http://www.scirp.org/journal/ns Natural Science, 2018, Vol. 10, (No. 9), pp: 370-384

mLysPTMpred: Multiple Lysine PTM Site Prediction Using Combination of SVM with Resolving Data Imbalance Issue

Md. Al Mehedi Hasan[1], Shamim Ahmad[2]

[1]Department of Computer Science and Engineering, Rajshahi University of Engineering and Technology, Rajshahi, Bangladesh; [2]Department of Computer Science and Engineering, University of Rajshahi, Rajshahi, Bangladesh

Correspondence to: Md. Al Mehedi Hasan, mehedi_ru@yahoo.com; Shamim Ahmad, shamim_cst@yahoo.com
Keywords: Multi-Label PTM Site Predictor, Sequence-Coupling Model, General PseAAC, Data Imbalance Issue, Different Error Costs, Support Vector Machine

Received: August 7, 2018 **Accepted:** September 27, 2018 **Published:** September 30, 2018

ABSTRACT

Post-translational modification (PTM) increases the functional diversity of proteins by introducing new functional groups to the side chain of amino acid of a protein. Among all amino acid residues, the side chain of lysine (K) can undergo many types of PTM, called K-PTM, such as "acetylation", "crotonylation", "methylation" and "succinylation" and also responsible for occurring multiple PTM in the same lysine of a protein which leads to the requirement of multi-label PTM site identification. However, most of the existing computational methods have been established to predict various single-label PTM sites and a very few have been developed to solve multi-label issue which needs further improvement. Here, we have developed a computational tool termed mLysPTMpred to predict multi-label lysine PTM sites by 1) incorporating the sequence-coupled information into the general pseudo amino acid composition, 2) balancing the effect of skewed training dataset by Different Error Cost method, and 3) constructing a multi-label predictor using a combination of support vector machine (SVM). This predictor achieved 83.73% accuracy in predicting the multi-label PTM site of K-PTM types. Moreover, all the experimental results along with accuracy outperformed than the existing predictor iPTM-mLys. A user-friendly web server of mLysPTMpred is available at http://research.ru.ac.bd/mLysPTMpred/.

1. INTRODUCTION

The structural and functional diversities of proteins as well as plasticity and dynamics of living cells are significantly dominated by the post-translational modifications (PTMs) [1]. PTMs are also responsible for expanding the genetic code and for regulating cellular physiology [2-4]. It changes the properties of a protein by proteolytic cleavage or adding a modifying group to one or more amino acids [5, 6].

In general, the side chain of lysine plays the key role in increasing the complexity of PTM network [7].

The lysine residue in proteins can experience various types of PTM, called K-PTM, such as methylation, acetylation, biotinylation, ubiquitination, ubiquitin-like modifications, propionylation, and butyrylation [7, 8]. Moreover, some lysine residues in proteins can undergo multiple K-PTMs which lead to the requirement of multi-label PTM site identification. This kind of multiplex lysine residues in proteins may have some exceptional functions that require special attention [9]. As a result, the identification of multiple K-PTM sites in proteins has become a vital question in cellular physiology and pathology, which in turns, helps in providing some valuable evidence for both biomedical research and drug development [6, 9].

However, the purely experimental technique such as mass spectrometry, peptide micro-array, liquid chromatography, etc., to determine the exact modified sites of protein is expensive as well as time-consuming, especially for large-scale datasets. In this context, it is highly demanded to use computational approaches to identify the K-PTM sites effectively and accurately [8]. Meanwhile, many computational methods have been developed to predict the modification sites in proteins for different K-PTM types which called single-label prediction [10-16]. According to our best knowledge, so far one computational tool has been developed to predict several different K-PTM types simultaneously for multiplex lysine residues. It is noted that various types of multi-label systems have been successfully studied in some other fields of computational biology [17, 18]. However, in order to meet the current demand to produce efficient high-throughput tools, additional effort is required to further improve the prediction quality [9].

In the development of computational classifier, one of the major challenges is to handle imbalance dataset problem [8, 15, 19, 20], as it is found in most of the dataset for this kind of prediction, the number negative subset is much larger than the corresponding positive subset [8, 15]. As the real world picture is that here the non K-type modification sites are always the majority compared with the K-type modification ones, so naturally the predictor should be biased to the non K-type modification sites. Here the problem is that, for this type of predictors may interpret many K-PTM sites as non K-PTM sites [21-23]. But, the information about the K-PTM sites is mostly desired than non K-PTM sites. As a result, it is crucial to find an effective solution to balance this kind of bias consequence.

The current study has been begun with an attempt to address the problems mentioned above and then tried to develop a more powerful predictor using combination of support vector machine which can be used to predict the multiple K-type modification sites of proteins. In this predictor, the Different Error Costs (DEC) method [24-26] has been used to resolve the data imbalance issue. It should be noted here that the features used in this predictor are extracted by using vectorized sequence-coupling model [27]. In the recent works, the performance of iPTM-mLys [9] on a large set of proteins has been studied in [9]. Therefore, in order to compare the performance of mLysPTMpred with the system iPTM-mLys [9], we use the exactly same dataset employing the commonly used stratified 5-fold cross-validation [9]. Since the information about the exact 5-way splits used in previous studies [9] is not available, so we have performed five complete runs of 5-fold-crossvalidation (*i.e.* 25 runs in total), where each complete run of 5-fold cross-validation uses a different 5-way split. The use of multiple runs with different splits helps to validate the stability and the statistical significance of the results. Finally, the average results of all metrics found from this study have been reported. Our experimental results indicate that mLysPTMpred achieves better results than those found from iPTM-mLys [9].

In order to launch a useful sequence-based statistical predictor for a biological system as demonstrated in a series of recent publications [8, 15, 28-35], the Chou's five-step rules [36] should be followed: 1) construct or select a valid benchmark dataset to train and test the predictor, 2) formulate the biological sequence samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted, 3) introduce or develop a powerful algorithm (or engine) to operate the prediction, 4) properly perform cross-validation tests to objectively evaluate its anticipated accuracy, and 5) establish a user-friendly webserver that is accessible to the public.

2. MATERIAL AND METHODS

2.1. Benchmark Dataset

iPTM-mLys's [9] benchmark dataset set has been used in this study. iPTM-mLys's dataset was de-

rived from the 1769 protein sequences from human. These 1769 protein sequences were collected from the web site at http://www.uniprot.org/ by giving various constrains such as 1) experimental assertion for evidence, 2) consider only human protein sequences, and 3) use keywords of "acetyllysine", "crotonyllysine", "methyllysine" or "succinyllysine" in the advance search option.

