PSSM-Suc: Accurately predicting succinylation using position specific scoring matrix into bigram for feature extraction

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

PII: S0022-5193(17)30207-2
DOI: 10.1016/j.jtbi.2017.05.005
Reference: YJTBI 9062

To appear in: *Journal of Theoretical Biology*

Received date: 30 January 2017
Revised date: 28 April 2017
Accepted date: 3 May 2017

Please cite this article as: Abdollah Dehzangi , Yosvany López , Sunil Pranit Lal , Ghazaleh Taherzadeh , Jacob Michaelson , Abdul Sattar , Tatsuhiko Tsunoda , Alok Sharma , PSSM-Suc: Accurately predicting succinylation using position specific scoring matrix into bigram for feature extraction, *Journal of Theoretical Biology* (2017), doi: 10.1016/j.jtbi.2017.05.005

Highlights

- New computational approach for predicting succinylation sites.
- Remarkable transformation of the position specific scoring matrix into bigram.
- Evolutionary information collected in bigram probabilities for accurate prediction.

PSSM-Suc: Accurately predicting succinylation using position specific scoring matrix into bigram for feature extraction

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

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

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

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

[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

[†] These authors contributed equally to this work.

[§] These authors jointly supervised this work.

[*] Corresponding authors.

Abstract

Post-Transcriptional Modification (PTM) is a process which contributes to diversify the proteome. Despite many recently reported PTMs with essential roles in the cellular functioning, lysine succinylation has emerged as a subject of particular interest. The experimental identification of succinylated lysines remains a costly and time-consuming process. As a result, computational predictors have been proposed recently for tackling this important issue. However, the performance of predictors is still very limited. In this paper, we propose a new predictor called PSSM-Suc which employs evolutionary information of amino acids for predicting succinylated lysine residues. Here we described each lysine residue through its profile bigrams which are extracted from position specific scoring matrices. We compare the performance of PSSM-Suc with other existing predictors using a widely used benchmark from PTM databases. PSSM-Suc showed a significant improvement in performance over previous predictors. Its sensitivity, accuracy and Matthews correlation coefficient were 0.8159, 0.8199 and 0.6396, respectively.

Keywords: protein sequences, amino acids, succinylation prediction.

Introduction

The chemical modification of proteins after being translated in the ribosome constitutes a relevant biological mechanism in the cell (Walsh et al., 2005; Xu and Chou, 2016). This modification implies the covalent addition of functional groups to the amino acids of proteins. A long list of post-translational modifications (PTM), which extends from methylation (Qiu et al., 2014) and ubiquitination (Qiu et al., 2015) to acetylation (Hou et al., 2014), has been unraveled. These PTMs have active roles in cellular activity making post-translational networks more complex entities. Recently, the scientific community has focused its attention on another PTM, which was first detected by mass spectrometry (Jensen, 2004) and protein sequence alignment, called succinylation (Weinert et al., 2013; Zhang et al., 2011). Succinylation contributes to important changes in the structure and function of proteins (Zhang et al., 2011). Succinylated enzymes have been reported to be involved in mitochondria and fatty acid metabolism (Park et al., 2013) whereas succinylated histones have shown possible roles in chromatin function (Xie et al., 2012).

The detection of succinylation sites has become a challenging task in order to better understand the succinylation mechanism. However, experimental detection has proved to be extremely costly, inefficient and consume a considerable amount of time. Therefore, computational tools able to accurately identify succinylation sites have been thought as an absolute necessity. Some of the predictors proposed so far use the composition of amino acids of proteins. iSuc-PseAAC used a support vector machine as a classifier and the peptide position-specific propensity along the pseudo amino acid composition (Xu et al., 2015b) as features. Likewise, iSuc-PseOpt (Jia et al., 2016b) and pSuc-Lys (Jia et al., 2016a) incorporated sequence-coupling effects into the amino acid composition. They further employed treatments such as the k-nearest neighbors cleaning and inserted hypothetical training samples for class imbalance. The former predictor (iSuc-PseOpt) used a random forest algorithm (Jia et al., 2016b) as a classifier, whereas the later predictor (pSuc-Lys) employed an ensemble of random forest classifiers (Jia et al., 2016a). SucPred designed a learning algorithm with only positive and unlabeled samples (Zhao et al., 2015), whereas SuccinSite provided a random forest classifier with three encoding schemes: k-spaced amino acid pairs, binary scoring and amino acid index properties for prediction (Hasan et al., 2016). SuccFind, on the other hand, used the information of evolution-derived sequences and an improved feature strategy for optimization purposes (Xu et al., 2015a). A previous study identified homologous succinylated proteins in several species and conserved orthologs for some proteins, highlighting the possible importance of evolutionary features (Zhen et al., 2016).

