American Journal of Biomedical and Life Sciences
2017; 5(6): 135-143
http://www.sciencepublishinggroup.com/j/ajbls
doi: 10.11648/j.ajbls.20170506.15
ISSN: 2330-8818 (Print); ISSN: 2330-880X (Online)

DV-iSucLys: Decision Voting to Improve Protein Lysine Succinylation Site Identification from Sequence Data

Md. Khaled Ben Islam[1, 2, *], Md. Nazrul Islam Mondal[1], Julia Rahman[1], Md. Al Mehedi Hassan[1]

[1]Department of Computer Science & Engineering, Rajshahi University of Engineering & Technology, Rajshahi, Bangladesh

[2]Department of Computer Science & Engineering, Pabna University of Science & Technology, Pabna, Bangladesh

Email address:

mdkhaledben@gmail.com (Md. K. B. Islam)

[*]Corresponding author

To cite this article:

Md. Khaled Ben Islam, Md. Nazrul Islam Mondal, Julia Rahman, Md. Al Mehedi Hassan. DV-iSucLys: Decision Voting to Improve Protein Lysine Succinylation Site Identification from Sequence Data. *American Journal of Biomedical and Life Sciences*. Vol. 5, No. 6, 2017, pp. 135-143. doi: 10.11648/j.ajbls.20170506.15

Received: September 8, 2017; **Accepted**: October 8, 2017; **Published**: November 30, 2017

Abstract: Protein Post Translation Modification identification is one of the important steps in conducting disease-associated mutation studies. Though multiple chemical alterations happen in a protein after translation, the addition of succinyl group to lysine residue plays a vital role in regulating cellular metabolism and thus disease. Use of a classification algorithm on some features, driven either from protein structural, physicochemical or even biochemical information becomes a common approach that can yield a satisfactory result up to a certain level. Although, researchers already developed many computational methods to identify whether a lysine residue modified with succinyl group after translation, most of them focused on the improvement either on a single decision using a single method or feature enrichment or even development of a benchmark dataset. Therefore, there still exists scope for further improvement to characterise lysine residues of a protein sequence by considering multiple predictors at a time. In this study, an ensemble based approach called DV-iSucLys has been designed to characterise the lysine residue by adapting three well known and conceptually different classifiers and ensembling their decisions. Also, a benchmark succinylation dataset was extracted from existing benchmark datasets and recently updated succinylation data from UniProt consortium to investigate the performance of the proposed approach as well as contribute to further research. Analysing rigorous cross-validation results show that DV-iSucLys can characterise succinyl lysine residue better than the existing predictors.

Keywords: Lysine Succinylation, AAC, CKSAAP, Binary Encoding, PSAAP, AAindex, Ensemble Classifier

1. Introduction

A small number of genes (20,000−25,000) operate human life by encoding multiple proteins from single gene. Among different mechanism of genetic code expedition, protein post-translational modification (PTM) is one of the most significant biological processes which extends the functional diversity of the proteome by the covalent addition of functional groups or proteins, proteolytic cleavage of regulatory subunits or degradation of entire proteins. More than 300 different types of PTMs are distinguished [1] in vivo. Among evolutionary conserved PTMs, Lysine succinylation is one of them which was first discovered to occur at the active site of homoserine trans-succinylase [2] and available in both eukaryotes and prokaryotes. The importance of lysine succinylation is immense in terms of changes in protein structure as well as function, regulation of the physicochemical property of protein, protein conformation space and protein stability. Nonetheless, the details about the full regulatory role of succinylation are still an elusive issue. Identification of succinylation sites is considered as the most challenging and crucial topics for the researchers, not only for addressing the mechanism and function of protein succinylation which is very useful for both biomedical research and drug development but also for the availability of enormous amount of protein sequence data by blessings of genome projects.

The traditional wet-lab experimental methods for

succinylation site prediction are expensive, laborious and face uncertain time boundary to meet the research demands, especially for large-scale datasets. Additionally, post-genomic era generates a huge amount of protein sequences which are helpful for computational techniques. As a result, the automated computational system is highly desirable to predict succinylation sites. Currently, some computational methods based on machine learning approaches are available to predict protein succinylation sites, and much progress has already been achieved in this direction.

Zhao et al. [3] developed SucPred, a support vector machine (SVM) based succinylation site predictor which used protein sequence based multiple feature encoding schemes (autocorrelation functions, grouped weight based encoding, positional weight amino acids composition and normalised van der Waals volume). Another SVM based predictor SuccFind [4] was developed using both sequences driven feature (k-space amino acid pairs) and evolution-derived information of sequence (amino acid index (AAindex) properties)). iSuc-PseAAC predictor [5] incorporated the peptide position specific propensity into the general form of PseAAC (Pseudo Amino Acid Composition) for training support vector machine. iSuc-PseOpt [6] integrated the sequence-coupling effects into the general pseudo amino acid composition, solved class imbalance problem and applied random forest algorithm for prediction. SuccinSite [7] used a random forest classifier, incorporating three sequence encoding features such as the composition of k-spaced amino acid pairs, binary encoding and amino acid index. pSuc-Lys [8] has incorporated the sequence-coupled information into the general pseudo amino acid composition and used ensemble random forest as a classifier. In ILSES [9] several physicochemical properties of succinylated sites have been extracted, namely the physicochemical property of the amino acids and a flexible neural tree has been employed as the classification model. SucStruct [16] use k-nearest neighbours cleaning method for imbalanced data and pruned decision tree for classification of succinylated sites based on structural features of amino acids. However, these predictors have shown poor sensitivity in detecting succinylated lysine residues. Therefore, additional efforts are still needed for improving the prediction.

