Improving succinylation prediction accuracy by incorporating the secondary structure via helix, strand and coil, and evolutionary information from profile bigrams

Abdollah Dehzangi[1☯], **Yosvany López**[2,3☯]*, **Sunil Pranit Lal**[4], **Ghazaleh Taherzadeh**[5], **Abdul Sattar**[5,6], **Tatsuhiko Tsunoda**[2,3,7‡], **Alok Sharma**[3,6,8‡]

1 Department of Computer Science, Morgan State University, Baltimore, Maryland, United States of America, **2** Department of Medical Science Mathematics, Medical Research Institute, Tokyo Medical and Dental University, Tokyo, Japan, **3** Laboratory for Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences, Yokohama, Kanagawa, Japan, **4** School of Engineering & Advanced Technology, Massey University, Palmerston North, New Zealand, **5** School of Information and Communication Technology, Griffith University, Queensland, Australia, **6** Institute for Integrated and Intelligent Systems, Griffith University, Queensland, Australia, **7** CREST, JST, Tokyo, Japan, **8** School of Engineering & Physics, University of the South Pacific, Suva, Fiji

☯ These authors contributed equally to this work.
‡ These authors also contributed equally to this work.
* yosvany.lopez.alvarez@gmail.com

Abstract

Post-translational modification refers to the biological mechanism involved in the enzymatic modification of proteins after being translated in the ribosome. This mechanism comprises a wide range of structural modifications, which bring dramatic variations to the biological function of proteins. One of the recently discovered modifications is succinylation. Although succinylation can be detected through mass spectrometry, its current experimental detection turns out to be a timely process unable to meet the exponential growth of sequenced proteins. Therefore, the implementation of fast and accurate computational methods has emerged as a feasible solution. This paper proposes a novel classification approach, which effectively incorporates the secondary structure and evolutionary information of proteins through profile bigrams for succinylation prediction. The proposed predictor, abbreviated as SSEvol-Suc, made use of the above features for training an AdaBoost classifier and consequently predicting succinylated lysine residues. When SSEvol-Suc was compared with four benchmark predictors, it outperformed them in metrics such as sensitivity (0.909), accuracy (0.875) and Matthews correlation coefficient (0.75).

Introduction

Post-translational modification (PTM) refers to the enzymatic modification of proteins [1]. As part of this biological mechanism, one or more amino acids of a protein interact with specific molecular groups. Such interaction functionally changes the amino acids, thereby impacting

the function of the entire protein. From the 20 amino acids that make up the genetic code, lysine has proven to be the most susceptible residue to PTM. It has been involved in many PTMs including methylation [2, 3], sumoylation [4], acetylation [5], glycation [6] and malonylation [7]. One of the recently identified PTMs is succinylation, which reportedly impacts the function and structure of proteins within biological interactions [8]. Succinylation mainly refers to the addition of a succinyl group to lysine residues. This molecular change alters the charge of the lysine to -1, thus introducing a large structural moiety. Succinylation occurs in both eukaryotic and prokaryotic cells, and is common in enzymes involved in mitochondrial metabolism, amino acid degradation, and fatty acid metabolism. It has been also observed in histones with functions in chromosome configuration and gene expression. Nevertheless, the role of succinylation in other biological reactions needs to be extensively clarified. Therefore, the identification of succinylation sites can provide detailed insights into the function of proteins and their biological interactions.

The identification of PTM sites has become a serious challenge in the last years [9]. In this direction, many bioinformatics methods have been proposed for detecting them within protein sequences [10–30]. Experimental methods like mass spectrometry remain the main technique for identifying lysine succinylation sites. However, these methods are costly and unable to keep up with the exponential growth of sequenced proteins. Consequently, there is an urgent demand for fast and accurate computational methods capable of predicting succinylation sites. In the past years, a wide range of prediction methods have been proposed to tackle this issue, but their performance is consistently limited. This limitation is more apparent for lysine succinylation than for other PTMs because of its recent discovery. Thus far, the pioneering methods proposed to solve this problem have been mainly focused on analyzing the protein sequence. Two of these methods are SucPred [31] and SuccFind [32]. The former is a semi-supervised machine learning-based method, which incorporates the sequence and physicochemical features into a support vector machine for classification. The later, however, introduced a more robust approach that considers information about the neighboring amino acids of succinylated and non-succinylated lysines to better discriminate between them. Another method, iSuc-PseAAC, employed a strategy that integrates the peptide position-specific propensity into the general form of pseudo amino acid composition for training a support vector machine [33]. Another method that incorporates sequence-coupling effects into the pseudo amino acid composition was iSuc-PseOpt [34]. It introduced the k-nearest neighbors strategy and hypothetical training samples in an attempt to ameliorate the imbalance between classes. Subsequently, a random forest algorithm was designed for prediction. SuccinSite also regarded a random forest classifier but with informative encoding features, such as the composition of k-spaced amino acid pairs, binary encoding and specific physicochemical attributes [35]. However, the above predictors showed a poor sensitivity when it comes to detecting succinylated lysine residues.

Studies related to protein subcellular localization [36], structure and function prediction [37, 38], and local structure and torsion angles prediction [39] have demonstrated that the structural and evolutionary information of proteins can significantly improve prediction performance. We previously proposed two different predictors: SucStruct [40] and PSSM-Suc [41], which corroborated the above premise. For instance, SucStruct used structural features like secondary structure and torsion angles [40], whereas PSSM-Suc transformed the evolutionary information of the position specific scoring matrix (PSSM) for succinylation prediction [41]. Both approaches trained a pruned decision tree for classification purposes, and outperformed state-of-the-art predictors which only relied on sequence and physicochemical attributes. These predictors clearly demonstrated that the use of powerful classifiers alongside evolutionary and structural attributes can significantly improve succinylation prediction.

In order to design an efficient sequence-based computational predictor for solving biological problems, a long list of studies [25, 26, 42–47] has made reference to a five-step rule [48]. This rule comprises the following steps: (1) the construction or selection of a correct dataset for training and testing a predictor, (2) the use of an accurate mathematical expression for transforming the biological sequence and considering the intrinsic correlation to future predictions, (3) the development of an exact algorithm for making predictions, (4) the proper use of statistical metrics for assessing the predictor accuracy, and (5) the design of a user-friendly server for making the predictor available to the public. These steps will be described in the subsequent sections.

Given the explosion of biological sequences, one of the most serious challenges is how to represent these sequences as discrete models or vectors while keeping the information related to the order of sequences. This problem is often caused by the intrinsic limitations of machine learning algorithms, which can only handle numerical vectors [16]. Besides any vector could lose the information of patterns in a sequence. In order to overcome the above limitations for protein sequences, the pseudo amino acid composition (PseAAC) [49] was proposed. Since its proposal, the concept known as Chou's PseAAC, has been widely used in the area of computational proteomics [9, 50]. PseAAC has been recently incorporated in three software: 'PseAAC-Builder', 'propy', and 'PseAAC-General'. The first two are aimed at creating models of Chou's special PseAAC, whereas the third one uses the Chou's general PseAAC [48]. These software considered the special modes of feature vectors in addition to high-level vectors such as 'functional domain', 'gene ontology' and 'sequential evolution', or 'PSSM' modes [48]. Due to the usefulness of PseAAC for dealing with protein/peptide sequences, a new concept coined pseudo k-tuple nucleotide composition [51], aimed at generating feature vectors from DNA/RNA sequences, was proposed. Recently, a new web server called 'Pse-in-One' [52] and its updated version 'Pse-in-One 2.0' [53], which facilitate the generation of feature vectors from protein/peptide or DNA/RNA sequences, were developed. Our study made use of secondary structure and evolutionary information for defining pseudo components and thus identifying succinylation sites.

