Hindawi
Scientific Programming
Volume 2020, Article ID 8847694, 8 pages
https://doi.org/10.1155/2020/8847694

Research Article

CirBiTree: Citrullination Site Inference Based on a Fuzzy Neural Network and Flexible Neural Tree

Chuandong Song and Haifeng Wang

School of Information Science and Engineering, Zaozhuang University, Zaozhuang 277100, China

Correspondence should be addressed to Haifeng Wang; 23384722@qq.com

Received 6 September 2020; Revised 27 October 2020; Accepted 30 October 2020; Published 28 November 2020

Academic Editor: Wenzheng Bao

Emerging evidence demonstrates that post-translational modification plays an important role in several human complex diseases. Nevertheless, considering the inherent high cost and time consumption of classical and typical in vitro experiments, an increasing attention has been paid to the development of efficient and available computational tools to identify the potential modification sites in the level of protein. In this work, we propose a machine learning-based model called CirBiTree for identification the potential citrullination sites. More specifically, we initially utilize the biprofile Bayesian to extract peptide sequence information. Then, a flexible neural tree and fuzzy neural network are employed as the classification model. Finally, the most available length of identified peptides has been selected in this model. To evaluate the performance of the proposed methods, some state-of-the-art methods have been employed for comparison. The experimental results demonstrate that the proposed method is better than other methods. CirBiTree can achieve 83.07% in sn%, 80.50% in sp, 0.8201 in F1, and 0.6359 in MCC, respectively.

1. Introduction

Human genome project has been successfully completed in the end of the twentieth century. More than 20,000 protein-coding genes have been reported. These coding genes construct the intact proteins in the biological processions. Nevertheless, this information can hardly cover the relationships among the proteins and the human biological processions [1, 2]. With the development of the proteomics, several types of post-translational modification (PTM) have been reported in the level of protein. These modifications have the ability to construct protein structure and maintain proteins' stability. According to the foundational protein composition, PTMs make contributions to translating peptides [3, 4]. A great number of PTMS can alter physiological activity. Meanwhile, several PTMs have reversible biological functions. It was noted that PTMs take part in several diseases. For instances, PTM enzymes are involved in neurodegeneration diseases, especially in patients with AD and Parkinson's disease [5–7]. So, having a good knowledge of PTMs is critical for achieving basic biology functions, the human diseases' detection, and drug target [8, 9]. It was pointed that an increasing number of modification sites can be identified with the methods of machine learning. Nevertheless, the majority of machine learning approaches and experimental ones are inherently expensive and time consuming. Therefore, constructing an accurate and effective identification algorithm seems to be an urgent issue in the field of computational biology.

Citrullination, which can be treated as a special type of deamination, is one of the most universal type in the level of post-translational modification [10, 11]. Citrullination has been reported in several biological processions, including cytoplasmic, nucleic, and membrane [12]. In order to have a good knowledge of the mechanisms of citrullination, one of the most significant steps can be regarded as the effective and accurately classification on the modification sites and nonmodification ones. It was pointed that several proteomics approaches, which include immune detection [13], colorimetric detection [14], and mass spectrometry [15, 16], should be utilized in this field. Nevertheless, these abovementioned methods'

experimental approaches can be regarded to be time consuming to some degree [17, 18].

With the development of machine learning and artificial intelligence, some methods in silicon have been widely utilized in the area of bioinformatics. It was pointed that computational tools, including phosphorylation [19], methylation [20], acetylation [21], ubiquitination [22], carbonylation [23–25], succinylation [26], malonylation, S-sulfenylation [27], and S-nitrosylation sites [28], have been proposed. Currently, Zhang et al. [29] initially proposed a computational approach to identification of such modification residues. Meanwhile, such work has the ability to remove some noise and redundant features [30, 31]. However, these subtle performances of such algorithms cannot be neglected. In order to design an effective and accurate algorithm to classify the citrullination sites in this work, we noted that the available features and the classification model can be regarded as basic elements in this classification problem.

The CirBiTree, whose full name is citrullination site identification with a fuzzy neural network and flexible neural tree, has been proposed in this work. First of all, we utilize the biprofile Bayesian to extract peptide sequence information. A flexible neural tree and fuzzy neural network are employed as the classification model in the second step. The most available length of identified peptides has been selected in the final step. To evaluate the performance of the proposed methods, some state-of-the-art methods have been used for comparison. CirBiTree can achieve 83.07% in sn%, 80.50% in sp, 0.8201 in F1, and 0.6359 in MCC, respectively, and the outlines are shown in Figure 1.

2. Materials and Methods

2.1. Dataset. In the work, we take advantage of the training dataset [29], established by Zhang et al., to train and test the proposed algorithm. The dataset contains 116 modification sites and 332 nonmodification ones in the level of citrullination. Meanwhile, each sample has been demonstrated as the style of peptide, whose center amino acid residue is the potential modification site. Therefore, the length of the peptide should be discussed in this experiment. According to such a situation, length ranges from 15 to 21 in the predicted peptide segments are chosen. So as to easily understand the length, we give an example in this section. A sample can be demonstrated as a peptide segment of length 21 in the employed dataset. In order to ensure the same length of each sample, some added residues 'X' can be filled in the positions.

