ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/289535873

SuccinSite: A computational tool for the prediction of protein succinylation sites by exploiting the amino acid…

Article *in* Molecular BioSystems · January 2016

DOI: 10.1039/c5mb00853k

CITATIONS
5

READS
400

4 authors:

Md. Mehedi Hasan
Kyushu Institute of Technology, Iizuka, Fuku…
15 PUBLICATIONS **13** CITATIONS
SEE PROFILE

Shiping Yang
China Agricultural University
7 PUBLICATIONS **15** CITATIONS
SEE PROFILE

Yuan Zhou
Peking University
36 PUBLICATIONS **221** CITATIONS
SEE PROFILE

Md Nurul Haque Mollah
University of Rajshahi
39 PUBLICATIONS **189** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

I am working on 2nd and 3rd generation sequencing analysis View project

All content following this page was uploaded by Md. Mehedi Hasan on 06 November 2017.

The user has requested enhancement of the downloaded file.

Molecular BioSystems

PAPER

Cite this: *Mol. BioSyst.*, 2016, 12, 786

SuccinSite: a computational tool for the prediction of protein succinylation sites by exploiting the amino acid patterns and properties†

Md. Mehedi Hasan,*[a] Shiping Yang,[a] Yuan Zhou[a] and Md. Nurul Haque Mollah[b]

Received 6th December 2015,
Accepted 17th December 2015

DOI: 10.1039/c5mb00853k

www.rsc.org/molecularbiosystems

Lysine succinylation is an emerging protein post-translational modification, which plays an important role in regulating the cellular processes in both eukaryotic and prokaryotic cells. However, the succinylation modification site is particularly difficult to detect because the experimental technologies used are often time-consuming and costly. Thus, an accurate computational method for predicting succinylation sites may help researchers towards designing their experiments and to understand the molecular mechanism of succinylation. In this study, a novel computational tool termed SuccinSite has been developed to predict protein succinylation sites by incorporating three sequence encodings, *i.e.*, *k*-spaced amino acid pairs, binary and amino acid index properties. Then, the random forest classifier was trained with these encodings to build the predictor. The SuccinSite predictor achieves an AUC score of 0.802 in the 5-fold cross-validation set and performs significantly better than existing predictors on a comprehensive independent test set. Furthermore, informative features and predominant rules (*i.e.* feature combinations) were extracted from the trained random forest model for an improved interpretation of the predictor. Finally, we also compiled a database covering 4411 experimentally verified succinylation proteins with 12 456 lysine succinylation sites. Taken together, these results suggest that SuccinSite would be a helpful computational resource for succinylation sites prediction. The web-server, datasets, source code and database are freely available at http://systbio.cau.edu.cn/SuccinSite/.

Introduction

Lysine succinylation has been recently found in widespread reversible protein post-translational modification (PTM) and is present in both eukaryotic and prokaryotic cells.[1–3] In the succinylation process, a succinyl group is transferred from succinyl-CoA to the specific ε-amino group of a lysine residue in the target protein. Since succinylation results in a more substantial alteration to the chemical structures of lysine in comparison with other types of lysine PTMs such as methylation and acetylation,[4] lysine succinylation has been proposed to promote more remarkable changes in protein structure and function, including regulation of the physicochemical properties of protein, protein conformational space and protein stability.[2,5,6] Nevertheless, the full regulatory role of succinylation is still an elusive issue.

The identification of succinylation sites is an essential step to address the mechanism and function of protein succinylation. A number of proteomic experiments have been performed to identify succinylated proteins based on the molecular signature of the succinylated sites.[1,4,7–9] However, the experimental identification of succinylation sites is inefficient; it is often time consuming and expensive. Therefore, efficient computational prediction methods are highly desirable and necessary. Currently, three computational methods have been proposed for succinylation sites prediction.[10–12] Zhao *et al.* proposed a support vector machine (SVM)-based computational predictor SucPred based on primary sequences information.[10] This predictor used multiple positional sequence encoding schemes, including auto-correlation functions, grouped weight based encoding, positional weight amino acids composition and normalized van der Waals volume to predict succinylation sites. Xu *et al.* developed another SVM-based predictor iSuc-PseAAC, by exploiting a single sequence encoding, *i.e.* pseudo amino acid composition.[11] More recently, Xu *et al.* also developed a SVM-based predictor SucFind.[12] It was constructed based on the feature composition of *k*-spaced amino acid pairs and one of the amino acid index (AAindex) properties. Nevertheless, there is still room for improvement in the performance of the predictors. First, SucPred and iSuc-PseAAC predictor datasets were compiled from the

[a] State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University, Beijing, 100193, China. E-mail: mehedicau@hotmail.com
[b] Laboratory of Bioinformatics, Department of Statistics, University of Rajshahi, Rajshahi 6205, Bangladesh

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c5mb00853k

This journal is © The Royal Society of Chemistry 2016

same lysine modification database.[13] In addition, the SuccFind predictor dataset was compiled from the lysine modification database and some relevant articles.[4,8,13] The training datasets of these three predictors are relatively small and miss a number of novel succinylation sites from the latest high-throughput proteomic assays. Second, these predictors rely mainly on position-dependent encodings and position-independent composition-based encodings. It is likely that the integration of these two types of encodings could result in an improved prediction performance. Finally, the SVM model is not straightforward enough and the underlying biological implications remain hard to interpret.

To improve the prediction performance, in this study, we proposed a novel predictor "SuccinSite" for succinylation sites prediction. First, we compiled a more comprehensive dataset containing 12 456 lysine succinylation sites from 4411 proteins. Second, we combined three informative encoding features, *i.e.*, composition of *k*-spaced amino acid pairs (CKSAAP), binary encoding and selected AAindex physicochemical features. Then, the random forest (RF) classifier was trained with these encodings to build the predictor. In particular, the composition-based CKSAAP encoding was combined with the other two position-dependent encodings. The CKSAAP encoding was initially introduced for protein crystallization prediction,[14] and it was used to solve a number of prediction tasks in bioinformatics such as the prediction of flexible/rigid region,[15] protein ubiquitination sites,[16] and protein *O*-glycosylation sites.[17] In this study, we found that CKSAAP was suitable for representing the sequence context surrounding the succinylation sites. To reflect the position-specific amino acid pattern of the surrounding succinylation sites, the binary encoding was also important for the prediction of succinylation sites. By further integrating of AAindex properties, the proposed method can achieve an even better performance on both of the cross-validation test and a large-scale independent test.

Finally, we analyzed the trained RF model and extracted the significant rules from it. The significant rules were helpful not only to straightforwardly interpret how the predictor exploits feature combinations, but also to understand the biological implications underlying the model. Thus, the proposed SuccinSite is likely to provide useful information about potential novel succinylation sites in query proteins.

Materials and methods

In brief, SuccinSite is a RF-based predictor, which was constructed using the combination of the three consecutive sequence encoding features. The summary of this proposed SuccinSite predictor is shown in Fig. 1.

Datasets

Annotations of succinylated sites were collected from multiple published articles.[1,4,5,7–9] These annotations were extracted from the UniProtKB/Swiss-Prot[18] and NCBI protein sequence database (http://www.ncbi.nlm.nih.gov/protein/). Among these collected succinylated proteins, some proteins were not included in the protein sequence database and some protein lysine positions did not match in line with the database. Thus, these proteins were removed from our study. Applying a 30% homology-reducing screening procedure using CD-HIT,[19] we obtained 2322 succinylation proteins with 5004 experimentally verified lysine succinylation sites. The experimentally validated lysine succinylated sites were considered as positive samples (*i.e.* succinylated sites), while all the remaining lysine residues were considered as negative samples (*i.e.* non-succinylated sites). Each site with lysine in the center was represented as a sequence fragment. Initially, 124 proteins were singled out as an independent test-set to evaluate and compare the performance of proposed SuccinSite with existing predictors, while the remaining dataset was used as a training set. As a result, 124 proteins with 254 succinylated sites and 2977 non-succinylated sites were obtained as an independent test-set. Then, remaining 2198 proteins with 4750 succinylated sites were utilized as positive training data and 9500 non-succinylated sites were randomly selected from the whole negative dataset as negative training set.

It can be noted that for the independent test set, the corresponding proteins with all positive and negative samples were used. Nonetheless, for the training set, only a 1 : 2 ratio of the positive *vs.* negative samples was assessed. The performance of all the training models were tested and assessed through 5-fold cross-validation tests.

