PLOS ONE

RESEARCH ARTICLE

Computational Identification of Protein Pupylation Sites by Using Profile-Based Composition of *k*-Spaced Amino Acid Pairs

Md. Mehedi Hasan[1☯], Yuan Zhou[1☯], Xiaotian Lu[1], Jinyan Li[2], Jiangning Song[3,4], Ziding Zhang[1]*

1 State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University, Beijing, 100193, China, **2** Advanced Analytics Institute and Centre for Health Technologies, University of Technology, Sydney, 81 Broadway, NSW 2007, Australia, **3** National Engineering Laboratory for Industrial Enzymes and Key Laboratory of Systems Microbial Biotechnology, Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin, 300308, China, **4** Monash Bioinformatics Platform and Department of Biochemistry and Molecular Biology, Faculty of Medicine, Monash University, Melbourne, VIC 3800, Australia

☯ These authors contributed equally to this work.
* zidingzhang@cau.edu.cn

Citation: Hasan MM, Zhou Y, Lu X, Li J, Song J, Zhang Z (2015) Computational Identification of Protein Pupylation Sites by Using Profile-Based Composition of *k*-Spaced Amino Acid Pairs. PLoS ONE 10(6): e0129635. doi:10.1371/journal.pone.0129635

Academic Editor: Yu Xue, Huazhong University of Science and Technology, CHINA

Received: February 5, 2015

Accepted: May 10, 2015

Published: June 16, 2015

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: MMH was financially supported by the China Scholarship Council (CSC). JL was supported by an ARC Discovery Project (DP130102124). JS is a recipient of the Hundred Talents Program of the Chinese Academy of Sciences (CAS).

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Prokaryotic proteins are regulated by pupylation, a type of post-translational modification that contributes to cellular function in bacterial organisms. In pupylation process, the prokaryotic ubiquitin-like protein (Pup) tagging is functionally analogous to ubiquitination in order to tag target proteins for proteasomal degradation. To date, several experimental methods have been developed to identify pupylated proteins and their pupylation sites, but these experimental methods are generally laborious and costly. Therefore, computational methods that can accurately predict potential pupylation sites based on protein sequence information are highly desirable. In this paper, a novel predictor termed as pbPUP has been developed for accurate prediction of pupylation sites. In particular, a sophisticated sequence encoding scheme [i.e. the profile-based composition of *k*-spaced amino acid pairs (pbCKSAAP)] is used to represent the sequence patterns and evolutionary information of the sequence fragments surrounding pupylation sites. Then, a Support Vector Machine (SVM) classifier is trained using the pbCKSAAP encoding scheme. The final pbPUP predictor achieves an AUC value of 0.849 in10-fold cross-validation tests and outperforms other existing predictors on a comprehensive independent test dataset. The proposed method is anticipated to be a helpful computational resource for the prediction of pupylation sites. The web server and curated datasets in this study are freely available at http://protein.cau.edu.cn/pbPUP/.

Introduction

The bacterial prokaryotic ubiquitin-like protein (Pup) is initially perceived as a small protein related to post-translational modifications (PTMs). Pup is an intrinsically unstructured protein consisting of 64 amino acids [1, 2]. In the tagging system referred as pupylation, this protein covalently attaches to target lysines for proteasomal degradation by forming isopeptide bonds [3–5]. In eukaryotes, the ubiquitin-proteasome degradation pathway was discovered in the late 1970's [6], while the analogous Pup-proteasome pathway was not identified in prokaryotes until recently [5, 7, 8]. To date, the proteasomal Pup has been discovered in the phyla Actinobacteria and Nitrospira species [9]. The evidence of Pup proteasome degradation pathway has been rapidly accumulating in both the *in vitro* [10, 11] and *in vivo* systems [12].

Pupylation and ubiquitylation are functionally identical but their enzymology is different. In general, ubiquitylation requires three types of enzymes: ubiquitin-activating enzymes, ubiquitin-conjugating enzymes, and ubiquitin ligases [13]. Comparatively, the pupylation process involves two enzymes: one is the deamidase of Pup (DOP) which deamidates the C-terminal glutamine of Pup to glutamate [14, 15], and the other is the proteasome accessory factor A (PafA) which proceeds the deamidase process by attaching Pup to a specific lysine [16, 17]. More specifically, pupylation enzymes are originated from bacterial organisms and show no homology to ubiquitylation enzymes [18, 19].

The Pup-proteasome degradation pathway plays a nutritional role under nitrogen starvation by recycling amino acids [20]. This proteasomal pathway is also critical for the virulence of bacteria [21, 22]. Therefore, identification of pupylated substrates is fundamentally important for understanding both the physiological and pathological mechanisms. A number of large-scale proteomic studies have been performed to identify pupylated proteins based on the molecular signature of pupylated sites [23–27]. Despite the increasing number of experimentally determined pupylated proteins, the underlying mechanism of protein pupylation specificity remains largely unknown [25]. On the other hand, large-scale experimental identification of pupylation substrates is laborious, time-consuming and costly. As an alternative, accurate and cost-effective prediction methods can be used to complement the experimental efforts.

Up to now, a few computational approaches have been developed to predict pupylation sites [28–31]. Xue and co-workers [30] proposed the first predictor named GPS-PUP, which was developed from their original Group-based Prediction System (GPS) with three procedures (i.e. weight training, motif length selection, and matrix mutation) for performance improvement. In 2013, Tung [29] used a training dataset collected from the PupDB database [32] and an encoding scheme called the composition of *k*-spaced amino acid pair (CKSAAP) to develop a predictor called iPUP. Support Vector Machine (SVM) together with a backward feature selection method was used to train the classifier. Both GPS-PUP [30] and iPUP [29] predictors yielded good performance for predicting pupylation sites. In particular, they achieved higher specificity, although their sensitivity was generally low. More recently, Chen et al. [31] developed a predictor PupPred based on balanced training data (1:1 ratio of positive to negative samples). To train the classifier, PupPred combined the k-nearest neighbor (KNN) algorithm with a variety of features including binary features, amino acid pairs, protein secondary structures, position-specific scoring matrix (PSSM) and physicochemical properties. They demonstrated that the encoding of amino acid pairs, the implementation of F-measures for feature selection and the SVM-based classifier contributed to the improved performance of PupPred.

However, the overall performance of the aforementioned three existing predictors is still not satisfying and there is enough room for improvement. To develop a machine learning-based predictor, it is important to devise an appropriate encoding scheme to represent the sequence fragments surrounding pupylation/non-pupylation sites. In the current study, we develop a

new SVM predictor named pbPUP based on an improved CKSAAP encoding, i.e. the profile-based composition of *k*-spaced amino acid pairs (pbCKSAAP). The traditional CKSAAP encoding has been widely and successfully used in diverse bioinformatics prediction tasks, such as the prediction of pupylation sites [29], flexible/rigid region [33], O-glycosylation sites [34], ubiquitination sites [35], palmitoylation sites [36], methylation sites [37] and phosphorylation sites [38]. Compared with the traditional CKSAAP encoding, the pbCKSAAP encoding scheme has the advantage of integrating the sequence evolutionary information from the profile (i.e. PSSM) generated by PSI-BLAST search. Originally developed for the prediction of membrane protein [39], pbCKSAAP has revealed more powerful performance in some applications such as the prediction of bacterial pathogen effectors [40].

In this study, the pbPUP predictor was constructed using the training dataset of iPUP [29]. An independent test dataset [25, 29] was used for making fair performance comparison among different methods. The results indicated that pbPUP achieved significantly improved performance on the independent tests compared with other existing methods. Moreover, we also conducted a series of computational analyses to provide in-depth understandings of the pbCKSAAP encoding. Finally, the proposed method pbPUP has been implemented as a web server. Taken together, the current study provides a useful tool for predicting pupylation sites as well as valuable insights into the important sequence patterns surrounding pupylation sites.

Materials and Methods

In brief, pbPUP is an SVM-based predictor, which was constructed using the pbCKSAAP sequence encoding scheme. An overview of the computational framework of the proposed pbPUP predictor is shown in Fig 1.

Data preparation

In this study, two datasets were used to train and assess the proposed pbPUP predictor, including the training dataset of iPUP (S1 Table) and an independent test dataset (S2 Table). The experimentally validated pupylation sites (lysine residues) were considered as positive samples, while all the remaining lysine residues that have not been verified as pupylation sites in these proteins were considered as negative samples (i.e. non-pupylation sites). Each site was represented as a sequence fragment with lysine (K) in the center. These two datasets are also summarized in Table 1.

The iPUP training dataset was previously compiled to train the iPUP predictor [29], which includes 162 pupylated proteins covering 183 positive and 2205 negative sites. The iPUP training dataset was also employed to train our pbPUP predictor. The numbers of positive and putative negative samples are highly imbalanced in the original iPUP training dataset (~1:12); this imbalance will hamper model training. Therefore, a relatively balanced dataset with a 1:2 ratio of positives to negatives (i.e. 183 positive sites and 366 randomly selected negative sites) was compiled to train our pbPUP predictor.

