International Journal of Statistical Sciences *ISSN 1683-5603*
Vol. 19, 2020, pp 125-136

Statistical Prediction of Lysine PTM Sites Mapping on *Homo sapiens* under the CKSAAP Encoding

Samme Amena Tasmia[1], Fee Faysal Ahmed[2], Md. Parvez Mosharaf[1], Md. Hadiul Kabir[1] and Md. Nurul Haque Mollah[1]*

[1]Bioinformatics Lab., Dept. of Statistics, Rajshahi University, Rajshahi-6205, Bangladesh

[2]Dept. of Mathematics, Jessore University of Science and Technology, Jessore, Bangladesh

[*]Correspondence should be addressed to Md. Nurul Haque Mollah (mollah.stat.bio@ru.ac.bd)

[Received Jan. 5, 2020; Revised January 20, 2020; Accepted February 10, 2020]

Abstract

Post-translational modification (PTM) refers to the covalent and enzymatic modification of proteins which plays a key role in protein conformation regulation and cellular function control. Identification of Lysine PTM sites can facilitate our understanding about the molecular mechanism accurately. However, traditional experimental approaches of PTM site predictions are labor-intensive and time-consuming. In this context, a more accurate computational method for predicting Succinylation sites is an urgent issue which can be useful for drug development. In this work, we developed a novel Succinylation site predictor called Succinpred, which is constructed including CD-HIT (removing 40% identity), 5-fold cross-validation, CKSAAP encoding, 1:2 ratio of the positive vs. negative samples, and AdaBoost. The performance of this method was measured with an accuracy of 87.5%, a MCC (Matthew Correlation Coefficient) of 79.8% and AUC (Area under the ROC Curve) of 0.883 using 5-fold cross validation on training dataset and an accuracy of 76.8%, a MCC of 65.6% and AUC of 0.833 on independent dataset. The proposed predictor also performed much better than the other existing predictors.

Keywords: Succinylation site, Feature Extraction, Sequence Analysis, CKSAAP and Machine Learning.

AMS Classification: 92B20.

1. Introduction

Post-translational modifications (PTMs) significantly dominate the structural and functional heterogeneity of proteins as well as malleability and dynamics of living cells (Xu et al., 2013). Not only that, PTMs are also accountable for enlarging the genetic code as well as for regulating cellular physiology (Wash et al., 2005; Witze et al., 2007).

Lysine (K) residue of a protein molecule is also known as succinylation site. Therefore, due to the advancement technologies on sequence analysis, the computational identification scheme of succinylation is needed before experimental verification. Until now, a few of computational methods focused on predicting the Succinylation. Despite all these efforts, the succinylation site prediction is still not accurate enough and more efficient algorithms are still desirable. Up to date, prediction of succinylation substrates have established in terms of bioinformatics implementations. Zhao et al. (2015) proposed a predictor SucPredwhich is based on Support Vector Machine (SVM) including four types of encoding methods (i.e. grouped weight-based encoding, auto-correlation functions, normalized vander Waals volume and position amino acids weight composition) were used. Another SVM-based predictor iSuc-PseAAC developed by Xu et al. (2015) which adopts the pseudo amino acid composition encoding scheme to improving the prediction performance. Xu et al. (2015) developed another predictor SuccFind based on SVM including amino acid composition (AAC), an amino acid index (AAindex) physicochemical properties and k-space amino acid pair composition (CKSAAP). Jea et al. (2016) developed two predictors (i.e. iSuc-PseOpt and pSuc-Lys) by using random forest (RF) classifiers with the general pseudo amino acid composition encoding. Lopez et al. (2017) developed a predictor SucStruct which is structure-based predictor using a decision tree classifier. Hasanet. al. (2016 & 2017) established two predictors (i.e. SuccinSite and SuccinSite2.0) including amino acid frequency and properties with combined RF classifier scores. The SuccinSite2.0 predictor builds up with seven species-specific and their generic model classifiers. This predictor used combination of two sequence features information including profile-based composition of k-spaced amino acid pairs (CKSAAP) and binary amino acid codes (BE) with a RF classifier. In this study, a novel predictor Succinpred has been developed for accurate identification of succinylation by the integrated of

complementary features. The Succinpred predictor achieved an AUC score of 0.833 on a cross-validation set and outperforms other existing methodologies on a comprehensive independent dataset. The Succinpred is suggested to be a helpful computational resource for biomedical researchers.