Feature encoding

In order to build an effective prediction model, we encoded each sequence fragment into a numeric vector, which was the crucial step to present the classifier and ensemble architecture. Thus, a high-quality sequence encoding method for keeping the generated code compact in dimensionality was necessary. Instead of employing a simple binary representation, three types of amino acid feature encodings were adopted, including CKSAAP, binary and AAindex encoding.

CKSAAP encoding

CKSAAP is one of the most classical encoding methods and was initially developed by Chen *et al.*[14] It has been widely used in several bioinformatics tasks.[15–17,20–23] The procedure of CKSAAP is briefly described as follows. If a sequence fragment was composed of a window size $2r + 1$ and 21 types of amino acids (including the gap (O)), it may contain $(21 \times 21) = 441$ types of amino acid pairs (*i.e.* AA, AC, AD,..., OO) for every single *k* (*k* denotes the space between two amino acids). For instance, "AXXA", whose *k*-space number is equal to 2. In this study, the optimal k_{max} was set at 5, indicating that $21 \times (k_{max} + 1) \times 21 = 2646$ different amino acid pairs were collected to calculate the corresponding feature vector for each sequence.

The feature vector was then calculated using the following equation:

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

where N_{total} is the length of the total composition residues (for example, if the fragment residue with a length *L* is 27 and

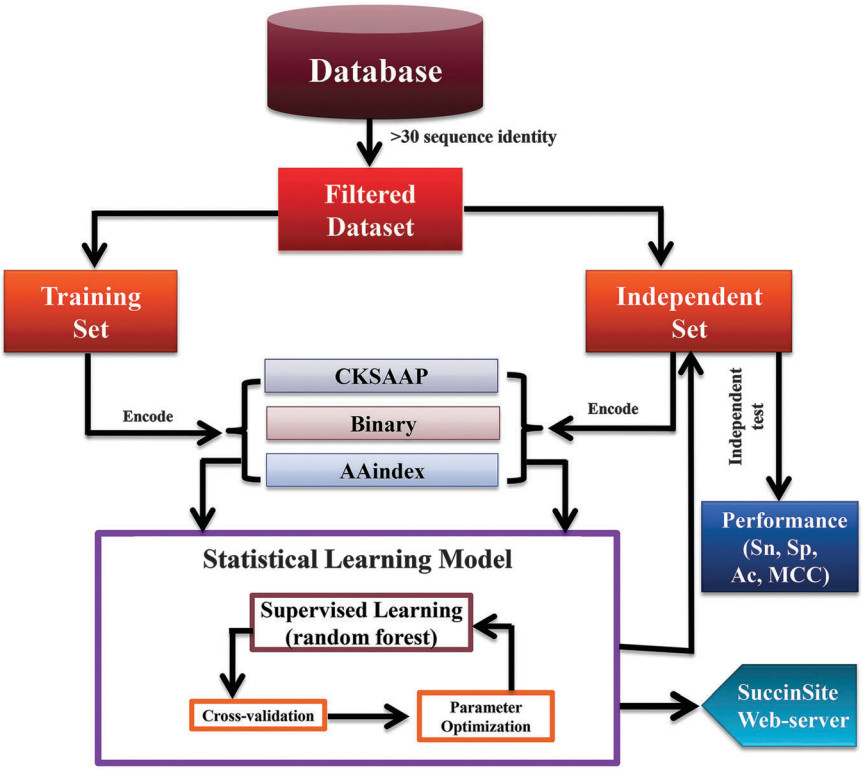

Fig. 1 The SuccinSite predictor pipeline. The datasets were collected from several published articles. After limiting a 30% sequence redundancy, the non-redundant proteins were obtained. Initially, the independent test datasets were randomly selected from the filtered proteins. Then, for the remaining proteins, a 1 : 2 ratio of succinylation *vs.* non-succinylation sites was selected to construct the training set. After encoding three types of features, the RF was utilized to build the classifier. Then, the best model was built after parameter optimization and performance evaluation. Finally, an online web server of SuccinSite was established.

k = 0, 1, 2, 3, 4, 5 then $N_{total} = L - k - 1$ will be 26, 25, 24, 23, 22 and 21, respectively). N_{AA}, N_{AC},. . ., N_{OO} represent the frequency of the amino acid pair within the fragment.

Binary encoding

In order to make a robust predictor, binary amino acid encoding was considered to calculate the positional information from the corresponding sequence fragments. In this study, 21 (including gap (O)) amino acids were transformed into numeric vectors by adopting a binary vector. The 21 types of residues were ordered as ACDEFGHIKLMNPQRSTVWYO. For adopting binary vector, in query proteins, A was represented as 100000000000000000000 and C as 010000000000000000000, and so on. The selected window size of surrounding succinylated sites was 27. For the query proteins of succinylation sites, the center position was always K. Thus, it was not considered to be taken into account. Finally, the feature vectors with a dimensionality (21 × 26) = 546 were obtained from the binary encoding.

AAindex encoding

The primary physicochemical and biochemical properties of the amino acids were extracted from AAindex database[24] (version 9.1). In this study, 544 amino acid biochemical property indices were considered. Among the 544 indices, 13 had an overrepresentation of zeroes and incomplete data, which were removed from our study. Finally, 531 physicochemical properties were considered as potential features for representing the protein sequence. If all the biochemical indices were taken into account, the dimension of feature vector would be $L \times 531$, where L is the window size. Owing to such high dimensional spaces, 350 most informative AAindex features were selected *via* the minimum redundancy maximum relevance (mRMR) feature selection approach.[25] Feature selection was performed to avoid the over-fitting problem based on the solely training samples during cross-validation tests.

Random forest classifier

RF classifier is a collection of decision tree classifiers, wherein each tree is trained with a randomly selected subset of samples. The decision tree is grown as follows. Suppose N samples are randomly selected with replacement from the F features, then the best split node is selected from F features. Finally, the decision tree is grown as large as possible without pruning. In the construction of the forest, it is generalized based on most votes given by all the individual trees; within the post for the error estimate it does not produce bias. It is relatively robust to noise and outliers.[26] As a supervised learning algorithm, it has been widely used in protein bioinformatics.[27–29] The predicted result of the RF was decided by voting among the number of trees, which contains two classes, either positive samples (succinylated sites) or negative samples (non-succinylated sites). In this study, the

 This journal is © The Royal Society of Chemistry 2016

RF algorithm was implemented using the 'RandomForest' R package.[30]

Feature selection

As mentioned in "Feature encoding", each investigated succinylated or non-succinylated fragment was encoded into as a high dimensional vector. Therefore, they may not equally contribute to determine the surrounding succinylated or non-succinylated sites. To address this, a popular feature selection method mRMR was adopted to distinguish them. It was first introduced by Peng *et al.*[25] and executed based on the criteria of Max-Relevance and Min-Redundancy. For more details on the mRMR method, please refer to the related reference.[31]

Rule extraction

A significant rule extraction strategy was applied to see how this feature works in combination within the train RF model. The RF can be represented by a set of rules for an individual tree. From the root to a leaf node is a rule for a tree. In this study, for extraction of the rules, only 100 decision trees were grown for each encoding to the train RF model, since rule extraction processes are very time consuming. We applied a method to find the rules that do not result in any wrong prediction while covering more positive instances (*i.e.* succinylation sites) as possible. More details on the rule extraction method can be found in the related literature.[27,32]

Performance assessment

To investigate the performance of the proposed SuccinSite predictor, we considered four widely used performance measures denoted as sensitivity (Sn), specificity (Sp), accuracy (Ac) and the Matthews correlation coefficient (MCC). To formulate these performance measures in our current context, let us consider a two-class prediction problem (binary classification), in which the outcomes (lysine succinylated or non-succinylated) are labeled either as positive (+) or negative (−). There are four possible outcomes from a binary classifier. If the predicted class is '+' and the actual value is also '+', then it is called a true positive (TP); however, if the actual value is '−' then it is said to be a false positive (FP). In contrast, a true negative (TN) occurred when both the prediction outcome and the actual value are '−', and false negative (FN) is when the prediction outcome is '−', while the actual value is '+'. The four outcomes are presented in the following formulas,

$$\mathrm{Sn} = \frac{n(\mathrm{TP})}{n(\mathrm{TP}) + n(\mathrm{FN})} \quad (2)$$

$$\mathrm{Sp} = \frac{n(\mathrm{TN})}{n(\mathrm{TN}) + n(\mathrm{FP})} \quad (3)$$

$$\mathrm{Ac} = \frac{n(\mathrm{TP}) + n(\mathrm{TN})}{n(\mathrm{TP}) + n(\mathrm{TN}) + n(\mathrm{FP}) + n(\mathrm{FN})} \quad (4)$$

$$\mathrm{MCC} = \frac{n(\mathrm{TP}) \times n(\mathrm{TN}) - n(\mathrm{FP}) \times n(\mathrm{FN})}{\sqrt{[n(\mathrm{TN}) + n(\mathrm{FN})][n(\mathrm{TP}) + n(\mathrm{FP})][n(\mathrm{TN}) + n(\mathrm{FP})][n(\mathrm{TN}) + n(\mathrm{FP})]}} \quad (5)$$

where n(TP) represents the number of correctly predicted succinylation windows, n(TN) is the number of correctly predicted non-succinylation windows, n(FP) is the number of incorrectly predicted succinylation windows and n(FN) is the number of incorrectly predicted non-succinylation windows.

