Genomics Proteomics Bioinformatics 18 (2020) 208–219

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

METHOD

SuccSite: Incorporating Amino Acid Composition and Informative *k*-spaced Amino Acid Pairs to Identify Protein Succinylation Sites

Hui-Ju Kao [1,#], Van-Nui Nguyen [2,#], Kai-Yao Huang [3,4], Wen-Chi Chang [5], Tzong-Yi Lee [3,4,*]

[1] *Department of Computer Science and Engineering, Yuan Ze University, Taoyuan 32003, Taiwan, China*
[2] *Department of Information Technology, University of Information and Communication Technology, Thai Nguyen 1000, Vietnam*
[3] *School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen 518172, China*
[4] *Warshel Institute for Computational Biology, The Chinese University of Hong Kong, Shenzhen 518172, China*
[5] *Institute of Tropical Plant Sciences, Cheng Kung University, Tainan 701, Taiwan, China*

Received 8 March 2018; revised 1 October 2018; accepted 11 October 2018
Available online 24 June 2020

Handled by Yu Xue

KEYWORDS

Protein succinylation;
Succinyl group;
Substrate specificity;
Amino acid composition;
k-spaced amino acid pair composition

Abstract **Protein succinylation** is a biochemical reaction in which a **succinyl group** (-CO-CH2-CH2-CO-) is attached to the lysine residue of a protein molecule. Lysine succinylation plays important regulatory roles in living cells. However, studies in this field are limited by the difficulty in experimentally identifying the substrate site specificity of lysine succinylation. To facilitate this process, several tools have been proposed for the computational identification of succinylated lysine sites. In this study, we developed an approach to investigate the **substrate specificity** of lysine succinylated sites based on **amino acid composition**. Using experimentally verified lysine succinylated sites collected from public resources, the significant differences in position-specific amino acid composition between succinylated and non-succinylated sites were represented using the Two Sample Logo program. These findings enabled the adoption of an effective machine learning method, support vector machine, to train a predictive model with not only the amino acid composition, but also the composition of *k*-spaced amino acid pairs. After the selection of the best model using a ten-fold cross-validation approach, the selected model significantly outperformed existing tools based on an independent dataset manually extracted from published research articles. Finally, the selected model was used to develop a web-based tool, SuccSite, to aid the study of protein succinylation. Two proteins were used as case studies on the website to demonstrate the effective prediction of succinyla-

* Corresponding author.
E-mail: leetzongyi@cuhk.edu.cn (Lee TY).
# Equal contribution.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

https://doi.org/10.1016/j.gpb.2018.10.010

tion sites. We will regularly update SuccSite by integrating more experimental datasets. SuccSite is freely accessible at http://csb.cse.yzu.edu.tw/SuccSite/.

Introduction

Post-translational modification (PTM) is a chemical form of regulation that occurs after protein translation. This post-translational regulation plays a vital role in a variety of cellular processes including signaling networks, protein degradation, gene transcriptional regulation, protein–protein interaction, and metabolic pathways. The attachment and removal of chemical groups catalyzed by enzymes underlie most PTMs [1]. Protein succinylation is the biochemical reaction in which a succinyl group (-CO-CH2-CH2-CO-) is attached to a lysine residue of a protein molecule. Succinyl coenzyme A (succinyl-CoA) is a cofactor for enzyme-mediated lysine succinylation [2,3]. Protein lysine succinylation plays important regulatory roles in living cells. For instance, a previous study [2] reported evidence for possible implications of docosahexomic acid (DHA) exposure in the central nervous system. In a related work [4], Kawai et al. demonstrated the ability of DHA to promote succinylation of lysine residues.

Mass spectrometry (MS), a high-throughput biotechnology, is widely utilized to determine a large amount of site-specific succinylated peptides [5,6]. Due to the labor-intensive experiments of MS-based proteomics in identifying succinylated sites, there is an increasing number of computational tools dedicated to predicting potential succinylated lysine residues for further functional analyses [7–10]. Succinylation is a site-specific modification that mainly occurs on lysine residues. The process requires substrate specificity—recognition of a succinylated site in accordance with the composition of surrounding residues. Therefore, increasing the precision of succinylation site prediction requires the detailed characterization of substrate specificity.

Table 1 provides a summary of published methods used for the identification of lysine succinylation sites based on protein sequences. For example, a web-based tool developed by Zhao et al. [8] incorporated support vector machine (SVM) with multiple feature-encoding schemes to identify succinylated sites. The iSuc-PseOpt [9] incorporates the random forest algorithm with *k*-nearest neighbors cleaning (KNNC) [11] and the Included Hypothetical Training Samples (IHTS) for the identification of protein succinylation sites. In addition, Xu et al. [7] developed an approach to predict succinylated lysines according to the biochemical property that PTM prefers a specific composition of amino acids around the substrate site. Various features, such as amino acid composition, a flexibility scalar, aromatic content, the net charge, beta entropy, hydrophobic moment, disorder information, and position-specific scoring matrix (PSSM), were also investigated [12]. Despite several approaches and tools for the identification of protein succinylation sites, the number, quality, and performance of datasets were insufficient to meet current demand. Moreover, the recent advancements in high-throughput biotechnologies increased the available data of experimentally verified succinylated sites. Therefore, we were motivated to develop a new approach for identifying protein succinylation sites primarily using the composition of amino acids [13] and the informative *k*-spaced amino acid pairs (KSAAPs) [14]. Additionally, other sequence-based features such as the composition of dipeptides [15] are also taken into account. After the selection of the best model based on the evaluation of cross-validation, the proposed model can provide a better independent testing result than existing online tools. Finally, the proposed model has been adopted to implement a web-based predictor, SuccSite, to accelerate the practical applications for functional proteomics.

Method

Data collection and preprocessing

With the advent of high-throughput MS or MS/MS-based proteomics in protein succinylation, numerous resources were developed for compiling experimentally confirmed PTM peptides including lysine succinylated sites based on manual

Table 1 Summary of the training datasets and learning methods of existing tools for the prediction of protein succinylation sites

Tool	Datasets	Method	Ref.
iSuc-PseAAC	CPLM, UniProtKB	SVM	[7]
SucPred	CPLM	SVM	[8]
iSuc-PseOpt	CPLM, UniProtKB	Random forest	[9]
pSuc-Lys	CPLM, UniProtKB	Random forest	[10]
SuccinSite	CPLM	SVM	[22]
SuccFind	CPLM, UniProtKB	SVM	[23]

Note : SVM, support vector machine. The method proposed in the current study is highlighted in bold.

