RESEARCH ARTICLE

Identify Lysine Neddylation Sites Using Bi-profile Bayes Feature Extraction *via* the Chou's 5-steps Rule and General Pseudo Components

Zhe Ju[1,*] and Shi-Yun Wang[1]

[1] *College of Science, Shenyang Aerospace University, Shenyang 110136, P.R. China*

ARTICLE HISTORY

Received: July 01, 2019
Revised: October 19, 2019
Accepted: November 07, 2019

DOI:
10.2174/1389202921666191223154629

Abstract: ***Introduction***: Neddylation is a highly dynamic and reversible post-translational modification. The abnormality of neddylation has previously been shown to be closely related to some human diseases. The detection of neddylation sites is essential for elucidating the regulation mechanisms of protein neddylation.

Objective: As the detection of the lysine neddylation sites by the traditional experimental method is often expensive and time-consuming, it is imperative to design computational methods to identify neddylation sites.

Methods: In this study, a bioinformatics tool named NeddPred is developed to identify underlying protein neddylation sites. A bi-profile bayes feature extraction is used to encode neddylation sites and a fuzzy support vector machine model is utilized to overcome the problem of noise and class imbalance in the prediction.

Results: Matthew's correlation coefficient of NeddPred achieved 0.7082 and an area under the receiver operating characteristic curve of 0.9769. Independent tests show that NeddPred significantly outperforms existing lysine neddylation sites predictor NeddyPreddy.

Conclusion: Therefore, NeddPred can be a complement to the existing tools for the prediction of neddylation sites. A user-friendly webserver for NeddPred is accessible at 123.206.31.171/NeddPred/.

Keywords: Post-translational modification, neddylation, feature extraction, fuzzy support vector machine, chou's 5-steps rule, pseudo components.

1. INTRODUCTION

NEDD8 is an 81 amino acid polypeptide, which is 60% identical and 80% similar to ubiquitin. The process of ubiquitin-like protein NEDD8 attaching substrate lysine *via* isopeptide bonds is known as neddylation [1]. Neddylation is a highly dynamic and reversible protein post-translational modification (PTM), which occurs similarly to ubiquitination and needs enzyme cascades involving E1, E2 and E3 [2]. Although neddylation relies on its own E1 and E2 enzymes, no NEDD8-specific E3 has yet been identified and it is possible that the neddylation system relies on E3 ligases with dual specificity [3]. Neddylation has been demonstrated to be essential to maintain the ubiquitin ligase activity of Cullin-Roc based E3 ligases, and affects cell-cycle regulation, transcriptional regulation and signal transduction indirectly [4]. Previous studies have shown that abnormal neddylation is strongly linked to some human diseases, such as cancer, Parkinson's disease and Alzheimer's disease [5-7]. Therefore, exploring the biological functions of neddylation will help to reveal the pathogenesis of the abovementioned diseases. However, compared with the ubiquitination that has been widely studied in the past two decades, the molecular mechanism and physiological functions of neddylation still not well characterized.

Accurate detection of neddylation sites is the biggest challenge to decipher the molecular mechanisms of neddylation. However, the experimental approaches are often time-consuming and expensive, it is crucial to develop computational methods to identify neddylation sites. The computational identification and analysis of PTM sites are gaining more attention in recent years [8-12]. Yavuz *et al.* [13] developed a predictor named NeddyPreddy to predict neddylation sites using a support vector machine based on various sequence properties, position-specific scoring matrices, and disorder. However, the prediction sensitivity of NeddyPreddy (75%) is not satisfactory.

In order to develop an accurate predictor for the identification of neddylation sites, the key is to seek an efficient feature extraction method to encode neddylation sites. Based on many aspects of assessments, we found bi-profile bayes (BPB) was more suitable for distinguishing between the neddylation sites and non-neddylation sites than split amino acid composition (SplitAAC), amino acid factors (AAF) amino acid composition (AAC) and binary encoding (BE) which are the widely used feature extraction techniques in PTM sites prediction. Therefore, the BPB was used to encode neddylation sites. Moreover, a fuzzy SVM algorithm is used to handle the class imbalance and noise problem in the neddylation sites training dataset. A novel predictor named NeddPred was constructed by combining the BPB with the fuzzy SVM. Feature analysis indicated that the residues in some positions around neddylation sites play a key role in predicting neddylation sites.

*Address correspondence to this author at the College of Science, Shenyang Aerospace University, Shenyang 110136, P.R. China;
Tel: +86 024 89723442; E-mail: juzhe1120@hotmail.com

2. MATERIALS AND METHODS

As demonstrated by a series of recent publications [14-25] and summarized in a comprehensive review [26], to develop a really useful predictor for PTMs site, one needs to follow Chou's 5-step rule: (a) collect valid PTMs sites to train the predictor; (b) encode the PTMs sites by effective feature extraction that can reflect their sequential pattern; (c) develop a robust algorithm to conduct the prediction; (d) properly perform cross-validation tests to objectively evaluate the effectiveness of the predictor; (e) establish a user-friendly and publicly accessible web-server for the predictor. Below, let us elaborate on how to deal with these five steps.

2.1. Dataset

Yavuz's training dataset, validation dataset and independent test dataset [13] were used to train and assess NeddPred. The training dataset consisted of 34 experimentally verified neddyllysine sites and 687 non-neddyllysine sites; the validation dataset consisted of 6 neddyllysine sites and 115 non-neddyllysine sites; and the independent test dataset consisted of 11 neddyllysine sites and 229 non-neddyllysine sites. According to Yavuz's work and our trials (section 3.1), the window size was selected as 21. The neddylated peptides were used as positive samples, while the non-neddylated peptides were used as negative samples. The training dataset and the independent test dataset are provided in Supplementary material **S1**.

2.2. Feature Extraction

It is well-known that how to express a biological sequence with a discrete model or a vector is one of the most difficult problems in computational biology. This is because the machine learning algorithms (such as "Optimization" algorithm [27], "Covariance Discriminant" algorithm [28, 29], "Nearest Neighbor" algorithm [30], and "Support Vector Machine" algorithm [31] can only handle vectors [32]. To avoid completely losing the sequence-pattern information for proteins, the pseudo amino acid composition (PseAAC) [26, 33] was proposed. The PseAAC has been widely used in the areas of bioinformatics [34-44]. As PseAAC has been widely and increasingly used, four powerful open-access software, called 'PseAAC' [45], 'PseAAC-Builder' [46], 'Propy' [47], and 'PseAAC-General' [48], were established to generate pseudo amino acid composition features. The former three are for generating various modes of Chou's special PseAAC [49]; while PseAAC-General is for those of Chou's general PseAAC such as "Functional Domain" mode, "Gene Ontology" mode, and "Sequential Evolution" or "PSSM" mode [26]. Encouraged by the successes of using PseAAC to deal with protein/peptide sequences, the concept of PseKNC (Pseudo K-tuple Nucleotide Composition) [50] was developed for generating various feature vectors for DNA/RNA sequences [51-53]. Particularly, recently a very powerful web-server called 'Pse-in-One' [54] and its updated version 'Pse-in-One2.0' [55] have been established that can be used to generate any desired feature vectors for protein/peptide and DNA/RNA sequences according to the need of users' studies.

