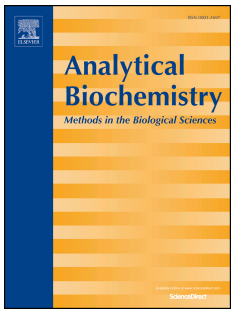
Accepted Manuscript



SucStruct: Prediction of succinylated lysine residues by using structural properties of amino acids

Yosvany López, Abdollah Dehzangi, Sunil Pranit Lal, Ghazaleh Taherzadeh, Jacob Michaelson, Abdul Sattar, Tatsuhiko Tsunoda, Alok Sharma

PII: S0003-2697(17)30147-1

DOI: 10.1016/j.ab.2017.03.021

Reference: YABIO 12662

To appear in: *Analytical Biochemistry*

Received Date: 20 December 2016

Revised Date: 13 March 2017

Accepted Date: 28 March 2017

Please cite this article as: Y. López, A. Dehzangi, S.P. Lal, G. Taherzadeh, J. Michaelson, A. Sattar, T. Tsunoda, A. Sharma, SucStruct: Prediction of succinylated lysine residues by using structural properties of amino acids, *Analytical Biochemistry* (2017), doi: 10.1016/j.ab.2017.03.021.

SucStruct: Prediction of Succinylated Lysine Residues by Using Structural Properties of Amino Acids

Yosvany López[a,b,‡], Abdollah Dehzangi[c,‡], Sunil Pranit Lal[d], Ghazaleh Taherzadeh[e], Jacob Michaelson[c], Abdul Sattar[e,f], Tatsuhiko Tsunoda[a,b,g,&], Alok Sharma[b,f,&]

[a] Department of Medical Science Mathematics, Medical Research Institute, Tokyo Medical and Dental University, Tokyo, Japan

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

[c] Department of Psychiatry, Carver College of Medicine, University of Iowa, Iowa, USA

[d] School of Engineering & Advanced Technology, Massey University, New Zealand

[e] School of Information and Communication Technology, Griffith University, Parklands Drive, Southport, Queensland 4215, Australia

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

[g] CREST, JST, Tokyo, 113-8510, Japan

[‡]First and corresponding authors

[&]Last authors

Abstract

Post-Translational Modification (PTM) is a biological reaction which contributes to diversify the proteome. Despite many modifications with important roles in the cellular activity, lysine succinylation has recently emerged as an important PTM mark. It alters the chemical structure of lysines, leading to remarkable changes in the structure and function of proteins. Given the huge amount of proteins being sequenced in the post-genome era, the experimental detection of succinylated residues remains expensive, inefficient and time-consuming. Therefore, the development of computational tools for accurately predicting succinylated lysines is an urgent necessity. To date, several approaches have been proposed but their sensitivity has been reportedly poor. In this paper, we propose an approach that utilizes structural features of amino acids to improve lysine succinylation prediction. Succinylated and non-succinylated lysines were first retrieved from 670 proteins and characteristics such as accessible surface area, backbone torsion angles, and local structure conformations were incorporated. We used the k-nearest neighbors cleaning for dealing with class imbalance and designed a pruned decision tree for classification. Our predictor, referred as SucStruct (Succinylation using Structural features), proved to significantly improve performance when compared to previous predictors, with sensitivity, accuracy and Mathew's correlation coefficient equal to 0.7334~0.7946, 0.7444~0.7608 and 0.4884~0.5240, respectively.

Keywords: lysine succinylation, structural features, protein sequences, amino acids, prediction.

Introduction

Post translational modification (PTM) refers to modification of amino acids along the protein sequence. It modifies the amino acids of a protein by covalently adding functional groups after translation is completed in ribosome [1]. There are different PTMs which play key roles in the function of cells [2]. For instance, lysine residues can undergo modifications such as methylation [3], acetylation [4], ubiquitination [5] and so on which contribute to the complexity of post-translational networks. Recently, a new

type of PTM coined succinylation has been detected in both eukaryotes and prokaryotes [6, 7]. Succinylation was identified through mass spectrometry [8] and protein sequence alignment, and was found to lead to remarkable changes in the structure and function of proteins [7]. Previous research has reported succinylation in enzymes with roles in mitochondria metabolism, amino acid degradation and fatty acid metabolism [9]. In addition, succinylation in histones suggested regulatory roles in the structure and function of chromatin [10] and in the 3D formation of human genome inside the nucleus. Therefore, the identification of succinylated residues has become an essential step in deciphering the succinylation mechanism.

Numerous proteomic experiments aimed to identify succinylated proteins considered the molecular mark of succinylated residues [11-13]. Nevertheless, the detection of these residues by experimental techniques has proved to be extremely expensive apart from being inefficient and time-consuming. Consequently, the development of computational tools capable of predicting succinylated lysines has become absolutely necessary. To date, several studies have been proposed for this purpose. SuccFind uses the sequence information derived from evolution and an enhanced feature strategy for optimization [14]. SucPred, on the other hand, makes use of a positive samples only and learning the algorithm with positive and unlabeled samples [15]. iSuc-PseAAC integrates the peptide position-specific propensity into the general form of pseudo amino acid composition for training a support vector machine [16]. iSuc-PseOpt incorporates the sequence-coupling effects into the general pseudo amino acid composition. It also employs k-nearest neighbors cleaning and insert hypothetical training samples as treatments to deal with class imbalance as well as uses a random forest algorithm for prediction task [17]. SuccinSite trains a random forest classifier with informative encoding features such as the composition of k-spaced amino acid pairs, binary encoding and selected physicochemical attributes [18]. However, these predictors have shown poor sensitivity in detecting succinylated lysine residues. Therefore, additional efforts are still needed for improving the prediction. Structural features were considered in a previous study [19] analyzed the secondary structure of

succinylated proteins and found that α-helix and coil accounted for >40% of succinylated residues.

In this paper, we propose a predictor named SucStruct, which considers a comprehensive set of structural features to discriminate succinylated from non-succinylated lysine residues. For this, we employed 670 proteins experimentally detected with succinylated residues [20, 21] and calculated features such as accessible surface area (ASA), backbone torsion angles, and probability of amino acid contribution to local structure conformations (helix, strand and coil). These characteristics were taken for the 15 upstream and downstream amino acids around succinylated and non-succinylated residues to create one training dataset [17]. We implemented the k-nearest neighbors cleaning [17] for dealing with class imbalance and finally used a pruned decision tree for classification task. SucStruct showed a considerable improvement in performance over previously proposed succinylation site predictors found in the literature [17, 18]. SucStruct successfully classifies succinylated lysine residues with 0.7334 sensitivity, 0.7444 accuracy and 0.4884 Mathew's correlation coefficient. When the features were ranked by information gain and the top ones were used for classification, the resulting sensitivity was as high as 0.7946, accuracy was 0.7608 and Mathew's correlation coefficient was 0.5240.

Materials and Methods

In this paper, we propose a machine learning based method called SucStruct to predict succinylation/non-succinylation binding sites. SucStruct uses nine structural features such as accessible surface area, backbone torsion angles and probability of amino acid contribution to local structure conformations (helix, strand and coil). These characteristics as well as the machine learning scheme used are explained in the following subsections.

Benchmark Dataset

Our benchmark dataset consists of proteins stored in two PTM databases [20, 21]. In total, we considered 670 proteins where each protein sequence could have one or more lysine residues. Each sequence was analyzed in order to extract succinylated and non-succinylated lysine residues. Consequently, 1,782 succinylation and 18,344 non-succinylation sites were retrieved. The following sections describe the different structural features computed from each of these proteins.