From the above exploration, it is evident that different researchers use different algorithms as well as distinct features. This study hypothesis that combining multiple decisions on a single issue will lower the chance of error. Considering this hypothesis, an ensemble based lysine residue classifier has been designed and tested.

Based on this hypothesis, major contributions of this work are-

Firstly, this study contemplates five commonly used feature extraction techniques for lysine residue characterisation related to sequence based, physicochemical and biochemical properties based information of proteins.

Secondly, a classifier named DV-iSucLys has been developed and compared the performance with three baseline classifiers such as K-nearest neighbour (KNN), Support Vector Machine (SVM) and Random Forest (RF) for succinylation site prediction regarding accuracy, sensitivity, specificity and MCC metrics.

Thirdly, a focus was given to the development of an updated benchmark succinylation protein dataset for further research.

2. Materials and Methods

2.1. Datasets

To construct a robust benchmark dataset, experimentally validated protein sequences with lysine succinylation site details were collected from SwissProt/UniProt (retrieved on 21 May 2017) [17] and Compendium of Protein Lysine Modifications database curated by CUCKOO Workgroup [18]. Initially, 897 proteins with 2523 verified succinylation site (i.e. positive site) from different species were extracted. All the non-succinylation sites (i.e. negative site in total 24669) were also extracted from the same protein sequences to maintain consistency. For formulating any post-translational modification site (PTM), a de-facto standard used by the researchers [19-22] is to extract the PTM site centred (in this experiment, lysine centred) peptide segments of an optimal size. These peptide segments can be expressed as-

$$\text{Peptide}_{\text{PtmSite}} = \text{R}_{-\text{n}}\text{R}_{-(\text{n}-1)} \dots \text{R}_{-2}\text{R}_{-1}\text{KR}_{+1}\text{R}_{+2} \dots \text{R}_{+(\text{n}-1)}\text{R}_{+\text{n}} \quad (1)$$

where, centred K represents the amino acid residue "lysine" and n represents the maximum length of each side of a considered PTM site, so as each peptide segments will be of size (2n + 1). Lower sized peptide segments were padded with non-existing amino acid residue 'O' to keep the consistency of window size of PTM sites.

A peptide segment, represented in the form (equation 1) is considered positive sample if it is centred "K" is experimentally verified as Succinylation site (suc site); otherwise it is considered as a negative sample. i.e.

$$\text{Peptide}_{\text{PtmSite}} \in \begin{cases} \text{P}^{+}_{\text{Suc}}, \text{if Centered K residue is verified as suc site} \\ \text{P}^{-}_{\text{Suc}}, \text{Otherwise} \end{cases}$$

where $\text{P}^{+}_{\text{Suc}}$ and $\text{P}^{+}_{\text{Suc}}$ represents positive and negative set of data respectively.

As the benchmark dataset contains both succinylation positive site and negative site, which was used for training and testing (subset by subset) the considered computational techniques, thus the benchmark dataset can be represented as-

$$\text{S} = \text{S}^{+} \cup \text{S}^{-}$$

where $\text{S}^{+} \in \text{P}^{+}_{\text{Suc}}$, $\text{S}^{-} \in \text{P}^{-}_{\text{Suc}}$

To develop a robust benchmark dataset, multiple experiments was carried out with different values of n (in equation 1) for selecting optimal peptide segment size i.e. (2n+1). By doing this, a set of benchmark datasets have been obtained.

