Hindawi
Complexity
Volume 2019, Article ID 1603867, 9 pages
https://doi.org/10.1155/2019/1603867

Research Article

LipoFNT: Lipoylation Sites Identification with Flexible Neural Tree

Wenzheng Bao,[1] **Bin Yang**,[2] **Rong Bao**,[1] **and Yuehui Chen**[3]

[1]*School of Information and Electrical Engineering, Xuzhou University of Technology, Xuzhou 221018, China*
[2]*School of Information Science and Engineering, Zaozhuang University, Zaozhuang 277160, China*
[3]*School of Information Science, University of Jinan, Jinan 250022, China*

Correspondence should be addressed to Bin Yang; batsi@126.com

Received 6 January 2019; Revised 25 April 2019; Accepted 4 June 2019; Published 14 July 2019

Academic Editor: Lingzhong Guo

Lysine lipoylation is a special type of posttranslational modification in both prokaryotes' and eukaryotes' proteomics researches. Such a modification takes part in several significant biological processions and plays a key role in the cellular level. In order to construct and design an accurate classification algorithm for identifying lipoylation sites in the protein level, the computational approaches should be taken into account in this field. Meanwhile, several factors plays different role in the identification of modification sites. Considering such a situation, the foundational elements of the effective identification of modification sites are the available feature description and the high effective classification. With these two elements, the distinguishing between the lipoylation samples and the nonlipoylation samples can be treated as a typical classification issue in the field of machine learning. In this work, we have proposed a method named LipoFNT, which employed the two featuring sets, including the Position-Specific Scoring Matrix and bi-profile Bayesian, as the classification features. And then, the flexible neural tree algorithm is utilized to deal with the imbalance classification issue in lipoylation modification sample dataset. The proposed method can achieve 81.07% in sn%, 80.29% in sp, 80.68% in Acc, 0.8076 in F1, and 0.6136 in MCC, respectively. Meanwhile, we have demonstrated the relationship between the lengths of peptide and identification of modification sites.

1. Introduction

Lysine lipoylation can be regarded as one of the most significant elements in the field of biology. Such a type of modifications has high conservation. Therefore, the lysine lipoylation is a special type of posttranslational modification in both prokaryotes' and eukaryotes' proteomics researches [1]. It was pointed out that lipoylation can be regarded as one special process, which is the covalent attachment of lipoic acid to 2-oxoacid dehydrogenase multienzyme complexes [2–5]. Such a type of modification is different from other PTM types, which depend on the local amino acid residues, in the level of protein sequence. Considering the high conservation of lipoylation modification, such a type of modification can hardly be influenced by the neighboring amino acid residues in the level of protein sequences [6]. It was known that the lysine lipoylation, which is one of the effective evolutionary processions, appears in various enzymes, including pyruvate dehydrogenase and other related enzymes, in many organisms, including bacteria and mammals [7–9]. Meanwhile, lipoylation plays the significant role in many key metabolic pathways and protein interactions [10]. With several years' efforts, some important researches have reported that the modification has some relationships with several human diseases. These diseases, including metabolic disorders, cancer, viral infection, and Alzheimer's disease [11–15], may cause some negative and harmful influences in the human being. Considering the above mentioned reasons, discovering the biological function of such a modification can be helpful and beneficial to understanding the causes of such mentioned serious diseases in some degree. Nevertheless, large numbers of lipoylation sites can hardly be effectively and accurately identified in this field. Without identification of such a modification sites, the molecular functions of lipoylation can hardly be discovered and researched. So, such an issue can be treated as one of the urgent topics in the related fields.

Lipoylation can be regarded as one of the rare but highly conserved lysine PTM types in the area of PTM researches. With the increasing development of lipoylation, some important issues have been reported. One of them is that there are merely four types of multimeric metabolic enzymes among the mammals. In these proteins, the majority of them are the core metabolic landscape. It was pointed that the dysregulation of such mitochondrial proteins may cause some human metabolic disorders in some degrees and even some diseases. Meanwhile, the most striking issue can be regarded as the lipoylation itself. Therefore, with further in-depth study of such high conserved lysine modification type, the addition or removing of such a modification is all evolutionarily conserved among the majority species in the level of protein. In short, such a modification can be treated as one of the most significant essential cofactors in the field of biology. So, we will demonstrate the biological functions and significances of such a modification. From these reasons, the significance of understanding the regulation of such a modification may be one of the necessary elements in the research of human diseases.

