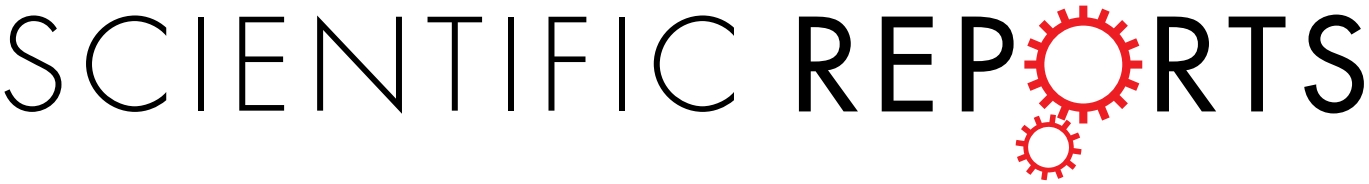

OPEN

iSuc-PseAAC: predicting lysine succinylation in proteins by incorporating peptide position-specific propensity

Received: 14 January 2015
Accepted: 01 April 2015
Published: 18 June 2015

Yan Xu[1], Ya-Xin Ding[1], Jun Ding[1], Ya-Hui Lei[1], Ling-Yun Wu[2] & Nai-Yang Deng[3]

Lysine succinylation in protein is one type of post-translational modifications (PTMs). Succinylation is associated with some diseases and succinylated sites data just has been found in recent years in experiments. It is highly desired to develop computational methods to identify the candidate proteins and their sites. In view of this, a new predictor called iSuc-PseAAC was proposed by incorporating the peptide position-specific propensity into the general form of pseudo amino acid composition. The accuracy is 79.94%, sensitivity 51.07%, specificity 89.42% and MCC 0.431 in leave-one-out cross validation with support vector machine algorithm. It demonstrated by rigorous leave-one-out on stringent benchmark dataset that the new predictor is quite promising and may become a useful high throughput tool in this area. Meanwhile a user-friendly web-server for iSuc-PseAAC is accessible at http://app.aporc.org/iSuc-PseAAC/ . Users can easily obtain their desired results without the need to understand the complicated mathematical equations presented in this paper just for its integrity.

Protein post-translational modification (PTM) is one of the most efficient biological mechanisms for expanding the genetic code and for regulating cellular physiology[1]. Lysine succinylation is one type of PTMs. The succinyllysine residue was initially identified by mass spectrometry and protein sequence alignment. The research further showed that lysine succinylation is evolutionarily conserved and responds to different physiological conditions[2]. Park *et al.*[3] identified 2565 succinylation sites on 779 proteins in 2013. They revealed potential impacts of lysine succinylation on enzymes involved in mitochondrial metabolism such as, amino acid degradation, the tricarboxylic acid cycle (TCA) and falty acid metabolism. SIRT5 has been found as the known enzyme to catalyze lysine desuccinylation[3,4]. Lysine succinylation is also present on histones, suggesting possible roles in regulating chromatin structures and functions[5]. Therefore, identifying the succinylated sites in proteins may provide useful information for biomedical research.

Identification of succinylation residues with experiments was mainly by means of mass spectrometry, which was expensive and laborious. Facing the avalanche of protein sequences generated in the post genomic age, it is a supplementary way to develop computational methods for timely and effectively identifying the succinylation residues in proteins.

There are not computational methods to identify lysine succinylation sites. The present study was devoted to develop a new predictor for identifying lysine succinylation in proteins incorporating the peptide position-specific propensity into the general form of pseudo amino acid composition. According to a comprehensive review[6], to develop a really useful predictor for a protein system, we usually need to

[1]Department of Information and Computer Science, University of Science and Technology Beijing, Beijing 100083, China. [2]Institute of Applied Mathematics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China. [3]College of Science, China Agricultural University, Beijing 100083, China. Correspondence and requests for materials should be addressed to Y.X. (email: xuyan@ustb.edu.cn)

No.	Positive	Negative
Homologous	2521	24128
Non-redundancy	1167	3553

Table 1. The number of positive and negative peptides in the benchmark dataset $\mathbb{S}$.

consider the following procedures: (a) select or construct a valid benchmark dataset to train and test the predictor; (b) represent the protein or peptide samples with an effective formulation that can truly reflect their intrinsic correlation with the target to be predicted; (c) introduce or develop a powerful algorithm or operation engine to conduct the prediction; (d) properly perform cross-validation tests to objectively evaluate the anticipated prediction accuracy; (e) establish a user-friendly web-server for the predictor that is accessible to the public.