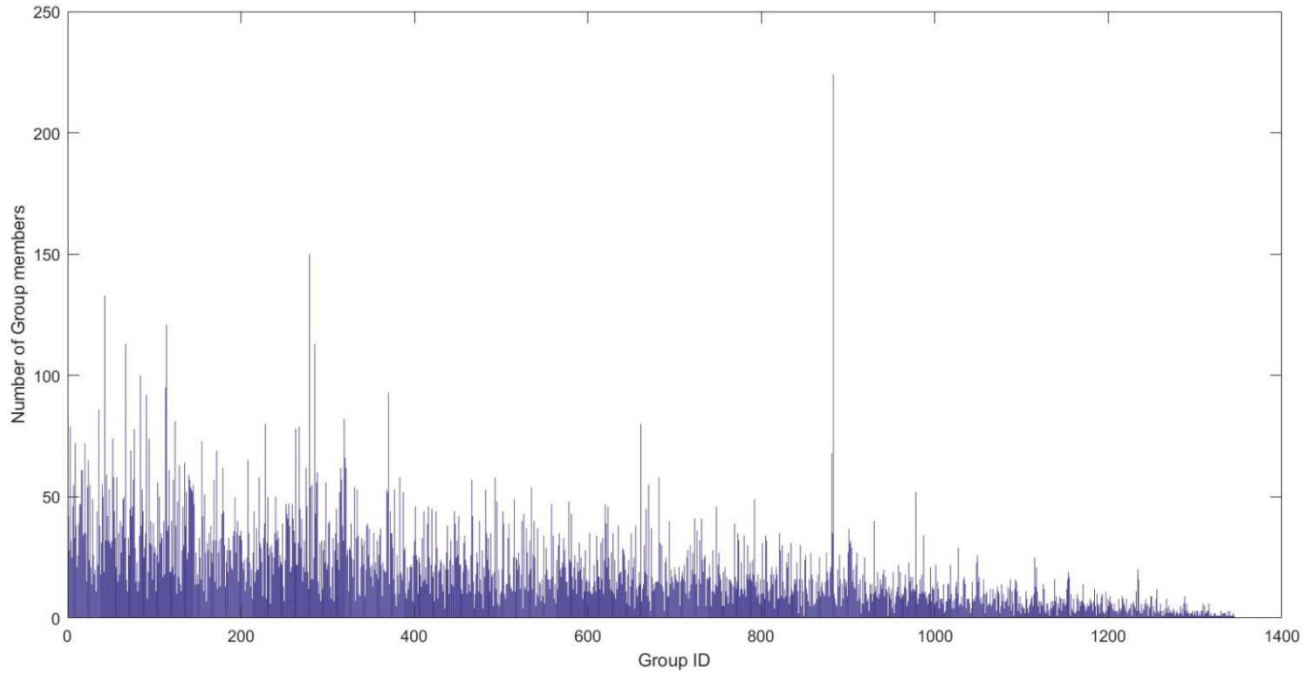

Figure 1. *Rate of Homology (No of homologous site in each homology group).*

In PTM Site prediction, the presence of homology and redundancy in the peptide segments may bias or overestimate the performance of the predictors, which is also mentioned by different PTM site researchers [19-23]. This study investigates this issue by considering less than 40% pairwise sequence identity in different initial benchmark datasets (for different segment size) using clustering. These investigations had indicated that data set with n=5, hence segment size = 2n+1 = 2*5+1=11 would be most promising. In Figure 1, for segment size 11, the rate of homologous site is presented in the form of a graph. It shows that, there was a significant number of homologous sites in the initial dataset and wipe of these sites were required for unbiased evaluation. For each group, the most representative segments was kept by prioritising the verified succinylation sites, as the number of non-verified and non-succinylation sites were initially so high. Finally, a non-homologous and non-redundant dataset was formed having 796 positive sites with corresponding negative sites in 2:1 ratio (random selection).

2.2. Features

Alphabetic Sequence of Amino Acids is a well-known representation of protein samples and peptides which can be directly derived from raw protein sequence. To use effectively in computational tools to predict the succinylation site, peptides should be converted into an effective mathematical expression. In most cases, it is called feature vector which can be represented as in equation (2), also suggested by Y. Xu et al. in a similar form in [5].

$$\text{Peptide, } P = (C_1, C_2, \ldots, C_d)^T \quad (2)$$

where, T = transpose operator and

d = vector's dimension (integer value).

The value of d as well as the components Cd depends on the technique of extracting the desired information from the protein or peptide sequences.

In our considered case, peptide P can be represented as-

$$P = (R_1, R_2, \ldots, R_6, \ldots, R_{11})^T \quad (3)$$

here, R_6 = K and

R_i (i = 1, 2, …, 11; i ≠ 6) = any of the 21 considered amino acid residue (any of 20 standard residue or non-existing amino acid residue 'O').

2.2.1. Amino Acid Composition (AAC)

Amino Acid Composition is a simple and commonly used method of feature extraction technique, based on calculating the proportion of each amino acid in peptide sequence. Mathematically, this technique can be expressed as-

$$X_i = \frac{\text{count(i)}}{N};\ i \in [1, \ldots\ldots\ldots, 20] \quad (4)$$

where, count(i) computes the number of occurrences of i^{th} amino acid within N length protein sequence.

This representation also disccused by L. Nanni, A. Lumini, and S. Brahnam [10].

2.2.2. CKSAAP

CKSAAP was developed by Chen et al. [11], and now it is widely used in many bioinformatics research. In our study, a protein sequence is fragmented by window size = 2n+1. The number of amino acid residues is 21; 20 basic amino acids and including gap represented by "O". Thus, total amino acid pairs are (21*21) =441 like as AA, AC, AD,..., OO for every single k (integer) where k denotes the space between two amino acids. For example, "AA" means k-space=0, "AXA" means

k-space=1, "AXXA" means k-space=2. In this work, k_{max} = 5 and it produced 21*(k_{max}+1)*21 = 2646 different amino acid pairs which were used in feature vector for each segment sequence. The feature vector was calculated using the following equation which was also used in the design of SuccinSite tool [7]:

$$\left(\frac{N_{AA}}{N_{Total}}, \frac{N_{AC}}{N_{Total}}, \ldots, \frac{N_{OO}}{N_{Total}}\right)_{441} \quad (5)$$

where,

N_{Total} = length of the total composition residues

N_{AA}, N_{AC},…, N_{OO} = frequency of amino acid pair within fragment

If the selected window size of sequence is n and k = 0, 1, 2, 3, 4, 5 then N_{Total} = n - k - 1.

