predSucc-Site: Lysine Succinylation Sites Prediction in Proteins by using Support Vector Machine and Resolving Data Imbalance Issue

Md. Al Mehedi Hasan
Department of Computer Science and Engineering
Rajshahi University of Engineering and Technology, Bangladesh

Shamim Ahmad
Department of Computer Science and Engineering
University of Rajshahi, Bangladesh

ABSTRACT
The lysine succinylation is found as an important post-translational modification where succinyle group is added to a lysine (K) residue of a protein molecule. It plays major role not only in regulating the cellular processes but also associated with some diseases. As a result, it requires an easiest way to detect succinylation modification in proteins. However, since the experimental technologies are costly and time-consuming, so it is quite hard to detect the succinylation modification timely at low cost to face the explosive growth of protein sequences in postgenomic age. In this context, an accurate computational method for predicting succinylation sites is an urgent issue which can be useful for drug development. In this study, a novel computational tool termed predSucc-Site has been developed to predict protein succinylation 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 Costs (DEC) method, and (3) constructing a predictor using support vector machine as classifier. The experimental result shows that the predSucc-Site predictor achieves an average AUC (area under curve) score of 0.97 in predicting lysine succinylation sites. All of the experimental results along with AUC of our system are found from the average of 5 complete runs of the 5-fold cross-validation and those results indicate significantly better performance of predSucc-Site than existing predictors. A user-friendly web server for the predSucc-Site is available at http://research.ru.ac.bd/predSucc-Site/

Keywords
Lysine Succinylation Sites Prediction, Sequence-coupling Model, General PseAAC, Data Imbalance Issue, Support Vector Machine

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]. Not only that, PTMs are also responsible for expanding the genetic code and for regulating cellular physiology as well [2, 3]. In general, the side chain of lysine plays the key role in increasing the complexity of PTM network [4]. The lysine residue in proteins can experience many types of PTMs, such as methylation, acetylation, biotinylation, ubiquitination, ubiquitin-like modifications, propionylation, and butyrylation, which lead to the remarkable complexity of PTM networks [4, 5].

Succinylation is an emerging posttranslational modification where a succinyl group (-CO-CH2-CH2-CO-) is added to a lysine residue of a protein molecule [6] and it plays an potential role in regulating protein conformation, function and physicochemical properties [7]. As a result, the identification of lysine succinylation 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 [5, 8].

However, the purely experimental technique to determine the exact modified sites of succinylated substrates 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 succinylated sites effectively and accurately [7]. Therefore, recently various types of computational classifiers have been developed to identify succinylation sites through different types of machine learning algorithms [1, 4, 5, 6, 7, 8, 9, 10]. However, in order to meet the current demand to produce efficient high-throughput tools, additional effort are required to enrich the prediction quality [5, 8].

In the development of computational classifier, one of the major challenges is to handle imbalance dataset problem [5], as it is found in most of the dataset of succinylation sites prediction, the number negative subset is much larger than the corresponding positive subset [5]. As the real world picture is that the non-succinylation sites are always the majority compared with the succinylation ones, so naturally the predictor should be biased to the non-succinylation sites. Here the problem is that, for this type of predictors may interpret many succinylation sites as non-succinylation sites [11, 12, 13]. But, the information about the succinylation sites is mostly desired than non-succinylation sites. As a result, it is crucial to find an effective solution to balance this kind of bias consequence.

