C-iSUMO: A sumoylation site predictor that incorporates intrinsic characteristics of amino acid sequences

C-iSUMO: A sumoylation site predictor that incorporates intrinsic characteristics of amino acid sequences

Yosvany López[a,*], Abdollah Dehzangi[b], Hamendra Manhar Reddy[c], Alok Sharma[c,d,e,*]

[a] Genesis Institute of Genetic Research, Genesis Healthcare Co., Tokyo, Japan

[b] Department of Computer Science, Morgan State University, Baltimore, Maryland, USA

[c] School of Engineering and Physics, University of the South Pacific, Suva, Fiji Islands

[d] Laboratory for Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences, Yokohama, Kanagawa, Japan

[e] Institute for Integrated and Intelligent Systems, Griffith University, Queensland, Australia

Email addresses:

YL: yosvany.lopez.alvarez@gmail.com

AS: alok.sharma@griffith.edu.au

Abstract

Post-translational modifications are considered important molecular interactions in protein science. One of these modifications is “sumoylation” whose computational detection has recently become a challenge. In this paper, we propose a new computational predictor which makes use

of the sine and cosine of backbone torsion angles and accessible surface area for predicting sumoylation sites. The aforementioned features were computed for all the proteins in our benchmark dataset, and a training matrix consisting of sumoylation and non-sumoylation sites was ultimately created. This training dataset was balanced by undersampling the majority class (non-sumoylation sites) using the NearMiss method. Finally, an AdaBoost classifier was used for discriminating between sumoylation and non-sumoylation sites. Our proposed predictor was called "C-iSumo" because of its effective use of circular functions. C-iSumo was compared with another predictor, which was outperformed in statistical metrics such as sensitivity (0.734), accuracy (0.746) and Matthews correlation coefficient (0.494).

Introduction

Once proteins are translated in the ribosome, they undergo a series of modifications commonly referred to as post-translational modifications (PTMs) [1]. These molecular marks modify the amino acids of a protein by covalently incorporating specific functional groups. A long list of PTMs, which differ from each other and play different roles at the cellular level, has been discovered. For instance, the amino acid lysine can be modified through methylation [2, 3], acetylation [4], ubiquitination and succinylation [5-9]. Each PTM contributes to the cellular complexity via complex post-translational networks. Recently, the detection of a new type of PTM coined sumoylation [10] has proven extremely difficult. Sumoylation is regarded a dynamic and reversible modification, which regulates a variety of cellular processes from cellular dynamics and plasticity to gene expression and genome stability maintenance[11, 12] [13, 14]. This PTM is produced by small ubiquitin-related modifiers, and its molecular structure is similar to that of ubiquitination. The detection of sumoylation sites is essential to drug development efforts due to their dramatic effect on protein binding. However, the detection of

sumoylation residues by experimental techniques is still considered expensive, inefficient and impractical. Therefore, the development of new computational approaches capable of accurately predicting sumoylation sites has turned out absolutely necessary. This has eventually prompted the proposal of extensive studies. For instance, a predictor called SUMOsp made use of a curated dataset along with two previously developed methods: GPS and MotifX, which had been initially designed for phosphorylation prediction [15]. Another approach called SUMOsp 2.0 utilized an improved group-based phosphorylation scoring algorithm for prediction purposes [16]. Likewise, a tool called GPS-SUMO was developed for predicting SUMO-interacting motifs by combining a group-based prediction algorithm and a particle swarm optimization approach [17]. All the above methods used a group-based phosphorylation scoring algorithm to tackle the sumoylation prediction challenge. Additionally, a statistical method called SUMOpre showed improvements in sumoylation prediction via the removal of homologs [18]. An approach called SUMOhydro considered amino acid hydrophobicity and a support vector machine trained on a non-redundant sumoylation set [19]. Another method named SUMO_LDA integrated three feature constructions, including AAIndex, position-specific amino acid propensity and the composition of *k*-space amino acid pairs with the general pseudo amino acid composition for predicting sumoylation sites [20]. Although some of the above methods have achieved good performance, they have mainly relied on sequence analysis for identification of sumoylation sites. For instance, SUMO_LDA made use of the pseudo amino acid composition for extracting evolutionary information. Moreover, SUMOsp and SUMOsp 2.0 used the putative motif ψ-K-X-E and evolutionary conservation information, as well as SUMO-interaction motifs for classification purposes. Because their use of protein sequence was limited to evolutionary

information and sequence motifs, the proposed predictor offers an alternative solution which includes structural information for sumoylation site identification.