In iPTM-mLys [9], according to Chou's scheme, a peptide sample was generally expressed by

$$P_\xi(\mathbb{K}) = R_{-\xi}R_{-(\xi-1)}\cdots R_{-2}R_{-1}\mathbb{K}R_1R_2\cdots R_{+(\xi-1)}R_{+\xi} \tag{1}$$

where the subscript ξ is an integer, $R_{-\xi}$ represents the ξ-th up stream amino acid residue from the center, the $R_{+\xi}$ represents the ξ-th downstream amino acid residue, and so forth.

The $(2\xi+1)$-tuple peptide sample $P_\xi(\mathbb{K})$ was further classified into the following two categories [9]

$$P_\xi(\mathbb{K}) \in \begin{cases} P_\xi^+(\mathbb{K}), \text{if its center is K} - \text{PTM site} \\ P_\xi^-(\mathbb{K}), \text{otherwise} \end{cases} \tag{2}$$

where $P_\xi^+(\mathbb{K})$ denotes a true K-PTM segment with K at its center, $P_\xi^-(\mathbb{K})$ a false K-PTM segment with K at its center, and the symbol $\in$ means "a member of" in the set theory.

In iPTM-mLys's work, $(2\xi+1)$-tuple peptide window was used to collect peptide segment that have K at the center. It should be mentioned here that if the upstream or downstream in a protein sequence is less than ξ or greater than L-ξ (L is the length of the protein sequence concerned) then the lacking amino acid has been filled with the same residue as its nearest one [9].

After applying some screening procedure based on some constraints on that collected peptide samples, for example, considering window size, keep only one when two or more samples share same sequence, iPTM-mLys finally constructed a benchmark dataset [9]. It should be mentioned here that iPTM-mLys constructed four bench mark dataset for "acetylation", "crotonylation", "methylation" and "succinylation", respectively using same the above mentioned procedure. The detail procedure about the construction of iPTM-mLys's benchmark dataset is explained in [9].

The four benchmark dataset $S_\xi(\mathbb{K})$ in iPTM-mLys's study was formulated as

$$\begin{cases} S_\xi(\text{acetylation}) = S_\xi^+(\text{acetylation}) \cup S_\xi^-(\text{acetylation}) \\ S_\xi(\text{crotonylation}) = S_\xi^+(\text{crotonylation}) \cup S_\xi^-(\text{crotonylation}) \\ S_\xi(\text{methylation}) = S_\xi^+(\text{methylation}) \cup S_\xi^-(\text{methylation}) \\ S_\xi(\text{succinylation}) = S_\xi^+(\text{succinylation}) \cup S_\xi^-(\text{succinylation}) \end{cases} \tag{3}$$

where the positive subset $S_\xi^+(\text{acetylation})$ contains only the peptide samples with their center residues K (Equation (3)) confirmed by experiments being able to be of acetylation, while the negative subset $S_\xi^-(\text{acetylation})$ only contains those samples unable to be of acetylation, and the symbol $\cup$ means union in the set theory. Likewise, the remaining three sub-equations in Equation (3) have exactly the same definition but refer to "crotonylation", "methylation" and "succinylation", respectively.

Using numeric values, in iPTM-mLys's study, the Equation (3) was formulated as

$$\begin{cases} S_\xi(1) = S_\xi^+(1) \cup S_\xi^-(1) \\ S_\xi(2) = S_\xi^+(2) \cup S_\xi^-(2) \\ S_\xi(3) = S_\xi^+(3) \cup S_\xi^-(3) \\ S_\xi(4) = S_\xi^+(4) \cup S_\xi^-(4) \end{cases} \tag{4}$$

where the numerical argument 1, 2, 3 or 4 denotes 'acetylation', 'crotonylation', 'methylation' or 'succinylation', respectively.

Note that, depending on some preliminary test, window size was selected as 27 (2*ξ + 1) in iPTM-mLys's study, where $\xi = 13$. Thus, the benchmark dataset obtained by iPTM-mLys for $S_{\xi=13}(1)$, $S_{\xi=13}(2)$, $S_{\xi=13}(3)$, and $S_{\xi=13}(4)$ are available at online supplementary materials

(http://research.ru.ac.bd/mLysPTMpred/) as Supporting Information. It should be mention that our published online supplementary materials are taken from iPTM-mLys's work [9]. A summary of this benchmark dataset is given in **Table 1**.

2.2. Feature Extraction

The appropriate features of protein sequences or samples plays very important roles for the prediction of PTM site, as a result it draws the much attention of scientist that how to select the core and essential features of protein samples but this task becomes harder as this types of features are either hidden or burred in the complicated protein sequences. As most existing machine learning algorithm can handle only vector but not sequence sample, one of the critical problem in bioinformatics is how to extract vector from biological sequence with keeping considerable sequence characteristics [7].

To avoid complete losing the sequence pattern information for a protein, the pseudo amino acid composition or PseAAC [37] was proposed. Ever since the concept of Chou's PseAAC was proposed, it has been widely used in nearly all the areas of computational proteomics [38, 39]). Because it has been widely and increasingly used, recently a very powerful web-server called "Pse-in-One" [40] and its updated version "Pse-in-One2.0" [41] 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.

In this paper, the incorporation of sequence-coupling model [27, 42] into Chou's general PseAAC [36] has been adopted to extract feature from peptide segment. Based on the concept of sequence-coupled information [27, 42] into the general PseAAC, the peptide sequence of Equation (1) can be formulated as

$$P_{\xi}\left(\circledast\right)=P_{\xi}^{+}\left(\circledast\right)-P_{\xi}^{-}\left(\circledast\right) \tag{5}$$

where

$$P_{\xi}^{+}\left(\circledast\right)=\begin{bmatrix} p_{-\xi}^{+}\left(R_{-\xi}\mid R_{-(\xi-1)}\right) \\ p_{-(\xi-1)}^{+}\left(R_{-(\xi-1)}\mid R_{-(\xi-2)}\right) \\ \vdots \\ p_{-2}^{+}\left(R_{-2}\mid R_{-1}\right) \\ p_{-1}^{+}\left(R_{-1}\right) \\ p_{+1}^{+}\left(R_{+1}\right) \\ p_{+2}^{+}\left(R_{+2}\mid R_{+1}\right) \\ \vdots \\ p_{+(\xi-1)}^{+}\left(R_{+(\xi-1)}\mid R_{+(\xi-2)}\right) \\ p_{+\xi}^{+}\left(R_{+\xi}\mid R_{+(\xi-1)}\right) \end{bmatrix} \tag{6}$$

and

Table 1. Summary of the four benchmark dataset.