An independent test dataset was also compiled to benchmark the prediction performance of different predictors. First, 20 pupylated proteins, originally used as the independent test data of iPUP, was directly used in our work. Moreover, we also collected 55 pupylated proteins from a recent work [25]. Among these 55 proteins, the lysine positions of four proteins did not match with the UniProt database (http://www.uniprot.org/). Thus, these four proteins were removed from our study. Finally, we obtained an independent dataset containing 71 proteins with 86 experimentally validated pupylation sites and 1136 putative non-pupylation sites. In the

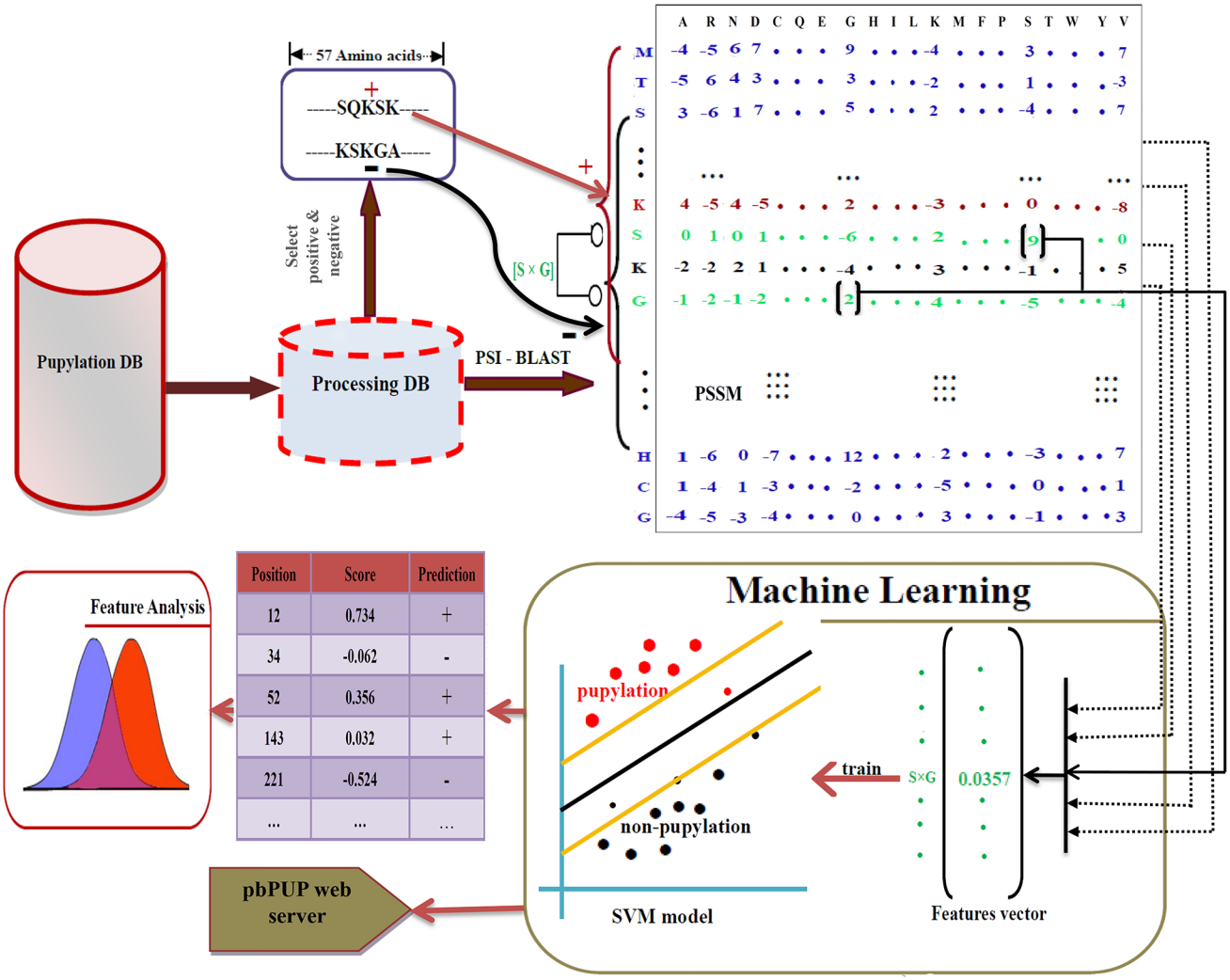

Fig 1. Overview of the proposed pbPUP predictor. The full-length sequence of a pupylated protein is first used to generate the PSSM profile by running PSI-BLAST search against the NCBI NR90 database. Meanwhile, the PSSM matrixes corresponding to pupylation and non-pupylation sites are extracted from the whole profile. The encoded profile-based features are used as the input to train a SVM classifier. After optimization of the SVM parameters, the best SVM model is constructed based on the 10-fold cross-validation performance. Finally, a web server pbPUP is implemented and made available for interested users to predict the potential pupylation sites from the submitted proteins.

doi:10.1371/journal.pone.0129635.g001

independent test, all the pupylation and non-pupylation sites were used to assess the performance of different predictors. We believe that the performance assessed using the highly imbalanced data could reflect the real applications of different predictors.

Table 1. The statistics of pupylated proteins and their pupylation sites used in this study.

	The iPUP training dataset	Independent test dataset
Number of pupylated proteins	162	71
Number of pupylation lysines	183	86
Number of non-pupylation lysines	2205 (366)	1136

Values in parentheses represent the number of sites used in this study.

doi:10.1371/journal.pone.0129635.t001

Encoding scheme of pbCKSAAP

The encoding scheme of pbCKSAAP has been used in previous studies [39, 40]. Briefly, a *k*-spaced amino acid pairs can be represented as $p_i\{k\}p_j(i, j = 1, 2, \ldots, 20)$, where p_i and p_j denote any two residues of the 20 amino acid types. When $k = 0$, $p_i\{k\}p_j$ stands for a dipeptide and a total of 20×20 = 400 different dipeptides should be taken into account. In this work, $k = 0, 1, 2, 3$ and 4 were jointly considered (i.e. $k_{max} = 4$). Thus, the feature vector of each pupylation/non-pupylation site has a dimensionality of 400×5 = 2000. To conduct the pbCKSAAP encoding, each protein sequence was searched by PSI-BLAST against the NCBI NR90 database (version of December 2010) to generate a profile (i.e. PSSM matrix). The e-value cutoff for the inclusion of new sequences and iteration times were set as 1.0×10^{-4} and 3, respectively. For each pupylation/non-pupylation site, the corresponding PSSM matrix was extracted from the whole profile. If an amino acid pair $p_i\{k\}p_j$ appears between the residue positions t and $t+k+1$ in the PSSM matrix, the composition score can be calculated using the following equation:

$$S_{i,j} = \sum^{\mathrm{N}} \max\{\min\{\mathrm{PSSM}(t, p_i), \mathrm{PSSM}(t + k + 1, p_j)\}, 0\} \quad (1)$$

where PSSM (t, p_i) denotes the score of amino acid p_i at the t^{th} row position of PSSM, PSSM $(t+k+1, p_j)$ stands for the score of amino acid p_j at the $(t+k+1)^{\mathrm{th}}$row position of PSSM, N means $p_i\{k\}p_j$ appears N times in the pupylation/non-pupylation site. Furthermore, we normalized $S_{i,j}$ using the following formula:

$$S'_{i,j} = \frac{S_{i,j}}{L - k - 1} \quad (2)$$

where L denotes the total length of sequence fragment, i.e. window size = L. Finally, we generated the pbCKSAAP encoding, which is a 2000-dimensional feature vector for each pupylation/non-pupylation site.

To investigate the evolutionary conservation of pupylated or non-pupylated sites, we calculated the average PSSM value (APV) of each position (i.e. the average of each row of the PSSM matrix) in the flanking sequence fragments of each pupylated/non-pupylated site. These APVs were further averaged. More specifically, because the optimal window size in this study was 57, the APVs of the positions [–28,–1] were averaged to obtain the APV of the upstream sites, while the APVs of the positions [+1,+28] were averaged to obtain the APV of the downstream sites.

Encoding scheme of CKSAAP

Compared with pbCKSAAP, the encoding scheme of CKSAAP is quite simple, which can be directly calculated from the sequence fragments of pupylation/non-pupylation sites. By effectively representing the short sequence motif information in protein sequences or fragments, CKSAAP is an important encoding scheme in many prediction tasks [29, 34–36, 38, 39]. In this work, we retrained the SVM model using the CKSAAP encoding scheme with the purpose of comparing the performance between pbCKSAAP and CKSAAP. To conduct a stringent comparison, the same window size and the same k_{max} value were adopted. Thus, a 2000-dimensional feature vector was also generated in the CKSAAP encoding scheme. More details about the CKSAAP encoding can be found in our previous studies [34, 35].

Feature selection

For a pupylation site, the proposed pbCKSAAP encoding represents its flanking sequence pattern in a comprehensive manner, resulting in a high-dimensional, partially redundant feature

vector. It is well known that there could be some key residues or motifs which contribute significantly to the identification of PTM sites [34, 41, 42]. However, it would be challenging to read-out the key residues or motifs directly from the high-dimensional feature vector of the pbCKSAAP encoding. Therefore, we employed a well-established dimensionality reduction method, Chi-Squared (χ^2) to characterize the top ranking features [39]. Let X be a feature with n possible values $x_1, x_2, \ldots, x_n$ with the probability $P(X = x_j) = p_j$. Then, for a dataset with c_{tot} positive samples and d_{tot} negative samples, the χ^2 score of this feature can be calculated using the following formula:

$$\chi^2 = \sum_{j=1}^{n} \left[\frac{(c_j - c_{tot} \cdot p_j)^2}{c_{tot} \cdot p_j} + \frac{(d_j - d_{tot} \cdot p_j)^2}{d_{tot} \cdot p_j} \right] \quad (3)$$

In addition to the aforementioned variables (p_j, c_{tot}, d_{tot}), c_j is the observed numbers of the positive samples whose feature value $X = x_j$, while d_j is the observed numbers of the positive samples whose feature value $X = x_j$. By definition, a larger value of χ^2 indicates that the corresponding feature has a greater impact on the discrimination capability of the predictor. More information about the χ^2 feature selection method can be found in the literature [39].