In this work, we propose a new predictor called "C-iSumo", which takes into consideration the sine and cosine of four torsion angles (ϕ, Ψ, θ and τ) along with accessible surface area for accurately predicting sumoylation sites. We used a benchmark dataset consisting of 448 proteins whose sumoylation sites were experimentally detected and annotated. Each sumoylation and non-sumoylation residue was described by its 15 upstream and 15 downstream amino acids, and subsequently summarized in a training dataset. To ameliorate the imbalance between classes, we employed the NearMiss method for undersampling the majority class (non-sumoylation sites) [21]. For classification purposes, we designed an AdaBoost classifier which is considered one of the best ensemble methods. When compared with existing state-of-the-art predictors, C-iSumo showed a significant improvement in performance with 0.734 sensitivity, 0.746 accuracy and 0.494 Matthews correlation coefficient.

Materials and Methods

In this paper, we describe a new machine learning-based predictor, which was able to effectively discriminate between sumoylation and non-sumoylation sites. The proposed predictor employed two main characteristics of proteins, namely, backbone torsion angles and accessible surface area. The description of lysine residues and computed features, as well as the machine learning approach are presented in the following subsections. A flowchart of the methodology can be found in Figure 1.

Dataset

The benchmark dataset was retrieved from the Compendium of Protein Lysine Modifications [22, 23], which comprises around 45,000 proteins distributed across 122 species. To avoid

overestimations because of sequence homology, we retained those proteins <40% similar for further analysis. The final dataset was composed of 448 proteins with experimentally annotated sumoylation residues. Subsequently, each protein sequence was analyzed and its annotated lysines were assigned to either of two sets. As a result, the positive set contained 780 sumoylation sites whereas the negative set comprised 21,353 non-sumoylation sites.

Accessible Surface Area

Accessible surface area (ASA) is a characteristic that provides the estimated accessibility area of an amino acid to a solvent in the 3D configuration of a protein. Therefore, the prediction of the ASA for individual amino acids tends to reveal essential information about the protein structure. To calculate the ASA, each protein sequence was analyzed with the tool SPIDER2 [24, 25], which outputs one value for each amino acid in the protein. Of note, SPIDER2 uses evolutionary, physicochemical, and sequence features for training a deep learning model [24, 25]. This means that any structural features, predicted by SPIDER2, will implicitly incorporate sequence-based information. Because the proposed predictor used the output of SPIDER2 as structural features (backbone torsion angles and accessible surface area), it has indirectly considered sequence information. This strategy allowed us to benefit from sequence-based characteristics without necessarily increasing the number of features.

Backbone Torsion Angles

Backbone torsion angles between nearby amino acids reportedly provide valuable information on the local structure of amino acids, somehow complementing the ASA. For a given amino acid, the backbone torsion angles ϕ and Ψ are predicted as continuous representations of the interaction between local amino acids along the protein backbone. Recent studies have also considered two new angles: θ which is formed between $C\alpha$ atoms ($C\alpha_{i-1}$- $C\alpha_i$- $C\alpha_{i+1}$), and τ which

rotates around the $C\alpha_i$- $C\alpha_{i+1}$ bond [26]. We run SPIDER2 [24, 25] for each protein sequence and consequently obtained four different numerical vectors (ϕ, Ψ, θ, and τ). These vectors were finally converted into sine and cosine functions.

It is worth noting that the above structural information was only extracted from protein sequences and not from the actual 3D structure of proteins. As previously stated, we predicted the secondary (local) structure of proteins using the software SPIDER 2.0 which is a fully sequence-based model, and used the resulting structure for sumoylation site prediction. Therefore, the proposed model solely relies on protein sequences and no additional information was regarded.