2.2.3. Binary Encoding

Binary amino acid encoding which featured in [7], calculates the positional information from the corresponding amino acids sequence fragments. In this study, all the 20 standard amino acid residues were considered and non-existing "O" residue was used for gap or padding purpose. In total, 21 amino acid residues are ordered as ACDEFGHIKLMNPQRSTVWYO. These amino acids were transformed into numeric values for adopting a binary vector. For example, A was represented as 100000000000000000000, C as 010000000000000000000 and so on.

Since the window size of the peptides to be encoded is always "K" centred, so encoding all the peptides using same binary coded values in the same position would not carry any significant PTM information. For this reason, centred "K" was not considered into account at the time of encoding. As a result, final feature vectors dimension of binary encoding is ((n - 1) * 21).

2.2.4. Amino Acid Index (AAindex)

Amino Acid Index representing different physicochemical and biological properties of amino acids are used as the informative feature in different predictors (e.g. SuccinSite [7] and SuccFind [4]). In this experiment, both Physicochemical and biochemical properties of amino acids were extracted from AAindex database, version 9.1 [12]. In this study, all the 544 biochemical and biological indices from AAindex database are taken into account. Overrepresentation of zeroes and incomplete data in those indices were pre-processed using zero-based replacement strategy. As a result, all the 544 physicochemical properties were considered as potential features for representing the PTM sites. Thus, the dimension of feature vector becomes (n * 544), where n is the window size.

2.2.5. Position Specific Amino Acid Propensity (PSAAP)

Position Specific Amino Acid Propensity (PSAAP) is a feature of incorporating Peptide Position Specific Propensity into the general form of PseAAC. This information presentation technique is formulated as the following matrix, presented in [5] as:

$$Z = \begin{bmatrix} z_{1,1} & \cdots & z_{1,n} \\ \vdots & \ddots & \vdots \\ z_{21,1} & \cdots & z_{21,n} \end{bmatrix}_{21 \times n} \quad (6)$$

where,

$$z_{i,j} = F^{+}(R_i \mid j) - F^{-}(R_i \mid j) \quad (7)$$

$$i = 1, 2, \ldots, 21 \text{ and } j = 1, 2, \ldots, n$$

$F^{+}(R_i \mid j)$ is the occurrence frequency of the i^{th} amino acid (i = 1, 2, …, 21) in the j^{th} column in the positive benchmark dataset S^{+}, while $F^{-}(R_i \mid j)$ is the corresponding occurrence frequency and derived from the negative benchmark dataset S^{-}. The centred amino acid K was excluded as it was the same in positive and negative peptides (samples) respectively. Thus, the components in Equation (1) can be uniquely defined by:

$$P \in \begin{cases} z_{1,u} \; When \; R_i = A \\ z_{2,u} \; When \; R_i = C \\ \quad . \\ \quad . \\ \quad . \\ z_{21,u} \; When \; R_i = X \end{cases} \quad (8)$$

where, u indicate a particular component.

To encode a potential site (i.e. a fragment of 11 amino acids), one 11-dimensional feature vector (X) was constructed by looking up the corresponding parameters from the above matrix, presented in Equation (6), which was further explained in the following example which was earlier used in [13].

If a succinylated "K" residue was presented by the following 27 residues long fragment instead of 11 length like Equation (9) -

$$\overbrace{\text{G E S}\ldots\text{V}} \; K \; \overbrace{\text{P}\ldots\text{P P D}} \quad (9)$$

Then, the corresponding feature vector (X) would derived as-

$$X = (x_1, x_2, \ldots, x_{26}) = \overbrace{z_{6,1}, z_{4,2}, \ldots, z_{18,12}}^{13\ Downstream} \times \overbrace{z_{13,13}, \ldots, z_{13,22}, z_{3,26}}^{13 Upstream} \quad (10)$$

In the above Equation (9), x_1 would encoded by glycine (G) in the first position of 13 downstream residues. Since G is alphabetically ranked at the sixth position among the 20 amino acids, the corresponding value of x_1 would $z_{6,1}$. Analogous to x_1, the values of x_2, x_3,…, x_{26} also could be obtained as described in Equation (10). Since the matrix Z reflects the position-specific amino acid propensity surrounding the phosphorylation sites, this encoding system is known as PSAAP feature.

2.3. Classifier

Choosing and designing effective classifier is a crucial step in succinylation sites prediction. For prediction of succinylation sites, a variety of machine learning algorithms have already used, namely Support Vector Machine (SVM), Random Forest (RF), Ensemble Random Forest and so on.

This study has considered three different types of algorithms as base algorithm among them – KNN, SVM and RF.