Attribute	PTM Type and Number of Samples			
	Ace	Cro	Met	Suc
	S(1)	S(2)	S(3)	S(4)
Positive	3991	115	127	1169
Negative	2403	6279	6267	5225

Ace, acetylation; Cro, crotonylation; Met, methylation; Suc, succinylation.

$$P_{\xi}^{-}(\circledast)=\begin{bmatrix} p_{-\xi}^{-}\left(R_{-\xi}\mid R_{-(\xi-1)}\right) \\ p_{-(\xi-1)}^{-}\left(R_{-(\xi-1)}\mid R_{-(\xi-2)}\right) \\ \vdots \\ p_{-2}^{-}\left(R_{-2}\mid R_{-1}\right) \\ p_{-1}^{-}\left(R_{-1}\right) \\ p_{+1}^{-}\left(R_{+1}\right) \\ p_{+2}^{-}\left(R_{+2}\mid R_{+1}\right) \\ \vdots \\ p_{+(\xi-1)}^{-}\left(R_{+(\xi-1)}\mid R_{+(\xi-2)}\right) \\ p_{+\xi}^{-}\left(R_{+\xi}\mid R_{+(\xi-1)}\right) \end{bmatrix} \tag{7}$$

In Equation (5) $p_{-\xi}^{+}\left(R_{-\xi}\mid R_{-(\xi-1)}\right)$ is the conditional probability of amino acid $R_{-\xi}$ occurring at the left 1st position (see Equation (1)) given that its closest right neighbor is $R_{-(\xi-1)}$, $p_{-(\xi-1)}^{+}\left(R_{-(\xi-1)}\mid R_{-(\xi-2)}\right)$ is the conditional probability of amino acid $R_{-(\xi-1)}$ occurring at the left 2nd position given that its closest right neighbor is $R_{-(\xi-2)}$, and so forth. It should be mentioned here that in Equation (6), only $p_{-1}^{+}\left(R_{-1}\right)$ and $p_{+1}^{+}\left(R_{+1}\right)$ are of non-conditional probability since the right neighbor of R_{-1} and the left neighbor of R_{+1} are always K. All these probability values can be easily derived from the positive benchmark dataset given in Supporting Information as done in [42]. Likewise, the components in Equation (7) are the same as those in Equation (6) except for that they are derived from the negative benchmark dataset given in Supporting Information.

2.3. SVM Classification

The modeling algorithm of SVM searches an optimal hyperplane with the maximum margin for separating two classes by finding a solution of the following constraint optimization problem [43-45]:

$$\begin{aligned} &\text{maximize}_{\alpha}\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j k\left(x_i,x_j\right) \\ &\text{subject to}:\ \sum_{i=1}^{n}y_i\alpha_i=0,\ 0\le\alpha_i\le C\ \ \text{for all}\ i=1,2,3,\cdots,n \end{aligned} \tag{8}$$

where $x_i\in R^p$ and $y_i\in\{-1,+1\}$ is the class label of x_i, $1\le i\le n$.

Finally, the discriminant function of SVM by involving the kernel function takes the following form

$$f\left(x\right)=\sum_{i}^{n}\alpha_i y_i k\left(x,x_i\right)+b \tag{9}$$

It noted here that a kernel function and its parameter have to be chosen to build a SVM classifier [43-45]. In this work, radial basis function kernel has been used to build SVM classifier which is defined below:

$K\left(x_i,x_j\right)=\exp\left(-\dfrac{\left\|x_i-x_j\right\|^2}{2\sigma^2}\right)$, σ is the width of the function.

2.4. Imbalance Data Management

Any data set that shows an unequal distribution between its classes can be considered imbalanced data set problem. The main challenge in imbalance problem is that the small classes are often more useful, but standard classifiers tend to be weighed down by the huge classes and ignore the tiny ones. Although SVMs work effectively with balanced datasets, they provide sub-optimal models with imbalanced datasets [24, 25]. The main reason for the SVM algorithm to be sensitive to class imbalance would be that the soft

margin objective function [43-45] assigns the same cost (*i.e.*, C) for both positive and negative misclassifications in the penalty term [26].

In this paper, we have used a Different Error Costs (DEC) method to handle imbalance dataset problem of K-PTM sites prediction. The Different Error Costs (DEC) method is a cost-sensitive learning solution proposed in [24] to overcome this problem in SVMs. In DEC method, the SVM soft margin objective function is modified to assign two misclassification costs, such that C^+ is the misclassification cost for positive class examples, while C^- is the misclassification cost for negative class examples. In our work, the following equations give the cost for the positive and negative classes

$$C^+ = \frac{C * N}{2 * N_1},\ C^- = \frac{C * N}{2 * N_2} \tag{10}$$

where N is the total number of instances, N_1 is the number of instances for positive class, and N_2 is the number of negative class.

2.5. Experimental Setting

In statistical prediction, there are three commonly used methods to derive the metric values for a predictor, these are, the independent dataset test, subsampling (e.g., K-fold cross validation) test, and jackknife test [15, 46]. These methods are often used for testing the accuracy of a statistical prediction algorithm. However, among those three methods, the jackknife test is deemed the most objective because it can always yield a unique result for a given benchmark data set, as reported in a comprehensive review [36]. Although the jackknife test has been increasingly and widely adopted by investigators to examine the power of various prediction methods, it takes huge computational time for a larger dataset.

In this study, we have used K-fold cross validation (subsampling) method to save the computational time. As the information about the exact 5-way splits of dataset used in previous studies is not published [9], therefore, in order to validate the stability and the statistical significance of our results, we have repeated the 5-fold cross validation for 5 times (*i.e.* 25 runs in total). It can be mentioned here that in each 5-fold cross validation the given training samples are randomly partitioned into 5 mutually exclusive sets of approximately equal size and approximately equal class distribution. Finally, we have reported the average results of all metrics in this study.

2.6. Measuring Metrics

According to the description of iPTM-mLys dataset in [9], we have a total of 6394 samples, of which 3991 are labeled with "acetylation", 115 with "crotonylation", 127 with "methylation", 1169 with "succinylation" and 1,750 with "non-K-PTM". However, in the above samples, some have two or more labels. It should be noted that in this study, we have considered {acetylation, crotonylation, methylation, succinylation, ∅} as class label set for a protein. Here ∅ is used to denote non-K-PTM. Since we are dealing with a multi-label system [47], so the metrics for a multi-label system will be used in this work instead of the conventional metrics defined for single-label systems [48-50].