Undersampling the Majority Class

Class imbalance is one of the problems commonly encountered in machine learning applications. To tackle it, two widely used strategies are oversampling and undersampling. Whereas oversampling the minority class often results in overfitted models, undersampling the majority class offers a practical solution for increasing the sensitivity of models. Though previous studies have introduced different approaches for balancing a training set [27], we used here a method called NearMiss. NearMiss removes samples from the majority class by considering average distance measures [21]. In this study, we used a version of NearMiss which chooses those instances of the majority class whose average distances to three closest instances of the minority class are the smallest. To do this, we employed the imbalanced-learn package of Python, which provides a wide range of methods for dealing with highly unbalanced datasets.

Describing the Lysine Residues

Each lysine residue was described in terms of its 15 upstream and 15 downstream amino acids (Fig. 2A). Previous studies have extensively investigated different window sizes [28-30],

however, all of them have consistently concluded that a 15-residue window provides relevant information about lysines. On the other hand, when a lysine residue was located near either terminus of the protein, and a gap of 15 amino acids was not possible to retrieve, we completed the feature vector by mirroring the entire side with missing amino acids (Fig. 2B).

Let us consider a peptide segment

$$S = \{R_{-15}, R_{-14}, \dots, R_{-2}, R_{-1}, L, R_1, R_2, \dots, R_{14}, R_{15}\} \quad (1)$$

which describes the lysine residue L and contains R_{-i} and R_i $(1 \leq i \leq 15)$ for upstream and downstream residues. Each lysine was thus represented by a peptide segment S composed of 31 residues. These amino acids were described by structural characteristics such as ASA, and the sine and cosine of the four torsion angles ϕ, Ψ, θ and τ. Each feature was represented by one numerical vector from which the segment S was extracted. As a result, each lysine was described as a 279-dimensional feature vector. It is worth noting that if the dimensionality of the feature space increases, feature selection or dimensionality reduction techniques should be used before applying any classifier. Although the length of the protein fragments could dramatically affect the predictor performance, we chose to use a 31-residue window for describing each lysine. This window size was chosen after previous analyses with different windows [6, 8, 9, 31], and has also been consistently supported by scientific publications [32-36].

AdaBoost

AdaBoost is an ensemble strategy consisting of multiple algorithms, which are combined by an adaptive boosting scheme. It combines the outputs of individual weak classifiers in order to produce a strong predictor [37-39]. Because individual classifier tends to poorly predict instances, this ensemble approach uses the training data to build several base models by

bootstrap sampling. Initially, one model is created and subsequently a second model, intended to correct the errors of the first model, is designed. This procedure continues until the training set is correctly predicted, or a specific number of models is reached. AdaBoost is mainly utilized to improve the performance of decision trees in binary classification problems, often encountered in the field of computational proteomics [40]. In this study, we used the scikit-learn library of Python and decision trees as base classifiers [41, 42].

Results and Discussion

C-iSumo includes two categories of structural features, namely, accessible surface area and backbone torsion angles. Both characteristics were used for describing each lysine residue and ultimately predicting those sumoylation sites. The following sections describe the comparison of the proposed method with an existing predictor in the literature.

Statistical Evaluation

When a new prediction approach is designed, it is extremely important to measure its performance. For this purpose, we considered four statistical metrics: sensitivity, specificity, accuracy, and Matthews correlation coefficient (MCC).

The above metrics are defined as follows,

$$sensitivity = \frac{PS_+}{PS_+ + PS_-} \tag{2}$$

$$specificity = \frac{NS_+}{NS_+ + NS_-} \tag{3}$$

$$accuracy = \frac{PS_+ + NS_+}{PS_+ + PS_- + NS_+ + NS_-} \tag{4}$$

$$MCC = \frac{(NS_+ \times PS_+) - (NS_- \times PS_-)}{\sqrt{(PS_+ + PS_-)(PS_+ + NS_-)(NS_+ + PS_-)(NS_+ + NS_-)}} \tag{5}$$

where PS_+ and PS_- represent the amount of sumoylation sites correctly and incorrectly classified, whereas NS_+ and NS_- indicate the amount of non-sumoylation sites correctly and incorrectly classified by the predictor.

