BMC Bioinformatics

RESEARCH

Open Access

pSuc-EDBAM: Predicting lysine succinylation sites in proteins based on ensemble dense blocks and an attention module

Jianhua Jia[1*], Genqiang Wu[1], Meifang Li[2] and Wangren Qiu[1]

*Correspondence:
jjh163yx@163.com

[1] Computer Department, Jingdezhen Ceramic University, Jingdezhen 333403, China
[2] Computer Department, Nanchang Institute of Technology, Nanchang 330044, China

Abstract

Background: Lysine succinylation is a newly discovered protein post-translational modifications. Predicting succinylation sites helps investigate the metabolic disease treatments. However, the biological experimental approaches are costly and inefficient, it is necessary to develop efficient computational approaches.

Results: In this paper, we proposed a novel predictor based on ensemble dense blocks and an attention module, called as pSuc-EDBAM, which adopted one hot encoding to derive the feature maps of protein sequences, and generated the low-level feature maps through 1-D CNN. Afterward, the ensemble dense blocks were used to capture feature information at different levels in the process of feature learning. We also introduced an attention module to evaluate the importance degrees of different features. The experimental results show that Acc reaches 74.25%, and MCC reaches 0.2927 on the testing dataset, which suggest that the pSuc-EDBAM outperforms the existing predictors.

Conclusions: The experimental results of ten-fold cross-validation on the training dataset and independent test on the testing dataset showed that pSuc-EDBAM outperforms the existing succinylation site predictors and can predict potential succinylation sites effectively. The pSuc-EDBAM is feasible and obtains the credible predictive results, which may also provide valuable references for other related research. To make the convenience of the experimental scientists, a user-friendly web server has been established (http://bioinfo.wugenqiang.top/pSuc-EDBAM/), by which the desired results can be easily obtained.

Keywords: Lysine succinylation, Post-translational modifications, Feature maps, Ensemble dense blocks, Feature learning, An attention module

Introduction

It has been discovered that succinylation is a novel protein post-translational modification, which is related to a variety of biological processes, including cancer progression and metastasis, and involves in the life activities such as glucose metabolism and amino acid metabolism through regulating the protease activity and gene expression [1]. Owing to the binding of lysine residues to succinyl group after succinylation, a

series of changes have taken place in the protein structure. Furthermore, succinylation changes the charge of lysine residues from +1 to −1, which further changes the physico-chemical properties of amino acids and enriches protein function [2]. Many related studies have fully displayed that succinylation may regulate multiple metabolic processes of organisms [3, 4], whose abnormalities are closely connected with emergence and development of human diseases, which include inflammation, tumors, cardiometabolic diseases, and so on [5, 6]. It can be seen that the importance of succinylation is self-evident, which has also attracted the atteintion of many researchers at home and broad [7].

Currently, a variety of biological experimental approaches have been come up with identifying succinylation sites, for instance, high-performance liquid chromatography assays, mass spectrometry, liquid chromatography-mass spectrometry, and so on [8]. In my opinion, these approaches are both costly and inefficient. Therefore, it is urgent to solve the shortcomings of biological experimental approaches by exploring a novel approach.

In recent years, thanks to the wide application of machine learning, a host of researchers have applied it to identify succinylation sites to solve the weakness of biological experimental approaches [9, 10]. Hasan et al. reviewed the latest advances regarding the current predictors, datasets, and online resources, which provided a useful guideline for developing effective succinylation site prediction tools [11]. Tasmia et al. updated the predictors, datasets, and online resources metioned in this review according to the development in recent years [12]. Xu et al. [13] built a succinylated site predictor based on SVM called iSuc-PseAAC in 2015, but the true distribution of the dataset is not fully taken into account. Jia et al. constructed some predictors including pSuc-Lys [14] and iSuc-PseOpt [15] in 2016; however, these classifiers ignored some important sequence information. Hasan et al. [16] built a predictor called SuccinSite based on the random forest (RF). In 2018, Dehzangi et al. [17] constructed the SSEvol-Suc predictor, which combined PSSM and the secondary structure with the AdaBoost by graph double-byte mapping, which is remarkably superior to previous predictors. Hasan et al. [18] constructed GPSuc by using logical regression (LR) combined with the output of different RF scores. Yosvany et al. [19] proposed a SVM-based predictor named Success with combining the structure and evolution information of amino acids with double-stranded maps. Zhu et al. [20] developed a RF-based predictor named Inspector combined with some sequence feature encoding schemes in 2020. To reduce the computational complexity, Zeng et al. [21] proposed a computational method named iSuc-ChiDT in 2022. Obviously, these approaches adopt manual feature selection, but it is difficult to find useful potential information [22]. Therefore, it is very necessary to explore a novel predictor which can automatically learn features to predict succinylation sites.

Along with the deepening of study, we found that deep learning (DL) can effectively overcome the shortcomings of the above problems, and can automatically learn useful features from the dataset. In 2020, Ning et al. [23] constructed the predictor named HybridSucc integrating a variety of information features, and adopts the penalized logistic regression algorithm and deep neural network (DNN) to make the model optimized. Thapa et al. [24] created a DL-based predictor called DeepSuccinylSite. Huang et al. [25] introduced long short-term memory (LSTM) and convolution neural network (CNN)

into DL methods in 2021. These predictors enrich the applications of DL in succinylation site prediction.

Based on the above review, we proposed a novel predictor using ensemble dense blocks and an attention module, called as pSuc-EDBAM. pSuc-EDBAM used one hot encoding to extract the initial protein sequence feature maps, and generated the low-level feature maps through 1-D CNN. Afterward, ensemble dense blocks [26] were adopted to capture the advanced features from the initial features. In addition, an attention module was used to evaluate the importance degrees of different protein sequence features, make every feature map weighted, and then improve the network abstraction ability to predict potential succinylation sites. The features were then matched with the softmax classifier to go on succinylation site prediction. To further illustrate the performance of pSuc-EDBAM, we performed ten-fold cross-validation on the training dataset and independent test on the testing dataset. According to experimental results, our model has yielded promising results and is superior to the existing predictors. The prediction and potential succinylation sites further show that pSuc-EDBAM is a powerful predictor for predicting unknown succinylation sites.