In this work, we propose a new predictor, SSEvol-Suc, which primarily integrates information about the best secondary structure and the PSSM for predicting succinylation sites [54, 55]. Our predictor combines both features and transforms them into profile bigrams [56] in order to describe each lysine residue. The k-nearest neighbors strategy was employed for reducing the imbalance between succinylation and non-succinylation sites [34]. An AdaBoost classifier was finally designed for discriminating between lysine residues. We compared the prediction results of SSEvol-Suc with those of iSuc-PseAAC [33], SuccinSite [35], iSuc-PseOpt [34] and pSuc-Lys [57]. SSEvol-Suc achieved remarkable results by outperforming all the above predictors. Its sensitivity, accuracy and Matthews correlation coefficient (MCC) were recorded at 0.909, 0.875 and 0.75, respectively.

Materials and methods

In this paper, we propose a novel predictor, SSEvol-Suc, which makes use of the secondary structure and the PSSM of proteins for accurately predicting succinylation sites [58–60]. These features were transformed into profile bigrams and employed for describing each lysine residue. The resulting matrix was then used for training an AdaBoost classifier and predicting succinylated lysines.

Benchmark dataset

The benchmark dataset was extracted from the Compendium of Protein Lysine Modifications (CPLM) [61, 62]. This compendium consists of over 45,000 proteins from 122 species, and 12

different annotated PTMs. In the CPLM, succinylation was the most abundant and diversely distributed mark across all the included species [62]. The original collection comprised 2,521 succinylation and 24,128 non-succinylation sites from 896 unique proteins. To avoid overestimations due to homology and be able to directly compare our results with those of previous studies, we removed those proteins with $\geq$ 40% pairwise sequential similarity. We then ended up with a benchmark dataset consisting of 670 unique proteins, where the longest and shortest proteins were 5,656 and 47 residues long and the average protein comprised 464 residues. The 1,782 succinylation and 18,344 non-succinylation sites located in such proteins were grouped into two mutually exclusive collections: positive and negative. The subsequent sections will introduce the structural and evolutionary features computed from the protein sequences.
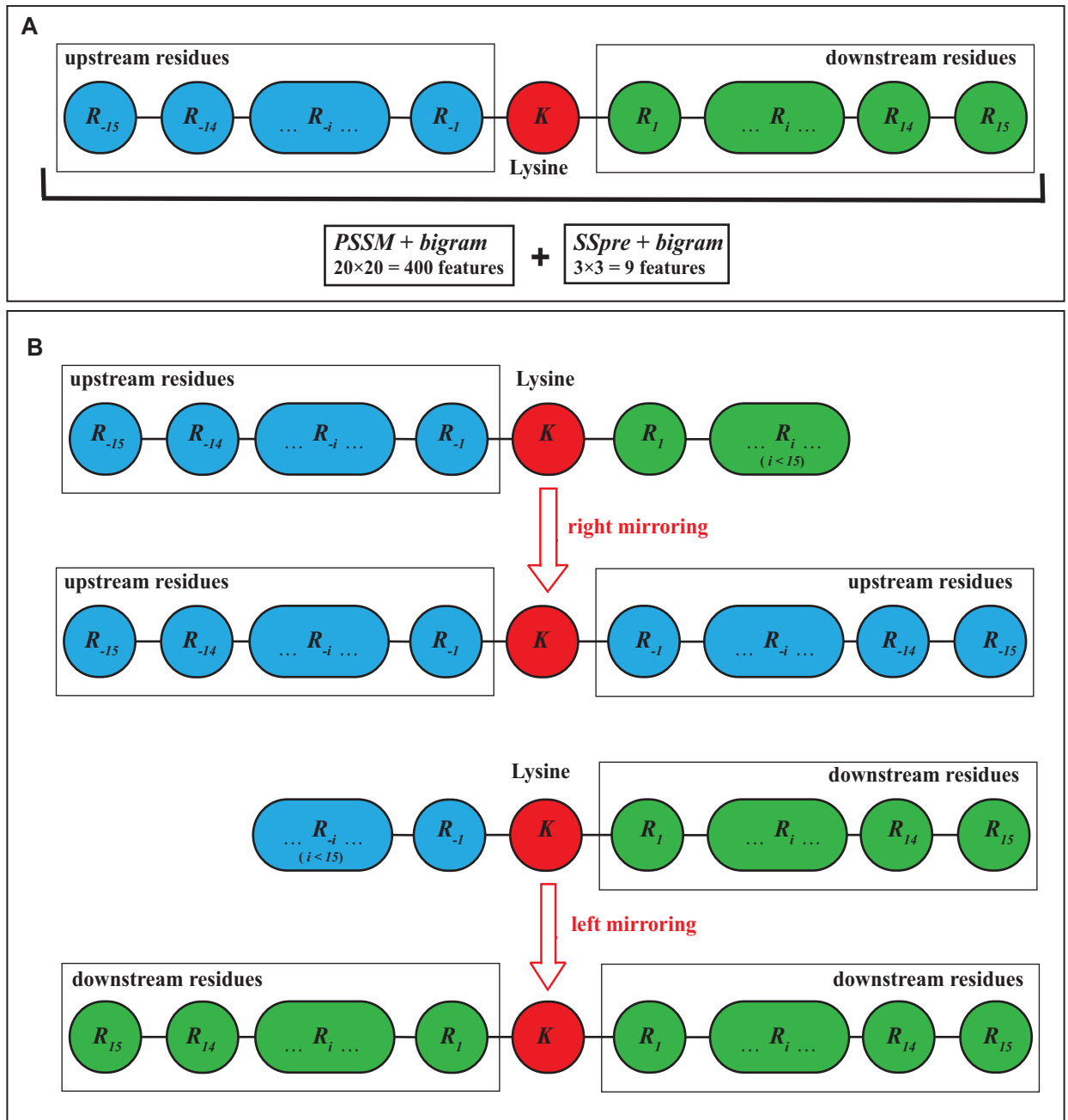
Secondary structure feature

The secondary structure of proteins provides accurate information about their local structure and how they fold into their general tertiary configuration. We predicted the secondary structure of each protein in our benchmark dataset with the tool SPIDER2 [63, 64]. SPIDER2 is one of the latest predictors aimed at computing the local structure of proteins. This software has been successfully used to compute the structural properties of proteins in sequence-based predictions of protein binding sites. Secondary structure indicates the contribution of each amino acid to specific local structures, namely, helix, strand, and coil for determining the local 3D structure of proteins. In other words, secondary structure determines the local structure of proteins by considering the local configuration of amino acids in the sequence. Therefore, its understanding can provide critical information about the function and folding of proteins. We run SPIDER2 on each protein sequence by providing all the sequences in FASTA format. SPIDER2 automatically detects a FASTA file, and for each sequence, it retrieves the local structure with the highest probability. This results in a matrix of size $L \times 3$, where L represents the protein length and the three columns indicate the transition probabilities to the three secondary structure conformations (helix, strand and coil). Hereafter, we refer to this matrix as *SSpre*.

