Journal of Theoretical Biology 374 (2015) 60–65

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Accurate in silico identification of protein succinylation sites using an iterative semi-supervised learning technique

Xiaowei Zhao [a,*], Qiao Ning [a], Haiting Chai [a], Zhiqiang Ma [b,**]

[a] School of Computer Science and Information Technology, Northeast Normal University, Changchun, 130117, China
[b] Key Laboratory of Intelligent Information Processing of Jilin Universities, Northeast Normal University, Changchun 130117, China

HIGHLIGHTS

- Positive samples only learning technique was applied to PTMs sites prediction problems to improve the performance.
- Accurate prediction of succinylation sites by using an iterative semi-supervised learning technique.
- The first succinylation sites online predictor was developed.

ARTICLE INFO

Article history:
Received 29 October 2014
Received in revised form
21 March 2015
Accepted 24 March 2015
Available online 2 April 2015

Keywords:
Succinylated proteins
Positive samples only learning
Multiple features

ABSTRACT

As a widespread type of protein post-translational modifications (PTMs), succinylation plays an important role in regulating protein conformation, function and physicochemical properties. Compared with the labor-intensive and time-consuming experimental approaches, computational predictions of succinylation sites are much desirable due to their convenient and fast speed. Currently, numerous computational models have been developed to identify PTMs sites through various types of two-class machine learning algorithms. These methods require both positive and negative samples for training. However, designation of the negative samples of PTMs was difficult and if it is not properly done can affect the performance of computational models dramatically. So that in this work, we implemented the first application of positive samples only learning (PSoL) algorithm to succinylation sites prediction problem, which was a special class of semi-supervised machine learning that used positive samples and unlabeled samples to train the model. Meanwhile, we proposed a novel succinylation sites computational predictor called SucPred (succinylation site predictor) by using multiple feature encoding schemes. Promising results were obtained by the SucPred predictor with an accuracy of 88.65% using 5-fold cross validation on the training dataset and an accuracy of 84.40% on the independent testing dataset, which demonstrated that the positive samples only learning algorithm presented here was particularly useful for identification of protein succinylation sites. Besides, the positive samples only learning algorithm can be applied to build predictors for other types of PTMs sites with ease. A web server for predicting succinylation sites was developed and was freely accessible at http://59.73.198.144:8088/SucPred/.

1. Introduction

Succinylation is a widespread type of protein post-translational modification, where the succinyl group (–CO–CH_2–CH_2–Co–) is attached to specific lysine residues of substrate proteins (Xie et al., 2012) and plays an important role in regulating protein conformation, function and physicochemical properties. Since identification of succinylated substrate proteins along with succinylation sites can provide valuable insights to understand the molecular mechanism of succinylation in biological systems, much interest has focused on this field, and large-scale proteomic technologies have been applied to identify succinylated proteins and succinylated sites (Weinert et al., 2013; Zhang et al., 2011). However, the experimental determination of exact modified sites of succinylated substrates is labor-intensive and time-consuming, especially for large-scale datasets. In this regard, the computational approaches which could effectively and accurately identify the succinylated sites are urgently needed.

Currently, numerous computational classifiers have been developed to identify PTM sites through various types of two-class machine learning algorithms, such as support vector machines (Li et al., 2011),

* Corresponding author. Tel./fax: +86 0431 84536338.
** Corresponding author.
E-mail addresses: zhaoxw303@nenu.edu.cn (X. Zhao), zhiqiang.ma967@gmail.com (Z. Ma).

http://dx.doi.org/10.1016/j.jtbi.2015.03.029
0022-5193/

Random Forests (Fu et al., 2014), Conditional Random Field (Xu et al., 2013), etc. This form of learning algorithms is usually called as "supervised learning". In the supervised learning, the performance of classifier highly depends on the training dataset which should include samples from both of two classes to be learned (Bhardwaj et al., 2010). However, for the protein succinylation sites prediction problem, the negative samples were difficult to define appropriately and most methods collected negative samples based on "accept or reject" rule, some of which were in fact false and will contribute to false negative prediction. Thus, under a machine learning perspective, the traditional supervised inference of new succinylation sites can not be applied here directly (Cerulo et al., 2013; Zhang and Zuo, 2008).