2. Materials and methods

To construct Succinpred site predictor we have collected 550 amino acid sequences fromUni-PortKB/Swiss-Port and were given in the published article (Hasan et. al., 2016). In this dataset, 1407 Succinylation sites in 550 proteins were experimentally detected.

Data Preparation

Experimentally identified succinylation data for *Homo sapiens* dataset was collected from uniportdatabase.We used CD-HIT with a 40% identity cutoff to remove the redundant sequences. Experimentally identified succinylated lysine residues were considered as succinylated sites (i.e., positive samples). On the other hand, the remaining lysine residues that have not been identified as positive samples (i.e. succinylated sites) in these proteins were considered as, non-succinylated sites (i.e. negative samples). Each site was defined as a peptide segment of 2w+1 length with lysine (K) in the center. Randomly selected non-succinylated sites were declared as negative samples based on an intimate assumption.

To construct a Succinpred predictor, the training and independent dataset was compiled using the AdaBoost, K-Nearest Neighbor and Naïve Bayes classifiers.A total of 550 proteins with 1407 positive samples (i.e. succinylated sites) and 18205 negative samples (i.e. non-succinylated sites) were obtained as an independent dataset in this study.

2.1 Sequence encoding strategy of CKSAAP

The composition of k spaced amino acid pairs (CKSAAP) encoding method was firstly introduced by Chen et at. (2007). It has been broadly using in the numerous bioinformatics work. If window size of a fragment is 2w+ 1 (where w = 1, 2, 3,…) and 21 types of amino acids (including the gap (O)), which may create (21× 21) = 441 types of amino acid pairs (i.e. AA, AC, AD,. . ., OO) for every single k (k denotes the space between two amino acids). For the optimal kmax =5, there are

21× (kmax + 1) × 21 = 2646 different amino acid pairs are created for each sequence. Then the feature vectors are calculated using the following equation:

$$\left(\frac{N_{AA}}{N_{total}},\frac{N_{AC}}{N_{total}},\frac{N_{AD}}{N_{total}},\ \cdots,\frac{N_{OO}}{N_{total}}\right)_{441} \quad (1)$$

where N_{total} is the length of the total composition residues (for example, if the fragment length w is 25 and k = 0, 1, 2, 3, 4, 5 then N_{total} = w - k -1 will be 28, 27, 26, 25, 24 and 23, respectively). N_{AA},N_{AC},...,N_{OO} are frequency of the amino acid pair within the fragment. More details are available somewhere.

2.2 Classification assessment

To build a better predictor for protein Succinylationsite prediction, we considered three popular classifiers (AdaBoost (ADA), Naïve Bayes (NB) & K-nearest Neighbor (KNN),) for a comparison based on the encoded protein sequences. For accessibility of the readers, let us get together those classifiers as follows:

ADA: The ADA classifier is a machine-learning algorithm which was reorganized using the R package 'fastAdaboost' as adaboost(formula, data(train), iter=10, nu=1, type = "discrete"). By fitting data to a logistic curve, logistic regression is used for identification of the probability of occurrence of an event. It is commonly used for predictor variables which are either numerical or categorical.

NB: Naïve Bayes is a predictive algorithm based on the statistical learning theory of Bayesian theorem. NB is a probabilistic classifier established on relating Bayes'theorem with individuality assumptions. The Naive Bayes classifier was implemented using the R package 'naiveBayes'asnaiveBayes(formula, data(train), laplace = 0, ..., subset, na.action = na.pass).