According to the function of lipoylation, lipoamide can be regarded as a cofactor central in the level of cellular metabolism [7, 16]. The lipoylation is presented as a conserved lysine PTM on essential multimeric metabolic complexes, and this function group needs some enzymatic activities among these protein complexes [17, 18]. For instance, both pyruvate dehydrogenase (PDH) and alpha-ketoglutarate (KDH) complexes own the ability to regulate distinct carbon entry points into the key metabolic pathway of TCA. For both the above mentioned complexes, lipoylation plays the critical role in proper enzyme functions. Meanwhile, removing such type of lysine modification may cause the inhabitation of their activities in some degree. It was reported that the evolutionary conservation of such type PTM of lipoylated enzymes can range from a variety of species and make some contributions in several core metabolic pathways in the level of organisms [8, 9]. Such theme of conservation can be treated as the lipoylated complexes [19, 20]. With the striking evolutionary conservation of such lysine rare modification, it was noted that these modified enzymes make great contributions to maintenance health and several serious diseases [12, 13, 21].

In order to better discover and know the molecular mechanisms of lipoylation, the main problem of identification of such a modification site can be treated as the classification issue, where positive samples and negative samples own different scales. There are some elements of this issue. Actually, some experimental methods and biological approaches have been proposed in this field. However, both the experimental and the biological ones can hardly meet the needs, which seem to be time-consuming and waste of resources in some degrees. Some PTM sites, including phosphorylation [22–24], S-nitrosylation [25–28], succinylation sites [29, 30], hydroxylation sites [31, 32], crotonylation [33, 34], sumoylation [35], glycosylation [36], ubiquitination [37], prenylation [38], carbonylation [39], and methylation [40–45], have successfully been classified with the methods in silico. From these successful instances, we can easily find out that several key elements of such a classification issue should be pointed out. These key elements include the feature evaluation, the model construction, the classification model selections, and the measurements of classification. On the other hand, the imbalance dataset, whose negative samples are far larger than the positive ones, should be considered.

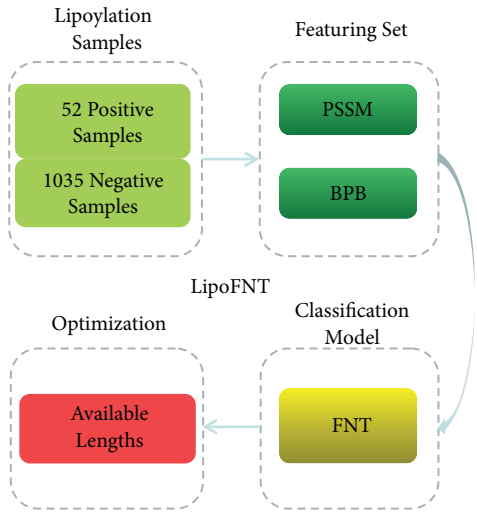

Figure 1: Outlines of LipoFNT.

In order to construct and design an accurate classification algorithm for identifying lipoylation sites in the protein level, as far as the researches covered, the foundational elements of the effective identification of modification sites are the available feature description and the high effective classification. With these two elements, distinguishing between the lipoylation samples and the nonlipoylation samples can be treated as a typical classification issue in the field of machine learning. In this work, we have employed the two featuring sets, including the Position-Specific Scoring Matrix (PSSM) and bi-profile Bayesian, as the classification features. And then, the flexible neural tree (FNT) algorithm is utilized to deal with the imbalance classification issue in lipoylation modification sample dataset. By combining other featuring sets and other machine learning models, we find out that the proposed method has better performances than other art-of-the-state methods in the field of PTM sites identification. What is more, we have demonstrated the relationship between the lengths of peptide and identification of modification sites. The steps can be shown in Figure 1. We will introduce such work in the following section step by step (http://121.250.173.184/).

2. Materials and Methods

2.1. Dataset. All employed protein sequences have been sourced from the UniProt database (http://www.uniprot.org/), which contains 576 lipoylated protein sequences. At the same time, the sequence high-similarity should be taken into account. Therefore, some necessary reduction redundancy should be proposed to deal with this problem.

These employed protein sequences, whose similarities are higher than 40%, should be removed with the tool of the CD-HIT program [58, 59]. With this procession, we achieve the nonredundant sample set, including 44 lipoylated proteins covering 52 lipoylation sites and 1035 nonmodification lysine sites. In order to reduce some unuseful protein segments, we utilized the sliding window to cover every lysine residue in the employed protein sequences. It was pointed that the scale of sliding window should be discussed in this work and we want to find the relationship between the scales of sliding windows and the classification performances. At the same time, some blank amino acid position may appear in the sliding windows. In order to deal with such phenomenon, the X amino acid stands for the blank amino acid position in the sample peptide segments.