To deal with the problems mentioned above, a special class of learning called semi-supervised learning is adopted to extract the authentic negative samples. Semi-supervised learning uses both labeled and unlabeled data to perform an otherwise learning task and it has tremendous practical value. In many tasks, there is a paucity of labeled data. The labels may be difficult to obtain because they require human annotators, special devices, or expensive and slow experiments. In this situation, the semi-supervised learning is usually utilized because it can potentially utilize both labeled and unlabeled data to achieve better performance than supervised learning. But one limit of this method is that the learner should absolutely certain there is some non-trivial relationship between labels and the unlabeled distribution (Jiang and Mcquay, 2012; Kundu et al., 2013). In this study we implemented the first application of positive samples only learning (PSoL) algorithm to succinylation sites prediction problem. This algorithm has been widely used to deal with diverse prediction topics in the field of bioinformatics (Liu et al., 2002; Wang et al., 2006; Liu and Liu, 2003), which reduces the succinylation sites prediction problem to the traditional supervised learning by selecting reliable negative samples from the unlabeled dataset. Meanwhile, we developed a novel succinylation sites computational predictor called SucPred, which used multiple feature descriptors to extract the most informative amino acid residue features, including the auto-correlation functions (ACF), the encoding based on grouped weight (EBGW), the normalized van der waals volume (VDWV) and the position weight amino acids composition (WAAC). Promising results were obtained by the SucPred predictor with an accuracy of 88.65% using 5-fold cross validation on the training dataset and an accuracy of 84.40% on the independent testing dataset, which demonstrated that the positive samples only learning algorithm presented here was particularly useful for identification of protein succinylation sites.

Compared with other types of PTMs sites predictors, the proposed predictor has the following features. Firstly, we constructed a new non-redundant dataset and developed the first predictor for succinylation sites identification. Secondly, we implemented the first application of positive samples only learning (PSoL) algorithm to PTMs sites prediction problem to select the reliable negative samples. Thirdly, we adopted multiple feature descriptors to extract the most informative amino acids features and showed how much important the roles of these features played in the prediction.

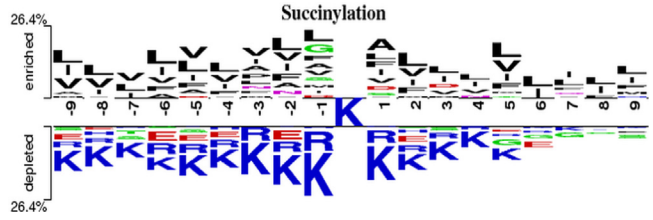

Fig. 1. Two Sample Logos of the compositional biases around succinylation lysine sites compared to unlabeled sites.

2. Materials and methods

2.1. Datasets

The succinylated proteins and succinylated sites used in this study were extracted from a freely accessible database named CPLM (Liu et al., 2014). In this study, for the purpose of developing a solid predictor, we constructed a new non-redundant dataset according to the following steps. (1) All succinylated proteins and succinylated sites with experimental evidences were extracted from CPLM. The initial dataset included 897 succinylated proteins with 2511 known succinylated sites; (2) After a homology-reducing screening procedure by using CD-HIT (Huang et al., 2010; Li and Godzik, 2006) to remove those proteins that had 35% sequence identity to any other, we then got 634 succinylated proteins with 1686 positive sites; (3) For the creation of positive dataset, all peptides centered with succinylated lysine were regarded as positive samples. For the unlabeled dataset, all the remaining peptides centered with non-succinylated lysine were regarded as unlabeled samples. Subsequently, similar to the development of other PTM site predictors (Hu et al., 2011; Zhao et al., 2011), the sliding window strategy was utilized to extract samples. By evaluating several different sizes and according to the two sample logo (see Fig. 1), a window size of 19 was adopted in this paper, with 9 residues located upstream and 9 residues located downstream of the lysine sites in the protein sequence; (4) Finally, 1436 positive peptide segments (succinylated sites) and 18958 unlabeled peptide segments constituted the training dataset. In the meantime, 250 positive peptide segments constituted the independent testing dataset (see Supporting information text S1).

2.2. Feature construction

When the succinylation training dataset was constructed, four types of sequence features were used to transform each sample into feature vectors, including the auto-correlation functions (ACF), the encoding based on grouped weight (EBGW), the normalized van der waals volume (VDWV) and the position weight amino acids composition (WAAC).