Structural Features

For each protein in the benchmark dataset we retrieved its sequence and computed nine features related to secondary structure, backbone and torsion angles, and accessible surface area. To do this, we used the recently developed toolbox SPIDER-2 [22] which proved to achieve good results when it comes to predicting protein secondary structure [23, 24], backbone and torsion angles [25, 26] and accessible surface area [25, 27, 28]. SPIDER-2 has also been successfully used to extract the structural properties of proteins in sequence-based prediction of protein binding sites [29, 30]. The descriptions of these structural properties are discussed in the subsequent sections.

Accessible Surface Area

ASA provides the estimated accessible area of a given amino acid to a solvent in the 3D configuration of a protein [31, 32]. Because of this, the predicted ASA of individual amino acids tends to provide essential information on the protein structure. For the ASA computation, we run SPIDER-2 for each protein sequence and retrieved its output as one estimated value for each amino acid in the protein. It is worth noting that SPIDER-2 only uses the primary sequence of proteins and such prediction is solely based on sequence information.

Secondary Structure

Secondary structure indicates the local 3D structure of proteins. In other words, each amino acid can contribute to one of the three defined local structures of a protein, namely, helix, strand and coil. The secondary structure of proteins can provide critical

information towards understanding their general 3D configuration. For each protein we run SPIDER-2 and predicted the probability of each amino acid contribution to the three local structure conformations: helix (*ph*), strand (*pe*), and coil (*pc*). Consequently, SPIDER-2 returns the local structure with highest probability in the form of one matrix whose size is $L \times 3$, where *L* represents the protein length and the three columns are the transition probabilities corresponding to the three secondary structure confirmations. For simplification, this matrix is indicated as *SSpre*.

Local Backbone Angles

The secondary structure provides discrete information regarding the local configuration of the amino acid with respect to the probability of their contribution to one of the three secondary structure elements. However, torsion angles between neighboring amino acids have been reported to provide important continuous information regarding the local structure of amino acids which complement ASA and the predicted secondary structure [26]. Due to the fact that the predicted secondary structure is a discrete output for a given amino acid, the backbone torsion angles ϕ and Ψ are predicted as continuous representatives of the local amino acid interaction along the protein backbone [33, 34]. Recent studies have also considered two new angles based on the dihedral angles θ between three Cα atoms (Cα_{i-1} - Cα_i - Cα_{i+1}) and τ rotated about the Cα_i - Cα_{i+1} bond [25]. For each protein sequence we again run SPIDER-2 to obtain the four angles. As a result, we obtained four different numerical vectors ϕ, Ψ, θ and τ.

Feature Extraction for Lysine Residues

We used the above structural features to describe each succinylated and non-succinylated lysine residue. Here we discuss the method of extracting features for each lysine residue. In order to extract features of each lysine residue, we utilized the 15 upstream and 15 downstream amino acids adjacent to the lysine residue K (Fig. 1A). If a lysine residue does not contain 15 adjacent amino acids (either up or down side) absent amino acids are then created by mirror effect [17] (Fig. 1B). The +/-15 residue window has proven to be the most promising segment for obtaining accurate lysine succinylation predictions [17]. To assess whether this reported window could also prove

effective for our study, we constructed two additional training datasets (using +/-5 and +/-10 residue windows) and trained the SucStruct predictor. Consequently, no significant improvement in performance was obtained (Supplementary material 1).

The segment of sequence S with 15 upstream, 15 downstream and lysine residue K can be expressed as follows:

$$S = \{A_{-15}, A_{-14}, \dots, A_{-2}, A_{-1}, K, A_1, A_2, \dots, A_{14}, A_{15}\} \quad (1)$$

where A_i (for $1 \leq i \leq 15$) are downstream amino acids and A_{-i} (for $1 \leq i \leq 15$) are upstream amino acids. It can be observed from Eq. (1) that in total 31 amino acids are used (including K) to define a lysine residue. A lysine residue corresponding to a segment of sequence S has class label y, where $y = \{0,1\}$; i.e., if S is a succinylated lysine residue then $y = 1$ and 0 if S is a non-succinylated residue. Furthermore, each amino acid A_i (for $-15 \leq i \leq 15$, where $A_0 = K$) is described by the following structural features:

$$A_i = \{ASA, ph, pe, pc, \phi, \psi, \theta, \tau\} \quad (2)$$

It should be noted here that the structural features $ASA, ph, pe, pc, \phi, \psi, \theta$ and τ are numeric and have a single value for each amino acid A_i. Therefore, amino acid A_i is represented by an 8-dimensional feature vector. In addition, we use the secondary structure feature $SSpre$ with bigram [35] (or $SSpre + bigram$) to represent the segment of sequence S. The $SSpre + bigram$ is a matrix of size 3×3. Therefore, each segment S (with 31 amino acids) is represented by 248 ($31 \text{ amino acids} \times 8$) structural features and 9 (3×3 matrix) other structural features extracted from profile bigrams. These features are employed to capture the structural information of a lysine (either succinylated or non-succinylated) residue corresponding to S.

Next, we discuss how to compute the $SSpre$ feature using bigram and how it can be utilized to represent the lysine residue corresponding to S. A protein sequence of length

L provides a $SSpre$ feature matrix M of size $L \times 3$, where the number of rows is L and the 3 columns depict the local structure conformations: helix, strand and coil. If an element of matrix M is denoted by m_{ij} then this element represents the transitional probability of j-th secondary structure at the i-th location of the protein sequence. If we consider a segment of protein sequence S (from Eq. (1)) then the corresponding feature matrix would be of size 31×3.

For a given amino acid in the protein sequence, *SSpre* produces the probability of formation of each secondary structure element. This matrix is then processed through a profile bigram [35] as follows:

$$B_{p,q} = \sum_{k=1}^{30} m_{k,p} m_{k+1,q} \text{ , where } 1 \leq p \leq 3 \text{ and } 1 \leq q \leq 3 \quad (3)$$

Eq. (3) returns 9 frequencies of occurrences $B_{p,q}$ ($p = 1,2,3$ and $q = 1,2,3$) for 9 bigram transitions. We define matrix B, which comprises all the elements $B_{p,q}$, as the bigram occurrence matrix of $SSpre$. In the literature, profile bigram has shown promising results for protein analysis problems and thus we have used this technique in this work [35-39]. The bigram of $SSpre$ (or simply $SSpre + bigram$) matrix B can be transformed into a 9-element feature vector F as

$$F = \left[B_{1,1}, \dots, B_{1,3}, B_{2,1}, \dots, B_{2,3}, B_{3,1}, \dots B_{3,3}\right]^T \quad (4)$$

where superscript T denotes transpose. Now we can completely define a lysine residue (corresponding to segment S) using 248 structural features and 9 other structural features derived from bigrams. Therefore, each lysine residue is defined as a vector of 257 structural features. After retrieving all the lysine residues from the benchmark dataset, we ended up with 1,782 succinylated residues ($y = 1$) and 18,344 non-succinylated residues ($y = 0$). This data was further processed to train and test our predictor SucStruct (see below).