Model training

In our study, SVM was used to build the classifiers to distinguish the pupylation sites from non-pupylation sites. As an efficient machine learning algorithm, SVM has been widely used in protein bioinformatics [43–48]. In this work, the LIBSVM package (http://www.csie.ntu.edu.tw/~cjlin/libsvm/) was used as an implementation of SVM to train the classifiers [49]. The kernel radial basis function (RBF) was selected and two parameters C and γ were optimized based on the training dataset through a grid search provided by the LIBSVM package. The ranges of both C and γ were set as $[2^{-7}, 2^8]$, which resulted in 225 grids. All the grids were evaluated based on 10-fold cross validation in order to find the optimal parameter combination of C and γ.

Model evaluation and cross validation

10-fold cross-validation tests were performed to assess the performance of our prediction model. In the 10-fold cross-validation tests, the training dataset was divided into10 subgroups with approximately equal size. At each cross-validation step, one subgroup was singled out as the test dataset to assess the performance of the classifier, while the classifier was trained using the remaining 9 subgroups. The performance of each cross-validation produced a single estimation and this procedure was repeated 10 times. To evaluate the model's performance, four measurements were calculated, including accuracy (Ac), sensitivity (Sn), specificity (Sp), and Matthews' correlation coefficient (MCC). The following formulae are used to calculate these measures:

$$\mathrm{Ac} = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

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

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

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TN + FN) \times (TP + FP) \times (TN + FP) \times (TN + FP)}} \quad (7)$$

Where *TP*, *FP*, *TN*, and *FN* represent the numbers of true positive, false positive, true negative and false negative, respectively. Furthermore, the receiver-operating characteristic (ROC) curve, which plots Sn against 1-Sp at different thresholds, was also employed for performance assessment. To further quantify the performance, the areas under the ROC curves (AUCs) were calculated by the pROC package in R software [50, 51].

Results and Discussion

Performance assessment on the training dataset

The iPUP training dataset was used to develop the pbPUP predictor. The ratio of positive to negative samples is nearly 1:12 in this dataset, which is highly imbalanced. It has been established that machine learning algorithms become computationally intractable and their accuracy is strongly affected due to the nature of the unbalanced datasets [52, 53]. To address this, many PTM site predictors employ a relatively balanced ratio of positives to negatives to train the classification models, including the prediction of pupylation sites as well [31, 54, 55]. In the current study, a 1:2 ratio of positives to negatives was used for the training dataset to develop the proposed SVM predictor.

The window size is an important factor of the prediction performance, which reflects the influence of surrounding residues on the discrimination of pupylation from non-pupylation sites. The window sizes ranging from 25 to 61were optimized based on the AUC values. For each window size, the SVM parameters were optimized through the grid search, and the corresponding AUC value was obtained from the 10-fold cross-validation test on the training set. As a result, the optimal window size of 57 (the corresponding optimal SVM parameters are $C = 8$ and $\gamma = 2$) was finally selected, though the performance increase with the window size ranging from 45–57 was not prominent (S1 Fig).

At the 90% specificity control (SVM score$\geq$0.0), pbPUP reached an accuracy of 76.06% (Sn = 48.15% and MCC = 0.44). Meanwhile, in terms of ROC curve (Fig 2A), pbPUP achieved an AUC value of 0.849. Furthermore, we also conducted 4-, 6-, and 8-fold cross-validation tests, and the corresponding AUC values were 0.829, 0.838 and 0.846, respectively. In summary, we conclude that pbPUP predictor provides a stable and promising performance in the cross-validation tests on the training dataset.

Performance comparison of pbPUP with three existing predictors on the independent dataset

To compare the performance of pbPUP and three other existing predictors (iPUP, GPS-PUP, and PupPred), we compiled an independent dataset covering 71 pupylated proteins, which contain 86 pupylation and 1136 putative non-pupylation sites. Among these proteins, 20 proteins (i.e. the independent test set used in iPUP) were extracted from the original article of iPUP [29] and 51 proteins were retrieved from a recent study [25]. Although pbPUP and these three predictors did not employ the same training dataset for predicting pupylation sites, the independent dataset can allow for a generally fair performance comparison. To make the comparison, the independent data were directly submitted to the respective web servers. Note that the authors of iPUP combined the training and testing datasets when constructing the server. In other words, there were 20 proteins already included in the training data of the iPUP server. Accordingly, it is not reasonable to submit these same 20 proteins again to the iPUP server. Instead, we assessed the performance of the iPUP predictor on these 20 proteins according to their original literature. The rest 51 proteins were submitted to the iPUP server and the prediction results on the 20 proteins from the published literature were further combined for making

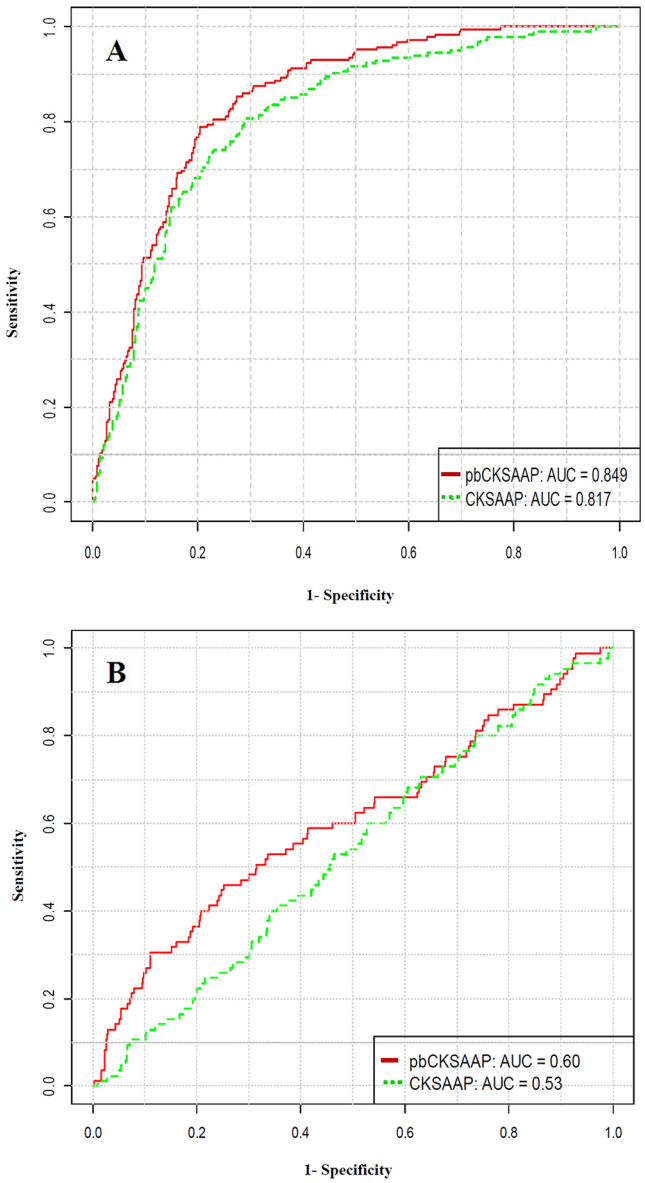

Fig 2. Performance comparison between pbCKSAAP and CKSAAP using ROC curves. (A) Performance comparison based on 10-fold cross-validation of the training dataset; (B) Performance comparison based on the independent test dataset.

doi:10.1371/journal.pone.0129635.g002

a comparison. Similar to the other predictors, we also reported the performance of pbPUP at High, Medium and Low confidence thresholds. To make a fair comparison, the thresholds of High, Medium and Low in pbPUP were set to ensure that the corresponding specificities were controlled at the same levels as those of GPS-PUP. As shown in Table 2, the pbPUP predictor achieved an improved performance with approximately 4%, 5%, and 3% higher MCC values under high, medium, and low confidence thresholds than iPUP (Table 2). The MCCs of the pbPUP predictor were nearly 7%, 8%, and 2% higher than the GPS-PUP predictor at high, medium, and low thresholds, respectively (Table 2). In addition, the pbPUP predictor achieved MCC values of almost 9%, 5%, and 5% higher than PupPred at high, medium, and low confidence thresholds (Table 2). The performance comparison results demonstrate that our

Table 2. The prediction performance of pbPUP and other existing predictors evaluated on the independent test dataset.

Predictor	Threshold[a]	Ac (%)	Sn (%)	Sp (%)	MCC (%)
GPS-PUP	High	83.89	19.76	88.74	6.73
	Medium	78.82	24.41	82.93	4.94
	Low	71.70	36.26	74.24	7.71
iPUP	High	81.13	29.06	84.90	9.56
	Medium	75.63	33.72	78.80	7.72
	Low	72.02	37.21	74.64	6.89
PupPred	High	88.93	9.19	94.77	4.33
	Medium	79.74	27.58	83.57	7.45
	Low	63.97	43.67	65.45	4.82
pbPUP	High	84.14	30.13	88.21	13.97
	Medium	78.72	37.65	81.79	12.46
	Low	70.15	44.70	72.05	9.38

[a]The threshold values of GPS-PUP, iPUP and PupPred were the same as those defined in the corresponding webservers. To make the performance comparison, the thresholds of High, Medium and Low in pbPUP were set as 0.06, 0.00 and -0.04, respectively. Thus, the corresponding specificities were controlled at the same levels as GPS-PUP.

doi:10.1371/journal.pone.0129635.t002

proposed pbPUP predictor provides a better or competitive performance with the other three existing predictors, indicating the encoding scheme of pbCKSAAP is very useful and powerful.