Despite the availability of several predictors, performance in detecting succinylated lysine residues remains a serious issue. Therefore, a better approach capable of using relevant characteristics of amino acids useful for discrimination information are still required. From the aforementioned predictors, only SuccFind incorporated evolutionary information, but poor sensitivity was achieved. In this paper, we propose a new predictor named as PSSM-Suc (Position Specific Scoring Matrix into bigram for Succinylation prediction), which efficiently utilizes evolutionary features for predicting succinylated lysines. In PSSM-Suc, the position-specific scoring matrix (PSSM) was computed for each protein. A segment comprising 15 amino acids upstream and downstream corresponding to each lysine residue was considered for feature extraction.

Thereafter, profile bigram (Sharma et al., 2013) on this segment was computed which is used to describe the features of lysine residue. Because there is no information available from the strict knowledge of primary sequences, PSSM-Suc was aimed at obtaining such information by analyzing each protein sequence. Hereafter, the concept of information related to succinylation sites refers to the analysis of sequences by the proposed predictor.

For experimentation, we used a benchmark dataset consisting of 670 proteins (Liu et al., 2011; Liu et al., 2014). This dataset has very large number of non-succinylated lysine residues compared to succinylated lysine residues. In order to reduce this imbalance, we used the k-nearest neighbors cleaning treatment (Jia et al., 2016b). Furthermore, a pruned decision tree was built for succinylation prediction. When compared with other existing predictors (Hasan et al., 2016; Jia et al., 2016a; Jia et al., 2016b), PSSM-Suc showed a significantly enhanced performance. It was able to successfully predict succinylation residues with 0.8159 sensitivity, 0.8199 accuracy and 0.6396 Matthews correlation coefficient. A result that to the best of our knowledge has not been achieved by any available predictor in the literature.

Materials and Methods

This study describes a novel predictor named PSSM-Suc which considers the PSSM of a protein along with the profile bigram (Sharma et al., 2013) of amino acids around lysines for predicting succinylated and non-succinylated lysine residues. The subsequent sections detail the benchmark that is used in this study and how the profile bigram (Sharma et al., 2013) is computed from PSSM for a segment of amino acids corresponding to lysine residue. Moreover, the design of decision tree for succinylation prediction is discussed.

Benchmark Dataset

The benchmark dataset used in this study was downloaded from two PTM databases (Liu et al., 2011; Liu et al., 2014). It consists of 670 protein sequences with annotated (succinylated and non-succinylated) lysine residues. We used all the protein sequences for computing the sequence identity of the dataset. To do this, we made use of the cd-hit program (Li and Godzik, 2006), which reported a maximum sequence identity of 48%. Moreover, 99% of the succinylation sites were located in proteins <40% similar. We further analyzed each sequence and retrieved its succinylated and non-succinylated lysines. As a result, 1,782 succinylation sites (positive set) and 18,344 non-succinylation sites (negative set) were obtained.

Evolutionary Information via Position Specific Scoring Matrix

PSSM provides information about the substitution probability of a given amino acid, based on its specific position along the protein sequence, with all 20 amino acids of the genetic code. To calculate these probabilities, PSIBLAST aligns a given protein with similar proteins in the protein data bank (Berman et al., 2000). We used all the protein sequences in our benchmark dataset to compute their respective PSSM, which was calculated as the output of PSIBLAST (Altschul et al., 1997). For each protein, PSSM produces two $L \times 20$ matrices where L represents the protein length and the columns indicate the 20 amino acids of the genetic code. One matrix is the log-odds whereas the other is the linear probability of amino acids. The substitution probability for a given amino acid was computed by normalizing the log-odds of its substitution scores across all the 20 amino acids of the genetic code. A linear model was used for converting the log-odds of the substitution scores to substitution probabilities (Bhagwat and Aravind, 2007). To produce PSSM, we run PSI-BLAST (version 2.3.0+) on non-redundant protein data bank (as of April 2016) (Berman et al., 2000) with a cut-off value (E) of 0.001 in three iterations. As a substitution matrix for generating the PSSM we used BLOSUM62. PSSM has been shown as an important feature in a wide range of problems, with promising results when compared to other protein features. Indeed, it has been the best feature to produce relevant evolutionary information (Dehzangi et al., 2014; Faraggi et al., 2012; Heffernan et al., 2015; McGuffin et al., 2000; Taherzadeh et al., 2016a; Taherzadeh et al., 2016b). Out of the two previous matrices we used the linear probabilities of amino acids.