The values of all of these measurements lie between 0 and 1, and a higher value represents a better prediction. In addition, we also used the receiver operating characteristics (ROC) curve (Sn *vs.* 1 − Sp plot) and area under the ROC curve (AUC) for calculating the performance of the proposed prediction.[33,34]

To evaluate the performance of the proposed predictor using the performance measures (Sn, Sp, Ac and MCC) as early mentioned earlier, a five-fold cross validation (CV) was used to investigate the performance of the proposed method based on the three encoding methods (CKSAAP, Binary and AAindex). For the 5-fold CV, the original datasets were randomly and equally divided into 5 subgroups. Among the 5 subgroups, one subgroup was singled out as the test dataset and the remaining 4 subgroups were considered as the training dataset. Then, we computed all the performance measures for each predictor. We replicated this procedure 5 times by changing the training and test datasets from the 5 subgroups. Finally, we computed the average value for each performance measure for each predictor.

Results and discussion

Prediction capabilities of the different encoding features with a RF classifier

In nature, the succinylation and non-succinylation datasets are highly unbalanced. It has been established that statistical learning algorithms become computationally intractable and the accuracy is strongly affected due to the nature of the unbalanced datasets. To address this issue, many PTM site prediction studies employ a relatively balanced ratio between the positive and negative samples during the training of the classifiers (*e.g.* the ratio of positives *versus* negatives is controlled at 1 : 1 or 1 : 2),[16,23,28] including the succinylation sites prediction as well.[10] After evaluating the different ratios of succinylation and non-succinylation peptides, a comparatively balanced training dataset (1 : 2 ratio of positive *vs.* negative) was used to develop the RF-based predictor of SuccinSite.

In order to evaluate the performance capability of the different encoding features, three single feature encoding-based models (CKSAAP, Binary and AAindex) and their combined model were constructed. The performance of these models was assessed using a 5-fold cross-validation. To obtain the RF score of the combined model, the corresponding RF scores for the CKSAAP, binary and AAindex encoding-based models were summed with the weight values at 0.55, 0.4 and 0.05, respectively. The AUC performance values on both the training and independent test datasets are given in Table 1. As indicated by the highest weight value, CKSAAP provides most accurate predictions of succinylation sites

Table 1 Performance of single and combined classifiers on the training and independent test data[a]

Datasets	Predictors	AUC
Training test	CKSAAP	0.770
	CKSAAP + Binary	0.798
	CKSAAP + Binary + AAindex	0.802
Independent test	CKSAAP	0.689
	CKSAAP + Binary	0.732
	CKSAAP + Binary + AAindex	0.738

[a] The weight values 0.55, 0.4, and 0.05 were used for corresponding CKSAAP, Binary and AAindex encoding, respectively.

(with AUC scores of 0.770 and 0.689 under the 5-fold cross-validation for the training and independent test datasets). These results indicate that the contribution of sequence composition is significant and acceptable for succinylation sites prediction. We also depicted the ROC curve for the RF classifiers (Fig. 2). It can be clearly observed that the combined model significantly outperforms the single CKSAAP-based model, achieving AUC scores of 0.802 and 0.738 (Fig. 2A and B) under the 5-fold cross-validation on the training and independent test datasets, respectively. Thus, the combined model was finally used as the proposed predictor (SuccinSite).

Another issue is the optimal size of the sequence windows flanking the succinylation and non-succinylation sites. The optimal window size was decided based on the combined AUC value of the consecutive three encoding methods mentioned above. The performance of each model with different window size is shown (Table S1, ESI†). We observed that the model with a window size of 27 was optimal to discriminate succinylation and non-succinylation sites.

A comparison of the proposed SuccinSite with existing methods

To compare the performance of SuccinSite with existing predictors (SucPred, iSuc-PseAAC and SuccFind[10–12]), we compiled an independent test dataset, including 124 succinylation proteins, which contain 254 succinylation and 2977 non-succinylation sites. In practice, 124 succinylation proteins were submitted to the existing online servers. The Sn, Sp, Ac and MCC performance measurements were used to assess the prediction capability of the servers. The Sp, Sn, and Ac in SucPred were 0.673, 0.271 and 0.643, respectively, and the corresponding MCC was only about −0.030 (Table 2). We noted that SucPred[10] mainly considered the positional sequence encoding schemes (*i.e.* the encodings that depict the amino acid position properties by the position along the sequence window), including auto-correlation functions, grouped weight based encoding, normalized van der Waals volume and the positional weight amino acids composition to predict the succinylation sites. The proposed SuccinSite combines the positional amino acid encoding, frequency based amino acid encoding and AAindex properties. For the training dataset, the amino acid propensities of surrounding succinylated sites compared to the non-succinylated sites were displayed by Two Sample Logos software[35] (Fig. 3). Briefly, in the two sample logo, only over- or under-represented residues at each position are plotted above and under the *X*-axis, respectively. The height of the letter was in proportion to the percentage of positive (if over-represented) or negative samples (if under-represented) harboring the corresponding residue. The *Y*-axis reports the cumulative percentage of these over-/under-represented residues. We can see that some amino acids are over-/underrepresented in the specific positions (Fig. 3), which indicates that the positional amino acid encoding was an efficient method to identify the succinylation sites. Nevertheless, positional

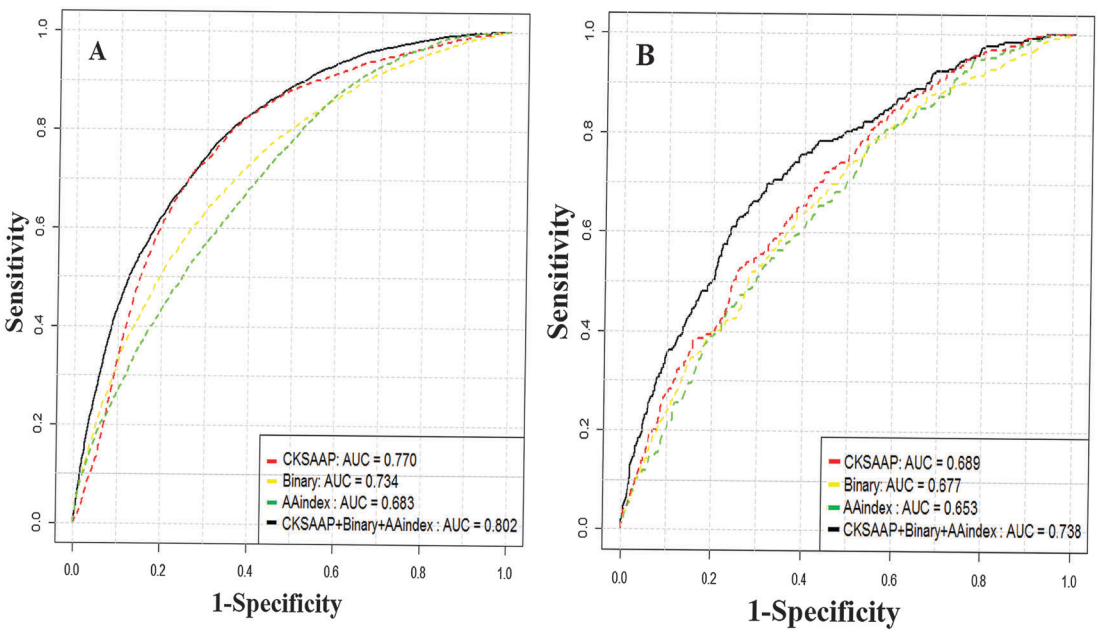

Fig. 2 The performance of single and combined ROC curves for CKSAAP, Binary and AAindex encoding. (A) Performance based on a 5-fold cross-validation of the training dataset and (B) performance based on the independent test dataset.