Interestingly, pbPUP and the other three existing predictors showed significantly lower performance on the independent data. Our analysis suggests that the sequence patterns of pupylation sites and surrounding regions in the training and independent datasets are highly different. The position-specific amino acid occurrences for the pupylation and putative non-pupylation sites in the training and independent datasets were visualized using the Two-Sample-Logo [56] (Fig 3). Generally, the amino acid pattern around the pupylation sites is somewhat camouflaged in the independent dataset (Fig 3B), because the independent data was collected from two distinct non-pathogenic bacteria *Escherichia coli* and *Corynebacterium glutamicum* [25, 29]. The pupylation data of the latter organism has never been considered by any of the predictors. Therefore, the collected independent dataset was novel and challenging. On one hand, by intensively exploiting evolutionary information, pbPUP could achieve better performance on these novel data. On the other hand, there might exist species-specific pupylation site patterns, similar to other PTM types such as acetylation [55]. Accordingly, more comprehensive predictors (e.g. species-specific predictors) need to be developed when more pupylation data become available in the future.

The influence of sequence redundancy on the predictive performance

The sequence redundancy might lead to the overestimation of prediction performance. Therefore, we adopted two approaches to remove the redundant sequences: 1) BLASTClust (http://www.ncbi.nlm.nih.gov/BLAST/docs/blastclust.html) was applied to remove redundant protein sequences with the 30% identity cutoff (i.e. redundancy removal at the protein level); 2) An in-house PERL script was used to remove redundant pupylated/non- pupylated peptides (also with 30% identity cutoff) at the peptide level. It is noteworthy that, as mentioned above, the authors of iPUP combined the training and testing datasets when they constructed the iPUP server. It is therefore not reasonable to submit any of the 20 proteins again to the iPUP server. To make a fair performance comparison, we had to keep these 20 proteins as they were (i.e., no

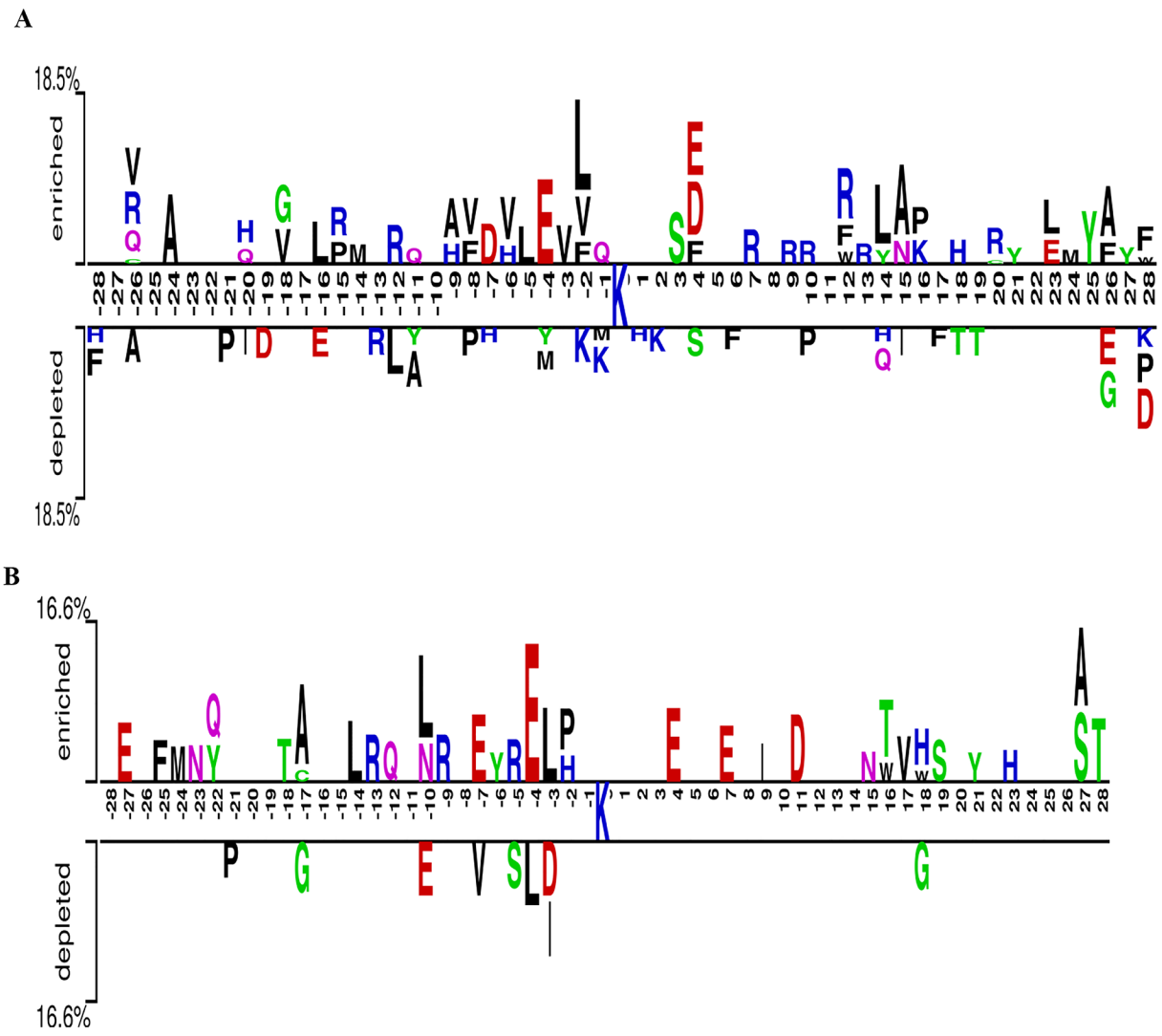

Fig 3. Sequence logo representations showing the amino acid occurrences between pupylation and putative non-pupylation sites. Only residues that were significantly enriched or depleted (*t*-test, $P<0.05$) flanking the centred pupylation sites are shown. Panel A represent the two-sample logo of the iPUP training dataset, while panel B plots the two-sample logo of the independent test dataset. The two-sample sequence logos were prepared using the web server http://www.twosamplelogo.org/.

doi:10.1371/journal.pone.0129635.g003

redundancy removal procedure was applied to these 20 proteins), and used the performance reported in their original literature to evaluate predictors' performance on these 20 proteins.

After removing the protein level sequence redundancy, we re-assembled a training dataset that contained 129 proteins with 149 pupylation and 298 non-pupylation sites (with the consistent 1:2 ratio of positives to negatives), and a testing dataset that contained 64 proteins with 76 pupylated and 1049 non-pupylation sites. As shown in S2 Fig, the overall performance of pbPUP in the 10-fold cross-validation decreased slightly (AUC = 0.841) after removal of the protein sequence redundancy. Furthermore, pbPUP could still achieve the best performance on the independent testing dataset (S3 Table). For example, when compared with iPUP, pbPUP achieved MCC values of approximately 4%, 3% and 1% higher under high, medium, and low thresholds, respectively. pbPUP also achieved at least a 2% MCC improvement compared with PupPred and GSP-PUP at any of the three confidence thresholds. These

performance comparison results prove that pbPUP predictor provides a better or competitive performance with the other three existing predictors on the independent test datasets, even after removal of the protein sequence redundancy.

In addition, we examined the predictors' performance after removing the peptide-level sequence redundancy. A training dataset including 148 pupylated sites and 296 non-pupylated sites were accordingly obtained (with the consistent 1:2 ratio of positives to negatives). Similar to the situation after protein-level sequence redundancy removal, there was only a small change of the overall cross-validation performance (AUC = 0.837). The independent test dataset after removal of the peptide-level sequence redundancy included 79 pupylated sites and 992 non-pupylated sites. On this dataset, pbPUP achieved the MCC values of 4%, 2%, 1% higher than iPUP at high, medium, and low confidence thresholds (S4 Table). Likewise, the MCC values of the pbPUP predictor was nearly 7%, 5%, and 1% better than the GPS-PUP predictor and 10%, 4%, and 2% better than PupPred at the corresponding thresholds (S4 Table). Altogether, we conclude that pbPUP predictor achieves a stable and competitive performance compared with other methods under both sequence-level and peptide-level sequence redundancy reduction conditions.

Comparison of the pbCKSAAP and CKSAAP encoding schemes

The CKSAAP encoding has been previously used for prediction of pupylation sites (i.e. the iPUP predictor) [29], and the aforementioned independent test has clearly shown that our pbPUP can outperform iPUP. Since the encoding schemes of pbCKSAAP and CKSAAP are developed based on a similar strategy, it is of particular interest to comprehensively compare these two encoding schemes. To this end, we re-trained the CKSAAP-based SVM model using the training dataset in this work. Note that the window size and SVM parameters were the same as those used for training pbPUP. Based on the 10-fold cross-validation tests, pbCKSAAP outperformed the conventional CKSAAP considerably (Fig 2A).The AUC value of pbCKSAAP was approximately 3% higher than that of CKSAAP. Moreover, pbCKSAAP achieved MCC, Ac, and Sn of about 4%, 2%, and 7% higher than CKSAAP, respectively, at the fixed Sp of 90%. In addition, on the independent test dataset, the pbCKSAAP method also achieved an AUC value of approximately 7% higher than CKSAAP for pupylation site prediction (Fig 2B). These results again suggest that pbCKSAAP achieved a significant performance improvement over CKSAAP for predicting pupylation sites.

