IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 20, 2019, accepted February 6, 2019, date of publication March 5, 2019, date of current version May 3, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2900275

IMKPse: Identification of Protein Malonylation Sites by the Key Features Into General PseAAC

WENZHENG BAO[1], BIN YANG[2], DE-SHUANG HUANG[3], DONG WANG[4], QI LIU[5], YUE-HUI CHEN[4], AND RONG BAO[1]

[1]School of Information and Electrical Engineering, Xuzhou University of Technology, Xuzhou 221018, China
[2]School of Information Science and Engineering, Zaozhuang University, Zaozhuang 277100, China
[3]Institute of Machine Learning and Systems Biology, School of Electronics and Information Engineering, Tongji University, Shanghai 201804, China
[4]School of Information Science, University of Jinan, Jinan 250022, China
[5]Affiliated Hospital, Xuzhou Medical University, Xuzhou 221006, China

Corresponding authors: Wenzheng Bao (baowz55555@126.com) and Rong Bao (baorong@xzit.edu.cn)

This work was supported by the grants of the National Science Foundation of China, under Grant 61873270 and Grant 61702445.

ABSTRACT Currently, lysine malonylation is treated as one of the most key protein post translational modification in the field of biology and lysine plays a significant role for the regulation of several biological processions. Therefore, accurately identification such modification type will make contributions to understanding their biological processions in this field. The experimental approaches to identify such type of modification sites are time-wasting and laborious in some degree. So, it is necessary and urgent to design and propose computational biology approaches to identify these sites. In this paper, we proposed the IMKPse model that utilized general PseAAC as the classification features and employed flexible neural tree as classification model. In order to deal with the overfitting problem, we utilized the independent datasets of each species. More specifically, such algorithm initially employed amino acid properties from the general PseAAC as the candidate features. With the comparison of candidate features, such a method has the ability to finding out the top five features among them. When evaluated on three data sets in testing set, IMKPse obtained MCC value of 0.9185, 0.9097, and 0.9525 in three species, including E.coli, M.musculus, and H.sapiens, respectively. Meanwhile, IMKPse obtained MCC value of 0.9149, 0.9060, and 0.9467, respectively, in the independent sets. In addition, then, we make some combinations among the top five features. The results demonstrate that the proposed algorithm has superior performances than other approaches. A user-friendly web resource of IMKPSE is available at http://121.250.173.184.

INDEX TERMS Post translational modification, amino acid residues identification, flexible neural tree.

I. INTRODUCTION

Protein post translational modifications (PTMs) are made to mature proteins when they have been translated from RNA sequences [1]–[3]. PTM is one of the most efficient biological mechanisms for expanding the genetic code and for regulating cellular physiolog. A lot of PTMs involve the chemical modification to a particular amino acid residue in the protein sequence. Modification at lysine residues in protein sequence have been extensively research about half century. Dysregulation of the lysine modification pathway is associated with several serious diseases, including cancers and malignant diseases [4], [5].

The latest researches report that malonylation proteins have influence on several important cellular functions in both eukaryotic and prokaryotic organisms [6]–[8]. Unfortunately, considering its dynamic property and pretty low abundance, it can hardly detect the exact substrates or sites [9]–[11]. Indeed, a major and ongoing influence is to validate the sites of Kmal, and to understand how malonylation's functions and activities in the related fields. A list of experimental approaches, such as mass spectrometry (MS), isotopic labeling, chemical probe, affinity enrichment and label free quantitative proteomics, have been widely utilized in this field [12], [13]. Nevertheless, the experimental identification of PTM sites is regarded as both expensive and resource-wasting. So, such issue is still a challenging task. With the development of sequence analysis, the computational identification of PTM play key role in this field [14]–[18]. During last few years, several PTM identification efforts in silico have been reported and such approach can be regarded as a

The associate editor coordinating the review of this manuscript and approving it for publication was Navanietha Krishnaraj Krishnaraj Rathinam.

novel method to deal with this challenging task [19]–[25]. On the one hand, several feature description methods, including Pseudo Amino Acid Composition (PseAAC) and Pseudo K-tuple Nucleotide Composition (PseKNC), have been proposed [26]–[30]. One of the most typical and classical methods is the Pseudo, whose own several web tools, including Pse-in-One 1.0 and its update version Pse-in-One 2.0, was proposed by Chou [31], [32]. Henceforth, PseAAC has been widely utilized in nearly all the areas of computational proteomics [33]. Considering the widely and increasingly utilization, several update tools, 'PseAAC-Builder', 'propy', and 'PseAAC-General', were established [34]–[36]. 'PseAAC-Builder' and 'propy' are working for Chou's special PseAAC and 'PseAAC-General' is working for Chou's general PseAAC [37]–[39]. It was pointed that PseKNC focuses on generating various feature vectors sequences in the DNA/RNA level. It was noted that some researches have been utilized these efforts [40], [41].

On the other hand, several identification tools of other types of PTM sites have been designed and proposed with machine learning approaches. For example, lots of such tools have been based on several typical machine learning tools, including neural networks, support vector machine, K-nearest neighbor and other related methods. From the comparison of the existing identification tools, it can be easily found that the sufficient samples, available features and special classification algorithm are the basic element of high performances of PTM sites identification [42]–[44].

Considering such elements, Chou has proposed the 5 steps to deal with these issues: initially, we select the valid benchmark datasets to evaluate the classification algorithm; secondly, we formulate the identified sequence samples with available mathematical expression; thirdly, we develop an algorithm to prediction the samples; nextly we evaluate the anticipated performances of the algorithm with properly cross-validation methods; lastly, we construct a user-friendly web-resource of this algorithm is accessible to the public [44]–[46]. So, we demonstrate the above mentioned operations step by step.

II. METHODS

A. DATA COLLECTION

There exist several main steps in the identification model:

Step I: The valid benchmark datasets should be selected to train and test the proposed classification model for different organisms separately.

Step II: A series of features which can make contribution to identification modification residues accurately.

Step III: An appropriate classification algorithm should be designed and developed with the issue on the malonylation modification sites prediction.

In order to construct an effective identification model, a novel non-redundant dataset of malonylation modification sites should be constructed. First of all, all of the experimental malonylation sites, including 1746 Kmal's identification sites from 595 proteins in E.coli, 3435 Kmal's identification sites from 1174 proteins in M.musculus, 4579 Kmal's identification sites from 1660 proteins in H.sapiens were collected by searching information containing the keywords of 'malonylated' or 'malonylation' from different database, including UniProtKB/SwissProt databases and CPLM databases as well as the relevant literatures. Meanwhile, E.coli, M.musculus and H.sapiens's data limitation of other organisms can hardly take into account in this thesis. The malonylaton of lysine are widely existed in the three employed species. Therefore, we utilized the E.coli, M.musculus and H.sapiens malonylations in this work.

TABLE 1. The selected protein sequence in each species.

Species	Positive Samples	Negative Samples
E.coli	1555	7853
M.musculus	3041	27499
H.sapiens	4039	53584

And then, the experimentally identified Kmal's malonylated modification sites have been defined as positive samples. At the same time, the same type of amino acid residue excluding known manolylated sites in the selected proteins has been regarded as the negative ones, which merely contain the non-maloylated modification sites.

The next step mainly focus on eliminating sequence redundancy and avoiding overestimates of the performance of machine learning-based classifiers has been selected to generate a non-redundant subset at a sequence identity level of 30%.

Finally, all of the sequences were truncated to 25-residue symmetrical windows (-12 to 12) which could have better performance to characterize the malonylated sites. It was pointed that the head or the end of these protein sequences can hardly meet the length of symmetrical windows the char "X" could be fulfilled in this protein segments.