The main contributions of the paper are summarized below: (1) An effective and novel predictor was proposed based on ensemble dense blocks and an attention module for succinylation sites prediction, called as pSuc-EDBAM. (2) An improved attention module was introduced into the prediction of succinylation sites, which improved the prediction ability of succinylation sites. (3) In this paper, our model is simple and easy to use, and features are automatically learned on the basis of feature map extracted by one hot encoding, which greatly improves the ability of succinylation site prediction. (4) The model built in this paper can broaden the thinking of other researchers and do better research. (5) A web-server has been provided at http://bioinfo.wugenqiang.top/pSuc-EDBAM/, by which the desired results may be easily obtained.

Materials and methods

In binary classification-based lysine succinylation site prediction studies, we labeled each potential site as a succinylated site or a non-succinylated site [27]. In particular, we extracted a protein sequence with length $L=2r+1$ with lysine (K) as the center, where r represents amino acid residues on each side. Firstly, we converted the input into numerical vectors by an encoding method, and then we trained pSuc-EDBAM based on the benchmark training dataset. Finally, its performance was evaluated by comparing it with other existing predictors.

Benchmark dataset

The benchmark dataset was gathered from the UniProtKB/Swiss-Prot database [28] and NCBI protein sequence database from Ning et al. [29]. In order to reduce the model deviations owing to the sequence homology, we used CD-HIT [30] to remove redundant sequences, and set the threshold to 0.3. Finally, 2322 proteins were retained as our final benchmark dataset, containing 5009 experimentally verified succinylation sites and 53,542 non-succinylation sites. To conveniently compare with other existing methods, 124 proteins were randomly separated from 2322 proteins as an independent testing dataset, and the remaining proteins were used as a training dataset. Table 1 lists the

Table 1 The specifies of the benchmark dataset

Original dataset	Number of proteins	Positive site	Negative site
Training dataset	2198	4755	50,549
Testing dataset	124	254	2977

specifies of the benchmark dataset. In order to facilitate further research by researchers, the benchmark dataset can be easily obtained from https://github.com/wugenqiang/pSuc-EBDAM/tree/main/dataset.

We adopted Chou's peptide formulation [31], each protein sequence can be defined as Eq. (1).

$$P_\delta(K) = R_{-\delta}R_{-(\delta-1)}\cdots R_{-2}R_{-1}KR_{+1}R_{+2}\cdots R_{+(\delta-1)}R_{+\delta} \tag{1}$$

where K denotes the lysine and δ denotes an integer, $R_{-\delta}$ denotes the δth amino acid residue to the left of K, and $R_{+\delta}$ denotes the δth amino acid residue to the right of K, so that $P_\delta(K)$ can define each protein sequence as two classes as shown in Eq. (2).

$$P_\delta(K) \in \begin{cases} P_\delta^+(K), & \textit{if the center is a succinylation site} \\ P_\delta^-(K), & \textit{otherwise} \end{cases} \tag{2}$$

It is not difficult to find that some protein sequences have some non-standard residues, such as "X", we adopted a better approach from Jia's paper [14], which made this part of amino acid residues filled via mirroring image, as defined in Eqs. (3) and (4).

(A) The mirror image of carbon-terminus

$$R_{+\delta}R_{+(\delta-1)}\cdots R_{+2}R_{+1} \underset{K}{\Longleftrightarrow} R_{+1}R_{+2}\cdots R_{+(\delta-1)}R_{+\delta} \tag{3}$$

(B) The mirror image of nitrogen-terminus

$$R_{-\delta}R_{-(\delta-1)}\cdots R_{-2}R_{-1} \underset{K}{\Longleftrightarrow} R_{-1}R_{-2}\cdots R_{-(\delta-1)}R_{-\delta} \tag{4}$$

The mirror image of carbon-terminus is on the left of "⇔" in Eq. (3), and the mirror image of nitrogen-terminus is on the right of "⇔" in Eq. (4); while the original sequence is on the other side, with "⇔" indicating the mirror image and K representing the lysine.

Sequence feature extraction via one hot encoding

One hot encoding is a common feature extraction method to reflect the types and positions of amino acid residues directly, which has been maturely applied to the process of protein feature extraction [32]. In this study, we applied this method to obtain feature maps from the protein sequences for further research. We listed the 20 amino acid residues in alphabetical sequence as ACDEFGHIKLMNPQRSTVWY and enforced the following rule: the ith amino acid residue was labeled as 1 in the ith position and 0 in the other positions, such as amino acid residue D was coded as 00100000000000000000, and then the protein sequence of length L can be converted into $L \times 20$ dimensional feature vector.

Model construction

We constructed a model to learn the deeply hidden features of succinylation sites efficiently in this study, called as pSuc-EDBAM. In this pSuc-EDBAM model, ensemble dense blocks [26] were adopted to obtain the advanced features. Afterward, we introduced an attention module to evaluate feature importance degrees. Eventually, the advanced features were input into the softmax classifier to predict succinylation sites. The framework of pSuc-EDBAM is shown in Fig. 1.

1-D CNN

The convolution neural network (CNN) is a common feed-forward network, which was proposed by LeCun et al. [33, 34] and has some notable advantages such as parameter sharing and local connectivity. In 1-D CNN, the CNN kernel moves in one direction to extract the protein sequence features, and the dimensions of input and output data are both two-dimensional, mainly used for sequence model, while in 2-D CNN, the CNN kernel moves in two directions to extract the protein sequence features, and the dimensions of input and output data are both three-dimensional, mainly used for image data [35].

Here, we used 1-D CNN to extract low-level features. Suppose a discrete sequence is $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_n]$, and the convolution kernel is $\beta = [\beta_1, \beta_2, \ldots, \beta_m]$. The 1-D CNN of α and β is expressed as Eq. (5).

$$\alpha * \beta = \left[\sum_{i=1}^{m} \alpha_{jd+i-1} b_i \right], \quad j = 1, 2, \ldots, k \tag{5}$$

where d represents the stride of convolution and k indicates the length of the output sequence features, which is the most integer less than or equal to $\frac{(n-m)}{d} + 1$.