2.2.1. Autocorrelation functions (ACF)

In order to represent the correlation and dependence of peptides surrounding the succinylation sites, the auto-correlation function was used to extract the sequence order information (Huang et al., 2012). To calculate the ACF, each residue in the peptide was replaced by its side chain interaction parameter index. Thus, each peptide was transformed into a numerical vector:

$$S = h_1 h_2 h_3 \ldots h_i \ldots h_L \quad (1)$$

where h_i represented the amino acid index of the residue at position i ($i=1, 2, \ldots, L$), and L represented the number of residues in the peptide. Therefore, the ACF r_n was defined as

$$r_n = \frac{1}{L-n}\sum_{i=1}^{L-n} h_i h_{i+n}, n = 1, 2, 3, \ldots, m \quad (2)$$

where m was an integer ($m=1, 2, \ldots, L-1$), and r_m meant the correlation of two amino acid separated by $m-1$ other amino acids. Then, each peptide was represented as an m dimension feature vector:

$$V = [r_1 r_2 r_3 \ldots r_m] \quad (3)$$

Here, $m=8$ was selected to extract the sequence order information.

2.2.2. Encoding based on grouped weight (EBGW)

Since physicochemical properties of amino acid residues have an important influence on protein structure and function, charged

properties and hydrophobicity were adopted in the following to represent the protein peptide. Firstly, 20 amino acid residues were divided into four distinct groups according to their charged character and hydrophobicity properties: the hydrophobic group $C_1=\{A,F,G,I,L,M,P,V,W\}$, the polar group $C_2=\{C,N,Q,S,T,Y\}$, the positively charged group $C_3=\{H,K,R\}$ and the negatively charged group $C_4=\{D,E\}$ (Shi et al., 2012). Then the amino acid residues were partitioned on the basis of the following disjoint groups: C_1+C_2 versus C_3+C_4, or C_1+C_3 versus C_2+C_4, or C_1+C_4 versus C_2+C_3. For a given protein peptide P (the length was L), three binary sequences (H_1,H_2,H_3) can be obtained using the following equations:

$$H_1=\begin{cases}1 & \text{if} \quad P(j)\in C_1+C_2\\ 0 & \text{if} \quad P(j)\in C_3+C_4\end{cases} \tag{4}$$

$$H_2=\begin{cases}1 & \text{if} \quad P(j)\in C_1+C_3\\ 0 & \text{if} \quad P(j)\in C_2+C_4\end{cases} \tag{5}$$

$$H_3=\begin{cases}1 & \text{if} \quad P(j)\in C_1+C_4\\ 0 & \text{if} \quad P(j)\in C_2+C_3\end{cases} \tag{6}$$

We divided each binary sequence obtained above into N sub-sequences increasing in length. The length of the kth sub-sequence was $(L/N)*k$, and the feature value was $sum(k)/(L/N)*k$, where $sum(k)$ represented the number of 1 in the kth sub-sequence. Therefore, each sequence had N dimension feature vector. There were three binary sequences in total, and in total $3*N$ dimension feature vectors. Preliminary tests on the training dataset indicated that $N=5$ was the appropriate number of sub-sequences for identifying succinylation sites.

2.2.3. Normalized van der Waals volume (VDWV)

Van der Waals volume of side groups is important for binding sites (Rudbeck et al., 2012), so we took it into account as a feature to predict succinylation sites. Each amino acid residue had a value of van der Waals volume as shown in Table 1 (Shi et al., 2012).

2.2.4. Position weight amino acids composition (WAAC)

We can come to the conclusion from the Fig. 1 that the sequence-order information between succinylation sequence and non-succinylation sequence was diverse, so we employed position weight amino acid composition to represent this difference (Shi et al., 2013). Given an amino acid residue r_i ($i=1, 2, 3\ldots, 20$), we calculated its value of WAAC by the following formula:

$$C_i=\frac{1}{L(L+1)}\sum_{j=-L}^{L}x_{i,j}(j+\frac{|j|}{L}) \tag{7}$$

where L indicated the number of residues located upstream or downstream from the central lysine site, and j denoted the position of the amino acid residue. $x_{i,j}=1$ if r_i was the jth residue of the protein peptide, on the contrary, $x_{i,j}=0$.