2.2. Biprofile Bayesian. The biprofile Bayesian feature set is an original type of an encoding approach in the field of bioinformatics [32]. The encoding approach is based on the statistical theories. For instance, an employed sample, which includes n length peptide segments, makes a predicted center modification residue, upstream side and downstream side. The potential predicted samples can be defined as two groups. These two groups include one negative sample group and one positive sample group. Therefore, we can give the definition that the sample in the positive group can be treated as the C_p and the sample in the negative group can be treated as the C_n. The C_p is the citrullination center site in the predicted sample, and the C_n is the noncitrullination center site in the predicted dataset. With statistical theories, each amino acid residue can be defined mutually independent, and the posterior's probability of the peptide for the two types can be shown as the following equations:

$$\begin{aligned} P\left(C_p|P\right) &= \frac{P\left(P|C_p\right)P\left(C_p\right)}{P\left(P\right)}, \\ &= \prod_{i=1}^{\text{length}} \frac{P\left(p_i|C_p\right)P\left(C_p\right)}{P\left(P\right)}, \end{aligned} \tag{1}$$

$$\begin{aligned} P\left(C_n|P\right) &= \frac{P\left(P|C_n\right)P\left(C_n\right)}{P\left(P\right)}, \\ &= \prod_{i=1}^{\text{length}} \frac{P\left(p_i|C_n\right)P\left(C_n\right)}{P\left(P\right)}. \end{aligned} \tag{2}$$

Then, we may update equations (1) to (2) into the index form as follows:

$$\begin{aligned} \log\left(P\left(C_p|P\right)\right) = &\sum_{i=1}^{\text{length}} \log\left(P\left(p_i|C_p\right)\right) - \log\left(P\left(P\right)\right) \\ &+ \log\left(P\left(C_p\right)\right), \end{aligned} \tag{3}$$

$$\begin{aligned} \log\left(P\left(C_n|P\right)\right) = &\sum_{i=1}^{\text{length}} \log\left(P\left(p_i|C_n\right)\right) - \log\left(P\left(P\right)\right) \\ &+ \log\left(P\left(C_n\right)\right). \end{aligned} \tag{4}$$

The prior distribution can follow the uniform distribution. Therefore, both the probability of negative samples and positive ones can be defined as equal. The distinguished function can be demonstrated as follows:

$$\begin{aligned} f\left(P\right) &= \text{sgn}\left(\log\left(P\left(C_p|P\right)\right) - \log\left(P\left(C_n|P\right)\right)\right), \\ &= \text{sgn} \sum_{i=1}^{\text{length}} \left(\log\left(P\left(p_i|C_p\right)\right) - \log\left(P\left(p_i|C_n\right)\right)\right), \end{aligned} \tag{5}$$

With Shao's approach, equation (5) can be redefined as follows:

$$f\left(P\right) = \text{sgn}\left(\overrightarrow{W} \cdot \overrightarrow{P}\right). \tag{6}$$

2.3. Flexible Neural Tree. The flexible neural tree (FNT), considered as a special neural network, has been proposed by Bao et al. [18, 33]. Such model has the ability to regulate the neural network with special strategies. FNT has been widely utilized in the field of machine learning. The main steps of FNT are shown in the following section.

First of all, the flexible neural tree utilizes instruction set to generate population with the following equations:

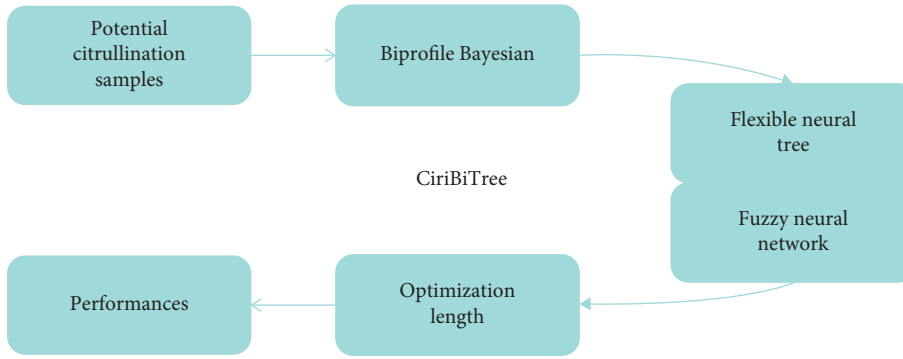

Figure 1: Outlines of CirBiTree.

$$\text{Instructor\_Set} = \text{Operation\_Set} \cup \text{Variable\_Set}, \tag{7}$$

$$\text{Operation\_Set} = \{+_1, +_2, \cdots, +_m\}, \tag{8}$$

$$\text{Variable\_Set} = \{x_1, x_2, \cdots, x_n\}, \tag{9}$$

where the instruction group consists of two operating subgroups, including the operation subgroup and the variable subgroup. The operation set $+_i$ includes several operation processions, and the variable set x_i includes several values. Then, the employed flexible activation function is described in the following equation:

$$f(m_i, n_i, x) = e^{-((x-m_i)/n_i)^2}. \tag{10}$$

In the next step, the output can be computed by the method of recursion in each neural node. For each operation set element $+i$, the total excitation can be calculated as follows:

$$\text{network}_i = \sum_{j=1}^{i} \omega_j \times y_j, \tag{11}$$

where $y_j \cdot (j = 1, 2, \ldots, i)$ is the input to node $+i$. The output of the node $+i$ is computed in as follows:

$$\text{out}_i = f(m_i, n_i, \text{network}_i) = e^{-((\text{network}_i - m_i)/n_i)^2}. \tag{12}$$