Methods

Benchmark Dataset. In this study the benchmark dataset was derived from the CPLM[7] which was a protein lysine modification database. There are 2521 lysine succinylation sites and 24128 non-succinylation sites in 896 unique proteins. The corresponding protein sequences were derived from Uniprot database[8]. For facilitating description later, let us adopt the Chou's peptide formulation which was used for signal peptide cleavage sites[9], and S-Nitrosylation site prediction[10]. According to Chou's scheme, a peptide with lysine (K) located at its center can be expressed as

$$\mathbf{P} = \mathrm{R}_{-\xi}\mathrm{R}_{-(\xi-1)}\cdots\mathrm{R}_{-2}\mathrm{R}_{-1}\mathbb{K}\mathrm{R}_{+1}\mathrm{R}_{+2}\cdots\mathrm{R}_{+(\xi-1)}\mathrm{R}_{+\xi} \tag{1}$$

where the subscript ξ is an integer, $\mathrm{R}_{-\xi}$ represents the ξ-th downstream amino acid residue from the center, R_{ξ} the ξ-th upstream amino acid residue, and so forth. A peptide $\mathbf{P}$ is classified into the following categories:

$$\mathbf{P} \in \begin{cases} \text{succinylated peptide}, & \text{if its center is a succinylation site} \\ \text{non}-\text{succinylated peptide}, & \text{otherwise} \end{cases} \tag{2}$$

Thus, the benchmark dataset can be formulated as

$$\mathbb{S} = \mathbb{S}^{+} \bigcup \mathbb{S}^{-} \tag{3}$$

where $\mathbb{S}^{+}$ contains the samples for the succinylated peptides only, $\mathbb{S}^{-}$ contains the non-succinylated peptides only (cf. Eq.2).

The parameter ξ in peptides was $\xi = 7$ after some preliminary trials and the sample extracted from proteins in this study was a $2\xi + 1 = 15$ tuple peptide. If the upstream or downstream in a peptide sample was less than ξ, the lacking residues were filled with the dummy code X. The experimental results would be overestimated if the benchmark dataset contained homology peptides. Those peptides that had $\geq 40\%$ pairwise sequence identity to any other were rigorously excluded from the benchmark datasets.

Finally, we obtained the benchmark dataset $\mathbb{S}$ containing $1167 + 3553 = 4720$ peptide samples in Table 1, of which 1167 were succinylated peptides belonging to the positive subset $\mathbb{S}^{+}$, and 3553 were non-succinylated peptides belonging to the negative subset $\mathbb{S}^{-}$. The peptide fragments as well as their succinylation or non-succinylation sites in proteins are given in the Supplementary Materials S1 and S2 for $\mathbb{S}^{+}$ and $\mathbb{S}^{-}$, respectively.

Feature Vector Construction. The peptides need to convert into effective mathematical expression (feature construction) which could reflect intrinsic correlation with the desired target in predicting the PTMs. The protein sequences are the most and important information to construct features. According to the review[6], the general form for a protein or peptide $\mathbf{P}$ can be formulated by

$$\mathbf{P} = [\psi_1 \ \ \psi_2 \ \ \cdots \ \ \psi_u \ \ \cdots \ \ \psi_\Omega]^{\mathbf{T}} \tag{4}$$

where $\mathbf{T}$ is the transpose operator and Ω is an integer to reflect the vector's dimension. The value of Ω as well as the components $\psi_u (u = 1, 2, \cdots, \Omega)$ in Eq.4 will depend on how to extract the desired information from the protein or peptide sequences. Below, let us describe how to extract the useful information from the benchmark dataset $\mathbb{S}$ to define the peptide samples via Eq.4.