Bi-profile bayes (BPB) is an effective feature coding method which can be covered by the general PseAAC. The BPB coding has been applied to many bioinformatics problems [56-60]. Here, BPB was used to encode neddylation sites. Given a sequence fragment $S = s_1 s_2 ... s_n$, where $s_j \ (j = 1,2,...,n)$ stands for one amino acid and *n* denotes the length of the sequence fragment. *S* belongs to category C_1 or C_{-1}, where C_1 and C_{-1} represent neddylation sites and non-neddylation sites, respectively. Based on Bayes' rule, assume that s_j (j=1,2,…,n) are mutually independent, the posterior probability of S for C_1 and C_{-1} can be given by:

$$P(c_1 | S) = P(S | c_1) P(c_1) / P(S) = \prod_{j=1}^{n} P(s_j | c_1) P(c_1) / P(S) \quad (1)$$

$$P(c_{-1} | S) = P(S | c_{-1}) P(c_{-1}) / P(S) = \prod_{j=1}^{n} P(s_j | c_{-1}) P(c_{-1}) / P(S) \quad (2)$$

Therefore, (1) and (2) can be rewritten as:

$$\log(P(c_1 | S)) = \sum_{j=1}^{n} \log(P(s_j | c_1)) - \log(P(S)) + \log(P(c_1)) \quad (3)$$

$$\log(P(c_{-1} | S)) = \sum_{j=1}^{n} \log(P(s_j | c_{-1})) - \log(P(S)) + \log(P(c_{-1})) \quad (4)$$

Assume that $P(c_1) = P(c_{-1})$, the decision function can be written as formula (5):

$$\begin{aligned} f(S) &= \text{sgn}(\log(P(c_1 | S)) - \log(P(c_{-1} | S))) \\ &= \text{sgn}(\sum_{j=1}^{n} \log(P(s_j | c_1)) - \sum_{j=1}^{n} \log(P(s_j | c_{-1}))) \end{aligned} \quad (5)$$

Based on the results of the literature [56], formula (5) can be rewritten as:

$$f(S) = \text{sgn}(w \bullet p) \quad (6)$$

where $\text{sgn}(x) = \begin{cases} 1, & \text{if } x \geq 0 \\ -1, & \text{if } x < 0 \end{cases}$; $w = (w_1, w_2, ..., w_n, w_{n+1}, ..., w_{2n})$ is weigh vector; $p = (p_1, p_2, ..., p_n, p_{n+1}, ..., p_{2n})$ is the posterior probability vector; $p_1, p_2, ..., p_n$ ($p_{n+1}, p_{n+2}, ..., p_{2n}$) represent the posterior probability of each amino acid at each position in the category *C*1 (*C*2). In this study, the posterior probability *p* was given by the frequency of each amino acid in the training peptides. Therefore, every training peptide was encoded by BPB encoding as 42-dimensional vectors.

2.3. Fuzzy Support Vector Machine

As one of the effective machine learning algorithms, SVM has been used in the detection of protein PTMs sites, such as succinylation sites [61], glycation sites [62], crotonylation sites [63], propionylation sites [38] and citrullination sites [9]. In the standard SVM model, each training sample was assigned to the same weight. However, there may be some noisy samples in the training dataset. Therefore, it is more reasonable to assign different weight values to different samples based on their importance and imbalance than to assign the same weight value. Here, the fuzzy SVM model was used to construct the classifier.

To facilitate the description, the training set is denoted as $\{(x_i, t_i), i = 1, 2, ..., l\}$. Assume that the first p training examples are positive (*i.e.,* $t_i = 1, i = 1, 2, ..., p$), and the rest $l - p$ training examples are negative (*i.e.,* $t_i = -1, i = p+1, p+2, ..., l$). The fuzzy SVM can be formulated as follows:

$$\min_{\omega,\xi} \frac{1}{2}\|\omega\|^2 + C^+ \sum_{i=1}^{p} s_i^+ \xi_i + C^- \sum_{i=p+1}^{l} s_i^- \xi_i$$

$$\text{s.t. } t_i(\omega \cdot \Phi(x_i) + b) \geq 1 - \xi_i \tag{7}$$

$$\xi_i \geq 0, \quad i = 1, 2, ..., l$$

where C^+ and C^- are the penalty factors; $\xi_i\ (i = 1, 2, ..., l)$ are slack variables, and $\Phi(x)$ is the non-linear mapping; s_i^+ and s_i^- are the fuzzy memberships.

As described in the literature [64], in this study, the fuzzy memberships are defined as follows:

$$s_i^+ = 1 - \frac{d_i^{cen+}}{\max\limits_j(d_j^{cen+}) + \delta},\ i = 1, 2, ..., p \tag{8}$$

$$s_i^- = 1 - \frac{d_i^{cen-}}{\max\limits_j(d_j^{cen-}) + \delta},\ i = p+1, \text{p}+2, ..., l \tag{9}$$

where $d_i^{cen+} = \| x_i - \frac{1}{p}\sum_{j=1}^{p} x_j \|, d_i^{cen-} = \| x_i - \frac{1}{l-p}\sum_{j=p+1}^{l} x_j \|$; δ is a very small positive value guaranteeing the value of fuzzy membership always higher than zero. The observation was treated more important and assigned higher fuzzy membership values when they were closer to their class center; whereas the observation was treated as less important (such as noises or outliers) and assigned lower fuzzy membership values when they were farther away from their class center.

Based on the results reported by Batuwita and Palade [65], to handle the problem of class imbalance in the prediction, the penalty factors C^+ and C^-were set to $\frac{(l-p)C}{p}$ and C, respectively. The Gaussian kernel function $K(x_i, x_j) = \Phi(x_i)^{\mathrm{T}}\Phi(x_j) = \exp(-\gamma \| x_i - x_j \|^2)$ is used in the fuzzy SVM [38, 63] The kernel parameter γ was selected from {2^{-10}, 2^{-9}, …, 2^0}; penalty parameters *C* was selected from {2^0, 2^1, …, 2^{12}}; the Libsvm-weights-3.20 package [66] was used to implement the fuzzy SVM models.

2.4. Cross-validation and Performance Assessment

K-fold cross-validation test, jackknife test and independent dataset test are often adopted to evaluate the performance of a predictor. As the jackknife test can always yield a unique result for a given training dataset, it is the most objective and least arbitrary among the above three test methods [26]. However, to reduce computational time, the 10-fold cross-validation was adopted to evaluate our model. Here, the 10-fold cross-validation is repeated 10 times.

Five widely-accepted measurements, including sensitivity (Sn), specificity (Sp), accuracy (ACC) and Matthew's correlation coefficient (MCC) and area under the receiver operating characteristic curve (AUC), were used to evaluate prediction performances of NeddPred. In accordance with Eq. (14) [18], they are defined as:

$$\begin{cases} \text{Sn} = 1 - N_-^+ / N^+ \\ \text{Sp} = 1 - N_+^- / N^- \\ \text{ACC} = 1 - (N_-^+ + N_+^-) / (N^+ + N^-) \\ \text{MCC} = (1 - (N_-^+ / N^+ + N_+^- / N^-)) / \sqrt{(1 + (N_+^- - N_-^+)/N^+)(1 + (N_-^+ - N_+^-)/N^-)} \end{cases} \tag{10}$$

Where N^+ is the total number of the neddylation sites investigated, while N_-^+ is the number of the sites incorrectly predicted as the non-neddylation sites, and N^- is the total number of the non-neddylation sites investigated, while N_+^- is the number of the non-neddylation sites incorrectly predicted as the neddylation sites. The AUC can measure the overall performance of a given prediction system. The closer the AUC is to 1, the better the prediction system is.