Design of the Decision Tree

Decision trees are non-parametric methods able to predict a class label by learning rules in the dataset. Although decision trees sometimes create biased models, they require small number of samples, handle categorical and numerical values and can easily explain a condition by boolean logic. Specifically, the C4.5 algorithm can efficiently deal with discrete and continuous values, handle incomplete data and solve over-fitting through pruning [40]. In this paper, we use the WEKA software for generating a pruned C4.5 decision tree [41]. We set a confidence threshold for pruning at 0.25 and a minimum number of 2 instances per leaf, only allowed binary splits and did not use the MDL correction for information gain on numeric attributes.

Results and Discussion

It is imperative to measure the performance of SucStruct using well-defined metrics. In the literature, the four following metrics are generally used: sensitivity, specificity, accuracy (Acc) and Mathew's correlation coefficient (MCC) [36, 42-46]. In this paper, we used these four metrics for assessing the ability of our predictor to discriminate succinylated lysine residues on the benchmark dataset [20, 21].

Metrics

Sensitivity or true positive rate is a metric to evaluate the predictor performance to correctly identify succinylated lysine residues. The higher the sensitivity, the better the predictor is in detecting succinylated lysines. This metric varies between 0 and 1, where 0 indicates the predictor is totally inaccurate and 1 represents a totally accurate predictor. Sensitivity metric can be defined as

$$Sensitivity = \frac{TP}{TP+FN} \quad (5)$$

where TP denotes true positives or the number of correctly identified samples and FN denotes false negatives or the number of incorrectly rejected samples.

Specificity or true negative rate evaluates the ability of the predictor to identify negative samples or non-succinylated lysines. This metric also varies between 0 (totally incorrect) and 1 (totally correct). Specificity can be defined as

$$Specificity = \frac{TN}{TN+FP} \quad (6)$$

where TN depicts true negatives or the number of correctly rejected samples and FP depicts false positives or the number of incorrectly accepted samples.

Accuracy (Acc) is equal to the total number of correctly classified samples (*C*) over the total number of samples (*N*). It varies between 0 (least accurate) and 1 (most accurate), and can be defined as

$$Acc = \frac{C}{N} \quad (7)$$

Mathew's correlation coefficient (MCC) gauges the classification quality of the model. It varies between -1 and 1, where 1 denotes a full positive classification correlation and -1 denotes a full negative classification correlation. MCC can be defined as

$$MCC = \frac{(TN \times TP) - (FN \times FP)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (8)$$

The ideal predictor is the one which performs the highest in all the four metrics. Most importantly, at least sensitivity should be higher compared to other predictors. A predictor with lower value of sensitivity depicts that it cannot correctly predict the succinylated lysine residues and therefore it is not suitable for lysine succinylation detection.

Evaluation methods

In order to evaluate the effectiveness of a model, cross-validation methods are generally used. The two most common cross-validation schemes are n-fold and

jackknife [47, 48]. Moreover, an independent test dataset is also used for evaluation purposes. Jackknife test is considered the least arbitrary and provides unique results for a benchmark dataset [49]. However, we adopted the n-fold cross-validation scheme in this work to perform analyses in a timely manner. The n-fold cross-validation procedure is carried out as follows:

Step 1: Partition samples randomly into n roughly equal segments.
Step 2: Hold-out one segment as validation data and the remaining $n-1$ segments as training data.
Step 3: Using training data, estimate the parameters of the predictor.
Step 4: Using validation data, compute all the metrics.
Step 5: Partition of samples and computation of the metrics (from Step 1 to Step 4) are done n times to evaluate the average of performance metrics.

In this work, we used 6-, 8- and 10-fold cross-validation. The results are discussed in the following sections.

Filtering of samples to reduce imbalanced data

Our study uses a benchmark dataset consisting of 670 proteins (see section Benchmark Dataset). Among these proteins, there are 1,782 succinylated (positive set) and 18,344 non-succinylated (negative set) lysine residues, which leads to highly imbalanced sets (with a ratio of more than 10:1). This makes sense from the biological viewpoint, however, training any machine learning classifier with such imbalance between classes could strongly bias the classification results. To resolve this, we performed the k-nearest neighbors cleaning treatment [17]. In doing so, we first computed the number of neighbors to be used as a threshold by dividing the number of negative and positive samples. Consequently, the calculated ratio was 18,344/1,782 $\approx$ 10.29. A threshold of $\mathrm{k} = 10$ was first used for imbalance removal. This was done by removing a negative sample which has one (or more) of nearest neighbors as a positive sample. Because the number of negative instances was still relatively larger than that of positive instances, we iteratively computed new thresholds by multiplying the initial cut-off (k =

10) with different integers and used the resulting thresholds for imbalance removal until negative and positive samples were almost similar in size. After the k-nearest neighbors cleaning treatment, the positive set contained 1,782 samples and the negative set consisted of 1,872 samples with a threshold of 40. Unlike previous studies [17] we did not introduce hypothetical samples for training but kept the above positive and negative sets. These two sets were finally used to perform n-fold cross-validation and evaluate our predictor SucStruct.

Comparison with benchmark predictors

We compared SucStruct with two recently proposed predictors iSuc-PseOpt [17] and SuccinSite [18]. Web servers of both predictors are also available and already trained so that protein sequences can be easily uploaded to identify succinylated lysine residues. To compare the performance of these predictors with SucStruct we manually uploaded all the protein sequences [20, 21] to these web servers and retrieved their predictions. It should be noted here that the web servers were trained using most of the sequences. Therefore, it is possible that some of the uploaded sequences were already used for training the predictors, thus biasing their results. For these predictors we report their performance on the validation set (the segment of samples that were held-out for testing) during the n-fold cross-validation procedure. This validation set was not used in estimating SucStruct parameters. Furthermore, since it was not clear which samples were used for training, we are not able to report the AUC (area under the curve) measure for both predictors. However, we reported the AUC of SucStruct with 6-, 8-, and 10-fold cross-validation.

In Table 1, we provide the comparison of iSuc-PseOpt [17], SuccinSite [18] and SucStruct. It can be clearly observed that SucStruct is outperforming both iSuc-PseOpt [17] and SuccinSite [18] on sensitivity, accuracy and MCC metrics. The sensitivity was improved significantly by 19.3% over that of previous methods. Furthermore, accuracy was improved by 6.5% whereas MCC was reasonably improved by 22.2%. This is a significant improvement over previous prediction methods. However, although

SuccinSite [18] reached very high specificity (0.9017) it obtained quite low sensitivity (0.3019); i.e., around 70% of succinylated lysine residues were not detected.

Table 1: Comparison of two predictors with SucStruct (highest values are shown using bold faces).

Method	Sensitivity	Specificity	Acc	MCC
iSuc-PseOpt [17]	0.6150	0.7788	0.6990	0.3998
SuccinSite [18]	0.3019	**0.9017**	0.6092	0.2556
SucStruct (6-fold cross-validation)	**0.7334**	0.7548	**0.7444**	**0.4884**
SucStruct (8-fold cross-validation)	0.7273	0.7495	0.7386	0.4769
SucStruct (10-fold cross-validation)	0.7194	0.7447	0.7323	0.4642

We have also computed the AUC for SucStruct. The higher the AUC the better the predictor would be. For 6-, 8- and 10-fold cross-validation, the AUCs were 0.720, 0.722, and 0.719 (Fig. 2), respectively.

It can be clearly observed that SucStruct predictor is able to achieve extremely promising results. This is because it is utilizing important structural features of proteins such as accessible surface area, backbone torsion angles (e.g. ϕ, ψ, θ and τ) and local structure conformations (helix, strand and coil). The computation of these characteristics was possible with the toolbox SPIDER-2 [22], and they have proved to be exceptionally useful in determining succinylated lysines.