After adopting one hot encoding to extract the feature map, we generated the low-level feature map through 1-D CNN, as shown in Eq. (6).

$$X^0 = \sigma(I * W + b) \tag{6}$$

where I denotes the feature map extracted from one hot encoding, W denotes the weight matrix, b is used to denote the bias term, σ is used to denote the exponential linear unit (ELU) activation function [36], and X^0 denotes the low-level feature map generated by the 1-D CNN.

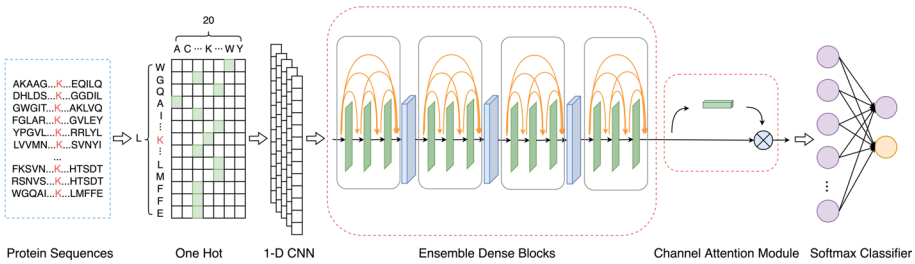

Fig. 1 The framework of the pSuc-EDBAM model

Ensemble dense blocks for further feature extraction

For the sake of extracting the advanced features of succinylation sites, we introduced the ensemble dense blocks, which have been shown to perform higher than traditional CNN. The structure of a dense block is expressed as Fig. 2.

The advanced feature representation of the low-level feature map was extracted by the dense block, which is expressed as Eq. (7).

$$X^l = \sigma([X^0; X^1; \ldots; X^{l-1}] * W' + b') \tag{7}$$

where X^{l-1} represents the feature map generated by the $(l-1)$th convolutional layer within the dense block. W' represents the weight matrix, b' denotes the bias term. The output value of the dense block is the concatenation of X^0, $X^1, \ldots$, and X^l.

The next step was to build the transition layer, which is expressed as Eq. (8).

$$X = \sigma([X^0; X^1; \ldots; X^l] * W'' + b'') \tag{8}$$

where W'' denotes the weight matrix, b'' represents the bias term, and X refers to the output value of the transition layer. Afterward, the average pooling was performed on X to reduce the risk of overfitting.

In the study, we integrated four dense blocks. What's more, Eq. (8) is not performed after the fourth implementation of Eq. (7) but replaces it with global average pooling. Ultimately, the advanced feature $X^{(seq)}$ was extracted after the above process.

An attention module for learning feature importance degrees

From my point of view, different features have different degrees of importance. Consequently, we introduced an attention module to learn feature importance degrees and make every feature map weighted. Here, we proposed the channel attention module, which is implemented via global average pooling, global max pooling, and two fully connected layers, which adds the global max pooling to increase the receptive field of the channel and the importance of learning characteristics more comprehensively based on SE [37] module. The structure of the channel attention module is described in Fig. 3.

For the advanced feature $X^{(seq)}$, the channel attention module used global average pooling and global max pooling to squeeze the space information of $X^{(seq)}$ into the

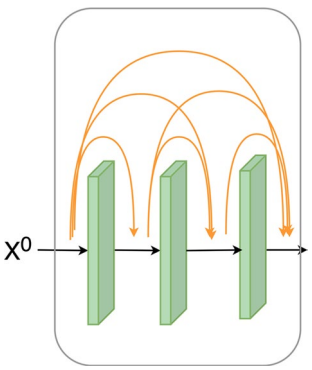

Fig. 2 The structure of a dense block

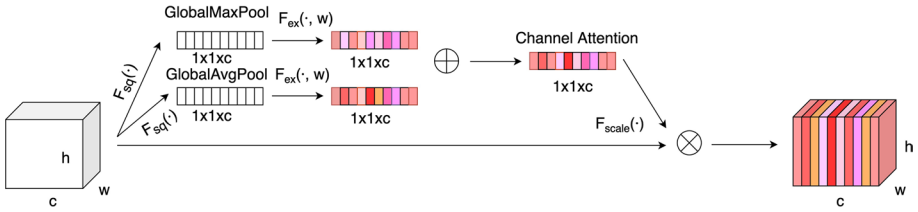

Fig. 3 The structure of the channel attention module

channel z, separately, which respectively used Eq. (9) to get the compressed result z_{avg} and z_{max}.

$$z = F_{sq}(X^{(seq)}) = \frac{1}{w \times h} \sum_{i=1}^{h} \sum_{j=1}^{w} X^{(seq)}(i,j) \tag{9}$$

where w is the width of $X^{(seq)}$ and h is the height of $X^{(seq)}$.

Thereafter, two fully connected layers were adopted to process z_{avg} and z_{max} respectively to obtain the channel information of $X^{(seq)}$ and learn the weight of $X^{(seq)}$, as described in Eq. (10), and then got s_{avg} and s_{max}, respectively.

$$s = F_{ex}(z, W) = \sigma(W_2 * \tau(W_1 * z)) \tag{10}$$

where s denotes the weight of $X^{(seq)}$, τ means a rectified linear unit (RELU) function and σ represents a sigmoid function. W_1 and W_2 are the parameter. To make the weight information of captured every feature map more comprehensive, we added s_{avg} and s_{max} to get the specific weight information of every feature map, as described in Eq. (11).

$$s = s_{avg} + s_{max} \tag{11}$$

Ultimately, the output value of the attention module can be got by scaling $X^{(seq)}$ with the activation described in Eq. (12).

$$X^{(seq)} = F_{scale}(X^{(seq)}, s) = s \cdot X^{(seq)} \tag{12}$$

where $F_{scale}(X^{(seq)}, s)$ indicates that each specific value of $X^{(seq)}$ is multiplied by the weight s.

Softmax classifier

On the basis of the advanced features, the softmax classifier was adopted to predict succinylation sites in this study, which received the advanced features as input, and then weighted summation and activation operations are performed to obtain the predicted results of succinylation sites, just as Eq. (13).

$$P(y = i|x) = \frac{e^{W_i^s * X + b_i^s}}{\sum_{j=1}^{2} e^{W_j^s * X + b_j^s}} \tag{13}$$

where W_i^s and W_j^s indicate the weight matrices, b_i^s and b_j^s indicate the bias terms, and x denotes the samples. $P(y = i|x)$ refers to the probability that x is predicted to be i.