To further compare pbCKSAAP with CKSAAP, the χ^2 feature selection method was applied to select the most important pbCKSAAP and CKSAAP features. In particular, we found that the average χ^2 feature score of pbCKSAAP features was much higher than that of CKSAAP features (Fig 4A). This suggests that the pbCKSAAP features contained more important information than the CKSAAP features. To make a stringent comparison, we used the same feature score cutoff (i.e. $\chi^2 \geq 3$) to select more informative features from both CKSAAP and pbCKSAAP sequence encodings. When this cutoff was applied, the number of selected pbCKSAAP features was 196, while the number of selected CKSAAP features was only 169 (Fig 4B). The number of common features shared by both pbCKSAAP and CKSAAP was 45 (Fig 4B). In summary, we conclude that pbCKSAAP contained more informative features than CKSAAP, which provides an important evidence to explain the better performance of pbCKSAAP.

Compared with CKSAAP, pbCKSAAP is able to capture the evolutionary information contained in the PSSM matrix, which may explain the performance difference between CKSAAP and pbCKSAAP. In other words, the better performance of pbCKSAAP suggests that the residue conservation patterns of pupylation sites are significantly different from those of non-pupylation sites. To support our speculation, we calculated the average PSSM

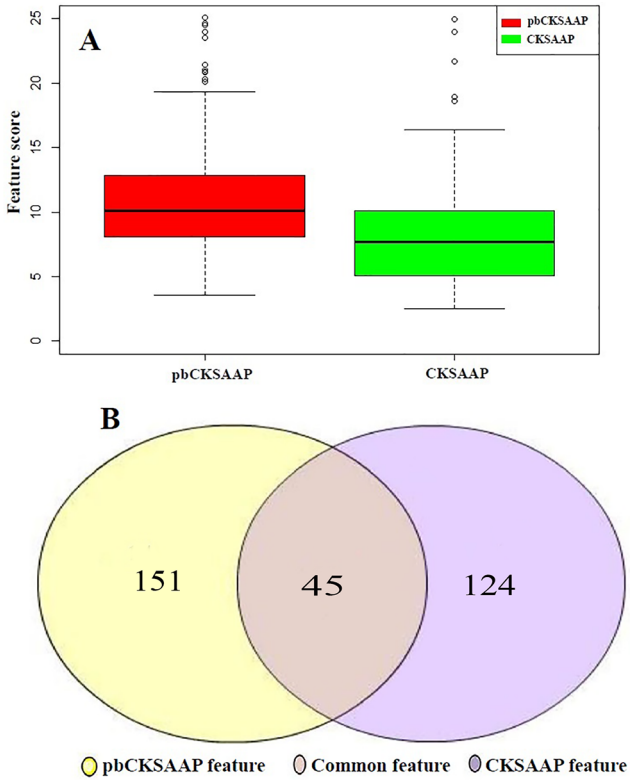

Fig 4. Comparison of the selected features in pbCKSAAP and CKSAAP using the χ^2 feature selection method. (A) Feature scores of pbCKSAAP and CKSAAP; (B) The numbers of selected features in pbCKSAAP and CKSAAP with the same feature selection score cutoff $\chi^2 \geq 3$.

doi:10.1371/journal.pone.0129635.g004

score (APV) of each residue surrounding pupylation and non-pupylation sites, as a useful indicator of residue conservation. The scores were calculated from each line of the PSSM matrix of the given sequences. In particular, the average PSSM values (APV) were summarized for the upstream (positions from -28 to -1), center (position 0 or central lysine) and downstream (positions from +1 to +28) regions surrounding pupylation sites. The evolutionary conservation scores of PSSM between pupylation and non-pupylation sites are illustrated in Fig 5. *P*-values were also calculated using the one-tailed *t*-test for residue positions in the upstream, center and downstream regions between pupylation and non-pupylation sequence fragments. As a result, we found that the *P*-values in the upstream and downstream regions were greater than 0.05 ($P = 0.333$ and 5.44×10^{-2}, respectively), which means that the two groups of samples were not significantly different. Nevertheless, certain adjacent amino acid positions surrounding pupylation sites had significantly higher APV scores, especially the upstream positions -25, -8, -3,-4, -1 and downstream positions +3, +4, +7, +8, +11, +15, +18, +22, +25 (S3 Fig). On the other hand, *P*-value in the center region of lysine position was also less than 0.05 ($P = 3.31\times10^{-3}$), which suggests that pupylation sites are relatively more conserved (Fig 5). Altogether, our results confirm that the local regions surrounding pupylation sites have more conserved sequence patterns than the non-pupylation counterparts, which might possibly explain why the pbCKSAAP scheme performed better than the simple CKSAAP scheme for this prediction task.

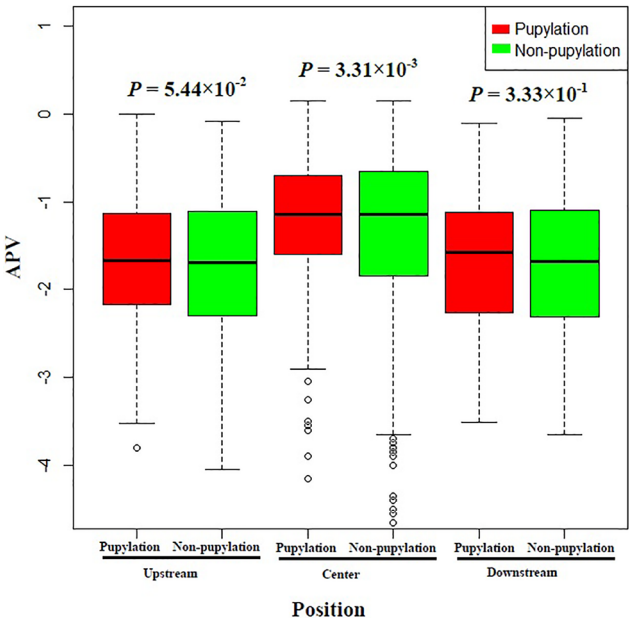

Fig 5. Box plots of the average PSSM values (APV) of amino acids positioned in the upstream, center, and downstream regions of pupylation and non-pupylation sites. Red color denotes pupylation sites, while green color denotes non-pupylation sites.

doi:10.1371/journal.pone.0129635.g005

Significant features of pbCKSAAP

As mentioned above, a well-established feature selection method χ^2 was used to select the most important features from the high-dimensional pbCKSAAP encoding that contributed to the performance. We performed multiple rounds of experiments to select appropriate feature sets; however, it turned out that there was no significant improvement in the corresponding performance using the selected features. Probably due to the fact that SVM has a good tolerance of high-dimensional input features, the feature selection did not result in a better SVM model, which is consistent with the observations in our previous studies. Therefore, feature selection was not utilized in our final predictor. Although the feature selection strategy did not lead to significant performance improvement, we identified the top ranked 30 amino acid pairs for the purpose of investigating the most significant residues and positions surrounding pupylation and non-pupylation sites. The top 30 residue pair scores and their corresponding positions are listed in S5 Table. These important features are also presented in a radar diagram (Fig 6). The feature 'N×××E' (i.e. 3-spaced residue pair of 'NE', where '×' stands for any residue) was the most important amino acid pair, representing the most enriched motif surrounding pupylation sites. Similarly, the feature 'AA' which represents a 0-spaced residue pair of 'AA' is the most important and enriched in the non-pupylated sites (Fig 6). Interestingly, the majority of the top 30 features contain charged residues such as K, R, H, E, and D (Fig 6), indicating that charged residues may play an important role in the recognition of pupylation sites. We also observed that amino acid pairs that cover all possible k values (i.e. $k = 0, 1, 2, 3$ and 4) were included as the most significant features (Fig 6), suggesting that all spaced amino acid pairs are necessary to make a collective contribution to the prediction of pupylation sites.