Either the set of conventional metrics copied from math books or the intuitive metrics derived from the Chou's symbols [67-69] are valid only for the single-label systems. For the multi-label systems, whose existence has become more frequent in system biology [70-75] system medicine [76, 77], and biomedicine [17], a completely different set of metrics, as previously defined [78] is needed.

3. RESULTS AND DISCUSSION

3.1. Performance of NeddPred

The optimal parameters (window size, penalty factor *C* and kernel parameter γ) of the proposed model were determined by the highest AUC value in 10-fold cross-validation performances. The proposed model achieved the highest AUC value of 0.9769 when using the window size 21, $C = 2^5$ and $\gamma = 0.5$. Therefore, the optimal window size was selected as 21. As shown in Table **1**, the predicted Sn, Sp, ACC and MCC values were 79.41%, 97.96%, 97.09% and 0.7082, respectively. Moreover, NeddPred was also implemented by the jackknife test with the optimal parameter obtained in the 10-fold cross-validation. NeddPred also achieved a satisfactory performance with a Se of 79.41%, an Sp of 97.09%, an ACC of 96.26%, an MCC of 0.6569 and an AUC of 0.9789.

To assess the performance of the fuzzy SVM, it was compared with the standard SVM and biased SVM [65]. The comparison results of the above SVM algorithms were shown in Table **2**. The fuzzy SVM reached the highest Sn, ACC and MCC values of 79.41%, 97.09% and 0.7082, respectively. Although the Sp value of the standard SVM (99.56%) was slightly higher than that of the fuzzy SVM (97.96%), the Sn value of the standard SVM (44.12%) was much lower than that of the standard SVM (79.41%). In short, the fuzzy SVM showed better results as compared with standard SVM and biased SVM.

Table 1. The 10-fold cross-validation results of NeddPred with different window sizes.

Window Size	Sn(%)	Sp(%)	ACC(%)	MCC	AUC
11	76.47	93.01	92.23	0.4853	0.9331
13	73.53	91.27	90.43	0.4259	0.9096
15	67.65	93.89	92.65	0.4554	0.9329
17	76.47	97.96	96.95	0.6893	0.9592
19	73.53	96.80	95.70	0.6039	0.9592
21	79.41	**97.96**	**97.09**	**0.7082**	**0.9769**
23	76.47	97.82	96.81	0.6800	0.9756
25	79.41	97.09	96.26	0.6569	0.9723
27	**82.35**	95.78	95.15	0.6138	0.9721

Table 2. Comparison of fuzzy SVM with standard SVM and biased SVM.

Method	Sn	Sp	ACC	MCC	AUC
Standard SVM	44.12	99.56	96.95	0.5936	0.9747
Biased SVM	79.41	97.38	96.53	0.6729	0.9716
Fuzzy SVM	79.41	97.96	97.09	0.7082	**0.9769**

3.2. Comparison of BPB with Other Feature Extraction Technologies

To demonstrate the effectiveness of BPB, it was compared with the most widely used feature extraction technologies in computational biology, including amino acid composition (AAC) [79], split amino acid composition (SplitAAC) [80], amino acid factors (AAF) [81], binary encoding (BE) [82] and composition of *k*-space amino acid pairs (CKSAAP) [83]. For comparison, CKSAAP with *k*=0, 1, 2, 3 and 4 was performed, and the peptide in SplitAAC was divided into three parts: 7 amino acids of N termini, 7 amino acids of C termini, and the region between these two termini. The performance of 10-fold cross-validation with various features was shown in Table **3**. The model with BPB reached the highest value of AUC. The results indicated that BPB encoding is more effective for extracting the sequence information around the neddylation sites than other encoding schemes.

3.3. Comparison of NeddPred with Existing Predictor

At present, only one predictor named NeddyPreddy [13] was proposed for the prediction of neddylation sites. It is considered that NeddPred and NeddyPreddy were both trained on Yavuz's dataset [13] which contained 34 neddylation sites and 687 non-neddylation sites. It is interesting to compare NeddPred with NeddyPreddy. As shown in Table **4**, NeddPred outperforms NeddyPreddy significantly, whether on the training dataset, validation set and independent test set. For example, NeddPred revealed about 26% higher MCC than NeddyPreddy. These results showed that NeddPred can predict more reliable neddylation sites from protein sequences than NeddyPreddy. The ROC curves for NeddPred by 10-fold cross-validation, jackknife test, validation set test and independent test are shown in Fig. (**1**). The results indicated that NeddPred can be an effective predictor for the prediction of neddylation sites. There are two factors for the improvement of NeddPred. One is the fuzzy SVM that can effectively handle the problem of the noise in the prediction of neddylation sites. Another factor is that the BPB feature used in NeddPred outperforms sequence properties, position-specific scoring matrices, and disorder used in NeddyPreddy.

3.4. Prediction Server of NeddPred

As pointed out previously [84], user-friendly and publicly accessible web-servers are the future direction for developing useful bioinformatics tools [85-89]. To provide convenience for the experimental scientists, NeddPred has been implemented as a web-server which was trained on all available data (training data, validation data and independent testing data, *i.e.*, 34+6+11=51 neddylation sites and 687+115+229=1031 non-neddylation sites) using the optimal parameters (window size 21, $C=2^5$ and $\gamma=0.5$). The web-server for NeddPred is now available at http://123.206.31.171/NeddPred/. As shown in Fig. (**2**), users can enter query protein sequences (FASTA) or batch-upload the query protein sequences (FASTA) as a file for the prediction. The CKSAAP_NeddSite server will output a CSV-formatted file with prediction results.

Table 3. The predictive performance of 10-fold cross-validation using various training features.

Feature	Sn(%)	Sp(%)	ACC(%)	MCC(%)	AUC(%)
AAC	77.73	38.24	75.87	0.0804	0.6319
SplitAAC	44.12	88.21	86.13	0.2017	0.7096
AAF	23.53	99.71	96.12	0.4212	0.6649
BE	26.47	99.27	95.84	0.3955	0.6912
CKSAAP	32.35	96.94	93.90	0.3015	0.7717
BPB	79.41	97.96	97.09	0.7082	**0.9769**

Table 4. Comparison of NeddPred with NeddyPreddy under different evaluation strategies.

Method	Evaluation Strategie	Sn	Sp	ACC	MCC	AUC
NeddyPreddy[1]	10-fold cross-validation	0.76	0.91	0.91	0.45	0.95
NeddPred		0.7941	0.9796	0.9709	0.7082	0.9769
NeddyPreddy[1]	Validation set	0.67	0.91	0.90	0.39	0.83
NeddPred		1.00	0.9913	0.9917	0.9218	1.00
NeddyPreddy[1]	Independent testing set	0.64	0.91	0.90	0.35	0.80
NeddPred		1.00	0.9520	0.9542	0.6899	1.00

[1] The corresponding results were obtained from the literature (Yavuz *et al.*, 2015).