Sensitivity and specificity evaluate the ability of a predictor to correctly detect sumoylation and non-sumoylation sites. Accuracy assesses how accurate a predictor is, and MCC measures the classification quality of a predictor. Interested readers should refer to these studies [43-61] for more detailed descriptions.

The ideal predictor should demonstrate a significant improvement across all the above metrics. If this is not possible, at least sensitivity should be improved when compared to previous approaches.

Validation Strategy

To evaluate the performance of a prediction model, two schemes known as n-fold cross-validation and jackknife [62] are often used. Both strategies are suitable due to the very limited number of available samples. In these schemes, a different test set is always employed for assessing the predictor. It is worth noting that an independent test data was used for the two predictors compared in this study, however, it was not employed during parameter learning. Concretely, both predictors used a similar dataset extracted from the "Compendium of Protein Lysine Modifications." The least arbitrary of the above strategies is jackknife, which returns unique outcomes for each benchmark dataset. In spite of this, we used here the n-fold cross-validation strategy for a faster processing time. The procedure was conducted in five steps:

Step 1. The dataset was randomly split into n subsets of equal size.

Step 2. One subset was retained for validation purposes whereas the remaining subsets were used

for training the predictor.

Step 3. The parameters of the predictor were estimated with the training subsets.

Step 4. The evaluation metrics were calculated on the validation subset.

Step 5. The above steps were repeated n times for computing each average metric across all the partitions.

To ensure that our predictor was fairly compared, we conducted 6-, 8- and 10-fold cross-validations for evaluating its performance. Because we do not have any feature extraction or feature selection, it is also important to emphasize that any parameter tuning was dramatically reduced. This contributed to eliminate any potential bias towards the proposed model.

Comparison of C-iSumo and pSumo-CD predictors

The pSumo-CD predictor [63] has reportedly achieved the best results for sumoylation prediction among other state-of-the-art predictors in the literature. Therefore, we decided to compare the C-iSumo predictor with it. It is worth noting that there is a limited number of approaches using computational methods for tackling the sumoylation prediction problem (some are reviewed in the introductory section). However, most of these studies do not have an available predictor, which makes it even harder to establish fair comparisons. One clear advantage of pSumo-CD [63] is its web-server, which constitutes the necessary tool that enables us to directly compare our predictor.

Having said that, we were able to submit the protein sequences to the trained pSumo-CD web-server for sumoylation site identification. To accurately compare both approaches based on their prediction results, all the sequences in our benchmark dataset were manually uploaded to the pSumo-CD web server [63], and the predicted sites were retrieved. It should be noted that the pSumo-CD web server was trained with part of the sequences in our dataset, which could

somehow contribute to favorably bias its performance. Moreover, the pSumo-CD performance was reported using the validation samples of the cross-validation scheme.

Table 1 depicts the comparison between C-iSumo and pSumo-CD predictors. As it clearly shows, the C-iSumo predictor outperformed the pSumo-CD method in statistical metrics such as sensitivity, accuracy and MCC. Although accuracy and MCC slightly improved by 4.2% and 6.7%, sensitivity achieved a significant improvement of 36.9%. This performance confirms the practical use of the C-iSumo predictor for detecting sumoylation sites in real scenarios. Though the specificity of the pSumo-CD method [63] turned out to be higher than that of the C-iSumo predictor, it is reasonable to assume that non-sumoylation sites outnumber sumoylation sites. Additionally, the Receiving Operating Characteristic for 6-, 8- and 10-fold cross-validations was drawn and the area under the curve was computed at 0.73, 0.75 and 0.74, respectively (Figure 3). Because of this, the detection of non-sumoylation sites could turn out to be much easier.

Table 1: Comparison of C-iSumo and pSumo-CD predictors.

Predictor	**Sensitivity**	**Specificity**	**Accuracy**	**MCC**
pSumo-CD [63]	0.536	**0.896**	0.716	0.463
C-iSumo (6-CV)	0.710	0.752	0.731	0.465
C-iSumo (8-CV)	**0.734**	0.757	**0.746**	**0.494**
C-iSumo (10-CV)	0.719	0.758	0.738	0.478

*The highest value of each metric is highlighted in bold.