KNN: In the KNN, several distance measures such as the Mahalanobis distance, Euclidean distance and the Hamming distance, are used to measures the nearest neighbors. i.e. if an instance x_I is x_i= $[x_i^1, \ldots, x_i^n]$, Where x_i^T denote the value of the r-th feature of instance x_i ,then the distance between two instance x_i and $x_j$is

$$D(x_i, x_j)=\sqrt{\sum_{r=1}^{n}(x_i^r - x_j^r)^2} \quad (2)$$

In this study, the KNN algorithm was administered using the R package 'knn3' as knn3(formula = formula(train), train, test, na.action = na.omit(), k = 50, distance =

2, kernel ="rectangular", ykernel = NULL, scale=TRUE, contrasts = c('unordered' = "contr.dummy", ordered = "contr.ordinal")).

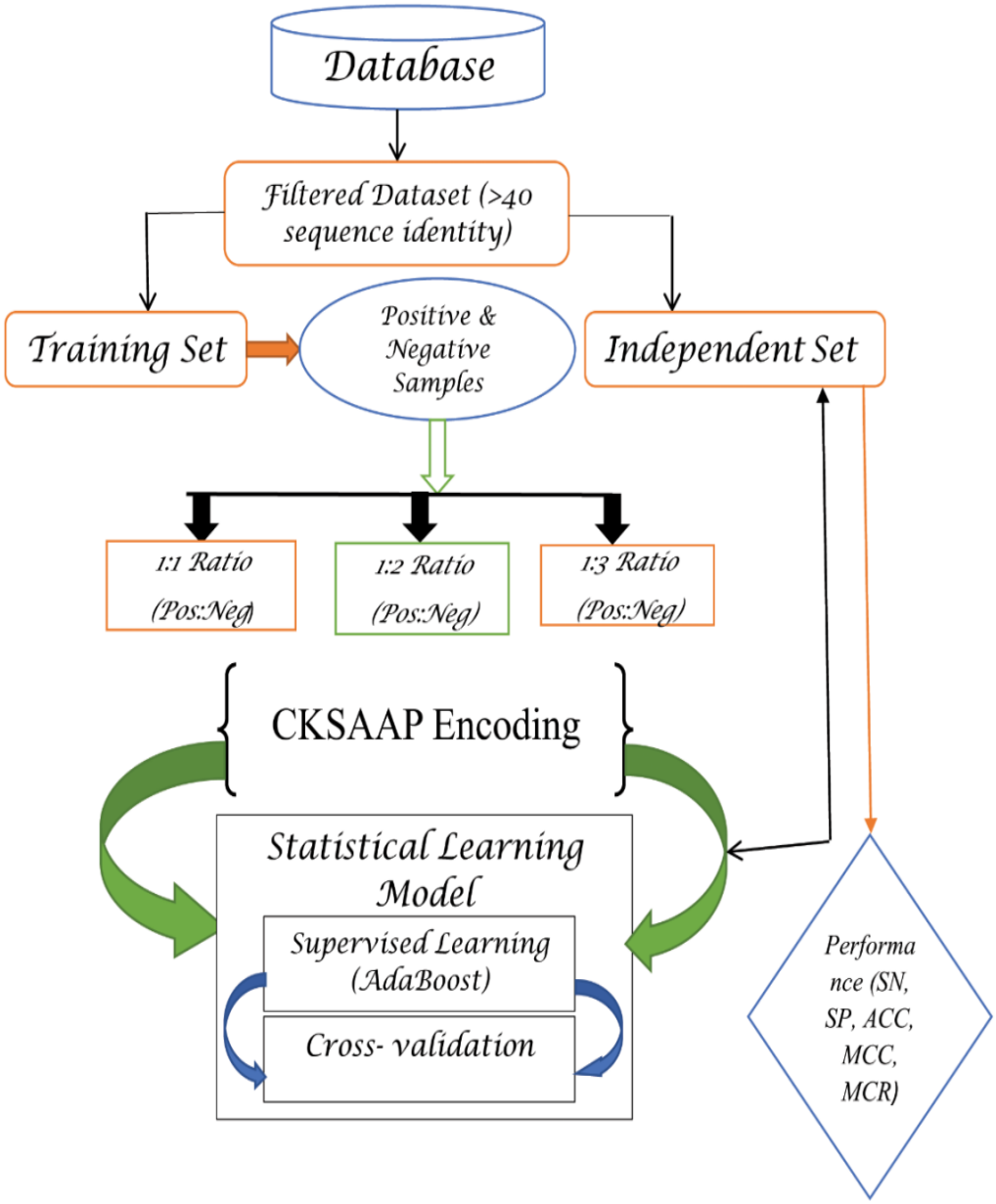

Figure 1: Overview of the study

2.3 Performance Assessment

In this study, six performance measurements have used, Sensitivity (SN), Specificity (SP), False Negative Rate (FNR), Accuracy (ACC), Misclassification Rate (MCR), and Mathew Correlation Coefficient (MCC). They are defined as below:

$$SN = \frac{n(TP)}{n(TP)+n(FN)}; 0 \leq SN \leq 1 \quad (3)$$

$$SP = \frac{n(TN)}{n(TN)+n(FP)}\ ;\ 0 \leq SP \leq 1 \quad (4)$$

$$FNR = \frac{n(FN)}{n(TP)+n(FN)};\ 0 \leq FNR \leq 1 (5)$$

$$ACC = \frac{n(TP)+n(TN)}{n(TP)+n(FP)+n(TN)+n(FN)}; 0 \leq ACC \leq 1 \quad (6)$$

$$MCR = \frac{n(FP)+n(FN)}{n(TP)+n(FP)+n(TN)+n(FN)}; 0 \leq MCR \leq 1 \quad (7)$$

$$MCC = \frac{(n(TP) \times n(TN)) - (n(FP) \times n(FN))}{\sqrt{\big(n(TP)+n(FN)\big) \times \big(n(TN)+n(FP)\big) \times \big(n(TP)+n(FP)\big) \times \big(n(TN)+n(FN)\big)}};$$
$$-1 \leq MCC \leq 1 \quad (8)$$