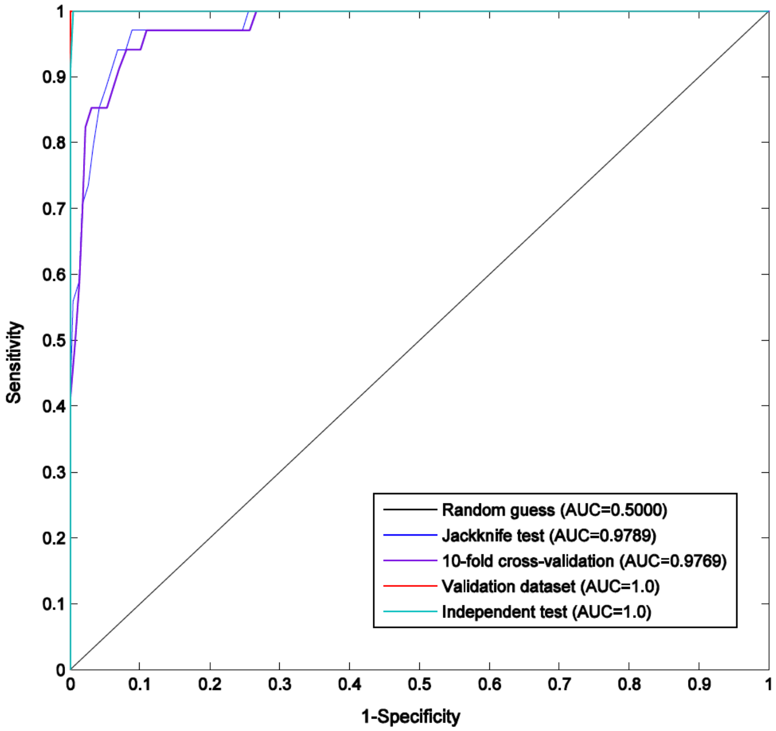
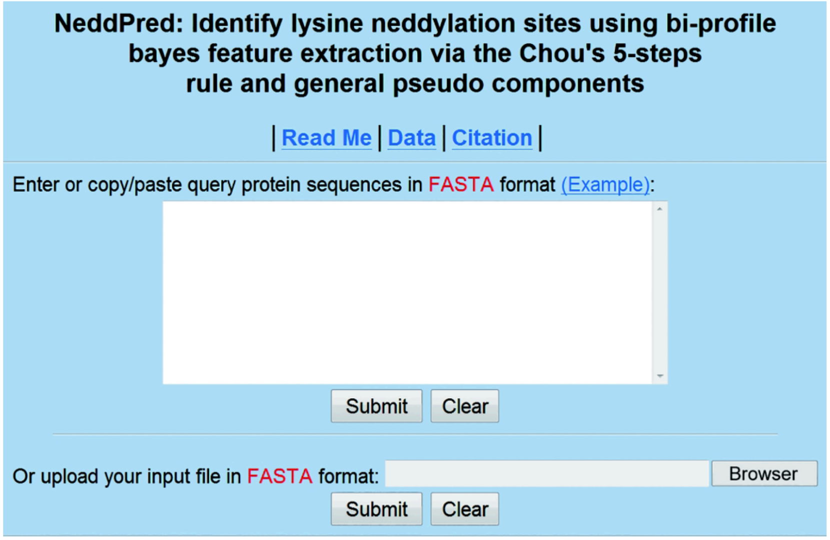


Fig. (1). The ROC curves of NeddPred with different evaluation strategies. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).



NeddPred: Identify lysine neddylation sites using bi-profile bayes feature extraction via the Chou's 5-steps rule and general pseudo components

Enter or copy/paste query protein sequences in FASTA format (Example):

Or upload your input file in FASTA format:

Fig. (2). The prediction interface of the web-server NeddPred. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

Table 5. The 42 BPB features ranked by the F-score method.

Order	Amino Acid Position	F-score	Order	Amino Acid Position	F-score
1	Pos_8[1]	0.5889	22	Neg_-1	0.0995
2	Pos_-4	0.3916	23	Pos_-9	0.0885
3	Pos_-3	0.3843	24	Neg_1	0.0832
4	Pos_1	0.3752	25	Pos_6	0.0744
5	Pos_-2	0.3665	26	Neg_10	0.0521
6	Pos_-7	0.3549	27	Neg_-5	0.0229
7	Pos_-5	0.3474	28	Neg_3	0.0195
8	Pos_-1	0.3353	29	Neg_-10	0.0124
9	Pos_7	0.3276	30	Neg_-4	0.0092
10	Pos_-10	0.2673	31	Neg_9	0.0063
11	Pos_5	0.2653	32	Neg_2	0.0056
12	Pos_10	0.2608	33	Neg_6	0.0023
13	Pos_-6	0.2416	34	Neg_-6	0.0012
14	Pos_4	0.2403	35	Neg_-8	0.0009
15	Pos_2	0.2290	36	Neg_-2	0.0005
16	Pos_3	0.2105	37	Neg_4	0.0005
17	Pos_-8	0.2005	38	Neg_-7	0.0004
18	Pos_9	0.1940	39	Neg_-9	0.0003
19	Neg_-3	0.1856	40	Neg_5	0.0000
20	Neg_7	0.1383	41	Pos_0	-1.0000
21	Neg_8	0.1133	42	Neg_0	-1.0000

[1] Pos_*i* and Neg_*j* mean position *i* in neddylated peptides and position *j* in non-neddylated peptides, respectively.

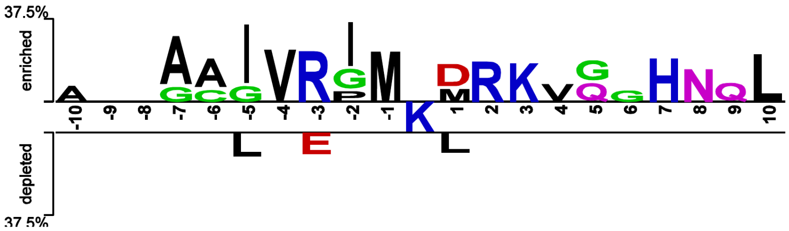


Fig. (3). Two Sample Logo of the position-specific residue composition around the 34 neddylation and 687 non-neddylation sites (t-test, P<0.05). (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

3.5. The Significant Features

As previously described, every lysine site in neddylated proteins was encoded as a 42-dimensional vector through BPB encoding. To clarify the contribution of different features for the prediction of neddylation sites, we used the F-score feature selection method to rank the 42 BPB features [90] (Table **5**). The higher the F-score of a feature is, the more important a feature will be.

Moreover, the position-specific residue composition of lysine-centric peptides was characterized by Two-Sample-Logo [91] in Fig. (**3**). As shown in Table **5** and Fig. (**3**), the 'Pos_8' feature was ranked at the top of the 42 BPB features, which imply that asparagine residue in position 8 around neddylation sites may play a key role in the identification of neddylation sites. The residues in positions (-4, -3, 1 and -2) around neddylation sites may play a relatively important role. The 42 BPB features ranked by the F-score may provide clues for deciphering the molecular mechanisms of neddylation.

CONCLUSION

In this paper, a bioinformatics tool named NeddPred was developed to identify neddylation sites using BPB encoding and fuzzy SVM. Experimental results showed that NeddPred yielded better performance than the existing neddylation sites predictor. Therefore, NeddPred will be a useful predictor for the accurate identification of neddylation sites. To provide convenience for researchers to study neddylation, a web-server for NeddPred was established. Feature analysis shows that BPB features at some positions may play a key role in the prediction of neddylation sites.

AUTHOR'S CONTRIBUTIONS

Z.J. wrote the manuscript and was involved in all the experimental steps. S.Y.W. constructed the online webserver of NeddPred. Both the authors approved the final version of this manuscript.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

HUMAN AND ANIMAL RIGHTS

No Animals/Humans were used for studies that are the basis of this research.

CONSENT FOR PUBLICATION

Not applicable.

AVAILABILITY OF DATA AND MATERIALS

The data supporting the findings of the article is available in the Bitbucket repository at https://bitbucket.org/asyavuz/neddypreddy, reference number [13].