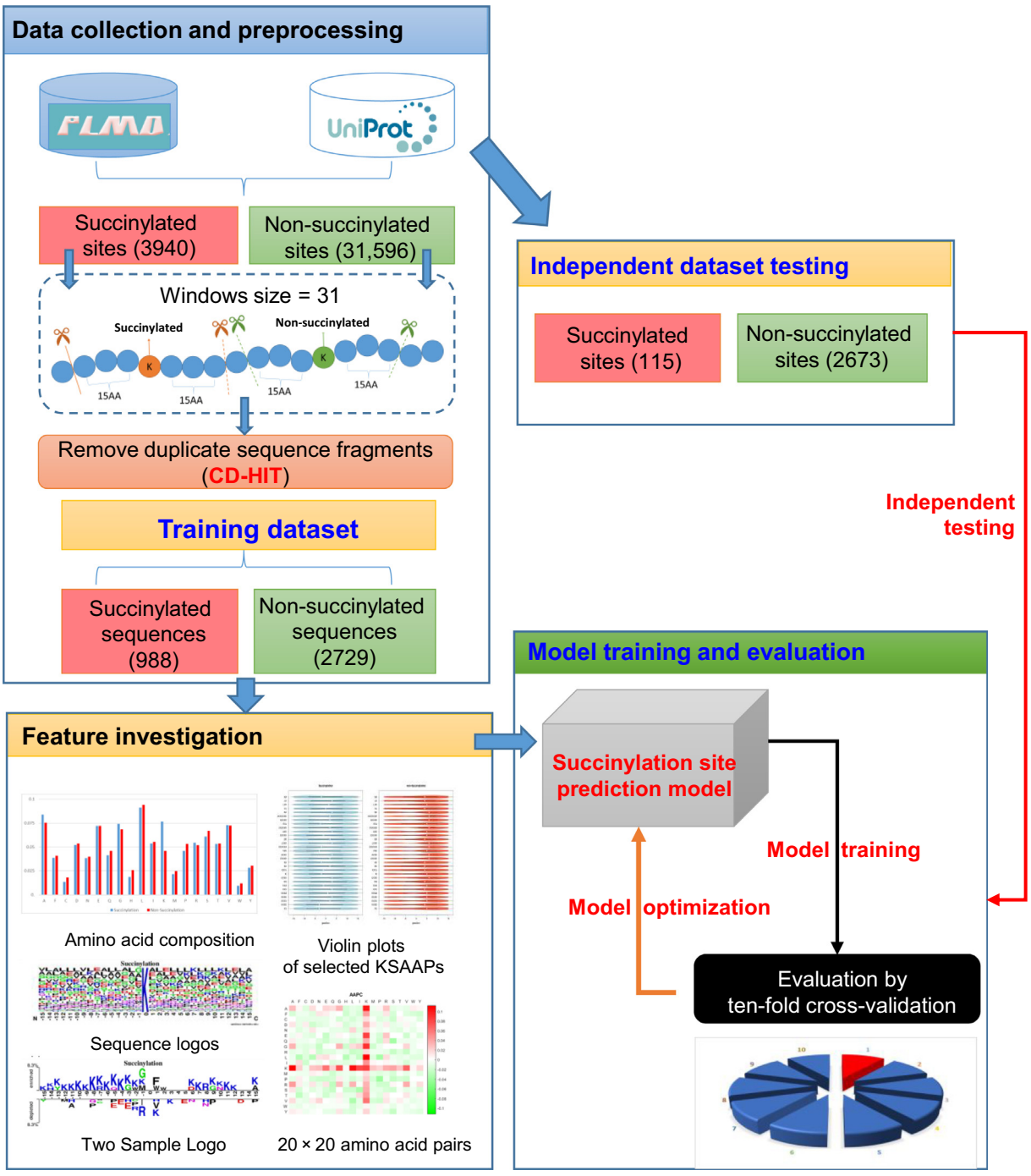


Figure 1 Flowchart of protein succinylation site prediction in this work

There are four major steps, including data collection and preprocessing, feature investigation, model training and evaluation, as well as independent dataset testing.

curation of MS/MS-related literature [16–18]. As presented in Figure 1, the experimentally verified lysine succinylated site dataset was extracted from UniProtKB [19] and CPLM [18]. In UniProtKB, the succinylated proteins are collected and filtered to remove non-experimental entries represented by the Evidence Code Ontology (ECO) codes "0000305", "0000250", "0000255", and "0000256" [20]. This results in 1382 experimentally verified lysine succinylated sites from 419 proteins. The CPLM database 2.0 [18] has 189,919 modified lysines from 45,748 proteins for 12 different types of lysine modifications. With the consideration of only experimentally confirmed lysine succinylated sites, 2558 sites are collected from 897 unique proteins. As a result, 1316 proteins with 3940 experimentally verified lysine succinylated sites are collected from UniProtKB and CPLM. After removing the duplicated and redundant data using the CD-HIT program [21] with a cut-off threshold of 40%, we obtain 1169 unique proteins with 2509 lysine succinylated sites (positive data). In order to prepare the training and independent testing datasets for model training and evaluation, respectively, we randomly

select 1000 unique proteins for the training dataset. The remaining 169 unique proteins are the independent testing dataset.

For the identification of lysine succinylation sites, a $(2n + 1)$-mer window size was adopted to extract fragmented sequences centered on modified sites with n left-hand and n right-hand neighboring amino acids. Given a specific number of succinylated proteins, the negative dataset is generated from non-succinylated sites, which are those fragmented sequences centered on lysine residues without annotation of succinylation. Related works [7–10,22,23] revealed that models trained using a 31-mer window size ($n = 15$) perform best in the prediction of lysine succinylation sites. Owing to the possibility of over-fitting originating from the training dataset, the predictive power of the generated models might be overestimated. Thus, the independent testing dataset, which is blind to the training dataset, is necessary for further evaluation of real cases. In addition, the fragmented sequences may be homologous among datasets used for model training. Therefore, the CD-HIT software is used again to eliminate fragmented sequences with high similarity between the training and testing datasets. As displayed in Table S1, based on a sequence identity threshold of 40%, the final dataset for model training consists of 998 positive and 2729 negative instances; the final dataset for independent testing contained 115 positive and 2673 negative instances. In this work, the positive and negative testing data are further utilized to compare the proposed model with other prediction schemes for predictive performance.

Amino acid composition

This study focuses on the sequence-based characterization of substrate site specificity for protein succinylation. Amino acid composition (AAC) is a typical attribute used to examine substrate site motifs. AAC determines the probability of amino acids occurring in the flanking region of PTM sites [24]. Given a fragmented sequence x with a 31-mer string length, $n_x(k)$ is the number of a specific amino acid, k, occurring in the fragment, where k denotes the 20 amino acids. Consequently, the probability $P_x(k)$ of specific amino acid k is [25]

$$P_x(k) = \frac{n_x(k)}{\sum_{k=1}^{20} n_x(k)} \quad k = 1, 2, \cdots, 20 \tag{1}$$

Then, the composition of the 20 amino acids can be transformed to a 20-dimensional numeric vector V_x for the fragmented sequence x:

$$V_x = [P_x(1), P_x(2), \cdots, P_x(20)] \tag{2}$$

In order to observe the position-specific AAC for lysine succinylated sites, WebLogo [26] is utilized to create frequency plots of sequence logos for visualizing the potential amino acid motifs surrounding succinylated sites (at position 0). In addition, a web-based program, Two Sample Logo [27], is adopted to further discover the differences in the position-specific composition of amino acids between succinylated sites (positive data) and non-succinylated sites (negative data).