For measuring the predictive capability and reliability for this kind of classification, a set metrics are usually used in the literature which are define below [47]:

$$\text{Aiming} = \frac{1}{N}\sum_{k=1}^{N}\left(\frac{\|Y_k \cap Z_k\|}{\|Z_k\|}\right)$$

$$\text{Coverage} = \frac{1}{N}\sum_{k=1}^{N}\left(\frac{\|Y_k \cap Z_k\|}{\|Y_k\|}\right)$$

$$\text{Accuracy} = \frac{1}{N}\sum_{k=1}^{N}\left(\frac{\|Y_k \cap Z_k\|}{\|Y_k \cup Z_k\|}\right)$$

$$\text{Absolute-True} = \frac{1}{N}\sum_{k=1}^{N}\Delta\left(Y_k, Z_k\right)$$

$$\text{Absolute-False} = \frac{1}{N}\sum_{k=1}^{N}\left(\frac{\left\|Y_k \cup Z_k\right\| - \left\|Y_k \cap Z_k\right\|}{M}\right) \tag{11}$$

where N is the total number of the samples concerned, M the total number of labels in the system, $\cup$ and $\cap$ the symbols are for the "union" and "intersection" in the set theory, $\|\|$ means the operator acting on the set therein to count the number of its elements, Y_k denotes the subset that contains all the labels experiment-observed for the k-th sample, Z_k represents the subset that contains all the labels predicted for the kth sample, and

$$\Delta\left(Y_k, Z_k\right) = \begin{cases} 1, \text{if all labels in } Z_k \text{ are identical with those in } Y_k \\ 0, \text{otherwise} \end{cases}$$

All of these metrics defined in this section have been successfully applied to study several multi-label systems, such as those in which a protein may stay in two or more different subcellular locations [51], or a membrane protein may have two or more different types [52], or an antimicrobial peptide may have two or more different types [53].

3. RESULTS AND DISCUSSION

3.1. Model Selection and Working Procedure of the Proposed System

In order to generate highly performing SVM classifiers capable of dealing with real data an efficient model selection is required [54]. Grid-search technique has been used to find the best model for SVM in this work. In our experiments, grid-search technique selects the values of parameters considering highest performance which is measured using a metric and then time if more than one position in search space has the same performance.

In this study, four SVM classifiers, one for each dataset, have been used for predicting the acetylation, crotonylation, methylation and succinylation sites. The model selection of each SVM classifiers has been done separately as binary classifier using the corresponding benchmark dataset given in **Table 1**. In this work, we have used RBF kernel for all SVM classifiers.

For radial basis function (RBF) kernel, to find the parameter value C (penalty term for soft margin) and σ (sigma), we have considered the value from 2^{-8} to 2^{8} for C and from 2^{-8} to 2^{8} for sigma as our searching space. Herein, the value of C will be used to find the misclassification cost of C^{+} and C^{-} defined in Equation (10). Since the information about the exact 5-way splits of dataset used in previous studies is not published [9], We have performed 5 times complete run of 5-fold cross-validations and each time we have selected the best parameter of the classifier depending on the value of AUC (area under curve). It should be noted here that AUC (area under the curve) is an important metric for single-label PTM site prediction [8, 15] which will be calculated from ROC curve (receiver operating characteristic curve). Finally, five sets of C and sigma (**Supplementary Tables S1**) have been selected from 5 times complete run of 5-fold cross-validations for each SVM classifier which is dedicated to a specific types of training dataset (acetylation, crotonylation, methylation, or succinylation).

After getting the four trained binary SVM classifier with appropriate values of C and sigma (**Supplementary Tables S1**), a multi-label predictor, named mLysPTMpred, has been developed by combing output from these four SVM classifiers, as shown in **Figure 1**. As we have repeated the 5-fold cross validation for 5 times for our mLysPTMpred, we have got five sets of values for all metrics defined in section 2.5. Finally, we have averaged our results in order to ensure unbiased model selection.

However, in order to train the system for the web server, we have used that value of C and sigma which appears most of the times as best model in 5 times complete run of 5-fold cross validation in each dataset. Note that, a random selection of the value of C and sigma has also been performed from 5 set of C and sigma of each dataset where "most of the times" criteria fail to select C and sigma. In this way, the selected C and sigma for each type of dataset is given in **Table 2**.

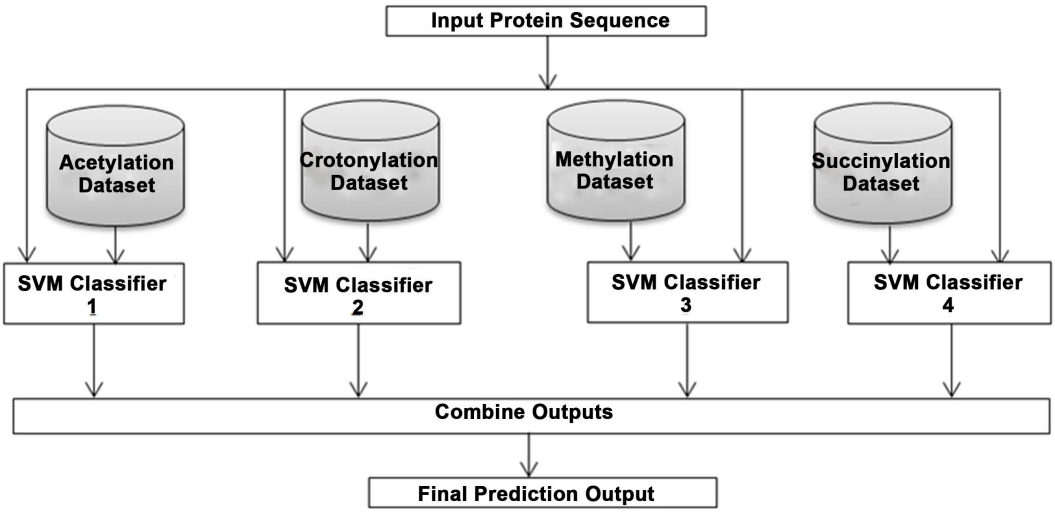

Figure 1. A flowchart to show how the mLysPTMpred predictor works.

Table 2. Selection of C and σ to train the system for web server.

Type of PTM	C	σ
acetylation	2^4	2^4
crotonylation	2^{-2}	2^7
methylation	2^0	2^4
succinylation	2^1	2^3

It can be mentioned here that all the trains and tests have been conducted on a standard machine of DELL Optiplex 390 with 8 GB RAM and Core-i3 processor running at 3.30 GHz. We have used Matlab 2014b version to implement our system where the *svmtrain* function of Matlab by default uses DEC with the same cost defined in Eq. (8) to handle imbalance situation.