Where n(TP) represents the number of correctly identified positive windows, n(TN) defined as the number of correctly identifiednegative windows, n(FP) represents the number of incorrectly identified positive windows and n(FN) defined as the number of incorrectly predicted negative windows. The higher value of all of these measurements represents a better prediction and the values lie between 0 to 1. In addition, we also showed the receiver operating characteristics (ROC) curve (Sensitivity vs. 1-Specificity plot) and area under the ROC curve (AUC) which calculates the performance of the proposed prediction.

3. Results and Discussion

3.1 Performance with Training Dataset

The final feature vectors were trained by the ADA, KNN and NB classifiers. The consecutive model parameters were optimized via training dataset based on the 5-

fold cross-validation. The proposed method provided the highest AUC values of 0.883 on the training dataset. The performances of training dataset for different ratios are given in Table 1, Table 2 and Table 3.

Table 1: The performances of different classifiers based on ratio 1:1 using training dataset at FPR=0.1

Classifiers	SN	SP	FNR	ACC	MCC	MCR	AUC	pAUC
ADA	0.768	0.901	0.232	0.789	0.724	0.211	0.854	0.109
KNN	0.428	0.901	0.585	0.623	0.265	0.377	0.696	0.054
NB	0.512	0.901	0.488	0.724	0.453	0.275	0.743	0.065

Table 2: The performances of different classifiers based on ratio 1:2 using training dataset at FPR=0.1

Classifiers	SN	SP	FNR	ACC	MCC	MCR	AUC	pAUC
ADA	**0.823**	**0.901**	**0.177**	**0.875**	**0.798**	**0.125**	**0.883**	**0.112**
KNN	0.509	0.901	0.491	0.652	0.318	0.351	0.732	0.058
NB	0.498	0.901	0.502	0.716	0.437	0.284	0.734	0.057