Feature extraction method

The PSSM forms the basis to characterize its succinylation and non-succinylation sites of proteins. Each lysine residue K was described based on its surrounding 15 downstream and 15 upstream amino acids (Fig. 1A). If one lysine residue was located near the N- or C-terminus of a protein and does not contain enough amino acids (either left or right), we considered the mirror effect for dealing with absent amino acids (Jia et al., 2016b) (Fig. 1B). This approach has been studied and used in previous works (Jia et al., 2016a; Jia et al., 2016b). These studies considered different numbers of amino acids around lysines and concluded that the most promising residue window consists of 15 residues upstream and downstream of the lysine. We further performed a similar analysis by regarding distinct windows and drew the same conclusions (Supplementary material 1). Although the use of mirror images for lysines might not be the correct way to tackle this problem, so far it is the best alternative to tackle this problem. For example, let us consider the sequence segment *S* which describes a lysine residue as

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

and consists of 15 amino acids left and 15 amino acids right to the lysine K. Each R_{-i} $(1 \leq i \leq 15)$ indicates an amino acid upstream of the lysine and R_i $(1 \leq i \leq 15)$ represents an amino acid downstream of it. As Eq. (1) shows there are 31 amino acids (including the lysine K) which represent one lysine residue. Thus, the complete sequence segment S describing a lysine belongs to either of two classes ($c = \{0,1\}$). If the represented lysine is succinylated then $c = 1$, but if the lysine is not succinylated then $c = 0$. Each segment *S* was represented by a vector of values from the PSSM. The PSSM was transformed to a frequency vector with bigram (Sharma et al., 2013) (or $PSSM + bigram$) for representing each segment. The transformation $PSSM + bigram$ results in a 20×20 matrix and every segment *S* of 31 amino acids was thus described by 400 $(20 \times 20 \text{ matrix})$ evolutionary features. This feature vector was used to capture the evolutionary information related to each lysine residue and its neighboring amino acids in the segment S.

The method of computing the PSSM using bigram and how $PSSM + bigram$ was used to represent each lysine residue is explained below. The PSSM of each protein sequence is a matrix M of size $L \times 20$, where L is the length of protein and the columns represent the 20 different amino acids of the genetic code. Each element m_{ij} of matrix M represents the transitional probability of j-th amino acid at i-th location within the protein sequence. The sequence segment S (from Eq. (1)) is then described as a 31×20 feature matrix. For each amino acid in the protein sequence PSSM produces the substitution probabilities of the 20 amino acids. Thus, matrix M is processed as a profile bigram (Sharma et al., 2013) by

$$B_{p,q} = \sum_{k=1}^{30} m_{k,p} m_{k+1,q} \quad \text{where } 1 \leq p \leq 20 \text{ and } 1 \leq q \leq 20 \qquad (2)$$

and returns 400 frequencies of $B_{p,q}$ ($p = 1,2,\dots,20$ and $q = 1,2,\dots,20$) for 400 bigram transitions. Let us define a bigram occurrence matrix B which will contain all the frequencies $B_{p,q}$. We used profile bigram because of its promising results in protein analysis problems (Dehzangi et al., 2015a; Dehzangi et al., 2015b; Paliwal et al., 2014; Sharma et al., 2013; Sharma et al., 2015). The bigram of $PSSM$ ($PSSM + bigram$) matrix B was finally converted into a feature vector F of 400 elements as

$$F = [B_{11}, \dots, B_{ij}, \dots, B_{20,20}]^T \text{ for } i = 1,2,\dots,20 \text{ and } j = 1,2,\dots,20 \qquad (3)$$

where transpose is denoted by superscript T.

Each lysine residue (segment S) was defined by a 400-dimensional vector of evolutionary features. This information was computed for each lysine residue in our benchmark dataset and one matrix of 1,782 ($c = 1$) and 18,344 ($c = 0$) samples was created. This matrix was further processed for class imbalance (see section “Reducing the imbalance between classes”) and used for training a pruned decision tree.

One advantage of bigram feature extraction technique is that it is a window-size independent method. In other words, it extracts 400-dimensional feature vector regardless of the window size adopted around a given amino acid. It was shown that using large window size can potentially provide us with more local discriminatory information for a given amino acid with respect to its neighboring residues (Jia et al., 2016a; Jia et al., 2016b). Therefore, using bigram enables us to increase the window size around a given amino acid without increasing the number of features. That is why we can increase the window size to 15 without increasing the number of features instead of using smaller window sizes.