Owing to succinylation site prediction may be considered as a problem of the binary classification, therefore $i = 0$ or $i = 1$.

The decision threshold refers to the decision that converts the prediction probability into the target class. In this paper, the threshold we use is the default value, set to 0.5. When the prediction probability is higher than 0.5 ($P > 0.5$), the site is predicted to be succinylated, and when the prediction probability is lower than 0.5 ($P < 0.5$), the site is predicted to be non-succinylated.

Model training

The pSuc-EDBAM was carried out based on Keras 2.8 (https://keras.io/) in this paper, which is a flexible, simple, and Python-based approach. We adopted the cross entropy as the loss function, which is shown in Eq. (14).

$$C = -\frac{1}{n}\sum_{j=1}^{n} y^j lnP(y^j = 1|x^j) + (1 - y^j)lnP(y^j = 0|x^j) \tag{14}$$

where n represents the number of training samples, x^j indicates jth input, and y^j represents the true label of x^j. The loss function was optimized by Adam optimizer [38], and the parameters were adjusted by the gradient descent method to minimize the loss function.

Additionally, L_2 regularization was adopted to weaken the negative influence of overfitting, we also used dropout [39] and early stopping [40] to further avoid overfitting. The ratio of positive samples to negative samples in our succinylated dataset is 1:11, which is very unbalanced. To weaken the influence of unbalanced dataset, we introduced the class weight and set the class weight ratio of positive samples to negative samples to 11:1. By this means, the pSuc-EDBAM model could improve the influence of positive samples, so as to further improve the recognition rate of succinylation sites.

Performance evaluation

Four common metrics were considered to evaluate the performance of pSuc-EDBAM reasonably as previously described [41], including sensitivity (Sn), specificity (Sp), accuracy (Acc), and Mathews Correlation Coefficient (MCC) [42], which are defined as Eq. (15).

$$\begin{cases} Sp = \frac{TN}{TN+FP} \\ Sn = \frac{TP}{TP+FN} \\ Acc = \frac{TP+TN}{TP+TN+FP+FN} \\ MCC = \frac{TP\times TN - FP\times FN}{\sqrt{(TP+FP)\times(TP+FN)\times(TN+FP)\times(TN+FN)}} \end{cases} \tag{15}$$

where TP, TN, FP, and FN denote true positive samples, true negative samples, false positive samples, and false negative samples, respectively. Sn was adopted to measure the proportion of predicting succinylation sites correctly, Sp measured the proportion of predicting non-succylation sites correctly, and Acc revealed the proportion of predicting sites correctly. When the distributions of samples are very imbalanced, MCC was considered to be the more noteworthy measure because it can more accurately reflect the quality of the model [43]. In general, the value of MCC is –1, indicating that the

prediction of succinylated sites is completely wrong; the value of MCC is 0, indicating that the prediction effect of succinylation sites is not better than that of random prediction, and the value of MCC is + 1, meaning that the prediction of succinylation sites is completely correct.

What's more, the receiver operating characteristic (ROC) curve was adopted to reveal the performance of the model. On this basis, we introduced the area under the ROC curve (AUC) to further intuitively explain the performance of the model. The higher the AUC, the better the overall performance of the model. ten-fold cross-validation was adopted to evaluate the robustness of pSuc-EDBAM and independent test was used to compare the performance of pSuc-EDBAM with the existing predictors.

Results and discussion

Select the best window size of succinylation sites

The size of the protein sequence directly determines the feature representation learned by the model, so the selection of its value has an important influence on the succinylation site prediction. To gain the best window size of the succinylation sites of the pSuc-EDBAM model, it is especially necessary to make full use of the automatic and efficient feature extraction of CNN. Based on the training dataset, we chose 19, 21, 23, 25, 27, 29, 31, and 33 as the window size, tested each window size by ten-fold cross-validation, and then averaged the experimental results. The experimental results are shown in Fig. 4. MCC value reaches maximum when the window size is 31, which indicates that 31 is the best window size of the succinylation sites of the pSuc-EDBAM model.

Sequence analysis of succinylation sites

This investigation analyzed the frequency of occurrence of 30 amino acid residues surrounding the succinylation site on fragment protein sequences to find the potential consensus motifs. Two Sample Logo [44] is an effective tool to find statistically noteworthy differences in position-specific symbol compositions between the succinylated and non-succinylated sites. To better distinguish succinylation sites and non-succinylation sites in the samples, we used Two Sample Logo to analyze the protein sequences and looked

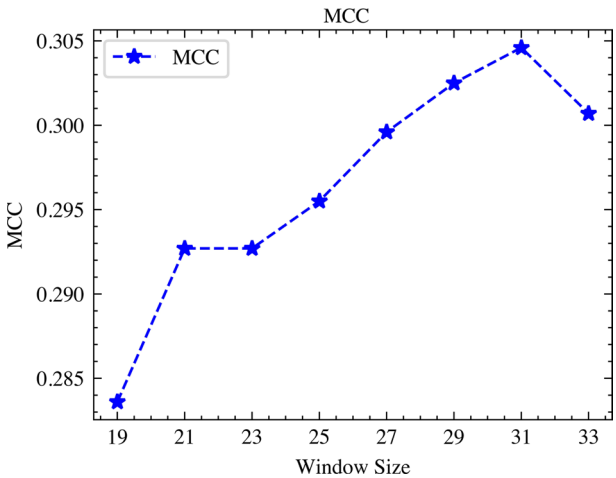

Fig. 4 The values of MCC under different window sizes are based on the training dataset

into the frequency and position differences of 20 amino acid residues near the succinylation sites and non-succinylation sites, as described in Fig. 5.

