively. Here ReLU is used as the activation function to classify in the model construction process. We use softmax function as the activation function of the network output layer. The softmax function can normalize the output value and convert the output results into the form of probability. The loss function in the output layer uses the binary cross entropy function, which can predict the difference between the real value and the predicted value, and judge the quality of the prediction models through the loss value. The formula for calculation is as follows:

$$\log loss(t, p) = -((1-p)\times\log(1-p) + t\times\log(p)) \quad (4)$$

The prediction framework of anti-tumor HSPs is shown in FIG.1. Our model is implemented in Keras framework with a Tensorflow backend.

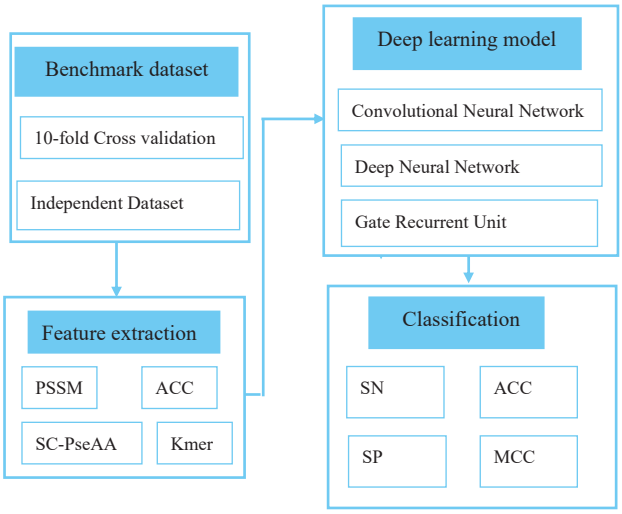

Figure 1. Overall framework for identifying anti-tumor HSPs.

C. Performance evaluation

To evaluate the prediction performance of the model, three statistical methods are generally used in the literature, such as K-fold cross-validation, independent test, and jackknife validation. In this study, we applied 10-fold cross-validation method for model optimization to find the optimal parameters, then adopted independent test to verify the reliability of model predictions. We employed the following evaluation indicators, namely accuracy (ACC), sensitivity (SN), specificity (SP) and Matthew's correlation coefficient (MCC), to appraise the generalization abilities of the prediction model. These indicators are calculated as follows:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

$$SN = \frac{TP}{TP + FN} \quad (6)$$

$$SP = \frac{TN}{TN + FP} \quad (7)$$

$$MCC = \frac{TP\times TN - FP\times FN}{\sqrt{(TP+FN)(TP+FP)(TN+FN)(TN+FP)}} \quad (8)$$

where TN is true negatives, TP is true positives, FN is false negatives, and FP is false positives.

III. EXPERIMENTAL RESULTS AND ANALYSIS

A. Hyperparameter selection

The proposed deep learning model for anti-tumor HSPs prediction contains several hyperparameters, such as the batch size, number of epochs and dropout value, and so on. In order to obtain the model with the best prediction results, we performed 10-fold cross-validation for hyperparameter tuning.

TABLE 1. Several Parameters of convolutional neural network and recommended values.

Parameter	Value
Batch size	10
Epoch	150
Optimizer	Adam
Activation function	Relu
Dropout rate	0.1

TAB.1 lists several central parameters of convolutional neural network. Among the four optimizers in Keras, Adam, Adadelta, Adagrad and Rmsprop were used. As shown in FIG. 2, Adam has yielded superior performance. FIG. 3 shown the model accuracy and error loss curves of Adam optimizer. We selected Adam as an optimizer to update the weight and optimize the convolutional neural network.

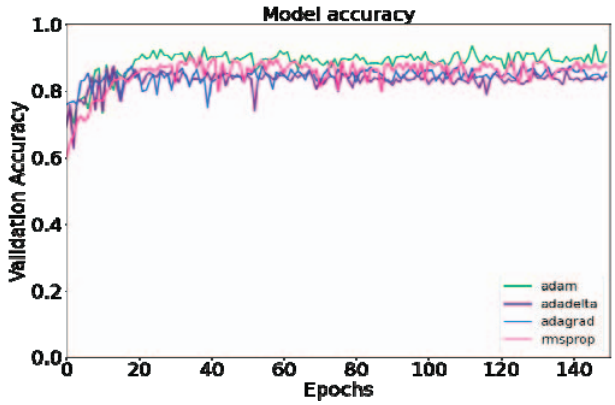

Figure 2 Model accuracy curves corresponding to four different optimizers.