2.4. Fuzzy Neural Network. In this section, we introduce a special type of fuzzy neural network, whose name is reinforced hybrid interval fuzzy neural networks (RHIFNNs). Such model can be employed as a classification model in the field of machine learning. In the proposed classification model, the membership intervals are obtained on a basis of the membership grades produced by the two methods being realized for different values of the fuzzy parameters.

$$\begin{aligned} u^i &= \min\left(u_{\text{ik}}^1, u_{\text{ik}}^2\right), \\ \bar{u}^i(x) &= \max\left(u_{\text{ik}}^1, u_{\text{ik}}^2\right), \end{aligned} \tag{13}$$

where u_{ik}^1 is the membership grade formed by the Fuzzy C means when being run for the fuzzy parameter m_1, while u_{ik}^1 is the membership grade produced by the Fuzzy C means with the value of the fuzzy parameter set to m_2.

The consequent part of fuzzy rule, $Y^p{}_i$ can be treated as an interval, $Y_i^p = [y_{il}^p, y_{ir}^p]$, where

$$\begin{aligned} y_{\text{il}}^p &= c_{i0}^p - s_{i0}^p + \sum_{j=1}^{n} c_{\text{ij}}^p x_j - \sum_{j=1}^{n} \left|x_j\right| s_{\text{ij}}^p, y_{\text{ir}}^p \\ &= c_{i0}^p + s_{i0}^p + \sum_{j=1}^{n} c_{\text{ij}}^p x_j + \sum_{j=1}^{n} \left|x_j\right| s_{\text{ij}}^p, \end{aligned} \tag{14}$$

where $i = 1, 2, \ldots, c$ and $p = 1, 2, \ldots, q$ are the indexes of fuzzy class in this model. The model output can be calculated as

$$g_p(x) = \frac{y_l^p + y_r^p}{2}. \tag{15}$$

3. Results and Discussions

3.1. Performance Measurements. In this classification problem, samples can be defined as two types, including the positive samples and the negative samples. Defined positive samples mean the peptide segments, whose center lysine residues have the acetylation modification. On the contrary, the defined negative samples mean the peptide segments, whose center lysine residues do not have the acetylation modification. According to the definition of the classified samples, they can cause the four results in the common situation. We can easily obtain these formulations, including sensitivity, specificity, accuracy, $F1$ scores, and MCC. Also, the detailed information is given as follows:

$$\text{Sn} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{16}$$

$$\text{Sp} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \tag{17}$$

$$F1 = \frac{2\text{TP}}{2\text{TP} + \text{FN} + \text{FP}}, \tag{18}$$

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

where P is the scale of positive samples and N is the scale of negative ones. T is a set of the true predicted result, and F is a set of the false predicted result.

Table 1 summarizes that several different types of features have been employed to be compared with the proposed method. All the abovementioned features, namely, binary encoding, AA composition, grouping AA composition, physicochemical properties, KNN Features, Secondary Tendency Structure, PSSM, and BPB, have been tested in the proposed method. Our approach can get the performances that the proposed method can achieve: 78.19% in sn%, 79.28% in sp, 0.7862 in $F1$, and 0.5747 in MCC, respectively.

Table 2 demonstrates several art-of-the-state tools and approaches that have been employed to be compared to the proposed algorithm. Meanwhile, the length is 15.

Table 3 shows several art-of-the-state methods' results. In particular, our proposed algorithm can achieve 80.09% in sn %, 78.86% in sp, 0.7960 in $F1$, and 0.5896 in MCC, respectively. Meanwhile, we find that some features have different functions in this type modification site classification.

From Table 4, several art-of-the-state tools and approaches have been employed to be compared the proposed algorithm, while the length is equal to 17.

From Table 5, it can be seen that the proposed method can achieve 81.01% in sn%, 80.09% in sp, 0.8064 in $F1$, and 0.6111 in MCC, respectively. Meanwhile, we find that some features have different functions in this type modification site classification.

From Table 6, several art-of-the-state tools and approaches have been employed to be compared the proposed algorithm, while the length is equal to 19.

From Table 7, the proposed method can achieve 83.07% in sn%, 80.50% in sp, 0.8201 in $F1$, and 0.6359 in MCC, respectively. Meanwhile, we find that some features have different functions in this type modification site classification.

From Table 8, it can be seen that several art-of-the-state tools and approaches have been employed to be compared the proposed algorithm, while the length is equal to 21. The ROC curves of the art-of-the-state methods have been demonstrated in Figure 2.

It was pointed that the compared features and art-of-the-state approaches have some good performances in this classification issue. The proposed method has the ability, which is more accurate, in these candidate lengths. Meanwhile, we can easily find out that the different lengths of the amino acid residue have the different performances. We can get the conclusion that the most available length among the employed candidate ones is 21. The distances of upstream and downstream are equal to 10.