2.2. Feature Construction. The first featuring set is the PSSM information of the identification of protein samples. With the development of the processing biological sequences in the field of bioinformatics, one of the most significant and challenging issues in this field is the method to express the biological sequences with different methods, including the discrete methods and the vector methods. However, these methods may keep some considerable sequence information and key pattern properties. It was pointed that the vector methods merely keep some foundational information and lose several sequence pattern in the level of protein. In order to avoid losing such information, the pseudo amino acid composition [60, 61] or PseAAC [62] was utilized in this work. Such a model has been widely utilized in the field of biological sequences, including protein level, DNA level and RNA level, and procession [63–66]. The "Pse-in-One" [67] and its updated version "Pse-in-One2.0" [68] can be treated as the most powerful tool in this area [68, 69].

The second one is the BPB feature set, which is a novel type of encode method [70]. When it comes to the BPB, such a feature depends on Bayesian's theories. So, a sample was given, which means peptide segments, that contains n length amino acid residues among it. The identified sample can be classified into two types, including the positive type and the negative one. Here, we define the positive type as the C_p and the negative type as the C_n. In detail, the C_p means the center lysine residue has the lipoylation modification in the identified peptide segment and the C_n stands for the fact that center lysine residue cannot be modified with the lipoylation in the classified peptide segment. With the rule of Bayesian's, assume the n amino acid residues are mutually independent; the posterior's probability of the peptide for the two types can be shown as

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

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

And then, we can redefine the above mentioned in

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

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

We assume the prior distribution can follow the uniform distribution. Therefore, the probability of negative samples and the probability of positive ones are equal. The decision function can be demonstrated in

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

According to the Shao's method, (5) can be redefined in

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

2.3. Flexible Neural Tree. Flexible neural tree, which can be regarded as one type of special alternative tree structural neural network, was proposed by Chen [71, 72]. The model owns the ability to construct the neural network with the tree structure. Such a type of neural network has been widely utilized in some classification issues in the field of machine learning. The main steps of such an algorithm can be demonstrated in the following section.

Initially, the utilizing instruction set for generating the foundational elements in the FNT model can be demonstrated in

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

$$Operation\_Set = \left\{+_1, +_2, \ldots, +_m\right\} \tag{8}$$

$$Variable\_Set = \left\{x_1, x_2, \ldots, x_n\right\} \tag{9}$$

where the instruction set contains two subsets, including the operation set and the variable one. The operation set $+_i$

includes several operation processions and the variable set x_i includes several values. At the same time, we can find out that the operation set mainly can be utilized in the nonleaf nodes and the variable set mainly can be utilized in the leaf nodes in the tree structure neural network. In other words, the variable set can be treated as the input of their neural node and the operation set can be regarded as the neural node in this model. And then, the employed flexible activation function is described in

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

Next, the output of each neural node can be calculated with the method of recursion. For each operation set element $+i$, the total excitation can be calculated in

$$network_i = \sum_{j=1}^{i} \omega_j \times y_j \tag{11}$$

where y_j ($j = 1\ 2\ \ldots\ i$) are the input to node $+i$. The output of the node $+i$ is then calculated in

$$out_i = f\left(m_i, n_i, network_i\right) = e^{-((network_i-m_i)/n_i)^2} \tag{12}$$

2.4. Performance Measurements. When it comes to the model performances, some well-known methods should be listed. In this work, some typical measurements, including sensitivity, specificity, accuracy, F1 scores, and Matthew's Correlation Coefficient (MCC) [73, 74], of the identification of modification sites issue should be listed. At the same time, the AUC [75] should also be employed to test the performance of imbalance classification problem and that is the negative samples size is much bigger than that of the positive ones.

In this classification problem, samples can be defined into two types, including the positive samples and the negative samples. According to the definition of the classified samples, they can cause the four results in the common situation. If the modification sample is classified as the modification one, this result can be named as *TP*, which stands for true positive. If the modification sample is classified as the nonmodification one, this result can be named as *FP*, which stands for false positive. With the concept, the nonmodificiation sample with classified modification one is the *TN* and the nonmodification sample with classified nonmodification is the *FN*. According to the number of *TP*, *TN*, *FP*, and *FN*, we can easily obtain these formulations, including sensitivity, specificity, accuracy, *F1* scores, and *MCC*. And the detailed information is shown in