Toally, the non-redundant datasets include 1555, 3041, 4039 positive sites and 7853, 27499, 53584 negative sites for E.coli, M.musculus and H.sapiens, respectively. The detailed information of these data shows in table 1. In order to overcome the overfitting problem, we make divisions of these dataset into three parts, which include the training sets, the testing sets and the independent sets. The former two sets make contributions to algorithm training and finding out the top five features in each species. The independent ones are utilized to show the performances of each species in constructed algorithm.

However, the selected length of peptides should be considered 3 types in the protein sequences. The first type is the segment normal distribution in the protein sequences. The second one is the segment in the head of the protein sequences and the last one is the segment in the end of the sequences. Considering these possible situations, the three type's peptides description method of the potential

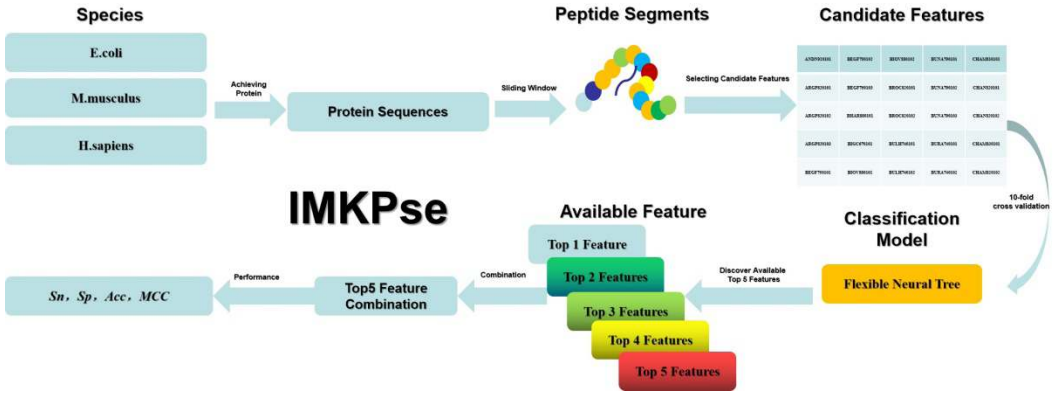

FIGURE 1. Outlines of this thesis.

classification segments may be defined as the following form.

$$\begin{aligned} &Potential\ Protein\ Segements \\ &= \begin{cases} X\ \cdots\ X + Protein\ Segements & (Head) \\ Protein\ Segements & (Normal) \\ Protein\ Segements + X \cdots X & (End) \end{cases} \end{aligned} \quad (1)$$

where, the $X \ldots X$ means the length can hardly meet the need of length of 15-peptids in the head or end situation. So, the length of $X \ldots X$ will highly depend on the length of protein segments. Therefore, the normal type can be treated as the special forms both the head type and the end ones. Given all that, the general description should be defined as the following form. While the segment belonging to the head type, the $X \cdots X_{Head}$ is non empty. While the segment belonging to the end type, the $X \cdots X_{End}$ is non empty. While the segment belonging to the normal type, both the $X \cdots X_{Head}$ and the $X \cdots X_{End}$ are empty. In one word, X can be treated as blank sites.

$$\begin{aligned} &Potential\ Protein\ Segements \\ &= X \cdots X_{Head} + Protein\ Segements + X \cdots X_{End} \end{aligned} \quad (2)$$

In total, the whole of predicted modification sites have been formulated by a general form in this work. Twenty-five types of the position specific amino acid propensity and sequence order information were utilized to convert peptide fragments into mathematical expressions for the feature construction. The predicted peptide segment has been demonstrated as the following form:

$$Potential\ Protein\ Segements = R_{-n} \cdots R_{-1} C R_1 \cdots R_n \quad (3)$$

where R_{i} can be any of the 20 native amino acids and the C is the center amino acid residue, which is lysine. When the variable i below the zero, it means the amino acid residue in the upstream. On the contrary, the variable i is a positive one, it means the amino acid residue in the downstream. Meanwhile, the value of blank amino acid properties in the head and end segments is defined as 0.

B. FEATURE REPRESENTATION

With the rocketing increasing of protein and other biology sequences, one of the most significant issues and most challenging tasks is how to demonstrate these sequences with a certain style. Unluckily, neither discrete nor vector style can hardly keep all the sequence-pattern information. Such two styles merely keep considerable sequence-order information or key pattern characteristic. PseACC has the ability to avoiding losing the sequence-pattern information.

In this work, 25 types of the properties amino acid residues among AAIndex dataset. And these feature can achieve by the Pse-in-One 2.0 software, which was designed by Bin Liu, have been employed as the classification features. It was pointed that these selected properties may play roles in the classification of the really modification sites in various degrees. So, such selected features may have their own contributions in the modification identification. Considering such situation, we establish an algorithm to select the top 5 properties among the 25 candidate ones in different species. The detailed steps of this algorithm demonstrated in the Fig 1. And the selected top 5 properties are regarded as the feature of the classification model, which is named Flexible Neural Tree.

C. FLEXIBLE NEURAL TREE ALGORITHM

The flexible neural tree algorithm, whose code can download from http://121.250.173.184, is a novel classification method. The model has a well performance in the field of classification [47]–[49]. Considering the specialty of the alternative tree, such model could be utilized in the feature selection. In this work, a tree-structural encoding approach to deal with specific instruction set has been selected for representing the neural network structure. The reason for selecting such representation is that the tree can be created and evolved utilizing the modified the construction of the neural network structure, whose ability to feature selection, in the algorithm [50], [51].

The utilized operational set F and terminal operational set T for construction the FNT model can be show as follows:

$$Instructor\_Set = \{+_2, \cdots, +_{Fn}, x_1, x_2, \cdots, x_{Fn}\} \quad (4)$$

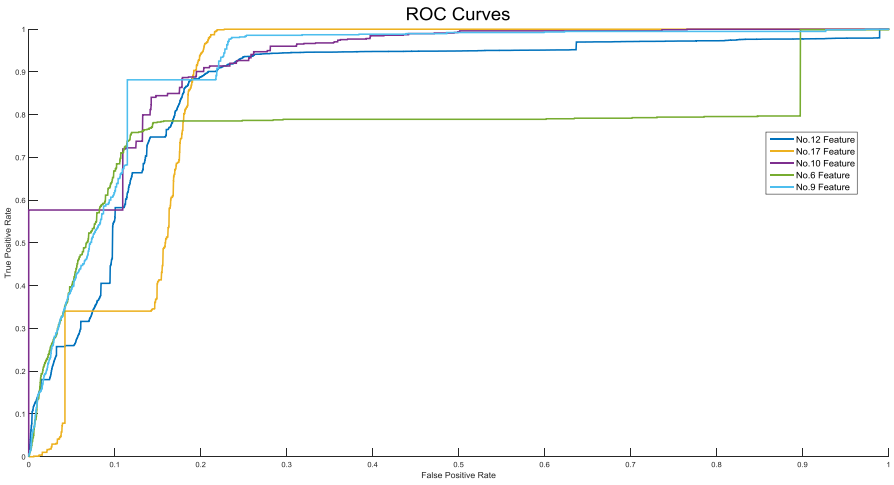

FIGURE 2. Top 5 features' ROC curves of E. coli.

where, $+a(i = 2, 3, \cdots, Fn)$ denote non-leaf nodes instructions and taking b arguments. $x_1, x_2, \cdots, x_{Fn}$ are leaf nodes instructions and without other arguments. The output of non-leaf nodes can be achieved with the flexible neuron model. From this principle, the instruction $+_i$ can be achieved with the same way of No i inputs neural node.