Pruned Decision Tree

Decision trees are non-parametric methods that create logical diagrams in a similar way to rule-based prediction algorithms. They can predict the class of an object by learning specific rules. Despite disadvantages such as the creation of biased models, they require little data and efficiently handle numerical and categorical information. This study uses the C4.5 algorithm which can deal with continuous and incomplete data while minimize overfitting through pruning (Quinlan, 1992). C4.5 algorithm uses gain ratios to build the decision tree. For instance, the gain of the training set Q after splitting on feature F is computed by

$$G(Q,F) = E(Q) - \sum_{i=1}^{n} P(F_i)\, E(Q_{F_i}) \tag{4}$$

where n is the number of all the different values of feature F, $P(F_i)$ is the frequency of samples with F_i value and $E(Q_{F_i})$ represents the subset of samples with F_i value. $E(Q)$ represents the information entropy of Q, which is calculated as follows

$$E(Q) = \sum_{j=1}^{2} -P(C_j) * \log_2 P(C_j) \tag{5}$$

where $P(C_j)$ indicates the frequency of class C_j.

We made use of the WEKA toolbox for creating a pruned decision tree (Hall et al., 2009). Pruning is an important technique aimed at reducing the size of decision trees. It eliminates those parts of the tree which do not significantly contribute to the classification of objects. By using pruning, we were able to reduce the complexity of our decision tree and improve the prediction accuracy by minimizing overfitting. It is worth mentioning that complexity here refers to the behavior of the pruned decision tree when entropy information becomes negative. The confidence cut-off for pruning was set at 0.25 and at least two instances per leaf were considered. Only binary splits were allowed and the MDL correction for information gain on numeric attributes was not enabled.

Statistical measures

One of the most important measures is sensitivity which assesses the proportion of succinylated lysine residues correctly classified by the predictor. A high sensitivity indicates that the predictor can accurately detect those positive instances (succinylated residues) in the dataset. In other words, a sensitivity equals to 1 reflects an accurate predictor whereas a value of 0 points to an inaccurate one. This metric is defined as

$$Sensitivity\ = \frac{SL_{+}}{SL_{+} +\ SL_{-}} \tag{6}$$

where SL_{+} represents the number of correctly predicted succinylated lysines and SL_{-} indicates the number of succinylated lysines incorrectly classified by the predictor.

Specificity, on the other hand, evaluates the proportion of correctly classified non-succinylated lysine residues. Likewise, a numeric value equals to 0 indicates a predictor unable to classify the negative instances (non-succinylated residues) while a specificity of 1 shows an accurate predictor capable of predicting such negative samples in the dataset. Specificity is calculated as

$$Specificity\ = \frac{NSL_{+}}{NSL_{+} + NSL_{-}} \tag{7}$$

where NSL_+ is the number of non-succinylated lysines correctly predicted as such and NSL_- represents the number of incorrectly predicted non-succinylated lysine residues.

The ability of any predictor to differentiate succinylated lysines from non-succinylated ones is also assessed by the accuracy measure. One predictor with zero accuracy is totally inaccurate whereas any accurate predictor would show an accuracy of 1. This measure is defined as

$$Accuracy = \frac{SL_+ + NSL_+}{SL + NSL} \tag{8}$$

where SL and NSL are the total numbers of succinylated and non-succinylated lysine residues, respectively.

The last statistical measure is the Matthews correlation coefficient (MCC). MCC is often used for binary classification and can be used when both classes have different sizes. It results in a correlation coefficient between predicted and observed instances. A predictor with MCC equals to 1 shows perfect correlation between prediction and observation whereas one with -1 MCC does not show any agreement whatsoever. This metric is defined by

$$MCC = \frac{(NSL_+ \times SL_+) - (NSL_- \times SL_-)}{\sqrt{(SL_+ + SL_-)(SL_+ + NSL_-)(NSL_+ + SL_-)(NSL_+ + NSL_-)}} \tag{9}$$

The best predictor should be able to achieve the highest performance in the four statistical measures. However, its performance in at least some of the measures should be higher than that of available predictors. A predictor with low sensitivity is unable to accurately predict succinylation sites and therefore cannot be used for succinylation prediction.

Validation scheme

Any new predictor also needs a validation method in order to assess its effectiveness. In the literature, there are several validation techniques available but the two most common ones are the n-fold cross-validation scheme and the jackknife (Alpaydin, 2014; Chou and Shen, 2008). In these cases, an independent test set is always used for assessment. Jackknife resampling model is less arbitrary than cross-validation and provides unique outcomes for a dataset (Hajisharifi et al., 2014). By following the same validation scheme of similar studies (Jia et al., 2016b; Xu et al., 2015b) we use the n-fold cross-validation technique in this study.

The n-fold cross-validation scheme was conducted as follows,

1. Split data samples into complementary segments (n subsets of approximately equal size).
2. Use $n-1$ subsets for training and the remaining subset for validation.
3. Adjust the predictor parameters with the $n-1$ subsets.
4. Compute the statistical measures on the validation subset.
5. Repeat steps 1 to 4 n times and compute the average of each statistical measure.