For the classification of succinylated lysine residues we designed a pruned decision tree. Using all the nine structural features, we generated a decision tree of 493 nodes for 10-fold cross-validation. This large number of nodes appeared to be the result of using 257-dimensional feature vectors. In order to illustrate a decision tree with smaller number of nodes, we further ranked the 257 features using an information gain ranking scheme. Our top ranked attributes of amino acids were related to secondary structure (pe and pc) and local backbone angles (τ and ψ). By retrieving the top two ranked features (pe_{14} and τ_{-11}) we generated a new decision tree of only 15 nodes (Fig. 3). Therefore, we could also simplify the initially large decision tree by performing feature

selection. All the four metrics were computed for this new decision tree and are reported in Table 2.

Table 2: Comparison of two benchmark predictors and SucStruct with selected features (highest values are shown using bold faces).

Method	Sensitivity	Specificity	Acc	MCC
iSuc-PseOpt [17]	0.6150	0.7788	0.6990	0.3998
SuccinSite [18]	0.3019	**0.9017**	0.6092	0.2556
SucStruct + feature selection (10-fold cross-validation)	**0.7946**	0.7286	**0.7608**	**0.5240**

It can be observed from Table 2 that sensitivity, accuracy and MCC were further increased whereas the recorded AUC was 0.807. These results are particularly significant and therefore can summarize that the structure of proteins plays a crucial role in discriminating between succinylated and non-succinylated lysines. It is then possible to improve the performance of current predictors by carefully using the described structural features. Further investigation in this direction would undoubtedly help to understand the lysine succinylation mechanism in a better manner. We will consider adding additional information, varied features or classifiers (i.e. ensemble of classifiers) in order to improve both sensitivity and specificity.

Biological insights

We further compared the sensitivity results of SucStruct and the two previously described predictors [17, 18]. This step was aimed at figuring out how well the succinylation sites of individual proteins were detected and the possible biological functions of such proteins (Supplementary material 2). Consequently, SucStruct was able to accurately detect succinylation sites in 132 proteins which were not discovered by iSuc-PseOpt [17] and SuccinSite [18] predictors. Most of these proteins only contained one or two succinylated lysines. Among the proteins with one succinylation site were the antioxidant enzyme methionine-R-sulfoxide reductase B2 (UniProtKB ID Q78J03) which is able to repair damaged methionine residues [50] and peptidylprolyl

isomerase (UniProtKB ID F6X9I3) which catalyzes the *cis–trans* isomerisation of peptide bonds to proline residues, folds synthesized proteins and is actively involved in the immune system [51]. Among the proteins with two succinylated residues were calcium-binding mitochondrial carrier protein SCaMC-1 (UniProtKB ID Q8BMD8) whose involvement in solute transport across the inner membrane of mitochondria and expression in testis have been studied [52]; Rab GDP dissociation inhibitor beta (UniProtKB ID Q61598) that results altered in mice subjected to prenatal stress [53]; and kynurenine aminotransferase III (UniProtKB ID Q71RI9) which catalyzes the transamination of kynurenine to kynurenic acid in mice [54].

Unlike predictors iSuc-PseOpt [17] and SuccinSite [18], SucStruct also predicted all the succinylation sites of acyl-coenzyme A synthetase ACSM1 (UniProtKB ID D3Z106) with a well-documented function in oxidation [55] and acyl-CoA dehydrogenases such as short-chain specific acyl-CoA dehydrogenase (UniProtKB ID Q07417), acyl-CoA dehydrogenase family member 11 (UniProtKB ID D3YTQ5) and long-chain specific acyl-CoA dehydrogenase (UniProtKB ID P51174). Acyl-CoA dehydrogenases have been reported to lead to mitochondrial dysfunction and alteration of fatty acid metabolism [56], catalyze the alpha and beta dehydrogenation of acyl-CoA esters in fatty acids and amino acids [57].

Additionally, SucStruct and iSuc-PseOpt [17] showed a similar performance with proteins like kynurenine/alpha-aminoadipate aminotransferase (UniProtKB ID Q9WVM8) which catalyzes the kynurenic acid synthesis [58] and heat shock protein 10 (UniProtKB ID Q64433) involved in the development of polycystic ovarian syndrome [59]. In a few cases, SucStruct and SuccinSite [18] produced similar results. For example, both predictors correctly identified the three modified lysines of heterogeneous nuclear ribonucleoprotein K (UniProtKB ID B2M1R6), a protein abundantly expressed in the nuclei of mouse oocytes [60]. Given the complex identification of succinylated residues there were 42 proteins whose succinylated lysines were not detected by any predictor. These proteins include peroxisomal sarcosine oxidase (UniProtKB ID Q9D826) which is known to be associated with the peroxisomal membrane [61].

In summary, we believe that the use of the mirror effect when a lysine residue is located near either terminus can somehow affect the sensitivity of SucStruct predictor. In doing

so, we are just using a copy of the structural features from the other side of the lysine which might not be biologically relevant. However, the combination of SucStruct with proposed predictors [17, 18] could have a significant impact on discovering more true lysine succinylation sites. Although previous approaches have consistently showed similar performance (high specificity and low sensitivity) the completely different performance of SucStruct could complement them.

For the benefit of the research community the information related to this study can be accessed at https://github.com/YosvanyLopez/SucStruct.

Conclusions

In this paper, we propose a new predictor (SucStruct) which combines structural features of proteins for classifying lysine succinylation sites. Characteristics such as accessible surface area, backbone torsion angles, and likelihood of amino acid contribution to local structure conformations proved to be important to discriminate between succinylated and non-succinylated residues. As previously reported, the k-nearest neighbors cleaning treatment helps to efficiently remove redundant negative samples. SucStruct showed a significant improvement in performance over previous benchmark predictors. The use of information gain to rank features also allows to considerably increase in the performance of the predictor in terms of sensitivity, accuracy and MCC. This further underscores the importance of certain secondary structures and local backbone angles in future studies on succinylation. Additionally, it would be interesting to further investigate the distribution of proteins with succinylation marks with respect to their subcellular localization. To the best of our knowledge, such studies have not been extensively proposed and they can provide important information on how biased these proteins could be with respect to their cellular location.

Acknowledgments

The authors acknowledge Jia *et al.* and Hasan *et al.* for allowing us to test our benchmark dataset with their predictors. YL and AS thank the members of Tsunoda

Laboratory for their comments and suggestions. This work was supported by the Grant-in-Aid for JSPS Fellows by the Japan Society for the Promotion of Science (15F15385) to Y.L.