$$Acc = \frac{TP + TN}{P + N} \tag{13}$$

$$Sn = \frac{TP}{TP + FN} \tag{14}$$

$$Sp = \frac{TN}{TN + FP} \tag{15}$$

$$F1 = \frac{2TP}{2TP + FN + FP} \tag{16}$$

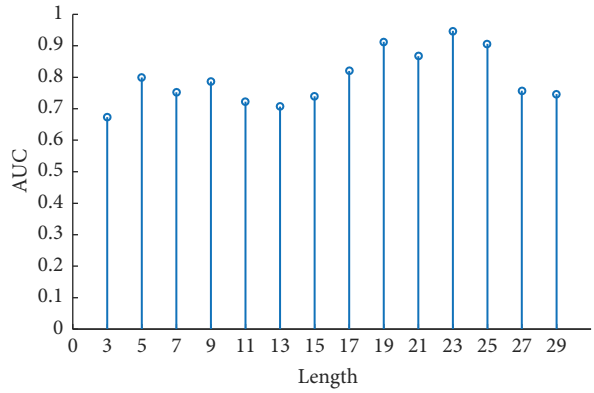

Figure 2: The ROC values of each length.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{17}$$

where P means the number of positive samples and N means the number of negative samples.

3. Results and Comparisons

3.1. Performance of LipoTree. In this section, we want to find out the available length of the sliding window in each sample. Meanwhile, the employ several lengths, which range from 3 to 29, whose center sites are lysine residues were pointed out. Therefore, the radius of each sample can be selected from 1 to 14. The ROC curves of each length can be demonstrated in Figure 2.

From Figure 2, we find out that the 14 employed lengths play different role in the classification of such medication type. At the same time, such classification issue can be treated as one of the typical imbalance classification issues in the field of machine learning. Considering such a situation, the ROC (receiver operating characteristic) curves can be known as one reasonable measurement to deal with such problem. It was pointed that while the length is equal to 23, the AUC value, which is the area under ROC curve, can reach the highest value. So, we can get the conclusion that such a length can be treated as the most available length among these employed lengths with the method of FNT and the feature of PSSM and BPB combination.

In order to demonstrate the performances of such algorithm, some typical feature descriptions have been employed to be compared with such an algorithm and several art-of-the-state methods have also been compared with such an algorithm in this field.

From Table 1, we can easily find that several typical feature description methods, including binary encoding, amino acid composition, grouping amino acid composition, physicochemical properties, KNN features, secondary tendency structure, Bi-gram [76], and Tri-gram [77], have been employed to be compared with proposed algorithm in this work. From Table 1, we can get the performances where the proposed method can achieve 81.07% in sn, 80.29%

Table 1: The Performances of Different Features.

Features	Sn(%)	Sp(%)	Acc(%)	F1	MCC
Binary Encoding	56.36	75.80	66.08	0.6243	0.3279
AA Composition	64.84	62.79	63.82	0.6418	0.2764
Grouping AA Composition	71.78	72.04	71.91	0.7187	0.4382
Physicochemical Properties	75.53	73.93	74.73	0.7493	0.4947
KNN Features	74.94	65.85	70.40	0.7168	0.4096
Secondary Tendency Structure	69.96	77.40	73.68	0.7266	0.4749
PSSM	71.20	79.39	75.30	0.7424	0.5076
BPB	72.81	78.51	75.66	0.7495	0.5140
Bi-gram	75.17	76.81	75.99	0.7579	0.5199
Tri-gram	77.28	78.27	77.78	0.7766	0.5555
Proposed Algorithm	81.07	80.29	80.68	0.8076	0.6136

Table 2: The Performances of Different Methods.

Method	Sn(%)	Sp(%)	Acc(%)	F1	MCC
DNABIND [46]	69.78	70.97	70.38	0.7020	0.4075
DNAbinder [46]	69.89	73.79	71.84	0.7128	0.4371
DBD-Threader [47]	57.79	94.71	76.25	0.7087	0.5649
DNA-Prot [47]	67.81	80.71	74.26	0.7249	0.4893
iDNA-Prot [48]	76.71	75.52	76.12	0.7626	0.5223
DBPPred [49]	79.37	74.82	77.10	0.7760	0.5425
PLMLA [50]	65.80	69.71	67.76	0.6711	0.3554
Phosida [51]	78.61	84.91	81.76	0.8117	0.6365
LysAcet [52]	77.50	75.14	76.32	0.7660	0.5265
EnsemblePail [53]	77.31	72.24	74.78	0.7540	0.4961
PSKAcePred [54]	71.20	69.87	70.54	0.7073	0.4107
BRABSB [55]	81.09	72.28	76.65	0.7762	0.5349
SSPKA [56]	75.81	79.57	77.69	0.7726	0.5542
SMOTE [57]	80.91	79.18	80.05	0.8022	0.6010
Proposed Algorithm	81.07	80.29	80.68	0.8076	0.6136

in sp, 80.68% in Acc, 0.8076 in F1, and 0.6136 in MCC, respectively. At the same time, we can get the conclusion that these typical and classical features play various roles in this classification issue. However, these features can hardly overcome the distance between the sensitivity and specialty in this classification issue.