 This journal is © The Royal Society of Chemistry 2016

Table 2 A comparison of SuccinSite (proposed) with existing predictors using an independent test set[a]

Measurement	SucPred	iSuc-PseAAC	SuccFind	SuccinSite
Sp	0.673	0.887	0.792	0.882
Sn	0.272	0.122	0.252	0.371
Ac	0.643	0.827	0.750	0.842
MCC	−0.030	0.013	0.029	0.199

[a] The threshold values of SucPred and iSuc-PseAAC were consistent with values defined in the servers. However, the threshold value was selected as 0.9 from the SuccFind server. The proposed SuccinSite threshold was controlled as the same 90% specificity of the trained model performance.

based encoding is not enough to accurately predict the succinylation sites, as demonstrated by the performance of binary encoding in SuccinSite (Table 1) and those of the sequence encoding scheme-based predictor, SucPred (Table 2). In addition, the SucPred predictor model was trained using a 1 : 1 ratio of the positive *vs.* negative samples from the corresponding peptide fragments. In the SucPred trained model, only the negative samples dissimilar to the positive samples were selected (when the Euclidean distance score was < -0.2). In addition, the authors[10] did not select any negative samples as independent testing data to check the false positive rate. In such a case, this predictor might somehow be biased and the resulting performance is not satisfying in our independent testing set. Similarly, we observed that the proposed SuccinSite prediction performance was significantly better than other two existing predictors, iSuc-PseAAC and SuccFind (Table 2). All the prediction results prove that the SuccinSite predictor on average was better or comparable with the existing predictors for succinylation sites prediction.

Selected informative features

The mRMR feature selection method was performed to analyze the most informative features of the surrounding succinylation substrates. Although the mRMR feature selection method did not result in an improvement in performance, it allowed us to collect the most important features from the corresponding encoding methods. The top 50 informative features were collected from each of the encoding based methods to investigate the most significant residues and position between the surrounding succinylation and non-succinylation sites (see Tables S2–S4, ESI†). The feature 'EXL' represents a 1-spaced residue pair of 'EL', where 'X' stands for any amino acid and enriched the most important features for CKSAAP (see Table S2, ESI†). The same representation was applied to the other *k*-spaced residue pairs. It was intriguing that all the *k*-space, *i.e.*, 0, 1, 2, 3, 4, and 5 amino acid pairs are contained in the top 50 significant features (see Table S2, ESI†). The result shows that all space amino acid pairs are important for succinylation sites prediction. Moreover, the majority of binary encoding of the top 50 features contain charged residues, such as "K, R, H, E, and D," and stay in both sides of the window positions (see Table S3, ESI†), indicating that charged residues and both sides of the window positions may play an important role in the recognition of succinylated proteins. It was also observed that the majority of the top 50 features in the AAindex encoding method contained negative charges and polar amino acids such as "D, E" and "N, Q" (see Table S4, ESI†). These results specify that negative charges and polar amino acid residues can serve as an explicit and enriched representation of the sequence patterns within the succinylation sites. Taken together, all the selected features imply that the amino acid frequency, positional amino acid, and AAindex properties are important to identify succinylation sites.

Extracted important rules obtained from the RF models

To identify the significant features from the succinylated protein, a rule extraction method was used (Material and method section). The top 10 significant rules were extracted in each encoding from all the training succinylation sites to see how the features are important for each encoding. The full description of each specific rule and the number of succinylation sites covered by the corresponding rule are given in Table 3–5. For each rule, "&" denotes the logical conjunction and $I(\cdot)$ indicates the

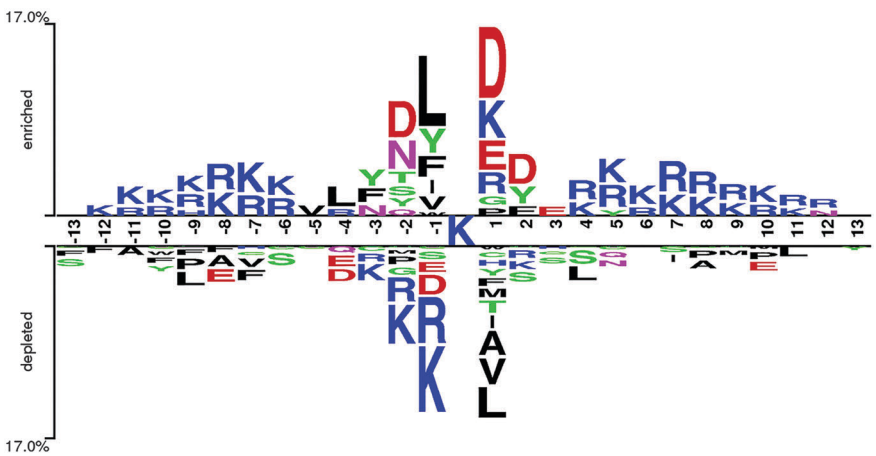

Fig. 3 The amino acid propensities of surrounding succinylation sites compared to non-succinylation sites, as displayed with the Two Sample Logos software. It also shows that the position between the compositional amino acids of the succinylated and non-succinylated peptides had a wide difference, especially those located in the positions from ~−13 to −1 and +1 to +13.

frequency of the pattern. As shown in the corresponding tables, it is likely that the same site can be covered by more than two rules. In the CKSAAP encoding method, the identified RF features rule had higher coverage than the other encodings (Table 3). This indicates that the most complicated encoding with many interactions between features partly explains why CKSAAP performs better than binary and AAindex encodings for succinylation sites prediction. In addition, we found that in the binary encoding method, some positions frequently co-occurred in the same rule (Table 4). For instance, some positions, such as −1, −2, −4, −6, −7, −10, −13, +1, +3, +6, +9, +12 and +13, frequently co-occurred in the same rule (Table 4), indicating that the amino acid propensities in the corresponding positions were correlated with the succinylation sites. It also suggests that for efficient classification, the RF algorithm was particularly powerful to explain the prediction results in the train model. Moreover, in the AAindex encoding method, a number of amino acid properties, such as "PALJ810101, QIAN880129, PRAM820102 and PALJ810113," were overrepresented in the same rule (Table 5); it means that those properties may be correlated with the succinylation sites in the corresponding AAindex. Altogether, these results determine the predictive ability of the constructed SuccinSite model and usefulness of the extracted rules based on the consecutive feature indices.

Case studies

To demonstrate the performance of SuccinSite under somehow realistic conditions, two recently experimentally identified succinylated proteins were tested, *i.e.*, *Homo sapiens*, Hepatitis C virus 5A2Q (PDB ID: 5A2Q, chain: A)[36] and the histone H1.5 in human 2RHI receptor (PDB ID: 2RHI, chain: A).[37] The probability of the succinylation sites (K66, K62, K106 and K120) was 0.459, 0.465, 0.485 and 0.384, as given by our SuccinSite server. Among the 4 lysine succinylation sites, the SuccinSite predictor was identified as 3 TP. The best rules to detect succinylation sites can also be extracted to explain the prediction results. The best rules are covered as follows:

I(AXXXN) > 0.023 and I(DXXXXXF) ≤ 0.024 and I(VXXXK) ≤ 0.022, I(RL) > 0.019 and I(IXXXV) ≤ 0.019 for the CKSAAP encoding as listed in Table 3; I(Y, −4) > 0.5, I(I, +5) ≤ 0.5 and I(R, −2) > 0.5 and I(K, −7) ≤ 0.5 for the binary encoding, as listed in Table 4 and I(R, +5, PALJ810110) ≤ 6.945 and I(G, −8, PARJ860101) ≤ 2.375, I(K, −9, PRAM820101) ≤ 0.165 and I(N, −4, QIAN880120) > 0.485 and I(T, +12, QIAN880119) > 0.085 for the AAindex encoding, as listed in Table 5.

In the abovementioned rule, while I(np) indicates the frequency of the amino acid pair np for CKSAAP, $I(n, w)$ indicates the amino acid n at position w for binary and $I(n, w, r)$ indicates the amino acid n at position w with the corresponding r position of the AAindex property.