In the construction procession of this algorithm, if a non-terminal instruction, i.e., $+_i$ $(i = 2, 3, 4, \cdots, N)$ is selected, i real values have been generated in random. Meanwhile, such parameters can be utilized for generation the connection weight between the node $+i$ and its children node in the tree structure. At the same time, two adjustable parameters, including a_i and b_i, can be randomly selected as the parameters of the algorithm's activation function. In this work the activation function employed is *tanh* that show in the following.

$$f(x, a, b) = a * \tanh(x) + b \tag{5}$$

where, the parameter a and b can be selected. The output of such neuron +n can be achieved as follows. The general excitation of +n is

$$network = \sum_{j \in N} x_j w_j \tag{6}$$

where, x_j $(j = 1, 2, \cdots, n)$ are treated as the input nodes, which is named $+_n$. The output nodes of the algorithm $+_n$ can be computed by

$$out_i = f(network, a, b) = a * \tanh(network) + b \tag{7}$$

The classical flexible neural tree algorithm can be demonstrated as Fig 2. The output of such algorithm can be calculated with the principle, which is followed by the left-to-right in the depth-first approach, recursively.

III. RESULTS AND DISCUSSIONS

By utilizing the candidate 25 types of amino acid residues' properties from the Pse-in-one 2.0, these features play different roles in the classification of the three species in this work and the detailed steps show in Fig 1.

A. PERFORMANCE OF KMAL IN DIFFERENT SPECIES

So as to provide the easier-to-understand approach to measure the identification performance, the classical criteria was available in this thesis. According to such criteria, the rates of correct identification for the modification samples in data set and the non-modification samples in data set are respectively treated as

$$Set^+ = \frac{S^+ - S^+_-}{S^+} \text{ the modification sites} \tag{8}$$

$$Set^- = \frac{S^- - S^-_+}{S^-} \text{ the non-modification sites} \tag{9}$$

where, S^+ means the total number of the modification sites investigated, whereas S^+_- the number of the modification sites incorrectly classified as the non-modification ones; S^- the total number of the non-modification sites investigated, whereas S^-_+ the number of the non-modification sites incorrectly classifies as the modification ones. The overall success identification rate is defined by

$$Sample\_Set = \frac{Set^+S^+ + Set^-S^-}{S^+ + S^-} = 1 - \frac{S^+_- + S^-_+}{S^+ + S^-} \tag{10}$$

It was pointed that while $Set^+ = Set^- = 1$ and $S^+_- = S^-_+ = 0$, when both the modification sites and the non-modification sites are classified, the overall success rate $Sample_{Set=1}$. Otherwise, the overall success rate is lower than 1.

On the other hand, it is noted that the following equation set is utilized for checking the performance of a classification algorithm.

$$Sn = \frac{TP}{TP + FN} \tag{11}$$

$$Sp = \frac{TN}{TN + FP} \tag{12}$$

TABLE 2. Performances of potential feature of *E.coli*in training set.

Feature	*Sn(%)*	*Sp(%)*	*Acc(%)*	*MCC*
1	99.67	82.69	91.18	0.8357
2	97.85	64.80	81.33	0.6638
3	95.47	90.32	92.90	0.8591
4	90.69	93.88	92.28	0.8461
5	93.23	79.28	86.25	0.7322
6	99.41	89.59	94.50	0.8943
7	96.99	59.16	78.08	0.6067
8	94.94	70.75	82.85	0.6770
9	96.99	92.12	94.56	0.8922
10	96.28	95.39	95.84	0.9167
11	96.84	80.05	88.44	0.7799
12	94.68	96.87	95.78	0.9158
13	99.76	84.71	92.23	0.8544
14	89.22	88.36	88.79	0.7758
15	92.87	87.59	90.23	0.8057
16	96.91	38.13	67.52	0.4332
17	94.38	94.67	94.53	0.8905
18	96.85	86.54	91.69	0.8383
19	97.43	90.56	94.00	0.8820
20	99.51	36.63	68.07	0.4649
21	96.90	60.59	78.74	0.6170
22	96.97	73.99	85.48	0.7291
23	98.01	50.33	74.17	0.5499
24	98.25	82.24	90.24	0.8154
25	99.06	47.90	73.48	0.5465

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{13}$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{14}$$

where, *TP* means the true positive; *TN* is the true negative; *FP* is the false positive and *FN* means the false negative. *Sn* is the abbreviation of sensitivity, *Sp* is the abbreviation of specificity, *Acc*means the accuracy and *MCC* is the abbreviation of Mathew's correlation coefficient. Meanwhile, the relationships among these parameters show in the following.

$$TP = S^{+} - S^{+}_{-} \tag{15}$$

$$TN = S^{-} - S^{-}_{+} \tag{16}$$

$$FP = S^{-}_{+} \tag{17}$$

$$FN = S^{+}_{-} \tag{18}$$

It was pointed that the Mathew's correlation coefficient is usually utilized in measuring of binary classifications. While $S^{+}_{-} = S^{-}_{+} = 0$, meaning that none of the modification samples in the positive data set and none of the non-modification samples in the negative data set were non-predicted, so we can get $MCC = 1$. While $S^{+}_{-} = 0.5 * S^{+}$ and $S^{-}_{+} = 0.5 * S^{-}$, we get $MCC = 0$, meaning no better than random prediction. While $S^{+}_{-} = S^{+}$ and $S^{-}_{+} = S^{-}$, $MCC = -1$ means total mismatching between prediction and observation.

With the above mentioned performances, we can evaluate the proposed method to identification such modification type.

TABLE 3. Performances of potential feature of *M.musculus* in training set.

Feature	*Sn(%)*	*Sp(%)*	*Acc(%)*	*MCC*
1	90.60	82.50	86.55	0.7333
2	91.86	64.13	78.00	0.5828
3	91.37	89.91	90.64	0.8130
4	90.39	93.55	91.97	0.8398
5	92.93	79.16	86.04	0.7278
6	91.30	89.09	90.20	0.8041
7	90.42	58.65	74.53	0.5175
8	94.59	70.28	82.44	0.6688
9	92.62	91.55	92.08	0.8417
10	91.66	96.93	94.29	0.8871
11	92.80	79.52	86.16	0.7297
12	96.72	94.50	95.61	0.9124
13	99.27	84.40	91.83	0.8461
14	88.87	87.62	88.25	0.7650
15	92.48	87.15	89.82	0.7975
16	91.95	37.35	64.65	0.3497
17	91.42	95.19	93.31	0.8667
18	91.63	85.99	88.81	0.7774
19	95.04	90.47	92.76	0.8561
20	92.90	36.31	64.60	0.3542
21	92.91	60.20	76.55	0.5620
22	91.66	73.41	82.54	0.6618
23	92.81	50.01	71.41	0.4738
24	94.38	81.76	88.07	0.7676
25	94.97	47.12	71.05	0.4794

TABLE 4. Performances of potential feature of *H.sapiens* in training set.