In this study, the lysine (K) amino acid residue was placed in the middle of the fragment protein sequences, and positions of the flanking amino acid residues were shown in the range from 1 to 31. The comparison between 4755 succinylation sites and 50,549 non-succinylation sites in Fig. 5 demonstrates that the amino acids including alanine (A), asparagine (D), leucine (L), arginine (R), and valine (V), had the highest ratios around the succinylation site, glutamicacid (E), lysine (K), serine (S) appeared more frequently around the non-succinylation site. Therefore, we concluded that the frequencies and positions of amino acid residues around the succinylation site and the non-succinylation site are different, and the analysis showed that the distance among amino acid characteristics in a sequence plays a vital role in distinguishing between succinylation sites and non-succinylation sites.

Performance of pSuc-EDBAM on the training dataset

To analyze the performance of pSuc-EDBAM, we conducted ten-fold cross-validation on the training dataset shown in Table 1. As described in Table 2, we find that the values of Sn, Sp, Acc, and MCC on the training dataset are very stable, and the fluctuation is relatively small, indicating that our proposed model has a certain stability, and effectively avoids the problem of overfitting. Therefore, the pSuc-EDBAM model has good performance on the training dataset.

Figure 6 shows the ROC curve of the pSuc-EDBAM on the training dataset, and the mean of AUC is 0.8295. The result indicates that our proposed pSuc-EDBAM has notable advantages including better stability.

Comparison with the existing predictors

To certify the effectiveness of our predictor named pSuc-EDBAM, we conducted independent test on the same independent testing dataset to objectively compare

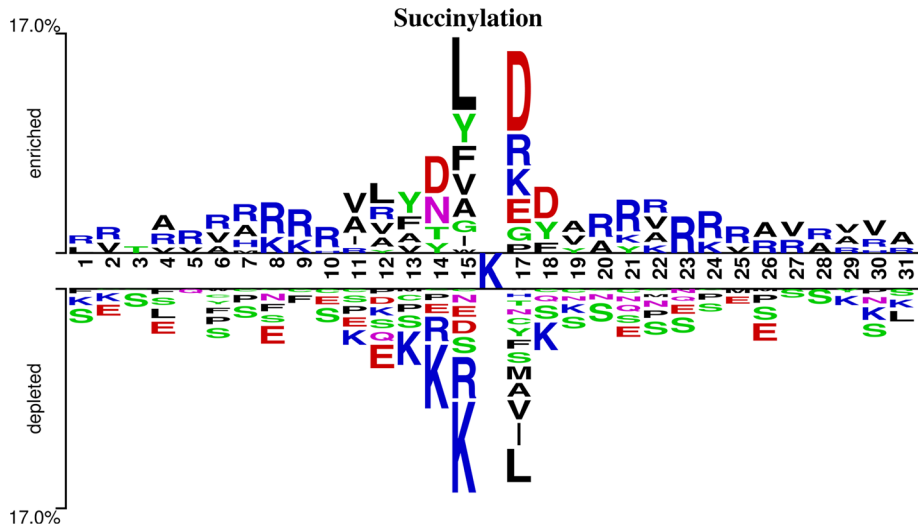

Fig. 5 A two-sample logo of succinylation sites against non-succinylation sites with $L = 31$

Table 2 Performance of pSuc-EDBAM on the training dataset

Fold times	Sn (%)	Sp (%)	Acc (%)	MCC
1	77.10	73.16	73.49	0.3043
2	76.68	74.44	74.63	0.3130
3	81.30	72.01	72.81	0.3190
4	76.26	72.36	72.70	0.2928
5	77.52	73.17	73.54	0.3069
6	73.89	72.74	72.84	0.2819
7	75.16	76.48	76.37	0.3224
8	79.79	69.73	70.60	0.2919
9	79.16	71.41	72.08	0.3015
10	72.21	77.25	76.82	0.3122
Mean ± STD	76.91 ± 2.60	73.28 ± 2.15	73.59 ± 1.80	0.3046 ± 0.0122

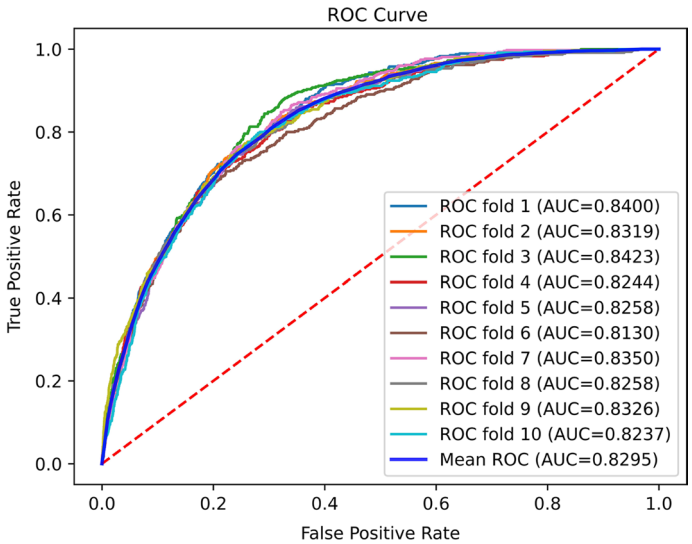

Fig. 6 The receiver operating characteristics (ROC) curve for pSuc-EDBAM on the training dataset. AUC denotes the area under the ROC curve

pSuc-EDBAM with existing predictors. Eight existing predictors were considered, including SuccinSite [16], SuccinSite2.0 [45], Success [19], pSuccE [29], GPSuc [18], Inspector [20], iSuc-ChiDT [21], and iSuccLys-BLS [46]. The details are shown in Table 3. SuccinSite was established based on RF which was trained with three combined encodings. SuccinSite2.0 was also constructed through RF, but it took the composition of profile-based amino acid and orthogonal binary features as the training data. Success was constructed by using the structural and evolutionary information of amino acids to extract protein features. PSuccE was another classifier combining multiple features with a feature selection scheme. GPSuc was established based on RF. iSuc-ChiDT was proposed to identify succinylation sites using statistical difference table encoding and the chi-square decision table classifier. iSuccLys-BLS was constructed using a broad learning system (BLS), which optimized the imbalanced training dataset using randomly labeling samples. The novel deep learning-based predictor pSuc-EDBAM was proposed in this

Table 3 Performance comparison of pSuc-EDBAM with other existing predictors on the independent testing dataset

Predictor	Sn (%)	Sp (%)	Acc (%)	MCC
SuccinSite	37.10	88.20	84.20	0.1990
SuccinSite2.0	45.40	88.20	84.80	0.2610
Success	14.20	86.80	81.10	0.0700
PSuccE	37.50	88.60	84.50	0.2040
GPSuc	49.90	88.30	85.30	0.2960
Inspector	69.30	71.70	71.50	0.2380
iSuc-ChiDT	70.47	66.27	68.30	0.2050
iSuccLys-BLS	72.30	68.90	69.20	0.2340
pSuc-EDBAM	75.59	74.13	74.25	0.2927

paper based on ensemble dense blocks and an attention module, which adopted one hot encoding to capture the protein sequence feature.