Amino acid pair composition

The dipeptides surrounding the succinylated sites are explored by calculating the occurring probability of each amino acid pair (dipeptide) around the substrate sites. The probabilities of 400 dipeptides are compared between succinylated and non-succinylated data to determine the significant dipeptides for model construction. For a fragmented sequence x, $f_k(x)$ represents the occurrence frequency of a specific amino acid pair. The occurrence frequency of an amino acid pair $pk(x)$ is defined as follows:

$$pk(x) = \frac{f_k(x)}{\sum_{i=1}^{400} f_i(x)} \quad i, k = 1, 2, \cdots, 400 \tag{3}$$

Then, the composition of the 400 amino acid pairs for a fragmented sequence x is

$$P(x) = [p_1(x), p_2(x), \cdots, p_{400}(x)] \tag{4}$$

In order to provide a better observation of amino acid pair composition (AAPC), a 20×20 matrix is illustrated with red and green colors to represent over-expression and under-expression of dipeptides, respectively, around succinylated sites. Along with generating sequence logos for lysine succinylated sites and creating the heatmap of AAPC between succinylated and non-succinylated sites, the different value of each amino acid pair as well as its P value are determined. All the dipeptides are ranked according to their P values, and the dipeptides having a probability difference value >0.02 and a P value <0.05 are selected as significant attributes for the classification between positive and negative instances.

Composition of informative KSAAPs

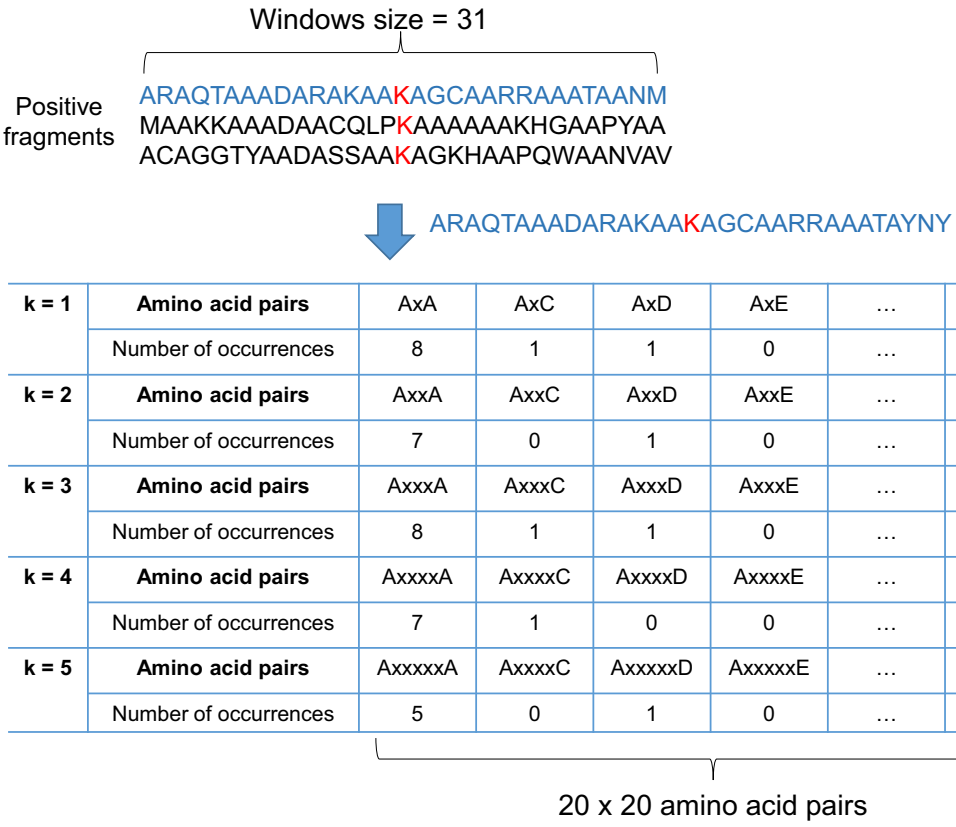
In recent years, the composition of k-spaced amino acid pairs (CKSAAPs) [28–30], represented as a numeric vector in the n-dimensional Gaussian space feature, has been widely adopted for the prediction of functional sites on proteins [14,16,24,31–36]. In this work, we utilize the CKSAAP-based encoding scheme to transform the fragmented sequences into n-dimensional numeric vectors. The CKSAAPs are extracted from the flanking amino acid sequences of succinylated sites. As presented in **Figure 2**, when $k = 1$, $[A_i x A_j]$ denotes the pair of amino acids A_i ($i = 1, \ldots, 20$, corresponding to 20 amino acids) and A_j ($j = 1, \ldots, 20$, corresponding to 20 amino acids) that are separated by one amino acid; when $k = 2$, $[A_i x x A_j]$ denotes the pair of amino acids A_i and A_j that are separated by two amino acids. Thus, $N([A_i x A_j])$ is the number of occurrences of the one-spaced amino acid pair $[A_i x A_j]$ in the training dataset and the conditional probability $P[A_i x A_j]$ is

$$P[A_i x A_j] = \frac{N([A_i x A_j])}{N([A_i x A_*])}, \tag{5}$$

where $N([A_i x A_*]) = \sum_{j=1,\cdots,20} N([A_i x A_j])$. The strength of the one-spaced amino acid pair $[A_i x A_j]$ between positive and negative datasets is given by

$$C[A_i x A_j] = \log \frac{P^+[A_i x A_j]}{P^-[A_i x A_j]}, \tag{6}$$

where $P^+[A_i x A_j]$ and $P^-[A_i x A_j]$ are the conditional probabilities of the one-spaced amino acid pair $[A_i x A_j]$ in positive and negative datasets, respectively. If $C[A_i x A_j] > 0$, then $[A_i x A_j]$



k = 1	**Amino acid pairs**	AxA	AxC	AxD	AxE	…	YxY
	Number of occurrences	8	1	1	0	…	1
k = 2	**Amino acid pairs**	AxxA	AxxC	AxxD	AxxE	…	YxxY
	Number of occurrences	7	0	1	0	…	0
k = 3	**Amino acid pairs**	AxxxA	AxxxC	AxxxD	AxxxE	…	YxxxY
	Number of occurrences	8	1	1	0	…	0
k = 4	**Amino acid pairs**	AxxxxA	AxxxxC	AxxxxD	AxxxxE	…	YxxxxY
	Number of occurrences	7	1	0	0	…	0
k = 5	**Amino acid pairs**	AxxxxxA	AxxxxC	AxxxxxD	AxxxxxE	…	YxxxxxY
	Number of occurrences	5	0	1	0	…	0

Figure 2 Composition of k-spaced amino acid pairs

Given 400 (20 × 20) amino acid pairs and five values for k (k = 1–5), there are 2000 attributes generated for the CKSAAP feature. The number of occurrences of each k-spaced amino acid pair is determined by sliding through the fragmented sequence one by one. CKSAAP, composition of k-spaced amino acid pair.

is enriched in the positive dataset; otherwise, the $[A_i x A_j]$ is depleted in the positive dataset if $C[A_i x A_j] < 0$. The high value of $|C[A_i x A_j]|$ indicates that the one-spaced amino acid pair $[A_i x A_j]$ is the more significant attribute for classifying between positive and negative datasets. Applying this approach, which is similar to previous work [28–30], for CKSAAP features, this study has examined the KSAAPs with k ranging from one to five. Given 20 × 20 amino acid pairs and five values for k, 2000 attributes are used to train the predictive model. However, the higher dimensions of feature vectors could induce a lower efficiency of model learning and evaluation. Therefore, all 2000 features should be tuned to obtain optimal CKSAAPs for providing better predictive performance.