A peptide $\mathbf{P}$ in Eq.1 can be simplified to a more convenient form given by

$$\mathbf{P} = \mathrm{R}_1\mathrm{R}_2\cdots\mathrm{R}_8\cdots\mathrm{R}_{14}\mathrm{R}_{15} \tag{5}$$

where $R_8 = K$, and $R_i (i = 1, 2, \cdots, 15; i \neq 8)$ can be any of the 20 native amino acids or the dummy code X. We use the numerical codes 1, 2, 3, ⋯, 20 to represent the 20 native amino acids according to the alphabetic order of their single letter code, and use 21 to represent the dummy amino acid X. A "Position Specific Amino Acid Propensity" (PSAAP) matrix $\mathbb{Z}$[10,11] was introduced according to the benchmark dataset $\mathbb{S}$.

$$\mathbb{Z} = \begin{bmatrix} z_{1,1} & z_{1,2} & \cdots & z_{1,14} \\ z_{2,1} & z_{2,2} & \cdots & z_{2,14} \\ \vdots & \vdots & \ddots & \vdots \\ z_{20,1} & z_{20,2} & & z_{20,14} \\ z_{21,1} & z_{21,2} & \cdots & z_{21,14} \end{bmatrix} \tag{6}$$

where the element

$$z_{i,j} = F^+(R_i|j) - F^-(R_i|j) \, (i = 1, 2, \cdots, 21; j = 1, 2, \cdots, 14) \tag{7}$$

$F^+(R_i|j)$ is the occurrence frequency of the i-th amino acid (i=1, 2, ⋯, 21) in the$j$-th column in the positive benchmark dataset $\mathbb{S}^+$ while $F^-(R_i|j)$ is the corresponding occurrence frequency but derived from the negative benchmark dataset $\mathbb{S}^-$. We deleted the center amino acid K as it was the same in positive and negative peptides (samples), respectively. Thus, the components in Eq.4 can be uniquely defined by

$$\psi_u = \begin{cases} z_{1,u} & \text{when} R_i = A \\ z_{2,u} & \text{when} R_i = C \\ \vdots & \vdots \quad [u = 1, 2, \cdots, \Omega(=14)] \\ z_{20,u} & \text{when} R_i = Y \\ z_{21,u} & \text{when} R_i = X \end{cases} \tag{8}$$

Prediction Algorithm. Support vector machine (SVM) is one of the most widely used machine learning algorithms in bioinformatics. The decision rule g(x) was obtained by solving a convex quadratic programming with kernel function. In this work, the kernel function was RBF (Radial Basis Function) kernel with parameter g=0.005. In order to obtain the probability output from SVM, i.e. the probability of that unlabeled input x belongs to a certain class, P(y=1|x), a logistic model was built to map the output g(x) of the SVM into estimated probabilities[12].

$$\Pr(y = 1|x) = P_{A,B}(g(x)) = \frac{1}{1 + \exp(A_* g(x) + B)} \tag{9}$$

Parameter A and B can be obtained by solving the following model

$$\begin{aligned} &\min_{A,B} - \sum_{i=1}^{N_+ + N_-} (t_i \log(p_i) + (1 - t_i) \log(1 - p_i)) \\ &s.\ t.\ t_i = \begin{cases} \dfrac{N_+ + 1}{N_+ + 2}, & y_i = +1 \\ \dfrac{1}{N_- + 2}, & y_i = -1 \end{cases} \\ &p_i = P_{A,B}(g(x_i)), \, i = 1, 2, \cdots, (N_+ + N_-) \end{aligned} \tag{10}$$

where N_+ and N_- represent the number of $\mathbb{S}^+$ and $\mathbb{S}^-$ during training process, respectively.

For a query peptide $\mathbf{P}$ as formulated by Eq.4, suppose $\Pr(y = 1|\mathbf{P})$ is its probability to the succinylated peptides. Thus, the prediction rule for the query peptide $\mathbf{P}$ can be formulated as

$$\mathbf{P} \in \begin{cases} \text{succinylated peptide}, & \text{if } \Pr(y = 1|\mathbf{P}) > \theta \\ \text{non-succinylated peptide}, & \text{otherwise} \end{cases} \tag{11}$$

The cutoff value θ is 0.35 for balancing the true positive and negative rate, unless an additional introduction is attached. The SVM algorithm is implemented by LIBSVM, a public and widely used SVM library.