The current study was begun with an attempt to address the problems mentioned above and then tried to develop a more powerful predictor using support vector machine, called ‘predSucc-Site’. In this predictor, the Different Error Costs (DEC) method [14, 15, 16] has been used to resolve the data imbalance issue. It should be noted here that the features used in that predictor are extracted by using vectorized sequence-coupling model [17]. In the recent works, the performance of iSuc-PseAAC [18], SucPred[7], pSuc-Lys [8], and iSuc-PseOpt [5] on a large set of proteins has been studied in [5, 8]. Therefore, in order to compare the performance of predSucc-Site with those systems (iSuc-PseAAC [18], SucPred[7], pSuc-Lys [8], iSuc-PseOpt [5]), we use the exactly same dataset employing the commonly used stratified 5-fold cross-validation [8]. Since the information about the exact 5-way splits used in previous studies [8] is not available, so we have performed 5 complete runs of the 5-fold-crossvalidation, where each complete run of 5-fold cross-

validation uses a different 5-way splits. 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 has been reported. Our experimental results indicate that predSucc-Site achieves significantly better results than those found from other top systems (iSuc-PseAAC [18], SucPred [7], pSuc-Lys [8], iSuc-PseOpt [5]).

In order to launch a useful sequence-based statistical predictor for a biological system, the Chou's five-step rules should be followed [19, 5, 8] (i) construct or select a valid benchmark dataset to train and test the predictor, (ii) formulate the biological sequence samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted, (iii) introduce or develop a powerful algorithm (or engine) to operate the prediction, (iv) properly perform cross-validation tests to objectively evaluate its anticipated accuracy, and (v) establish a user-friendly webserver that is accessible to the public.

2. MATERIAL AND METHODS

2.1 Benchmark Dataset

pSuc-Lys's benchmark dataset set has been used in this study which was derived from the CPLM, a protein lysine modification database [8, 20]. CPLM contains 2521 lysine succinylation sites and 24128 non-succinylation sites determined from 896 proteins [20]. In pSuc-Lys, all of the corresponding protein sequences were derived from the UniProt [21] database. In pSuc-Lys's work, $(2\xi + 1)$-tuple peptide window was used to collect peptide segment that had K at the center from these 896 proteins. 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 its mirrorimage in pSuc-Lys [8]. After applying some screening procedure based on some constraints on that dataset, for example, considering window size, <= 40% pairwise sequence identity to any other peptides, pSuc-Lys finally extracted a filtered training dataset. Detail description of screening procedure is explained in [8].

The final dataset of pSuc-Lys consisted of 1167 lysine succinylation sites and 3553 non-succinylation sites. It can be noted here that sliding window method was used to encode every lysine residue K of that dataset since succinylation only occurred in lysine residues K. According to [5, 8], window size has been selected as 31 (2*ξ+1) in our study, where ξ=15. The detailed sequences for the 1167 samples in the positive subset ($S^+_{\xi=15}$) and those for the 3553 samples in the negative subset ($S^-_{\xi=15}$) are available at online supplementary materials (http://research.ru.ac.bd/predSucc-Site/). Thus, the benchmark dataset set S for the current study can be formulated as

$$S_{\xi=15} = S^+_{\xi=15} \cup S^-_{\xi=15} \qquad (1)$$

2.2 Feature Extraction

The appropriate features of protein sequences or samples plays very important roles for the prediction of succinylation site, as a result it draws the much attention of scientist that how to select the core and essential features of protein samples. Moreover, 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. In order to avoid complete losing the sequence pattern information for protein, in this paper, the Chou's general PseAAC [19, 22] has been adopted to extract feature from peptide segment using sequence-coupling model [17, 23] which has been described briefly below.

According to Chou's scheme, a peptide with lysine (K) located at its center can be generally expressed by

$$P_\xi(\mathbb{K}) = R_{-\xi}R_{-(\xi-1)} \dots \dots R_{-2}R_{-1}\mathbb{K}\,R_1R_2 \dots \dots . R_{+(\xi-1)}R_{+\xi} \qquad (2)$$

where the center K represents "lysine", 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 $P_\xi(\mathbb{K})$ can be further classified into the following two categories:

$$P_\xi(\mathbb{K}) \in \begin{cases} P^+_\xi(\mathbb{K}), & \text{if its center is a succinylation site} \\ P^-_\xi(\mathbb{K}), & \text{otherwise} \end{cases} \qquad (3)$$