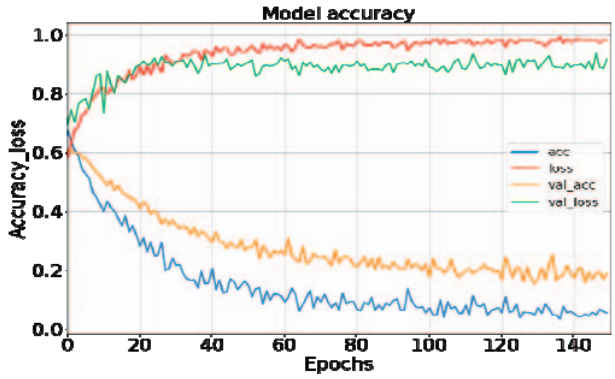

Figure 3 Model accuracy and error rate using Adam optimizer.

B. Prediction performance of different feature extraction algorithm

Feature extraction refers to converting the original data into a new set of feature vectors through data transformation or data mapping, so as to find effective features for the prediction model. It directly affects the result of model prediction. Protein sequence character information is transformed into numerical vector information by feature extraction method, and then classified and predicted by classification method. In the paper, four feature extraction methods (kmer, ACC, SC-PseAAC and PSSM) were used to extract protein sequence features. Among the four feature extraction methods, Kmer, ACC and SC-PseAAC are generated from the web server called 'Pse-in-One 2.0'. PSSM is obtained by PSI-BLAST tool. The maximum number of iterations is set to 3, and the threshold value of E is set to 0.001.

TABLE 2. Comparison of prediction results obtained with different feature extraction methods on the 10-fold cross-validation.

Feature extraction	ACC (%)	SN (%)	SP (%)	MCC (%)
ACC	90.40	89.00	92.73	85.51
Kmer	88.93	83.50	92.00	79.05
SC-PseAAC	89.60	86.50	91.67	78.45
PSSM	**93.83**	**92.00**	**94.33**	**87.40**

As shown in TAB.2, for the 10-fold cross-validation dataset, the performance of the PSSM feature extraction algorithm is the best. The PSSM method achieved an ACC of 93.83%, a SN of 92.00%, a SP of 94.33%, and an MCC of 87.40%. The better values in the TAB.2 are shown in bold face. The ACC of PSSM is 93.83%, which is 3.43%, 4.90%, and 4.23% higher than that of ACC, Kmer and SC-PseAAC, respectively. In addition to PSSM achieving good performance, the second is feature extraction method ACC. On the independent dataset, we could notice that performance of PSSM algorithm is also the best, far exceeding ACC, Kmer and SC-PseAAC methods.

C. Performance comparisons with other classification tools

To evaluate the model performance, we compared the CNN model with deep neural network (DNN), and gated recurrent unit (GRU), on the same datasets. TAB.3 displays the performances of three models on the 10-fold cross-validation, which concludes that the CNN model has a better performance than DNN and GRU. In the 10-fold cross-validation dataset, CNN achieved 92.80% value of SN, 93.33% value of SP, 93.76% value of ACC and 86.39% value of MCC, respectively. Compared to the other two models, CNN shared the highest SN, ACC and MCC values on 10-fold cross-validation dataset. As shown in TAB.3, GRU scored the highest SP value. However, its corresponding SN value is lower than the CNN model, since GRU predicts too many positive samples as negative.

As shown in FIG.4, the ACC, SN, and MCC values of CNN were also higher than the corresponding values measured for DNN and GRU on the independent dataset. The ACC, SN, and MCC values of CNN were 86.27%, 91.11%, and 79.56%, respectively, higher than that of DNN and GRU on the independent dataset. From the above results, it is clear that proposed CNN model has the excellent performance compared with other deep learning models.

TABLE 3. Prediction results obtained with different classification methods on the 10-fold cross-validation.