While SuccinSite considers sequence features only, we do not rule out the possibility that the protein structural features may also be helpful. For example, we predicted the disordered tendency scores of succinylation and non-succinylation sites using IUPRED.[38] We found that the succinylation sites were more likely to be disordered (unstructured) than the non-succinylation sites (t-test, $p < 0.01$). We further randomly

Table 3 The extracted rules collected from the lysine succinylation sites in CKSAAP encoding[a]

No.	Description of rule extraction for CKSAAP	No. of samples covered by rule
1	I(RV) ≤ 0.019 & I(VXC) ≤ 0.02 & I(VXXXK) ≤ 0.022 & I(GXXXXXR) ≤ 0.024 & I(YXXR) ≤ 0.021 & I(HXXXR) ≤ 0.022 & I(KXY) ≤ 0.02 & I(SXXY) ≤ 0.021 & I(MXXXXY) ≤ 0.023 & I(KXI) ≤ 0.02 & I(VXXXXXC) ≤ 0.024 & I(KXXXXR) ≤ 0.023 & I(AXXXN) ≤ 0.023 & I(LK) ≤ 0.096 & I(PQ) ≤ 0.019 & I(HXXL) ≤ 0.062 & I(WXXG) ≤ 0.021 & I(NXK) ≤ 0.02 & I(LK) ≤ 0.019 & I(IXXXV) ≤ 0.019 & I(EXXK) > 0.146	221
2	I(KXF) ≤ 0.02 & I(YXXK) ≤ 0.021 & I(KXXXXXF) ≤ 0.071 & I(RI) ≤ 0.058 & I(PK) ≤ 0.058 & I(KL) > 0.019 & I(LXXXXXK) > 0.119	193
3	I(RL) ≤ 0.019 & I(RI) ≤ 0.019 & I(KXXE) ≤ 0.021 & I(DXXR) ≤ 0.021 & I(RT) ≤ 0.019 & I(SXXXXXT) ≤ 0.024 & I(DT) ≤ 0.019 & I(PXK) > 0.1 & I(VXXXXXP) ≤ 0.024	177
4	I(GXR) ≤ 0.02 & I(RXXXXD) ≤ 0.023 & I(RXY) > 0.06	147
5	I(RV) ≤ 0.019 & I(GXXL) ≤ 0.021 & I(YXXXXW) ≤ 0.023 & I(LXXC) ≤ 0.021 & I(GXXXL) ≤ 0.065 & I(NXXK) ≤ 0.021 & I(RXXXXV) ≤ 0.023 & I(IXXK) ≤ 0.021 & I(YK) ≤ 0.019 & I(GK) ≤ 0.019 & I(DXXK) ≤ 0.021 & I(KXL) ≤ 0.02 & I(GXXK) > 0.104	140
6	I(CS) ≤ 0.019 & I(AXXXXXA) ≤ 0.341 & I(FXXXXXS) ≤ 0.024 & I(MXC) ≤ 0.02 & I(SC) ≤ 0.019 & I(YXXXXXT) ≤ 0.024 & I(MS) ≤ 0.019 & I(FXXXK) ≤ 0.022 & I(KXY) ≤ 0.02 & I(GR) ≤ 0.019 & I(RV) ≤ 0.019 & I(TXK) ≤ 0.02 & I(VD) ≤ 0.096 & I(VXR) ≤ 0.06& I(GXP) ≤ 0.06 & I(SXXXXXY) ≤ 0.024 & I(DXXXXXF) ≤ 0.024 & I(TXXR) ≤ 0.021 & I(KXXK) ≤ 0.104 & I(RT) ≤ 0.019 & I(PXXXXL) ≤ 0.023 & I(YXXK) ≤ 0.021 & I(KXW) ≤ 0.019 & I(KXXXXK) > 0.227	132
7	I(DXXT) ≤ 0.021 & I(SXC) ≤ 0.02 & I(KXI) ≤ 0.019 & I(KXY) ≤ 0.02 & I(IXXR) ≤ 0.021 & I(AXXXXE) ≤ 0.370 & I(NXR) ≤ 0.02 & I(KXXXV) ≤ 0.022 & I(LD) ≤ 0.019 & I(YXXXXXK) ≤ 0.024 & I(AXXXXA) ≤ 0.114 & I(RXXXA) ≤ 0.022 & I(TXXG) ≤ 0.021 &I(VXXXXXG) > 0.119	106
8	I(KXXXXF) ≤ 0.023 & I(GXXXXXY) ≤ 0.024 & I(NXXXE) ≤ 0.065 & I(TH) ≤ 0.019 & I(MXXXXXK) ≤ 0.071 & I(FXXK) ≤ 0.021 & I(TXS) ≤ 0.06 & I(VXXXC) ≤ 0.022 & I(TI) ≤ 0.019 & I(AXXXY) ≤ 0.022 & I(RXY) ≤ 0.02 & I(LXXS) ≤ 0.021 & I(AXXXXXF) ≤ 0.205 & I(KXY) ≤ 0.02 & I(KXXXD) ≤ 0.022 & I(CXXXXXK) ≤ 0.024 & I(KXXXXP) > 0.114	101
9	I(SXXXXA) ≤ 0.023 & I(KD) ≤ 0.019 & I(TXC) ≤ 0.02 & I(YK) ≤ 0.019 & I(FXXC) ≤ 0.021 & I(RK) ≤ 0.019 & I(RV) ≤ 0.019 & I(VXR) ≤ 0.02 & I(AXL) ≤ 0.02 & I(TXR) ≤ 0.02 & I(PXXXXQ) ≤ 0.068 & I(RXXXA) ≤ 0.022 & I(KN) ≤ 0.019 & I(DXXR) > 0.063	87
10	I(GXR) ≤ 0.02 & I(KR) ≤ 0.019 & I(KXD) ≤ 0.02 & I(IK) ≤ 0.019 & I(RXXXXW) ≤ 0.023 & I(FS) ≤ 0.019 & I(NXK) ≤ 0.02 & I(MXR) ≤ 0.02 & I(AXXXXA) ≤ 0.370 & I(DXK) ≤ 0.02 & I(RXXXXXA) ≤ 0.024 & I(KXXXXA) ≤ 0.023 & I(WXXXD) ≤ 0.022 & I(PXR) ≤ 0.02 & I(KXXXQ) > 0.065	84

[a] For instance, the feature 'KXE' represents the 1-spaced residue (any amino acid) pair of 'KE', where X stands for any amino acid. The same representation was applied to other k-spaced residue pairs. For each rule, "&" denotes the logical conjunction and I(np) indicates the frequency of amino acid pair np.

 This journal is © The Royal Society of Chemistry 2016

Table 4 The extracted rules collected from the lysine succinylation sites in binary encoding[a]