3.2. Comparison with the Existing Methods

The values of the five metrics (cf. Equation (11)) obtained by the current mLysPTMpred predictor for multi-label lysine PTM site are given in the Table 3. These values are the average result of 5 times complete run of 5-fold cross-validation on the benchmark dataset given in Supporting Information. Moreover, standard deviations of each metrics of 5 times complete run of 5-fold cross validation are shown in parentheses.

Table 3 also includes the corresponding rates achieved by iPTM-mLys [9], the one existing predictors for identifying the multiple lysine PTM site in the aforesaid benchmark dataset. It should be mentioned here that the performance of iPTM-mLys [9] as shown in Table 3 are noted from [9].

In Equation (11), the first four metrics are completely opposite to the last one. For the former, the higher the rate is, the better the multi-label predictor's performance will be; for the latter, the lower the rate is, the better its performance will be [9]. The rate of "Absolute-False" or "Hamming-Loss" [47] for our predictor is 6.66% which is about half than iPTM-mLys. So, the average ratio of the completely wrong hits over the total prediction events for mLysPTMpred is significantly lower than iPTM-mLys.

Among the five metrics in Equation (11), the most strict and harsh one is the "Absolute-True". According to [9], very few multilabel predictors in biology could reach over 50% for the absolute true rate. However, the absolute-true rate achieved by mLysPTMpred can reach over 80% as shown in Table 3.

Also, among the same five metrics, the most important is the "Accuracy", the average ratio of the correctly predicted labels over the total labels including correctly and incorrectly predicted ones as well as

Table 3. A comparison of the proposed predictor with the existing methods on the same dataset.

Predictor	Aiming (%)	Coverage (%)	Accuracy (%)	Absolute-True (%)	Absolute-False (%)
iPTM-mLys	69.78	74.54	68.37	60.92	13.40
mLysPTMpred	84.82 (±0.0022)	86.56 (±0.0021)	83.73 (±0.0024)	79.73 (±0.0029)	6.66 (±0.00009)

those real labels but are missed out during the prediction. The mLysPTMpred achieves 83.73% accuracy which is considerable amount of higher than iPTM-mLys. In addition, the rate of "Aiming" or "Precision" [47] and the rate of "Coverage" or "Recall" [47] achieved by mLysPTMpred also better than iPTM-mLys.

Therefore, it is obvious from **Table 3**, mLysPTMpred has performed remarkably better over iPTM-mLys [9] in all types of metrics measurement. To provide an intuitive comparison, a bar chart to represent **Table 3** is shown **Figure 2**. Therefore, it is projected that mLysPTMpred may become a useful and higher throughput tool in multiple lysine PTM sites predictions.

In iPTM-mLys, an example protein sequence (Q16778) has been used to validate their findings. It should be noted here that the example protein sequence (Q16778) is also available in our site under Example button. For making comparison, we have not changed the sequence as example. The prediction result using this sequence (Q16778) from mLysPTMpred and the actual experimental result of this sequence is reported in **Table 4**. By putting this predicted result in equation 11, we have got the rate of aiming = 88.33%, coverage = 87.50%, accuracy = 85.83%, absolute-true = 80.00% and absolute-false = 6.00% which is similar to the rates obtained by the cross-validation tests as given in **Table 3**.

Why can the proposed method enhance the prediction quality so significantly? First, the coupling effects among the amino acids around the target sites have been taken into account via the conditional probability as done by many investigators in successfully enhancing the prediction quality in some applications [27, 31, 35, 42]. Second, the predictor used Different Error Costs (DEC) method to balance the effect of skewed training dataset and hence many false prediction events produced by imbalanced and skewed training datasets can be avoided as established in some recent studies [7, 8, 15, 31, 35].

3.3. Web Server

To attract more users especially for the convenience of experimental scientists and enhance the value of practical application, a user-friendly web-server for mLysPTMpred has been established at http://research.ru.ac.bd/mLysPTMpred/. In order to get the predicted result, users are required to submit protein sequence through the input text box in our site. The input sequence should follow the FASTA format. An example of a sequence of FASTA format is available under example button in our published site. Moreover, in order to get batch prediction, users are required to enter desired batch input file in the FASTA format. Noted that, the benchmark dataset used to train and test the mLysPTMpred predictor are available under Supporting Information button.

4. CONCLUSIONS

In this article, we have designed a simple and efficient predictor mLysPTMpred for predicting multiple lysine PTM sites. Experimental results show that our method is very promising and can be a useful tool for prediction of multiple lysine PTM site. The mLysPTMpred has achieved remarkably higher success rates in comparison with the existing predictors (iPTM-mLys) in this area. We believe that the approach and formulations proposed in this article for multi-label K-PTM can be used to study other multi-label PTM systems such as C-PTM, R-PTM and S-PTM for the corresponding multi-label PTM sites at Cys, Arg and Ser residues, respectively.

For convenience of the experimental scientists, we have established a user-friendly web server and a step by step guide has been provided about how to use this web server. It provides an easier way to obtain the desired results without knowing the mathematical details. We have projected that the mLysPTMpred

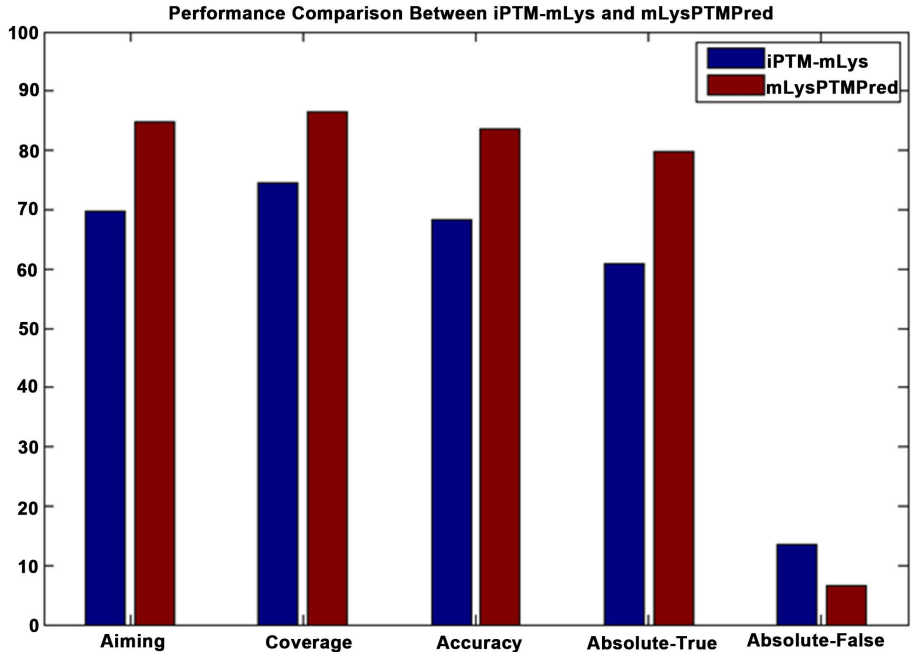

Figure 2. Performance Comparison Between iPTM-mLys and mLysPTMpred.