The predictor established via the above procedures is called **iSuc-PseAAC**, where "i" stands for the 1st character of "identify", "Suc" for "succinylation", and "PseAAC" for that the general form of pseudo

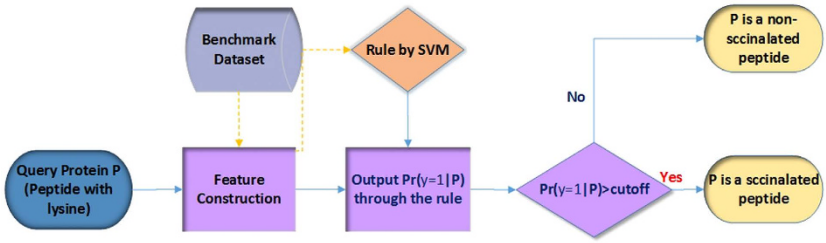

Figure 1. A flowchart of the iSuc-PseAAC predictor.

Cross-validation	Sen (%)	Spe (%)	Acc (%)	AUC	MCC
10-fold	50.65 ± 0.63	89.67 ± 0.27	80.02 ± 0.27	0.782 ± 0.003	0.432 ± 0.007
8-fold	50.25 ± 0.90	89.65 ± 0.34	79.91 ± 0.27	0.782 ± 0.002	0.428 ± 0.007
6-fold	49.95 ± 0.62	89.70 ± 0.35	79.87 ± 0.35	0.781 ± 0.002	0.426 ± 0.009
LOO	51.07	89.42	79.94	0.782	0.431

Table 2. The 10-fold, 8-fold and 6-fold cross-validation results by the predictor on the benchmark dataset $\mathbb{S}$. The experiments have been executed 30 times for every cross-validation and the results were the mean ± standard variation.

amino acid composition was used to formulate the peptide sequences. A flowchart of the predictor was given in Fig. 1 to illustrate how **iSuc-PseAAC** worked during the process of prediction.

Four metrics for measuring prediction quality. To measure the performance of the predictor **iSuc-PseAAC**, four usual metrics were adopted as in[10,13–16] and they are defined as

$$\begin{cases} \text{Sen} = \dfrac{\text{TP}}{\text{TP} + \text{FN}} \\ \text{Spe} = \dfrac{\text{TN}}{\text{TN} + \text{FP}} \\ \text{Acc} = \dfrac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \\ \text{MCC} = \dfrac{(\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \end{cases} \quad (12)$$

where TP (true positive) denotes the number of succinylated peptides correctly predicted, TN (true negative) the numbers non-succinylated peptides correctly predicted, FP (false positive) the non-succinylated incorrectly predicted as the succinylated peptides, and FN (false negative) the succinylated peptides incorrectly predicted as the non-succinylated peptides. Sen, Spe, Acc, and MCC are the sensitivity, specificity, accuracy and the Mathew's correlation coefficient[17], respectively. The ROC curve (receiver operating characteristic curve) which shows the trade-off between sensitivity and specificity is also been examined. AUC (area under the curve) is also another indicator in practical application. It is instructive to point out that the metrics as defined in Eqs.12 are valid for single-label systems; for multi-label systems a set of more complicated metrics should be used as given in[18].

Results and Discussion

Leave-one-out Cross Validation. The cross validation methods are often used to examine the quality of a predictor and its effectiveness in PTMs. The independent dataset test, subsampling or K-fold (such as 6-fold, 8-fold, or 10-fold) cross validation test and leave-one-out (LOO) test are the most cross validations. The K-fold cross validation was used for its less computational time and often been performed many times for different subsampling combinations followed by averaging their outcomes as done by investigators for PTM site predictions[19–22]. The LOO test is the least arbitrary that can always yield a unique result for a given benchmark dataset. Therefore, it has been widely recognized and increasingly utilized to examine the quality of various predictors (see, e.g.,[18,23–25]). Accordingly, in this study the LOO and K-fold cross validation were adopted to evaluate the accuracy of the current predictor. The 10-fold, 8-fold and 6-fold cross validations have been executed for 30 times to avoid the bias. Their results obtained by iSuc-PseAAC on the benchmark dataset were listed in Table 2.