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

It is obvious from Eq. (2) that when ξ=15 the corresponding peptide contains $(2\xi + 1) = 31$ amino acid residues; that is, it can be reduced to

$$P_\xi(\mathbb{K}) = R_1R_2 \dots \dots R_{14}R_{15}\mathbb{K}\,R_{16}R_{17} \dots \dots . R_{29}R_{30} \qquad (4)$$

Thus, according to the general form of PseAAC [19], the samples in the positive subset $S^+_{\xi=15}$ and negative subset $S^-_{\xi=15}$ of Eq. (1) can be respectively formulated as

$$P^+ = [\theta^+_1\theta^+_2 \dots \theta^+_u \dots \theta^+_\pi]^T \qquad (5)$$

and

$$P^- = [\theta^-_1\theta^-_2 \dots \theta^-_u \dots \theta^-_\pi]^T \qquad (6)$$

where T is the transpose operator and π is an integer to reflect the dimension of the PseAAC vector. The value of π, as well as the components θ^+_u and θ^-_u $(u = 1, 2, \dots, \dots, \pi)$ therein, will depend on how to extract the desired information from the peptide samples in Eq. (4). In this study, to make P^+ better reflect the intrinsic correlation with the lysine succinylation sites, the components in Eq. (5) are defined by the following sequence-coupling factors via the conditional probability approach as originally proposed in Refs. [17, 23] for predicting the HIV protease cleavage sites in proteins:

$$\theta^+_u = \begin{cases} p^+(R_1|R_2), & if\ u = 1 \\ p^+(R_2|R_3), & if\ u = 2 \\ \vdots & \vdots \\ p^+(R_{15}), & if\ u = 15 \\ p^+(R_{16}), & if\ u = 16 \\ \vdots & \vdots \\ p^+(R_{29}|R_{28}), & if\ u = 29 \\ p^+(R_{30}|R_{29}), & if\ u = 30 \end{cases} \quad \pi = 30 \qquad (7)$$

where $p^+(R_1|R_2)$ is the conditional probability of amino acid R_1 occurring at the first position given that its right neighbor in the peptide sample (cf. Eq. (4)) is R_2, $p^+(R_2|R_3)$ is the conditional probability of amino acid R_2 occurring at the second position given that its right neighbor is R_3, and so forth. Note that in the above equation only $p^+(R_{15})$ and $p^+(R_{16})$ are of nonconditional probability given that the left neighbor of R_{15} and the rightneighbor of R_{16} are always K. All of these probability values can be easily derived from the

positive benchmark dataset given in Supporting Information S1 in the supplementary material as done in Ref. [23]. Similarly, the components in Eq. (6) are defined by

$$\theta_u^- = \begin{cases} p^-(R_1|R_2), & if\ u = 1 \\ p^-(R_1|R_2), & if\ u = 2 \\ \vdots & \vdots \\ p^-(R_{15}), & if\ u = 15 \\ p^-(R_{16}), & if\ u = 16 \\ \vdots & \vdots \\ p^-(R_{29}|R_{28}), & if\ u = 29 \\ p^-(R_{30}|R_{29}), & if\ u = 30 \end{cases} \quad \pi = 30 \qquad (8)$$

where the probability values are derived from the corresponding negative benchmark dataset as given in Supporting Information S2.