In order to demonstrate the performances of the CiBiTree, some art-of-the-state machine learning methods, including random forest, neural network, support vector machine (SVM), and k nearest neighbor (KNN), have been employed to be compared with it. The ROC curves of the different machine learning methods have been demonstrated in Figure 3.

From Table 9, we can easily find out that the proposed method has better performance than other machine learning methods in this field.

Table 1: The performances of different features in length 15.

Features	Sn (%)	Sp (%)	$F1$	MCC
Binary encoding	55.36	75.25	0.6147	0.3123
AA composition	62.87	60.53	0.6214	0.2341
Grouping AA composition	68.33	71.94	0.6959	0.4030
Physicochemical properties	72.06	73.04	0.7241	0.4510
KNN features	74.37	65.70	0.7128	0.4023
Secondary tendency structure	66.47	77.09	0.7019	0.4380
PSSM	67.75	76.75	0.7095	0.4469
BPB	71.06	76.29	0.7297	0.4741
Proposed algorithm	78.19	79.28	0.7862	0.5747

Table 2: The performances of different methods in length 15.

Method	Sn (%)	Sp (%)	$F1$	MCC
DNABIND [34]	69.27	67.93	0.6881	0.3720
DNAbinder [34]	68.37	73.68	0.7024	0.4211
DBD-Threader [35]	54.49	93.31	0.6761	0.5187
DNA-Prot [35]	64.96	79.49	0.7005	0.4492
iDNA-Prot [36]	73.26	73.07	0.7319	0.4633
DBPPred [37]	77.01	72.28	0.7523	0.4934
PLMLA [38]	65.67	69.11	0.6682	0.3480
Phosida [39]	75.55	83.34	0.7862	0.5908
LysAcet [40]	74.14	73.71	0.7398	0.4785
EnsemblePail [41]	74.87	70.17	0.7315	0.4509
PSKAcePred [42]	68.47	67.60	0.6817	0.3607
BRABSB [43]	78.41	69.87	0.7520	0.4846
SSPKA [44]	74.40	78.69	0.7603	0.5313
Proposed algorithm	78.19	79.28	0.7862	0.5747

Table 3: The performances of different features in length 17.

Features	Sn (%)	Sp (%)	$F1$	MCC
Binary encoding	55.51	74.89	0.6146	0.3099
AA composition	64.63	62.28	0.6388	0.2691
Grouping AA composition	71.63	71.39	0.7154	0.4301
Physicochemical properties	74.88	72.33	0.7394	0.4723
KNN features	73.69	64.62	0.7049	0.3847
Secondary tendency structure	69.52	76.49	0.7203	0.4612
PSSM	70.44	77.23	0.7291	0.4778
BPB	72.52	76.99	0.7418	0.4956
Proposed algorithm	80.09	78.86	0.7960	0.5896

Table 4: The performances of different methods in length 17.

Method	Sn (%)	Sp (%)	$F1$	MCC
DNABIND [34]	69.45	68.59	0.6915	0.3804
DNAbinder [34]	69.23	73.05	0.7058	0.4231
DBD-Threader [35]	56.88	92.74	0.6931	0.5316
DNA-Prot [35]	66.65	78.75	0.7094	0.4574
iDNA-Prot [36]	75.46	74.53	0.7511	0.5000
DBPPred [37]	78.66	73.34	0.7662	0.5208
PLMLA [38]	65.62	69.51	0.6692	0.3516
Phosida [39]	78.42	84.77	0.8099	0.6331
LysAcet [40]	77.17	73.76	0.7587	0.5095
EnsemblePail [41]	76.22	70.21	0.7400	0.4652
PSKAcePred [42]	70.87	67.44	0.6968	0.3833
BRABSB [43]	80.03	71.94	0.7692	0.5214
SSPKA [44]	75.49	78.09	0.7649	0.5360
Proposed algorithm	80.09	78.86	0.7960	0.5896

Table 5: The performances of different features in length 19.

Features	Sn (%)	Sp (%)	$F1$	MCC
Binary encoding	56.36	75.56	0.6234	0.3252
AA composition	64.77	62.79	0.6413	0.2756
Grouping AA composition	71.75	71.85	0.7178	0.4359
Physicochemical properties	75.36	73.73	0.7475	0.4910
KNN features	74.87	65.63	0.7157	0.4068
Secondary tendency structure	69.85	77.38	0.7258	0.4736
PSSM	71.17	79.29	0.7418	0.5062
BPB	72.68	78.45	0.7484	0.5121
Proposed algorithm	81.01	80.09	0.8064	0.6111

Table 6: The performances of different methods in length 19.