Table 4. Comparison between the predicted and experimental results on protein Q16778.

Sites	Predicted Result				Experimental Result			
	Ace	Cro	Met	Suc	Ace	Cro	Met	Suc
6	Yes	Yes	Yes	No	Yes	Yes	No	No
12	Yes	Yes	No	No	Yes	Yes	No	No
13	Yes	Yes	No	No	Yes	Yes	No	No
16	Yes	Yes	No	No	Yes	Yes	No	No
17	Yes	Yes	No	No	Yes	Yes	No	No
21	Yes	Yes	No	No	Yes	Yes	No	No
24	Yes	Yes	No	No	Yes	Yes	No	No
25	No	No	No	No	No	No	No	No
28	No	No	No	No	No	No	No	No
29	No	No	No	No	No	No	No	No
31	No	No	No	No	No	No	No	No
35	No	Yes	No	No	No	Yes	No	No
44	No	No	No	No	No	No	No	No
47	No	No	Yes	No	No	No	Yes	No
58	No	No	Yes	No	No	No	Yes	No
86	Yes	No	Yes	No	Yes	No	Yes	No
109	No	No	Yes	No	No	No	Yes	No
117	Yes	No	No	No	No	No	No	No
121	Yes	No	No	No	No	No	No	No
126	Yes	No	No	No	No	No	No	No

will become a very useful and higher throughput tool to deal with both single- and multi-label PTM systems.

As the current mLysPTMpred has been developed to study the multi-label system for only four different K-PTM types, we will try to add more types of K-PTM and include more new sequences [55] in future in our system and an announcement about the new version will be provided at the website http://research.ru.ac.bd/mLysPTMpred/.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest regarding the publication of this paper.

REFERENCES

1. Xu, Y., Ding, J., Wu, L.Y. and Chou, K.C. (2013) iSNO-PseAAC: Predict Cysteine S-Nitrosylation Sites in Proteins by Incorporating Position Specific Amino Acid Propensity into Pseudo Amino Acid Composition. *PLoS ONE*, **8**, e55844. https://doi.org/10.1371/journal.pone.0055844
2. Walsh, C.T., Garneau-Tsodikova, S. and Gatto, G.J. (2005) Protein Posttranslational Modifications: The Chemistry of Proteome Diversifications. *Angewandte Chemie International Edition*, **44**, 7342-7372. https://doi.org/10.1002/anie.200501023
3. Witze, E.S., Old, W.M., Resing, K.A. and Ahn, N.G. (2007) Mapping Protein Post-Translational Modifications with Mass Spectrometry. *Nature Methods*, **4**, 798-806. https://doi.org/10.1038/nmeth1100
4. Xu, Y., Wang, Z., Li, C. and Chou, K.C. (2017) iPreny-PseAAC: Identify C-Terminal Cysteine Prenylation Sites in Proteins by Incorporating Two Tiers of Sequence Couplings into PseAAC. *Medicinal Chemistry*, **13**, 544-551. https://doi.org/10.2174/1573406413666170419150052
5. Mann, M. and Jensen, O.N. (2003) Proteomic Analysis of Post-Translational Modifications. *Nature Biotechnology*, **21**, 255-261. https://doi.org/10.1038/nbt0303-255
6. Xu, Y., Wang, X., Wang, Y., Tian, Y., Shao, X., Wu, L.Y. and Deng, N. (2014) Prediction of Posttranslational Modification Sites from Amino Acid Sequences with Kernel Methods. *Journal of Theoretical Biology*, **344**, 78-87. https://doi.org/10.1016/j.jtbi.2013.11.012
7. Zhang, Z., Tan, M., Xie, Z., Dai, L., Chen, Y. and Zhao, Y. (2011) Identification of Lysine Succinylation as a New Post-Translational Modification. *Nature Chemical Biology*, **7**, 58-63. https://doi.org/10.1038/nchembio.495
8. Jia, J., Liu, Z., Xiao, X., Liu, B. and Chou, K.C. (2016) iSuc-PseOpt: Identifying Lysine Succinylation Sites in Proteins by Incorporating Sequence-Coupling Effects into Pseudo Components and Optimizing Imbalanced Training Dataset. *Analytical Biochemistry*, **497**, 48-56. https://doi.org/10.1016/j.ab.2015.12.009
9. Qiu, W.R., Sun, B.Q., Xiao, X., Xu, Z.C. and Chou, K.C. (2016) iPTM-mLys: Identifying Multiple Lysine PTM Sites and Their Different Types. *Bioinformatics*, **32**, 3116-3123. https://doi.org/10.1093/bioinformatics/btw380
10. Deng, W., Wang, Y., Ma, L., Zhang, Y., Ullah, S. and Xue, Y. (2016) Computational Prediction of Methylation Types of Covalently Modified Lysine and Arginine Residues in Proteins. *Briefings in Bioinformatics*, **18**, 647-658. https://doi.org/10.1093/bib/bbw041
11. Hasan, M.M., Yang, S., Zhou, Y. and Mollah, M.N.H. (2016) SuccinSite: A Computational Tool for the Prediction of Protein Succinylation Sites by Exploiting the Amino Acid Patterns and Properties. *Molecular BioSystems*, **12**, 786-795. https://doi.org/10.1039/C5MB00853K
12. Xu, Y., Ding, Y.X., Ding, J., Wu, L.Y. and Xue, Y. (2016) Mal-Lys: Prediction of Lysine Malonylation Sites in Proteins Integrated Sequence-Based Features with mRMR Feature Selection. *Scientific Reports*, **6**, 38318. https://doi.org/10.1038/srep38318

13. Jiang, M. and Cao, J.Z. (2016) Positive-Unlabeled Learning for Pupylation Sites Prediction. *BioMed Research International*, **2016**, Article ID 4525786. https://doi.org/10.1155/2016/4525786

14. Wuyun, Q., Zheng, W., Zhang, Y.P., Ruan, J.S. and Hu, G. (2016) Improved Species-Specific Lysine Acetylation Site Prediction Based on a Large Variety of Features Set. *PLoS ONE*, **11**, e0155370. https://doi.org/10.1371/journal.pone.0155370