References

[1] C.T. Walsh, S. Garneau-Tsodikova, G.J.G. Jr., Protein Posttranslational Modifications: The Chemistry of Proteome Diversifications, Angewandte Chemie International Edition, 44 (2005) 7342–7372.
[2] Y. Xu, K.-C. Chou, Recent Progress in Predicting Posttranslational Modification Sites in Proteins, Current Topics in Medicinal Chemistry, 16 (2016) 591-603.
[3] W.-R. Qiu, X. Xiao, W.-Z. Lin, K.-C. Chou, iMethyl-PseAAC: Identification of Protein Methylation Sites via a Pseudo Amino Acid Composition Approach, BioMed Research International, 2014 (2014) 947416.
[4] T. Hou, G. Zheng, P. Zhang, J. Jia, J. Li, L. Xie, C. Wei, Y. Li, LAceP: Lysine Acetylation Site Prediction Using Logistic Regression Classifiers, PLoS ONE, 9 (2014) e89575.
[5] W.-R. Qiu, X. Xiao, W.-Z. Lin, K.-C. Chou, iUbiq-Lys: prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a gray system model, Journal of Biomolecular Structure & Dynamics, 33 (2015) 1731-1742.
[6] Brian T. Weinert, C. Schölz, Sebastian A. Wagner, V. Iesmantavicius, D. Su, Jeremy A. Daniel, C. Choudhary, Lysine Succinylation Is a Frequently Occurring Modification in Prokaryotes and Eukaryotes and Extensively Overlaps with Acetylation, Cell Reports, 4 (2013) 842-851.
[7] Z. Zhang, M. Tan, Z. Xie, L. Dai, Y. Chen, Y. Zhao, Identification of lysine succinylation as a new post-translational modification, Nature Chemical Biology, 7 (2011) 58–63.
[8] O.N. Jensen, Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry, Current Opinion in Chemical Biology, 8 (2004) 33-41.
[9] J. Park, Y. Chen, Daniel X. Tishkoff, C. Peng, M. Tan, L. Dai, Z. Xie, Y. Zhang, Bernadette M.M. Zwaans, Mary E. Skinner, David B. Lombard, Y. Zhao, SIRT5-Mediated Lysine Desuccinylation Impacts Diverse Metabolic Pathways, Molecular Cell, 50 (2013) 919–930.
[10] Z. Xie, J. Dai, L. Dai, M. Tan, Z. Cheng, Y. Wu, J.D. Boeke, Y. Zhao, Lysine Succinylation and Lysine Malonylation in Histones, Molecular & Cellular Proteomics, 11 (2012) 100-107.
[11] X. Li, X. Hu, Y. Wan, G. Xie, X. Li, D. Chen, Z. Cheng, X. Yi, S. Liang, F. Tan, Systematic Identification of the Lysine Succinylation in the Protozoan Parasite Toxoplasma gondii, Journal of Proteome Research, 13 (2014) 6087-6095.
[12] G. Colak, Z. Xie, A.Y. Zhu, L. Dai, Z. Lu, Y. Zhang, X. Wan, Y. Chen, Y.H. Cha, H. Lin, Y. Zhao, M. Tan, Identification of Lysine Succinylation Substrates and the Succinylation Regulatory Enzyme CobB in Escherichia coli, Molecular & Cellular Proteomics, 12 (2013) 3509-3520.
[13] M. Yang, Y. Wang, Y. Chen, Z. Cheng, J. Gu, J. Deng, L. Bi, C. Chen, R. Mo, X. Wang, F. Ge, Succinylome Analysis Reveals the Involvement of Lysine Succinylation in Metabolism in Pathogenic Mycobacterium tuberculosis, Molecular & Cellular Proteomics, 14 (2015) 796-811.
[14] H.-D. Xu, S.-P. Shi, P.-P. Wen, J.-D. Qiu, SuccFind: a novel succinylation sites online prediction tool via enhanced characteristic strategy, Bioinformatics, 31 (2015) 3748-3750.

[15] X. Zhao, Q. Ning, H. Chai, Z. Ma, Accurate in silico identification of protein succinylation sites using an iterative semi-supervised learning technique, Journal of Theoretical Biology, 374 (2015) 60–65.
[16] Y. Xu, Y.-X. Ding, J. Ding, Y.-H. Lei, L.-Y. Wu, N.-Y. Deng, iSuc-PseAAC: predicting lysine succinylation in proteins by incorporating peptide position-specific propensity, Scientific Reports, 5 (2015) 10184.
[17] J. Jia, Z. Liu, X. Xiao, B. Liu, K.-C. Chou, iSuc-PseOpt: Identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset, Analytical Biochemistry, 497 (2016) 48-56.
[18] M.M. Hasan, S. Yang, Y. Zhou, M.N.H. Mollah, SuccinSite: a computational tool for the prediction of protein succinylation sites by exploiting the amino acid patterns and properties, Molecular BioSystems, 12 (2016) 786-795.
[19] S. Zhen, X. Deng, J. Wang, G. Zhu, H. Cao, L. Yuan, Y. Yan, First Comprehensive Proteome Analyses of Lysine Acetylation and Succinylation in Seedling Leaves of Brachypodium distachyon L., Scientific Reports, 6 (2016) 31576.
[20] Z. Liu, Y. Wang, T. Gao, Z. Pan, H. Cheng, Q. Yang, Z. Cheng, A. Guo, J. Ren, Y. Xue, CPLM: a database of protein lysine modifications, Nucleic Acids Research, 42 (2014) D531-D536.
[21] Z. Liu, J. Cao, X. Gao, Y. Zhou, L. Wen, X. Yang, X. Yao, J. Ren, Y. Xue, CPLA 1.0: an integrated database of protein lysine acetylation, Nucleic Acids Research, 39 (2011) D1029-D1034.
[22] R. Heffernan, K. Paliwal, J. Lyons, A. Dehzangi, A. Sharma, J. Wang, A. Sattar, Y. Yang, Y. Zhou, Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning, Scientific Reports, 5 (2015) 11476.
[23] E. Faraggi, T. Zhang, Y. Yang, L. Kurgan, Y. Zhou, SPINE X: Improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles, Journal of Computational Chemistry, 33 (2012) 259-267.
[24] L.J. McGuffin, K. Bryson, D.T. Jones, The PSIPRED protein structure prediction server, Bioinformatics, 16 (2000) 404-405.
[25] J. Lyons, A. Dehzangi, R. Heffernan, A. Sharma, K. Paliwal, A. Sattar, Y. Zhou, Y. Yang, Predicting Backbone Cα Angles and Dihedrals from Protein Sequences by Stacked Sparse Auto-Encoder Deep Neural Network, Journal of Computational Chemistry, 35 (2014) 2040-2046.
[26] E. Faraggi, Y. Yang, S. Zhang, Y. Zhou, Predicting Continuous Local Structure and the Effect of Its Substitution for Secondary Structure in Fragment-Free Protein Structure Prediction, Structure, 17 (2009) 1515-1527.
[27] R. Heffernan, A. Dehzangi, J. Lyons, K. Paliwal, A. Sharma, J. Wang, A. Sattar, Y. Zhou, Y. Yang, Highly accurate sequence-based prediction of half-sphere exposures of amino acid residues in proteins, Bioinformatics, 32 (2016) 843-849.
[28] Y. Yang, R. Heffernan, K. Paliwal, J. Lyons, A. Dehzangi, A. Sharma, J. Wang, A. Sattar, Y. Zhou, SPIDER2: A Package to Predict Secondary Structure, Accessible Surface Area, and Main-Chain Torsional Angles by Deep Neural Networks, in: Y. Zhou, A. Kloczkowski, E. Faraggi, Y. Yang (Eds.) Prediction of Protein Secondary Structure, Springer New York2016, pp. 55-63.
[29] G. Taherzadeh, Y. Zhou, A.W.-C. Liew, Y. Yang, Sequence-Based Prediction of Protein-Carbohydrate Binding Sites Using Support Vector Machines, Journal of Chemical Information and Modeling, 56 (2016) 2115-2122.