Method	Sn (%)	Sp (%)	$F1$	MCC
DNABIND [34]	69.64	70.86	0.7007	0.4051
DNAbinder [34]	69.75	73.56	0.7110	0.4334
DBD-Threader [35]	57.63	94.66	0.7073	0.5630
DNA-Prot [35]	67.72	80.64	0.7240	0.4877
iDNA-Prot [36]	76.69	75.48	0.7623	0.5218
DBPPred [37]	79.32	74.79	0.7756	0.5416
PLMLA [38]	65.61	69.49	0.6691	0.3513
Phosida [39]	78.58	84.77	0.8109	0.6346
LysAcet [40]	77.33	75.00	0.7644	0.5234
EnsemblePail [41]	77.20	72.20	0.7532	0.4946
PSKAcePred [42]	70.99	69.66	0.7052	0.4065
BRABSB [43]	81.07	72.12	0.7760	0.5341
SSPKA [44]	75.72	79.48	0.7717	0.5524
Proposed algorithm	81.01	80.09	0.8064	0.6111

Table 7: The performances of different features in length 21.

Features	Sn (%)	Sp (%)	$F1$	MCC
Binary encoding	57.25	77.00	0.6352	0.3493
AA composition	65.01	63.51	0.6452	0.2852
Grouping AA composition	72.31	72.11	0.7224	0.4443
Physicochemical properties	75.80	74.22	0.7521	0.5003
KNN features	75.35	66.29	0.7208	0.4181
Secondary tendency structure	70.49	78.43	0.7340	0.4907
PSSM	72.12	79.41	0.7485	0.5167
BPB	72.92	78.56	0.7504	0.5157
Proposed algorithm	83.07	80.50	0.8201	0.6359

Table 8: The performances of different methods in length 21.

Method	Sn (%)	Sp (%)	$F1$	MCC
DNABIND [34]	71.87	71.78	0.7184	0.4365
DNAbinder [34]	70.97	74.70	0.7232	0.4571
DBD-Threader [35]	58.87	95.52	0.7208	0.5846
DNA-Prot [35]	68.56	81.27	0.7321	0.5024
iDNA-Prot [36]	78.70	76.20	0.7773	0.5492
DBPPred [37]	80.19	75.19	0.7823	0.5545
PLMLA [38]	66.05	70.64	0.6760	0.3673
Phosida [39]	80.33	85.15	0.8231	0.6556
LysAcet [40]	78.36	76.00	0.7745	0.5437
EnsemblePail [41]	77.84	72.47	0.7581	0.5039
PSKAcePred [42]	72.09	70.33	0.7146	0.4243
BRABSB [43]	81.30	73.06	0.7809	0.5455
SSPKA [44]	76.10	80.55	0.7783	0.5670
Proposed algorithm	83.07	80.50	0.8201	0.6359

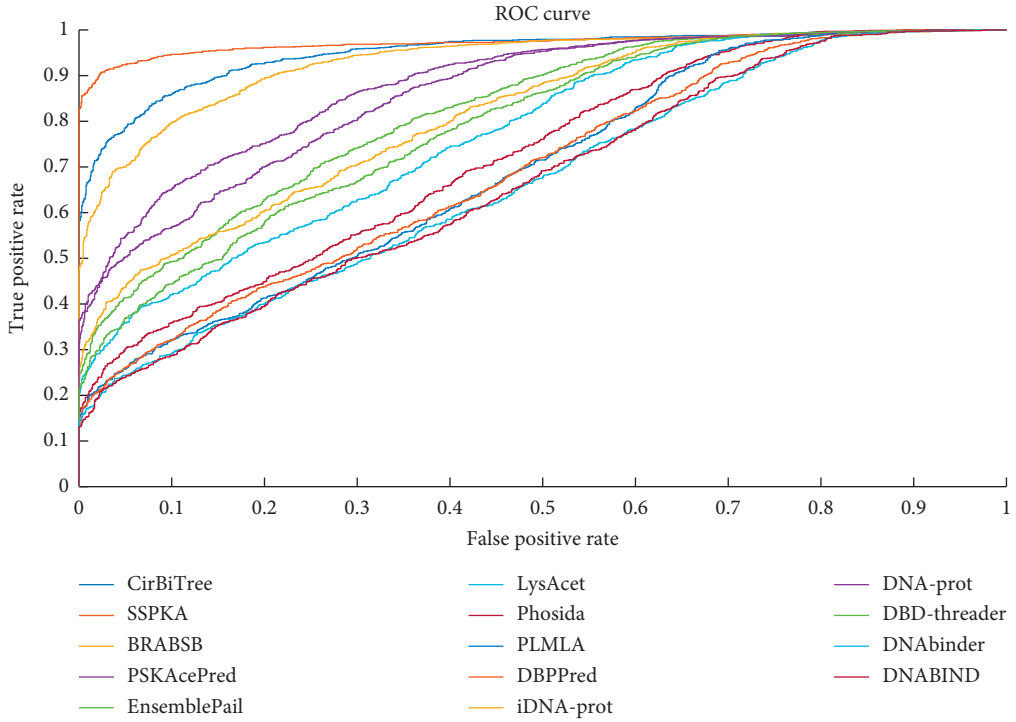

Figure 2: The ROC curves of art-of-the-state methods.