2.3.1. k-Nearest Neighbor (KNN)

Among the non-parametric machine learning methods used for classification and regression, K-Nearest Neighbour (KNN) is a technically simplest one. In both cases, the input consists of the K closest training examples in the feature space. Here, KNN has been used as a classifier. The output is a class membership. In its simplest form, classification is performed based on majority voting. A particular target instance is classified by a majority vote of its neighbours. The target instance is assigned to the class most common among its k nearest neighbours. If k = 1, then the instance is simply assigned to the class of that single nearest neighbour. If k >= 3, then the instance is assigned to the class of major nearest neighbours. The value of k is normally odd. KNN is a type of instance-based learning, or lazy learning, where the function is only approximated locally, and all computation is deferred until classification. The closest peptides of the target peptide, p^t is defined as-

$$NN_{set}(p^t) = \{q^i : d(q^i, p^t) \leq d_{(k)}\} \quad (11)$$

$$i = 1,2, \ldots, N$$

where,

$$d(q^i, p^t) = ||q^i - p^t||$$

i.e. $d(q^i, p^t)$ is the distance between q^i and p^t in Euclidean space and $d_{(k)}$ is the k^{th} order statistic of $\{d(q^i, p^t)\}^n$.

2.3.2. Support Vector Machine (SVM)

Support Vector Machine (SVM) is one of the most popular supervised learning algorithms that are used for classification, regression and outlier's detection. In this experiment, SVM is used for classification. In general, Support Vector Machine is a method of obtaining the optimal boundary of two sets in a vector space independently on the probabilistic distributions of training vectors in the sets. Its fundamental idea is to locate the boundary that is most distant from the vectors nearest to the boundary in both of the sets. It is useful for the linear boundary. In case of nonlinear boundary, kernel trick is introduced where a deformation of the vector space itself to a higher dimensional space is occurred. As kernel function crucially influence the classification of SVM, Redial Basis Function (RBF) was used as a kernel in this study. For each target peptide P_t, the decision about succinylation sites will be made by SVM as:

$$S_j(x^t) = \sum_{i=1}^{n} \alpha_i y_i K(x^i, x^t) + b \quad (12)$$

where, $K(x^i, x^t) = exp(-\gamma ||x^i - x^t||^2)$

$$y_i \in (+1, -1),$$

$$\gamma = \frac{1}{2\sigma^2}$$

σ is the width of the function,

α_i is the Lagrange multipliers.

The prediction rule for query peptide P can be formulated as-

$$P \in \begin{cases} \text{succinylated peptide}, if\ y_i = +1 \\ \text{nonsuccinylated peptide}, otherwise \end{cases} \quad (13)$$

2.3.3. Random Forest (RF)

As a widely used supervised learning algorithm in bioinformatics, Random forest is an ensemble of decision trees which act as both classification and regression tree. These decision trees are constructed and trained by different bootstrap samples of original data. When a new object comes for classification, each tree of the forest gives their opinion about its class and output is formed based on the majority voting of class. It is relatively robust to noise and outliers. This technique also used in other domains like Intrusion detection [14]. Like M. A. M. Hasan et al. [14], its considered workflow can be described as-

1) From the Training of n samples draw n_{tree} bootstrap samples.

2) For each of the bootstrap samples, grow classification tree with the following modification:

At each node, rather than choosing the best split among all predictors, randomly sample m_{try} of the predictors and choose the best split among those variables. The tree is grown to the maximum size and not pruned back.

3) Predict new data by aggregating the predictions of the n_{tree} trees (i.e., majority votes for classification).

3. Implementation

3.1. Selection of K, Kernel Function and Parameters

Parameter selection for the classifier is one of the important issues for getting the best performance of it. For SVM, RBF kernel function has been chosen to test the developed datasets. In RBF kernel, the best combination of γ and regularisation coefficient c is responsible for obtaining highest accuracy. In this work, it was observed that higher accuracy for the training dataset in KNN and SVM classifiers were obtained when k, γ and c were certain values. These parameters were selected from the sets for k = {1, 2, 3,..., 20}, $\gamma = \{2^{-8}, 2^{-7}, \ldots, 2^0, 2^1, \ldots, 2^8\}$ and $c = \{2^{-2}, 2^{-1}, 2^0, 2^1, 2^2\}$. In case of random forest, three different parameters also tuned. The number of trees (n_{tree}), number of descriptors randomly sampled as candidates for splitting at each node (m_{try}) and minimum node size were selected from n_{tree} = {10, 20,…, 100}, $m_{try} = \{1, \ldots, \sqrt{feature\ dimension}\}$ and node size = {1, 2, 3} respectively for better accuracy.

3.2. Performance Metrics

In this work, four well-defined metrics were used to measure the performance of succinylated site prediction. Sensitivity (Sen), Specificity (Spe), Accuracy (Acc) and Mathews correlation coefficient (MCC) were used to assess the competence of different approaches including the proposed approach on the benchmark dataset. These metrices

were considered in this study to consistently compare the performance of different PTM identification tools. Formulation of these metrices were also adapted from those state-of-the-art techniques specially from SucPred [3], Succfind [4], iSuc-PseOpt [6], SuccinSite [7] and SucStruct [16].