The ratio of positive samples to negative samples in the independent testing dataset studied in this paper is about 1:11, and the dataset is extremely unbalanced. In the study of this kind of unbalanced data, Sn and MCC are the main indicators to consider, and the improvement of these two indicators is particularly important. As can be seen in Table 3, based on the independent testing dataset, the pSuc-EDBAM gained higher values of Sn, which is 3.29% higher than the current best predictor named iSuccLys-BLS [45]. We know that Sn means the proportion of all positive samples that are correctly predicted as positive samples, and then evaluates the predictor's performance in predicting positive samples. Therefore, the novel predictor pSuc-EDBAM with the higher Sn value is more significant and practical for predicting succinylation sites. We find that the earliest predictors have high Sp and low Sn, such as SuccinSite, SuccinSite2.0, Success, pSuccE, and GPSuc. This is because these predictors did not take into account the imbalance of the data or find a reliable method to solve the problem of unbalanced data, which caused the recognition of these predictors to the positive samples was not obvious. When noticing the unbalanced distribution of the dataset, the values of Sn and Sp would tend to be balanced. It means that the predictor taking into account the data distribution is of more practical significance, such as Inspector, iSuc-ChiDT, iSuccLys-BLS, and our proposed predictor called pSuc-EDBAM. It is found that our predictor is significantly superior to Inspector, iSuc-ChiDT, and iSuccLys-BLS in all metrics.

When AUC is nearer 1, the performance of predictor is better. Figure 7 shows the ROC curve of the pSuc-EDBAM on the independent testing dataset, and AUC is 0.8201. The result indicates that our proposed novel predictor has more advantages and better stability. Therefore, it is expected that pSuc-EDBAM may be a more representative and meaningful tool in succinylation site prediction.

Implementation of the pSuc-EDBAM predictor and user guide

An effective predictor can be beneficial for researchers to study the protein succinylation sites. In this study, an open online web-based predictor named pSuc-EDBAM is designed to analyze protein succinylation sites efficiently, which can be accessed

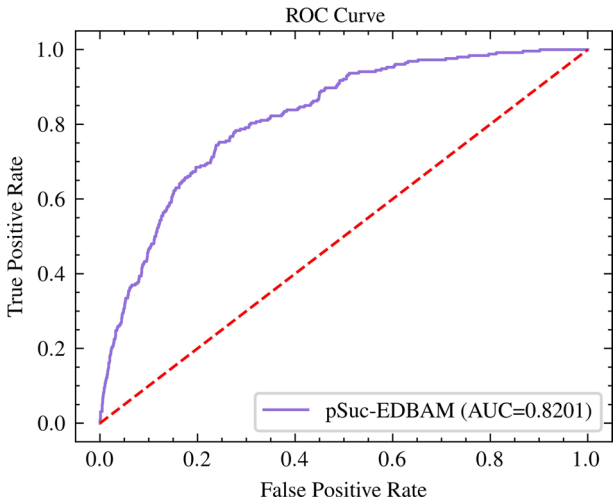

Fig. 7 The receiver operating characteristics (ROC) curve for pSuc-EDBAM on the independent testing dataset. AUC denotes the area under the ROC curve

at http://bioinfo.wugenqiang.top/pSuc-EDBAM/. To make the convenience of most researchers, we supply a user guide below:

Step 1: Open the homepage of the pSuc-EDBAM predictor as described in Fig. 8. You can click on the web page button such as "Help" or "More info..." to look up the profile of the pSuc-EDBAM web server.
Step 2: Enter a single protein sequence according to the prompt, which is required to be in FASTA format. Click on "example" button and you can see the example of a protein sequence in FASTA format.
Step 3: Click on "Submit" button after inputing the protein sequence and you can obtain the predicted results of the succinylation sites.
Step 4: The users of the pSuc-EDBAM predictor can upload files by the "Browse" button and these files must be in FASTA format. Then you need to leave the project name and your email address so that we can send the predicted results of the succinylation sites to you in a timely manner.

Conclusion

In this paper, we proposed a novel predictor called pSuc-EDBAM for succinylation site prediction, which used ensemble dense blocks and an atteintion module. The efficiency of the pSuc-EDBAM was demonstrated by ten-fold cross-validation on the training dataset and independent test on the testing dataset.

Although pSuc-EDBAM has shown strong robustness in predicting succinylation sites, it still has some weakness. In the course of continuous learning, we have learned that the deep learning is regarded as a black box, which may not be explained in biological processes [47]. In the following work, we will take biological interpretation into consideration and apply more effective attention modules including the