Inspired by the concept of discriminant function that has been successfully used by many previous investigators to predict the specificity of GalNAc transferase [24], cysteine S-nitrosylation sites [25], hydroxyproline and hydroxylysine [26], tight turns and their types [27], and nitrotyrosine sites [28], here we use the discriminant PseAAC vector to represent a peptide sample; that is, P of Eq. (4) is finally formulated as a 30-D (30-dimensional) vector given by

$$P = [\theta_1 \theta_2 \ldots \theta_u \ldots \theta_{30}]^T \qquad (9)$$

where

$$\theta_u = (\theta_u^+ - \theta_u^-), \quad u = 1,2,\ldots,\ldots,30 \qquad (10)$$

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 [29, 30, 31]:

$$\text{maximize}_\alpha \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i\, y_j k(x_i, x_j)$$

Subject to:

$$\sum_{i=1}^{n} y_i \alpha_i = 0\ ,\ 0 \le \alpha_i \le C$$

$$\text{for all } i = 1,2,3,\ldots\ldots,n \qquad (11)$$

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

$$f(x) = \sum_i^n \alpha_i y_i k(x, x_i) + b \qquad (12)$$

It noted here that a kernel function and its parameter have to be chosen to build a SVM classifier [29, 30, 31]. In this work, radial basis function kernel has been used to build SVM classifier which is defined below:

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

2.3.1 Imbalance Data Management for SVM

Any dataset that shows an unequal distribution between its classes can be considered imbalanced dataset problem. Although SVMs work effectively with balanced datasets, they provide sub-optimal models with imbalanced datasets [14, 15, 16]. In this paper, we have used a Different Error Costs (DEC) method to handle imbalance dataset problem for succinylation sites prediction. The Different Error Costs (DEC) method is a cost-sensitive learning solution proposed in [14] to overcome imbalance dataset 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. The following equations give the cost for the positive and negative classes

$$C^+ = \frac{C*N}{2*N_1}, \qquad C^- = \frac{C*N}{2*N_2} \qquad (13)$$

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.

2.4 Experimental Setting

In statistical prediction, there are three commonly used methods to derive the metric values for a predictor: the independent dataset test, subsampling (e.g., k-fold cross-validation) test, and jackknife test [8, 32]. In this study, we have used cross validation methods to save the computational time. As the information about the exact 5-way splits of dataset used in previous studies is not published [8], 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. 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.

2.5 Measuring Metrics

For measuring the success rates for this kind of classification, a set of four metrics is usually used in the literature: (i) overall accuracy or Acc, (ii) Mathew's correlation coefficient or MCC, (iii) sensitivity or Sn, and (iv) specificity or Sp [18, 33, 34].

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

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

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \qquad (14)$$

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

where TP (true positive) denotes the number of succinylated peptides correctly predicted, TN (true negative) the numbers non-succinylated peptides correctly predicted, FP (false positive) the non-succinylated incorrectly predicted as the succinylated peptides, and FN (false negative) the succinylated peptides incorrectly predicted as the non-succinylated peptides.

AUC (area under the curve) is also another indicator in practical application which will be calculated from ROC curve (receiver operating characteristic curve). It is instructive to point out that the metrics as defined in Eq. (14) are valid for single-label systems, for multi-label systems a set of more complicated metrics should be used [18].

3. RESULTS AND DISCUSSION

3.1 Model Selection for SVM

In order to generate highly performing SVM classifiers capable of dealing with real data an efficient model selection is required. In our experiment, grid-search technique has been used to find the best model for SVM. For radial basis function (RBF) kernel based SVM, 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 (8). The selected C and sigma of 5 complete runs of the 5-fold cross-validation on the benchmark data set is shown in Table 1. Finally, we have averaged our results in order to ensure unbiased model selection. It should be mentioned that we have used $C=2^{-3}$ and $\sigma = 2^{3}$ to trained the system for the web server, because most of the times, the best model is found for the value of $C=2^{-3}$ and $\sigma = 2^{3}$.

Table 1. Selected C and σ of 5 complete runs of 5-fold cross-validation for RBF kernel based SVM.

No. of Completes Run	**C**	σ
1st	2^{-2}	2^{3}
2nd	2^{-3}	2^{3}
3rd	2^{-3}	2^{3}
4th	2^{-3}	2^{3}
5th	2^{-2}	2^{3}

3.2 Comparison with the Existing Methods

The values of the four metrics (cf. Eq. (14)) obtained by the current predSucc-Site predictor are given in the Table 2. These values are the average result of 5 complete runs of 5-fold cross-validation on the benchmark dataset given in Supporting information S1 and Supporting information S2. Moreover, standard deviations of each metrics of 5 complete runs of the 5-fold cross-validation are shown in parentheses.