Although the SVM framework what pbPUP used is a black-box computational model, the above analyses have provided clues for interpreting the biological knowledge of the pbCKSAAP encoding scheme. That is, the pbCKSAAP encoding is able to represent and depict the weakly conserved motifs hidden in the surrounding sequences of pupylation sites. Three important

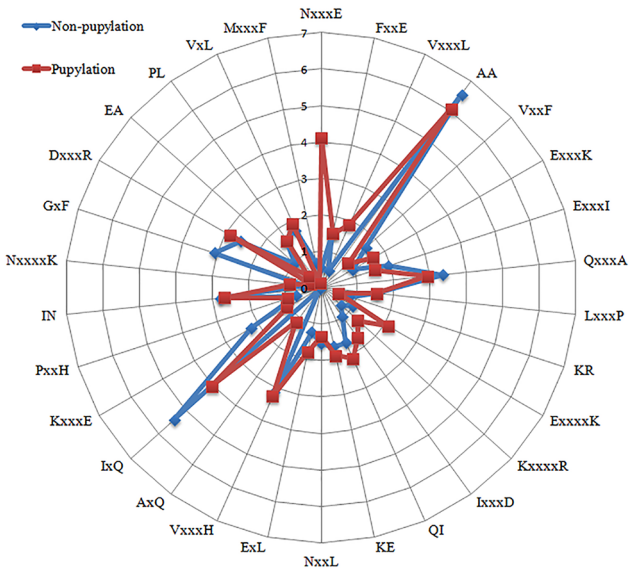

Fig 6. Top 30 amino acid pairs selected by the χ^2 feature selection method. Red color denotes pupylation sites, while blue color denotes non-pupylation sites. The radar diagram is represented by the composition of each residue pair whose length is proportional to the composition of pbCKSAAP features.

doi:10.1371/journal.pone.0129635.g006

properties of this encoding should be highlighted. The first one is the usage of *k*-spaced amino acid pair. As a sketch of sequence motif, *k*-spaced amino acid pair could better reflect the coordinated pairs of residues surrounding the pupylation sites. Indeed, as shown in Fig 6 and S5 Table, amino acid pairs covering all possible *k* values (i.e. $k = 0, 1, 2, 3$ and 4) and almost all of the 20 amino acids (except rare amino acids like C, W, Y) could be found in the list of the top 30 most-informative features. These results indicate that the spectrum of possible *k*-spaced amino acid pairs could serve as an enriched and explicit representation of the sequence patterns. The second key property of the pbCKSAAP encoding is the usage of position-independent composition encoding. We mapped the top 30 informative amino acid pairs onto the pupylated peptides in both the training and testing datasets, respectively (Fig 7). It is obvious that most of them did not exhibit concentrated distributions, but were instead dispersed along the peptide fragment. Even for the amino acid pairs that showed obvious concentrated distributions (e.g. FxxE and KxxxxK), their distributions were still somehow different in the training and testing samples (Fig 7). For instance, the distribution of FxxE shifted towards the downstream in the testing samples, while the distribution of KxxxxK shifted towards the upstream in the training samples. Therefore, in this situation, the position-independent encoding might be able to better describe the sequence patterns than a position-dependent encoding. On the other hand, it is also noticeable that pbCKSAAP did not completely disregard other informative position-dependent amino acid patterns. For example, a conserved enrichment of E at positions -4 and +4 was observed in the flanking sequences of pupylation sites (Fig 3). Accordingly, the amino acid pairs ExxxK and KxxxE were ranked among the top features (Fig 6) and exhibited conserved positional distributions in the training and testing samples (Fig 7). Last but not least, pbCKSAAP embedded the evolutionary information into its encoding. Our previous analysis has shown that several positions flanking the pupylation sites were slightly more conserved than the corresponding positions of non-pupylation sites (S3 Fig). pbCKSAAP took advantage of this weak conservation pattern to prioritize the weakly conserved amino acid pairs. To characterize the pbCKSAAP-specific features and CKSAAP-specific features (Fig 4B),

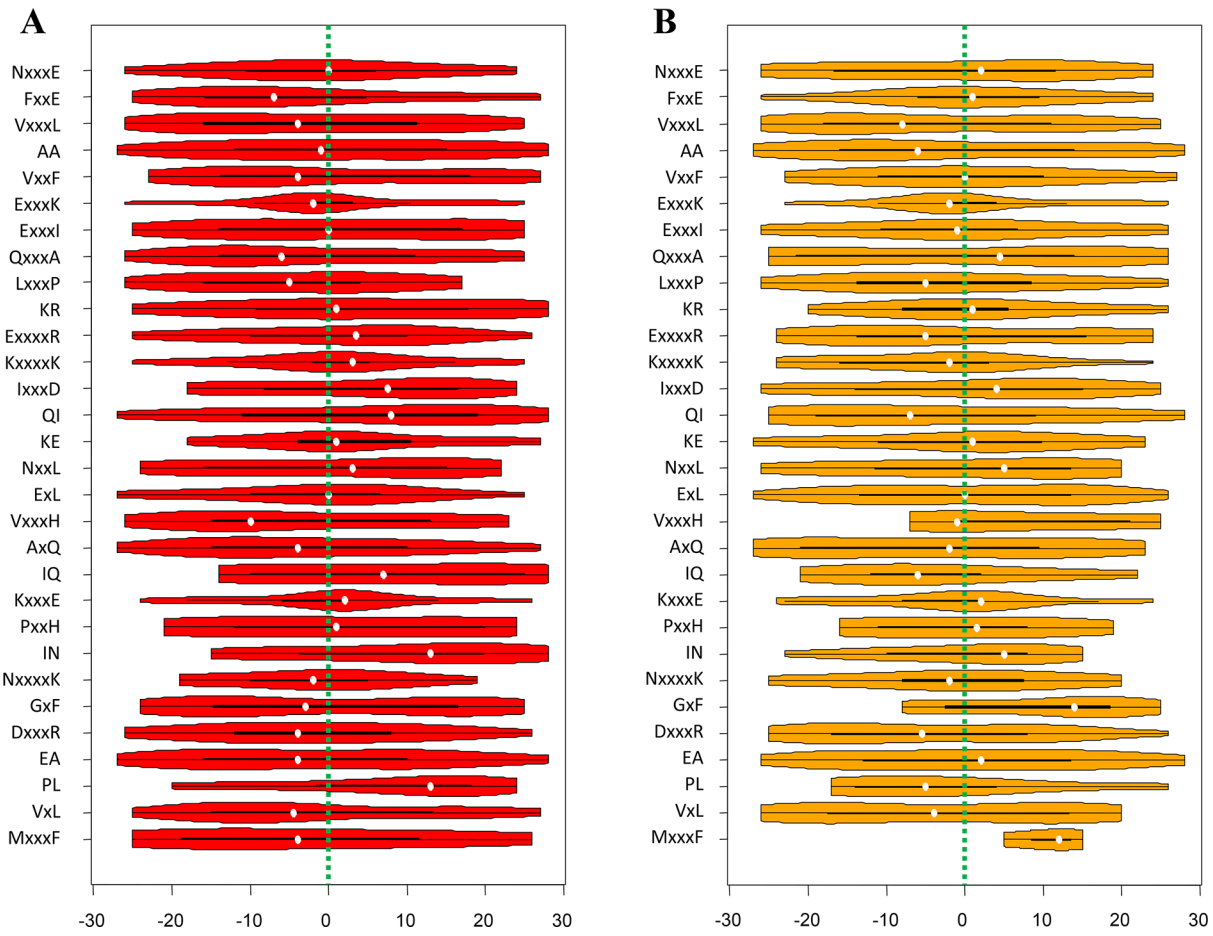

Fig 7. The violin plots illustrating the positional distributions of the top 30 amino acid pairs of the pbCKSAAP encoding on the pupylated peptides. (A) The distributions on the pupylated peptides from the training samples; (B) The distributions on the pupylated peptides from the independent testing samples. The white dots indicate the median values, the black boxes indicate the ranges between 1st quartiles and 3rd quartiles, while the outskirt violin-like shapes denote the probability destiny plots. For clarity, green dashed lines indicating the position of the central lysines are also added.

doi:10.1371/journal.pone.0129635.g007

we compared the numbers of their matched pupylated peptides on the independent testing dataset. As shown in S4 Fig, pbCKSAAP-specific features generally matched more pupylated peptides than CKSAAP-specific features. Especially, the fraction of zero-matched features of pbCKSAAP-specific features was significantly smaller than that of CKSAAP-specific features, indicating that pbCKSAAP is able to extract weakly conserved amino acid pairs to achieve more accurate prediction performance.

Web server implementation

As an implementation of our method, a web server of pbPUP (profile-based pupylation site predictor) has been made available at http://protein.cau.edu.cn/pbPUP/ to the research community. The web server was developed using Perl, CGI scripts, PHP and HTML. The input and exemplar output web pages of the server are shown in S5A and S5B Fig, respectively. In the input web page, users can submit their query sequence by pasting it into the text box. After submitting the query sequence to the server, it will initially generate lysine fragments of all candidate pupylation sites. Simultaneously, the server will generate the PSSM matrix of the query sequence by performing PSI-BLAST search and calculate the pbCKSAAP encodings for all the

generated fragments. Finally, the server will calculate the prediction scores of all the fragments with the assistance of SVM classifier. After the submission job is completed, the server will return the prediction result in the output webpage, consisting of the job ID, the query protein name, residue position, and the SVM score of the predicted pupylation sites in a tabular form. Note that the current pbPUP server predicts pupylation sites at the 90% specificity control. Users can also view the results in the text format. The prediction results will be generated for all candidate lysine residues of the submitted sequence. User will receive a job ID and can save this ID for future query. Our server stores this job ID for one month.

Conclusion

In this study, we have developed an efficient approach termed as pbPUP for improving the prediction of protein pupylation sites. Benchmarking experiments based on cross-validation and independent tests have shown that pbPUP provides a competitive performance compared with several existing methods. We have also shown that the proposed sequence encoding scheme pbCKSAAP outperformed the conventional CKSAAP encoding scheme. Our analysis suggests that the pbCKSAAP encoding is able to capture important sequence evolutionary information, which plays an important role for the performance improvement. Moreover, we performed feature selection experiments to characterize the contributive features and facilitate better understanding and interpretation of our prediction model. Computational analyses also demonstrate that our proposed method can be used as a powerful tool for understanding the mechanism of protein pupylation. Finally, we have also implemented a user-friendly web server for the research community, which is freely available at http://protein.cau.edu.cn/pbPUP/.