Table 3: The performances of different classifiers based on ratio 1:3 using training dataset at FPR=0.1

Classifiers	SN	SP	FNR	ACC	MCC	MCR	AUC	pAUC
ADA	0.725	0.901	0.275	0.756	0.687	0.244	0.816	0.102
KNN	0.499	0.901	0.501	0.641	0.293	0.362	0.727	0.051
NB	0.445	0.901	0.555	0.687	0.381	0.313	0.701	0.045

SN: Sensitivity; SP: Specificity; FNR: False Negetive Rate; ACC: Accuracy; MCC: Matthew's Correlation coefficient; AUC: Area under the ROC Curve; pAUC: partial AUC. ADA: AdaBoost; KNN: K-Nearest Neighbor; NB: Naïve Bayes.

3.2 Performance with Independent Test Dataset

In order to evaluate the practical prediction ability of the final prediction model, a large independent data set was constructed. Figure2 contains the results of independent data set and represents the average prediction results for 5- fold CV given the highest performance than other three classifiers using 1:2 shuffled protein sequences as negative data set. Compared with other three ratio of dataset, the model based on the negative data set 1:2 ratio gives the best performance. The

average prediction of AUC is 0.833 respectively, which indicate that this method is successful in predicting Succinylation sites in human.

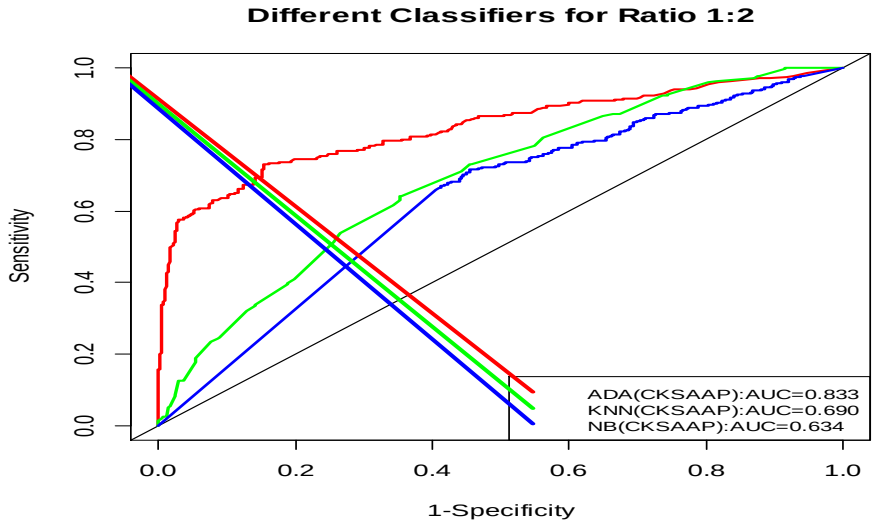

Figure2: Comparison between different classifiers with 1:2 ratio based on independent dataset

3.3 Proposed prediction model

Best performance was observed after combining the multiple encoding methods including CKSAAP encoding, AdaBoost, ratio 1:2, and window size 25 using 5-fold cross validation to represent the succinylation. The performance indexes of the proposed predictor in terms of specificity, sensitivity, accuracy, and MCC were summarized in the training and independent datasets (Table 4).

Table 4: The performances of proposed method using training and independent samples

Predictors	Training	Independent
Sensitivity	0.823	0.657
Specificity	0.901	0.901
ACC	0.875	0.768
MCC	0.798	0.656
AUC	0.883	0.833
pAUC(at 0.1 FPR)	0.112	0.105

3.4 Comparison of the proposed predictor with existing predictors

To compare the performance of proposed method (Succinpred) with existing predictors (SucPred, iSuc-PseAAC, SuccFind, iSuc-PseOpt, pSuc-Lys, SuccinSite, SuccinSite2.0, and GPSuc).we compared with SucPred, iSuc-PseAAC, SuccFind, iSuc-PseOpt, pSuc-Lys, SuccinSite, SuccinSite2.0, and GPSucto test the prediction was accepted here. In Table 2, we can see that the SN, SP, ACC, & MCC are achieved by the proposed method which is significantly improved than other predictors. So, we can undoubtedly suggest that our proposed method provides the most accurate predictions of succinylation sites than any other existing predictors.