From Table 2, we can get the information that several art-of-the-state methods, which include DNABIND, DNAbinder, DBD-Threader, DBPPred, and other approaches in this field, have been compared with the proposed algorithm. From the comparison, we can get the result that BRABSB can get the highest performance in sensitivity and the *Phosida* can play the most available results in specificity. It was pointed that the proposed algorithm can get the most ideal performances, while the length is equal to 23.

4. Conclusions and Discussions

In this study, a novel predictor named LipoFNT was developed to predict lysine lipoylation sites with the elements of bi-profile bayes feature encoding and flexible neural tree algorithm. As far as we are concerned, this is the first time flexible neural tree has been utilized in the classification of the lipoylation samples and nonlipoylation samples. Experimental results and performances showed that LipoFNT achieved an excellent performance and could be a useful bioinformatics algorithm for accurate identification of lipoylation sites.

From the above research, we can find out that there are 3 candidate lengths among all the employed lengths in this work. The top 3 lengths are 19, 23, and 25. In this section, we will discuss the performances of the top 3 lengths. And some art-of-the-state methods and features can be compared in this work. And the detailed information is shown in Tables S1–S6. It is shown that each sample can be calculated by the F-score with the BPB features [78, 79], which can be demonstrated in Table 3. With the candidate lengths, we can find that the most available length is 23. In this length, the proposed method can achieve well performances.

Meanwhile, some significant elements of lipoylated lysine site identification should be taken into account. First of all, the reasonable and effective features should be discovered and described in this classification issue. The features mainly have important influences on the sample valuation.

Table 3: The BPB features ranked by F-score method.

Order	Amino Acid	F-score	Order	Amino Acid	F-score
1	P8	3.8179	18	N4	0.9171
2	P10	3.5179	19	N11	0.9002
3	P9	3.4917	20	N-5	0.8227
4	P7	3.2817	21	N-7	0.8007
5	P14	3.0281	22	N13	0.7258
6	P12	2.8719	23	N-9	0.7091
7	P11	2.6971	24	N-3	0.6172
8	P2	2.2071	25	N-6	0.5817
9	P3	2.1718	26	N-9	0.5618
10	P5	1.9771	27	N2	0.4281
11	P6	1.7881	28	N-1	0.3171
12	P-10	1.6817	29	N-12	0.2812
13	P-4	1.5171	30	N-5	0.0017
14	P-9	1.2171	31	N1	0.0007
15	N5	1.1881	32	N-2	0.0002
16	N9	1.0117	33	P0	-1
17	N7	1.0021	34	N0	-1

The second step is the speed and accurate classification model. The classification model may have the ability to overcome some shortcomings and limitations on the features. In other words, the construction classification model may reduce some redundant and useless features and more effectively utilize some key features in the classification model. The last but not least step is the available sample length selection. The available length can reduce some useless features and low-useful neighbor amino acid residues influences.

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 that there are no conflicts of interest.

Authors' Contributions

Wenzheng Bao conceived the method. Rong Bao designed the method. Yuehui Chen conducted the experiments, Wenzheng Bao wrote the main manuscript text, and Bin Yang designed the website of this algorithm. All authors reviewed the manuscript.

Acknowledgments

This work was supported by the grants of the National Science Foundation of China, Nos. 61873270 and 61702445, and the grant from the Ph.D. Programs Foundation of Ministry of Education of China (No. 20120072110040).

Supplementary Materials

The supplementary material includes Tables S1 to S6 and Figures S1 to S14. *(Supplementary Materials)*

References

[1] M. G. Posner, A. Upadhyay, S. J. Crennell et al., "Post-translational modification in the archaea: structural characterization of multi-enzyme complex lipoylation," *Biochemical Journal*, vol. 449, no. 2, pp. 415–425, 2013.

[2] J. Collins, T. Zhang, S. W. Oh, R. Maloney, and J. Fu, "DNA-crowded enzyme complexes with enhanced activities and stabilities," *Chemical Communications*, vol. 53, no. 97, pp. 13059–13062, 2017.

[3] T. Tietjen and R. G. Wetzel, "Extracellular enzyme-clay mineral complexes: enzyme adsorption, alteration of enzyme activity, and protection from photodegradation," *Aquatic Ecology*, vol. 37, no. 4, pp. 331–339, 2003.