The Table 2 also includes the corresponding rates achieved by iSuc-PseAAC [18], SucPred[7], pSuc-Lys [8], and iSuc-PseOpt [5], the four existing predictors for identifying the lysine succinylation sites for the aforesaid dataset. It should be mentioned here that the performance of iSuc-PseAAC [18], SucPred[7], pSuc-Lys [8], iSuc-PseOpt [5] as shown in Table 2 are noted from [5, 8].

It is obvious from the Table 2, predSucc-Site has performed remarkably better over iSuc-PseAAC, SucPred, pSuc-Lys and iSuc-PseOpt while considering Acc, MCC, and Sn. It indicates that, the proposed new predictor has produced over all better accuracy, sensitivity, and stability. Although the achieved Sp by SucPred, pSuc-Lys and iSuc-PseOpt is higher than that by our predictor, the gap between its Sn and Sp is very large (48% for SucPred, 19% for pSuc-Lys, 27% for iSuc-PseOpt). Which implies that the results achieved by SucPred, pSuc-Lys and iSuc-PseOpt contain many false negative events [35] and hence its higher achieved Sp rate is problematic [8]. Since the information about the succinylation sites is mostly desired than non-succinylation sites [5, 8] from the biological point of view, predSucc-Site will be more preferable than such type of problematic predictors [8]. Moreover, it can be noted here that, the programs such as BLAST and FASTA have been widely used in genomic and proteomic analysis or prediction based on similarity search [36]. But, unfortunately these programs are helpless while facing sequences having low similarities, especially, when more and more orphan genes are discovered. Therefore, in the case of orphan genes or low-similar proteins, it is urgent to develop a statistical predictive model. From this biological viewpoint, this study is very meaningful and important [36].

Table 2. A comparison of the proposed predictor with the existing methods

Method	Acc(%)	MCC	Sn(%)	Sp(%)	AUC
iSuc-PseAAC	79.98	0.4370	50.63	89.68	0.7823
SucPred	85.32	0.5710	49.13	97.17	0.8933
pSuc-Lys	90.83	0.7695	76.79	95.97	0.9325
iSuc-PseOpt**	87.86	0.7193	69.38	96.86	0.9475
predSucc-Site	92.00 (±0.07)	0.8029 (±0.0023)	93.42 (±0.35)	91.47 (±0.04)	0.9788 (±0.0001)

** Taken the best result of iSuc-PseOpt from Table 1 of [5]

However, in order to make a more consistent comparison of our predSucc-Site with SucPred, pSuc-Lys and iSuc-PseOpt in the case of specificity metric, two thresholds of specificity 95% and 97% have been taken into consideration and values of other metrics of predSucc-Site at those level of specificities have been calculated too. In this case, to find out the performance of predSucc-Site, a single time run of the 5-fold cross-validation by considering $C=2^{-3}$ and $\sigma = 2^{3}$ has been used. It can be noted here that class discrimination value of SVM classifier have been found -0.26 and -0.48 for 95% and 97% specificity respectively. All of the findings of predSucc-Site by considering these two thresholds of specificity are shown in Table 3. From Table 2 and 3, it is clear that predSucc-Site has also produced better results than other predictors in those two thresholds levels of specificity.

Table 3. Values of Acc, MCC, and Sn of predSucc-Site for Two different levels of Specificity

Sp (%)	Acc (%)	MCC	Sn (%)
95.16	93.54	0.8285	88.60
97.02	93.39	0.8186	82.35

The area under the ROC curve is called AUC (area under the curve). The greater the AUC value is, the better the predictor will be [37, 38]. As we can see from Table 2, the value of AUC clearly indicates that the proposed predictor is better than iSuc-PseAAC [18], SucPred[7], pSuc-Lys [8], and iSuc-PseOpt [5]. Therefore, it is projected that predSucc-Site may become a useful and higher throughput tool in succinylation sites prediction.