Table 5: The performance comparison with existing methods on the independent set

Prediction methods	SP	SN	ACC	MCC
SucPred	0.673	0.272	0.643	-0.030
iSuc-PseAAC	0.887	0.122	0.827	0.013
SuccFind	0.792	0.252	0.750	0.029
iSuc-PseOpt	0.758	0.303	0.722	0.038
pSuc-Lys	0.826	0.224	0.779	0.036
SuccinSite	0.882	0.371	0.842	0.199
Succinsite2.0	0.884	0.457	0.850	0.263
GPSuc	0.883	0.499	0.853	0.296
Proposed	**0.901**	**0.657**	**0.768**	**0.656**

Moreover, the performances indexes were found robust in the independent datasets. In addition, the proposed model Sccinpred greatly outperformed the others existing algorithms (Table 5). Indeed, all performance measures in the Succinpred were higher than those of the other methods, thus indicating the superiority of the Succinpred in succinylation prediction.

3.5 Sequence specificity of Succinylation site

By Two Sample Logos software, the amino acid predilection of neighboringsuccinylation sites related to the non- succinylation sites, as shown for the training dataset Figure 4. It shows that positive samples represented residues at each location or negative samples represented residues at each location were

plotted below and under the X-axis respectively. In the following calculation and operation, we selected 25-mer (-12, +12) window size and Fig. 4 shows the position-specific difference of amino acid compositions between Succinylation sites and non-Succinylation sites.

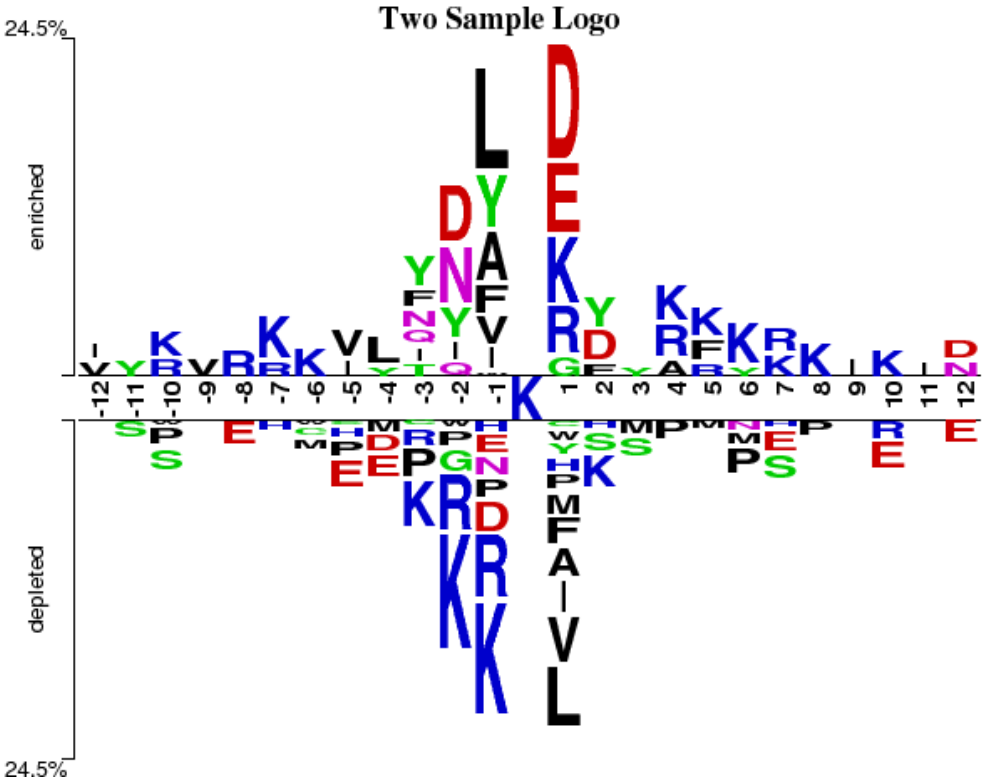

Figure4: The amino acid propensities of surrounding Succinylation sites compared to non-Succinylation sites, as displayed with the Two Sample Logos software (Vacie et al., 2006).