In order to extract informative features prior to the construction of the predictive model, each attribute (*e.g.*, KSAAPs) is evaluated according to the index score calculated by the minimum redundancy–maximum relevance (mRMR) [37] algorithm, which ranks all attributes according to each attribute's relevance value corresponding to the dataset as well as each attribute's redundancy index among all 2000 examined KSAAPs. An attribute having maximum relevance and minimum redundancy will contain the best discriminating power between positive and negative instances [38]. The scoring function of mRMR is

$$score_j = I(f_j, c) - \frac{1}{m}\sum_{i=1}^{m} I(f_i, f_j) \quad (7)$$

wherein $f_j \subset S_n$, $f_i \subset S_m$, $S_m = S - S_n$ in which S_m, S_n, and S are the attribute sets (m and n were the attribute sizes). The classification variable c stands for two classes corresponding to positive and negative datasets in this work. Additionally, the mutual information $I(x, y)$ is defined as

$$I(x, y) = p(x, y)\log\frac{p(x, y)}{p(x)p(y)}dxdy, \quad (8)$$

where $p(x, y)$, $p(x)$, and $p(y)$ are regarded as the probabilistic density functions between attributes x and y. Herein, all the KSAPPs were examined by the mRMR criteria. Furthermore, an incremental strategy for extracting useful features, called sequential forward selection (SFS), is applied to conduct a final CIKSAAP with best predictive performance. There are five main steps in this investigation:

1. Choose a classifier (*e.g.*, SVM) and determine an evaluation benchmark (*e.g.*, ten-fold cross-validation).
2. Among all unselected attributes, choose the attribute (KSAAPs) with lowest mRMR index score and combine it into the set of selected attributes.
3. Construct the classifier based on the set of selected attributes.
4. Evaluate the predictive performance of the constructed classifier based on the evaluation benchmark.
5. Repeat steps 2–4 until a sufficient number (default 30) of attributes has been selected, or until predictive performance has been optimized.

Model construction for succinylation site prediction

This study involves a binary classification between succinylated and non-succinylated sites on lysine residues. The positive and negative datasets are labeled with $+1$ and -1, respectively, for the two classes. The training dataset is $X = \{x^t, c^t\}$ where $c^t = +1$ if $x^t \in$ positive dataset and $c^t = -1$ if $x^t \in$ negative dataset. Thus, w and w_0 were identified such that

$$w^T x^t + w_0 \geq +1 \; for \; c^t = +1 \text{ and } w^T x^t + w_0 \leq -1 \; for c^t = -1,$$

which can be rewritten as

$$c^t(w^T x^t + w_0) \geq +1$$

This can be used to find an optimal separating hyperplane that can maximize the margin between the two classes [39]. The distance of x^t to the discriminating hyperplane is

$$\frac{|w^T x^t + w_0|}{\|w\|}$$

and the distance should be higher than a specific value h:

$$\frac{c^t(w^T x^t + w_0)}{\|w\|} \geq h, \; \forall t \; and \; c^t \in \{+1, -1\}$$

The SVM is an advanced algorithm used to identify a hyperplane between two classes with maximum margin based on n-dimensional vector space [39] with an attempt to maximize h, however, an unlimited number of possible values could be elucidated by tuning w. Hence, the $h\|w\|$ is defined as one to minimize $\|w\|$ using the following solution [35]:

$$min \frac{1}{2}\|w\|^2 \; subject \; to \; c^t\left(w^T x^t + w_0\right) \geq +1, \forall t$$

The SVM can determine a hyperplane for discriminating between succinylated and non-succinylated instances with maximal margin in a vector space containing n dimensions (size of attribute set). The sequence-based features are transformed into numeric vectors in an n-dimensional vector space, which are the input values for SVM. A SVM public resource, LIBSVM [40], has been downloaded and installed for iterative training of multiple SVMs in accordance with various feature sets. In the machine learning problem, if the best discriminant is nonlinear, instead of enabling nonlinear modeling, we can map all n-dimensional vectors to new vector space with higher dimension m, where $m > n$, using nonlinear kernel functions. As demonstrated in previous methods [31,41–44], the radial basis function (RBF) is typically chosen as the specified kernel function on learning for SVM models. The RBF function is as follows:

$$K(x^t, x) = \exp\left[-\frac{\|x^t - x\|^2}{2s^2}\right]$$

where x^t is the center and s is the radius, which should be provided by the programmer. When using LIBSVM, cost (c) and gamma (r) are two supporting parameters used to optimize the radius of kernel function and softness of hyperplane, respectively. To achieve the feasible values of gamma (r) and cost (c) in model learning, an optimization program, written in Python, was provided by LIBSVM.

Performance measurement

In this work, the ten-fold cross-validation (10-fold CV) method is performed to measure classifying power of the constructed SVM models. In 10-fold CV, all positive and negative training instances are split into ten subsets with approximately equal data size. After obtaining ten subsets, nine are used as the training dataset, whereas the remaining one subset is used as the test dataset. Each subset, selected from the ten subsets, is regarded as the test dataset until all ten subsets are tested in a 10-fold CV. The performance of the trained models is estimated according to the following metrics:

$$\text{Sensitivity (Sn)} = \frac{TP}{TP + FN},$$

$$\text{Specificity (SP)} = \frac{TN}{TN + FP},$$

$$\text{Accuracy (Acc)} = \frac{TP + TN}{TP + FP + TN + FN},$$

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

in which the predictions of true positives, false negatives, true negatives, and false positives are denoted as TP, FN, TN, and FP, respectively. Accuracy is typically chosen as a benchmark for determining the best predictive model. However, in this investigation, accuracy is not a good benchmark because the size of positive and negative datasets is skewed [36]. Given unbalanced positive and negative datasets in this study, the MCC is used as a reasonable benchmark for taking both prediction rate of TP (sensitivity) and prediction rate of TN (specificity) into account. After 10-fold CV evaluation, the SVM classifier containing the best MCC value is regarded as the best predictor. Finally, a testing dataset, which is independent from the training dataset, is utilized to examine the best model and compare the predicted results with other available online tools.

Results and discussion

Composition of amino acids and dipeptides around succinylated sites

The AAC is a feasible scheme to explore the potential motif of conserved residues around the succinylated sites based on the fragments with 31-mer sequence length. When comparing the AAC between positive and negative datasets, the residues having significant differences are useful attributes for succinylated site prediction. **Figure 3** shows that, for succinylated sites, the positively charged lysine residue appears to have the highest frequency around the substrate sites. In addition to AAC, the position-specific AAC neighboring the succinylated sites can be displayed using the frequency plots of WebLogo [26]. As illustrated in **Figure 4A**, no amino acid has significantly high frequency near the succinylated sites, but the slightly prominent amino acid residues include leucine, lysine, alanine, and valine. Without conserved motifs observed in the frequency plot, the Two Sample Logo [27] program was further