No.	Description of rule extraction for Binary	No. of samples covered by rule
1	$I(K,-10) \le 0.5$ & $I(P,+10) \le 0.5$ & $I(F,+5) \le 0.5$ & $I(P,-7) \le 0.5$ & $I(K,-8) \le 0.5$ & $I(A,-2) \le 0.5$ & $I(F,-12) \le 0.5$ & $I(H,-10) \le 0.5$ & $I(K,-9) \le 0.5$ & $I(G,-4) \le 0.5$ & $I(V,-10) \le 0.5$ & $I(R,-10) \le 0.5$ & $I(Y,-4) \le 0.5$ & $I(L,+7) \le 0.5$ & $I(V,-13) \le 0.5$ & $I(K,-1) \le 0.5$ & $I(K,-4) \le 0.5$ & $I(N,-13) > 0.5$ & $I(H,-2) \le 0.5$ & $I(L,+4) \le 0.5$ & $I(R,-2) \le 0.5$ & $I(I,+13) \le 0.5$ & $I(Y,-6) \le 0.5$ & $I(Y,8) \le 0.5$ & $I(I,+5) \le 0.5$ & $I(Q,+5) \le 0.5$ & $I(Q,+8) \le 0.5$ & $I(F,-12) \le 0.5$ & $I(I,+11) \le 0.5$ & $I(N,-10) \le 0.5$ & $I(P,+13) \le 0.5$ & $I(M,+5) \le 0.5$ & $I(G,+3) \le 0.5$ & $I(A,+1) \le 0.5$ & $I(I,+12) \le 0.5$ & $I(L,+1) \le 0.5$ & $I(L,+13) \le 0.5$ & $I(K,+6) \le 0.5$ & $I(A,+7) \le 0.5$ & $I(H,+1) \le 0.5$ & $I(E,-8) \le 0.5$ & $I(Y,-4) \le 0.5$ & $I(N,-4) \le 0.5$ & $I(M,10) \le 0.5$ & $I(W,-2) \le 0.5$ & $I(F,-4) \le 0.5$ & $I(K,+10) \le 0.5$ & $I(D,-6) \le 0.5$ & $I(V,+2) \le 0.5$ & $I(A,+11) \le 0.5$ & $I(Q,-12) \le 0.5$ & $I(A,-6) \le 0.5$	27
2	$I(K,-7) \le 0.5$ & $I(D,-13) \le 0.5$ & $I(F,+12) \le 0.5$ & $I(E,+1) > 0.5$ & $I(D,-4) > 0.5$ & $I(D,+3) \le 0.5$ & $I(V,+10) \le 0.5$ & $I(I,-6) \le 0.5$ & $I(L,+11) \le 0.5$ & $I(R,-10) \le 0.5$ & $I(N,+12) \le 0.5$ & $I(A,-1) \le 0.5$ & $I(M,-3) \le 0.5$ & $I(A,+1) \le 0.5$ & $I(G,+6) \le 0.5$ & $I(A,+1) \le 0.5$	23
3	$I(D,-13) \le 0.5$ & $I(D,-4) \le 0.5$ & $I(K,-4) \le 0.5$ & $I(E,+1) \le 0.5$ & $I(R,+10) \le 0.5$ & $I(R,-10) \le 0.5$ & $I(L,+7) \le 0.5$ & $I(R,-10) \le 0.5$ & $I(V,-1) \le 0.5$ & $I(R,-4) \le 0.5$ & $I(F,-12) \le 0.5$ & $I(M,+8) \le 0.5$ & $I(R,+10) \le 0.5$ & $I(L,-1) \le 0.5$ & $I(Y,-4) \le 0.5$ & $I(K,-10) \le 0.5$ & $I(K,+6) > 0.5$ & $I(K,+9) \le 0.5$ & $I(D,-6) \le 0.5$ & $I(K,-10) \le 0.5$ & $I(R,+9) \le 0.5$ & $I(Y,+6) \le 0.5$ & $I(I,+12) > 0.5$ & $I(Y,+5) \le 0.5$ & $I(G,-13) \le 0.5$ & $I(R,-9) \le 0.5$ & $I(L,+10) \le 0.5$ & $I(S,-8) \le 0.5$	22
4	$I(C,-9) \le 0.5$ & $I(R,+6) \le 0.5$ & $I(D,-4) \le 0.5$ & $I(E,+4) \le 0.5$ & $I(F,+12) \le 0.5$ & $I(D,-13) \le 0.5$ & $I(E,+1) \le 0.5$ & $I(K,+6) > 0.5$ & $I(K,-4) \le 0.5$ & $I(R,+9) \le 0.5$ & $I(L,-8) > 0.5$ & $I(G,-4) \le 0.5$ & $I(F,-4) \le 0.5$ & $I(D,-6) \le 0.5$ & $I(P,+4) \le 0.5$ & $I(L,+1) \le 0.5$ & $I(K,-1) \le 0.5$ & $I(N,+3) \le 0.5$ & $I(N,+9) \le 0.5$ & $I(Y,+3) \le 0.5$ & $I(L,+1) \le 0.5$ & $I(I,+12) \le 0.5$	21
5	$I(D,-4) \le 0.5$ & $I(K,+6) > 0.5$ & $I(L,-5) \le 0.5$ & $I(W,-11) \le 0.5$ & $I(R,-7) \le 0.5$ & $I(C,-12) \le 0.5$ & $I(R,+3) \le 0.5$ & $I(R,+9) \le 0.5$ & $I(N,-13) \le 0.5$ & $I(G,+12) \le 0.5$ & $I(K,-8) \le 0.5$ & $I(K,+13) \le 0.5$ & $I(Y,+11) \le 0.5$ & $I(K,-10) \le 0.5$ & $I(Y,-1) \le 0.5$ & $I(A,+4) \le 0.5$ & $I(I,-6) > 0.5$ & $I(V,+6) \le 0.5$ & $I(R,-13) \le 0.5$ & $I(L,+13) \le 0.5$ & $I(G,-4) \le 0.5$ & $I(P,-5) \le 0.5$	21
6	$I(R,+10) \le 0.5$ & $I(D,-13) > 0.5$ & $I(Y,-6) \le 0.5$ & $I(G,-4) \le 0.5$ & $I(E,-5) \le 0.5$ & $I(V,+13) \le 0.5$ & $I(D,+3) \le 0.5$ & $I(N,+12) \le 0.5$ & $I(M,+2) \le 0.5$ & $I(V,-10) \le 0.5$ & $I(R,+3) \le 0.5$ & $I(Q,-12) \le 0.5$ & $I(Y,8) \le 0.5$ & $I(I,+13) \le 0.5$ & $I(Y,-9) \le 0.5$ & $I(K,-7) \le 0.5$ & $I(E,-8) \le 0.5$ & $I(D,-7) \le 0.5$ & $I(S,-8) \le 0.5$ & $I(F,-12) \le 0.5$ & $I(R,-13) > 0.5$ & $I(H,+13) \le 0.5$ & $I(C,+8) \le 0.5$ & $I(N,+1) \le 0.5$ & $I(P,-7) \le 0.5$ & $I(G,-2) \le 0.5$	20
7	$I(Y,+2) \le 0.5$ & $I(R,+3) \le 0.5$ & $I(R,-2) \le 0.5$ & $I(A,+13) \le 0.5$ & $I(W,-11) \le 0.5$ & $I(K,+6) \le 0.5$ & $I(D,-4) \le 0.5$ & $I(C,-6) \le 0.5$ & $I(V,+9) \le 0.5$ & $I(K,+6) \le 0.5$ & $I(L,-5) \le 0.5$ & $I(R,+12) \le 0.5$ & $I(E,+1) \le 0.5$ & $I(C,-1) \le 0.5$ & $I(P,-2) \le 0.5$ & $I(F,-9) \le 0.5$ & $I(R,-4) \le 0.5$ & $I(K,-13) \le 0.5$ & $I(R,+10) \le 0.5$ & $I(K,-10) \le 0.5$ & $I(D,-13) > 0.5$ & $I(L,-8) \le 0.5$ & $I(Q,-6) \le 0.5$ & $I(E,-8) \le 0.5$ & $I(N,+6) \le 0.5$ & $I(D,-6) \le 0.5$ & $I(N,+3) \le 0.5$ & $I(C,+5) \le 0.5$ & $I(E,-11) \le 0.5$ & $I(F,-4) \le 0.5$ & $I(Y,-3) \le 0.5$ & $I(C,+8) \le 0.5$ & $I(Q,-9) \le 0.5$ & $I(I,-3) \le 0.5$ & $I(N,+9) \le 0.5$ & $I(N,+6) \le 0.5$ & $I(K,-7) \le 0.5$ & $I(T,+5) > 0.5$ & $I(Q,+8) \le 0.5$	19
8	$I(R,+3) \le 0.5$ & $I(E,+1) \le 0.5$ & $I(D,-4) \le 0.5$ & $I(K,+6) > 0.5$ & $I(D,-10) \le 0.5$ & $I(R,+9) \le 0.5$ & $I(K,+13) \le 0.5$ & $I(I,+2) \le 0.5$ & $I(N,+9) \le 0.5$ & $I(R,-1) \le 0.5$ & $I(N,-7) \le 0.5$ & $I(C,+11) \le 0.5$ & $I(N,+9) \le 0.5$ & $I(G,-7) \le 0.5$ & $I(N,+12) \le 0.5$ & $I(L,-8) \le 0.5$ & $I(D,-6) > 0.5$ & $I(M,-3) \le 0.5$ & $I(S,-5) \le 0.5$ & $I(S,+7) \le 0.5$ & $I(G,+6) \le 0.5$ & $I(S,-8) \le 0.5$ & $I(V,+9) \le 0.5$	19
9	$I(Y,+11) \le 0.5$ & $I(C,+2) \le 0.5$ & $I(A,-11) \le 0.5$ & $I(L,-5) \le 0.5$ & $I(Y,+2) \le 0.5$ & $I(C,-12) \le 0.5$ & $I(F,+12) > 0.5$ & $I(Y,-6) \le 0.5$ & $I(D,+6) \le 0.5$ & $I(K,+6) \le 0.5$ & $I(K,+10) \le 0.5$ & $I(R,+3) \le 0.5$ & $I(K,-7) \le 0.5$ & $I(G,-4) \le 0.5$ & $I(R,-1) > 0.5$ & $I(V,-11) \le 0.5$ & $I(P,+4) \le 0.5$ & $I(R,+12) \le 0.5$ & $I(K,-4) \le 0.5$ & $I(E,+10) \le 0.5$	17
10	$I(K,-4) \le 0.5$ & $I(K,-4) \le 0.5$ & $I(K,-4) \le 0.5$ & $I(K,-13) \le 0.5$ & $I(K,-9) > 0.5$ & $I(L,+13) \le 0.5$ & $I(C,-9) \le 0.5$ & $I(A,-11) \le 0.5$ & $I(F,+12) \le 0.5$ & $I(N,+12) \le 0.5$ & $I(E,-2) \le 0.5$ & $I(D,-4) \le 0.5$ & $I(K,+9) \le 0.5$ & $I(D,+3) \le 0.5$ & $I(A,-1) \le 0.5$ & $I(Y,+11) \le 0.5$ & $I(L,+1) \le 0.5$ & $I(Q,-9) \le 0.5$ & $I(R,+6) \le 0.5$ & $I(V,-1) \le 0.5$ & $I(D,-13) > 0.5$	17

[a] For each rule, "&" denotes the logical conjunction and $I(n, w)$ indicates the amino acid n at position w.

picked three succinylation sites that could be mapped to the known protein structures. All of these succinylation sites settled near disordered fragments (which are shown as chain breaks in cartoon illustrations) or inside the random coils (see Fig. S1, ESI†). In contrast, their flanking non-succinylation sites (within the ±50 residues in the protein sequence) were more frequently located in helixes and sheets. Therefore, it is plausible that the succinylation sites tend to locate in the more unstructured regions of proteins to enhance their availability. This tendency should be validated on a comprehensive protein structure dataset, which is, however, beyond the scope of this study.