3.2.1. Sensitivity

Sensitivity, also known as true positive rate, is measured by correctly identified succinylated lysine residues. This metric varies between 0 and 1, where 0 indicates the predictor is inaccurate, and 1 represents a totally accurate predictor. The higher sensitivity is proportional to the best prediction of succinylated lysines. Mathematically, it can be expressed as:

$$\text{Sensitivity} = \frac{TP}{TP+FN} \tag{14}$$

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

3.2.2. Specificity

Specificity, also known as true negative rate, is measured by correctly identified negative samples or non-succinylated lysine residues. This metric also varies between 0(totally incorrect) and 1(totally correct). Specificity metric can be formulated as:

$$\text{Sensitivity} = \frac{TN}{TN+FP} \tag{15}$$

where, TN depicts true negatives or the number of correctly rejected samples from succinylated sample set, and FP depicts false positives or the number of incorrectly accepted samples as succinylated.

3.2.3. Accuracy (Acc)

Accuracy (Acc) is the ratio of total number of correctly classified samples (C) and the total number of samples (N). i.e.

$$\text{Acc} = \frac{C}{N} \tag{16}$$

It varies between 0 (least accurate) and 1 (most accurate).For the best succinylation predictor, it will be 1.

3.2.4. Mathews Correlation Coefficient (MCC)

Mathews 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 expressed as:

$$\text{MCC} = \frac{(TP*TN)-(FN*FP)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{17}$$

3.3. Cross-Validation

Cross-validation becomes a de-facto standard and effective method to estimate the performance and effectiveness of a statistical prediction model. In the area of post-translational modification, there are three commonly used cross-validation techniques - independent dataset test, jackknife test and k-fold cross-validation test. Among these techniques, K-fold cross-validation is considered as technique to approximately estimate prediction error without bias under much more complicated situations, mentioned in [15], but with lower computational complexity compared to the other methods. Considering this fact as well as to maintain consistency with existing approaches in this area, in this study, 5-fold cross validation strategy has been considered to evaluate the performance of the proposed approach.

4. Result and Discussion

In this study, the main goal was to investigate as well as develop a predictor to improve the performance of succinylation post-translational modification using conceptually different well-known classifiers. For post-translational modification information source, this study relies on a sequence derived, physicochemical and biochemical based features which are commonly used in this area. For each information source, lysine residue was characterised using each classifier, i.e. KNN, SMV, Random Forest as well as the considered ensemble based DV-iSucLys. This study has found that, in most of the cases, DV-iSucLys over-perform each base classifier or almost align with them which are shown in Figure 2. This result indicates that none of the base algorithms can efficiently utilise all the available protein information independently for succinyl lysine characterisation.

A deeper inspection of the experimental data presented in this article has showed that the proposed method DV-iSucLys has achieved highest overall accuracy 75.4% when PSAAP was used as the information source. This performance has aligned with basic Random Forest algorithm with a slight increase in specificity and MCC. In case of sensitivity, no remarkable improvement has been achieved, still satisfactory. This result might provide some clue that, position information of amino acid residue in sequence data can be utilised in a better way to identify actual non-succinylation site or non-verified site.

However, for identifying the lysine residues which are actually modified by succinyl group and which are not, DV-iSucLys can use the physicochemical or biological information and exceeds the baseline classifiers. It happens not only in case of overall accuracy but also in case of some other performance metrics presented in Table 1, Table 2, Table 3 and in table 4.

In addition, it also shows from the experimental data that, raw frequency based information source of either amino acids or di-peptides was not benicial for ensemble of conceptually different strategy i.e. KNN (neighbor based lazy learning), SVM (kernel based model learning) and Random Forest (decision tree based learning). It reveals from the fact that performace of DV-iSucLys is not improved for frequency based information sources like CKSAAP, Binary Encoding than base classifiers. Though, in case of similar information

source i.e. direct frequecy of amino acids (AAC), a little bit of improvement has been marked in terms of overall accuracy for DV-iSucLys. A closer observation of the evaluation metrices data expose that, different base algorithms handles frequency data far differtly. This raised the convergence issue in decision voting and results in poor performance in ensembling.

Table 1. *Performance matrices of five features using SVM.*

Features	Sensitivity	Specificity	Accuracy	MCC
PSAAP	0.758	0.709	0.741	0.449
CKSAAP	0.927	0.402	0.752	0.402
AAC	0.879	0.435	0.731	0.354
Binary Encoding	0.857	0.490	0.735	0.373
AAindex	0.822	0.530	0.724	0.363

Table 2. *Performance matrices of five features using KNN.*

Features	Sensitivity	Specificity	Accuracy	MCC
PSAAP	0.925	0.349	0.733	0.346
CKSAAP	0.994	0.008	0.665	0.007
AAC	0.941	0.254	0.712	0.279
Binary Encoding	0.981	0.050	0.671	0.088
AAindex	0.911	0.239	0.687	0.204

Table 3. *Performance matrices of five features using RF.*

Features	Sensitivity	Specificity	Accuracy	MCC
PSAAP	0.886	0.490	0.754	0.415
CKSAAP	0.925	0.379	0.743	0.376
AAC	0.868	0.480	0.739	0.380
Binary Encoding	0.911	0.337	0.719	0.309
AAindex	0.957	0.261	0.725	0.322