FUNDING

This work was supported by the National Natural Science Foundation of China (No. 11701390); the Natural Science Foundation of Liaoning Province (No. 2019-BS-187); and the Scientific Research Fund Project in Liaoning Province Department of Education (No. JYT19027).

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

Declared none.

SUPPLEMENTARY MATERIAL

Supplementary material is available on the publisher's website along with the published article.

REFERENCES

[1] Jones, J.; Wu, K.; Yang, Y.; Guerrero, C.; Nillegoda, N.; Pan, Z.Q.; Huang, L. A targeted proteomic analysis of the ubiquitin-like modifier nedd8 and associated proteins. *J. Proteome Res.,* **2008**, *7*(3), 1274-1287. [http://dx.doi.org/10.1021/pr700749v] [PMID: 18247557]

[2] Rabut, G.; Peter, M. Function and regulation of protein neddylation. 'Protein modifications: beyond the usual suspects' review series. *EMBO Rep.,* **2008**, *9*(10), 969-976. [http://dx.doi.org/10.1038/embor.2008.183] [PMID: 18802447]

[3] Herrmann, J.; Lerman, L.O.; Lerman, A. Ubiquitin and ubiquitin-like proteins in protein regulation. *Circ. Res.,* **2007**, *100*(9), 1276-1291. [http://dx.doi.org/10.1161/01.RES.0000264500.11888.f0] [PMID: 17495234]

[4] Xirodimas, D.P. Novel substrates and functions for the ubiquitin-like molecule NEDD8. *Biochem. Soc. Trans.,* **2008**, *36*(Pt 5), 802-806. [http://dx.doi.org/10.1042/BST0360802] [PMID: 18793140]

[5] Yao, W.T.; Wu, J.F.; Yu, G.Y.; Wang, R.; Wang, K.; Li, L.H.; Chen, P.; Jiang, Y.N.; Cheng, H.; Lee, H.W.; Yu, J.; Qi, H.; Yu, X.J.; Wang, P.; Chu, Y.W.; Yang, M.; Hua, Z.C.; Ying, H.Q.; Hoffman, R.M.; Jeong, L.S.; Jia, L.J. Suppression of tumor angiogenesis by targeting the protein neddylation pathway. *Cell Death Dis.,* **2014**, *5*(2), e1059. http://dx.doi.org/10.1038/cddis.2014.21 PMID: 24525735

[6] Chen, Y.; Neve, R.L.; Liu, H. Neddylation dysfunction in Alzheimer's disease. *J. Cell. Mol. Med.,* **2012**, *16*(11), 2583-2591. [http://dx.doi.org/10.1111/j.1582-4934.2012.01604.x] [PMID:

22805479]

[7] Choo, Y.S.; Vogler, G.; Wang, D.; Kalvakuri, S.; Iliuk, A.; Tao, W.A.; Bodmer, R.; Zhang, Z. Regulation of parkin and PINK1 by neddylation. *Hum. Mol. Genet.,* **2012**, *21*(11), 2514-2523. [http://dx.doi.org/10.1093/hmg/dds070] [PMID: 22388932]

[8] Akbar, S.; Hayat, M. iMethyl-STTNC: Identification of N^6-methyladenosine sites by extending the idea of SAAC into Chou's PseAAC to formulate RNA sequences. *J. Theor. Biol.,* **2018**, *455*, 205-211. [http://dx.doi.org/10.1016/j.jtbi.2018.07.018] [PMID: 30031793]

[9] Ju, Z.; Wang, S.Y. Prediction of citrullination sites by incorporating k-spaced amino acid pairs into Chou's general pseudo amino acid composition. *Gene,* **2018**, *664*, 78-83. http://dx.doi.org/10.1016/j.gene.2018.04.055 PMID: 29694908

[10] Hussain, W.; Khan, Y.D.; Rasool, N.; Khan, S.A.; Chou, K.C. SPalmitoylC-PseAAC: A sequence-based model developed *via* Chou's 5-steps rule and general PseAAC for identifying S-palmitoylation sites in proteins. *Anal. Biochem.,* **2019**, *568*, 14-23. [http://dx.doi.org/10.1016/j.ab.2018.12.019] [PMID: 30593778]

[11] Li, F.; Zhang, Y.; Purcell, A.W.; Webb, G.I.; Chou, K.C.; Lithgow, T.; Li, C.; Song, J. Positive-unlabelled learning of glycosylation sites in the human proteome. *BMC Bioinformatics,* **2019**, *20*(1), 112. [http://dx.doi.org/10.1186/s12859-019-2700-1] [PMID: 30841845]

[12] Wang, L.; Zhang, R.; Mu, Y. Fu-SulfPred: Identification of protein s-sulfenylation sites by fusing forests *via* Chou's general PseAAC. *J. Theor. Biol.,* **2019**, *461*, 51-58. [http://dx.doi.org/10.1016/j.jtbi.2018.10.046] [PMID: 30365947]

[13] Yavuz, A.S.; Sözer, N.B.; Sezerman, O.U. Prediction of neddylation sites from protein sequences and sequence-derived properties. *BMC Bioinformatics,* **2015**, *16*(Suppl. 18), S9. [http://dx.doi.org/10.1186/1471-2105-16-S18-S9] [PMID: 26679222]

[14] Hussain, W.; Khan, Y.D.; Rasool, N.; Khan, S.A.; Chou, K.C. SPrenylC-PseAAC: A sequence-based model developed *via* Chou's 5-steps rule and general PseAAC for identifying S-prenylation sites in proteins. *J. Theor. Biol.,* **2019**, *468*, 1-11. [http://dx.doi.org/10.1016/j.jtbi.2019.02.007] [PMID: 30768975]

[15] Xiao, X.; Min, J.L.; Lin, W.Z.; Liu, Z.; Cheng, X.; Chou, K.C. iDrug-Target: predicting the interactions between drug compounds and target proteins in cellular networking *via* benchmark dataset optimization approach. *J. Biomol. Struct. Dyn.,* **2015**, *33*(10), 2221-2233. [http://dx.doi.org/10.1080/07391102.2014.998710] [PMID: 25513722]

[16] Liu, B.; Fang, L.; Wang, S.; Wang, X.; Li, H.; Chou, K.C. Identification of microRNA precursor with the degenerate K-tuple or Kmer strategy. *J. Theor. Biol.,* **2015**, *385*, 153-159. [http://dx.doi.org/10.1016/j.jtbi.2015.08.025] [PMID: 26362104]

[17] Liu, Z.; Xiao, X.; Yu, D.J.; Jia, J.; Qiu, W.R.; Chou, K.C. pRNAm-PC: Predicting N(6)-methyladenosine sites in RNA sequences *via* physical-chemical properties. *Anal. Biochem.,* **2016**, *497*, 60-67. [http://dx.doi.org/10.1016/j.ab.2015.12.017] [PMID: 26748145]

[18] Chen, W.; Feng, P.M.; Lin, H.; Chou, K.C. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. *Nucleic Acids Res.,* **2013**, *41*(6), e68. http://dx.doi.org/10.1093/nar/gks1450 PMID: 23303794

[19] Chen, W.; Feng, P.M.; Deng, E.Z.; Lin, H.; Chou, K.C. iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. *Anal. Biochem.,* **2014**, *462*, 76-83. [http://dx.doi.org/10.1016/j.ab.2014.06.022] [PMID: 25016190]

[20] Chen, W.; Feng, P.; Yang, H.; Ding, H.; Lin, H.; Chou, K.C. iRNA-AI: identifying the adenosine to inosine editing sites in RNA sequences. *Oncotarget,* **2017**, *8*(3), 4208-4217. [http://dx.doi.org/10.18632/oncotarget.13758] [PMID: 27926534]