Table 1
The value of van der Waals volume for 20 amino acid residues.

Amino acid	A/C	D/E	F/G	H/I	K/L	M/N	P/Q	R/S	T/V	W/Y
VDWV value	1.00	2.78	5.89	4.66	4.77	4.43	2.72	6.13	2.60	8.08
	2.43	3.78	0.00	4.00	4.00	2.95	3.95	1.60	3.00	6.47

2.3. SVM classifiers

Support vector machine (SVM) is a popular machine learning algorithm based on statistical learning theory. SVM looks for a rule that best maps each member of training dataset into the correct classification (Vapnik, 1998; Tung and Ho, 2008), and it has been widely used in bioinformatics community. For actual implementation, LIBSVM package (version 3.0) (Chang and Lin, 2011) with radial basis kernels (RBF) is used, where the kernel width parameter γ represents how the samples are transformed to a high dimensional space. Grid search strategy based on 5-fold cross-validation is utilized to find the optimal parameters C and $\gamma\in\{2^{-7}, 2^{-6}, \ldots, 2^{8}\}$, so that a total number of 256 grids are evaluated.

2.4. Positive samples only learning

At present, most methods collected negative samples based on "accept or reject" rule and all the remaining peptides centered with non-succinylated lysine were regarded as 'negative samples' (Liu et al., 2011; Shao et al., 2012; Chang et al., 2009; Wang et al., 2009). Then, they took random samples from the unlabeled dataset and assumed they were negative samples. However, some of the 'negative samples' in the training dataset were in fact positive and would affect the prediction results. Therefore, in this study we implemented the first application of positive samples only learning (PSoL) algorithm to succinylation sites prediction problem to determine a reliable negative training dataset.

The PSoL algorithm was described by six steps detailed as follows:

Algorithm 1. PSoL: positive samples only learning

Input: The positive training dataset P contains N_l positive samples. The unlabeled dataset U contains N_u samples with unknown labels.

Operation:

Initial negative dataset selection

Step 1. Select a point x_i from the unlabeled dataset that is most dissimilar to the positive dataset P according to $\max_{x_i\subset U} d(x_i,P)$. The minimum Euclidean distance is used as the distance between a single data point and the positive dataset, given as $d(x_i,P)=\min_{x_j\in P}\|x_i-x_j\|$.

Step 2. The rest of negative dataset N are chosen incrementally from the unlabeled dataset U using the following two formulas (1) the maximum dissimilarity to the positive set $\max_{x_i\subset(U-N)} d(x_i,P)$; (2) the maximum distance to the current negative dataset $\max_{x_i\subset(U-N)}\sum_{x_j\in N} d(x_i,x_j)$. When the specified size of N is reached, Step2 stops; We set the initial negative training dataset: $N_{train}=N$ and subtract these samples from the current unlabeled dataset: $U\leftarrow U-N_{train}$.

Negative dataset expansion

Step 3. Construct a SVM classifier f based on the training dataset $P+N_{train}$ with RBF as the kernel.

Step 4. Apply f to classify the unlabeled samples in U. The predicted valued of a sample in U is denoted as $f(x_i)$.

Step 5. Remove a subset of n examples from U with high confidence: $U\leftarrow U-N_{pred}$; add these samples to the current negative set: $N_{train}\leftarrow N_{train}+N_{pred}$.

Step 6. Repeat Step 3–5 until the size of N_{train} reaches a predefined number, and returns the final classifier f.

This algorithm works in a straightforward way. For the details of this algorithm, we can refer to (Kundu et al., 2013).

2.5. Performance assessment

Three cross validation methods are often used to examine a predictor for its effectiveness: independent dataset test, subsampling test, and jackknife test (Chou and Zhang, 1995). Of these three test methods, the jackknife test is deemed as the most objective one (Chou and Shen, 2008), since the outcome obtained by it is always unique for a given benchmark dataset. However, to reduce the computational time, 5-fold cross validation test is commonly used instead of jackknife test. In the 5-fold cross validation, the dataset is divided into 5 equal subsets, out of which 4 subsets are used for training and the remaining one for testing. This procedure is repeated 5 times and the final prediction result is the average accuracy of the 5 testing subsets. In this study, 5-fold cross validation and independent test methods were chosen for evaluating the proposed predictor.