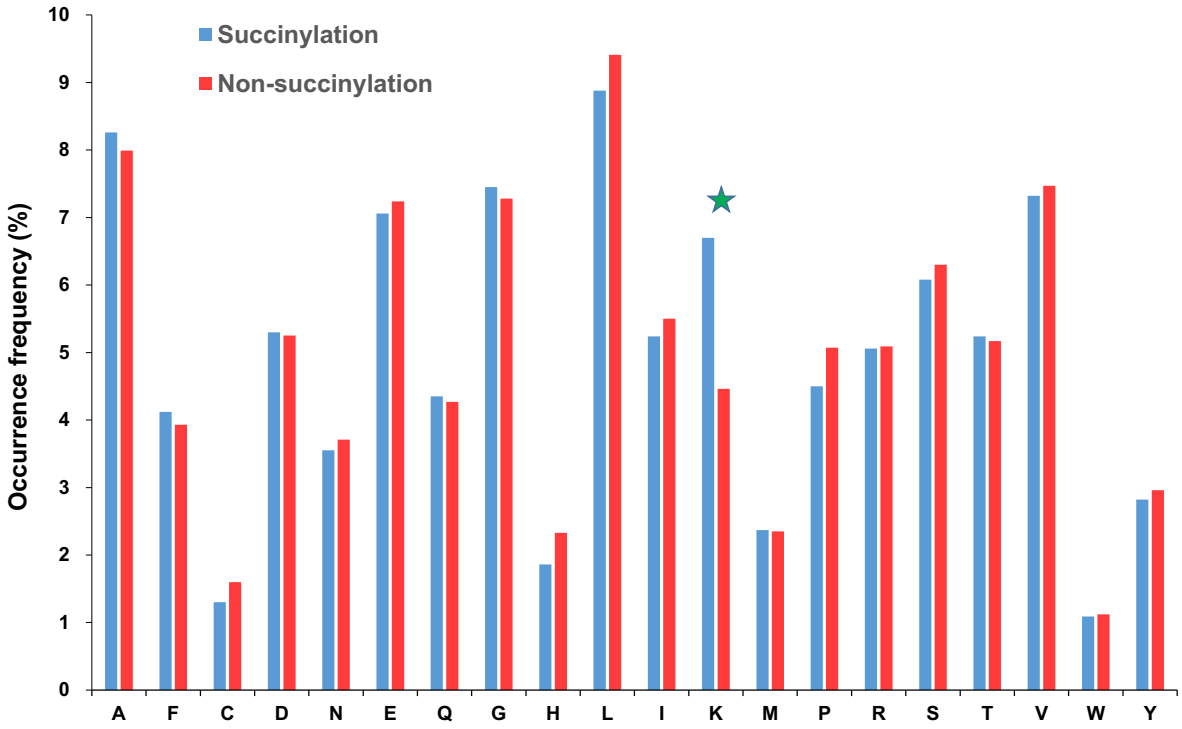


Figure 3 Comparison of amino acid composition between succinylated sites and non-succinylated sites
This investigation shows that the positively charged lysine residue (indicated with star) is abundant within the neighborhood of succinylated sites (in blue), compared to non-succinylated sites (in red).

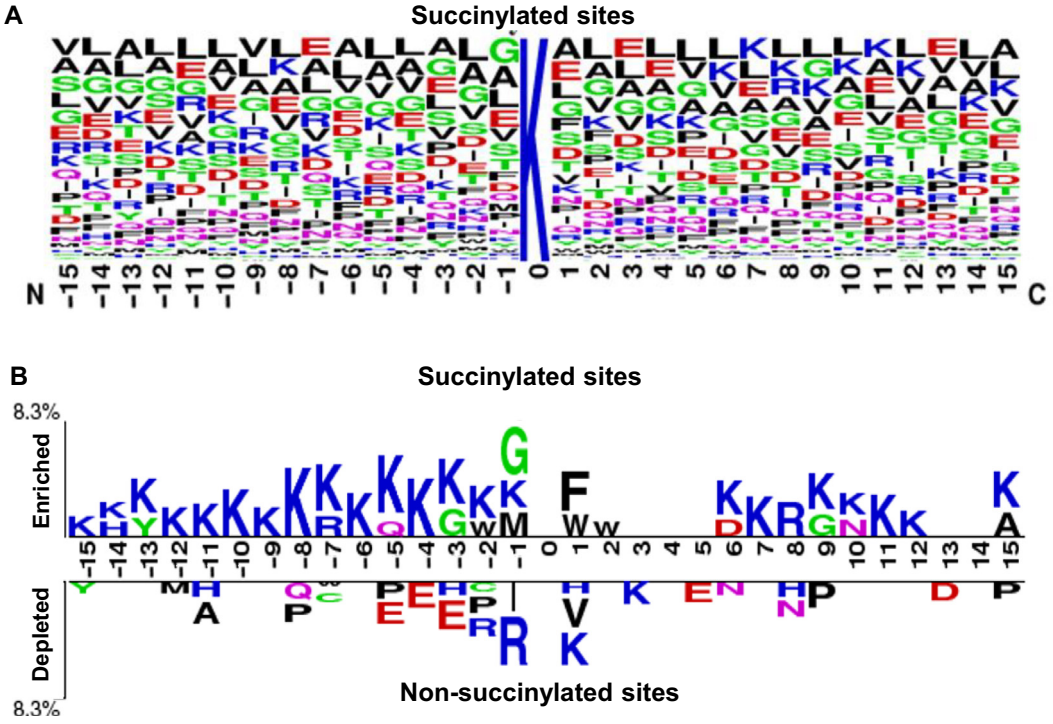


Figure 4 Position-specific amino acid composition of succinylated sites
A. Frequency plot of substrate sequences. **B.** Two Sample Logo of substrate sequences.

applied to compare the differences of position-specific AAC between flanking regions of succinylated and non-succinylated sites. As displayed in Figure 4B, the most conserved motifs appear to be associated with charged residues, in particular the positively charged lysine residue at positions −15~−1, +6~+12. In addition, the depletion of negatively charged amino acids, such as glutamic acid, at positions −5, −4, −3, and +5, is predictive. The results reveal that amino acids situated further away in the sequence but closer in the three-dimensional structure to the succinylated sites had notable differences between succinylated sites and non-succinylated sites.

The amino acid pairs surrounding lysine succinylated sites are also explored using the detection of remarkable amino acid pairs with significant differences between positive and negative datasets. In this investigation, a 20 × 20 matrix is adopted to represent the over-expressed and under-expressed amino acid pairs as red and green colors, respectively. As shown in Figure S1, the dipeptides involving a lysine residue in first position, such as KA, KE, KG, and KL, are over-represented around succinylated sites. Interestingly, the dipeptides involving lysine residue in the second position, such as AK, GK, and LK, are also over-represented around succinylated sites. By sorting the amino acid pairs according to their P values, the dipeptides with $P < 0.05$ and with a probability difference >2% are extracted and combined into an attribute set with statistical significance.

Investigation of informative k-spaced amino acid pairs

To support the identification of lysine succinylation sites, we have counted and ranked the frequencies of all k-spaced amino acid pairs that appeared in the positive and negative training datasets. In this study, top 30 significant KSAAPs, based on sequential forward selection in accordance with their mRMR scores, are selected for the identification of succinylated sites. Figure S2 provides violin plots of selected KSAAPs with their corresponding distributions, ranging from −15 to 15, around succinylated and non-succinylated sites (position 0). This investigation has indicated that most of the selected KSAAPs prefer to locate in the upstream and downstream regions of succinylated sites, whereas in non-succinylated sites these selected KSAAPs have concentrated distribution only in the downstream region. For instance, KA prefers to locate in the upstream and downstream regions of succinylated sites, but for the non-succinylated sites it only prefers to locate in the downstream region. Due to the difference of KSAAPs' distributions between succinylated and non-succinylated sequences, these top 30 KSAPPs are then incorporated into the construction of SVM models.