[21] Lin, H.; Deng, E.Z.; Ding, H.; Chen, W.; Chou, K.C. iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. *Nucleic Acids Res.,* **2014**, *42*(21), 12961-12972. [http://dx.doi.org/10.1093/nar/gku1019] [PMID: 25361964]

[22] Feng, P.M.; Chen, W.; Lin, H.; Chou, K.C. iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. *Anal. Biochem.,* **2013**, *442*(1), 118-125. [http://dx.doi.org/10.1016/j.ab.2013.05.024] [PMID: 23756733]

[23] Ding, H.; Deng, E.Z.; Yuan, L.F.; Liu, L.; Lin, H.; Chen, W.; Chou, K.C. iCTX-type: a sequence-based predictor for identifying the types of conotoxins in targeting ion channels. *BioMed Res. Int.,* **2014**, *2014,* 286419. [http://dx.doi.org/10.1155/2014/286419] [PMID: 24991545]

[24] Khan, Y.D.; Jamil, M.; Hussain, W.; Rasool, N.; Khan, S.A.; Chou, K.C. pSSbond-PseAAC: Prediction of disulfide bonding sites by integration of PseAAC and statistical moments. *J. Theor. Biol.,* **2019**, *463*, 47-55. [http://dx.doi.org/10.1016/j.jtbi.2018.12.015] [PMID: 30550863]

[25] Jia, J.; Li, X.; Qiu, W.; Xiao, X.; Chou, K.C. iPPI-PseAAC(CGR): Identify protein-protein interactions by incorporating chaos game representation into PseAAC. *J. Theor. Biol.,* **2019**, *460*, 195-203. [http://dx.doi.org/10.1016/j.jtbi.2018.10.021] [PMID: 30312687]

[26] Chou, K.C. Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.,* **2011**, *273*(1), 236-247. [http://dx.doi.org/10.1016/j.jtbi.2010.12.024] [PMID: 21168420]

[27] Zhang, C.T.; Chou, K.C. An optimization approach to predicting protein structural class from amino acid composition. *Protein Sci.,* **1992**, *1*(3), 401-408. [http://dx.doi.org/10.1002/pro.5560010312] [PMID: 1304347]

[28] Chou, K.C.; Elrod, D.W. Bioinformatical analysis of G-protein-coupled receptors. *J. Proteome Res.,* **2002**, *1*(5), 429-433. [http://dx.doi.org/10.1021/pr025527k] [PMID: 12645914]

[29] Chou, K.C.; Cai, Y.D. Prediction and classification of protein subcellular location-sequence-order effect and pseudo amino acid composition. *J. Cell. Biochem.,* **2003**, *90*(6), 1250-1260. [http://dx.doi.org/10.1002/jcb.10719] [PMID: 14635197]

[30] Hu, L.; Huang, T.; Shi, X.; Lu, W.C.; Cai, Y.D.; Chou, K.C. Predicting functions of proteins in mouse based on weighted protein-protein interaction network and protein hybrid properties. *PLoS One,* **2011**, *6*(1), e14556. [http://dx.doi.org/10.1371/journal.pone.0014556] [PMID: 21283518]

[31] Cai, Y.D.; Feng, K.Y.; Lu, W.C.; Chou, K.C. Using LogitBoost classifier to predict protein structural classes. *J. Theor. Biol.,* **2006**, *238*(1), 172-176. [http://dx.doi.org/10.1016/j.jtbi.2005.05.034] [PMID: 16043193]

[32] Chou, K.C. Impacts of bioinformatics to medicinal chemistry. *Med. Chem.,* **2015**, *11*(3), 218-234. [http://dx.doi.org/10.2174/1573406411666141229162834] [PMID: 25548930]

[33] Chou, K.C. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics,* **2005**, *21*(1), 10-19. [http://dx.doi.org/10.1093/bioinformatics/bth466] [PMID: 15308540]

[34] Dehzangi, A.; Heffernan, R.; Sharma, A.; Lyons, J.; Paliwal, K.; Sattar, A. Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou′s general PseAAC. *J. Theor. Biol.,* **2015**, *364*, 284-294. [http://dx.doi.org/10.1016/j.jtbi.2014.09.029] [PMID: 25264267]

[35] Behbahani, M.; Mohabatkar, H.; Nosrati, M. Analysis and comparison of lignin peroxidases between fungi and bacteria using three different modes of Chou's general pseudo amino acid composition. *J. Theor. Biol.,* **2016**, *411*, 1-5. [http://dx.doi.org/10.1016/j.jtbi.2016.09.001] [PMID: 27615149]

[36] Kabir, M.; Hayat, M. iRSpot-GAEnsC: identifing recombination spots *via* ensemble classifier and extending the concept of Chou's PseAAC to formulate DNA samples. *Mol. Genet. Genomics,* **2016**, *291*(1), 285-296. [http://dx.doi.org/10.1007/s00438-015-1108-5] [PMID: 26319782]

[37] Meher, P.K.; Sahu, T.K.; Saini, V.; Rao, A.R. Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC. *Sci. Rep.,* **2017**, *7*, 42362. [http://dx.doi.org/10.1038/srep42362] [PMID: 28205576]

[38] Ju, Z.; He, J.J. Prediction of lysine propionylation sites using biased SVM and incorporating four different sequence features into Chou's PseAAC. *J. Mol. Graph. Model.,* **2017**, *76*, 356-363. [http://dx.doi.org/10.1016/j.jmgm.2017.07.022] [PMID: 28763688]

[39] Yu, B.; Li, S.; Qiu, W.Y.; Chen, C.; Chen, R.X.; Wang, L.; Wang, M.H.; Zhang, Y. Accurate prediction of subcellular location of apoptosis proteins combining Chou's PseAAC and PsePSSM based on wavelet denoising. *Oncotarget,* **2017**, *8*(64), 107640-107665. [http://dx.doi.org/10.18632/oncotarget.22585] [PMID: 29296195]

[40] Ahmad, J.; Hayat, M. MFSC: Multi-voting based feature selection for classification of Golgi proteins by adopting the general form of Chou's PseAAC components. *J. Theor. Biol.,* **2019**, *463*, 99-109. [http://dx.doi.org/10.1016/j.jtbi.2018.12.017] [PMID: 30562500]

[41] Contreras-Torres, E. Predicting structural classes of proteins by incorporating their global and local physicochemical and conformational properties into general Chou's PseAAC. *J. Theor. Biol.,* **2018**, *454*, 139-145. [http://dx.doi.org/10.1016/j.jtbi.2018.05.033] [PMID: 29870696]

[42] Zhang, S.; Liang, Y. Predicting apoptosis protein subcellular localization by integrating auto-cross correlation and PSSM into Chou's PseAAC. *J. Theor. Biol.,* **2018**, *457*, 163-169. [http://dx.doi.org/10.1016/j.jtbi.2018.08.042] [PMID: 30179589]

[43] Tahir, M.; Hayat, M.; Khan, S.A. iNuc-ext-PseTNC: an efficient ensemble model for identification of nucleosome positioning by extending the concept of Chou's PseAAC to pseudo-tri-nucleotide composition. *Mol. Genet. Genomics,* **2019**, *294*(1), 199-210. [http://dx.doi.org/10.1007/s00438-018-1498-2] [PMID: 30291426]