Feature	*Sn(%)*	*Sp(%)*	*Acc(%)*	*MCC*
1	90.00	80.07	85.04	0.7042
2	72.77	69.79	71.28	0.4258
3	88.03	97.39	92.71	0.8580
4	92.65	68.82	80.74	0.6329
5	97.27	96.68	96.98	0.9395
6	93.13	62.52	77.83	0.5846
7	73.74	51.77	62.76	0.2615
8	98.52	36.14	67.33	0.4435
9	70.43	88.62	79.53	0.6005
10	96.68	50.38	73.53	0.5309
11	95.80	57.61	76.71	0.5779
12	91.64	60.95	76.30	0.5526
13	90.64	69.15	79.90	0.6122
14	86.81	92.54	89.68	0.7948
15	89.68	68.25	78.97	0.5931
16	90.50	60.74	75.62	0.5367
17	84.21	57.13	70.67	0.4294
18	90.85	80.03	85.44	0.7130
19	89.67	27.05	58.36	0.2145
20	89.60	47.14	68.37	0.4058
21	90.56	53.57	72.07	0.4750
22	90.75	59.87	75.31	0.5322
23	88.66	59.55	74.11	0.5039
24	86.54	93.36	89.95	0.8009
25	89.62	53.14	71.38	0.4593

So from the table 2 to 4, it is easily to find that the 25 type's candidate features play the various roles in the classification of the Kmal in E. It was pointed that the whole 25 types of

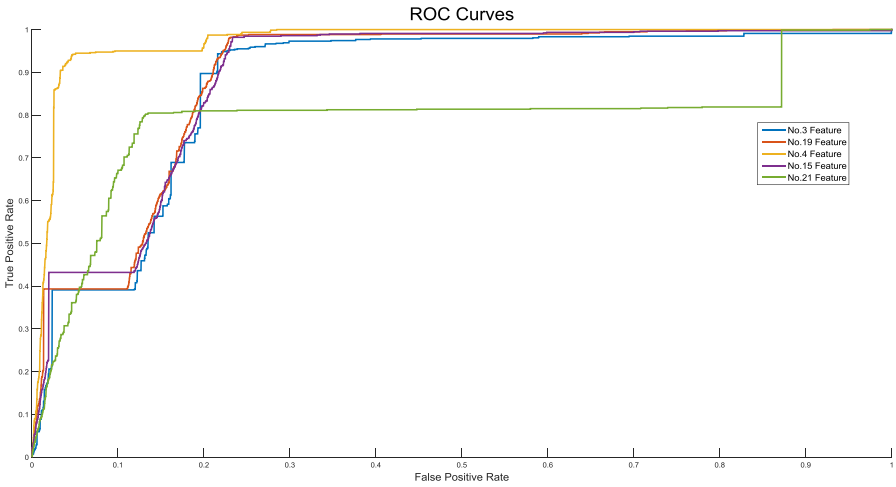

FIGURE 3. Top 5 features' ROC curves in M.musculus.

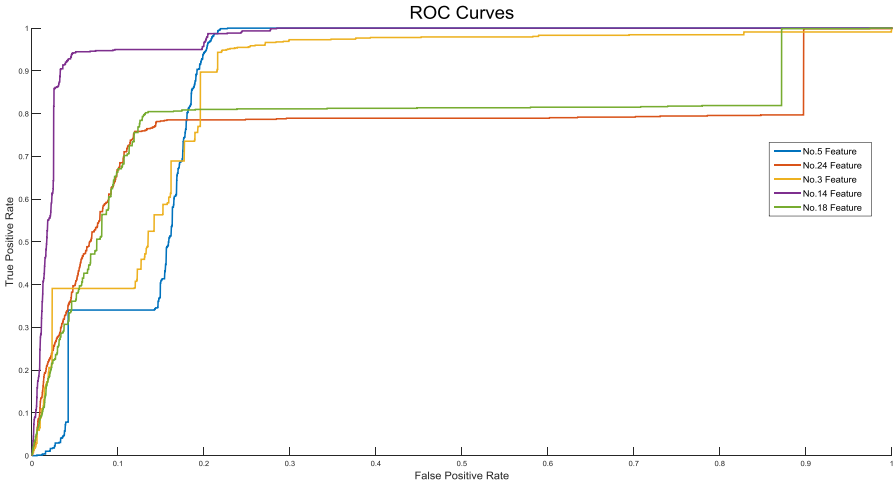

FIGURE 4. Top 5 features' ROC curves in H.sapiens.

properties have stabilities in their testing performances. With these supplementary, we can easily overcome the overfitting problem in this work. Meanwhile, these evaluation indicators on the two-type classification demonstrate the various functions in the field of identification lysine modification sites in this type of species. So from the table 2, it is easily to find that the No. features play the key role in the classification of the Kmal in M. Meanwhile, these evaluation indicators on the two-type classification demonstrate the various functions in the field of identification lysine modification sites in this type of species. From the table 2, we could obvious find out that the Sn parameter can range from 89.22% to 99.76%. The second parameter's scope can range from 36.63% to 96.87%. The Acc is from 67.52% to 95.78%. On the other hand, MCC value is from 0.4649 to 0.9158. So, the top 5 feature index is 17, 9, 6, 12 and 10 in E and the top five features and the top 5 roc curves show in Fig 2.

So from the table 3, it is easily to find that the No. features play the key role in the classification of the Kmal in M. Meanwhile, these evaluation indicators on the two-type classification demonstrate the various functions in the field of identification lysine modification sites in this type of species. From the table 3, we could obvious find out that the Sn

TABLE 5. The Top 5 features in each species.

SPECIES	No.
E.	17
	9
	6
	10
	12
H.	13
	19
	17
	10
	12
M.	18
	14
	24
	3
	5

TABLE 6. Performances of different methods of *E.coli*in testing set.

Method	*Sn(%)*	*Sp(%)*	*Acc(%)*	*MCC*
DNABIND[52]	65.78	67.97	66.88	0.3376
DNAbinder[52]	56.87	63.79	60.33	0.2071
DBD-Threader[53]	22.79	94.71	58.75	0.2519
DNA-Prot[53]	67.81	53.71	60.76	0.2174
iDNA-Prot[41]	65.71	65.72	65.72	0.3143
DBPPred[54]	75.37	72.87	74.12	0.4826
PLMLA[55]	60.80	64.70	62.70	0.2550
Phosida[56]	70.61	54.90	62.70	0.2580
LysAcet[57]	27.50	76.50	52.00	0.0450
EnsemblePail[58]	27.50	66.70	47.10	-0.0640
PSKAcePred[59]	41.20	60.80	51.00	0.0200
BRABSB[60]	51.00	60.80	55.90	0.1180
SSPKA[61]	54.90	76.50	65.70	0.3210
NN+Top1	84.03	85.98	85.01	0.7002
NN+Com-Top2	82.43	83.42	82.93	0.6585
NN+ Com-Top3	83.72	83.75	83.74	0.6747
NN+ Com-Top4	84.23	85.72	84.98	0.6996
NN+ Com-Top5	84.35	85.91	85.13	0.7027
RF+Top1	80.64	82.59	81.62	0.6324
RF+ Com-Top2	79.04	80.03	79.54	0.5907
RF+ Com-Top3	80.33	80.36	80.35	0.6069
RF+ Com-Top4	80.84	82.33	81.59	0.6318
RF+ Com-Top5	80.96	82.52	81.74	0.6349
SVM+Top1	82.73	84.68	83.71	0.6742
SVM + Com-Top2	81.13	82.12	81.63	0.6325
SVM + Com-Top3	82.42	82.45	82.44	0.6487
SVM + Com-Top4	82.93	84.42	83.68	0.6736
SVM + Com-Top5	83.05	84.61	83.83	0.6767
FNT+Top1	94.82	96.77	95.80	0.9161
FNT+ Com-Top2	93.22	94.21	93.72	0.8743
FNT+ Com-Top3	94.51	94.54	94.53	0.8905
FNT+ Com-Top4	95.02	96.51	95.77	0.9154
FNT+ Com-Top5	95.14	96.7	95.92	0.9185

Notes: In this table, the Com-Top2 means the combination of top 1 and top 2 features, whose size is the twice of the top 1. The Com-Top3 is the three times of top 1, which include top 1, 2 and 3 features. The Com-Top4 is the four times of top 1, which include top 1, 2, 3 and 4 features. The Com-Top5 contains the whole top 5 features of each species.