In order to evaluate the proposed predictor, four measurements are used: sensitivity (*Sn*), specificity (*Sp*), accuracy (*Ac*) and Matthew's correlation coefficient (*MCC*). They are defined by the following formulas:

$$Sn = \frac{TP}{TP+FN}$$

$$Sp = \frac{TN}{TN+FP}$$

$$Ac = \frac{TP+TN}{TP+TN+FP+FN}$$

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

where *TP*, *TN*, *FP* and *FN* stand for the number of true positive, true negative, false positive and false negative, respectively. In this study, due to an uncertainty about the negative samples' classes, only sensitivity was used for performance evaluation. During training, 5-fold cross validation was used to optimize the parameters.

3. Results and discussion

For the initial negative dataset selection, the specified size of *N* was 500 (approximately 35% of the total 1436 positive sites). The training dataset $P+N_{train}$ was then used to build the first binary classifier that was tested on the *U* dataset. To insure the quality of the negative dataset, we selected samples from *U* set which satisfied $f(x_i) \leq -0.2$ ($x_i \in U$). When the number of negative samples was the same as the number of the positive samples, negative dataset expansion stopped. This final model built in the last iteration was then used to classify each sample in the independent test dataset.

3.1. Investigation of different features

As described above, four types of features were used: autocorrelation function (ACF), encoding based on grouped weight (EBGW), normalized van der Waals volume (VDWV) and position weight amino acids composition (WAAC). Here we analyzed the distinction of ACF, EBGW, VDWV and WAAC features between succinylation sites (*P*) and non-succinylation sites (N_{train}).

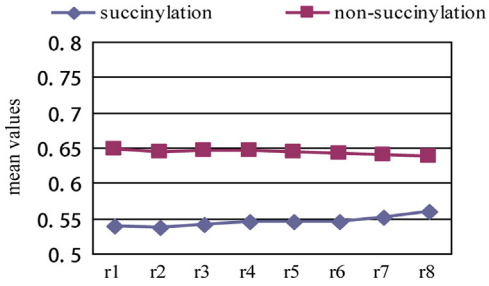

Fig. 2. Mean values of the ACF feature around succinylation and non-succinylation sites.

3.2. ACF feature analysis

ACF feature represents the sequence order information around succinylation sites. We calculated the mean values of ACF feature around succinylation sites and non-succinylation sites, as can be seen in Fig. 2. From Fig. 2, it is obvious that all the mean values of ACF feature around non-succinylation sites were higher than those around succinylation sites, which reflected that succinylation lysine sites had better correlation and more dependence than non-succinylation sites. The above analysis showed that the residue interactions in the adjacent sequences of lysine sites were distinctly different between the succinylation sites and the non-succinylation sites.

3.3. EBGW feature analysis

To illustrate whether succinylation sites and non-succinylation sites had distinct physicochemical properties, we calculated the differences in the distribution of four significant types of amino acid residues. The distribution of four types of amino acid residues formed three binary sequences, and the feature values of these three binary sequences were shown in Fig. 3. In this feature, feature values of H_1 indicate the ratios of hydrophobic and polar amino acids around lysine sites, H_2 indicate the ratios of hydrophobic and positively charged amino acids, and H_3 indicate the ratios of hydrophobic and negatively charged amino acids. It can be seen from Fig. 3 that the feature values of H_1, H_2, H_3 around succinylation sites were all higher than those around non-succinylation sites, especially H_1 and H_2, which showed that more hydrophobic, positively charged and polar amino acids existed around succinylation sites than non-succinylation sites. These observations reflected that the hydrophobic positively, charged and polar residues may have the important influence on succinylation. These analyses indicated that the physicochemical properties of the flanking sequence around succinylation sites and non-succinylation sites were distinct.