In this study, we conducted three (6-, 8- and 10-fold) cross-validation schemes for assessing our predictor PSSM-Suc.

Results and Discussion

Each proposed predictor requires to be rigorously assessed in order to gauge how well it performs. For this purpose, we used four statistical metrics: sensitivity, specificity, accuracy and Matthews correlation coefficient (Chen et al., 2015; Dehzangi et al., 2015a; Ding et al., 2014; Liu et al., 2015a; Liu et al., 2015b; Xiao et al., 2015), which have been widely used in the literature. The subsequent sections describe the treatment of the imbalance between classes and the classification results of the pruned decision tree. The overall performance of PSSM-Suc in predicting succinylation residues is also discussed for the above four metrics.

Reducing the imbalance between classes

We analyzed the protein sequences of our benchmark dataset, and ended up with a number of succinylation sites (positive set) much smaller than that of non-succinylation sites (negative set). This difference brings forward a high imbalance between classes, which could lead to completely biased classification results. Reducing the imbalance between classes is a crucial step in machine learning studies. It allows us to remove those redundant instances before reaching the classification stage. For dealing with such imbalance we employed the k-nearest neighbor technique which is very popular in pattern recognition and was reintroduced for protein attribute prediction by Chou (2011). In order to balance both classes, we removed redundant negative samples by using the *k*-nearest neighbors cleaning treatment (Jia et al., 2016b). We first computed the Euclidean distance between all the samples in the benchmark dataset. An initial cut-off was calculated by dividing the number of negative instances and positive instances. For example, the negative set consisted of 18,344 samples and the positive set comprised 1,782 samples which results in a ratio of 10.29. Thus, $k = 10$ was initially used for reducing class imbalance. In other words, we eliminated a negative sample as long as one of its 10 nearest neighbors (based on the Euclidean distance between the negative sample and others in the entire dataset) is a positive sample. After this first filtering, the imbalance between classes remained so that we further computed new thresholds by multiplying the initial cut-off ($k = 10$) with different integers and repeatedly used the resulting cut-offs until both sets were almost similar in size. The above procedure reduced the initial 18,344 negative samples to 1,643 with a threshold of 80 (negative samples were removed when at least one positive sample was part of its 80 nearest neighbors). It is worth noting that the positive instances, which could affect the predictor sensitivity, are not affected and remained as 1,782. Both sets were then employed to carry out n-fold cross-validation and assess the predictor performance.

PSSM-Suc versus available predictors

Our predictor PSSM-Suc was compared with three recently proposed predictors: iSuc-PseOpt (Jia et al., 2016b), SuccinSite (Hasan et al., 2016) and pSuc-Lys (Jia et al., 2016a). These three methods provided trained web servers which can be easily used for predicting unknown succinylation sites. For comparison purposes, we manually pasted our protein sequences on each of these web servers. We then analyzed the prediction results to evaluate each predictor performance with the same dataset used for training our decision tree. It is worth noting that the web servers were previously trained with some of the same protein sequences utilized for performance assessment. In addition, due to the fact that the web servers were already trained we could just calculate performance on the validation set (only those test sets produced by the cross-validation scheme). Consequently, the area under the curve (AUC) could not be computed for iSuc-PseOpt (Jia et al., 2016b), SuccinSite (Hasan et al., 2016) and pSuc-Lys (Jia et al., 2016a) predictors. On the contrary, we calculated the AUC of PSSM-Suc for each of the three cross-validation trials (6-, 8- and 10-fold).

Table 1 shows the comparison of predictors iSuc-PseOpt (Jia et al., 2016b), SuccinSite (Hasan et al., 2016) and pSuc-Lys (Jia et al., 2016a) with PSSM-Suc. It is clearly noticeable that an improvement in performance for PSSM-Suc over iSuc-PseOpt (Jia et al., 2016b), SuccinSite (Hasan et al., 2016) and pSuc-Lys (Jia et al., 2016a) on sensitivity, accuracy and MCC. These three metrics (sensitivity, accuracy and MCC) significantly improved by 32.7%, 13.64% and 36.0%, respectively. This is a substantial improvement over three benchmark predictors. Though SuccinSite (Hasan et al., 2016) achieved high specificity (0.9057), its ability to correctly predict succinylated lysines dropped to 0.3019 sensitivity. In other words, about 70% of succinylation residues are not detected.

Table 1: Performance of three benchmark predictors and PSSM-Suc. The highest values in each metric are highlighted in bold.