Evolutionary feature

Evolutionary information provides valuable insights into structural, functional and sequential similarities among proteins based on how they evolved [65]. PSSM describes the substitution probability of each amino acid in a protein with all the amino acids of the genetic code. This matrix was computed with the alignment toolbox PSI-BLAST [66], which aligns each protein to similar proteins in the Protein Data Bank [67]. We run PSI-BLAST on all the proteins in our benchmark dataset and retrieved the corresponding PSSM. For each protein, PSI-BLAST produces two $L \times 20$ matrices, where L is the protein length and the 20 columns indicate the amino acids of the genetic code. The running of PSI-BLAST was conducted on non-redundant proteins in the Protein Data Bank, with a cutoff (E) of 0.001 and three iterations. From these matrices, we used the normalized matrix which comprises the substitution probabilities of amino acids.

Lysine residues as profile bigrams

The structural and evolutionary features were used to describe each succinylated and non-succinylated lysine residue. Lysines (*K*) were described by considering their adjacent 15 upstream and 15 downstream amino acids (Fig 1A) [34]. If a lysine residue did not contain 15 amino acids (either upstream or downstream), we mirrored the missing peptide stretch (Fig 1B). The sequence segment *S* consisting of 15 upstream and 15 downstream residues in addition to the

Fig 1. Schematic representation of a lysine residue and its surrounding amino acids. (A) lysine with 15 residues on both sides, (B) lysine with missing residues to the right and left.

lysine K was expressed as

$$S = \{R_{-15}, R_{-14}, \ldots, R_{-2}, R_{-1}, K, R_1, R_2, \ldots, R_{14}, R_{15}\} \tag{1}$$

where R_{-i} and R_i (for $1 \leq i \leq 15$) are upstream and downstream amino acids, respectively. It can be observed from Eq (1) that 31 amino acids (including K) were used for defining each lysine residue. Accordingly, each lysine represented by the sequence segment S was labeled. In

other words, the segment S comprising a succinylation site was labeled as 1 whereas that describing a non-succinylation site was labeled as 0.

To describe each lysine, the submatrices *SSpre* and *PSSM* around the lysine were retrieved and transformed into frequency vectors of bigrams (*PSSM* + *bigram* and *SSpre* + *bigram*). These transformations resulted in two matrices of sizes 20 × 20 (for *PSSM* + *bigram*) and 3 × 3 (for *SSpre* + *bigram*). Each segment S was finally described by a 409-feature vector. This feature vector was then used to capture the structural and evolutionary information about the lysine represented by the segment S.

The bigram method has shown promising results when it comes to exploring discriminatory information [36, 56, 68–70], so that we used it here. The scheme for transforming the submatrices *SSpre* and *PSSM* into frequency vectors is explained below. The *PSSM* matrix M of size L × 20 and the *SSpre* matrix N of size L × 3 were used to construct a feature vector. Each element m_{ij} and n_{ij} of the matrices M and N, respectively, represents the transitional probability of the j -th amino acid/secondary structure conformation at i-th position in the protein sequence. The sequence segment S (Eq (1)) was described by two matrices of sizes 31 × 20 (for *PSSM*) and 31 × 3 (for *SSpre*). The matrices M and N were processed as profile bigrams [56] by

$$B_{p,q} = \sum_{k=1}^{30} m_{k,p} m_{k+1,q} \tag{2}$$

and

$$B'_{r,s} = \sum_{k=1}^{30} n_{k,r} n_{k+1,s} \tag{3}$$

where $1 \leq p,q \leq 20$ for the matrix M and $1 \leq r,s \leq 3$ for the matrix N.

Thus, the matrix B, which represents *PSSM* + *bigram* and its elements $B_{p,q}$ (for p = 1,2,..., 20 and q = 1,2,...,20), will be a 20 × 20 matrix. Similarly, the matrix B', which represents *SSpre* + *bigram* and consists of elements $B'_{r,s}$, (for r = 1,...,3 and s = 1,...,3) will be of size 3 × 3. Subsequently, the matrices B and B' can be transformed as

$$F = [B_{1,1}, B_{1,2} \ldots, B_{1,20}, B_{2,1}, \ldots B_{20,1}, \ldots B_{20,20}, B'_{1,1}, B'_{1,2} \ldots, B'_{3,3}]^T \tag{4}$$

where T is the transpose. Therefore, the matrix B will have 400 transitional probabilities and the matrix B' will comprise 9 transitional probabilities. Eq (4) is the feature vector, which contains 409 transitional probabilities and results from the *PSSM* + *bigram* and *SSpre* + *bigram* matrices. In other words, each lysine residue was defined by a 409-dimensional vector of structural and evolutionary features.

This information was computed for all the lysine residues in our benchmark dataset, resulting in a training matrix of 1,782 succinylation sites (*label* = 1) and 18,344 non-succinylation sites (*label* = 0). Such a matrix was further processed to reduce the imbalance between classes, and ultimately used for training an AdaBoost classifier (refer to the following section).

One advantage of the bigram method is its window-size independent nature. For instance, it extracts 400- and 9-dimensional feature vectors regardless of the window size adopted around lysine residues. Thereby, the bigram method enables us to enlarge the window around lysines without necessarily increasing the number of features.

AdaBoost classifier

Adaptive Boosting (AdaBoost) is a meta-classifier, which iteratively applies a base learner and adjusts its parameters to build a strong ensemble classifier [71]. The base classifier, usually a decision tree, is first applied to the training dataset. The weights are then iteratively adjusted by increasing the weight for misclassified samples. This procedure continues until changes in

the weights become trivial. Finally, AdaBoost combines the base classifiers across all the iterations to build the final predictor [72]. Decision trees are usually used as base classifiers because they can reflect larger changes due to their sensitivity to weight adjustments [72, 73]. AdaBoost has been successfully used in studies related to protein folding, attaining promising results that emphasize its applicability to protein science [74, 75]. We utilized the Weka implementation of the AdaBoost algorithm [76] with 1,000 iterations. Decision stumps, which are one-level decision trees, were used as weak classifiers.

Results and discussion

Any predictor, aimed at predicting succinylation sites, must have its performance assessed. In this work, we evaluated the performance of SSEvol-Suc in terms of four different statistical metrics: sensitivity, specificity, accuracy and Matthews correlation coefficient [15, 36, 77–80]. The following sections will discuss these metrics in addition to aspects such as class imbalance and predictor performance.