3.4. VDWV feature analysis

Vander Waals volume of side groups is important for binding sites. Therefore, we thought about the VDWV feature of the residues around succinylation sites and non-succinylation sites (the window size is 19 residues, $-9 \leq K \leq 9$), as shown in Fig. 4. For $+1$, $+2$, $+3$, $+4$, $+5$, -1, -2, -3, -4, -5 and -6 positions, the mean values of VDWV of residues surrounding succinylation sites were lower than those of residues surrounding non-succinylation sites, especially for

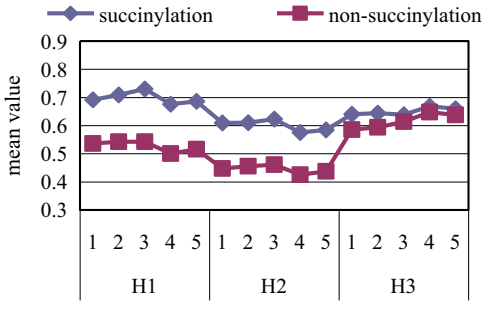

Fig. 3. The mean values of EBGW feature. H_1 includes the values of the first binary sequence, H_2 includes the values of the second binary sequence, and H_3 includes the values of the third binary sequence.

the residues at position −1 and +1. While the values of position at +6, +7, +8, +9, −7, −8 and −9 were almost equal for succinylation and non-succinylation, which indicated that the closer to the lysine, the values of VDWV feature between succinylation and non-succinylation were more distinct. We concluded from the above observations that the binding status of succinylation and non-succinylation was different.

3.5. WAAC feature analysis

Position weight amino acids composition is used to extract the position information of amino acid residues around succinylation sites. To analysis position, we adopted a web-based tool Two Sample Logo to present the compositional biased between succinylation and non-succinylation sites (Fig. 1). For each amino acid, we extracted its position information by its position in the protein sequence fragment. As we can see from Fig. 5, for some amino acids such as P, A and G, the WAAC feature values were higher of succinylation than non-succinylation. However, for some amino acids such as E, H, and K, the WAAC feature values were lower of succinylation than non-succinylation. In conclusion, from Figs. 1 and 5, the position information of amino acid residues surrounding succinylation sites was widely divergent.

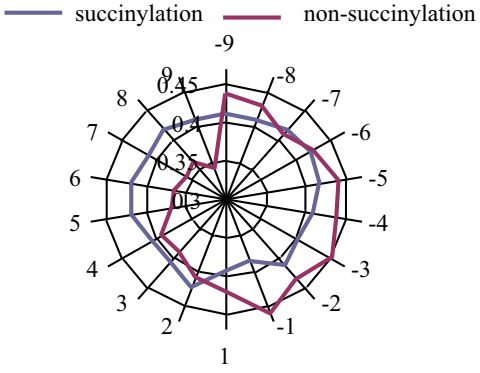

Fig. 4. The mean values of normalized van der Waals volume (VDWV) of residues around succinylation sites and non-succinylation sites.

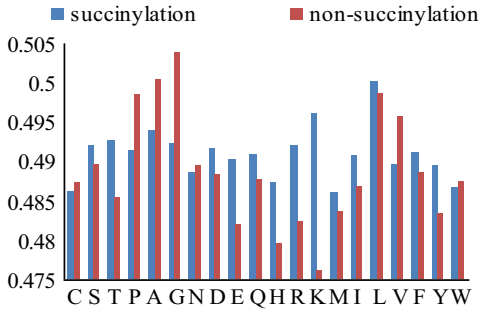

Fig. 5. The mean values of WAAC feature for 20 types of amino acids.

Table 2
The comparison of the performance of models trained with different positive to negative samples radios.

The radio (P/N)	Sensitivity	Specificity	Accuracy	*MCC*
1:1	0.8969	0.8766	0.8865	0.773
1:2	0.7432	0.9475	0.8745	0.723
1:3	0.4913	0.9717	0.8532	0.571

3.6. The performance of the proposed predictor

Based on the four types of features and the window size 19, the performance of models trained with different positive to negative samples radios using 5-fold cross validation was shown in Table 2. With the increase of relative number of the negative samples, the specificity of the models increased, instead the sensitivity kept decreasing. This was because the number of negative samples was bigger than that of positive samples, and more negative samples will cause the model preferentially to predict negative samples correctly to maximize the accuracy (Shi et al., 2013). In order to avoid this phenomenon and reduced the false positive rate, the ratio 1:1 was adopted to construct the optimal predictive model. We can see that from Table 2, the final model got the accuracy of 88.65%, showed that this model learnt to identify succinylation sites with a high accuracy. These results revealed that the combination of multiple features could incorporate more information of succinylation.