3.3 Protocol Guide

To attract more users especially for the convenience of experimental scientists [39, 40] and enhance the value of practical application, a user-friendly web-server for mLysPTMpred has been established at http://research.ru.ac.bd/predSucc-Site/. 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 predSucc-Site predictor are available under Supporting Information button.

4. CONCLUSION

In this article, we have designed a simple and efficient predictor predSucc-Site for predicting succinylation sites. Experimental results show that our method is very promising and can be a useful tool for prediction of succinylation sites. The predSucc-Site has achieved remarkably higher success rates in comparison with the existing predictors in this area. In addition to it, we have established a user-friendly web server and provided a step-by-step guide for convenience of the experimental scientists. It provides as easier way to obtain the desired results without knowing the mathematical details. We have projected that the predSucc-Site will become a very efficient and higher throughput tool for predicting of protein succinylation sites.

5. REFERENCES

[1]. Xu, Y., Ding, J., Wu, L. Y., 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(2), e55844.

[2]. Walsh, C. T., Garneau-Tsodikova, S., Gatto, G. J., 2005. Protein posttranslational modifications: the chemistry of proteome diversifications. Angewandte Chemie International Edition, 44(45), 7342-7372.

[3]. Witze, E. S., Old, W. M., Resing, K. A., & Ahn, N. G. (2007). Mapping protein post-translational modifications with mass spectrometry. Nature Methods, 4(10), 798-806.

[4]. Zhang, Z., Tan, M., Xie, Z., Dai, L., Chen, Y., Zhao, Y., 2011. Identification of lysine succinylation as a new post-translational modification. Nature Chemical Biology, 7(1), 58-63.

[5]. Jia, J., Liu, Z., Xiao, X., Liu, B., 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.

[6]. Xie, Z., Dai, J., Dai, L., Tan, M., Cheng, Z., Wu, Y., Zhao, Y., 2012. Lysine succinylation and lysine malonylation in histones. Molecular & Cellular Proteomics, 11(5), 100-107.

[7]. Zhao, X., Ning, Q., Chai, H., Ma, Z., 2015. Accurate in silico identification of protein succinylation sites using an iterative semi-supervised learning technique. Journal of Theoretical Biology, 374, 60-65.

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

[9]. Hasan, M. M., Yang, S., Zhou, Y., 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(3), 786-795.

[10]. Xu, H. D., Shi, S. P., Wen, P. P., Qiu, J. D., 2015. SuccFind: a novel succinylation sites online prediction tool via enhanced characteristic strategy. Bioinformatics, 31(23), 3748-3750.

[11]. Liu, Z., Xiao, X., Qiu, W. R., Chou, K. C., 2015. iDNA-Methyl: identifying DNA methylation sites via pseudo trinucleotide composition. Analytical Biochemistry, 474, 69-77.

[12]. Sun, Y., Wong, A. K., Kamel, M. S., 2009. Classification of imbalanced data: a review. International Journal of Pattern Recognition and Artificial Intelligence, 23(04), 687-719.

[13]. Xiao, X., Min, J. L., Lin, W. Z., Liu, Z., Cheng, X., 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(10), 2221-2233.

[14]. Veropoulos, K., Campbell, C., Cristianini, N., 1999. Controlling the sensitivity of support vector machines. In Proceedings of the International Joint Conference on Artificial Intelligence, pp. 55-60.

[15]. Hasan, M. A. M., Li, J., Ahmad, S., 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.

[16]. Hasan, M. A. M., Ahmad, S., 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.

[17]. Chou, K. C., 1993. A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins. Journal of Biological Chemistry, 268(23), 16938-16948.