TABLE 7. Performances of different methods of *M.musculus*in testing set.

Method	*Sn(%)*	*Sp(%)*	*Acc(%)*	*MCC*
DNABIND[52]	62.70	64.36	63.53	0.2706
DNAbinder[52]	58.08	65.48	61.78	0.2363
DBD-Threader[53]	26.26	92.06	59.16	0.2433
DNA-Prot[53]	69.03	58.24	63.63	0.2742
iDNA-Prot[41]	68.98	66.33	67.65	0.3532
DBPPred[54]	78.15	74.25	76.20	0.5244
PLMLA[55]	50.96	51.85	51.41	0.0281
Phosida[56]	58.87	54.53	56.70	0.1342
LysAcet[57]	42.92	66.53	54.72	0.0972
EnsemblePail[58]	51.00	75.72	63.36	0.2758
PSKAcePred[59]	51.00	65.61	58.31	0.1680
BRABSB[60]	63.19	58.37	60.78	0.2159
SSPKA[61]	64.39	66.38	65.39	0.3078
NN+Top1	77.72	75.21	76.46	0.5295
NN+ Com-Top2	76.49	73.15	74.82	0.4967
NN+ Com-Top3	76.50	75.71	76.11	0.5222
NN+ Com-Top4	77.68	71.40	74.54	0.4918
NN+ Com-Top5	75.47	72.98	74.22	0.4846
RF+Top1	91.42	88.75	90.08	0.8019
RF+ Com-Top2	89.99	86.87	88.43	0.7690
RF+ Com-Top3	90.73	89.08	89.90	0.7981
RF+ Com-Top4	91.63	84.87	88.25	0.7668
RF+ Com-Top5	89.31	86.47	87.89	0.7581
SVM+Top1	95.73	93.00	94.37	0.8877
SVM + Com-Top2	94.18	90.95	92.56	0.8517
SVM + Com-Top3	94.57	93.39	93.98	0.8796
SVM + Com-Top4	95.85	88.84	92.34	0.8489
SVM + Com-Top5	92.85	90.73	91.79	0.8360
FNT+Top1	96.59	94.36	95.47	0.9097
FNT+ Com-Top2	95.40	92.40	93.90	0.8783
FNT+ Com-Top3	95.91	94.48	95.19	0.9040
FNT+ Com-Top4	97.03	90.53	93.78	0.8775
FNT+ Com-Top5	94.24	91.75	92.99	0.8601

parameter can range from 88.87% to 99.27%. The second parameter's scope can range from 37.35% to 96.93%. The Acc is from 64.60% to 95.61%. On the other hand, MCC value is from 0.3497 to 0.9124. So, the top 5 feature index is 13, 19, 17, 10 and 12 in M and the top five features and the top 5 roc curves show in Fig 3.

So from the above table 4, it is easily to find that the No. features play the key role in the classification of the Kmal in M. Meanwhile, these evaluation indicators on the two-type classification demonstrate the various functions in the field of identification lysine modification sites in this type of species. From the table 4, we could obvious find out that the

TABLE 8. Performances of different methods of *H.sapiens*in testing set.

Method	*Sn(%)*	*Sp(%)*	*Acc(%)*	*MCC*
DNABIND[52]	65.75	67.34	66.55	0.3309
DNAbinder[52]	57.89	66.88	62.39	0.2487
DBD-Threader[53]	27.30	90.59	58.94	0.2311
DNA-Prot[53]	66.76	60.73	63.74	0.2754
iDNA-Prot[41]	67.55	65.77	66.66	0.3332
DBPPred[54]	79.76	73.81	76.79	0.5367
PLMLA[55]	63.02	66.25	64.63	0.2928
Phosida[56]	55.33	58.28	56.81	0.1362
LysAcet[57]	50.33	61.55	55.94	0.1195
EnsemblePail[58]	45.73	61.74	53.73	0.0756
PSKAcePred[59]	55.32	55.78	55.55	0.1110
BRABSB[60]	61.23	66.29	63.76	0.2756
SSPKA[61]	48.22	72.47	60.35	0.2133
NN+Top1	60.29	58.37	59.33	0.1867
NN+ Com-Top2	54.46	51.94	53.20	0.0640
NN+ Com-Top3	53.02	54.37	53.70	0.0739
NN+ Com-Top4	56.00	54.41	55.21	0.1042
NN+ Com-Top5	57.56	52.40	54.98	0.0997
RF+Top1	79.18	77.21	78.20	0.5640
RF+ Com-Top2	73.31	70.86	72.08	0.4418
RF+ Com-Top3	71.87	73.22	72.54	0.4509
RF+ Com-Top4	74.88	73.23	74.06	0.4812
RF+ Com-Top5	76.42	71.26	73.84	0.4774
SVM+Top1	96.25	94.28	95.26	0.9054
SVM + Com-Top2	90.36	87.89	89.12	0.7827
SVM + Com-Top3	88.92	90.31	89.61	0.7924
SVM + Com-Top4	91.92	90.26	91.09	0.8219
SVM + Com-Top5	93.47	88.30	90.88	0.8188
FNT+Top1	98.59	96.64	97.62	0.9525
FNT+ Com-Top2	92.72	90.25	91.49	0.8300
FNT+ Com-Top3	91.31	92.67	91.99	0.8398
FNT+ Com-Top4	94.29	92.65	93.47	0.8695
FNT+ Com-Top5	95.83	90.65	93.24	0.8659

TABLE 9. Performances of different methods of *E.coli*in independent set.

Method	*Sn(%)*	*Sp(%)*	*Acc(%)*	*MCC*
DNABIND[52]	65.62	67.91	66.76	0.3354
DNAbinder[52]	56.84	63.75	60.29	0.2064
DBD-Threader[53]	22.67	94.33	58.50	0.2438
DNA-Prot[53]	67.72	53.64	60.68	0.2157
iDNA-Prot[41]	65.60	65.73	65.66	0.3133
DBPPred[54]	75.21	72.64	73.93	0.4787
PLMLA[55]	60.65	64.37	62.51	0.2504
Phosida[56]	70.56	54.65	62.60	0.2553
LysAcet[57]	27.42	76.44	51.93	0.0442
EnsemblePail[58]	27.48	66.57	47.02	-0.0647
PSKAcePred[59]	41.00	60.61	50.80	0.0164
BRABSB[60]	50.96	60.42	55.69	0.1143
SSPKA[61]	54.84	76.43	65.64	0.3203
NN+Top1	83.96	85.62	84.79	0.6959
NN+Com-Top2	82.37	83.17	82.77	0.6554
NN+ Com-Top3	83.62	83.59	83.60	0.6721
NN+ Com-Top4	84.13	85.67	84.90	0.6981
NN+ Com-Top5	84.15	85.76	84.95	0.6992
RF+Top1	80.43	82.42	81.43	0.6286
RF+ Com-Top2	78.99	79.99	79.49	0.5898
RF+ Com-Top3	80.18	80.14	80.16	0.6033
RF+ Com-Top4	80.76	82.26	81.51	0.6303
RF+ Com-Top5	80.88	82.40	81.64	0.6328
SVM+Top1	82.64	84.47	83.55	0.6711
SVM + Com-Top2	80.93	82.04	81.49	0.6298
SVM + Com-Top3	82.32	82.35	82.33	0.6467
SVM + Com-Top4	82.74	84.18	83.46	0.6693
SVM + Com-Top5	82.95	84.53	83.74	0.6749
FNT+Top1	94.67	96.42	95.55	0.9111
FNT+ Com-Top2	93.08	93.81	93.44	0.8689
FNT+ Com-Top3	94.38	94.24	94.31	0.8862
FNT+ Com-Top4	94.86	96.39	95.62	0.9126
FNT+ Com-Top5	95.00	96.48	95.74	0.9149

Sn parameter can range from 72.77% to 98.52%. The second parameter's scope can range from 36.14% to 97.39%. The Acc is from 58.36% to 96.98%. On the other hand, MCC value is from 0.2615 to 0.9395. So, the top 5 feature index is 18, 14, 24, 3 and 5 in H and the top five features show and the top 5 roc curves show in Fig 4. Meanwhile, all the top 5 features of the selected species show in table 5.