Cross-validation evaluation of characteristics flanking succinylated sites

To obtain the optimal window lengths that generate the best performance, we have investigated and assessed various window lengths using 10-fold CV. In accordance with the difference of position-specific AACs between succinylated and non-succinylated sites as well as our preliminary evaluation, the window size of 31 (−15 ~ +15; with the centered residue at lysine) provides the best performance in the prediction. Based on the investigated features, the corresponding SVM models are built to determine the effectiveness of those features in the identification of succinylated sites. As displayed in Table S2, the AAC-based SVM model has 64.6% accuracy and an MCC value of 0.27. The AAPC-based SVM model yields 63.2% accuracy and an MCC value of 0.24. In addition, the CKSAAP-based SVM model ($K = 5$) obtains 61.9% accuracy and an MCC value of 0.22.

In a binary classification between succinylated and non-succinylated sites, it is feasible to incorporate two or more different attribute sets in modeling. Therefore, in addition to the single attribute set, a hybrid combination of different attribute sets are also considered in this study. Based on the three attribute sets that were investigated (AAC, AAPC, and CKSAAP), four combinations ("AAC + AAPC", "AAC + CIKSAAP", "AAPC + CIKSAAP", and "AAC + AAPC + CIKSAAP") are analyzed for the identification of succinylated sites. Table S2 presents the performance of the hybrid features-based models when evaluated using 10-fold CV. The results reveal that most hybrid features-based models can obtain better performance, wherein the "AAC + CIKSAAP"-based model performs the best, with 71.4% accuracy and an MCC value of 0.40. Thus, the hybrid features of AAC and CIKSAAP yield the most promising predictions. In addition, the ROC curve is generated to compare different predictive models. As displayed in Figure S3, the results show the SVM model trained from the combination of AAC and CKSAAP attribute sets gave the best predictive power.

Evaluation of the selected models using independent test set

In the classification between positive and negative instances, there is a potential risk to over-estimate the predictive performance because of over-fitting in model training. Therefore, an independent testing dataset was constructed to assess the model's ability and stability in practice. As mentioned previously, the independent testing dataset comprised 115 positive and 2673 negative sites. To assess the practical ability of our proposed model, the comparison between our model and several existing prediction tools is performed using the testing dataset. As displayed in **Table 2** and Figure S4, our proposed model achieves higher values on both accuracy and MCC value, reaching 82.9% accuracy and a MCC value of 0.18. In addition, to provide an overview of the models' predictive ability, ROC curves are used to compare our proposed model with existing succinylation sites prediction tools in independent testing. As displayed in **Figure 5**, our proposed model outperforms other available prediction tools.

A web-based predictor of the proposed method

An effective prediction tool can help biologists save time and accelerate the functional study of protein succinylation sites. In this work, a web-based predictor (called SuccSite) is designed for users to analyze protein succinylation sites efficiently. Figure S5 shows the main functions such as predict, documentation, and dataset of SuccSite. **Figure 6** shows the prediction information (the prediction results with a bar-chart of AAC for each fragmented sequence having succinylated sites). To demonstrate the effectiveness of the SuccSite predictor, two case studies are provided on the website. The

Table 2 Performance comparison of SuccSite and other existing tools using an independent testing dataset

Name	Sensitivity	Specificity	Accuracy	MCC
SuccSite	**50.43%**	**84.32%**	**82.93%**	**0.18**
iSuc-PseAAC	8.6%	94.9%	72.6%	0.07
SucPred	N/A	N/A	N/A	N/A
iSuc-PseOpt	60.00%	49.78%	50.58%	0.05
pSuc-Lys	60.00%	49.78%	50.58%	0.05
SuccinSite	27.83%	73.13%	71.55%	0.00
SuccFind	31.9%	68.7%	48.7%	0.01

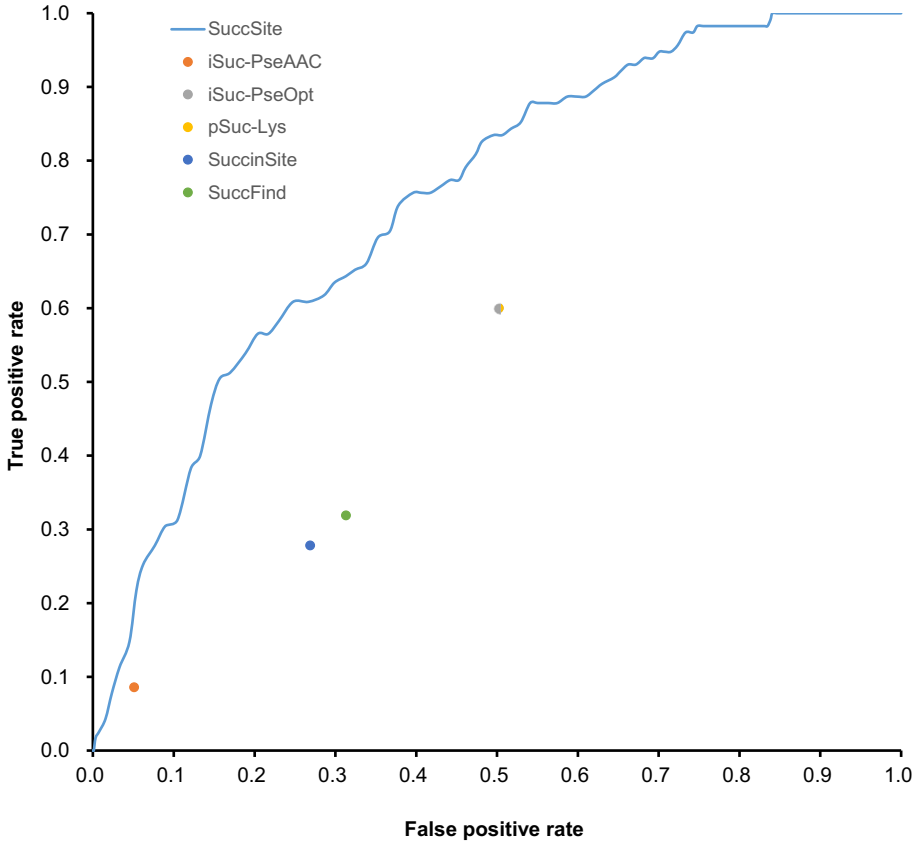


Figure 5 Comparison of ROC curves between SuccSite and other succinylation site prediction tools

first case study predicts succinylation sites for the ES1 protein homolog, mitochondrial (UniProtID: ES1_MOUSE). The ES1 consists of 266 AA residues, including 19 lysine residues. Six lysine residues are experimentally verified as succinylated substrate sites at positions 149, 155, 162, 186, 201, and 221. As displayed in Figure 6, the SuccSite can predict five succinylation sites at positions 149, 152, 155, 162, and 201. However, position 152 has not yet been experimentally confirmed as a succinylated site. Hence, the estimating TP, FN, FP, and TN of the SuccSite are 4, 2, 1, and 12, respectively. The SuccSite can achieve 84% accuracy, 67% sensitivity, and 92% specificity for this case study. The second case study is the predic-