15. Jia, J., Liu, Z., Xiao, X., Liu, B. and Chou, K.C. (2016) iCar-PseCp: Identify Carbonylation Sites in Proteins by Monto Carlo Sampling and Incorporating Sequence Coupled Effects into General PseAAC. *Oncotarget*, **7**, 34558-34570. https://doi.org/10.18632/oncotarget.9148

16. Qiu, W.R., Xiao, X., Lin, W.Z. and Chou, K.C. (2015) iUbiq-Lys: Prediction of Lysine Ubiquitination Sites in Proteins by Extracting Sequence Evolution Information via a Gray System Model. *Journal of Biomolecular Structure and Dynamics*, **33**, 731-1742. https://doi.org/10.1080/07391102.2014.968875

17. Xiao, X., Cheng, X., Su, S., Mao, Q. and Chou, K.C. (2017) pLoc-mGpos: Incorporate Key Gene Ontology Information into General PseAAC for Predicting Subcellular Localization of Gram-Positive Bacterial Proteins. *Natural Science*, **9**, 330. https://doi.org/10.4236/ns.2017.99032

18. Cheng, X., Zhao, S.G., Xiao, X. and Chou, K.C. (2017) iATC-mHyb: A Hybrid Multi-Label Classifier for Predicting the Classification of Anatomical Therapeutic Chemicals. *Oncotarget*, **8**, 58494-58503. https://doi.org/10.18632/oncotarget.17028

19. Jia, J., Liu, Z., Xiao, X., Liu, B. and Chou, K.C. (2016) iPPBS-Opt: A Sequence-Based Ensemble Classifier for Identifying Protein-Protein Binding Sites by Optimizing Imbalanced Training Datasets. *Molecules*, **21**, 95. https://doi.org/10.3390/molecules21010095

20. Xiao, X., Min, J.L., Lin, W.Z., Liu, Z., Cheng, X. and Chou, K.C. (2015) iDrug-Target: Predicting the Interactions between Drug Compounds and Target Proteins in Cellular Networking via Benchmark Dataset Optimization Approach. *Journal of Biomolecular Structure and Dynamics*, **33**, 2221-2233. https://doi.org/10.1080/07391102.2014.998710

21. Liu, Z., Xiao, X., Qiu, W.R. and Chou, K.C. (2015) iDNA-Methyl: Identifying DNA Methylation Sites via Pseudo Trinucleotide Composition. *Analytical Biochemistry*, **474**, 69-77. https://doi.org/10.1016/j.ab.2014.12.009

22. Sun, Y., Wong, A.K. and Kamel, M.S. (2009) Classification of Imbalanced Data: A Review. *International Journal of Pattern Recognition and Artificial Intelligence*, **23**, 687-719. https://doi.org/10.1142/S0218001409007326

23. Nath, A. and Karthikeyan, S. (2016) Enhanced Prediction and Characterization of CDK Inhibitors Using Optimal Class Distribution. *Interdisciplinary Sciences*: *Computational Life Sciences*, **9**, 292-303.

24. Veropoulos, K., Campbell, C. and Cristianini, N. (1999) Controlling the Sensitivity of Support Vector Machines. *Proceedings of the International Joint Conference on Artificial Intelligence*, Stockholm, 31 July-6 August 1999, 55-60.

25. Hasan, M.A.M., Li, J., Ahmad, S. and Molla, M.K.I. (2017) predCar-Site: Carbonylation Sites Prediction in Proteins Using Support Vector Machine with Resolving Data Imbalanced Issue. *Analytical Biochemistry*, **525**, 107-113. https://doi.org/10.1016/j.ab.2017.03.008

26. Batuwita, R. and Palade, V. (2010) Efficient Resampling Methods for Training Support Vector Machines with Imbalanced Datasets. *The* 2010 *International Joint Conference on Neural Networks*, Barcelona, 1-8.

27. Chou, K.C. (1993) A Vectorized Sequence-Coupling Model for Predicting HIV Protease Cleavage Sites in Proteins. *Journal of Biological Chemistry*, **268**, 16938-16948.

28. Hasan, M.A.M., Ahmad, S. and Molla, M.K.I. (2017) iMulti-HumPhos: A Multi-Label Classifier for Identifying Human Phosphorylated Proteins Using Multiple Kernel Learning Based Support Vector Machine. *Molecular BioSystems*, **13**, 1608-1618. https://doi.org/10.1039/C7MB00180K

29. Ju, Z. and He, J.J. (2017) Prediction of Lysine Propionylation Sites Using Biased SVM and Incorporating Four Different Sequence Features into Chou's PseAAC. *Journal of Molecular Graphics and Modelling*, **76**, 356-363. https://doi.org/10.1016/j.jmgm.2017.07.022

30. Chen, P., Hu, S., Zhang, J., Gao, X., Li, J., Xia, J. and Wang, B. (2016) A Sequence-Based Dynamic Ensemble Learning System for Protein Ligand-Binding Site Prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **13**, 901-912. https://doi.org/10.1109/TCBB.2015.2505286

31. Qiu, W.R., Zheng, Q.S., Sun, B.Q. and Xiao, X. (2016) Multi-iPPseEvo: A Multi-Label Classifier for Identifying Human Phosphorylated Proteins by Incorporating Evolutionary Information into Chou's General PseAAC via Grey System Theory. *Molecular Informatics*, **36**, Article ID: 1600085.

32. Wang, X., Yan, R., Li, J. and Song, J. (2016) SOHPRED: A New Bioinformatics Tool for the Characterization and Prediction of Human S-Sulfenylation Sites. *Molecular BioSystems*, **12**, 2849-2858. https://doi.org/10.1039/C6MB00314A

33. Hu, J., Li, Y., Yang, J.Y., Shen, H.B. and Yu, D.J. (2016) GPCR-Drug Interactions Prediction Using Random Forest with Drug-Association-Matrix-Based Post-Processing Procedure. *Computational Biology and Chemistry*, **60**, 59-71. https://doi.org/10.1016/j.compbiolchem.2015.11.007

34. Hu, J., Han, K., Li, Y., Yang, J.Y., Shen, H.B. and Yu, D.J. (2016) TargetCrys: Protein Crystallization Prediction by Fusing Multi-View Features with Two-Layered SVM. *Amino Acids*, **48**, 2533-2547. https://doi.org/10.1007/s00726-016-2274-4