In order to further demonstrate the reliability of the model that achieved the best accuracy in cross-validation, an independent testing dataset was used to test the PSoL prediction model. Due to an uncertainty about the negative samples' classes in the independent testing dataset, only sensitivity was used for performance evaluation. The detailed results were shown in Table 3. 211 of these 250 positive samples in the independent testing dataset were identified as positive and only 39 were identified as negative giving an accuracy of 84.40% with a sensitivity of 84.40%. We also compared PSoL with another widely adopted negative samples selection method: the negative samples were collected based on the "accept or reject" rule. As can be seen in Table 3, the PSoL method outperformed "Accept or reject" method both on the training dataset and independent testing dataset. The promising performance demonstrated that the positive samples only learning algorithm presented here was particularly useful for identification of protein succinylation sites.

4. Conclusion

In this work, we implement the first application of positive samples only learning (PSoL) algorithm to succinylation sites prediction problem. Experimental results have shown that our method is very promising and can be a useful tool to identification of succinylation sites. This work also indicated that PSoL was a useful way for the generation of negative training samples with high confidence. In general, the above proposed method can be easily applied to build predictors for other types of PTMs sites. Although the results obtained here were very promising, further

Table 3
The comparison of PSoL and "Accept or reject" methods based on two datasets.

Methods	Datasets	Sensitivity	Specificity	Accuracy	*MCC*
PSoL	Training dataset	0.8969	0.8766	0.8865	0.773
	Independent testing dataset	0.8440	–	0.8440	–
Accept or reject	Training dataset	0.4849	0.6458	0.5639	0.1324
	Independent testing dataset	0.524	–	0.524	–

investigation was needed to further clarify the mechanism of a succinylation process.

Acknowledgments

This research is partially supported by National Natural Science Foundation of China (61403077), the Research Fund for the Doctoral Program of Higher Education of China (20130043110016), the Fundamental Research Funds for the Central Universities (14QNJJ029), the Science and Technology Development Project of Jilin Province (201201069, 20150520061JH), and the Postdoctoral Science Foundation of China (2014M550166).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.jtbi.2015.03.029.