Table 4. *Performance matrices of five features for DV-iSucLys predictor.*

Features	Sensitivity	Specificity	Accuracy	MCC
PSAAP	0.878	0.505	0.754	0.417
CKSAAP	0.955	0.271	0.727	0.327
AAC	0.916	0.382	0.741	0.363
Binary Encoding	0.923	0.309	0.719	0.303
AAindex	0.928	0.334	0.730	0.338

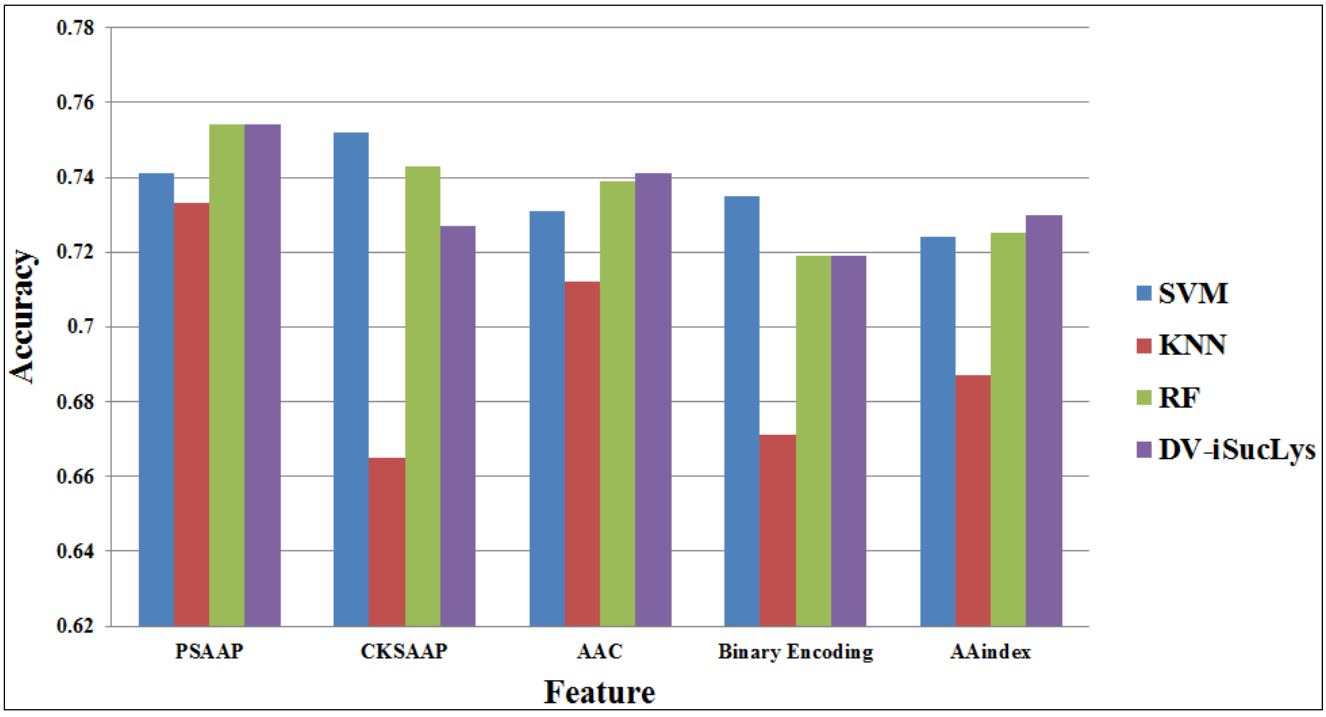

Figure 2. *Comparison of three classifiers for five features based on accuracy.*

Table 5. *A comparison with existing predictors.*

Method	Sensitivity	Specificity	Accuracy	MCC
SucPred	0.272	0.673	0.643	-0.030
SuccFind	0.252	0.792	0.750	0.029
iSuc-PseOpt	0.615	0.778	0.699	0.399
SuccinSite	0.3019	0.9017	0.6092	0.2556
SucStruct (6-fold)	0.7334	0.7548	0.7444	0.4884
Considered method, DV-iSucLys (for PSAAP feature)	0.878	0.505	0.754	0.417

In addition to this, data in Table 5 shows that the sensitivity and accuracy of DV-iSucLys predictor based on PSAAP feature is higher than all previous predictors. From this observation, it can be a hint that combination of decision fusion based approach with direct residue position based information source can be considered as a useful alternative way of succinylation site identification.

5. Conclusion

In this study, a computationally simple lysine residue characterisation approach has been evaluated with the primary motivation to combine conceptually different classifiers using majority voting rule, with the target to balance out their individual weakness. This experiment attempts a thorough exposition of the topic, i.e. characterising succinyl lysine residue from different commonly used information sources like pure protein sequence based, physicochemical as well as biochemical properties of amino acids.

This investigation shows that the proposed approach outperforms other approaches in this area in case of overall accuracy and sensitivity when compared with existing approaches. This result indicates that the chance of correctly identifying true succinylation site is higher than the others. In addition to the proposed approach, an updated benchmark succinylation modification dataset has been developed for different species by extracting PTM information existing dataset as well as from UniProt knowledgebase, which can be used for further research.