The computed metrics (Table 1) give a clear evidence of the promising results achieved by the C-iSumo predictor which have not been attained by any other predictor in the literature. This is due to the incorporation of key structural features, such as accessible surface area, and the sine

and cosine functions of four backbone torsion angles (ϕ, Ψ, θ and τ). Such features were computed with the tool SPIDER2 [24] and they appear to be useful for the difficult task of sumoylated lysine detection. Therefore, it seems possible to create better computational predictors by using the above structural characteristics given their importance in discriminating between sumoylation and non-sumoylation sites.

The feature matrices used in this study can be accessed at https://github.com/YosvanyLopez/C-iSumo.

Conclusions

In this paper, we proposed a new computational approach able to accurately predict sumoylation residues. The method, called "C-iSumo", combines two essential characteristics related to protein structure, namely, accessible surface area and the sine and cosine of four torsion angles. An under-sampling strategy was employed for dealing with the imbalance between classes, and an ensemble of decision trees, AdaBoost, was finally designed for classification purposes. The proposed method was compared to another benchmark predictor (pSumo-CD), outperforming it in metrics such as sensitivity, accuracy, and Matthews correlation coefficient.

Competing interests

The authors declare that they do not have any competing interests.

References

1. Knorre DG, Kudryashova NV, Godovikova TS. Chemical and Functional Aspects of Posttranslational Modification of Proteins. Acta Naturae. 2009;1(3):29-51.
2. Comb DG, Sarkar N, Pinzino CJ. The Methylation of Lysine Residues in Protein. The Journal of Biological Chemistry. 1966;241(8).
3. Martin C, Zhang Y. The diverse functions of histone lysine methylation. Nature Reviews Molecular Cell Biology. 2005;6:838-49.
4. Drazic A, Myklebust LM, Ree R, Arnesen T. The world of protein acetylation. Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics. 2016;1864(10):1372-401.