[44] Chou, K.C. An unprecedented revolution in medicinal chemistry driven by the progress of biological science. *Curr. Top. Med. Chem.,* **2017**, *17*(21), 2337-2358. [http://dx.doi.org/10.2174/1568026617666170414145508] [PMID: 28413951]

[45] Shen, H.B.; Chou, K.C. PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition. *Anal. Biochem.,* **2008**, *373*(2), 386-388. [http://dx.doi.org/10.1016/j.ab.2007.10.012] [PMID: 17976365]

[46] Du, P.; Wang, X.; Xu, C.; Gao, Y. PseAAC-Builder: a cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions. *Anal. Biochem.,* **2012**, *425*(2), 117-119. [http://dx.doi.org/10.1016/j.ab.2012.03.015] [PMID: 22459120]

[47] Cao, D.S.; Xu, Q.S.; Liang, Y.Z. propy: a tool to generate various modes of Chou's PseAAC. *Bioinformatics,* **2013**, *29*(7), 960-962. [http://dx.doi.org/10.1093/bioinformatics/btt072] [PMID: 23426256]

[48] Du, P.; Gu, S.; Jiao, Y. PseAAC-General: fast building various modes of general form of Chou's pseudo-amino acid composition for large-scale protein datasets. *Int. J. Mol. Sci.,* **2014**, *15*(3), 3495-3506. [http://dx.doi.org/10.3390/ijms15033495] [PMID: 24577312]

[49] Chou, K.C. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. *Curr. Proteomics,* **2009**, *6*(4), 262-274. [http://dx.doi.org/10.2174/157016409789973707]

[50] Chen, W.; Lei, T.Y.; Jin, D.C.; Lin, H.; Chou, K.C. PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition. *Anal. Biochem.,* **2014**, *456*, 53-60. [http://dx.doi.org/10.1016/j.ab.2014.04.001] [PMID: 24732113]

[51] Chen, W.; Feng, P.; Ding, H.; Lin, H.; Chou, K.C. iRNA-Methyl: Identifying N(6)-methyladenosine sites using pseudo nucleotide composition. *Anal. Biochem.,* **2015**, *490*, 26-33. [http://dx.doi.org/10.1016/j.ab.2015.08.021] [PMID: 26314792]

[52] Liu, B.; Yang, F.; Huang, D.S.; Chou, K.C. iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC. *Bioinformatics,* **2018**, *34*(1), 33-40. [http://dx.doi.org/10.1093/bioinformatics/btx579] [PMID: 28968797]

[53] Tahir, M.; Tayara, H.; Chong, K.T. iRNA-PseKNC(2methyl): Identify RNA 2′-O-methylation sites by convolution neural network and Chou's pseudo components. *J. Theor. Biol.,* **2019**, *465*, 1-6. [http://dx.doi.org/10.1016/j.jtbi.2018.12.034] [PMID: 30590059]

[54] Liu, B.; Liu, F.; Wang, X.; Chen, J.; Fang, L.; Chou, K.C. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res.,* **2015**, *43*(W1), W65-71. [http://dx.doi.org/10.1093/nar/gkv458] [PMID: 25958395]

[55] Liu, B.; Wu, H. Pse-in-One 2.0: An improved package of web servers for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nat. Sci.,* **2017**, *9*, 67-91. [http://dx.doi.org/10.4236/ns.2017.94007]

[56] Shao, J.; Xu, D.; Tsai, S.N.; Wang, Y.; Ngai, S.M. Computational identification of protein methylation sites through bi-profile Bayes feature extraction. *PLoS One,* **2009**, *4*(3), e4920. [http://dx.doi.org/10.1371/journal.pone.0004920] [PMID: 19290060]

[57] Song, J.; Tan, H.; Shen, H.; Mahmood, K.; Boyd, S.E.; Webb, G.I.; Akutsu, T.; Whisstock, J.C. Cascleave: towards more accurate prediction of caspase substrate cleavage sites. *Bioinformatics,* **2010**, *26*(6), 752-760. [http://dx.doi.org/10.1093/bioinformatics/btq043] [PMID: 20130033]

[58] Wang, Y.; Zhang, Q.; Sun, M.A.; Guo, D. High-accuracy prediction of bacterial type III secreted effectors based on position-specific amino acid composition profiles. *Bioinformatics,* **2011**, *27*(6), 777-784. [http://dx.doi.org/10.1093/bioinformatics/btr021] [PMID: 21233168]

[59] Jia, C.; Liu, T.; Chang, A.K.; Zhai, Y. Prediction of mitochondrial proteins of malaria parasite using bi-profile Bayes feature extraction. *Biochimie,* **2011**, *93*(4), 778-782. [http://dx.doi.org/10.1016/j.biochi.2011.01.013] [PMID: 21281691]

[60] Jia, C.Z.; Liu, T.; Wang, Z.P. O-GlcNAcPRED: a sensitive predictor to capture protein O-GlcNAcylation sites. *Mol. Biosyst.,* **2013**, *9*(11), 2909-2913. [http://dx.doi.org/10.1039/c3mb70326f] [PMID: 24056994]

[61] Xu, Y.; Ding, Y.X.; Ding, J.; Lei, Y.H.; Wu, L.Y.; Deng, N.Y. iSuc-PseAAC: predicting lysine succinylation in proteins by incorporating peptide position-specific propensity. *Sci. Rep.,* **2015**, *5*, 10184. [http://dx.doi.org/10.1038/srep10184] [PMID: 26084794]

[62] Xu, Y.; Li, L.; Ding, J.; Wu, L.Y.; Mai, G.; Zhou, F. Gly-PseAAC: Identifying protein lysine glycation through sequences. *Gene,* **2017**, *602*, 1-7. [http://dx.doi.org/10.1016/j.gene.2016.11.021] [PMID: 27845204]

[63] Qiu, W.R.; Sun, B.Q.; Tang, H.; Huang, J.; Lin, H. Identify and analysis crotonylation sites in histone by using support vector machines. *Artif. Intell. Med.,* **2017**, *83*, 75-81. d [http://dx.doi.org/10.1016/j.artmed.2017.02.007] [PMID: 28283358]

[64] Lin, C.F.; Wang, S.D. Fuzzy support vector machines. *IEEE Trans. Neural Netw.,* **2002**, *13*(2), 464-471. [http://dx.doi.org/10.1109/72.991432] [PMID: 18244447]

[65] Batuwita, R.; Palade, V. Class imbalance learning methods for support vector machines. *Imbalanced Learning: Foundations, Algorithms, and Applications*; He, H.; Ma, Y., Eds.; John Wiley Hoboken, NJ, **2013**, pp. 83-96. [http://dx.doi.org/10.1002/9781118646106.ch5]

[66] Chang, C.C.; Lin, C.J. LIBSVM: A library for support vector machines. *Acm T. Intel. Syst. Tec.,* **2011**, *2*(3), 1-27. [http://dx.doi.org/10.1145/1961189.1961199]

[67] Chou, K.C. Prediction of protein signal sequences and their cleavage sites. *Proteins,* **2001**, *42*(1), 136-139. [http://dx.doi.org/10.1002/1097-0134(20010101)42:1<136::AID-PROT130>3.0.CO;2-F] [PMID: 11093267]

[68] Chou, K.C. Using subsite coupling to predict signal peptides. *Protein Eng.,* **2001**, *14*(2), 75-79. [http://dx.doi.org/10.1093/protein/14.2.75] [PMID: 11297664]

[69] Chou, K.C. Prediction of signal peptides using scaled window. *Peptides,* **2001**, *22*(12), 1973-1979. [http://dx.doi.org/10.1016/S0196-9781(01)00540-X] [PMID: 11786179]