References

[1] B. N. Sobolev, A. V. Veselovsky, and V. V. Poroikov, "Prediction of protein post-translational modifications: main trends and methods," Russian Chemical Reviews: Russian Academy of Sciences and Turpion Ltd, vol. 83(2), pp. 143-154, 2014.

[2] Rosen and R. et al., "Probing the active site of homoserine trans-succinylase," FEBS Lett., vol. 577, pp. 386-392, 2004.

[3] X. Zhao, Q. Ning, H. Chai, and Z. Ma, "Accurate in silico identification of protein succinylation sites using an iterative semi-supervised learning technique," Journal of Theoretical Biology, vol. 374, pp. 60-65, 2015.

[4] H. D. Xu, S. P. Shi, P. P. Wen, and J. D. Qiu, "SuccFind: a novel succinylation sites online prediction tool via enhanced characteristic strategy," Bioinformatics, vol. 31(23), pp. 3748-3750, 2015.

[5] Y. Xu et al., "iSuc-PseAAC: predicting lysine succinylation in proteins by incorporating peptide positionspecific propensity," Scientific Reports, vol. 5, 2015.

[6] J. Jia et al., "iSuc-PseOpt: Identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset," Analytical Biochemistry, vol. 497, pp. 48-56, 2016.

[7] A. M. Hasan, S. Yang, Y. Zhoua, and M. N. H. Mollahb, "SuccinSite: a computational tool for the prediction of protein succinylation sites by exploiting the amino acid patterns and properties," Molecular BioSystems, vol. 12(3), pp. 786-795, 2016.

[8] J. Jia et al., "pSuc-Lys: predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach," Journal of Theoretical Biology, vol. 394, pp. 223-230, 2016.

[9] W. Bao, L. Zhu, and D. S. Huang, "ILSES: Identification lysine succinylation-sites with ensemble classification." In IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2016.

[10] L. Nanni, A. Lumini, and S. Brahnam, "An empirical study of different approaches for protein classification," The Scientific World Journal, 2014.

[11] K. Chen, L. Kurgan, and M. Rahbari, "Prediction of protein crystallization using collocation of amino acid pairs," Biochemical and Biophysical Research Communications, vol. 355(3), pp. 764-769, 2007.

[12] S. Kawashima et al., "AAindex: amino acid index database, progress report 2008," Nucleic Acids Research, vol. 36(D202-5), 2008.

[13] Y. R. Tang, Y. Z. Chen, C. A. Canchaya, and Z. Zhang, "GANNPhos: a new phosphorylation site predictor based on a genetic algorithm integrated neural network," Protein Engineering, Design and Selection, vol. 20(8), pp. 405-412, 2007.

[14] M. A. M. Hasan, M. Nasser, S. Ahmad and K. I. Molla, "Feature Selection for Intrusion Detection Using Random Forest," Journal of Information Security, vol. 7, pp. 129-140, 2016.

[15] S. Wang and S. Liu, "Protein Sub-Nuclear Localization Based on Effective Fusion Representations and Dimension Reduction Algorithm LDA." International Journal of Molicular Science, vol. 16(12), pp. 30343-30361, 2015.

[16] Y. López et al., "SucStruct: Prediction of succinylated lysine residues by using structural properties of amino acids," Analytical Biochemistry, vol. 527, pp. 24-32, 2017.

[17] The UniProt Consortium, "UniProt: the universal protein knowledgebase," Nucleic Acids Research; vol. 45, 2016, (D1): D158-D169. doi: 10.1093/nar/gkw1099.

[18] Z. Liu et al. “CPLM: a database of protein lysine modifications.” Nucleic Acids Res. Vol. 42, pp. D531–D536, 2016.

[19] W. R. Qiu, B. Q. Sun, X. Xiao, Z. C. Xu, K. C. Chou, iPTM-mLys: identifying multiple lysine PTM sites and their different types. Bioinformatics, 32(20), pp. 3116-3123, 2016.

[20] Z. Ju, J. J. He, "Prediction of lysine propionylation sites using biased SVM and incorporating four different sequence features into Chou’s PseAAC", Journal of Molecular Graphics and Modelling, vol. 76, pp. 356-363, 2017.

[21] W. R. Qiu, Q. S. Zheng, B. Q. Sun, X. Xiao, “Multi-iPPseEvo: A Multi-label Classifier for Identifying Human Phosphorylated Proteins by Incorporating Evolutionary Information into Chou′ s General PseAAC via Grey System Theory”, Molecular Informatics, 36(3), 2017.

[22] H. Long, M. Wang, H. Fu, “Deep Convolutional Neural Networks for Predicting Hydroxyproline in Proteins” Current Bioinformatics, 12(3), pp. 233-238, 2017.

[23] M. A. M. Hasan, S. Ahmad, M. K. I. Molla, "iMulti-HumPhos: a multi-label classifier for identifying human phosphorylated proteins using multiple kernel learning based support vector machines", Molecular Bio Systems, vol. 13, pp. 1608-1618, 2017.