Predictor	**Sensitivity**	**Specificity**	**Accuracy**	**MCC**
iSuc-PseOpt (Jia et al., 2016b)	0.615	0.7888	0.6984	0.4086
SuccinSite (Hasan et al., 2016)	0.3019	**0.9057**	0.5915	0.2581
pSuc-Lys (Jia et al., 2016a)	0.587	0.8673	0.7215	0.4703
PSSM-Suc (6-fold cross-validation)	0.8036	0.8058	0.8047	0.609
PSSM-Suc (8-fold cross-validation)	0.8086	0.804	0.8064	0.6124
PSSM-Suc (10-fold cross-validation)	**0.8159**	0.8241	**0.8199**	**0.6396**

The AUC of PSSM-Suc for 6-, 8- and 10-fold cross-validations was recorded at 0.784, 0.794 and 0.817, respectively.

We further checked whether the mirroring approach also proves true for this study. To do this, we created an independent training set in which the mirroring of lysines was not regarded and trained the decision tree. As a result, the performance of the PSSM-Suc predictor did not improve for any of the assessed metrics (Supplementary material 2). This further indicates that the use of a mirror image might somehow help to keep information necessary for classification.

These promising results clearly illustrate the ability of PSSM-Suc to correctly predict succinylated and non-succinylated lysine residues. This is because PSSM-Suc is effectively using important evolutionary information hidden in protein sequences. Such information is stored in the PSSM of each amino acid segment around lysines, which when converted to one matrix of bigram occurrences seems to be an essential characteristic for detecting modified lysines. Additionally, pruned decision trees and its efficient use of data also contributes to improved outcomes. In summary, the combination of $PSSM + bigram$ somehow appears to reveal important evolutionary information around lysine residues, which could be used to predict which lysines are succinylated and which are not.

Our decision tree script and feature matrix used for training can be accessed at https://github.com/YosvanyLopez/SucEvol. Further information can be provided upon request.

Conclusions

This paper describes a new predictor coined PSSM-Suc, which uses an efficient combination of $PSSM + bigram$ for succinylation prediction. The evolutionary information hidden in PSSMs and transformed to bigram occurrences proves to be an important feature to be taken into consideration. The k-nearest neighbors cleaning treatment can be used to further remove redundant samples and balance the dataset. One fairly balanced dataset along with a pruned decision tree appeared to significantly improve PSSM-Suc performance over available benchmark predictors. In the future, we intend to further explore the use of a 31-residue window for describing lysine residues.

Acknowledgements

The authors thank Jia *et al.* and Hasan *et al.* for making current predictors available, which helped them to establish fair comparisons. YL and AS acknowledge the useful comments and suggestions by the members of Tsunoda Laboratory.

References

Alpaydin, E., 2014. Introduction to Machine Learning. The MIT Press.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., Lipman, D. J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25, 3389-3402.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., Bourne, P. E., 2000. The Protein Data Bank. Nucleic Acids Research 28, 235-242.

Bhagwat, M., Aravind, L., 2007. PSI-BLAST Tutorial. In: NH, B., (Ed.), Comparative Genomics, Vol. 1 and 2. Humana Press, Totowa (NJ), pp. 177-186.

Chen, W., Feng, P., Ding, H., Lin, H., Chou, K.-C., 2015. iRNA-Methyl: Identifying N6-methyladenosine sites using pseudo nucleotide composition. Analytical Biochemistry 490, 26–33.

Chou, K.-C., 2011. Some remarks on protein attribute prediction and pseudo amino acid composition. Journal of Theoretical Biology 273, 236-247.

Chou, K.-C., Shen, H.-B., 2008. Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. Nature Protocols 3, 153-162.

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

Dehzangi, A., Heffernan, R., Sharma, A., Lyons, J., Paliwal, K., Sattar, A., 2015a. Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou's general PseAAC. Journal of Theoretical Biology 364, 284-294.

Dehzangi, A., Sohrabi, S., Heffernan, R., Sharma, A., Lyons, J., Paliwal, K., Sattar, A., 2015b. Gram-positive and gram-negative subcellular localization using rotation forest and physicochemical-based features. BMC Bioinformatics 16, S1.

Ding, H., Deng, E.-Z., Yuan, L.-F., Liu, L., Lin, H., Chen, W., Chou, K.-C., 2014. iCTX-Type: A Sequence-Based Predictor for Identifying the Types of Conotoxins in Targeting Ion Channels. BioMed Research International 2014, 286419.

Faraggi, E., Zhang, T., Yang, Y., Kurgan, L., Zhou, Y., 2012. 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, 259-267.

Hajisharifi, Z., Piryaiee, M., Beigi, M. M., Behbahani, M., Mohabatkar, H., 2014. Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test. Journal of Theoretical Biology 341, 34-40.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H., 2009. The WEKA Data Mining Software: An Update. SIGKDD Explorations 11, 10-18.