Online web server

To apply our method easily, a web server has been made to serve the user community. Our web server, SuccinSite (succinylation sites predictor), was implemented with Perl language, CGI scripts, MySQL, PHP and html. The server input and output pages are shown (see Fig. S2, ESI†). In the input page, users can directly submit their query sequence by pasting it into the text box. The server will generate all putative lysine succinylation site fragments and exert the predictions to obtain the combined RF scores for all of the putative sites. The server output page will show the results, including job ID, protein name, lysine position and RF score for the combined model, and the justification of the succinylation sites in a tabular form. Users can also view the results in text format. For future queries, the user will receive a job ID and can submit this ID again to see the prediction results within one month. The user-friendly web-server is freely accessible at http://systbio.cau.edu.cn/SuccinSite/.

Conclusions

In this article, we designed a simple and efficient predictor SuccinSite for identifying succinylation sites. For class prediction

Table 5 The extracted rules collected from the lysine succinylation sites in AAindex encoding[a]

No.	Description of rule extraction for AAindex	No. of samples covered by rule
1	I(P, +11, QIAN880136) > 5.235 & I(M, −9, PRAM820103) ≤ 6.945 & I(P, +4, PRAM900103) ≤ 5.51 & I(M, −11, QIAN880132) > −0.2 & I(W, +4, PALJ810101) ≤ 2.17 & I(M, +12, QIAN880138) ≤ −5.65 & I(Y, +13, PARJ860101) ≤ 1.49 & I(K, −9, PRAM820101) ≤ 0.165 & I(R, +5, PALJ810110) ≤ 6.945 & I(H, +11, PALJ810113) ≤ 0.299 & I(G, −7, PLIV810101) > 2.185 & I(T, +4, PALJ810101) > 1.44 & I(K, −10, PONP800105) ≤ −1.68 & I(G, −8, PARJ860101) ≤ 2.375 & I(G, −3, PALJ810103) > 0.665	21
2	I(Q, +10, PRAM820101) ≤ 1.85 & I(A, +1, QIAN880124) ≤ 6.035 & I(Y, −7, QIAN880139) > 0.315 & I(R, +5, PALJ810110) ≤ 6.16 & I(Q, −13, PARJ860101) ≤ 1.275 & I(R, +5, 232) ≤ 5.655 & I(R, −10, PTIO830101) ≤ 0.435 & I(N, −4, QIAN880120) > 0.485 & I(W, −8, PRAM820102) ≤ 4.3 & I(R, −6, PRAM820102) > −1.245 & I(M, +12, QIAN880127) ≤ 1.505 & I(T, +13, PRAM900102) ≤ 1.125 & I(H, +11, PALJ810113) > −0.0285 & (G, −7, −225) > 0.785 & I(E, +4, PRAM820102)≤4.105	12
3	I(V, −6, PONP800101) > 10.5 & I(T, +3, QIAN880126) > 1.035 & I(Q, +13, RICJ880116) ≤ −1.225 & I(T, +12, QIAN880119) > 0.085 & I(R, +5, PALJ810110) ≤ 8.665 & I(W, +2, PRAM820101) ≤ 0.35 & I(L, −9, QIAN880129) ≤ 0.8 & I(Y, −7, QIAN880138) ≤ 0.55 & I(E, +11, QIAN880129) ≤ 0.04 & I(H, +11, PALJ810113) ≤ 0.3265 & I(H, +11, PALJ810113) > 0.2175	11
4	I(E, +13, PLIV810101) ≤ 1.6 & I(G, −1, QIAN880126) ≤ −0.905 & I(W,+4, PALJ810101) ≤ 2.295 & I(Q, +13, PALJ810115) ≤ −2.125 & I(I, −2, QIAN880129) > −0.675 & I(D, −5, PALJ810109) > 0.625 & I(K, +4, PALJ810101) ≤ 2.105 & I(K, +4, QIAN880126) ≤ 68.5 & I(H, +11, PALJ810113) ≤ 0.099 & I(H, +11, PALJ810113) > 0.0215 & I(P, +5, PRAM820103) ≤ 0.745 & I(F, −1, PONP800107)≤0.335	11
5	I(P, +6, PONP800108) > 1.135 & I(P, +6, PONP800108) ≤ −0.09185 & I(T, +4, PALJ810101) ≤ 2.295 & I(F, −6, QIAN880126) ≤ 1.175 & I(N, +9, PALJ810101) > 1.045 & I(D, −3, PONP800102) ≤ 1.135 & I(N, +9, PALJ810101) ≤ 1.165 & I(H, +8, PONP800104)≤7.01	10
6	I(E, +13, PLIV810101) ≤ 1.6 & I(K, −1, QIAN880129) ≤ −0.07 & I(A, −7, PRAM900102) > 38 & I(K, +4, QIAN880126) ≤ 79.5 & I(F, +4, QIAN880134) > 26.515 & I(T, +13, PRAM900102) ≤0.9465 & I(D, +10, PALJ810101) ≤ 0.9305 & I(D, +10, PALJ810101)≤0.772 & I(W, −7 QIAN880138) > 5.87 & I(N, +9, PALJ810101) ≤ 1.365 & I(R, +13, QIAN880131) ≤ 0.55	10
7	I(D, −7, QIAN880133) ≤ 2.205 & I(A, −8, PTIO830101) ≤ −0.215 & I(P, −1, QIAN880121) ≤ 0.93 & I(L, +10, PONP800104) > 2.455 & I(Q, +4, QIAN880130) > 0.055 & I(Y, +7, PALJ810115) > 0.45 & I(H, −1, OOBM850103) ≤ 0.5 & I(I, +9, PALJ810102) > 5.625 & I(C, −8, QIAN880133) > −0.55 & I(W, +6, PALJ810115) ≤ 88.5 & I(R, +5, PALJ810110) ≤ 5.99 & I(H, +10, PALJ810113) ≤ 0.2345 & I(P, +11, QIAN880136) ≤1.655	10
8	I(K, −4, QIAN880127) > −0.385 & I(Q, −6, PLIV810101) ≤ 12.45 & I(M, −5, PONP800103) > 5.65 & I(W, +8, QIAN880134) ≤ 1.185 & I(I, −2, QIAN880126) > −0.705 & I(Q, +7, PALJ810116) > 20.5 & I(L, −7, PONP800104) > 9.1 & I(G, −6, PONP800105) > 0.975 & I(G, +6, QIAN880129) > −0.37 & I(F, −8, PRAM900104) > 0.515 & I(M, −7, QIAN880120) > 0.375 & I(T, −6, QIAN880127) > 4.04 & I(P, −11, PRAM900101) > 0.93 & I(M, −3, QIAN880129) ≤ 0.45 & I(L, +1, PALJ810109) > 1.96 & I(G, −8, QIAN880125) > 0.19 & I(E, −5, QIAN880129) > 0.86	10
9	I(K, −4, QIAN880127) ≤ −0.385 & I(Y, −5, PONP800102) ≤ 81.7 & I(K, −10, PONP800105) > −2.455 & I(P, −8, QIAN880124) ≤ 1.28 & I(D, −11, QIAN880122) > −8.32 & I(F, −13, QIAN880125) > −1.15 & I(Q, −3, PRAM820101) > 1.095 & I(V, −13, QIAN880123) > 0.58 & I(L, +1, PALJ810109) > 1.673 & I(P, +13, PRAM820102) > 1.113	9
10	I(W, −8, PRAM820102) ≤ 0.4645 & I(V, −8, PRAM900104) > 0.185 & I(P, +5, PRAM820103) ≤ 0.735 & I(M, +13, QIAN880125) > 0.31 & I(Q, +3, QIAN880133) > 0.1925 & I(L, +6, PTIO830101) ≤ 5.935 & I(Y, +12, QIAN880125) ≤ 0.8 & I(G, −6, PONP800105) ≤ 0.515 & I(V, −8, PRAM900102) > 0.781 & I(L, +10, QIAN880127) ≤ −8.266	9