5. Zhang Z, Tan M, Xie Z, Dai L, Chen Y, Zhao Y. Identification of lysine succinylation as a new post-translational modification. Nature Chemical Biology. 2011;7(1):58-63.
6. Lopez Y, Sharma A, Dehzangi A, Lal SP, Taherzadeh G, Sattar A, et al. Success: evolutionary and structural properties of amino acids prove effective for succinylation site prediction. Bmc Genomics. 2018;19. doi: ARTN 923
10.1186/s12864-017-4336-8. PubMed PMID: WOS:000422886100011.
7. Dehzangi A, Lopez Y, Lal S, Taherzadeh G, Sattar A, Tsunoda T, et al. Improving succinylation prediction accuracy by incorporating the secondary structure via helix, strand and coil, and evolutionary information from profile bigrams. PLoS One. 2018;13(2):e0191900.
8. Dehzangi A, Lopez Y, Lal SP, Taherzadeh G, Michaelson J, Sattar A, et al. PSSM-Suc: Accurately predicting succinylation using position specific scoring matrix into bigram for feature extraction. Journal of Theoretical Biology. 2017;425:97-102. doi: 10.1016/j.jtbi.2017.05.005. PubMed PMID: WOS:000403743800009.
9. Lopez Y, Dehzangi A, Lal SP, Taherzadeh G, Michaelson J, Sattar A, et al. SucStruct: Prediction of succinylated lysine residues by using structural properties of amino acids. Anal Biochem. 2017;527:24-32. doi: 10.1016/j.ab.2017.03.021. PubMed PMID: WOS:000400954600005.
10. Lamoliatte F, Caron D, Durette C, Mahrouche L, Maroui MA, Caron-Lizotte O, et al. Large-scale analysis of lysine SUMOylation by SUMO remnant immunoaffinity profiling. Nat Commun. 2014;5:5409.
11. Han Z-J, Feng Y-H, Gu B-H, Li Y-M, Chen H. The post-translational modification, SUMOylation, and cancer (Review). International Journal of Oncology. 2018;52:1081-94.
12. Stefano BD, Hochedlinger K. Novel Roles for SUMOylation in Cellular Plasticity. Trends in Cell Biology. 2018;28(12):P971-3.
13. Heideker J, Perry JJP, Boddy MN. Genome Stability Roles of SUMO-targeted Ubiquitin Ligases. DNA Repair (Amst). 2009;8(4):517-24.
14. Zilio N, Eifler-Olivi K, Ulrich HD. Functions of SUMO in the Maintenance of Genome Stability. In: Wilson VG, editor. SUMO Regulation of Cellular Processes. Advances in Experimental Medicine and Biology. 963: Springer; 2017. p. 51-87.
15. Xue Y, Zhou F, Fu C, Xu Y, Yao X. SUMOsp: a web server for sumoylation site prediction. Nucleic Acids Res. 2006;34(Suppl 2):W254-W7.
16. Ren J, Gao X, Jin C, Zhu M, Wang X, Shaw A, et al. Systematic study of protein sumoylation: Development of a site-specific predictor of SUMOsp 2.0. Proteomics. 2009;9(12):3409-12.
17. Zhao Q, Xie Y, Zheng Y, Jiang S, Liu W, Mu W, et al. GPS-SUMO: a tool for the prediction of sumoylation sites and SUMO-interaction motifs. Nucleic Acids Res. 2014;42(W1):W325-W30.
18. Xu J, He Y, Qiang B, Yuan J, Peng X, Pan X-M. A novel method for high accuracy sumoylation site prediction from protein sequences. BMC Bioinformatics. 2008;9:8.
19. Chen Y-Z, Chen Z, Gong Y-A, Ying G. SUMOhydro: A Novel Method for the Prediction of Sumoylation Sites Based on Hydrophobic Properties. PLoS ONE. 2012;7(6):e39195.
20. Xu Y, Ding Y-X, Deng N-Y, Liu L-M. Prediction of sumoylation sites in proteins using linear discriminant analysis. Gene. 2016;576(1):99-104.
21. Yen S-J, Lee Y-S. Under-Sampling Approaches for Improving Prediction of the Minority Class in an Imbalanced Dataset. In: Huang D-S, Li K, Irwin GW, editors. Intelligent Control and Automation Lecture Notes in Control and Information Sciences. 344. Berlin, Heidelberg: Springer; 2006.
22. Liu Z, Cao J, Gao X, Zhou Y, Wen L, Yang X, et al. CPLA 1.0: an integrated database of protein lysine acetylation. Nucleic Acids Res. 2011;39(Database issue):D1029-D34.
23. Liu Z, Wang Y, Gao T, Pan Z, Cheng H, Yang Q, et al. CPLM: a database of protein lysine modifications. Nucleic Acids Res. 2014;42(Database issue):D531-D6.
24. Yang Y, Heffernan R, Paliwal K, Lyons J, Dehzangi A, Sharma A, et al. SPIDER2: A Package to Predict Secondary Structure, Accessible Surface Area, and Main-Chain Torsional Angles by Deep Neural