Evaluation metrics

The first metric, sensitivity, was used to evaluate the proportion of correctly predicted succinylation sites. If the predictor is able to accurately detect succinylation sites in the dataset, a high sensitivity will be achieved. For instance, a predictor with a sensitivity of 1 is able to accurately detect positive (succinylation) sites whereas that with a sensitivity of 0 fails to detect these sites.

The second metric, specificity, assesses the predictor ability to correctly detect non-succinylation sites. Similarly, a specificity of 1 presents a predictor able to classify all the negative sites whereas a specificity of 0 points to a predictor unable to detect them.

The third metric, accuracy, evaluates the predictor ability to discriminate between succinylation and non-succinylation sites. The predictor with an accuracy of 1 is an accurate one while that with an accuracy of 0 is regarded an inaccurate predictor.

The fourth metric, Matthews correlation coefficient (MCC), is often used in binary classification when the classes have different sizes. A perfect correlation between observed and predicted instances is indicated by a MCC of 1 whereas a perfect anticorrelation is confirmed by a MCC of -1.

These four metrics can be summarized as

$$sensitivity = 1 - \frac{N_-^+}{N^+} \tag{5}$$

$$specificity = 1 - \frac{N_+^-}{N^-} \tag{6}$$

$$accuracy = 1 - \frac{N_-^+ + N_+^-}{N^+ + N^-} \tag{7}$$

$$MCC = \frac{1 - \left(\frac{N_-^+}{N^+} + \frac{N_+^-}{N^-}\right)}{\sqrt{\left(1 + \frac{N_+^- - N_-^+}{N^+}\right)\left(1 + \frac{N_-^+ - N_+^-}{N^-}\right)}} \tag{8}$$

where N^+ and N_-^+ represent the total amount of positive (succinylation) sites and the number of positive sites misclassified by the predictor. Likewise, N^- and N_+^- indicate the total amount

of negative (non-succinylation) sites and the number of negative sites misclassified by the predictor.

A promising predictor should ideally outperform in the above statistical metrics. In any case, it should achieve a high performance in at least one of the statistics. Conversely, a predictor with a low sensitivity will be clearly displaying an inability to accurately predict succinylation sites.

Validation scheme

For assessing the performance of any predictor, the use of an appropriate validation scheme is absolutely necessary. Several validation schemes, including the n-fold cross-validation and the jackknife, have been proposed [81, 82]. While the jackknife resampling model turns out to be the least arbitrary and yield unique results for a dataset [83], the cross-validation strategy has been extensively used to evaluate previous predictors [33, 34]. Therefore, we also used the cross-validation scheme here for establishing a fair comparison with state-of-the-art predictors.

The cross-validation technique was carried out as follows,

1. The initial dataset was split into n different subsets of equal size.

2. The predictor was trained on the $n - 1$ subsets and tested on the remaining fold.

3. The predictor parameters were adjusted with the $n - 1$ subsets.

4. The four statistical metrics (sensitivity, specificity, accuracy and MCC) were calculated on the test fold.

5. Steps 1 to 4 were repeated n times and the average of each statistical metric was computed.

In this study, we assessed the performance of SSEvol-Suc with 6-, 8- and 10-fold cross-validations.

Dataset balancing

After retrieving the succinylated and non-succinylated lysines from each protein sequence, we obtained a number of non-succinylation (negative) sites greater than that of succinylation (positive) sites. Although such a difference makes sense from a biological viewpoint, it could strongly bias any computational predictor. Because of this, the elimination of class imbalances in training datasets proves critical in pattern recognition studies for achieving bias-free classifications. It is worth noting that different techniques have been proposed for balancing datasets. Though the upsampling of the positive set might further improve the predictor performance as previously suggested [84], we chose to downsample the negative set in order to avoid introducing artificial training instances. Therefore, we used the k-nearest neighbors classifier [34]. To do this, we initially calculated the Euclidean distance between all the instances (lysine residues) in our benchmark dataset. Subsequently, we set a threshold of 10, which indicates the number of neighbors to be regarded. This cutoff, used for ameliorating the imbalance between classes, was intended to provide a better comparison with benchmark predictors [34], which have utilized the same value for dataset balancing. It was computed as the division between the amount of negative (18,344) and positive (1,782) lysines. As a result, those non-succinylation sites, whose 10 nearest neighbors included at least one succinylation site, were removed. However, this initial filtering did not completely eliminate the imbalance so that new cutoffs were computed. These thresholds were calculated by multiplying the initial threshold ($k = 10$) by different integers. The computation procedure was repeatedly carried out until both sets

(succinylation and non-succinylation sites) were balanced. Consequently, the number of negative instances was reduced to 1,604 sites with a cutoff of 60 (i.e., non-succinylation sites, whose 60 nearest neighbors comprised at least one succinylation site, were eliminated). The remaining sets were then used to perform cross-validation and evaluate the performance of the proposed predictor.

Comparison of SSEvol-Suc and current predictors

The proposed predictor, SSEvol-Suc, was compared with four state-of-the-art predictors: iSuc-PseAAC [33], iSuc-PseOpt [34], SuccinSite [35] and pSuc-Lys [57]. These four predictors were implemented into user-friendly web servers for succinylation site prediction. Thereby, we manually uploaded all the protein sequences to the web servers and retrieved their predictions for performance assessment. These web servers were previously trained on part of our sequence dataset that is why we could only compute their performances on the validation set. While the area under the curve (AUC) of iSuc-PseAAC [33], iSuc-PseOpt [34], SuccinSite [35] and pSuc-Lys [57] could not be computed, that of SSEvol-Suc was calculated for 6-, 8- and 10-fold cross-validations.

As shown in Table 1, SSEvol-Suc represents a significant improvement over the four predictors: iSuc-PseAAC [33], iSuc-PseOpt [34], SuccinSite [35] and pSuc-Lys [57]. SSEvol-Suc outperformed the previous predictors in statistics such as sensitivity, accuracy and MCC. For instance, sensitivity, accuracy and MCC significantly improved by 47.8%, 21.7% and 60.3%, respectively, when compared to the highest value of each metric. These results clearly indicate a considerable improvement (i.e., an increase in succinylation prediction accuracy) over current predictors. It is worth noting that although the specificity (0.906) of SuccinSite [35] remained high, its sensitivity (0.302) was remarkably low, leaving approximately 70% of succinylation residues undetected. In addition, the AUC of SSEvol-Suc for 6-, 8- and 10-fold cross-validations was 0.941, 0.938 and 0.942, respectively (Fig 2). These AUC values show that the predictor performance was not significantly affected when 6- and 10-fold cross-validations were conducted. However, the AUC value tended to slightly decrease when 8-fold cross-validation was performed.