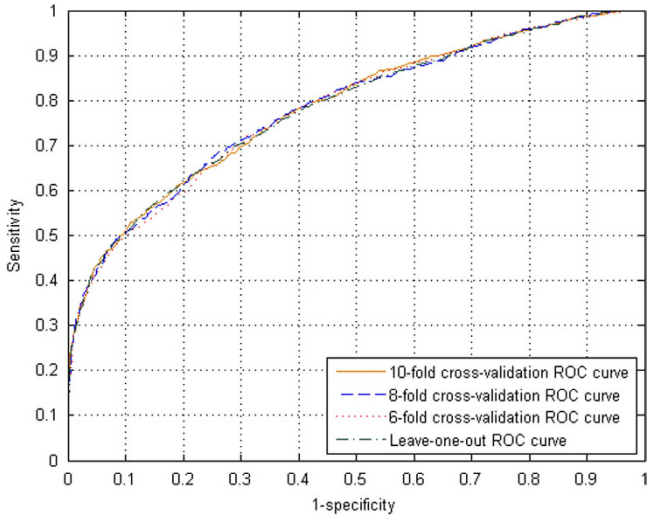

Figure 2. The ROC curves for the LOO test and 6-, 8-, 10-fold cross-validations.

Example ;B1XBY6;Length 234

MAKLTKRMRVIREKVDATKQYDINEAIALLKELATAKFVESVDVAVNLGIDARKSDQNVRGATVLPHGTGRSVRV
AVFTQGANAEAAKAAGAELVGMEDLADQIKKGEMNFDVVIASPDAMRVVGQLGQVLGPRGLMPNPKVGTVTPNVA
EAVKNAKAGQVRYRNDKNGIIHTTIGKVDFDADKLKENLEALLVALKKAKPTQAKGVYIKKVSISTTMGAGVAVD
QAGLSASVN

Position	Peptide	Posterior probability score	Cutoff
31	NEAIALLKELATAKF	0.3884	0.35
105	EDLADQIKKGEMNFD	0.3834	0.35
154	PNVAEAVKNAKAGQV	0.4770	0.35
197	EALLVALKKAKPTQA	0.3502	0.35

Figure 3. The predicted results of the predictor iSuc-PseAAC.

As we can see from Table 2, the overall accuracies for the lysine succinylation was (80.02 ± 0.27)% and its sensitivity (50.65 ± 0.63)%, specificity (89.67 ± 0.27)%, MCC (0.432 ± 0.007) and the AUC (0.782 ± 0.003) in 10-fold cross validation. The AUC were (0.782 ± 0.002) and (0.781 ± 0.002) in 8-fold and 6-fold cross validation, respectively. In LOO test the accuracy was 79.94%, sensitivity 51.07%, specificity 89.42% and AUC 0.782. The ROC curves in Fig.2 were intensive which illustrated the robust of the predictor iSuc-PseAAC. All these results in cross validations and LOO test were approximate. (in Table 2 and Fig.2).

As pointed out in[26], and emphasized in a series of recent publication (see, e.g.,[27,28]), another key in developing a practically useful prediction method is to establish a user-friendly and publicly accessible web-server. In view of this, the web server for **iSuc-PseAAC** has been established that can be freely accessible at http://app.aporc.org/iSuc-PseAAC/. Users can easily get the desired result by using **iSuc-PseAAC** without the need to follow the complicated mathematical equations presented in this paper. Either type or copy/paste the query protein sequences into the input box or upload your input files. The protein sequences should be in FASTA format. Click on the Submit button to see the predicted results in Fig.3. For example, protein B1XBY6 has lysine succinylation 105, 154, 186 and 197 sites, and the predictor iSuc-PseAAC has successfully predicted 31, 105, 154 and 197 sites. Protein E9Q5L3 has three succinylation sites (70, 278 and 284) and iSuc-PseAAC has successfully predicted 278 and 284 sites. Click on the Data button to download the benchmark dataset.

References

1. Walsh, C. T., Garneau-Tsodikova, S. & Gatto, G. J., Jr. Protein posttranslational modifications: the chemistry of proteome diversifications. *Angew. Chem. Int. Ed. Engl.* **44,** 7342–7372 (2005).
2. Zhang, Z. *et al.* Identification of lysine succinylation as a new post-translational modification. *Nat. Chem. Biol.* **7,** 58–63 (2011).
3. Park, J. *et al.* SIRT5-mediated lysine desuccinylation impacts diverse metabolic pathways. *Mol. Cell* **50,** 919–930 (2013).
4. Du, J. *et al.* Sirt5 is a NAD-dependent protein lysine demalonylase and desuccinylase. *Science* **334,** 806–809 (2011).
5. Xie, Z. *et al.* Lysine succinylation and lysine malonylation in histones. *Mol. Cell Proteomics* **11,** 100–107 (2012).
6. Chou, K. C. Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.* **273,** 236–247 (2011).
7. Liu, Z. *et al.* CPLM: a database of protein lysine modifications. *Nucleic Acids Res.* **42,** D531–536 (2014).
8. Uniprot, C. The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.* **38,** D142–148 (2010).