B. COMPARISON WITH OTHER METHODS

In order to evaluate the performance of the top 5 features, we make a combination of these top 5 features in each species. On the one hand, we compare the flexible neural tree with other typical machine learning approaches. On the other hand, some state-of-art amino acid sequence classification methods, including DBD-Threader, iDNA-Prot and other similar ones have been compared with the proposed algorithm. The detailed comparisons demonstrate in table 6, table 7 and table 8. Meanwhile, it was pointed that the top 5

TABLE 10. Performances of different methods of *M.musculus*in independent set.

Method	*Sn(%)*	*Sp(%)*	*Acc(%)*	*MCC*
DNABIND[52]	62.67	64.33	63.50	0.2701
DNAbinder[52]	57.85	65.39	61.62	0.2331
DBD-Threader[53]	26.04	92.01	59.03	0.2402
DNA-Prot[53]	68.83	58.15	63.49	0.2713
iDNA-Prot[41]	68.91	66.19	67.55	0.3512
DBPPred[54]	78.00	74.08	76.04	0.5212
PLMLA[55]	50.95	51.70	51.33	0.0266
Phosida[56]	58.76	54.25	56.51	0.1303
LysAcet[57]	42.84	66.36	54.60	0.0947
EnsemblePail[58]	50.96	75.42	63.19	0.2720
PSKAcePred[59]	50.96	65.41	58.18	0.1653
BRABSB[60]	63.08	58.06	60.57	0.2117
SSPKA[61]	64.37	66.30	65.33	0.3068
NN+Top1	77.57	74.99	76.28	0.5258
NN+ Com-Top2	76.37	73.06	74.71	0.4946
NN+ Com-Top3	76.33	75.50	75.91	0.5182
NN+ Com-Top4	77.51	71.18	74.34	0.4878
NN+ Com-Top5	75.31	72.96	74.13	0.4828
RF+Top1	91.41	88.67	90.04	0.8011
RF+ Com-Top2	89.97	86.80	88.39	0.7681
RF+ Com-Top3	90.65	88.87	89.76	0.7953
RF+ Com-Top4	91.50	84.60	88.05	0.7628
RF+ Com-Top5	89.15	86.36	87.75	0.7554
SVM+Top1	95.63	92.75	94.19	0.8841
SVM + Com-Top2	93.98	90.73	92.35	0.8475
SVM + Com-Top3	94.39	93.39	93.89	0.8778
SVM + Com-Top4	95.61	88.65	92.13	0.8446
SVM + Com-Top5	92.72	90.61	91.66	0.8334
FNT+Top1	96.51	94.07	95.29	0.9060
FNT+ Com-Top2	95.37	92.40	93.89	0.8781
FNT+ Com-Top3	95.76	94.33	95.04	0.9010
FNT+ Com-Top4	96.84	90.39	93.61	0.8741
FNT+ Com-Top5	94.13	91.60	92.87	0.8576

TABLE 11. Performances of different methods of *H.sapiens*in independent set.

Method	*Sn(%)*	*Sp(%)*	*Acc(%)*	*MCC*
DNABIND[52]	65.36	67.08	66.22	0.3245
DNAbinder[52]	57.73	66.69	62.21	0.2452
DBD-Threader[53]	26.90	90.40	58.65	0.2239
DNA-Prot[53]	66.52	60.45	63.49	0.2702
iDNA-Prot[41]	67.53	65.45	66.49	0.3299
DBPPred[54]	79.67	73.66	76.67	0.5343
PLMLA[55]	62.65	65.97	64.31	0.2864
Phosida[56]	55.09	58.11	56.60	0.1320
LysAcet[57]	50.25	61.20	55.72	0.1152
EnsemblePail[58]	45.56	61.39	53.47	0.0703
PSKAcePred[59]	55.01	55.67	55.34	0.1068
BRABSB[60]	61.13	66.03	63.58	0.2720
SSPKA[61]	47.84	72.23	60.03	0.2069
NN+Top1	60.17	58.14	59.15	0.1831
NN+ Com-Top2	53.99	51.57	52.78	0.0557
NN+ Com-Top3	52.88	54.26	53.57	0.0714
NN+ Com-Top4	55.61	54.28	54.94	0.0989
NN+ Com-Top5	57.46	52.35	54.91	0.0983
RF+Top1	79.03	76.82	77.92	0.5586
RF+ Com-Top2	73.26	70.59	71.93	0.4387
RF+ Com-Top3	71.58	73.02	72.30	0.4460
RF+ Com-Top4	74.53	72.96	73.75	0.4750
RF+ Com-Top5	76.14	71.03	73.59	0.4723
SVM+Top1	96.03	94.01	95.02	0.9006
SVM + Com-Top2	90.03	87.66	88.85	0.7771
SVM + Com-Top3	88.59	90.01	89.30	0.7860
SVM + Com-Top4	91.57	90.04	90.81	0.8162
SVM + Com-Top5	93.15	87.88	90.51	0.8114
FNT+Top1	98.11	96.55	97.33	0.9467
FNT+ Com-Top2	92.61	90.21	91.41	0.8284
FNT+ Com-Top3	90.95	92.62	91.79	0.8358
FNT+ Com-Top4	94.17	92.62	93.40	0.8680
FNT+ Com-Top5	95.77	90.48	93.12	0.8637

features of each spiece have many combination types. So in this thesis, we utilized the five types of combination, including top 1 (21 dimensions features), top 2 (42 dimensions features), top 3 (63 dimensions features), top 4 and top 5. The above mentioned tables demonstrate the detail performances of these combinations. Meanwhile, these comparisons show the independent sets of each species in table 9, 10 and 11.

IV. CONCLUSIONS

A great deal of information and knowledge about protein sequences with malonylated has been accumulated to date. There are still numerous undiscovered and unsolvable issues and events on the classification issue in the field of machine learning. Currently, the rocketing numbers of protein sequences have been sequenced with the High-throughput technology and methods. However, the discovering of the

properties of the amino acid level, peptide level and protein level can hardly meet the need of identification the function and structure in the field of proteomics, biostatics, bioinformatics and other similar omics. It was pointed that the size of negative samples is much larger than the positive ones. Therefore, it is a classical issue, which is a non-balanced classification problem, in the machine learning and classification. It is hard to regard that all segments carry similar structures before they bind to the component of the lysine malonylated modification sites.

Notes: In this table, the Com-Top2 means the combination of top 1 and top 2 features, whose size is the twice of the top 1. The Com-Top3 is the three times of top 1, which include top 1, 2 and 3 features. The Com-Top4 is the four times of top 1, which include top 1, 2, 3 and 4 features. The Com-Top5 contains the whole top 5 features of each species.