Furthermore, we randomly created 100 negative sets of 1,782 samples each and trained the AdaBoost classifier to properly sample the non-succinylation space. Nevertheless, the average statistical metrics for 6-, 8- and 10-fold cross-validations did not dramatically vary (S1 File).
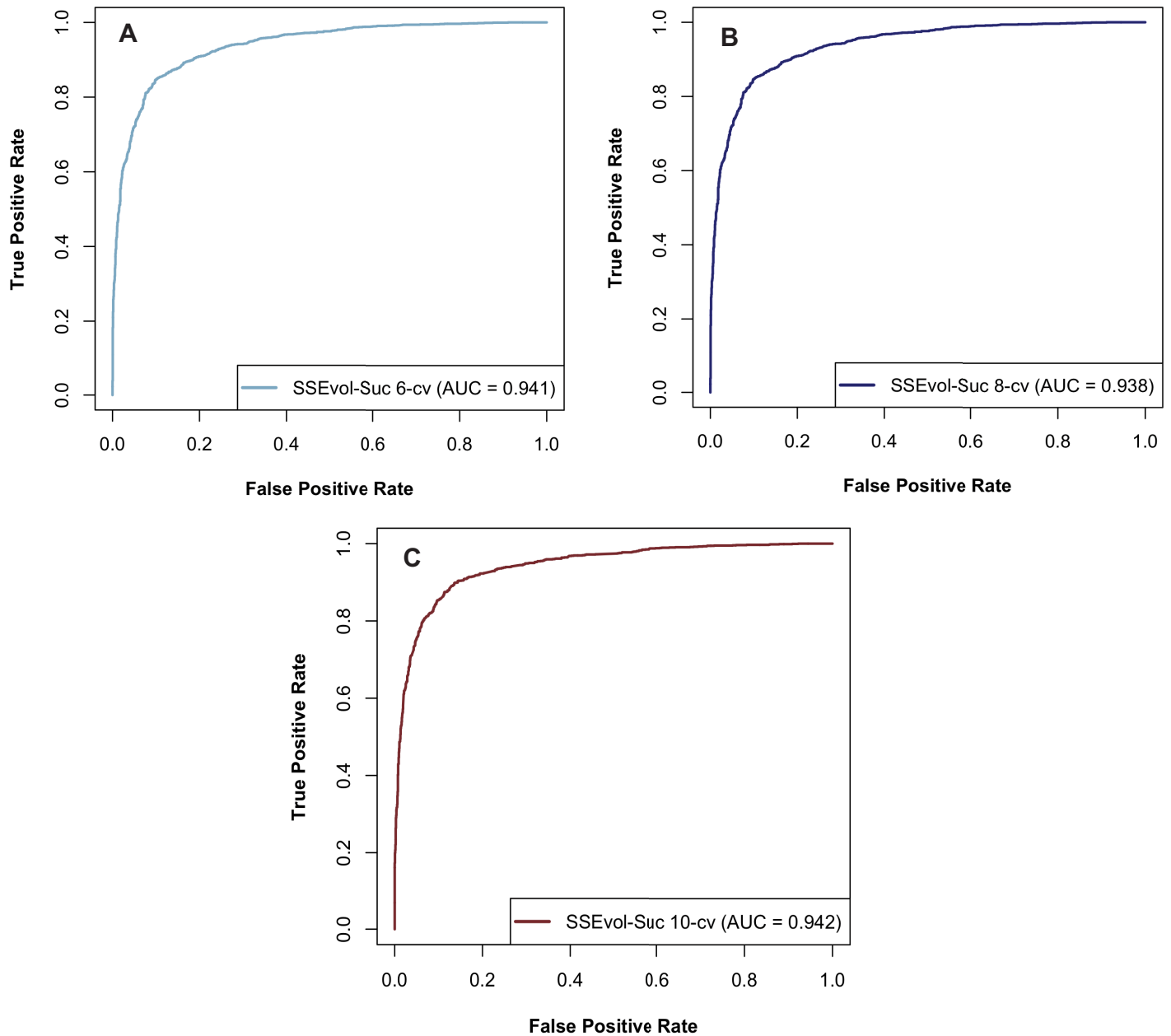
The above results illustrate the applicability of SSEvol-Suc when it comes to discriminating between succinylation and non-succinylation sites. These could be achieved by the effective combination of secondary structure and evolutionary information about proteins. The information on each peptide segment around a lysine was transformed into matrices of profile

Table 1. Comparison of SSEvol-Suc and state-of-the-art predictors.

Method	Sensitivity	Specificity	Accuracy	MCC	AUC
iSuc-PseAAC [33]	0.163	0.873	0.500	0.052	-
iSuc-PseOpt [34]	0.615	0.782	0.694	0.401	-
SuccinSite [35]	0.302	0.906*	0.588	0.258	-
pSuc-Lys [57]	0.587	0.866	0.719	0.468	-
SSEvol-Suc (6-CV)	0.900	0.835	0.870	0.739	0.941
SSEvol-Suc (8-CV)	0.905	0.836	0.872	0.745	0.938
SSEvol-Suc (10-CV)	0.909*	0.837	0.875*	0.750*	0.942*

*Highest value of this metric.

Fig 2. Receiver operating characteristic of SSEvol-Suc for (A) 6-, (B) 8- and (C) 10-fold cross-validations.

bigrams, and finally combined into a feature vector of transitional probabilities for classification purposes. This transformation appears to be essential to detect succinylated lysines and improve the sensitivity of SSEvol-Suc. Besides the AdaBoost classifier also contributed to such prediction outcomes. In summary, the use of one single vector, which combines *PSSM + bigram* and *SSpre + bigram*, seems to retain necessary information about lysine residues and therefore enables us to accurately detect succinylation sites.

Structural and evolutionary information has been previously considered in two computational predictors [40, 41]. For instance, SucStruct included the *SSpre* feature [40] whereas PSSM-Suc only regarded information about the *PSSM* [41]. Although evolutionary information allowed us

to discriminate lysines, better results are achieved when both types of characteristics are combined rather than independently used.

As stated in [85] and demonstrated in numerous studies [10–14, 17, 20–22, 25, 26, 29, 30, 42–44, 47, 86, 87], the availability of user-friendly web servers should be the next step in every computational predictor in order to enhance its impact [9]. To accomplish this, we will intend to build such a web server in the future so that the scientific community could significantly benefit from the proposed predictor.

Additional material related to this study can be downloaded from https://github.com/YosvanyLopez/SSEvol-Suc.

Conclusions

In this paper, we present a novel predictor called SSEvol-Suc which effectively uses a combination of *PSSM + bigram* and *SSpre + bigram* for predicting succinylated lysine residues. The secondary structure and evolutionary information about proteins was processed using profile bigrams and further integrated into a single vector for classification purposes. The k-nearest neighbors technique was utilized for removing redundant instances, which were finally input into an AdaBoost classifier for succinylation site prediction. When compared with previous approaches, the sensitivity, accuracy and MCC of the proposed predictor significantly improved by 47.8%, 21.7% and 60.3%, respectively. In spite of the significant performance of SSEvol-Suc, it is worth emphasizing that machine learning techniques do not help us understand why some lysines are succinylated and others are not. This is the main disadvantage of such techniques, which do not provide much scientific knowledge.