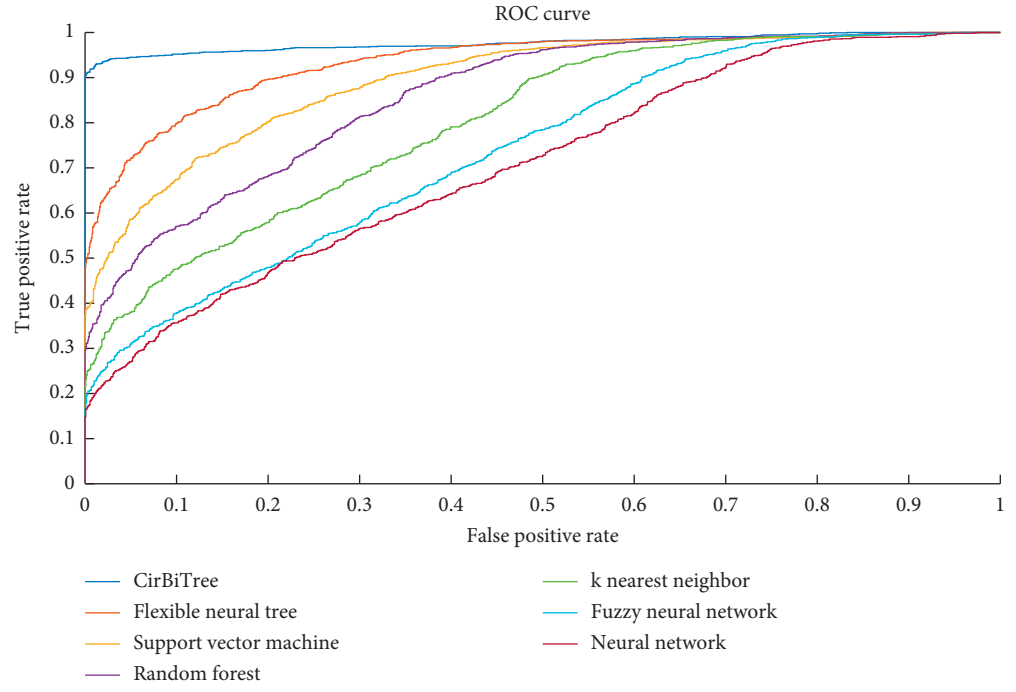

Figure 3: The ROC curves of classification algorithms in length 21.

Table 9: The performances of different methods.

Length	Model	Sn (%)	Sp (%)	$F1$	MCC
17	Neural network	50.47	58.27	0.5252	0.0877
	Fuzzy neural network	62.17	60.17	0.6155	0.2234
	k Nearest neighbor	62.87	64.28	0.6332	0.2715
	Random forest	68.28	69.28	0.6862	0.3756
	Support vector machine	64.28	78.28	0.6912	0.4298
	Flexible neural tree	72.28	68.28	0.7086	0.4059
	CirBiTree	80.09	78.86	0.7960	0.5895
19	Neural network	52.81	62.82	0.5559	0.1571
	Fuzzy neural network	64.82	62.91	0.6421	0.2774
	k Nearest neighbor	60.28	68.28	0.6279	0.2865
	Random forest	70.21	71.28	0.7059	0.4149
	Support vector machine	71.82	72.28	0.7199	0.4410
	Flexible neural tree	75.92	74.82	0.7550	0.5074
	CirBiTree	81.01	80.09	0.8064	0.6110
21	Neural network	62.87	65.81	0.6381	0.2869
	Fuzzy neural network	65.28	61.82	0.6417	0.2712
	k Nearest neighbor	60.28	63.25	0.6119	0.2354
	Random forest	70.95	75.28	0.7252	0.4627
	Support vector machine	76.28	74.82	0.7573	0.5111
	Flexible neural tree	79.28	75.28	0.7773	0.5460
	CirBiTree	83.07	80.50	0.8202	0.6359

4. Conclusions and Discussions

In this study, a novel predictor named CirBiTree has been designed to predict citrullination residues with the classification model based on a fuzzy neural network and flexible neural tree algorithm. As far we are concerned, it is the first time these abovementioned classification algorithms are utilized to the classification of the citrullination samples and noncitrullination ones. Experimental results and performances demonstrated that CirBiTree achieved an excellent performance and could be a useful bioinformatics tool to accurate identification of citrullination sites.

At the same time, several key elements of citrullination sites predicition issue should be considered. First of all, the effective description and the available features' discovery can be regarded as one of the most important elements to deal with such classification issue. On the one hand, several classical and typical methods should be utilized in this field. On the other hand, some potential information should be found with the deep learning approaches. Secondly, the high-effective classification algorithms should be proposed in the field of machine learning and artificial intelligence. With the development of deep learning, the deep learning methods can be utilized in this field. Meanwhile, it was pointed that the real-time capability should be taken into account in the model construction.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflicts of interest.

Authors' Contributions

C.S. conceived the method, designed the method, designed the website of this algorithm, and conducted the experiments, and H.W. wrote the main manuscript text. All authors reviewed the manuscript.

References

[1] M. Mann and O. N. Jensen, "Proteomic analysis of post-translational modifications," *Nature Biotechnology*, vol. 21, no. 3, pp. 255–261, 2003.

[2] E. Appella and C. W. Anderson, "Post-translational modifications and activation of p53 by genotoxic stresses," *European Journal of Biochemistry*, vol. 268, no. 10, pp. 2764–2772, 2001.

[3] G. Walsh and R. Jefferis, "Post-translational modifications in the context of therapeutic proteins," *Nature Biotechnology*, vol. 24, no. 10, pp. 1241–1252, 2006.