Classification method	ACC (%)	SN (%)	SP (%)	MCC (%)
CNN	**93.76**	**92.80**	93.33	**86.39**
DNN	76.40	78.00	88.67	74.06
GRU	80.60	82.50	**98.67**	79.53

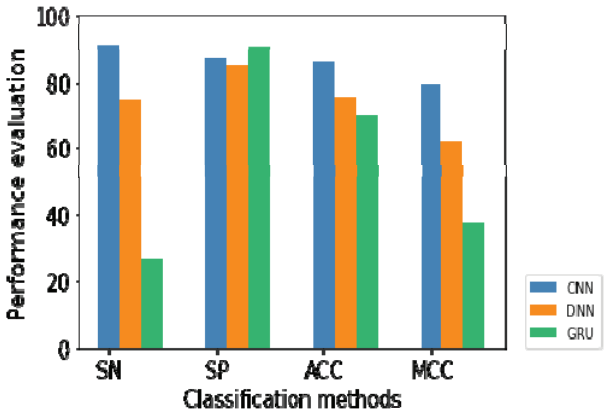

Figure 4. Prediction results obtained with different classification methods on the independent test.

IV. CONCLUSION

In this study, we have presented an effective prediction method for the discrimination of anti-tumor HSPs through convolutional neural network and PSSM feature extraction method, which enhanced prediction performance. Compared with other DL methods, our model has obtained significant improvements according to evaluation indicators. The superior performance of the constructed model for anti-tumor HSPs identification is due to several reasons as follows: (1) optimize the most critical hyperparameters. (2) extract protein sequence features using PSSM algorithm to obtain the best optimized feature set. (3) construct a deep neural network framework to effectively learn vital protein features. Constructing accurate classification model of anti-tumor HSPs is extremely important for understanding molecular functions of HSPs and designing drugs for tumor related diseases. At the same time, this study is helpful to promote the further research and application of deep learning in biomedical field.

ACKNOWLEDGMENT

This work is supported by the Qing Lan Project of Jiangsu Province (Year 2020 Young and Mid-aged Academic Leaders), the Jiangsu Overseas Visiting Scholar Program for University Prominent Young and Mid-aged Teachers and Presidents (Year 2019).

REFERENCES

[1] M. Regimbeau, J. Abrey, V. Vautrot, S. Causse, J. Gobbo,and C. Garrodo. Heat shock proteins and exosomes in cancer theranostics. Seminars in Cancer Biology, 2021, 12pages.

[2] Z. Albakova, M. K. S. Siam, P. K. Sacitharan, R. H. Ziganshin , D. Y. Ryazantsev, and A. M. Sapozhnikov. Extracellular heat shock proteins and cancer: New perspectives. Translational Oncology, 2021, 14:100995.

[3] J. Wu, T. Liu, Z. Rios, Q. Mei, X. Lin, and S. Cao. Heat shock proteins and cancer. Trends in Pharmacological sciences, 2017, 38(3):226-256.

[4] H. Zheng, L. Liu , H. Zhang, et al. Dendritic cells pulsed with placental gp96 promote tumor-reactive immune responses. PLoS One, 2019, 14(1):e0211490.

[5] M.I.Y. Elmallah, M. Cordonnier, V. Vautrot, G. Chanteloup, C. Garrido, and J. Gobbo. Membrane-anchored heat-shock protein 70 (Hsp70) in cancer. Cancer Letters, 2020, 469:134-141.

[6] P. Feng, W. Chen, H. Lin, K. Chou. iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. Analytical biochemistry, 2013, 442(1):118-125.

[7] S. Ahmad, M. Kabir, and M. Hayat. Identification of heat shock protein families and J-protein types by incorporating dipeptide composition into Chou's general PseAAC. Computer Methods and Programs in Biomedicine, 2015, 122(2):165-174.

[8] X. Jing, and F. Li. Identifying heat shock protein families from imbalanced data by using combined features. Computational and mathematical methods in medicine, 2020, Article ID 8894478, 11 pages.

[9] S. Min, H. Kim, B. Lee, and S. Yoon. Protein transfer learning improves identification of heat shock protein families. PLoS ONE, 2021,16(5):e0251865.

[10] Y. Lecun, Y. Bengio, and G. Hinton. Deep Learning. Nature, 2015, 521:436-444.