Hasan, M. M., Yang, S., Zhou, Y., Mollah, M. N. H., 2016. SuccinSite: a computational tool for the prediction of protein succinylation sites by exploiting the amino acid patterns and properties. Molecular BioSystems 12, 786-795.

Heffernan, R., Paliwal, K., Lyons, J., Dehzangi, A., Sharma, A., Wang, J., Sattar, A., Yang, Y., Zhou, Y., 2015. Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. Scientific Reports 5, 11476.

Hou, T., Zheng, G., Zhang, P., Jia, J., Li, J., Xie, L., Wei, C., Li, Y., 2014. LAceP: Lysine Acetylation Site Prediction Using Logistic Regression Classifiers. PLoS ONE 9, e89575.

Jensen, O. N., 2004. Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry. Current Opinion in Chemical Biology 8, 33-41.

Jia, J., Liu, Z., Xiao, X., Liu, B., Chou, K.-C., 2016a. pSuc-Lys: Predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach. Journal of Theoretical Biology 394, 223-230.

Jia, J., Liu, Z., Xiao, X., Liu, B., Chou, K.-C., 2016b. iSuc-PseOpt: Identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset. Analytical Biochemistry 497, 48-56.

Li, W., Godzik, A., 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22, 1658-1659.

Liu, B., Fang, L., Wang, S., Wang, X., Li, H., Chou, K.-C., 2015a. Identification of microRNA precursor with the degenerate K-tuple or Kmer strategy. Journal of Theoretical Biology 385, 153-159.

Liu, Z., Xiao, X., Qiu, W.-R., Chou, K.-C., 2015b. iDNA-Methyl: Identifying DNA methylation sites via pseudo trinucleotide composition. Analytical Biochemistry 474, 69-77.

Liu, Z., Cao, J., Gao, X., Zhou, Y., Wen, L., Yang, X., Yao, X., Ren, J., Xue, Y., 2011. CPLA 1.0: an integrated database of protein lysine acetylation. Nucleic Acids Research 39, D1029-D1034.

Liu, Z., Wang, Y., Gao, T., Pan, Z., Cheng, H., Yang, Q., Cheng, Z., Guo, A., Ren, J., Xue, Y., 2014. CPLM: a database of protein lysine modifications. Nucleic Acids Research 42, D531-D536.

McGuffin, L. J., Bryson, K., Jones, D. T., 2000. The PSIPRED protein structure prediction server. Bioinformatics 16, 404-405.

Paliwal, K. K., Sharma, A., Lyons, J., Dehzangi, A., 2014. A Tri-Gram Based Feature Extraction Technique Using Linear Probabilities of Position Specific Scoring Matrix for Protein Fold Recognition. IEEE Transactions on NanoBioscience 13, 44-50.

Park, J., Chen, Y., Tishkoff, Daniel X., Peng, C., Tan, M., Dai, L., Xie, Z., Zhang, Y., Zwaans, Bernadette M. M., Skinner, Mary E., Lombard, David B., Zhao, Y., 2013. SIRT5-Mediated Lysine Desuccinylation Impacts Diverse Metabolic Pathways. Molecular Cell 50, 919–930.

Qiu, W.-R., Xiao, X., Lin, W.-Z., Chou, K.-C., 2014. iMethyl-PseAAC: Identification of Protein Methylation Sites via a Pseudo Amino Acid Composition Approach. BioMed Research International 2014, 947416.

Qiu, W.-R., Xiao, X., Lin, W.-Z., Chou, K.-C., 2015. 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, 1731-1742.

Quinlan, J. R., 1992. C4.5: Programs for Machine Learning. Morgan Kaufmann, San Francisco, California, USA.

Sharma, A., Lyons, J., Dehzangi, A., Paliwal, K. K., 2013. A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition. Journal of Theoretical Biology 320, 41-46.

Sharma, R., Dehzangi, A., Lyons, J., Paliwal, K., Tsunoda, T., Sharma, A., 2015. 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, 915-926.

Taherzadeh, G., Zhou, Y., Liew, A. W.-C., Yang, Y., 2016a. Sequence-Based Prediction of Protein-Carbohydrate Binding Sites Using Support Vector Machines. Journal of Chemical Information and Modeling 56, 2115-2122.

Taherzadeh, G., Yang, Y., Zhang, T., Liew, A. W.-C., Zhou, Y., 2016b. Sequence-based prediction of protein–peptide binding sites using support vector machine. Journal of Computational Chemistry 37, 1223-1229.

Walsh, C. T., Garneau-Tsodikova, S., Jr., G. J. G., 2005. Protein Posttranslational Modifications: The Chemistry of Proteome Diversifications. Angewandte Chemie International Edition 44, 7342–7372.