References

Bhardwaj, N., Gerstein, M., Lu, H., 2010. Genome-wide sequence-based prediction of peripheral proteins using a novel semi-supervised learning technique. BMC Bioinform. 11, S6–S15.
Cerulo, L., Pduano, V., Zoppoli, P., Ceccarelli, M., 2013. A negative selection heuristic to predict new transcriptional targets. BMC Bioinform. 14, S3–S10.
Chang, C.C., Lin, C.J., 2011. LIBSVM: a library for support vector machine. ACM Trans. Intell. Syst. Technol. 2, 1–27.
Chang, W.C., Lee, Z.Y., Shien, D.M., 2009. Incorporating support vector machine for identifying protein tyrosine sulfation sites. J. Comput. Chem. 30, 2526–2537.
Chou, K.C., Shen, H.B., 2008. Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. Nat. Protoc. 3, 153–162.
Chou, K.C., Zhang, C.T., 1995. Review: Prediction of protein structural classes. Crit. Rev. Biochem. Mol. Biol. 30, 275–349.
Fu, L., Xie, H.L., Xu, X.R., Yang, H.J., Nie, X.D., 2014. Combining random forest with multi-amino acid features to identify protein palmitoylation sites. Chemom. Intell. Lab. 135, 208–212.
Hu, L.L., Li, Z., Wang, K., Niu, S., Shi, X.H., Cai, Y.D., Li, H.P., 2011. Prediction and analysis of protein methylarginine and methyllysine based on multisequence features. Biopolymers 96, 763–771.
Huang, S.Y., Shi, S.P., Qiu, J.D., 2012. PredSulSite: prediction of protein tyrosine sulfation sites with multiple features and analysis. Anal. Biochem. 428, 16–23.
Huang, Y., Niu, B., Gao, Y., Fu, L., Li, W., 2010. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics 26, 680–682.
Jiang, Q., Mcquay, L.J., 2012. Predicting protein function by multi-label correlated semi-supervised learning. IEEE/ACM Trans. Comput. Biol. Bioinform. 9, 1059–1069.
Kundu, K., Costa, F., Huber, M., Reth, M., Backofen, R., 2013. Semi-supervised prediction of SH2-Peptide interactions from imbalanced high-throughput data. PLoS ONE 8, e62732.
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.
Li, Z.C., Zhou, X., Dai, Z., Zou, X.Y., 2011. Identification of protein methylation sites by coupling improved ant colony optimization algorithm and support vector machine. Anal. Chim. Acta 703, 163–171.
B. Liu, W.S. Lee, P.S. Yu, X. Li, Partially supervised classification of text documents, San Francisco, USA, 2002.
X. Liu, B. Liu, Learning to classify texts using positive and unlabeled data, San Francsisco, USA, 2003.
Liu, Z.X., Gao, J., Ma, Q., Gao, X.J., Ren, J., Xue, Y., 2011. GPS-YNO2: computational prediction of tyrosine nitration sites in proteins. Mol. Biosyst. 7, 1197–1204.
Liu, Z.X., Wang, Y.B., Gao, T.S., Pan, Z.C., Cheng, H., Yang, Q., Cheng, Z.Y., Guo, A.Y., Xue, Y., 2014. CPLM: a database of protein lysine modifications. Nucleic Acids Res. 42, D531–D536.
Rudbeck, M.E., Nilsson, L.S., Barth, A., 2012. Influence of the molecular environment on phosphorylated amino acid models: a density functional theory study. J. Phys. Chem. B 116, 2751–2757.
Shao, J.L., Xu, D., Hu, L.D., Kwan, Y.W., Wang, Y.F., Kong, X.Y., Ngai, S.M., 2012. Systematic analysis of human lysine acetylation proteins and accurate prediction of human lysine acetylation through bi-relative adapted binomial score Bayes feature representation. Mol. Biosyst. 8, 2964–2973.
Shi, S.P., Qiu, J.D., Sun, X.Y., 2012. PLMLA: prediction of lysine methylation and lysine acetylation by combining multiple features. Mol. Biosyst. 8, 1520–1527.
Shi, S.P., Qiu, J.D., Sun, X.Y., 2012. PMeS: prediction of methylation sites based on enhanced feature encoding scheme. PLoS one 7, e38772.
Shi, S.P., Sun, X.Y., Qiu, J.D., 2013. The prediction of palmitoylation site locations using a multiple feature extraction method. J. Mol. Graph. Model. 40, 125–130.
Shi, S.P., Sun, X.Y., Qiu, J.D., Suo, S.B., Chen, X., Huang, S.Y., Liang, R.P., 2013. The prediction of palmitoylation site locations using a multiple feature extraction method. J. Mol. Graph. Model. 40, 125–130.
Tung, C.W., Ho, S.Y., 2008. Computational identification of ubiquitylation sites from protein sequences. BMC Bioinform. 9, 310–320.
Vapnik, V., 1998. Statistical Learning Theory. Wiley, New York.
Wang, C., Ding, C., Meraz, R.F., Holbrook, S.R., 2006. PSoL: a positive sample only learning algorithm for finding non-coding RNA genes. Bioinformatics 22, 2590–2596.
Wang, X.B., Wu, L.Y., Wang, Y.C., Deng, N.Y., 2009. Prediction of palmitoylation sites using the composition of k-spaced amino acid pairs. Protein Eng. Des. Sel. 22, 707–712.
Weinert, B.T., Scholz, C., Wagner, S.A., Iesmantavicius, V., Su, D., Daniel, J.A., Choudhary, C., 2013. Lysine succinylation is a frequently occurring modification in prokaryotes and eukaryotes and extensively overlaps with acetylation. Cell Rep. 4, 842–851.
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. Mol. Cell. Proteomics 11, 100–107.
Xu, Y., Ding, J., Wu, L.Y., Chou, K.C., 2013. iSNO-PseAAC: predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. PLoS ONE 8, e55844.
B. Zhang, W. Zuo, 2008. Learning from positive and unlabeled samples: a survey. In: Proceedings of the International Symposiums on Information Processing (ISIP), 10, pp. 650–654.
Zhang, Z., Tan, M., Xie, Z., Dai, L., Chen, Y., Zhao, Y., 2011. Identification of lysine succinylation as a new post-translational modification. Nat. Chem. Biol. 7, 58–63.
Zhao, X.W., Li, X.T., Ma, Z.Q., Yin, M.H., 2011. Prediction of lysine ubiquitylation with ensemble classifier and feature selection. Int. J. Mol. Sci. 12, 8347–8361.