Supporting Information

S1 Table. List of training data.
(XLSX)

S2 Table. List of independent data.
(XLSX)

S3 Table. The prediction performance of pbPUP and other existing predictors on the independent test dataset after the removal of protein-level sequence redundancy.
(DOC)

S4 Table. The prediction performance of pbPUP and other existing predictors on the independent test dataset after the removal of peptide-level sequence redundancy.
(DOC)

S5 Table. The most important features and their corresponding feature selection scores.
(DOCX)

S1 Fig. AUC values for different window sizes based on 10-fold cross-validation tests.
(DOCX)

S2 Fig. ROC curves after the application of different sequence redundancy removal methods (at either protein- or peptide-level), according to 10-fold cross-validation tests.
(DOCX)

S3 Fig. Average PSSM values (APV) at different positions of positive and negative fragments. P-values were calculated using the one-tailed t-test. *, $P<0.01$.
(DOCX)

S4 Fig. The distribution of matched pupylated peptides of the selected amino acid pair features in the testing dataset. The matched pupylated peptides of pbCKSAAP-specific features and CKSAAP-specific features were considered, respectively.
(DOCX)

S5 Fig. (A) The input page of the pbPUP server. Users can paste the query sequence into the text box and submit the prediction job; (B) The output page of the pbPUP server, which provides an example output of the prediction result for the query sequence.
(DOCX)

Acknowledgments

The authors would like to thank Dr. Zhen Chen, Xuhan Liu, and Hong Li at the China Agricultural University for helpful discussions of this work.

Author Contributions

Conceived and designed the experiments: ZZ YZ MMH. Performed the experiments: MMH XL. Analyzed the data: MMH YZ XL JL JS ZZ. Wrote the paper: MMH YZ JL JS ZZ.

References

1. Liao S, Shang Q, Zhang X, Zhang J, Xu C, Tu X. Pup, a prokaryotic ubiquitin-like protein, is an intrinsically disordered protein. The Biochemical journal. 2009; 422(2):207–15. Epub 2009/07/08. doi: 10.1042/BJ20090738 PMID: 19580545.
2. Chen X, Solomon WC, Kang Y, Cerda-Maira F, Darwin KH, Walters KJ. Prokaryotic ubiquitin-like protein pup is intrinsically disordered. Journal of molecular biology. 2009; 392(1):208–17. Epub 2009/07/18. doi: 10.1016/j.jmb.2009.07.018 PMID: 19607839; PubMed Central PMCID: PMC2734869.
3. Burns KE, Darwin KH. Pupylation: A signal for proteasomal degradation in *Mycobacterium tuberculosis*. Sub-cellular biochemistry. 2010; 54:149–57. doi: 10.1007/978-1-4419-6676-6_12 PMID: 21222280.
4. DeMartino GN. PUPylation: something old, something new, something borrowed, something Glu. Trends in biochemical sciences. 2009; 34(4):155–8. doi: 10.1016/j.tibs.2008.12.005 PMID: 19282181.
5. Pearce MJ, Mintseris J, Ferreyra J, Gygi SP, Darwin KH. Ubiquitin-like protein involved in the proteasome pathway of *Mycobacterium tuberculosis*. Science. 2008; 322(5904):1104–7. Epub 2008/10/04. doi: 10.1126/science.1163885 PMID: 18832610; PubMed Central PMCID: PMC2698935.
6. Etlinger JD, Goldberg AL. A soluble ATP-dependent proteolytic system responsible for the degradation of abnormal proteins in reticulocytes. Proceedings of the National Academy of Sciences of the United States of America. 1977; 74(1):54–8. PMID: 264694; PubMed Central PMCID: PMC393195.
7. Burns KE, Darwin KH. Pupylation versus ubiquitylation: tagging for proteasome-dependent degradation. Cellular microbiology. 2010; 12(4):424–31. doi: 10.1111/j.1462-5822.2010.01447.x PMID: 20109157; PubMed Central PMCID: PMC3647454.
8. Ikeda F, Dikic I. Atypical ubiquitin chains: new molecular signals. 'Protein Modifications: Beyond the Usual Suspects' review series. EMBO reports. 2008; 9(6):536–42. doi: 10.1038/embor.2008.93 PMID: 18516089; PubMed Central PMCID: PMC2427391.
9. Iyer LM, Burroughs AM, Aravind L. Unraveling the biochemistry and provenance of pupylation: a prokaryotic analog of ubiquitination. Biology direct. 2008; 3:45. Epub 2008/11/05. doi: 10.1186/1745-6150-3-45 PMID: 18980670; PubMed Central PMCID: PMC2588565.
10. Kraut DA, Matouschek A. Pup grows up: in vitro characterization of the degradation of pupylated proteins. The EMBO journal. 2010; 29(7):1163–4. doi: 10.1038/emboj.2010.40 PMID: 20372178; PubMed Central PMCID: PMC2857470.
11. Imkamp F, Striebel F, Sutter M, Ozcelik D, Zimmermann N, Sander P, et al. Dop functions as a depupylase in the prokaryotic ubiquitin-like modification pathway. EMBO reports. 2010; 11(10):791–7. doi: 10.1038/embor.2010.119 PMID: 20798673; PubMed Central PMCID: PMC2948181.
12. Sutter M, Striebel F, Damberger FF, Allain FH, Weber-Ban E. A distinct structural region of the prokaryotic ubiquitin-like protein (Pup) is recognized by the N-terminal domain of the proteasomal ATPase

Mpa. FEBS letters. 2009; 583(19):3151–7. Epub 2009/09/19. doi: 10.1016/j.febslet.2009.09.020 PMID: 19761766.

13. Goldberg AL. Nobel committee tags ubiquitin for distinction. Neuron. 2005; 45(3):339–44. Epub 2005/02/08. doi: 10.1016/j.neuron.2005.01.019 PMID: 15694320.

14. Yun HY, Tamura N, Tamura T. *Rhodococcus* prokaryotic ubiquitin-like protein (Pup) is degraded by deaminase of pup (Dop). Bioscience, biotechnology, and biochemistry. 2012; 76(10):1959–66. Epub 2012/10/11. doi: 10.1271/bbb.120458 PMID: 23047115.

15. Striebel F, Imkamp F, Sutter M, Steiner M, Mamedov A, Weber-Ban E. Bacterial ubiquitin-like modifier Pup is deamidated and conjugated to substrates by distinct but homologous enzymes. Nature structural & molecular biology. 2009; 16(6):647–51. Epub 2009/05/19. doi: 10.1038/nsmb.1597 PMID: 19448618.

16. Sutter M, Damberger FF, Imkamp F, Allain FH, Weber-Ban E. Prokaryotic ubiquitin-like protein (Pup) is coupled to substrates via the side chain of its C-terminal glutamate. Journal of the American Chemical Society. 2010; 132(16):5610–2. Epub 2010/04/02. doi: 10.1021/ja910546x PMID: 20355727.

17. Guth E, Thommen M, Weber-Ban E. Mycobacterial ubiquitin-like protein ligase PafA follows a two-step reaction pathway with a phosphorylated pup intermediate. The Journal of biological chemistry. 2011; 286(6):4412–9. Epub 2010/11/18. doi: 10.1074/jbc.M110.189282 PMID: 21081505; PubMed Central PMCID: PMC3039397.

18. Striebel F, Imkamp F, Ozcelik D, Weber-Ban E. Pupylation as a signal for proteasomal degradation in bacteria. Biochimica et biophysica acta. 2014; 1843(1):103–13. Epub 2013/04/06. doi: 10.1016/j.bbamcr.2013.03.022 PMID: 23557784.

19. Barandun J, Delley CL, Weber-Ban E. The pupylation pathway and its role in mycobacteria. BMC biology. 2012; 10:95. Epub 2012/12/04. doi: 10.1186/1741-7007-10-95 PMID: 23198822; PubMed Central PMCID: PMC3511204.

20. Elharar Y, Roth Z, Hermelin I, Moon A, Peretz G, Shenkerman Y, et al. Survival of mycobacteria depends on proteasome-mediated amino acid recycling under nutrient limitation. The EMBO journal. 2014; 33(16):1802–14. Epub 2014/07/06. doi: 10.15252/embj.201387076 PMID: 24986881.

21. Darwin KH. Prokaryotic ubiquitin-like protein (Pup), proteasomes and pathogenesis. Nature reviews Microbiology. 2009; 7(7):485–91. Epub 2009/06/02. doi: 10.1038/nrmicro2148 PMID: 19483713; PubMed Central PMCID: PMC3662484.

22. Salgame P. PUPylation provides the punch as *Mycobacterium tuberculosis* battles the host macrophage. Cell host & microbe. 2008; 4(5):415–6. Epub 2008/11/11. doi: 10.1016/j.chom.2008.10.009 PMID: 18996341; PubMed Central PMCID: PMC3202434.

23. Cerda-Maira FA, McAllister F, Bode NJ, Burns KE, Gygi SP, Darwin KH. Reconstitution of the *Mycobacterium tuberculosis* pupylation pathway in *Escherichia coli*. EMBO reports. 2011; 12(8):863–70. Epub 2011/07/09. doi: 10.1038/embor.2011.109 PMID: 21738222; PubMed Central PMCID: PMC3147258.

24. Festa RA, McAllister F, Pearce MJ, Mintseris J, Burns KE, Gygi SP, et al. Prokaryotic ubiquitin-like protein (Pup) proteome of *Mycobacterium tuberculosis*. PloS one. 2010; 5(1):e8589. Epub 2010/01/13. doi: 10.1371/journal.pone.0008589 PMID: 20066036; PubMed Central PMCID: PMC2797603.