[18]. Xu, Y., Ding, Y. X., Ding, J., Lei, Y. H., Wu, L. Y., Deng, N. Y., 2015. iSuc-PseAAC: predicting lysine succinylation in proteins by incorporating peptide position-specific propensity. Scientific Reports, 5.

[19]. Chou, K. C., 2011. Some remarks on protein attribute prediction and pseudo amino acid composition. Journal of Theoretical Biology, 273(1), 236-247.

[20]. Liu, Z., Wang, Y., Gao, T., Pan, Z., Cheng, H., Yang, Q., ..., Xue, Y., 2014. CPLM: a database of protein lysine modifications. Nucleic Acids Research, 42(D1), D531-D536.

[21]. UniProt Consortium., 2010. The universal protein resource (UniProt) in 2010. Nucleic acids research, 38(suppl 1), D142-D148.

[22]. Chou, K.C., 2005. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21,10–19..

[23]. Chou, K. C., 1996. Prediction of human immunodeficiency virus protease cleavage sites in proteins. Analytical Biochemistry, 233(1), 1-14.

[24]. Chou, K. C., 1995. A sequence-coupled vector-projection model for predicting the specificity of GalNAc-transferase. Protein Science, 4(7), 1365-1383.

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

[26]. Xu, Y., Wen, X., Shao, X. J., Deng, N. Y., Chou, K. C., 2014. iHyd-PseAAC: Predicting hydroxyproline and hydroxylysine in proteins by incorporating dipeptide position-specific propensity into pseudo amino acid composition. International Journal of Molecular Sciences, 15(5), 7594-7610.

[27]. Chou, K. C., 2000. Prediction of tight turns and their types in proteins. Analytical Biochemistry, 286(1), 1-16.

[28]. Xu, Y., Wen, X., Wen, L. S., Wu, L. Y., Deng, N. Y., Chou, K. C., 2014. iNitro-Tyr: Prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition. PloS one, 9(8), e105018.

[29]. Scholkopf, B., Smola, A. J., 2001. Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press.

[30]. Hasan, M. A. M., Ahmad, S., Molla, M. K. I., 2017. Protein subcellular localization prediction using multiple kernel learning based support vector machine. Molecular BioSystems, 13(4), 785-795.

[31]. Hasan, M. A. M., Ahmad, S., Molla, M. K. I., 2017. Protein Subcellular Localization Prediction using Support Vector Machine with the Choice of Proper Kernel", BioTechnologia vol. 98(2), 85-96.

[32]. Chou, K. C., Shen, H. B., 2007. Recent progress in protein subcellular location prediction. Analytical Biochemistry, 370(1), 1-16.

[33]. Xue, Y., Zhou, F., Fu, C., Xu, Y., Yao, X., 2006. SUMOsp: a web server for sumoylation site prediction. Nucleic Acids Research, 34(suppl 2), W254-W257.

[34]. Chen, Y. Z., Chen, Z., Gong, Y. A., Ying, G., 2012. SUMOhydro: a novel method for the prediction of sumoylation sites based on hydrophobic properties. PLoS One, 7(6), e39195.

[35]. Chen, J., Liu, H., Yang, J., Chou, K. C., 2007. Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. Amino Acids, 33(3), 423-428.

[36]. Tang, H., Zou, P., Zhang, C., Chen, R., Chen, W., Lin, H., 2016. Identification of apolipoprotein using feature selection technique. Scientific Reports, 6.

[37]. Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861-874.

[38]. Davis, J., Goadrich, M., 2006. The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd international conference on Machine learning (pp. 233-240). ACM.

[39]. Chen, W., Lin, H., Chou, K. C., 2015. Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences. Molecular BioSystems, 11(10), 2620-2634.

[40]. Chou, K. C., 2015. Impacts of bioinformatics to medicinal chemistry. Medicinal Chemistry, 11(3), 218-234.