[4] S. Westermann and K. Weber, "Post-translational modifications regulate microtubule function," *Nature Reviews Molecular Cell Biology*, vol. 4, no. 12, pp. 938–948, 2003.

[5] J. N. Keller, K. B. Hanni, and W. R. Markesbery, "Impaired proteasome function in alzheimer's disease," *Journal of Neurochemistry*, vol. 75, no. 1, pp. 436–439, 2001.

[6] R. B. Maccioni, J. P. Muñoz, and L. Barbeito, "The molecular bases of alzheimer's disease and other neurodegenerative disorders," *Archives of Medical Research*, vol. 32, no. 5, pp. 367–381, 2001.

[7] A. Ishigami and N. Maruyama, "Importance of research on peptidylarginine deiminase and citrullinated proteins in age-related disease," *Geriatrics & Gerontology International*, vol. 10, 2010.

[8] P. Mangat, N. Wegner, P. J. Venables, and J. Potempa, "Bacterial and human peptidylarginine deiminases: targets for inhibiting the autoimmune response in rheumatoid arthritis?" *Arthritis Research & Therapy*, vol. 12, no. 3, p. 209, 2010.

[9] A. Schwenzer, X. Jiang, T. R. Mikuls et al., "Identification of an immunodominant peptide from citrullinated tenascin-C as a major target for autoantibodies in rheumatoid arthritis," *Annals of the Rheumatic Diseases*, vol. 75, no. 10, pp. 1876–1883, 2016.

[10] A. Brill, T. A. Fuchs, A. S. Savchenko et al., "Neutrophil extracellular traps promote deep vein thrombosis in mice," *Journal of Thrombosis and Haemostasis*, vol. 10, no. 1, pp. 136–144, 2012.

[11] W. J. Van Venrooij and G. J. M. Pruijn, "Citrullination: a small change for a protein with great consequences for rheumatoid arthritis," *Arthritis Research*, vol. 2, no. 4, pp. 249–251, 2000.

[12] Q. Guo, M. T. Bedford, and W. Fast, "Discovery of peptidylarginine deiminase-4 substrates by protein array: antagonistic citrullination and methylation of human ribosomal protein S2," *Molecular BioSystems*, vol. 7, no. 7, pp. 2286–2295, 2011.

[13] S. Wang and Y. Wang, "Peptidylarginine deiminases in citrullination, gene regulation, health and pathogenesis," *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, vol. 1829, no. 10, pp. 1126–1135, 2013.

[14] K. L. Bicker, V. Subramanian, A. A. Chumanevich, L. J. Hofseth, and P. R. Thompson, "Seeing Citrulline:

development of a phenylglyoxal-based probe to visualize protein citrullination," *Journal of the American Chemical Society*, vol. 134, no. 41, pp. 17015–17018, 2012.

[15] M. Stensland, A. Holm, A. Kiehne, and B. Fleckenstein, "Targeted analysis of protein citrullination using chemical modification and tandem mass spectrometry," *Rapid Communications in Mass Spectrometry*, vol. 23, no. 17, pp. 2754–2762, 2009.

[16] M. Hermansson, K. Artemenko, E. Ossipova et al., "MS analysis of rheumatoid arthritic synovial tissue identifies specific citrullination sites on fibrinogen," *Proteomics-Clinical Applications*, vol. 4, no. 5, pp. 511–518, 2010.

[17] W. Bao, B. Yang, D.-S. Huang et al., "IMKPse: identification of protein malonylation sites by the key features into general PseAAC," *IEEE Access*, vol. 7, pp. 54073–54083, 2019.

[18] W. Bao, D. Wang, and Y. Chen, "Classification of protein structure classes on flexible neutral tree," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 14, no. 5, pp. 1122–1133, 2017.

[19] W.-R. Qiu, X. Xiao, Z.-C. Xu, and K.-C. Chou, "iPhos-PseEn: identifying phosphorylation sites in proteins by fusing different pseudo components into an ensemble classifier," *Oncotarget*, vol. 7, no. 32, pp. 51270–51283, 2016.

[20] W. Qiu, B. Sun, X. Xiao, Z. Xu, J. Jia, and K. Chou, "iKcr-PseEns: identify lysine crotonylation sites in histone proteins with pseudo components and ensemble classifier," *Genomics*, vol. 110, pp. 239–246, 2017.

[21] J. Gao, X. Tao, J. Zhao, Y. Feng, Y. Cai, and N. Zhang, "Computational prediction of protein epsilon lysine acetylation sites based on a feature selection method," *Combinatorial Chemistry & High Throughput Screening*, vol. 20, 2017.

[22] Y. Cai, T. Huang, L. Hu, X. Shi, L. Xie, and Y. Li, "Prediction of lysine ubiquitination with mRMR feature selection and analysis," *Amino Acids*, vol. 42, no. 4, pp. 1387–1395, 2012.

[23] M. A. M. Hasan, J. Li, S. Ahmad, and M. K. I. Molla, "predCarsite: carbonylation sites prediction in proteins using support vector machine with resolving data imbalanced issue," *Analytical Biochemistry*, vol. 525, pp. 107–113, 2017.