35. Jia, J., Zhang, L., Liu, Z., Xiao, X. and Chou, K.C. (2016) pSumo-CD: Predicting Sumoylation Sites in Proteins with Covariance Discriminant Algorithm by Incorporating Sequence-Coupled Effects into General PseAAC. *Bioinformatics*, **32**, 3133-3141. https://doi.org/10.1093/bioinformatics/btw387

36. Chou, K.C. (2011) Some Remarks on Protein Attribute Prediction and Pseudo Amino Acid Composition. *Journal of Theoretical Biology*, **273**, 236-247. https://doi.org/10.1016/j.jtbi.2010.12.024

37. Chou, K.C. (2004) Using Amphiphilic Pseudo Amino Acid Composition to Predict Enzyme Subfamily Classes. *Bioinformatics*, **21**, 10-19. https://doi.org/10.1093/bioinformatics/bth466

38. Behbahani, M., Mohabatkar, H. and Nosrati, M. (2016) Analysis and Comparison of Lignin Peroxidases between Fungi and Bacteria Using Three Different Modes of Chou's General Pseudo Amino Acid Composition. *Journal of Theoretical Biology*, **411**, 1-5. https://doi.org/10.1016/j.jtbi.2016.09.001

39. Meher, P.K., Sahu, T.K., Saini, V. and Rao, A.R. (2017) Predicting Antimicrobial Peptides with Improved Accuracy by Incorporating the Compositional, Physico-Chemical and Structural Features into Chou's General PseAAC. *Scientific Reports*, **7**, Article No. 42362. https://doi.org/10.1038/srep42362

40. Liu, B., Liu, F., Wang, X., Chen, J., Fang, L. and Chou, K.C. (2015) Pse-in-One: A Web Server for Generating Various Modes of Pseudo Components of DNA, RNA, and Protein Sequences. *Nucleic Acids Research*, **43**, W65-W71. https://doi.org/10.1093/nar/gkv458

41. Liu, B. and Wu, H. (2017) Pse-in-One 2.0: An Improved Package of Web Servers for Generating Various m Odes of Pseudo Components of DNA, RNA, and Protein Sequences. *Natural Science*, **9**, 67-91. https://doi.org/10.4236/ns.2017.94007

42. Chou, K.C. (1996) Prediction of Human Immunodeficiency Virus Protease Cleavage Sites in Proteins. *Analytical Biochemistry*, **233**, 1-14. https://doi.org/10.1006/abio.1996.0001

43. Vapnik, V.N. (1999) The Nature of Statistical Learning Theory. Second Edition, Springer, New York.

44. Scholkopf, B. and Smola, A.J. (2001) Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge.

45. Hasan, M.A.M., Ahmad, S. and Molla, M.K.I. (2017) Protein Subcellular Localization Prediction Using Multiple

Kernel Learning Based Support Vector Machine. *Molecular BioSystems*, **13**, 785-795. https://doi.org/10.1039/C6MB00860G

46. Ju, Z., Cao, J.Z. and Gu, H. (2016) Predicting Lysine Phosphoglycerylation with Fuzzy SVM by Incorporating k-Spaced Amino Acid Pairs into Chou's General PseAAC. *Journal of Theoretical Biology*, **397**, 145-150. https://doi.org/10.1016/j.jtbi.2016.02.020
47. Chou, K.C. (2013) Some Remarks on Predicting Multi-Label Attributes in Molecular Biosystems. *Molecular Biosystems*, **9**, 1092-1100. https://doi.org/10.1039/c3mb25555g
48. Xu, Y., Ding, Y.X., Deng, N.Y. and Liu, L.M. (2016) Prediction of Sumoylation Sites in Proteins Using Linear Discriminant Analysis. *Gene*, **576**, 99-104. https://doi.org/10.1016/j.gene.2015.09.072
49. Liu, B., Liu, Y., Jin, X., Wang, X. and Liu, B. (2016) iRSpot-DACC: A Computational Predictor for Recombination Hot/Cold Spots Identification Based on Dinucleotide-Based Auto-Cross Covariance. *Scientific Reports*, **6**, Article No. 33483. https://doi.org/10.1038/srep33483
50. Liao, Z., Ju, Y. and Zou, Q. (2016) Prediction of G Protein-Coupled Receptors with SVM-Prot Features and Random Forest. *Scientifica*, **2016**, Article ID: 8309253. https://doi.org/10.1155/2016/8309253
51. Lin, W.Z., Fang, J.A., Xiao, X. and Chou, K.C. (2013) iLoc-Animal: A Multi-Label Learning Classifier for Predicting Subcellular Localization of Animal Proteins. *Molecular BioSystems*, **9**, 634-644. https://doi.org/10.1039/c3mb25466f
52. Huang, C. and Yuan, J.Q. (2013) A Multilabel Model Based on Chou's Pseudo-Amino Acid Composition for Identifying Membrane Proteins with both Single and Multiple Functional Types. *The Journal of Membrane Biology*, **246**, 327-334. https://doi.org/10.1007/s00232-013-9536-9
53. Xiao, X., Wang, P., Lin, W.Z., Jia, J.H. and Chou, K.C. (2013) iAMP-2L: A Two-Level Multi-Label Classifier for Identifying Antimicrobial Peptides and Their Functional Types. *Analytical Biochemistry*, **436**, 168-177. https://doi.org/10.1016/j.ab.2013.01.019
54. Hasan, M.A.M., Ahmad, S. and Molla, M.K.I. (2017) Protein Subcellular Localization Prediction Using Support Vector Machine with the Choice of Proper Kernel. *BioTechnologia*, **98**, 85-96. https://doi.org/10.5114/bta.2017.68307
55. Xu, H., Zhou, J., Lin, S., Deng, W., Ying, Z. and Yu, X. (2017) PLMD: An Updated Data Resource of Protein Lysine Modifications. *Journal of Genetics & Genomics*, **44**, 243-250. https://doi.org/10.1016/j.jgg.2017.03.007

Table S1. Selected C and σ of 5 times run of 5 folds cross-validations for RBF kernel.

No. of Completes Run	Type of PTM							
	Acetylation		Crotonylation		Methylation		Succinylation	
	C	σ	C	σ	C	σ	C	σ
1st	2^4	2^4	2^{-1}	2^4	2^0	2^4	2^1	2^3
2nd	2^4	2^4	2^0	2^6	2^4	2^4	2^1	2^3
3rd	2^1	2^1	2^2	2^7	2^0	2^4	2^5	2^4
4th	2^1	2^1	2^{-2}	2^5	2^4	2^4	2^1	2^3
5th	2^4	2^4	2^{-2}	2^7	2^5	2^4	2^5	2^4