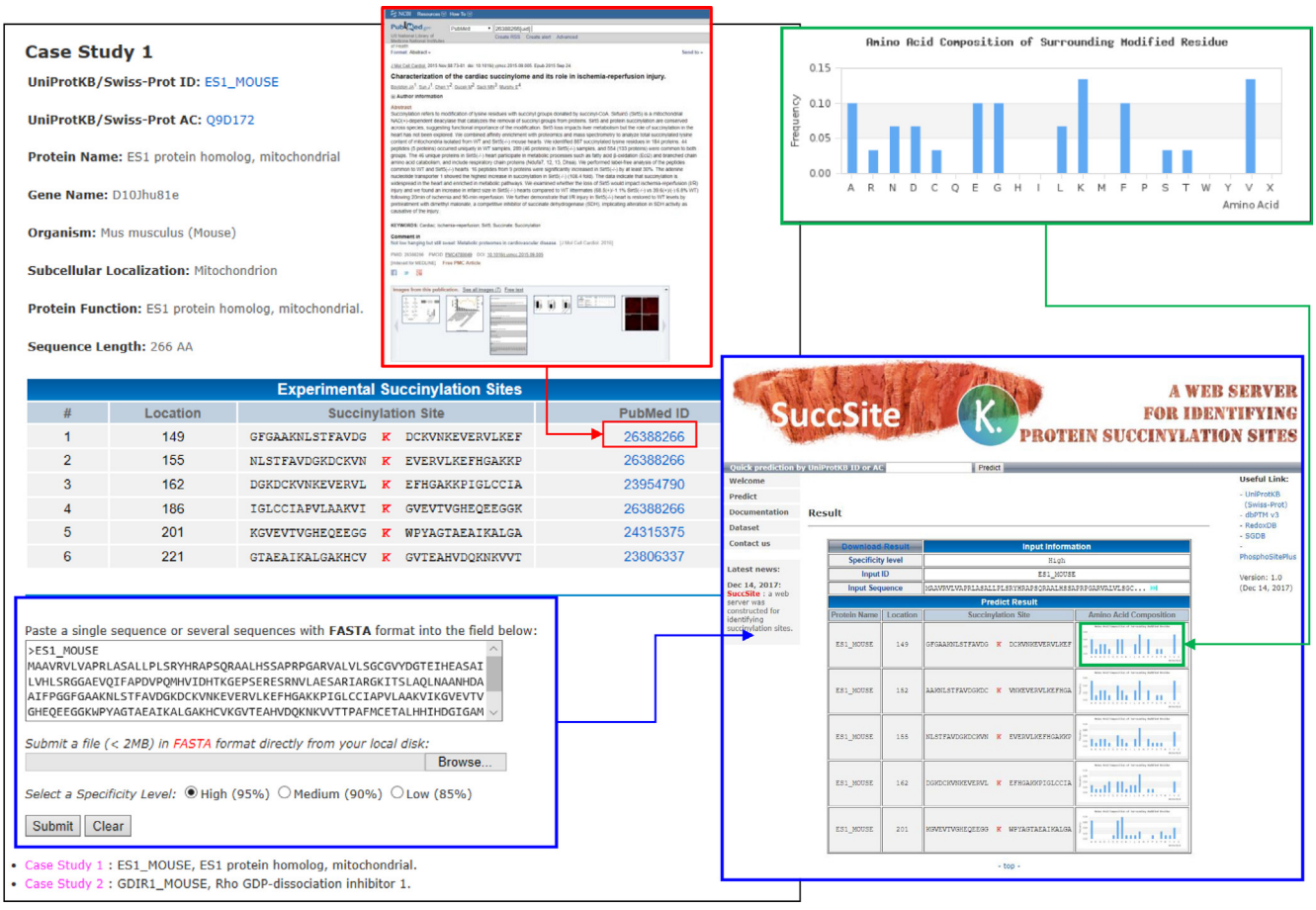


Figure 6 A case study of succinylation site prediction on ES1 protein homolog, mitochondrial
The prediction result includes the predicted positions of succinylation sites, flanking amino acids (from −15 to 15), and amino acid composition.

tion of Rho GDP-dissociation inhibitor 1 (UniProt ID: GDIR1_MOUSE), which consists of 204 AA residues, including 19 lysine residues. Two lysine residues are experimentally verified as succinylated substrate sites at positions 52 and 141. The SuccSite can predict a succinylation site at 141. Hence, the estimating TP, FN, FP, and TN were 1, 1, 0, and 17, respectively. The SuccSite yields 95% accuracy and 100% specificity for this case study.

With reference to the case study of Wang et al. [45], the computational identification of the top 10 potential succinylation sites has been conducted to determine novel succinylated lysines for biochemical communities. This investigation reveals that these potential succinylation sites occur in different proteins. As displayed in Table S3, the potential substrate site having the highest score (0.720) is at lysine 167 of histone H1 protein (UniProt ID: I7HFT9_MOUSE). Interestingly, this site is succinylated, as reported in a previous study [46]. The literature evidence indicates the reliability of the proposed method, SuccSite.

Conclusion

This work develops a new predictor, SuccSite, to investigate and identify lysine succinylation sites based on AAC and CKSAAPs. Pipelined analyses of various attributes in the neighborhood of succinylated sites are performed on the large-scale succinyl-proteome data. The Two Sample Logo investigation has revealed that the most remarkable finding is the enrichment of lysine residues within the flanking regions of succinylated sites. According to the 10-fold CV evaluation, the proposed method could yield a promising performance. The independent testing performed demonstrates that the selected SVM model (AAC + CIKSAAP) is comparable to other existing prediction tools. We believe that our proposed approach will help facilitate the determination of succinylated targets on lysine residues of proteins. In addition, to support research involved in the characterization of lysine succinylated sites, a web-based tool named SuccSite has been designed and implemented. The SuccSite is free for use and will be updated regularly.

Availability

SuccSite is available at http://csb.cse.yzu.edu.tw/SuccSite/.

Authors' contributions

HJK and VNN carried out the data collection and curation, participated in the bioinformatics analyses, and drafted the manuscript. HJK and KYH carried out the web tool imple-

mentation. WCC participated in the design of the study and performed the draft revision. TYL conceived of the study, and participated in its design and coordination and helped to revise the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors have declared no competing interests.

Acknowledgments

The authors sincerely appreciate the Warshel Institute for Computational Biology, School of Life and Health Sciences, The Chinese University of Hong Kong, Shenzhen, China for financially supporting this research.

Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gpb.2018.10.010.