[24] X. Cheng, X. Xiao, and K.-C. Chou, "pLoc-mEuk: predict subcellular localization of multi-label eukaryotic proteins by extracting the key GO information into general PseAAC," *Genomics*, vol. 110, no. 1, pp. 50–58, 2018.

[25] W. Bao, C.-A. Yuan, Y. Zhang et al., "Mutli-features prediction of protein translational modification sites," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 15, no. 5, pp. 1453–1460, 2018.

[26] J. Jia, Z. Liu, X. Xiao, B. Liu, and K.-C. Chou, "iSuc-PseOpt: identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset," *Analytical Biochemistry*, vol. 497, pp. 48–56, 2016.

[27] Y. Xu, Z. Wang, C. Li, and K. Chou, "iPreny-PseAAC: identify C-terminal cysteine prenylation sites in proteins by incorporating two tiers of sequence couplings into PseAAC," *Medicinal Chemistry*, vol. 13, pp. 544–551, 2017.

[28] B.-Q. Li, L.-L. Hu, S. Niu, Y.-D. Cai, and K.-C. Chou, "Predict and analyze S-nitrosylation modification sites with the mRMR and IFS approaches," *Journal of Proteomics*, vol. 75, no. 5, pp. 1654–1665, 2012.

[29] Q. Zhang, X. Sun, K. Feng et al., "Predicting citrullination sites in protein sequences using mRMR method and random forest algorithm," *Combinatorial Chemistry & High Throughput Screening*, vol. 20, pp. 164–173, 2017.

[30] W. Bao, B. Yang, R. Bao, and Y. Chen, "LipoFNT: lipoylation sites identification with flexible neural tree," *Complexity*, vol. 2019, p. 9, 2019.

[31] W. Bao, B. Yang, D. Li, Z. Li, Y. Zhou, and R. Bao, "CMSENN: computational modification sites with ensemble neural network," *Chemometrics and Intelligent Laboratory Systems*, vol. 185, pp. 65–72, 2019.

[32] J. Shao, D. Xu, S. Tsai, Y. Wang, and S. Ngai, "Computational identification of protein methylation sites through Bi-profile Bayes feature extraction," *PLoS One*, vol. 4, 2009.

[33] W. Bao, Y. Chen, and D. Wang, "Prediction of protein structure classes with flexible neural tree," *Bio-medical Materials and Engineering*, vol. 24, no. 6, pp. 3797–3806, 2014.

[34] A. Szilágyi and J. Skolnick, "Efficient prediction of nucleic acid binding function from low-resolution protein structures," *Journal of Molecular Biology*, vol. 358, no. 3, pp. 922–933, 2006.

[35] K. K. Kumar, G. Pugalenthi, and P. N. Suganthan, "DNA-Prot: identification of DNA binding proteins from protein sequence information using random forest," *Journal of Biomolecular Structure and Dynamics*, vol. 26, no. 6, pp. 679–686, 2009.

[36] W. Z. Lin, J. A. Fang, X. Xiao, and K. C. Chou, "iDNA-Prot: identification of DNA binding proteins using random forest with grey model," *PLoS One*, vol. 6, Article ID e24756, 2011.

[37] L. Song, D. Li, X. Zeng, Y. Wu, L. Guo, and Q. Zou, "nDNA-prot: identification of DNA-binding proteins based on unbalanced classification," *BMC Bioinformatics*, vol. 15, no. 1, p. 298, 2014.

[38] S.-P. Shi, J.-D. Qiu, X.-Y. Sun, S.-B. Suo, S.-Y. Huang, and R.-P. Liang, "PLMLA: prediction of lysine methylation and lysine acetylation by combining multiple features," *Molecular Biosystems*, vol. 8, no. 5, pp. 1520–1527, 2012.

[39] G. Florian, R. Shubin, C. Chunaram, C. Jürgen, and M. Matthias, "Predicting post-translational lysine acetylation using support vector machines," *Bioinformatics*, vol. 26, p. 1666, 2010.

[40] L. Songling, L. Hong, L. Mingfa, S. Yu, X. Lu, and L. Yixue, "Improved prediction of lysine acetylation by support vector machines," *Protein & Peptide Letters*, vol. 16, 2009.

[41] Y. Xu, X.-B. Wang, J. Ding, L.-Y. Wu, and N.-Y. Deng, "Lysine acetylation sites prediction using an ensemble of support vector machine classifiers," *Journal of Theoretical Biology*, vol. 264, no. 1, pp. 130–135, 2010.

[42] S. B. Suo, J. D. Qiu, S. P. Shi et al., "Position-Specific analysis and prediction for protein lysine acetylation based on multiple features," *PLoS One*, vol. 7, Article ID e49108, 2012.

[43] J. Shao, D. Xu, L. Hu et al., "Systematic analysis of human lysine acetylation proteins and accurate prediction of human lysine acetylation through bi-relative adapted binomial score Bayes feature representation," *Molecular Biosystems*, vol. 8, no. 11, pp. 2964–2973, 2012.

[44] Y. Li, M. Wang, H. Wang et al., "Accurate in silico identification of species-specific acetylation sites by integrating protein sequence-derived and functional features," *Scientific Reports*, vol. 4, p. 5765, 2014.