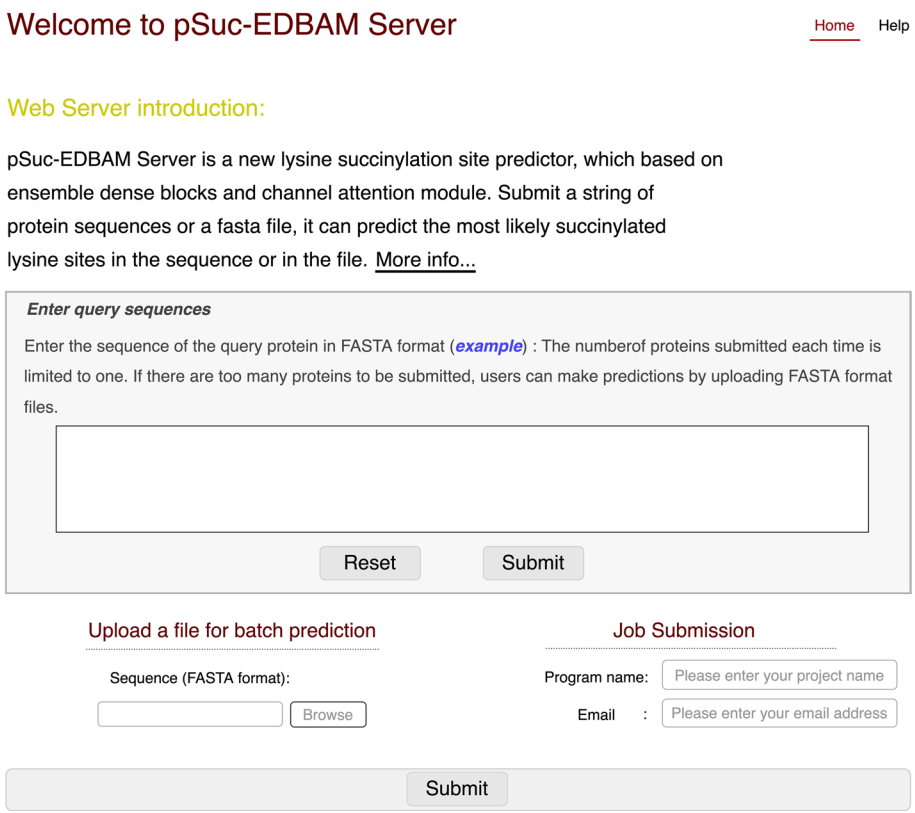

Fig. 8 The homepage of the pSuc-EDBAM predictor at http://bioinfo.wugenqiang.top/pSuc-EDBAM/

convolutional block attention module (CBAM) [48], external attention (EA) [49], and so on, through which will have more meaningful gains in the following experiments.

Considering these together, although further improvement should be conducted as new dataset are available, the pSuc-EDBAM will provide useful information for further experimental manipulation. With the upgrading of technology and the rapid development of proteomics research technology, new research approaches emerge in endlessly, which will bring great convenience to the medical field. It is helpful to further reveal the regulation mechanism of succinylation and provide new ideas for the biomedical research.

Acknowledgements
The authors are grateful for the constructive comments and suggestions made by the reviewers.

Author contributions
JJ and GW conceived and designed the experiments; GW implemented the feature extraction, model construction, model training, and performance evaluation. GW and ML drafted the manuscript and revised the manuscript. JJ and WQ supervised this study. All authors contributed to the content of this paper, and approved the final manuscript.

Funding
This work was partially supported by the National Nature Science Foundation of China (Nos. 61761023, 62162032 and 31760315), the Natural Science Foundation of Jiangxi Province, China (Nos. 20202BABL202004 and 20202BAB202007), the Scientific Research Plan of the Department of Education of Jiangxi Province (GJJ190695). These funders had no role in the study design, data collection and analysis, decision to publish or preparation of manuscript.

Availability of data and materials
The dataset and source code used in this study can be easily derived from https://github.com/wugenqiang/pSuc-EDBAM.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 16 August 2022 Accepted: 25 October 2022
Published online: 31 October 2022

References

1. Wang Y, Guo YR, Liu K, Yin Z, Liu R, Xia Y, et al. KAT2A coupled with the alpha-KGDH complex acts as a histone H3 succinyltransferase. Nature. 2017;552(7684):273–7.
2. Papanicolaou KN, O'Rourke B, Foster DB. Metabolism leaves its mark on the powerhouse: recent progress in post-translational modifications of lysine in mitochondria. Front Physiol. 2014;5:301.
3. Park J, Chen Y, Tishkoff DX, Peng C, Tan M, Dai L, et al. SIRT5-mediated lysine desuccinylation impacts diverse metabolic pathways. Mol Cell. 2013;50(6):919–30.
4. Rardin MJ, He W, Nishida Y, Newman JC, Carrico C, Danielson SR, et al. SIRT5 regulates the mitochondrial lysine succinylome and metabolic networks. Cell Metab. 2013;18(6):920–33.
5. Alleyn M, Breitzig M, Lockey R, Kolliputi N. The dawn of succinylation: a posttranslational modification. Am J Physiol Cell Physiol. 2018;314(2):C228–32.
6. Ao C, Yu L, Zou Q. Prediction of bio-sequence modifications and the associations with diseases. Brief Funct Genom. 2021;20(1):1–18.
7. Peng C, Lu Z, Xie Z, Cheng Z, Chen Y, Tan M, et al. The first identification of lysine malonylation substrates and its regulatory enzyme. Mol Cell Proteomics. 2011;10(12):M111 012658.
8. Lind C, Gerdes R, Hamnell Y, Schuppe-Koistinen I, Lowenhielm H, Holmgren A, et al. Identification of S-glutathionylated cellular proteins during oxidative stress and constitutive metabolism by affinity purification and proteomic analysis. Arch Biochem Biophys. 2002;406(2):229–40.
9. Chen Z, Liu X, Li F, Li C, Marquez-Lago T, Leier A, et al. Large-scale comparative assessment of computational predictors for lysine post-translational modification sites. Brief Bioinform. 2019;20(6):2267–90.
10. Li F, Fan C, Marquez-Lago TT, Leier A, Revote J, Jia C, et al. PRISMOID: a comprehensive 3D structure database for post-translational modifications and mutations with functional impact. Brief Bioinform. 2020;21(3):1069–79.
11. Hasan MM, Khatun MS, Kurata H. Large-scale assessment of bioinformatics tools for lysine succinylation sites. Cells. 2019;8(2).
12. Tasmia SA, Kibria MK, Islam MA, Khatun MS, Haque Mollah MN. A comprehensive comparative review of protein sequence based computational prediction models of lysine succinylation sites. Curr Protein Pept Sci. 2022.
13. Xu Y, Ding YX, Ding J, Lei YH, Wu LY, Deng NY. iSuc-PseAAC: predicting lysine succinylation in proteins by incorporating peptide position-specific propensity. Sci Rep. 2015;5:10184.
14. Jia J, Liu Z, Xiao X, Liu B, Chou KC. pSuc-Lys: predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach. J Theor Biol. 2016;394:223–30.
15. Jia J, Liu Z, Xiao X, Liu B, Chou KC. iSuc-PseOpt: identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset. Anal Biochem. 2016;497:48–56.
16. Hasan MM, Yang S, Zhou Y, Mollah MN. SuccinSite: a computational tool for the prediction of protein succinylation sites by exploiting the amino acid patterns and properties. Mol Biosyst. 2016;12(3):786–95.
17. Dehzangi A, Lopez Y, Lal SP, Taherzadeh G, Sattar A, Tsunoda T, et al. Improving succinylation prediction accuracy by incorporating the secondary structure via helix, strand and coil, and evolutionary information from profile bigrams. PLoS ONE. 2018;13(2): e0191900.
18. Hasan MM, Kurata H. GPSuc: global prediction of generic and species-specific succinylation sites by aggregating multiple sequence features. PLoS ONE. 2018;13(10): e0200283.
19. Lopez Y, Sharma A, Dehzangi A, Lal SP, Taherzadeh G, Sattar A, et al. Success: evolutionary and structural properties of amino acids prove effective for succinylation site prediction. BMC Genom. 2018;19(Suppl 1):923.
20. Zhu Y, Jia C, Li F, Song J. Inspector: a lysine succinylation predictor based on edited nearest-neighbor undersampling and adaptive synthetic oversampling. Anal Biochem. 2020;593: 113592.
21. Zeng Y, Chen Y, Yuan Z. iSuc-ChiDT: a computational method for identifying succinylation sites using statistical difference table encoding and the chi-square decision table classifier. BioData Min. 2022;15(1):3.
22. Wang D, Zeng S, Xu C, Qiu W, Liang Y, Joshi T, et al. MusiteDeep: a deep-learning framework for general and kinase-specific phosphorylation site prediction. Bioinformatics. 2017;33(24):3909–16.
23. Ning W, Xu H, Jiang P, Cheng H, Deng W, Guo Y, et al. HybridSucc: a Hybrid-learning Architecture for General and Species-specific Succinylation Site Prediction. Genom Proteomics Bioinform. 2020;18(2):194–207.
24. Thapa N, Chaudhari M, McManus S, Roy K, Newman RH, Saigo H, et al. DeepSuccinylSite: a deep learning based approach for protein succinylation site prediction. BMC Bioinform. 2020;21(Suppl 3):63.
25. Huang G, Shen Q, Zhang G, Wang P, Yu ZG. LSTMCNNsucc: a bidirectional LSTM and CNN-based deep learning method for predicting lysine succinylation sites. Biomed Res Int. 2021;2021:9923112.

26. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: 2017. p. 2261–2269.
27. Wang D, Liang Y, Xu D. Capsule network for protein post-translational modification site prediction. Bioinformatics. 2019;35(14):2386–94.
28. UniProt C. Ongoing and future developments at the Universal Protein Resource. Nucleic Acids Res. 2011;39(Database issue):D214–219.
29. Ning Q, Zhao X, Bao L, Ma Z, Zhao X. Detecting Succinylation sites from protein sequences using ensemble support vector machine. BMC Bioinform. 2018;19(1):237.
30. Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics. 2010;26(5):680–2.
31. Chou K. Prediction of signal peptides using scaled window. Peptides. 2001;22(12):1973–1979.
32. Jia J, Wu G, Qiu W. pSuc-FFSEA: predicting lysine succinylation sites in proteins based on feature fusion and stacking ensemble algorithm. Front Cell Dev Biol. 2022;10.
33. Lecun Y, Boser B, Denker J, Henderson D, Howard R, Hubbard W, et al. Backpropagation applied to handwritten zip code recognition. Neural Comput. 1989;1(4):541–51.
34. Lecun Y, Bottou L. Gradient-based learning applied to document recognition. Proc IEEE. 1998;86(11):2278–324.
35. Hasan MM, Tsukiyama S, Cho JY, Kurata H, Alam MA, Liu X, et al. Deepm5C: a deep-learning-based hybrid framework for identifying human RNA N5-methylcytosine sites using a stacking strategy. Mol Ther. 2022;30(8):2856–67.
36. Clevert D-A, Unterthiner T, Hochreiter S, editors. Fast and accurate deep network learning by exponential linear units (ELUs). ICLR; 2016.
37. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: 2018. p. 7132–7141.
38. Kingma D, Ba J. Adam: a method for stochastic optimization. Comput Sci. 2014.
39. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res. 2014;15(1):1929–58.
40. Yao Y, Rosasco L, Caponnetto A. On early stopping in gradient descent learning. Constr Approx. 2007;26(2):289–315.
41. Li S, Yu K, Wu G, Zhang Q, Wang P, Zheng J, et al. pCysMod: prediction of multiple cysteine modifications based on deep learning framework. Front Cell Dev Biol. 2021;9: 617366.
42. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Inf Process Manag. 2009;45(4):427–37.
43. Boughorbel S, Jarray F, El-Anbari M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. PLoS ONE. 2017;12(6): e0177678.
44. Vacic V, Iakoucheva LM, Radivojac P. Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. Bioinformatics. 2006;22(12):1536–7.
45. Hasan MM, Khatun MS, Mollah MNH, Yong C, Guo D. A systematic identification of species-specific protein succinylation sites using joint element features information. Int J Nanomed. 2017;12:6303–15.
46. Jia J, Shen Y, Qiu W. Identifying lysine succinylation sites in proteins by broad learning system and optimizing imbalanced training dataset via randomly labeling samples. Wuhan Univ J Nat Sci. 2021;26(01):81–8.
47. Ma J, Yu MK, Fong S, Ono K, Sage E, Demchak B, et al. Using deep learning to model the hierarchical structure and function of a cell. Nat Methods. 2018;15(4):290–8.
48. Woo S, Park J, Lee JY, Kweon IS. CBAM: convolutional block attention module. Cham: Springer; 2018.
49. Guo MH, Liu ZN, Mu TJ, Hu SM. Beyond self-attention: external attention using two linear layers for visual tasks. 2021.

Publisher's Note