[4] T. E. McAllister, T.-L. Yeh, M. I. Abboud et al., "Non-competitive cyclic peptides for targeting enzyme–substrate complexes," *Chemical Science*, vol. 9, no. 20, pp. 4569–4578, 2018.

[5] L. J. Reed, "From lipoic acid to multi-enzyme complexes," *Protein Science: A Publication of The Protein Society*, vol. 7, no. 1, pp. 220–224, 1998.

[6] N. G. Wallis and R. N. Perham, "Structural dependence of post-translational modification and reductive acetylation of the lipoyl domain of the pyruvate dehydrogenase multienzyme complex," *Journal of Molecular Biology*, vol. 236, no. 1, pp. 209–216, 1994.

[7] L. J. Reed, "A trail of research from lipoic acid to α-keto acid dehydrogenase complexes," *The Journal of Biological Chemistry*, vol. 276, no. 42, pp. 38329–38336, 2001.

[8] J. E. Cronan, X. Zhao, and Y. Jiang, "Function, attachment and synthesis of lipoic acid in Escherichia coli," *Advances in Microbial Physiology*, vol. 50, pp. 103–146, 2005.

[9] M. D. Spalding and S. T. Prigge, "Lipoic acid metabolism in microbial pathogens," *Microbiology and Molecular Biology Reviews*, vol. 74, no. 2, pp. 200–228, 2010.

[10] E. A. Rowland, C. K. Snowden, and I. M. Cristea, "Protein lipoylation: an evolutionarily conserved metabolic regulator of health and disease," *Current Opinion in Chemical Biology*, vol. 42, pp. 76–85, 2018.

[11] C. B. Pocernich and D. A. Butterfield, "Acrolein inhibits NADH-linked mitochondrial enzyme activity: implications for Alzheimer's disease," *Neurotoxicity Research*, vol. 5, no. 7, pp. 515–519, 2003.

[12] J. Munger, S. U. Bajad, H. A. Coller, T. Shenk, and J. D. Rabinowitz, "Dynamics of the cellular metabolome during human cytomegalovirus infection," *PLoS Pathogens*, vol. 2, no. 12, pp. 1165–1175, 2006.

[13] M. C. Sugden, "PDC deletion: the way to a man's heart disease," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 295, no. 3, pp. H917–H919, 2008.

[14] G. Huang, F. Cui, F. Yu et al., "Sirtuin-4 (SIRT4) is downregulated and associated with some clinicopathological features in gastric adenocarcinoma," *Biomedicine & Pharmacotherapy*, vol. 72, pp. 135–139, 2015.

[15] M. Miyo, H. Yamamoto, M. Konno et al., "Tumour-suppressive function of SIRT4 in human colorectal cancer," *British Journal of Cancer*, vol. 113, no. 3, pp. 492–499, 2015.

[16] C. S. Tsai, M. W. Burgett, and L. J. Reed, "α-keto acid dehydrogenase complexes XX. a kinetic study of the pyruvate dehydrogenase complex from bovine kidney," *The Journal of Biological Chemistry*, vol. 248, no. 24, pp. 8348–8352, 1973.

[17] Z. H. Zhou, D. B. McCarthy, C. M. O'Connor, L. J. Reed, and J. K. Stoops, "The remarkable structural and functional organization of the eukaryotic pyruvate dehydrogenase complexes," *Proceedings of the National Acadamy of Sciences of the United States of America*, vol. 98, no. 26, pp. 14802–14807, 2001.

[18] R. N. Perham, "Swinging arms and swinging domains in multifunctional enzymes: Catalytic machines for multistep reactions," *Annual Review of Biochemistry*, vol. 69, pp. 961–1004, 2000.

[19] R. A. Mathias, T. M. Greco, A. Oberstein et al., "Sirtuin 4 is a lipoamidase regulating pyruvate dehydrogenase complex activity," *Cell*, vol. 159, no. 7, pp. 1615–1625, 2014.

[20] E. A. Rowland, T. M. Greco, C. K. Snowden, A. L. McCabe, T. J. Silhavy, and I. M. Cristea, "Sirtuin lipoamidase activity is conserved in bacteria as a regulator of metabolic enzyme complexes," *mBio*, vol. 8, no. 5, 2017.

[21] M. A. Ansari, H. M. Abdul, G. Joshi, W. O. Opii, and D. A. Butterfield, "Protective effect of quercetin in primary neurons against Aβ(1–42): relevance to Alzheimer's disease," *The Journal of Nutritional Biochemistry*, vol. 20, no. 4, pp. 269–275, 2009.

[22] 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.