Networks. In: Zhou Y, Kloczkowski A, Faraggi E, Yang Y, editors. Prediction of Protein Secondary Structure. Methods in Molecular Biology. 1484: Springer New York; 2016. p. 55-63.
25. Heffernan R, Paliwal K, Lyons J, Dehzangi A, Sharma A, Wang J, et al. Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. Scientific Reports. 2015;5:11476.
26. Lyons J, Dehzangi A, Heffernan R, Sharma A, Paliwal K, Sattar A, et al. Predicting Backbone Cα Angles and Dihedrals from Protein Sequences by Stacked Sparse Auto-Encoder Deep Neural Network. Journal of Computational Chemistry. 2014;35(28):2040-6.
27. Xiao X, Min J-L, Lin W-Z, Liu Z, Cheng X, Chou K-C. iDrug-Target: predicting the interactions between drug compounds and target proteins in cellular networking via benchmark dataset optimization approach. Journal of Biomolecular Structure and Dynamics. 2015;33(10):2221-33.
28. Jia J, Liu Z, Xiao X, Liu B, Chou K-C. iSuc-PseOpt: Identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset. Anal Biochem. 2016;497:48-56.
29. Liu Z, Xiao X, Qiu W-R, Chou K-C. iDNA-Methyl: Identifying DNA methylation sites via pseudo trinucleotide composition. Anal Biochem. 2015;474:69-77.
30. Chen W, Feng P, Ding H, Lin H, Chou K-C. iRNA-Methyl: Identifying N6-methyladenosine sites using pseudo nucleotide composition. Anal Biochem. 2015;490:26–33.
31. Jia JH, Liu Z, Xiao X, Liu BX, Chou KC. pSuc-Lys: Predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach. Journal of Theoretical Biology. 2016b;394:223-30. doi: 10.1016/j.jtbi.2016.01.020. PubMed PMID: WOS:000379888800020.
32. Chandra AA, Sharma A, Dehzangi A, Tsunoda T. EvolStruct-Phogly: incorporating structural properties and evolutionary information from profile bigrams for the phosphoglycerylation prediction. BMC Genomics. 2019;19(Suppl 9):984. doi: 10.1186/s12864-018-5383-5. PubMed PMID: 30999859.
33. Chandra A, Sharma A, Dehzangi A, Ranganathan S, Jokhan A, Chou KC, et al. PhoglyStruct: Prediction of phosphoglycerylated lysine residues using structural properties of amino acids. Sci Rep. 2018;8(1):17923. doi: 10.1038/s41598-018-36203-8. PubMed PMID: 30560923; PubMed Central PMCID: PMCPMC6299098.
34. Reddy HM, Sharma A, Dehzangi A, Shigemizu D, Chandra AA, Tsunoda T. GlyStruct: glycation prediction using structural properties of amino acid residues. BMC Bioinformatics. 2019;19(Suppl 13):547. doi: 10.1186/s12859-018-2547-x. PubMed PMID: 30717650.
35. Hussain W, Khan YD, Rasool N, Khan SA, Chou KC. SPrenylC-PseAAC: A sequence-based model developed via Chou's 5-steps rule and general PseAAC for identifying S-prenylation sites in proteins. J Theor Biol. 2019;468:1-11. doi: 10.1016/j.jtbi.2019.02.007. PubMed PMID: 30768975.
36. Xu Y, Ding YX, Ding J, Lei YH, Wu LY, Deng NY. iSuc-PseAAC: predicting lysine succinylation in proteins by incorporating peptide position-specific propensity. Sci Rep. 2015;5:10184. doi: 10.1038/srep10184. PubMed PMID: 26084794; PubMed Central PMCID: PMCPMC4471726.
37. Freund Y, Schapire RE. A Short Introduction to Boosting. Journal of Japanese Society for Artificial Intelligence. 1999;14(5):771-80.
38. Freund Y, Schapire RE, editors. Experiments with a New Boosting Algorithm. Thirteenth International Conference on Machine Learning; 1996.
39. Freund Y, Schapire RE. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. Journal of Computer and System Sciences. 1997;55:119-39.
40. Niu B, Cai Y-D, Lu W-C, Li G-Z, Chou K-C. Predicting Protein Structural Class with AdaBoost Learner. Protein & Peptide Letters. 2006;13(5):489-92.
41. Buitinck L, Louppe G, Blondel M, Pedregosa F, Mueller A, Grisel O, et al. API design for machine learning software: experiences from the scikit-learn project. ECML PKDD Workshop: Languages for Data Mining and Machine Learning2013. p. 108-22.

42. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825-30.
43. Chen W, Tang H, Ye J, Lin H, Chou K-C. iRNA-PseU: Identifying RNA pseudouridine sites. Molecular Therapy – Nucleic Acids. 2016;5:e332.
44. Cheng X, Xiao X, Chou K-C. pLoc-mVirus: Predict subcellular localization of multi-location virus proteins via incorporating the optimal GO information into general PseAAC. Gene. 2017;628:315-21.
45. Feng P, Ding H, Yang H, Chen W, Lin H, Chou K-C. iRNA-PseColl: Identifying the Occurrence Sites of Different RNA Modifications by Incorporating Collective Effects of Nucleotides into PseKNC. Molecular Therapy - Nucleic Acids. 2017;7:155-63.
46. Liu B, Wang S, Long R, Chou K-C. iRSpot-EL: identify recombination spots with an ensemble learning approach. Bioinformatics. 2017;33(1):35-41.
47. Qiu W-R, Sun B-Q, Xiao X, Xu Z-C, Jia J-H, Chou K-C. iKcr-PseEns: Identify lysine crotonylation sites in histone proteins with pseudo components and ensemble classifier. Genomics (In Press). 2017.
48. Cheng X, Zhao S-G, Lin W-Z, Xiao X, Chou K-C. pLoc-mAnimal: predict subcellular localization of animal proteins with both single and multiple sites. Bioinformatics. 2017;33(22):3524-31.
49. Qiu W-R, Jiang S-Y, Sun B-Q, Xiao X, Cheng X, Chou K-C. iRNA-2methyl: Identify RNA 2'-O-methylation Sites by Incorporating Sequence-Coupled Effects into General PseKNC and Ensemble Classifier. Med Chem. 2017;13(8):734-43.
50. Cheng X, Xiao X, Chou K-C. pLoc-mEuk: Predict subcellular localization of multi-label eukaryotic proteins by extracting the key GO information into general PseAAC. Genomics. 2018;110(1):50-8.
51. Ehsan A, Mahmood K, Khan YD, Khan SA, Chou K-C. A Novel Modeling in Mathematical Biology for Classification of Signal Peptides. Scientific Reports. 2018;8:1039.
52. Feng P, Yang H, Ding H, Lin H, Chen W, Chou K-C. iDNA6mA-PseKNC: Identifying DNA N6-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC. Genomics (In Press). 2018.
53. Liu B, Yang F, Huang D-S, Chou K-C. iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC. Bioinformatics. 2018;34(1):33-40.
54. Cheng X, Xiao X, Chou K-C. pLoc-mHum: predict subcellular localization of multi-location human proteins via general PseAAC to winnow out the crucial GO information. Bioinformatics (In Press). 2017.
55. Lin H, Deng E-Z, Ding H, Chen W, Chou K-C. iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. Nucleic Acids Res. 2014;42(21):12961-72.
56. Liu B, Yang F, Chou K-C. 2L-piRNA: A Two-Layer Ensemble Classifier for Identifying Piwi-Interacting RNAs and Their Function. Molecular Therapy - Nucleic Acids. 2017;7:267-77.
57. Liu B, Long R, Chou K-C. iDHS-EL: identifying DNase I hypersensitive sites by fusing three different modes of pseudo nucleotide composition into an ensemble learning framework. Bioinformatics. 2016;32(16):2411-8.
58. Liu B, Fang L, Long R, Lan X, Chou K-C. iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. Bioinformatics. 2016;32(3):362-9.
59. Sharma R, Bayarjargal M, Tsunoda T, Patil A, Sharma A. MoRFPred-plus: Computational Identification of MoRFs in Protein Sequences using Physicochemical Properties and HMM profiles. Journal of Theoretical Biology. 2018;437:9-16. doi: 10.1016/j.jtbi.2017.10.015. PubMed PMID: WOS:000417228400002.
60. Uddin MR, Sharma A, Farid MD, Rahman MM, Dehzangi A, Shatabda S. EvoStruct-Sub: An accurate Gram-positive protein subcellular localization predictor using evolutionary and structural features. Journal of Theoretical Biology. 2018;443:138-46.

61. Shatabda S, Saha S, Sharma A, Dehzangi A. iPHLoc-ES: Identification of bacteriophage protein locations using evolutionary and structural features. Journal of Theoretical Biology. 2017;435:229-37.
62. Alpaydin E. Introduction to Machine Learning. Third ed: The MIT Press; 2014.
63. Jia J, Zhang L, Liu Z, Xiao X, Chou K-C. pSumo-CD: Predicting sumoylation sites in proteins with covariance discriminant algorithm by incorporating sequence-coupled effects into general PseAAC. Bioinformatics. 2016;32(20):3133-41.

Figures

Figure 1. Flowchart of the proposed methodology.

Figure 2. Description of a lysine residue by its surrounding amino acids. A) 15 upstream and 15 downstream amino acids are regarded, B) the side with missing amino acids was mirrored by taking into consideration the other side of the lysine.

Figure 3. Receiving Operating Characteristic for A) 6-fold B) 8-fold and C) 10-fold cross-validations.