[70] Cheng, X.; Xiao, X.; Chou, K.C. pLoc-mPlant: predict subcellular localization of multi-location plant proteins by incorporating the optimal GO information into general PseAAC. *Mol. Biosyst.,* **2017**, *13*(9), 1722-1727. [http://dx.doi.org/10.1039/C7MB00267J] [PMID: 28702580]

[71] Cheng, X.; Xiao, X.; Chou, K.C. pLoc-mVirus: predict subcellular localization of multi-location virus proteins *via* incorporating the optimal GO information into general PseAAC. *Gene,* **2017**, *628*, 315-321. [http://dx.doi.org/10.1016/j.gene.2017.07.036] [PMID: 28728979]

[72] Cheng, X.; Xiao, X.; Chou, K.C. pLoc-mGneg: predict subcellular localization of Gram-negative bacterial proteins by deep gene ontology learning *via* general PseAAC. *Genomics,* **2017**, *110*(4), 231-239. [http://dx.doi.org/10.1016/j.ygeno.2017.10.002] [PMID: 28989035]

[73] Cheng, X.; Xiao, X.; Chou, K.C. pLoc-mEuk: Predict subcellular localization of multi-label eukaryotic proteins by extracting the key GO information into general PseAAC. *Genomics,* **2018**, *110*(1), 50-

58.
[http://dx.doi.org/10.1016/j.ygeno.2017.08.005] [PMID: 28818512]

[74] Cheng, X.; Zhao, S.G.; Lin, W.Z.; Xiao, X.; Chou, K.C. pLoc-m Animal: predict subcellular localization of animal proteins with both single and multiple sites. *Bioinformatics,* **2017**, *33*(22), 3524-3531.
[http://dx.doi.org/10.1093/bioinformatics/btx476] [PMID: 29036535]

[75] Xiao, X.; Cheng, X.; Su, S.; Nao, Q. pLoc-mGpos: incorporate key gene ontology information into general PseAAC for predicting subcellular localization of Gram-positive bacterial proteins. *Nat. Sci.,* **2017**, *9*, 331-349.
[http://dx.doi.org/10.4236/ns.2017.99032]

[76] Cheng, X.; Zhao, S.G.; Xiao, X.; Chou, K.C. iATC-mISF: a multi-label classifier for predicting the classes of anatomical therapeutic chemicals. *Bioinformatics,* **2017**, *33*(3), 341-346.
[http://dx.doi.org/10.1093/bioinformatics/btx387] [PMID: 28172617]

[77] Cheng, X.; Zhao, S.G.; Xiao, X.; Chou, K.C. iATC-mHyb: a hybrid multi-label classifier for predicting the classification of anatomical therapeutic chemicals. *Oncotarget,* **2017**, *8*(35), 58494-58503.
[http://dx.doi.org/10.18632/oncotarget.17028] [PMID: 28938573]

[78] Chou, K.C. Some remarks on predicting multi-label attributes in molecular biosystems. *Mol. Biosyst.,* **2013**, *9*(6), 1092-1100.
[http://dx.doi.org/10.1039/c3mb25555g] [PMID: 23536215]

[79] Nakashima, H.; Nishikawa, K. Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. *J. Mol. Biol.,* **1994**, *238*(1), 54-61.
[http://dx.doi.org/10.1006/jmbi.1994.1267] [PMID: 8145256]

[80] Wan, S.; Mak, M.W.; Kung, S.Y. Ensemble linear neighborhood propagation forpredicting subchloro plast localization of multi-location proteins. *J. Proteome Res.,* **2016**, *15*(12), 4755-4762.
[http://dx.doi.org/10.1021/acs.jproteome.6b00686] [PMID: 27766879]

[81] Atchley, W.R.; Zhao, J.; Fernandes, A.D.; Drüke, T. Solving the protein sequence metric problem. *Proc. Natl. Acad. Sci. USA,* **2005**, *102*(18), 6395-6400.
[http://dx.doi.org/10.1073/pnas.0408677102] [PMID: 15851683]

[82] Sagara, J.I.; Shimizu, S.; Kawabata, T.; Nakamura, S.; Ikeguchi, M.; Shimizu, K. The use of sequence comparison to detect 'identities' in tRNA genes. *Nucleic Acids Res.,* **1998**, *26*(8), 1974-1979.
[http://dx.doi.org/10.1093/nar/26.8.1974] [PMID: 9518491]

[83] Chen, Y.Z.; Tang, Y.R.; Sheng, Z.Y.; Zhang, Z. Prediction of mucin-type O-glycosylation sites in mammalian proteins using the composition of k-spaced amino acid pairs. *BMC Bioinformatics,* **2008**, *9*(1), 101.
[http://dx.doi.org/10.1186/1471-2105-9-101] [PMID: 18282281]

[84] Chou, K.C.; Shen, H.B. Recent advances in developing web-servers for predicting protein attributes. *Nat. Sci.,* **2009**, *1*(2), 63-92.
[http://dx.doi.org/10.4236/ns.2009.12011]

[85] Cheng, X.; Xiao, X.; Chou, K.C. pLoc-mHum: predict subcellular localization of multi-location human proteins *via* general PseAAC to winnow out the crucial GO information. *Bioinformatics,* **2018**, *34*(9), 1448-1456.
[http://dx.doi.org/10.1093/bioinformatics/btx711] [PMID: 29106451]

[86] Cheng, X.; Xiao, X.; Chou, K.C. pLoc_bal-mGneg: Predict subcellular localization of Gram-negative bacterial proteins by quasi-balancing training dataset and general PseAAC. *J. Theor. Biol.,* **2018**, *458*, 92-102. d
[http://dx.doi.org/10.1016/j.jtbi.2018.09.005] [PMID: 30201434]

[87] Cheng, X.; Xiao, X.; Chou, K.C. pLoc_bal-mPlant: predict subcellular localization of plant proteins by general PseAAC and balancing training dataset. *Curr. Pharm. Des.,* **2018**, *24*(34), 4013-4022. e
[http://dx.doi.org/10.2174/1381612824666181119145030] [PMID: 30451108]

[88] Chou, K.C.; Cheng, X.; Xiao, X. pLoc_bal-mHum: Predict subcellular localization of human proteins by PseAAC and quasi-balancing training dataset. *Genomics,* **2019**, *111*(6), 1274-1282.
[http://dx.doi.org/10.1016/j.ygeno.2018.08.007] [PMID: 30179658]

[89] Xiao, X.; Cheng, X.; Chen, G.; Mao, Q.; Chou, K.C. pLoc_bal-mVirus: predict subcellular localization of multi-label virus proteins by PseAAC and IHTS treatment to balance training dataset. *Med. Chem.,* **2019**, *15*(5), 496-509.
[http://dx.doi.org/10.2174/1573406415666181217114710] [PMID: 30556503]

[90] Chen, Y.W.; Lin, C.J. Combining svms with various feature selection strategies. *Feature Extraction*; Guyon, I.; Nikravesh, N.; Gunn, S.; Zadeh, L., Eds.; Springer: Berlin, Germany, **2006**, pp. 315-324.
[http://dx.doi.org/10.1007/978-3-540-35488-8_13]

[91] Vacic, V.; Iakoucheva, L.M.; Radivojac, P. Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. *Bioinformatics,* **2006**, *22*(12), 1536-1537.
[http://dx.doi.org/10.1093/bioinformatics/btl151] [PMID: 16632492]