Limitations

The author used PTM sites datasets from the published databases. That was published in non-Asian country. If we generate *H. Sapiens* sequences from Bangladesh, our biological research will be more effective can improve different varieties of *H. Sapiens*.

4. Conclusions

Accurate identification of the succinylation sites prediction could hopefully decipher the molecular mechanisms of succinylation related biological processes. Through some researchers have focused on this problem, the overall accuracy of prediction is still not satisfactory. In this paper, different types of classifiers were used to find out the most important predictors and we make a comparing among the methods. Finally, we observed that, using CKSAAP encoding, AdaBoost classifier, 1:2 ratio for 5-fold cross validation perform better than the others.

Acknowledgements: The authors sincerely acknowledge Dr. Md. Nurul Haque Mollah (Professor of Statistics Department in Rajshahi University) of providing the necessary suggestions and facilities throughout the study.

References

[1] Chen, K., Kurgan, L. and Rahbari, M. (2007). Prediction of protein crystallization using collocation of amino acid pairs, Biochem. Biophys. Res. Commun. 007; 355: 764–769.

[2] Dehzangi, A., Lopez, Y., Lal, S. P., Taherzadeh, G., Michaelson, J. and Sattar, A. (2017). PSSM-Suc: Accurately predicting succinylation using position specific scoring matrix into bigram for feature extraction, Journal of theoretical biology, 425:97–102.

[3] Hasan, M. M., Yang, S., Zhou, Y. and Mollah, M. N. (2016). SuccinSite: a computational tool for the prediction of protein succinylation sites by exploiting the amino acid patterns and properties, Molecular bioSystems, 12(3):786–95.

[4] Hasan, M. M., Khatun, M. S., Mollah, M. N. H., Yong, C. and Guo, D. (2017). A systematic identification of species-specific protein succinylation sites using joint element features information, International journal of nanomedicine, 12:6303–15.

[5] 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, Anal Biochem, 497:48–56.

[6] Lopez, Y., Dehzangi, A., Lal, S. P., Taherzadeh, G., Michaelson, J. and Sattar, A. (2017). SucStruct: Prediction of succinylated lysine residues by using structural properties of amino acids, Analytical biochemistry, 527:24–32.

[7] Vacic, V., Iakoucheva, L. M., Radivojac, P. (2006).Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments, Bioinformatics, 22: 1536–1537.

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

[9] 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(10), 798-806.

[10] Wen, P. P, Shi, S. P., Xu, H. D., Wang, L. N. and Qiu, J. D. (2016). Accurate in silico prediction of species-specific methylation sites based on information gain feature optimization, Bioinformatics, 32(20):3107–15.

[11] 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(2), e55844.

[12] Xu, Y., Ding, Y. X., Ding, J., Lei, Y. H., Wu, L. Y. and Deng, N. Y. (2015). iSuc-PseAAC: predicting lysine succinylation in proteins by incorporating peptide position-specific propensity, Scientific reports, 5:10184.

[13] Xu, H. D., Shi, S. P., Wen, P. P. and Qiu, J. D. (2015). SuccFind: a novel succinylation sites online prediction tool via enhanced characteristic strategy, Bioinformatics, https://doi.org/10.1093/bioinformatics/btv439 PMID: 26261224.

[14] 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(1), 58-63.

[15] Zhao, X., Ning, Q., Chai, H. and Ma, Z. (2015). Accurate in silico identification of protein succinylation sites using an iterative semi-supervised learning technique, Theoretical biology, 374:60–5.