[23] 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 Informatics*, vol. 36, 2017.

[24] 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," *Analytical Biochemistry*, vol. 550, pp. 109–116, 2018.

[25] Y. Xu, J. Ding, L. Y. Wu, and K. C. Chou, "iSNO-PseAAC: predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition," *PLoS ONE*, vol. 8, no. 2, Article ID e55844, 2013.

[26] Y. Xu, X.-J. Shao, L.-Y. Wu, N.-Y. Deng, and K.-C. Chou, "ISNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins," *PeerJ*, vol. 1, article e171, 2013.

[27] C. Jia, X. Lin, and Z. Wang, "Prediction of protein S-nitrosylation sites based on adapted normal distribution bi-profile Bayes and Chou's pseudo amino acid composition," *International Journal of Molecular Sciences*, vol. 15, no. 6, pp. 10410–10423, 2014.

[28] J. Zhang, X. Zhao, P. Sun, and Z. Ma, "PSNO: predicting cysteine S-nitrosylation sites by incorporating various sequence-derived features into the general form of Chou's PseAAC," *International Journal of Molecular Sciences*, vol. 15, no. 7, pp. 11204–11219, 2014.

[29] 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.

[30] J. Jia, Z. Liu, X. Xiao, B. Liu, and K.-C. Chou, "pSuc-Lys: predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach," *Journal of Theoretical Biology*, vol. 394, pp. 223–230, 2016.

[31] Y. Xu, X. Wen, X.-J. Shao, N.-Y. Deng, and K.-C. Chou, "iHyd-PseAAC: predicting hydroxyproline and hydroxylysine in proteins by incorporating dipeptide position-specific propensity into pseudo amino acid composition," *International Journal of Molecular Sciences*, vol. 15, no. 5, pp. 7594–7610, 2014.

[32] W.-R. Qiu, B.-Q. Sun, X. Xiao, Z.-C. Xu, and K.-C. Chou, "IHyd-PseCp: identify hydroxyproline and hydroxylysine in proteins by incorporating sequence-coupled effects into general PseAAC," *Oncotarget* , vol. 7, no. 28, pp. 44310–44321, 2016.

[33] 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, 2018.

[34] Z. Ju and J.-J. He, "Prediction of lysine crotonylation sites by incorporating the composition of k -spaced amino acid pairs into Chou's general PseAAC," *Journal of Molecular Graphics and Modelling*, vol. 77, pp. 200–204, 2017.

[35] 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, 2016.

[36] H.-L. Xie, L. Fu, and X.-D. Nie, "Using ensemble SVM to identify human GPCRs N-linked glycosylation sites based on the general form of Chou's PseAAC," *Protein Engineering, Design and Selection*, vol. 26, no. 11, pp. 735–742, 2013.

[37] W. R. Qiu, X. Xiao, W. Z. Lin, and K. C. Chou, "iUbiq-Lys: prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a gray system model," *Journal of Biomolecular Structure Dynamics*, vol. 33, no. 12, 2015.

[38] 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 Chemistry*, vol. 13, no. 6, pp. 544–551, 2017.

[39] 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, 2016.

[40] W.-R. Qiu, X. Xiao, W.-Z. Lin, and K.-C. Chou, "iMethyl-PseAAC: identification of protein methylation sites via a pseudo amino acid composition approach," *Biomed Research International*, vol. 2014, Article ID 947416, 12 pages, 2014.

[41] W. Chen, P. Feng, H. Ding, H. Lin, and K.-C. Chou, "iRNA-Methyl: identifying N(6)-methyladenosine sites using pseudo nucleotide composition," *Analytical Biochemistry*, vol. 490, pp. 26–33, 2015.

[42] W. Chen, H. Tang, J. Ye, H. Lin, and K.-C. Chou, "iRNA-PseU: Identifying RNA pseudouridine sites," *Molecular Therapy - Nucleic Acids*, vol. 5, p. e332, 2016.

[43] 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," *Analytical Biochemistry*, vol. 497, pp. 60–67, 2016.

[44] W.-R. Qiu, S.-Y. Jiang, B.-Q. Sun, X. Xiao, X. Cheng, and K.-C. Chou, "iRNA-2methyl: identify RNA 2'-O-methylation sites by incorporating sequence-coupled effects into general PseKNC and ensemble classifier," *Medicinal Chemistry*, vol. 13, no. 8, pp. 734–743, 2017.

[45] 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, 2017.

[46] 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.

[47] 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.

[48] 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.

[49] 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, article 298, 2014.

[50] 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.