9. Boeckmann, B. *et al.* The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31,** 365–370 (2003).
10. Xu, Y., Ding, J., Wu, L. Y. & Chou, K. C. 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 (2013).
11. Tang, Y. R., Chen, Y. Z., Canchaya, C. A. & Zhang, Z. GANNPhos: a new phosphorylation site predictor based on a genetic algorithm integrated neural network. *Protein Eng. Des. Sel.* **20,** 405–412 (2007).
12. Platt, J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers* **10,** 61–74 (1999).
13. Xue, Y., Zhou, F., Fu, C., Xu, Y. & Yao, X. SUMOsp: a web server for sumoylation site prediction. *Nucleic Acids Res.* **34,** W254–257 (2006).
14. Chen, Y. Z., Chen, Z., Gong, Y. A. & Ying, G. SUMOhydro: a novel method for the prediction of sumoylation sites based on hydrophobic properties. *PLoS One* **7,** e39195 (2012).
15. Ren, J. *et al.* Systematic study of protein sumoylation: Development of a site-specific predictor of SUMOsp 2.0. *Proteomics* **9,** 3409–3412 (2009).
16. Xu, J. *et al.* A novel method for high accuracy sumoylation site prediction from protein sequences. *BMC Bioinformatics* **9,** 8 (2008).
17. Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure* **405,** 442–451 (1975).
18. Chen, W., Feng, P. M., Lin, H. & Chou, K. C. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. *Nucleic Acids Res.* **41,** e68 (2013).
19. Kim, J. H., Lee, J., Oh, B., Kimm, K. & Koh, I. Prediction of phosphorylation sites using SVMs. *Bioinformatics* **20,** 3179–3184 (2004).
20. Wong, Y. H. *et al.* KinasePhos 2.0: a web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns. *Nucleic Acids Res.* **35,** W588–594 (2007).
21. Chang, W. C. *et al.* Incorporating support vector machine for identifying protein tyrosine sulfation sites. *J. Comput. Chem.* **30,** 2526–2537 (2009).
22. Shao, J. L., Xu, D., Tsai S., Wang, Y. F. & Ngar, S. Computational Identification of Protein Methylation Sites through Bi-Profile Bayes Feature Extraction. *PLoS One* **4,** e4920 (2009).
23. Fan, G. L. & Li, Q. Z. Predicting protein submitochondria locations by combining different descriptors into the general form of Chou's pseudo amino acid composition. *Amino Acids* **43,** 545–555 (2012).
24. Sahu, S. S. & Panda, G. A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction. *Comput. Biol. Chem.* **34,** 320–327 (2010).
25. Sun, X. Y. *et al.* Identifying protein quaternary structural attributes by incorporating physicochemical properties into the general form of Chou's PseAAC via discrete wavelet transform. *Mol. Biosyst.* **8,** 3178–3184 (2012).
26. Chou, K.-C. & Shen, H. B. REVIEW : Recent advances in developing web-servers for predicting protein attributes. *Natural Science* **01,** 63–92 (2009).
27. Liu, Z. *et al.* GPS-PUP: computational prediction of pupylation sites in prokaryotic proteins. *Mol. Biosyst.* **7,** 2737–2740 (2011).
28. Xue, Y. *et al.* GPS-SNO: computational prediction of protein S-nitrosylation sites with a modified GPS algorithm. *PLoS One* **5,** e11290 (2010).

Acknowledgements

This work is supported by the Natural Science Foundation of China (No.11301024, No.11371365, No.31201002, No. 11131009).

Author Contributions

Y.X. designed and supervised experiments. Y.D. and Y.L. performed experiments and data analysis. J.D. and L.W. developed the online webserver. Y.X. and N.D. wrote the manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Xu, Y. *et al.* iSuc-PseAAC: predicting lysine succinylation in proteins by incorporating peptide position-specific propensity. *Sci. Rep.* doi: 10.1038/srep10184 (2015).