Supporting information

S1 File. Performance of the AdaBoost classifier on randomly created negative sets using 6-, 8- and 10-fold cross-validations.
(XLSX)

Acknowledgments

Author Contributions

Conceptualization: Abdollah Dehzangi, Yosvany López.

Formal analysis: Alok Sharma.

Funding acquisition: Tatsuhiko Tsunoda.

Investigation: Yosvany López, Tatsuhiko Tsunoda.

Methodology: Alok Sharma.

Supervision: Alok Sharma.

Writing – original draft: Yosvany López.

Writing – review & editing: Abdollah Dehzangi, Yosvany López, Sunil Pranit Lal, Ghazaleh Taherzadeh, Abdul Sattar, Tatsuhiko Tsunoda, Alok Sharma.

References

1. Knorre DG, Kudryashova NV, Godovikova TS. Chemical and Functional Aspects of Posttranslational Modification of Proteins. Acta Naturae. 2009; 1(3):29–51. PMID: 22649613

2. Comb DG, Sarkar N, Pinzino CJ. The Methylation of Lysine Residues in Protein. The Journal of Biological Chemistry. 1966; 241(8):1857–62. PMID: 5329588

3. Martin C, Zhang Y. The diverse functions of histone lysine methylation. Nature Reviews Molecular Cell Biology. 2005; 6(11):838–49. https://doi.org/10.1038/nrm1761 PMID: 16261189

4. 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. Nature Communications. 2014; 5:5409. https://doi.org/10.1038/ncomms6409 PMID: 25391492

5. 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.

6. Ansari NA, Moinuddin, Ali R. Glycated Lysine Residues: A Marker for Non-Enzymatic Protein Glycation in Age-Related Diseases. Disease Markers. 2011; 30(6):317–24. https://doi.org/10.3233/DMA-2011-0791 PMID: 21725160

7. Hirschey MD, Zhao Y. Metabolic Regulation by Lysine Malonylation, Succinylation, and Glutarylation. Molecular & Cellular Proteomics. 2015; 14(9):2308–15.

8. 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. https://doi.org/10.1038/nchembio.495 PMID: 21151122

9. Chou K-C. An Unprecedented Revolution in Medicinal Chemistry Driven by the Progress of Biological Science. Current Topics in Medicinal Chemistry. 2017; 17(21):2337–58. https://doi.org/10.2174/1568026617666170414145508 PMID: 28413951

10. Xu Y, Ding J, Wu L-Y, Chou K-C. iSNO-PseAAC: Predict Cysteine S-Nitrosylation Sites in Proteins by Incorporating Position Specific Amino Acid Propensity into Pseudo Amino Acid Composition. PLoS ONE. 2013; 8(2):e55844. https://doi.org/10.1371/journal.pone.0055844 PMID: 23409062

11. Xu Y, Shao X-J, Wu L-Y, Deng N-Y, Chou K-C. iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins. PeerJ. 2013; 1:e171. https://doi.org/10.7717/peerj.171 PMID: 24109555

12. Qiu W-R, Xiao X, Lin W-Z, Chou K-C. iMethyl-PseAAC: Identification of Protein Methylation Sites via a Pseudo Amino Acid Composition Approach. BioMed Research International. 2014; 2014:947416. https://doi.org/10.1155/2014/947416 PMID: 24977164

13. Xu Y, Wen X, Shao X-J, Deng N-Y, Chou K-C. iHyd-PseAAC: Predicting Hydroxyproline and Hydroxylysine in Proteins by Incorporating Dipeptide Position-Specific Propensity into Pseudo Amino Acid Composition. International Journal of Molecular Sciences. 2014; 15(5):7594–610. https://doi.org/10.3390/ijms15057594 PMID: 24857907

14. Xu Y, Wen X, Wen L-S, Wu L-Y, Deng N-Y, Chou K-C. iNitro-Tyr: Prediction of Nitrotyrosine Sites in Proteins with General Pseudo Amino Acid Composition. PLoS ONE. 2014; 9(8):e105018. https://doi.org/10.1371/journal.pone.0105018 PMID: 25121969

15. Chen W, Feng P, Ding H, Lin H, Chou K-C. iRNA-Methyl: Identifying N6-methyladenosine sites using pseudo nucleotide composition. Analytical Biochemistry. 2015; 490:26–33. https://doi.org/10.1016/j.ab.2015.08.021 PMID: 26314792

16. Chou K-C. Impacts of Bioinformatics to Medicinal Chemistry. Medicinal Chemistry. 2015; 11(3):218–34. PMID: 25548930

17. Qiu W-R, Xiao X, Lin W-Z, Chou K-C. 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. 2015; 33(8):1731–42. https://doi.org/10.1080/07391102.2014.968875 PMID: 25248923

18. Chen W, Tang H, Ye J, Lin H, Chou K-C. iRNA-PseU: Identifying RNA pseudouridine sites. Molecular Therapy–Nucleic Acids. 2016; 5:e332. https://doi.org/10.1038/mtna.2016.37 PMID: 28427142

19. Jia J, Liu Z, Xiao X, Liu B, Chou K-C. iCar-PseCp: identify carbonylation sites in proteins by Monte Carlo sampling and incorporating sequence coupled effects into general PseAAC. Oncotarget. 2016; 7 (23):34558–70. https://doi.org/10.18632/oncotarget.9148 PMID: 27153555

20. 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. https://doi.org/10.1093/bioinformatics/btw387 PMID: 27354696

21. Qiu W-R, Sun B-Q, Xiao X, Xu Z-C, Chou K-C. iHyd-PseCp: Identify hydroxyproline and hydroxylysine in proteins by incorporating sequence-coupled effects into general PseAAC. Oncotarget. 2016; 7 (28):44310–21. https://doi.org/10.18632/oncotarget.10027 PMID: 27322424

22. Qiu W-R, Sun B-Q, Xiao X, Xu Z-C, Chou K-C. iPTM-mLys: identifying multiple lysine PTM sites and their different types. Bioinformatics. 2016; 32(20):3116–23. https://doi.org/10.1093/bioinformatics/btw380 PMID: 27334473

23. Qiu W-R, Xiao X, Xu Z-C, Chou K-C. iPhos-PseEn: Identifying phosphorylation sites in proteins by fusing different pseudo components into an ensemble classifier. Oncotarget. 2016; 7(32):51270–83. https://doi.org/10.18632/oncotarget.9987 PMID: 27323404

24. Xu Y, Chou K-C. Recent Progress in Predicting Posttranslational Modification Sites in Proteins. Current Topics in Medicinal Chemistry. 2016; 16(6):591–603. PMID: 26286211

25. 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. https://doi.org/10.1016/j.omtn.2017.03.006 PMID: 28624191

26. Liu L-M, Xu Y, Chou K-C. iPGK-PseAAC: Identify Lysine Phosphoglycerylation Sites in Proteins by Incorporating Four Different Tiers of Amino Acid Pairwise Coupling Information into the General PseAAC. Medicinal Chemistry. 2017; 13(6):552–9. https://doi.org/10.2174/1573406413666170515120507 PMID: 28521678

27. 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. Medicinal Chemistry. 2017; 13(8):734–43. https://doi.org/10.2174/1573406413666170623082245 PMID: 28641529

28. Qiu W-R, Jiang S-Y, Xu Z-C, Xiao X, Chou K-C. iRNAm5C-PseDNC: identifying RNA 5-methylcytosine sites by incorporating physical-chemical properties into pseudo dinucleotide composition. Oncotarget. 2017; 8(25):41178–88. https://doi.org/10.18632/oncotarget.17104 PMID: 28476023

29. Qiu W-R, Sun B-Q, Xiao X, Xu D, Chou K-C. iPhos-PseEvo: Identifying Human Phosphorylated Proteins by Incorporating Evolutionary Information into General PseAAC via Grey System Theory. Molecular Informatics. 2017; 36(5–6):1600010.