[30] G. Taherzadeh, Y. Yang, T. Zhang, A.W.-C. Liew, Y. Zhou, Sequence-based prediction of protein–peptide binding sites using support vector machine, Journal of Computational Chemistry, 37 (2016) 1223-1229.
[31] L. Lins, A. Thomas, R. Brasseur, Analysis of accessible surface of residues in proteins, Protein Science, 12 (2003) 1406–1417.
[32] B.-B. Pan, F. Yang, Y. Ye, Q. Wu, C. Li, T. Huber, X.-C. Su, 3D structure determination of a protein in living cells using paramagnetic NMR spectroscopy, Chemical Communications, 52 (2016) 10237-10240.
[33] O. Dor, Y. Zhou, Real-SPINE: An integrated system of neural networks for real-value prediction of protein structural properties, Proteins: Structure, Function, and Bioinformatics, 68 (2007) 76-81.
[34] B. Xue, O. Dor, E. Faraggi, Y. Zhou, Real-value prediction of backbone torsion angles, Proteins: Structure, Function, and Bioinformatics, 72 (2008) 427-433.
[35] A. Sharma, J. Lyons, A. Dehzangi, K.K. Paliwal, A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition, Journal of Theoretical Biology, 320 (2013) 41-46.
[36] A. Dehzangi, R. Heffernan, A. Sharma, J. Lyons, K. Paliwal, A. Sattar, Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou's general PseAAC, Journal of Theoretical Biology, 364 (2015) 284-294.
[37] K.K. Paliwal, A. Sharma, J. Lyons, A. Dehzangi, A Tri-Gram Based Feature Extraction Technique Using Linear Probabilities of Position Specific Scoring Matrix for Protein Fold Recognition, IEEE Transactions on NanoBioscience, 13 (2014) 44-50.
[38] A. Dehzangi, S. Sohrabi, R. Heffernan, A. Sharma, J. Lyons, K. Paliwal, A. Sattar, Gram-positive and gram-negative subcellular localization using rotation forest and physicochemical-based features, BMC Bioinformatics, 16 (2015) S1.
[39] R. Sharma, A. Dehzangi, J. Lyons, K. Paliwal, T. Tsunoda, A. Sharma, Predict Gram-Positive and Gram-Negative Subcellular Localization via Incorporating Evolutionary Information and Physicochemical Features Into Chou's General PseAAC, IEEE Transactions on NanoBioscience, 14 (2015) 915-926.
[40] J.R. Quinlan, C4.5: Programs for Machine Learning, First Edition ed., Morgan Kaufmann, San Francisco, California, USA, 1992.
[41] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA Data Mining Software: An Update, SIGKDD Explorations, 11 (2009) 10-18.
[42] Z. Liu, X. Xiao, W.-R. Qiu, K.-C. Chou, iDNA-Methyl: Identifying DNA methylation sites via pseudo trinucleotide composition, Analytical Biochemistry, 474 (2015) 69-77.
[43] W. Chen, P. Feng, H. Ding, H. Lin, K.-C. Chou, iRNA-Methyl: Identifying N6-methyladenosine sites using pseudo nucleotide composition, Analytical Biochemistry, 490 (2015) 26–33.
[44] B. Liu, L. Fang, S. Wang, X. Wang, H. Li, K.-C. Chou, Identification of microRNA precursor with the degenerate K-tuple or Kmer strategy, Journal of Theoretical Biology, 385 (2015) 153-159.
[45] H. Ding, E.-Z. Deng, L.-F. Yuan, L. Liu, H. Lin, W. Chen, K.-C. Chou, iCTX-Type: A Sequence-Based Predictor for Identifying the Types of Conotoxins in Targeting Ion Channels, BioMed Research International, 2014 (2014) 286419.

[46] X. Xiao, J.-L. Min, W.-Z. Lin, Z. Liu, X. Cheng, K.-C. Chou, iDrug-Target: predicting the interactions between drug compounds and target proteins in cellular networking via benchmark dataset optimization approach, Journal of Biomolecular Structure and Dynamics, 33 (2015) 2221-2233.
[47] K.-C. Chou, H.-B. Shen, Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms, Nature Protocols, 3 (2008) 153-162.
[48] E. Alpaydin, Introduction to Machine Learning, Third ed., The MIT Press2014.
[49] Z. Hajisharifi, M. Piryaiee, M.M. Beigi, M. Behbahani, H. Mohabatkar, Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test, Journal of Theoretical Biology, 341 (2014) 34-40.
[50] F.L. Aachmann, G.-H. Kwak, R.D. Conte, H.-Y. Kim, V.N. Gladyshev, A. Dikiy, Structural and Biochemical Analysis of Mammalian Methionine Sulfoxide Reductase B2, Proteins, 79 (2011) 3123-3131.
[51] P.E. Shaw, Peptidyl-prolyl isomerases: a new twist to transcription, EMBO Reports, 3 (2002) 521-526.
[52] I. Amigo, J. Traba, J. Satrústegui, A.d. Arco, SCaMC-1Like a Member of the Mitochondrial Carrier (MC) Family Preferentially Expressed in Testis and Localized in Mitochondria and Chromatoid Body, PLoS One, 7 (2012) e40470.
[53] R. Orlando, M. Borro, M. Motolese, G. Molinaro, S. Scaccianoce, A. Caruso, L.d. Nuzzo, F. Caraci, F. Matrisciano, A. Pittaluga, J. Mairesse, M. Simmaco, R. Nisticò, J.A. Monn, F. Nicoletti, Levels of the Rab GDP dissociation inhibitor (GDI) are altered in the prenatal restrain stress mouse model of schizophrenia and are differentially regulated by the mGlu2/3 receptor agonists, LY379268 and LY354740, Neuropharmacology, 86 (2014) 133-144.
[54] Q. Han, H. Robinson, T. Cai, D.A. Tagle, J. Li, Biochemical and Structural Properties of Mouse Kynurenine Aminotransferase III, Molecular and Cellular Biology, 29 (2009) 784-793.
[55] T. Fujino, Y.A. Takei, H. Sone, R.X. Ioka, A. Kamataki, K. Magoori, S. Takahashi, J. Sakai, T.T. Yamamoto, Molecular Identification and Characterization of Two Medium-chain Acyl-CoA Synthetases, MACS1 and the Sa Gene Product, The Journal of Biological Chemistry, 276 (2001) 35961-35966.
[56] W. Wang, A.-W. Mohsen, G. Uechi, E. Schreiber, M. Balasubramani, B. Day, M.M. Barmada, J. Vockley, Complex changes in the liver mitochondrial proteome of short chain acyl-CoA dehydrogenase deficient mice, Molecular Genetics and Metabolism, 112 (2014) 30-39.
[57] Z. Swigoňová, A.-W. Mohsen, J. Vockley, Acyl-CoA Dehydrogenases: Dynamic History of Protein Family Evolution, Journal of Molecular Evolution, 69 (2009) 176-193.
[58] Q. Han, T. Cai, D.A. Tagle, J. Li, Structure, expression, and function of kynurenine aminotransferases in human and rodent brains, Cellular and Molecular Life Sciences, 67 (2010) 353-368.
[59] K.-K. Zhao, Y.-G. Cui, Y.-Q. Jiang, J. Wang, M. Li, Y. Zhang, X. Ma, F.-Y. Diao, J.-Y. Liu, Effect of HSP10 on apoptosis induced by testosterone in cultured mouse ovarian granulosa cells, European Journal of Obstetrics & Gynecology and Reproductive Biology, 171 (2013) 301-306.
[60] P. Zhang, N. Wang, X. Lin, L. Jin, H. Xu, R. Li, H. Huang, Expression and localization of heterogeneous nuclear ribonucleoprotein K in mouse ovaries and preimplantation embryos, Biochemical and Biophysical Research Communications, 471 (2016) 260-265.