25. Kuberl A, Franzel B, Eggeling L, Polen T, Wolters DA, Bott M. Pupylated proteins in *Corynebacterium glutamicum* revealed by MudPIT analysis. Proteomics. 2014; 14(12):1531–42. Epub 2014/04/17. doi: 10.1002/pmic.201300531 PMID: 24737727.

26. Watrous J, Burns K, Liu WT, Patel A, Hook V, Bafna V, et al. Expansion of the mycobacterial "PUPylome". Molecular bioSystems. 2010; 6(2):376–85. Epub 2010/01/23. doi: 10.1039/b916104j PMID: 20094657; PubMed Central PMCID: PMC2846642.

27. Poulsen C, Akhter Y, Jeon AH, Schmitt-Ulms G, Meyer HE, Stefanski A, et al. Proteome-wide identification of mycobacterial pupylation targets. Molecular systems biology. 2010; 6:386. Epub 2010/07/16. doi: 10.1038/msb.2010.39 PMID: 20631680; PubMed Central PMCID: PMC2925521.

28. Zhao X, Dai J, Ning Q, Ma Z, Yin M, Sun P. Position-specific analysis and prediction of protein pupylation sites based on multiple features. BioMed research international. 2013; 2013:109549. Epub 2013/09/26. doi: 10.1155/2013/109549 PMID: 24066285; PubMed Central PMCID: PMC3770009.

29. Tung CW. Prediction of pupylation sites using the composition of *k*-spaced amino acid pairs. Journal of theoretical biology. 2013; 336:11–7. Epub 2013/07/23. doi: 10.1016/j.jtbi.2013.07.009 PMID: 23871866.

30. Liu Z, Ma Q, Cao J, Gao X, Ren J, Xue Y. GPS-PUP: computational prediction of pupylation sites in prokaryotic proteins. Molecular bioSystems. 2011; 7(10):2737–40. Epub 2011/08/19. doi: 10.1039/c1mb05217a PMID: 21850344.

31. Chen X, Qiu JD, Shi SP, Suo SB, Liang RP. Systematic analysis and prediction of pupylation sites in prokaryotic proteins. PloS one. 2013; 8(9):e74002. Epub 2013/09/11. doi: 10.1371/journal.pone.0074002 PMID: 24019945; PubMed Central PMCID: PMC3760804.

32. Tung CW. PupDB: a database of pupylated proteins. BMC bioinformatics. 2012; 13:40. Epub 2012/03/20. doi: 10.1186/1471-2105-13-40 PMID: 22424087; PubMed Central PMCID: PMC3314583.

33. Chen K, Kurgan LA, Ruan J. Prediction of flexible/rigid regions from protein sequences using *k*-spaced amino acid pairs. BMC structural biology. 2007; 7:25. Epub 2007/04/18. doi: 10.1186/1472-6807-7-25 PMID: 17437643; PubMed Central PMCID: PMC1863424.

34. Chen YZ, Tang YR, Sheng ZY, Zhang Z. Prediction of mucin-type O-glycosylation sites in mammalian proteins using the composition of *k*-spaced amino acid pairs. BMC bioinformatics. 2008; 9:101. Epub 2008/02/20. doi: 10.1186/1471-2105-9-101 PMID: 18282281; PubMed Central PMCID: PMC2335299.

35. Chen Z, Chen YZ, Wang XF, Wang C, Yan RX, Zhang Z. Prediction of ubiquitination sites by using the composition of *k*-spaced amino acid pairs. PloS one. 2011; 6(7):e22930. Epub 2011/08/11. doi: 10.1371/journal.pone.0022930 PMID: 21829559; PubMed Central PMCID: PMC3146527.

36. Wang XB, Wu LY, Wang YC, Deng NY. Prediction of palmitoylation sites using the composition of *k*-spaced amino acid pairs. Protein engineering, design & selection: PEDS. 2009; 22(11):707–12. Epub 2009/09/29. doi: 10.1093/protein/gzp055 PMID: 19783671.

37. Zhang W, Xu X, Yin M, Luo N, Zhang J, Wang J. Prediction of methylation sites using the composition of *K*-spaced amino acid pairs. Protein and peptide letters. 2013; 20(8):911–7. Epub 2013/01/02. PMID: 23276225.

38. Zhao X, Zhang W, Xu X, Ma Z, Yin M. Prediction of protein phosphorylation sites by using the composition of *k*-spaced amino acid pairs. PloS one. 2012; 7(10):e46302. Epub 2012/10/31. doi: 10.1371/journal.pone.0046302 PMID: 23110047; PubMed Central PMCID: PMC3478286.

39. Chen K, Jiang Y, Du L, Kurgan L. Prediction of integral membrane protein type by collocated hydrophobic amino acid pairs. Journal of computational chemistry. 2009; 30(1):163–72. Epub 2008/06/21. doi: 10.1002/jcc.21053 PMID: 18567007.

40. Dong X, Zhang YJ, Zhang Z. Using weakly conserved motifs hidden in secretion signals to identify type-III effectors from bacterial pathogen genomes. PloS one. 2013; 8(2):e56632. Epub 2013/02/26. doi: 10.1371/journal.pone.0056632 PMID: 23437191; PubMed Central PMCID: PMC3577856.

41. Weinert BT, Wagner SA, Horn H, Henriksen P, Liu WR, Olsen JV, et al. Proteome-wide mapping of the *Drosophila* acetylome demonstrates a high degree of conservation of lysine acetylation. Science signaling. 2011; 4(183):ra48. Epub 2011/07/28. doi: 10.1126/scisignal.2001902 PMID: 21791702.

42. Zhou Y, Liu S, Song J, Zhang Z. Structural propensities of human ubiquitination sites: accessibility, centrality and local conformation. PloS one. 2013; 8(12):e83167. Epub 2013/12/19. doi: 10.1371/journal.pone.0083167 PMID: 24349449; PubMed Central PMCID: PMC3859641.

43. Yan RX, Si JN, Wang C, Zhang Z. DescFold: a web server for protein fold recognition. BMC bioinformatics. 2009; 10:416. Epub 2009/12/17. doi: 10.1186/1471-2105-10-416 PMID: 20003426; PubMed Central PMCID: PMC2803855.

44. Song J, Tan H, Shen H, Mahmood K, Boyd SE, Webb GI, et al. Cascleave: towards more accurate prediction of caspase substrate cleavage sites. Bioinformatics. 2010; 26(6):752–60. Epub 2010/02/05. doi: 10.1093/bioinformatics/btq043 PMID: 20130033.

45. Si JN, Yan RX, Wang C, Zhang Z, Su XD. TIM-Finder: a new method for identifying TIM-barrel proteins. BMC structural biology. 2009; 9:73. Epub 2009/12/17. doi: 10.1186/1472-6807-9-73 PMID: 20003393; PubMed Central PMCID: PMC2803183.

46. Chen Z, Zhou Y, Zhang Z, Song J. Towards more accurate prediction of ubiquitination sites: a comprehensive review of current methods, tools and features. Briefings in bioinformatics. 2014. Epub 2014/09/13. doi: 10.1093/bib/bbu031 PMID: 25212598.

47. Chen Z, Zhou Y, Song J, Zhang Z. hCKSAAP_UbSite: improved prediction of human ubiquitination sites by exploiting amino acid pattern and properties. Biochimica et biophysica acta. 2013; 1834(8):1461–7. Epub 2013/04/23. doi: 10.1016/j.bbapap.2013.04.006 PMID: 23603789.

48. Wang M, Zhao XM, Tan H, Akutsu T, Whisstock JC, Song J. Cascleave 2.0, a new approach for predicting caspase and granzyme cleavage targets. Bioinformatics. 2014; 30(1):71–80. Epub 2013/10/24. doi: 10.1093/bioinformatics/btt603 PMID: 24149049.

49. Chang CC, Lin CJ. LIBSVM: A Library for Support Vector Machines. ACM transactions on intelligent systems and technology2011; 2(3). doi: Artn 27 doi: 10.1145/1961189.1961199 PMID: ISI:000208617000010.

50. Gribskov M, Robinson NL. Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. Computers & chemistry. 1996; 20(1):25–33. doi: 10.1016/S0097-8485(96)80004-0 PMID: ISI:A1996UA22000003.

51. Centor RM. Signal detectability—the use of ROC curves and their analyses. Medical decision making. 1991; 11(2):102–6. doi: 10.1177/0272989x9101100205 PMID: ISI:A1991FF53200005.

52. Provost F. Machine learning from imbalanced data sets 101. AAAI Workshop on learning from imbalanced data set. 2000:1–3.

53. Lin C-J CY-W. Combining SVMs with various feature selection strategies. NIPS 2003 feature selection challenge. 2003:1–10.

54. Shi SP, Qiu JD, Sun XY, Suo SB, Huang SY, Liang RP. PLMLA: prediction of lysine methylation and lysine acetylation by combining multiple features. Molecular bioSystems. 2012; 8(5):1520–7. Epub 2012/03/10. doi: 10.1039/c2mb05502c PMID: 22402705.

55. Li Y, Wang M, Wang H, Tan H, Zhang Z, Webb GI, et al. Accurate *in silico* identification of species-specific acetylation sites by integrating protein sequence-derived and functional features. Scientific reports. 2014; 4:5765. Epub 2014/07/22. doi: 10.1038/srep05765 PMID: 25042424; PubMed Central PMCID: PMC4104576.

56. Vacic V, Iakoucheva LM, Radivojac P. Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. Bioinformatics. 2006; 22(12):1536–7. Epub 2006/04/25. doi: 10.1093/bioinformatics/btl151 PMID: 16632492.