30. Xu Y, Wang Z, Li C, Chou K-C. iPreny-PseAAC: Identify C-terminal Cysteine Prenylation Sites in Proteins by Incorporating Two Tiers of Sequence Couplings into PseAAC. Medicinal Chemistry. 2017; 13(6):544–51. https://doi.org/10.2174/1573406413666170419150052 PMID: 28425870

31. Zhao X, Ning Q, Chai H, Ma Z. Accurate in silico identification of protein succinylation sites using an iterative semi-supervised learning technique. Journal of Theoretical Biology. 2015; 374:60–5. https://doi.org/10.1016/j.jtbi.2015.03.029 PMID: 25843215

32. Xu H-D, Shi S-P, Wen P-P, Qiu J-D. SuccFind: a novel succinylation sites online prediction tool via enhanced characteristic strategy. Bioinformatics. 2015; 31(23):3748–50. https://doi.org/10.1093/bioinformatics/btv439 PMID: 26261224

33. Xu Y, Ding Y-X, Ding J, Lei Y-H, Wu L-Y, Deng N-Y. iSuc-PseAAC: predicting lysine succinylation in proteins by incorporating peptide position-specific propensity. Scientific Reports. 2015; 5:10184. https://doi.org/10.1038/srep10184 PMID: 26084794

34. 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. Analytical Biochemistry. 2016; 497:48–56. https://doi.org/10.1016/j.ab.2015.12.009 PMID: 26723495

35. Hasan MM, Yang S, Zhou Y, Mollah MNH. SuccinSite: a computational tool for the prediction of protein succinylation sites by exploiting the amino acid patterns and properties. Molecular BioSystems. 2016; 12(3):786–95. https://doi.org/10.1039/c5mb00853k PMID: 26739209

36. Dehzangi A, Heffernan R, Sharma A, Lyons J, Paliwal K, Sattar A. Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou's general PseAAC. Journal of Theoretical Biology. 2015; 364:284–94. https://doi.org/10.1016/j.jtbi.2014.09.029 PMID: 25264267

37. Dehzangi A, Paliwal K, Lyons J, Sharma A, Sattar A. Proposing a highly accurate protein structural class predictor using segmentation-based features. BMC Genomics. 2014; 15(Suppl 1):S2.

38. Sharma A, Dehzangi A, Lyons J, Imoto S, Miyano S, Nakai K, et al. Evaluation of Sequence Features from Intrinsically Disordered Regions for the Estimation of Protein Function. PLOS ONE. 2014; 9(2): e89890. https://doi.org/10.1371/journal.pone.0089890 PMID: 24587103

39. 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. https://doi.org/10.1002/jcc.23718 PMID: 25212657

40. López 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. Analytical Biochemistry. 2017; 527:24–32. https://doi.org/10.1016/j.ab.2017.03.021 PMID: 28363440

41. Dehzangi A, López 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. https://doi.org/10.1016/j.jtbi.2017.05.005 PMID: 28483566

42. Chen W, Feng P, Yang H, Ding H, Lin H, Chou K-C. iRNA-AI: identifying the adenosine to inosine editing sites in RNA sequences. Oncotarget. 2017; 8(3):4208–17. https://doi.org/10.18632/oncotarget.13758 PMID: 27926534

43. 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. https://doi.org/10.1093/bioinformatics/btw539 PMID: 27531102

44. 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. https://doi.org/10.1016/j.omtn.2017.04.008 PMID: 28624202

45. Niu B, Zhang M, Du P, Jiang L, Qin R, Su Q, et al. Small molecular floribundiquinone B derived from medicinal plants inhibits acetylcholinesterase activity. Oncotarget. 2017; 8(34):57149–62. https://doi.org/10.18632/oncotarget.19169 PMID: 28915661

46. Su Q, Lu W, Du D, Chen F, Niu B, Chou K-C. Prediction of the aquatic toxicity of aromatic compounds to tetrahymena pyriformis through support vector regression. Oncotarget. 2017; 8(30):49359–69. https://doi.org/10.18632/oncotarget.17210 PMID: 28467816

47. Xiao X, Cheng X, Su S, Mao Q, Chou K-C. pLoc-mGpos: Incorporate Key Gene Ontology Information into General PseAAC for Predicting Subcellular Localization of Gram-Positive Bacterial Proteins. Natural Science. 2017; 9(9):330–49.

48. Chou K-C. Some remarks on protein attribute prediction and pseudo amino acid composition. Journal of Theoretical Biology. 2011; 273(1):236–47. https://doi.org/10.1016/j.jtbi.2010.12.024 PMID: 21168420

49. Chou K-C. Prediction of Protein Cellular Attributes Using Pseudo-Amino Acid Composition. PROTEINS: Structure, Function, and Genetics. 2001; 43(3):246–55.

50. Saini H, Raicar G, Sharma A, Lal S, Dehzangi A, Lyons J, et al. Probabilistic expression of spatially varied amino acid dimers into general form of Chou's pseudo amino acid composition for protein fold recognition. Journal of Theoretical Biology. 2015; 380:291–8. https://doi.org/10.1016/j.jtbi.2015.05.030 PMID: 26079221

51. Chen W, Lei T-Y, Jin D-C, Lin H, Chou K-C. PseKNC: A flexible web server for generating pseudo K-tuple nucleotide composition. Analytical Biochemistry. 2014; 456:53–60. https://doi.org/10.1016/j.ab.2014.04.001 PMID: 24732113

52. Liu B, Liu F, Wang X, Chen J, Fang L, Chou K-C. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nucleic Acids Research. 2015; 43 (Web Server issue):W65–W71. https://doi.org/10.1093/nar/gkv458 PMID: 25958395

53. Liu B, Wu H, Chou K-C. Pse-in-One 2.0: An Improved Package of Web Servers for Generating Various Modes of Pseudo Components of DNA, RNA, and Protein Sequences. Natural Science. 2017; 9(4):67–91.

54. Dehzangi A, Paliwal K, Lyons J, Sharma A, Sattar A. Enhancing Protein Fold Prediction Accuracy Using Evolutionary and Structural Features. PRIB 2013: Pattern Recognition in Bioinformatics. 2013. pp 196–207.