Systematic analysis of the Kmal sites along with information on these sites could be utilized by identifying the modified sites from the amino acid residues' properties. However, even the same post translation modification maybe fit the distinguish features in different species. In other word, some features can get ideal results in one species. Nevertheless, such features can hardly meet the need of the other species. Considering the above mentioned situation, several key information and features on the identification of the malonylation sites of different species can be achieved and caught in this work.

On the other hand, another key result of this research is demonstrated that the candidate features and properties may play various roles, including the supporter features, the opponent features and the neutral features, in this classification work. So, each selected type of candidate feature will try to find out the fittest features of identification malonylation sites in the certain species.

Here, it was pointed that unbalanced datasets, which the negative samples can reach about 7 times than the positive ones, present a hottest topic in the field of machine learning classification. In our work, the unbalanced datasets will try to avoid the imbalance influences with the preprocess steps, which the positive samples replicate themselves until the size of positive samples can generally reach the scale of the negative ones in only testing set. For future research, other properties and features, not merely the AAIndex database, will be employed and utilized to deal with the different species modification sites' identification issue. On the other hand, several novel technology and method, such as the deep neural network, should be widely utilized in such modification site and other similar modification sites in the field of machine learning and bioinformatics.

In a word, the selection the fittest features of identification modification sites seem to be one of the most important tasks in the issue of identification modification sites. Therefore, in the following work, several more reliable measurement systems should be constructed. On the other hand, the discovery of the combination of the various features and properties should pay more attention in this field.

COMPETING INTERESTS

The authors declare no competing interests.

REFERENCES

[1] T. Kouzarides, "Chromatin modifications and their function," *Cell*, vol. 128, no. 4, pp. 693–705, Feb. 2007.
[2] M. Mann and O. N. Jensen, "Proteomic analysis of post-translational modifications," *Nature Biotechnol.*, vol. 21, no. 3, pp. 255–261, Mar. 2003.
[3] C. Dai and W. Gu, "p53 post-translational modification: Deregulated in tumorigenesis," *Trends Mol. Med.*, vol. 16, no. 11, pp. 528–536, Nov. 2010.
[4] A. J. Ruthenburg, H. Li, D. J. Patel, and C. D. Allis, "Multivalent engagement of chromatin modifications by linked binding modules," *Nature Rev. Mol. Cell Biol.*, vol. 8, no. 12, pp. 983–994, Dec. 2007.
[5] J. Wysocka *et al.*, "A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling," *Nature*, vol. 442, no. 7098, pp. 86–90, Jul. 2006.
[6] J. Wysocka *et al.*, "WDR5 associates with histone H3 methylated at K4 and is essential for H3 K4 methylation and vertebrate development," *Cell*, vol. 121, no. 6, pp. 859–872, Jun. 2005.
[7] L. Zeng and M. Zhou, "Bromodomain: An acetyl-lysine binding domain," *FEBS Lett.*, vol. 513, no. 1, pp. 124–128, Feb. 2002.
[8] T. Jenuwein and C. D. Allis, "Translating the Histone Code," *Science*, vol. 293, no. 5532, pp. 1074–1080, Aug. 2001.
[9] R. Marmorstein and S. Y. Roth, "Histone acetyltransferases: Function, structure, and catalysis," *Current Opinion Genet. Develop.*, vol. 11, no. 2, pp. 155–161, Apr. 2001.
[10] A. M. Bode and Z. Dong, "Post-translational modification of p53 in tumorigenesis," *Nature Rev. Cancer*, vol. 4, no. 10, pp. 793–805, Oct. 2004.
[11] G. Walsh and R. Jefferis, "Post-translational modifications in the context of therapeutic proteins," *Nature Biotechnol.*, vol. 24, no. 10, pp. 1241–1252, Oct. 2006.
[12] S. Westermann and K. Weber, "Post-translational modifications regulate microtubule function," *Nature Rev. Mol. Cell Biol.*, vol. 4, no. 12, pp. 938–947, Dec. 2003.
[13] C. Janke and J. C. Bulinski, "Post-translational regulation of the microtubule cytoskeleton: Mechanisms and functions," *Nature Rev. Mol. Cell Biol.*, vol. 12, no. 12, pp. 773–786, Dec. 2011.
[14] Y. Xu, X. Shao, L. Wu, N. Deng, and K. Chou, "iSNO-AAPair: Incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins," *PeerJ*, vol. 1, p. e171, Oct. 2013.
[15] W. Qiu, X. Xiao, W. Lin, and K. Chou, "iMethyl-PseAAC: Identification of protein methylation sites via a pseudo amino acid composition approach," *BioMed Res. Int.*, vol. 2014, Aug. 2014, Art. no. 947416.
[16] Y. Xu, X. Wen, X. Shao, N. Deng, and K. Chou, "iHyd-PseAAC: Predicting hydroxyproline and hydroxylysine in proteins by incorporating dipeptide position-specific propensity into pseudo amino acid composition," *Int. J. Mol. Sci.*, vol. 15, no. 5, pp. 7594–7610, May 2014.
[17] X. Xiao, H. Ye, Z. Liu, J. Jia, and K. Chou, "iROS-gPseKNC: Predicting replication origin sites in DNA by incorporating dinucleotide position-specific propensity into general pseudo nucleotide composition," *Oncotarget*, vol. 7, no. 23, pp. 34180–34189, Jun. 2016.
[18] W. Chen, H. Tang, J. Ye, H. Lin, and K. Chou, "iRNA-PseU: Identifying RNA pseudouridine sites," *Mol. therapy. Nucleic Acids*, vol. 5, p. e332, Jan. 2016.
[19] J. Jia, Z. Liu, X. Xiao, B. Liu, and K. C. Chou, "iCar-PseCp: Identify carbonylation sites in proteins by Monte Carlo sampling and incorporating sequence coupled effects into general PseAAC," *Oncotarget*, vol. 7, no. 23, pp. 34558–34570, Jun. 2016.
[20] J. Jia, L. Zhang, Z. Liu, X. Xiao, and K. C. Chou, "pSumo-CD: Predicting sumoylation sites in proteins with covariance discriminant algorithm by incorporating sequence-coupled effects into general PseAAC," *Bioinformatics*, vol. 32, no. 20, pp. 3133–3141, Jun. 2016.
[21] Z. Liu, X. Xiao, D. J. Yu, J. Jia, W. R. Qiu, and K. C. Chou, "pRNAm-PC: Predicting N^6-methyladenosine sites in RNA sequences via physical-chemical properties," *Anal. Biochemistry*, vol. 497, pp. 60–67, Mar. 2016.
[22] W. R. Qiu, B. Q. Sun, X. Xiao, Z. C. Xu, and K. C. Chou, "iPTM-mLys: Identifying multiple lysine PTM sites and their different types," *Bioinformatics*, vol. 32, no. 20, pp. 3116–3123, Jun. 2016.

[23] 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, Aug. 2016.

[24] P. Feng, H. Ding, H. Yang, W. Chen, H. Lin, and K. C. Chou, "iRNA-PseColl: Identifying the occurrence sites of different RNA modifications by incorporating collective effects of nucleotides into PseKNC," *Mol. Therapy-Nucleic Acids*, vol. 7, pp. 155–163, Jun. 2017.

[25] W. Bao, Z. Huang, C. A. Yuan, and D. S. Huang, "Pupylation sites prediction with ensemble classification model," *Int. J. Data Mining Bioinf.*, vol. 18, no. 2, pp. 91–104, 2017.

[26] W. R. Qiu, S. Y. Jiang, Z. C. Xu, X. Xiao, and K. C. Chou, "iRNAm5C-PseDNC: Identifying RNA 5-methylcytosine sites by incorporating physical-chemical properties into pseudo dinucleotide composition," *Oncotarget*, vol. 8, no. 25, pp. 41178–41188, Jun. 2017.