[51] F. Gnad, S. Ren, C. Choudhary, J. Cox, and M. Matthias, "Predicting post-translational lysine acetylation using support vector machines," *Bioinformatics*, vol. 26, no. 13, pp. 1666–1668, 2010.

[52] 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.

[53] 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.

[54] S. Suo, J. Qiu, S. Shi et al., "Position-specific analysis and prediction for protein lysine acetylation based on multiple features," *PLoS ONE*, vol. 7, no. 11, Article ID e49108, 2012.

[55] 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.

[56] 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, no. 1, Article ID 5765, 2015.

[57] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[58] W. Li and A. Godzik, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics*, vol. 22, no. 13, pp. 1658-1659, 2006.

[59] Y. Huang, B. Niu, Y. Gao, L. Fu, and W. Li, "CD-HIT Suite: a web server for clustering and comparing biological sequences," *Bioinformatics*, vol. 26, no. 5, Article ID btq003, pp. 680–682, 2010.

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

[61] K. Chou, "Prediction of protein cellular attributes using pseudo-amino acid composition," *Proteins: Structure, Function, and Bioinformatics*, vol. 43, no. 3, pp. 246–255, 2010.

[62] K.-C. Chou, "Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes," *Bioinformatics*, vol. 21, no. 1, pp. 10–19, 2005.

[63] S. Muthu Krishnan, "Using Chou's general PseAAC to analyze the evolutionary relationship of receptor associated proteins (RAP) with various folding patterns of protein domains," *Journal of Theoretical Biology*, vol. 445, pp. 62–74, 2018.

[64] Y. Liang and S. Zhang, "Identify Gram-negative bacterial secreted protein types by incorporating different modes of PSSM into Chou's general PseAAC via Kullback-Leibler divergence," *Journal of Theoretical Biology*, vol. 454, pp. 22–29, 2018.

[65] J. Mei and J. Zhao, "Analysis and prediction of presynaptic and postsynaptic neurotoxins by Chou's general pseudo amino acid composition and motif features," *Journal of Theoretical Biology*, vol. 447, pp. 147–153, 2018.

[66] M. S. Rahman, S. Shatabda, S. Saha, M. Kaykobad, and M. S. Rahman, "DPP-PseAAC: a DNA-binding protein prediction model using Chou's general PseAAC," *Journal of Theoretical Biology*, vol. 452, pp. 22–34, 2018.

[67] 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 Research*, vol. 43, no. 1, pp. W65–W71, 2015.

[68] B. Liu, H. Wu, and K. Chou, "Pse-in-One 2.0: an improved package of web servers for generating various modes of pseudo components of DNA, RNA, and protein sequences," *Natural Science*, vol. 9, pp. 67–91, 2017.

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

[70] J. Shao, D. Xu, S.-N. Tsai, Y. Wang, and S.-M. Ngai, "Computational identification of protein methylation sites through Bi-profile Bayes feature extraction," *PLoS ONE*, vol. 4, no. 3, 2009.

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

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

[73] W. Bao, Z. Huang, C.-A. Yuan, and D.-S. Huang, "Pupylation sites prediction with ensemble classification model," *International Journal of Data Mining and Bioinformatics*, vol. 18, no. 2, pp. 91–104, 2017.

[74] W. Bao, Z.-H. You, and D.-S. Huang, "CIPPN: computational identification of protein pupylation sites by using neural network," *Oncotarget*, vol. 8, no. 65, pp. 108867–108879, 2017.

[75] W. Bao, B. Yang, Z. Li, and Y. Zhou, "LAIPT: lysine acetylation site identification with polynomial tree," *International Journal of Molecular Sciences*, vol. 20, article 113, 2018.

[76] A. Sharma, J. Lyons, A. Dehzangi, and K. K. Paliwal, "A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition," *Journal of Theoretical Biology*, vol. 320, pp. 41–46, 2013.

[77] K. K. Paliwal, A. Sharma, J. Lyons, and A. Dehzangi, "A tri-gram based feature extraction technique using linear probabilities of position specific scoring matrix for protein fold recognition," *IEEE Transactions on NanoBioscience*, vol. 13, no. 1, pp. 44–50, 2014.

[78] Z. Ju and S.-Y. Wang, "Predicting lysine lipoylation sites using bi-profile bayes feature extraction and fuzzy support vector machine algorithm," *Analytical Biochemistry*, vol. 561-562, pp. 11–17, 2018.

[79] Z. Ju and J.-J. He, "Prediction of lysine propionylation sites using biased SVM and incorporating four different sequence features into Chou's PseAAC," *Journal of Molecular Graphics and Modelling*, vol. 76, pp. 356–363, 2017.