Weinert, Brian T., Schölz, C., Wagner, Sebastian A., Iesmantavicius, V., Su, D., Daniel, Jeremy A., Choudhary, C., 2013. Lysine Succinylation Is a Frequently Occurring Modification in

Prokaryotes and Eukaryotes and Extensively Overlaps with Acetylation. Cell Reports 4, 842-851.

Xiao, X., Min, J.-L., Lin, W.-Z., Liu, Z., Cheng, X., Chou, K.-C., 2015. 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, 2221-2233.

Xie, Z., Dai, J., Dai, L., Tan, M., Cheng, Z., Wu, Y., Boeke, J. D., Zhao, Y., 2012. Lysine Succinylation and Lysine Malonylation in Histones. Molecular & Cellular Proteomics 11, 100-107.

Xu, H.-D., Shi, S.-P., Wen, P.-P., Qiu, J.-D., 2015a. SuccFind: a novel succinylation sites online prediction tool via enhanced characteristic strategy. Bioinformatics 31, 3748-3750.

Xu, Y., Chou, K.-C., 2016. Recent Progress in Predicting Posttranslational Modification Sites in Proteins. Current Topics in Medicinal Chemistry 16, 591-603.

Xu, Y., Ding, Y.-X., Ding, J., Lei, Y.-H., Wu, L.-Y., Deng, N.-Y., 2015b. iSuc-PseAAC: predicting lysine succinylation in proteins by incorporating peptide position-specific propensity. Scientific Reports 5, 10184.

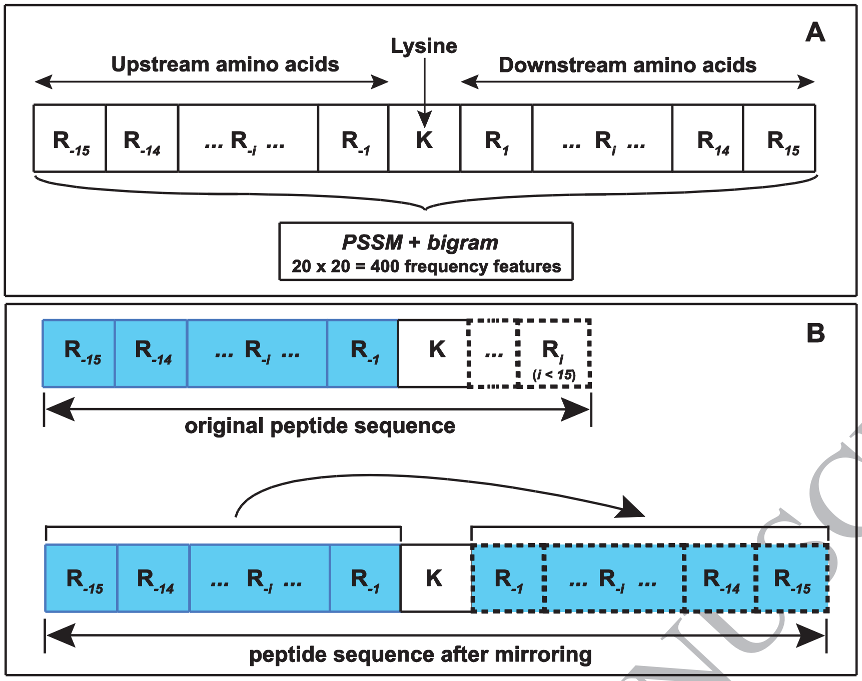
Zhang, Z., Tan, M., Xie, Z., Dai, L., Chen, Y., Zhao, Y., 2011. Identification of lysine succinylation as a new post-translational modification. Nature Chemical Biology 7, 58–63.

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

Zhen, S., Deng, X., Wang, J., Zhu, G., Cao, H., Yuan, L., Yan, Y., 2016. First Comprehensive Proteome Analyses of Lysine Acetylation and Succinylation in Seedling Leaves of Brachypodium distachyon L. Scientific Reports 6, 31576.

FIGURES

Fig. 1. Illustration of how the amino acids surrounding each lysine were taken into consideration. (A) lysine residue with 15 upstream and downstream amino acids. (B) lysine residue with missing downstream amino acids.


A
Upstream amino acids
Lysine
Downstream amino acids
R-15
R-14
... R-i ...
R-1
K
R1
... Ri ...
R14
R15
PSSM + bigram
20 x 20 = 400 frequency features
B
R-15
R-14
... R-i ...
R-1
K
...
Ri
(i < 15)
original peptide sequence
R-15
R-14
... R-i ...
R-1
K
R-1
... R-i ...
R-14
R-15
peptide sequence after mirroring