ORCID

0000-0001-6109-0319 (Kao, HJ)
0000-0002-8004-8068 (Nguyen, VN)
0000-0001-9855-1035 (Huang, KY)
0000-0002-0347-1516 (Chang, WC)
0000-0001-8475-7868 (Lee, TY)

References

[1] Marquez J, Lee SR, Kim N, Han J. Post-translational modifications of cardiac mitochondrial proteins in cardiovascular disease: not lost in translation. Korean Circ J 2016;46:1–12.
[2] Zhang Z, Tan M, Xie Z, Dai L, Chen Y, Zhao Y. Identification of lysine succinylation as a new post-translational modification. Nat Chem Biol 2011;7:58–63.
[3] Benit P, Letouze E, Rak M, Aubry L, Burnichon N, Favier J, et al. Unsuspected task for an old team: succinate, fumarate and other Krebs cycle acids in metabolic remodeling. Biochim Biophys Acta 2014;1837:1330–7.
[4] Kawai Y, Fujii H, Okada M, Tsuchie Y, Uchida K, Osawa T. Formation of Nepsilon-(succinyl)lysine *in vivo*: a novel marker for docosahexaenoic acid-derived protein modification. J Lipid Res 2006;47:1386–98.
[5] Ong SE, Mann M. Mass spectrometry-based proteomics turns quantitative. Nat Chem Biol 2005;1:252–62.
[6] Xie Z, Dai J, Dai L, Tan M, Cheng Z, Wu Y, et al. Lysine succinylation and lysine malonylation in histones. Mol Cell Proteomics 2012;11:100–7.
[7] 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.
[8] Zhao X, Ning Q, Chai H, Ma Z. Accurate in silico identification of protein succinylation sites using an iterative semi-supervised learning technique. J Theor Biol 2015;374:60–5.
[9] Jia J, Liu Z, Xiao X, Liu B, Chou KC. 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.
[10] Jia J, Liu Z, Xiao X, Liu B, Chou KC. pSuc-Lys: predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach. J Theor Biol 2016;394:223–30.
[11] Zhang Z. Introduction to machine learning: k-nearest neighbors. Ann Transl Med 2016;4:218.
[12] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–402.
[13] Sahu SS, Panda G. A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction. Comput Biol Chem 2010;34:320–7.
[14] Hasan MM, Zhou Y, Lu X, Li J, Song J, Zhang Z. Computational identification of protein pupylation sites by using profile-based composition of k-spaced amino acid pairs. PLoS One 2015;10:e0129635.
[15] Park KJ, Kanehisa M. Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. Bioinformatics 2003;19:1656–63.
[16] Huang KY, Su MG, Kao HJ, Hsieh YC, Jhong JH, Cheng KH, et al. dbPTM 2016: 10-year anniversary of a resource for post-translational modification of proteins. Nucleic Acids Res 2016;44:D435–46.
[17] Su MG, Huang KY, Lu CT, Kao HJ, Chang YH, Lee TY. topPTM: a new module of dbPTM for identifying functional post-translational modifications in transmembrane proteins. Nucleic Acids Res 2014;42:D537–45.
[18] 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:D531–6.
[19] Boutet E, Lieberherr D, Tognolli M, Schneider M, Bansal P, Bridge AJ, et al. UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to sse the entry view. Methods Mol Biol 2016;1374:23–54.
[20] Dimmer EC, Huntley RP, Alam-Faruque Y, Sawford T, O'Donovan C, Martin MJ, et al. The UniProt-GO annotation database in 2011. Nucleic Acids Res 2012;40:D565–70.
[21] Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics 2010;26:680–2.
[22] Hasan MM, Yang S, Zhou Y, Mollah MN. SuccinSite: a computational tool for the prediction of protein succinylation sites by exploiting the amino acid patterns and properties. Mol Biosyst 2016;12:786–95.
[23] Xu HD, Shi SP, Wen PP, Qiu JD. SuccFind: a novel succinylation sites online prediction tool via enhanced characteristic strategy. Bioinformatics 2015;31:3748–50.
[24] Huang HD, Lee TY, Tzeng SW, Horng JT. KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites. Nucleic Acids Res 2005;33:W226–9.
[25] Sahu SS, Panda G. A novel featue representation method based on Chou's pseudo amino acid composition for protein structural class prediction. Comput Biol Chem 2010;34:320–7.
[26] Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res 2004;14:1188–90.
[27] Vacic V, Iakoucheva LM, Radivojac P. Two sample logo: a graphical representation of the differences between two sets of sequence alignments. Bioinformatics 2006;22:1536–7.
[28] Wang XB, Wu LY, Wang YC, Deng NY. Prediction of palmitoylation sites using the composition of k-spaced amino acid pairs. Protein Eng Des Sel 2009;22:707–12.
[29] Chen Z, Chen YZ, Wang XF, Wang C, Yan RX, Zhang Z. Prediction of ubiquitination sites by using the composition of k-spaced amino acid pairs. PLoS One 2011;6:e22930.
[30] Chen Z, Zhou Y, Song J, Zhang Z. hCKSAAP_UbSite: improved prediction of human ubiquitination sites by exploiting amino acid pattern and properties. Biochim Biophys Acta 2013;1834:1461–7.

[31] Wong YH, Lee TY, Liang HK, Huang CM, Wang TY, Yang YH, et al. KinasePhos 2.0: a web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns. Nucleic Acids Res 2007;35:W588–94.
[32] Lee TY, Huang HD, Hung JH, Huang HY, Yang YS, Wang TH. dbPTM: an information repository of protein post-translational modification. Nucleic Acids Res 2006;34:D622–7.
[33] Lu CT, Huang KY, Su MG, Lee TY, Bretana NA, Chang WC, et al. DbPTM 3.0: an informative resource for investigating substrate site specificity and functional association of protein post-translational modifications. Nucleic Acids Res 2013;41: D295–305.
[34] Zien A, Ratsch G, Mika S, Scholkopf B, Lengauer T, Muller KR. Engineering support vector machine kernels that recognize translation initiation sites. Bioinformatics 2000;16:799–807.
[35] Byvatov E, Schneider G. Support vector machine applications in bioinformatics. Appl Bioinformatics 2003;2:67–77.
[36] Dennis Jr G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol 2003;4:P3.
[37] Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. J Bioinform Comput Biol 2005;3:185–205.
[38] Lv H, Han J, Liu J, Zheng J, Liu R, Zhong D. Carspred: a computational tool for predicting carbonylation sites of human proteins. PLoS One 2014;9:e111478.
[39] Vapnik VN. An overview of statistical learning theory. IEEE Trans Neural Netw 1999;10:988–99.
[40] Chang CC, Lin CJ. LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol 2011;2:1–27.
[41] Kumari B, Kumar R, Kumar M. PalmPred: an SVM based palmitoylation prediction method using sequence profile information. PLoS One 2014;9:e89246.
[42] Lu CT, Chen SA, Bretana NA, Cheng TH, Lee TY. Carboxylator: incorporating solvent-accessible surface area for identifying protein carboxylation sites. J Comput Aided Mol Des 2011;25:987–95.
[43] Lee TY, Chen SA, Hung HY, Ou YY. Incorporating distant sequence features and radial basis function networks to identify ubiquitin conjugation sites. PLoS One 2011;6:e17331.
[44] Chang WC, Lee TY, Shien DM, Hsu JB, Horng JT, Hsu PC, et al. Incorporating support vector machine for identifying protein tyrosine sulfation sites. J Comput Chem 2009;30:2526–37.
[45] Wang M, Jiang Y, Xu X. A novel method for predicting post-translational modifications on serine and threonine sites by using site-modification network profiles. Mol Biosyst 2015;11:3092–100.
[46] Balachandran R, Thampatty P, Enrico A, Rinaldo C, Gupta P. Human immunodeficiency virus isolates from asymptomatic homosexual men and from AIDS patients have distinct biologic and genetic properties. Virology 1991;180:229–38.