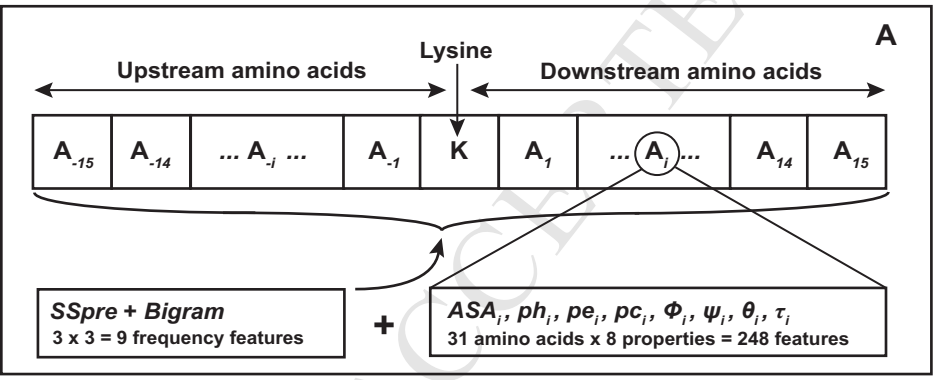
[61] M. Chikayama, M. Ohsumi, S. Yokota, Enzyme cytochemical localization of sarcosine oxidase activity in the liver and kidney of several mammals, Histochemistry and Cell Biology, 113 (2000) 489-495.

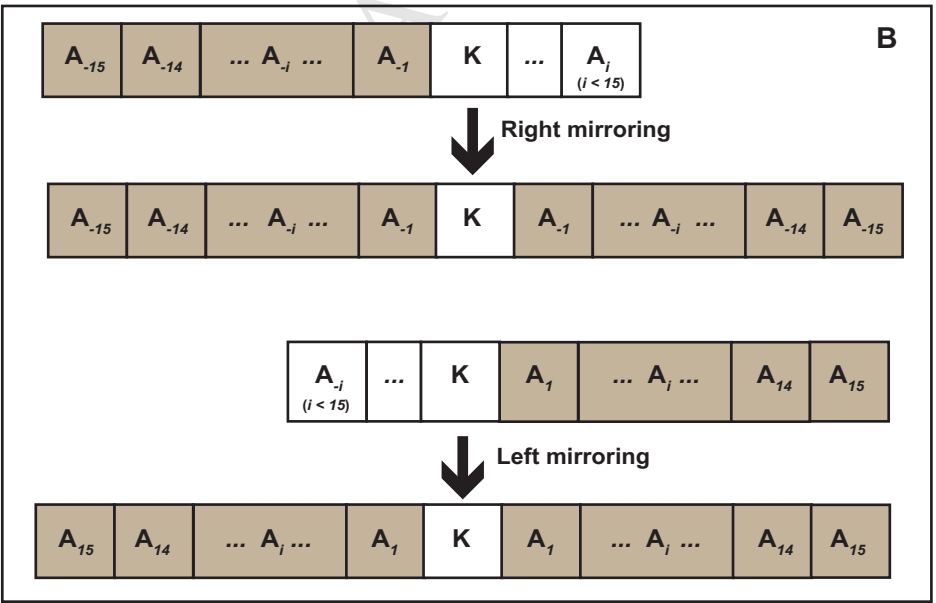
Figures

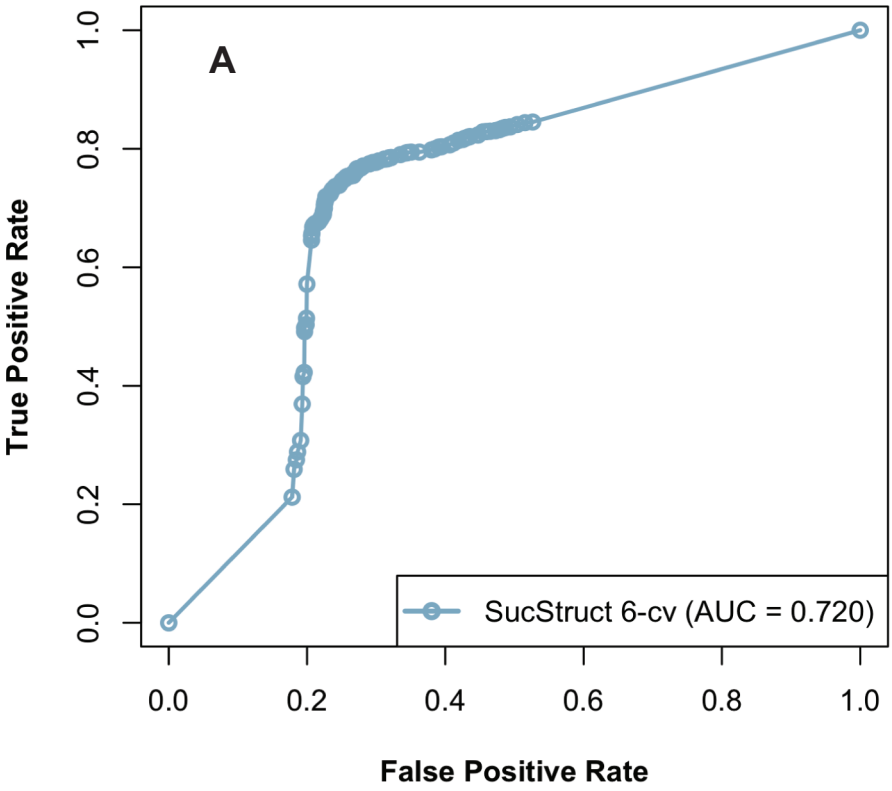
Fig. 1. Schematic illustration of how the amino acids adjacent to a lysine residue were regarded. (A) lysine with all the 15 amino acids upstream and downstream. (B) lysine with missing downstream (right mirroring) and upstream (left mirroring) amino acids.

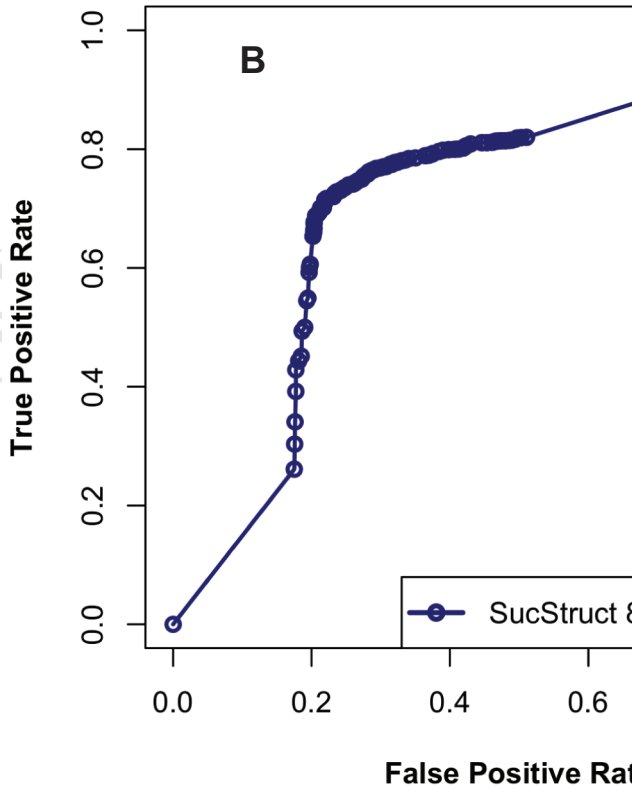
Fig. 2. Receiver operating characteristic of SucStruct for (A) 6-, (B) 8- and (C) 10-fold cross-validations.