[a] For each rule, "&" denotes the logical conjunction and, while $I(n, w, r)$ indicates the amino acid n at position w with the corresponding r position of AAindex property.

in the independent dataset, we observed that our proposed predictor performed better than the existing predictors. Moreover, a feature selection and rule extraction method were carried out to identify the significant rules from the RF model, which helps to better understand the important rules that underlie the succinylated proteins. The data analysis results demonstrated that the proposed method might be helpful to understand succinylation as well as the mechanisms of protein succinylation. In addition, we also provide a new succinylation sites database that contains 4411 experimentally verified succinylated proteins with 12 456 lysine succinylation sites. Although SuccinSite obtained a fairly good performance, there are still some spaces for improvement. In the future, we would like to pay more attention to make an organism specific predictor for improving the performance of succinylation sites prediction.

Acknowledgements

We would like to thank Professor Ziding Zhang from the China Agricultural University, Professor Jiangning Song from the Monash University, Australia and Dr Xiao-Feng Wang from Shanxi Normal University for their helpful discussions and comments. MMH was financially supported by the Chinese Scholarship Council (CSC) and NHM was a recipient of the Higher Education Quality Enhancement Project (CP-3603, W2-R3).

References

1 B. T. Weinert, C. Scholz, S. A. Wagner, V. Iesmantavicius, D. Su, J. A. Daniel and C. Choudhary, *Cell Rep.*, 2013, **4**, 842–851.
2 Z. Xie, J. Dai, L. Dai, M. Tan, Z. Cheng, Y. Wu, J. D. Boeke and Y. Zhao, *Mol. Cell. Proteomics*, 2012, **11**, 100–107.

 This journal is © The Royal Society of Chemistry 2016

3 M. Tan, C. Peng, K. A. Anderson, P. Chhoy, Z. Xie, L. Dai, J. Park, Y. Chen, H. Huang, Y. Zhang, J. Ro, G. R. Wagner, M. F. Green, A. S. Madsen, J. Schmiesing, B. S. Peterson, G. Xu, O. R. Ilkayeva, M. J. Muehlbauer, T. Braulke, C. Muhlhausen, D. S. Backos, C. A. Olsen, P. J. McGuire, S. D. Pletcher, D. B. Lombard, M. D. Hirschey and Y. Zhao, *Cell Metab.*, 2014, **19**, 605–617.

4 X. Li, X. Hu, Y. Wan, G. Xie, X. Li, D. Chen, Z. Cheng, X. Yi, S. Liang and F. Tan, *J. Proteome Res.*, 2014, **13**, 6087–6095.

5 Z. Zhang, M. Tan, Z. Xie, L. Dai, Y. Chen and Y. Zhao, *Nat. Chem. Biol.*, 2011, **7**, 58–63.

6 R. Rosen, D. Becher, K. Buttner, D. Biran, M. Hecker and E. Z. Ron, *FEBS Lett.*, 2004, **577**, 386–392.

7 G. Colak, Z. Xie, A. Y. Zhu, L. Dai, Z. Lu, Y. Zhang, X. Wan, Y. Chen, Y. H. Cha, H. Lin, Y. Zhao and M. Tan, *Mol. Cell. Proteomics*, 2013, **12**, 3509–3520.

8 J. Park, Y. Chen, D. X. Tishkoff, C. Peng, M. Tan, L. Dai, Z. Xie, Y. Zhang, B. M. Zwaans, M. E. Skinner, D. B. Lombard and Y. Zhao, *Mol. Cell*, 2013, **50**, 919–930.

9 M. Yang, Y. Wang, Y. Chen, Z. Cheng, J. Gu, J. Deng, L. Bi, C. Chen, R. Mo, X. Wang and F. Ge, *Mol. Cell. Proteomics*, 2015, **14**, 796–811.

10 X. Zhao, Q. Ning, H. Chai and Z. Ma, *J. Theor. Biol.*, 2015, **374**, 60–65.

11 Y. Xu, Y. X. Ding, J. Ding, Y. H. Lei, L. Y. Wu and N. Y. Deng, *Sci. Rep.*, 2015, **5**, 10184.

12 H. D. Xu, S. P. Shi, P. P. Wen and J. D. Qiu, *Bioinformatics*, 2015, **31**, 3748–3750.

13 Z. Liu, Y. Wang, T. Gao, Z. Pan, H. Cheng, Q. Yang, Z. Cheng, A. Guo, J. Ren and Y. Xue, *Nucleic Acids Res.*, 2014, **42**, D531–D536.

14 K. Chen, L. Kurgan and M. Rahbari, *Biochem. Biophys. Res. Commun.*, 2007, **355**, 764–769.

15 K. Chen, L. A. Kurgan and J. Ruan, *BMC Struct. Biol.*, 2007, **7**, 25.

16 Z. Chen, Y. Z. Chen, X. F. Wang, C. Wang, R. X. Yan and Z. Zhang, *PLoS One*, 2011, **6**, e22930.

17 Y. Z. Chen, Y. R. Tang, Z. Y. Sheng and Z. Zhang, *BMC Bioinf.*, 2008, **9**, 101.

18 C. UniProt, *Nucleic Acids Res.*, 2011, **39**, D214–D219.

19 Y. Huang, B. Niu, Y. Gao, L. Fu and W. Li, *Bioinformatics*, 2010, **26**, 680–682.

20 Z. Chen, Y. Zhou, Z. Zhang and J. Song, *Briefings Bioinf.*, 2015, **16**, 640–657.

21 K. Chen, L. A. Kurgan and J. Ruan, *J. Comput. Chem.*, 2008, **29**, 1596–1604.

22 Z. Chen, Y. Zhou, J. Song and Z. Zhang, *Biochim. Biophys. Acta*, 2013, **1834**, 1461–1467.

23 M. M. Hasan, Y. Zhou, X. Lu, J. Li, J. Song and Z. Zhang, *PLoS One*, 2015, **10**, e0129635.

24 S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama and M. Kanehisa, *Nucleic Acids Res.*, 2008, **36**, D202–205.

25 H. Peng, F. Long and C. Ding, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2005, **27**, 1226–1238.

26 L. BREIMAN, *Machine Learning*, 2001, **45**, 5–32.

27 C. Li, X. F. Wang, Z. Chen, Z. Zhang and J. Song, *Mol. BioSyst.*, 2015, **11**, 354–360.

28 Y. Li, M. Wang, H. Wang, H. Tan, Z. Zhang, G. I. Webb and J. Song, *Sci. Rep.*, 2014, **4**, 5765.

29 Y. Zhou, S. Liu, J. Song and Z. Zhang, *PLoS One*, 2013, **8**, e83167.

30 A. Liaw, M. Wiener, *R News*, 2002, **2**, 18–22.

31 Y. Cai, T. Huang, L. Hu, X. Shi, L. Xie and Y. Li, *Amino Acids*, 2012, **42**, 1387–1395.

32 X. Wang, Y. Zhou and R. Yan, *Mol. BioSyst.*, 2015, **11**, 1794–1801.

33 M. Gribskov and N. L. Robinson, *Comput. Chem.*, 1996, **20**, 25–33.

34 R. M. Centor, *Medical Decision Making*, 1991, 11, 102–106.

35 V. Vacic, L. M. Iakoucheva and P. Radivojac, *Bioinformatics*, 2006, **22**, 1536–1537.

36 N. Quade, D. Boehringer, M. Leibundgut, J. van den Heuvel and N. Ban, *Nat. Commun.*, 2015, **6**, 7646.

37 H. Li, W. Fischle, W. Wang, E. M. Duncan, L. Liang, S. Murakami-Ishibe, C. D. Allis and D. J. Patel, *Mol. Cell*, 2007, **28**, 677–691.

38 Z. Dosztanyi, V. Csizmok, P. Tompa and I. Simon, *Bioinformatics*, 2005, **21**, 3433–3434.

View publication stats