55. Dehzangi A, Paliwal K, Lyons J, Sharma A, Sattar A. A Segmentation-Based Method to Extract Structural and Evolutionary Features for Protein Fold Recognition. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2014; 11(3):510–9. https://doi.org/10.1109/TCBB.2013.2296317 PMID: 26356019

56. Sharma A, Lyons J, Dehzangi A, Paliwal K. A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition. Journal of Theoretical Biology. 2013; 320:41–6. https://doi.org/10.1016/j.jtbi.2012.12.008 PMID: 23246717

57. Jia J, Liu Z, Xiao X, Liu B, Chou K-C. pSuc-Lys: Predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach. Journal of Theoretical Biology. 2016; 394:223–30. https://doi.org/10.1016/j.jtbi.2016.01.020 PMID: 26807806

58. Sharma A, Paliwal KK, Dehzangi A, Lyons J, Imoto S, Miyano S. A strategy to select suitable physicochemical attributes of amino acids for protein fold recognition. BMC Bioinformatics. 2013; 14:233. https://doi.org/10.1186/1471-2105-14-233 PMID: 23879571

59. Dehzangi A, Sharma A, Lyons J, Paliwal KK, Sattar A. A mixture of physicochemical and evolutionary-based feature extraction approaches for protein fold recognition. International Journal of Data Mining and Bioinformatics. 2015; 11(1):115–38. PMID: 26255379

60. Paliwal KK, Sharma A, Lyons J, Dehzangi A. Improving protein fold recognition using the amalgamation of evolutionary-based and structural based information. BMC Bioinformatics. 2014; 15(Suppl 16):S12.

61. 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 Research. 2011; 39(Database issue):D1029–34. https://doi.org/10.1093/nar/gkq939 PMID: 21059677

62. Liu Z, Wang Y, Gao T, Pan Z, Cheng H, Yang Q, et al. CPLM: a database of protein lysine modifications. Nucleic Acids Research. 2014; 42(Database issue):D531–6. https://doi.org/10.1093/nar/gkt1093 PMID: 24214993

63. 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.

64. 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. https://doi.org/10.1038/srep11476 PMID: 26098304

65. Dehzangi A, Paliwal K, Lyons J, Sharma A, Sattar A. Exploring Potential Discriminatory Information Embedded in PSSM to Enhance Protein Structural Class Prediction Accuracy. PRIB 2013: Pattern Recognition in Bioinformatics. 2013. pp. 208–19.

66. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research. 1997; 25 (17):3389–402. PMID: 9254694

67. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Research. 2000; 28(1):235–42. PMID: 10592235

68. Paliwal KK, Sharma A, Lyons J, Dehzangi A. A Tri-Gram Based Feature Extraction Technique Using Linear Probabilities of Position Specific Scoring Matrix for Protein Fold Recognition. IEEE Transactions on NanoBioscience. 2014; 13(1):44–50. https://doi.org/10.1109/TNB.2013.2296050 PMID: 24594513

69. Dehzangi A, Sohrabi S, Heffernan R, Sharma A, Lyons J, Paliwal K, et al. Gram-positive and gram-negative subcellular localization using rotation forest and physicochemical-based features. BMC Bioinformatics. 2015; 16(Suppl 4):S1.

70. Sharma R, Dehzangi A, Lyons J, Paliwal K, Tsunoda T, Sharma A. Predict Gram-Positive and Gram-Negative Subcellular Localization via Incorporating Evolutionary Information and Physicochemical Features Into Chou's General PseAAC. IEEE Transactions on NanoBioscience. 2015; 14(8):915–26. https://doi.org/10.1109/TNB.2015.2500186 PMID: 26584499

71. Freund Y, Schapire RE. A Short Introduction to Boosting. Journal of Japanese Society for Artificial Intelligence. 1999; 14(5):771–80.

72. Freund Y, Schapire RE. Experiments with a New Boosting Algorithm. Thirteenth International Conference on Machine Learning. 1996.

73. Friedman J, Hastie T, Tibshirani R. Additive Logistic Regression: A Statistical View of Boosting. The Annals of Statistics. 2000; 28(2):337–407.

74. Dehzangi A, Paliwal K, Sharma A, Dehzangi O, Sattar A. A Combination of Feature Extraction Methods with an Ensemble of Different Classifiers for Protein Structural Class Prediction Problem. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2013; 10(3):564–75. https://doi.org/10.1109/TCBB.2013.65 PMID: 24091391

75. Dehzangi A, Phon-Amnuaisuk S, Dehzangi O. Enhancing Protein Fold Prediction Accuracy Using an Ensemble of Different Classifiers. Australian Journal of Intelligent Information Processing Systems. 2010; 10(4):32–40.

76. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA Data Mining Software: An Update. SIGKDD Explorations. 2009; 11(1):10–8.

77. Liu Z, Xiao X, Qiu W-R, Chou K-C. iDNA-Methyl: Identifying DNA methylation sites via pseudo trinucleotide composition. Analytical Biochemistry. 2015; 474:69–77. https://doi.org/10.1016/j.ab.2014.12.009 PMID: 25596338

78. Liu B, Fang L, Wang S, Wang X, Li H, Chou K-C. Identification of microRNA precursor with the degenerate K-tuple or Kmer strategy. Journal of Theoretical Biology. 2015; 385:153–9. https://doi.org/10.1016/j.jtbi.2015.08.025 PMID: 26362104

79. Ding H, Deng E-Z, Yuan L-F, Liu L, Lin H, Chen W, et al. iCTX-Type: A Sequence-Based Predictor for Identifying the Types of Conotoxins in Targeting Ion Channels. BioMed Research International. 2014; 2014:286419. https://doi.org/10.1155/2014/286419 PMID: 24991545

80. 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. https://doi.org/10.1080/07391102.2014.998710 PMID: 25513722

81. Chou K-C, Shen H-B. Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. Nature Protocols. 2008; 3(2):153–62. https://doi.org/10.1038/nprot.2007.494 PMID: 18274516

82. Alpaydin E. Introduction to Machine Learning. Third ed: The MIT Press; 2014.

83. Hajisharifi Z, Piryaiee M, Beigi MM, Behbahani M, Mohabatkar H. Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test. Journal of Theoretical Biology. 2014; 341:34–40. https://doi.org/10.1016/j.jtbi.2013.08.037 PMID: 24035842

84. Jia J, Liu Z, Xiao X, Liu B, Chou K-C. iPPBS-Opt: A Sequence-Based Ensemble Classifier for Identifying Protein-Protein Binding Sites by Optimizing Imbalanced Training Datasets. Molecules. 2016; 21(1): E95. https://doi.org/10.3390/molecules21010095 PMID: 26797600

85. Chou K-C, Shen H-B. Recent advances in developing web-servers for predicting protein attributes. Natural Science. 2009; 1(2):63–92.

86. Cheng X, Xiao X, Chou K-C. pLoc-mGneg: Predict subcellular localization of Gram-negative bacterial proteins by deep gene ontology learning via general PseAAC. Genomics. 2017 (In Press).

87. 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. 2017; 110(1): 50–8. https://doi.org/10.1016/j.ygeno.2017.08.005 PMID: 28818512