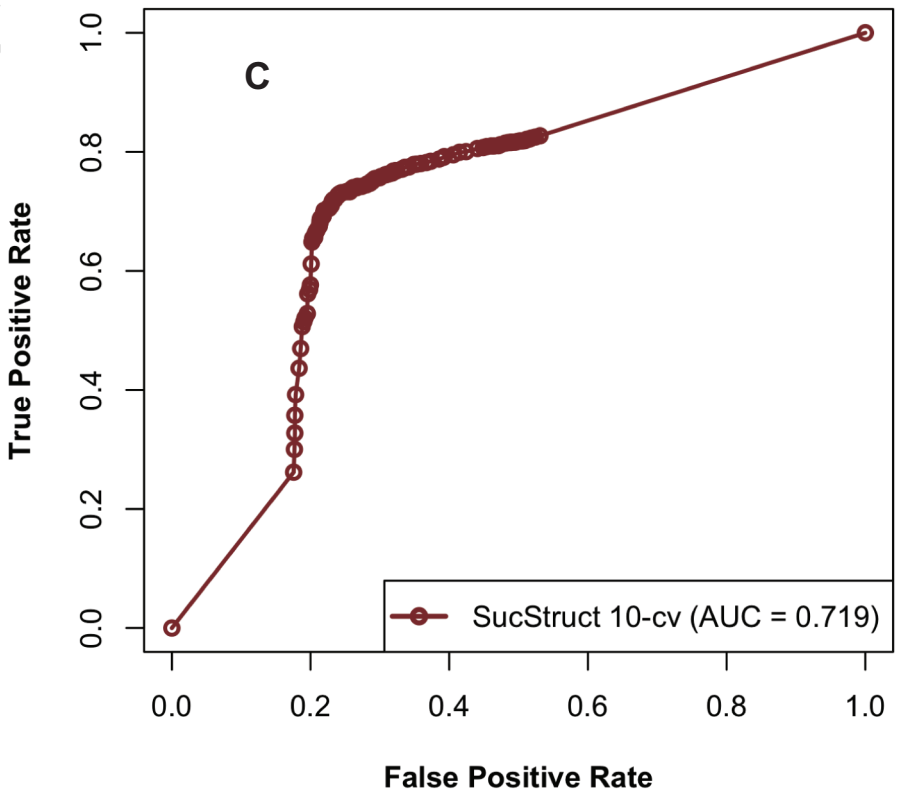
Fig. 3. Decision tree using the top two ranked features. Green and pink nodes represent succinylated and non-succinylated lysine residues, respectively.

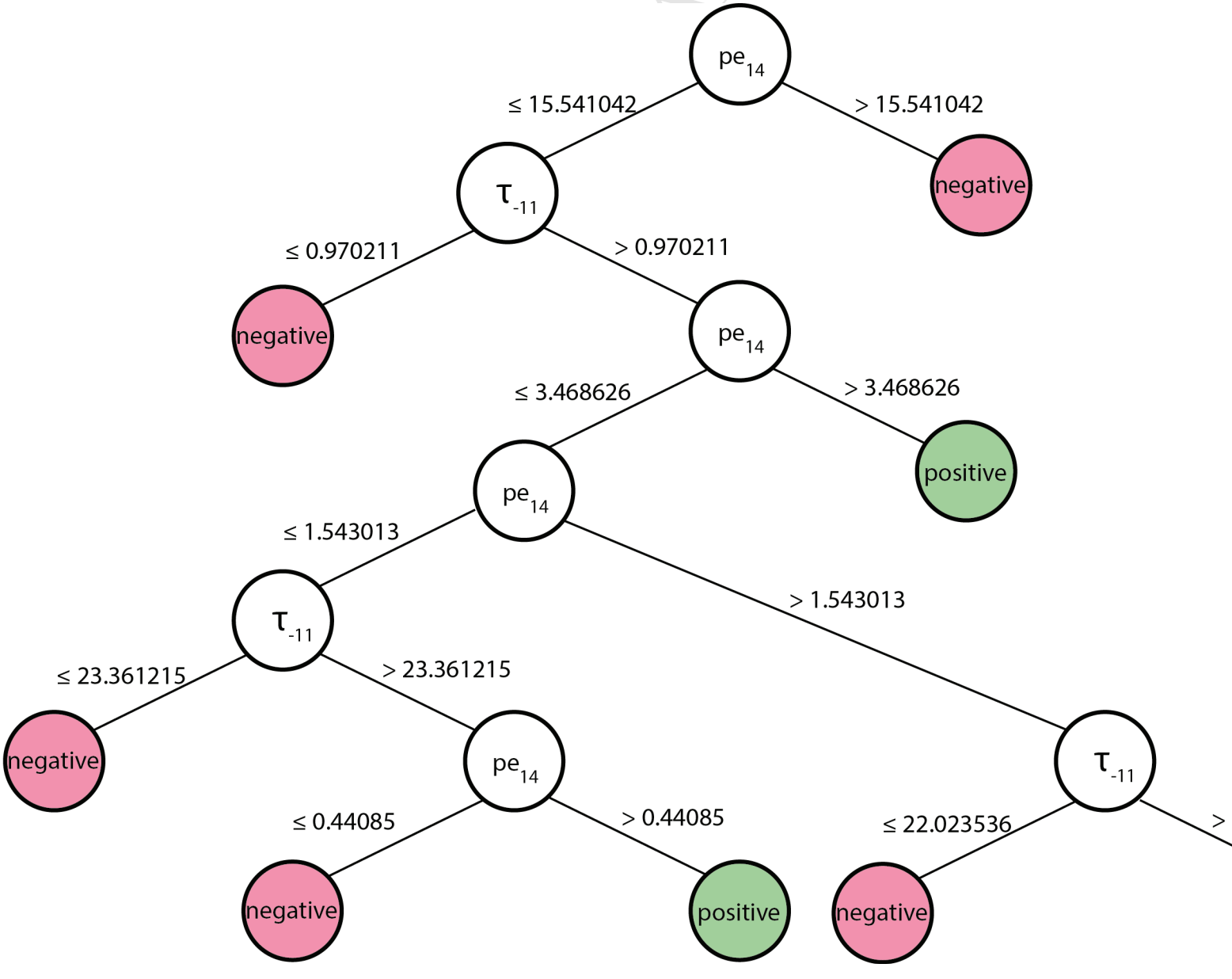











pe14
≤ 15.541042
> 15.541042
τ-11
negative
≤ 0.970211
> 0.970211
negative
pe14
≤ 3.468626
> 3.468626
positive
pe14
≤ 1.543013
> 1.543013
τ-11
≤ 23.361215
> 23.361215
negative
pe14
τ-11
≤ 0.44085
> 0.44085
≤ 22.023536
negative
positive
negative