[27] W. R. Qiu, B. Q. Sun, X. Xiao, D. Xu, and K. C. Chou, "iPhos-PseEvo: Identifying human phosphorylated proteins by incorporating evolutionary information into general PseAAC via grey system theory," *Molecular Informat.*, vol. 36, nos. 5–6, May 2017, Art. no. 1600010.

[28] W. R. Qiu, B. Q. Sun, X. Xiao, Z. C. Xu, J. H. Jia, and K. C. Chou, "iKcr-PseEns: Identify lysine crotonylation sites in histone proteins with pseudo components and ensemble classifier," *Genomics*, vol. 110, no. 5, pp. 239–246, Sep. 2017.

[29] Y. Xu, Z. Wang, C. Li, and K. C. Chou, "iPreny-PseAAC: Identify C-terminal cysteine prenylation sites in proteins by incorporating two tiers of sequence couplings into PseAAC," *Medicinal Chem.*, vol. 13, no. 6, pp. 544–551, Sep. 2017.

[30] W. Bao, Z. Jiang, and D. S. Huang, "Novel human microbe-disease association prediction using network consistency projection," *BMC Bioinf.*, vol. 18, no. 16, p. 543, Dec. 2017.

[31] K. C. Chou, "Prediction of human immunodeficiency virus protease cleavage sites in proteins," *Anal. Biochemistry*, vol. 233, no. 1, pp. 1–14, Jan. 1996.

[32] Y. D. Khan, N. Rasool, W. Hussain, S. A. Khan, and K. C. Chou, "iPhosT-PseAAC: Identify phosphothreonine sites by incorporating sequence statistical moments into PseAAC," *Anal. Biochemistry*, vol. 550, pp. 109–116, Jun. 2018.

[33] B. Liu, F. Liu, X. Wang, J. Chen, L. Fang, and K. C. Chou, "Pse-in-One: A web server for generating various modes of pseudo components of DNA, RNA, and protein sequences," *Nucleic Acids Res.*, vol. 43, pp. W65–W71, May 2015.

[34] K. C. Chou, "Impacts of bioinformatics to medicinal chemistry," *Medicinal Chem.*, vol. 11, no. 3, pp. 218–234, May 2015.

[35] L. F. Yuan, C. Ding, S. H. Guo, W. Chen, and H. Lin, "Prediction of the types of ion channel-targeted conotoxins based on feature selection techniques," *J. Biomath.*, to be published.

[36] K. C. Chou, "An unprecedented revolution in medicinal chemistry driven by the progress of biological science," *Current Topics Medicinal Chem.*, vol. 17, no. 21, pp. 2337–2358, Aug. 2017.

[37] K. C. Chou, "Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology," *Current Proteomics*, vol. 6, no. 4, pp. 262–274, Dec. 2009.

[38] C. Wei, L. Hao, and K. C. Chou, "Pseudo nucleotide composition or PseKNC: An effective formulation for analyzing genomic sequences," *Mol. Biosyst.*, vol. 11, no. 10, pp. 2620–2634, 2015.

[39] K. C. Chou, "Prediction of signal peptides using scaled window," *Peptides*, vol. 22, no. 12, pp. 1973–1979, Dec. 2001.

[40] Y. Xu, Y. X. Ding, J. Ding, Y. H. Lei, L. Y. Wu, and N. Y. Deng, "iSuc-PseAAC: Predicting lysine succinylation in proteins by incorporating peptide position-specific propensity," *Sci. Rep.*, vol. 5, no. 1, p. 10184, Jun. 2015.

[41] 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, no. 9, 2011, Art. no. e24756.

[42] W. Chen, P. M. Feng, H. Lin, and K. C. Chou, "iRSpot-PseDNC: Identify recombination spots with pseudo dinucleotide composition," *Nucleic Acids Res.*, vol. 41, no. 6, p. e68, Jan. 2013.

[43] X. Cheng, X. Xiao, and K. C. Chou, "pLoc-mPlant: Predict subcellular localization of multi-location plant proteins by incorporating the optimal GO information into general PseAAC," *Gene*, vol. 13, no. 9, pp. 1722–1727, 2017.

[44] X. Cheng, S. G. Zhao, W. Z. Lin, X. Xiao, and K. C. Chou, "pLoc-mAnimal: Predict subcellular localization of animal proteins with both single and multiple sites," *Bioinformatics*, vol. 33, no. 22, pp. 3524–3531, Jul. 2017.

[45] X. Cheng, X. Xiao, and K. C. Chou, "pLoc-mGneg: Predict subcellular localization of Gram-negative bacterial proteins by deep gene ontology learning via general PseAAC," *Genomics*, vol. 110, no. 4, pp. 231–239, Jul. 2017.

[46] C. Xiang, X. Xuan, 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, Jan. 2017.

[47] B. Wenzheng, C. Yuehui, and W. Dong, "Prediction of protein structure classes with flexible neural tree," *Bio-Medical Mater. Eng.*, vol. 24, no. 6, pp. 3797–3806, Jan. 2014.

[48] W. Bao, D. Wang, and Y. Chen, "Classification of protein structure classes on flexible neutral tree," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 14, no. 5, pp. 1122–1133, Sep. 2017.

[49] Y. Chen, B. Yang, J. Dong, and A. Abraham, "Time-series forecasting using flexible neural tree model," *Inf. Sci.*, vol. 174, nos. 3–4, pp. 219–235, Aug. 2005.

[50] Y. Chen, A. Abraham, and B. Yang, "Hybrid flexible neural-tree-based intrusion detection systems," *Int. J. Intell. Syst.*, vol. 22, no. 4, pp. 337–352, Apr. 2007.

[51] Y. Chen, A. Abraham, and B. Yang, "Feature selection and classification using flexible neural tree," *Neurocomputing*, vol. 70, no. 1, pp. 305–313, Dec. 2006.

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

[53] K. K. Kumar, G. Pugalenthi, and P. N. Suganthan, "DNA-Prot: Identification of DNA binding proteins from protein sequence information using random forest," *J. Biomol. Struct. Dyn.*, vol. 26, no. 6, pp. 679–686, Jun. 2009.

[54] 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 Bioinf.*, vol. 15, no. 1, p. 298, Dec. 2014.

[55] 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," *Mol. Biosyst.*, vol. 8, no. 5, pp. 1520–1527, 2012.

[56] G. Florian, R. Shubin, C. Chunaram, J. Cox, and M. Matthias, "Predicting post-translational lysine acetylation using support vector machines," *Bioinformatics*, vol. 26, no. 13, pp. 1666–1668, May 2010.

[57] L. Songling, L. Hong, L. Mingfa, S. Yu, X. Lu, and L. Yixue, "Improved prediction of lysine acetylation by support vector machines," *Protein Peptide Lett.*, vol. 16, no. 8, pp. 977–983, 2009.

[58] 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," *J. Theor. Biol.*, vol. 264, no. 1, pp. 130–135, 2010.

[59] S. B. Suo *et al.*, "Position-specific analysis and prediction for protein lysine acetylation based on multiple features," *Plos One*, vol. 7, no. 11, 2012, Art. no. e49108.

[60] J. Shao *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," *Mol. Biosyst.*, vol. 8, no. 11, pp. 2964–2973, 2012.

[61] Y. Li *et al.*, "Accurate in silico identification of species-specific acetylation sites by integrating protein sequence-derived and functional features," *Sci. Rep.*, vol. 4, p. 5765, Jul. 2014.

Authors' photographs and biographies not available at the time of publication.