molecules

MDPI

Article

Recognition of Protein Pupylation Sites by Adopting Resampling Approach

Tao Li [1,2], Yan Chen [1], Taoying Li [1,*] and Cangzhi Jia [3]

1 School of Transportation Management, Dalian Maritime University, Dalian 116026, China; litao@wti.ac.cn (T.L.); chenyan@dlmu.edu.cn (Y.C.)
2 China Waterborne Transport Research Institute, Beijing 100088, China
3 College of Science, Dalian Maritime University, Dalian 116026, China; cangzhijia@dlmu.edu.cn
* Correspondence: litaoying@dlmu.edu.cn; Tel.: +86-155-6680-2152

Received: 13 October 2018; Accepted: 22 November 2018; Published: 27 November 2018

Abstract: With the in-depth study of posttranslational modification sites, protein ubiquitination has become the key problem to study the molecular mechanism of posttranslational modification. Pupylation is a widely used process in which a prokaryotic ubiquitin-like protein (Pup) is attached to a substrate through a series of biochemical reactions. However, the experimental methods of identifying pupylation sites is often time-consuming and laborious. This study aims to propose an improved approach for predicting pupylation sites. Firstly, the Pearson correlation coefficient was used to reflect the correlation among different amino acid pairs calculated by the frequency of each amino acid. Then according to a descending ranked order, the multiple types of features were filtered separately by values of Pearson correlation coefficient. Thirdly, to get a qualified balanced dataset, the K-means principal component analysis (KPCA) oversampling technique was employed to synthesize new positive samples and Fuzzy undersampling method was employed to reduce the number of negative samples. Finally, the performance of our method was verified by means of jackknife and a 10-fold cross-validation test. The average results of 10-fold cross-validation showed that the sensitivity (Sn) was 90.53%, specificity (Sp) was 99.8%, accuracy (Acc) was 95.09%, and Matthews Correlation Coefficient (MCC) was 0.91. Moreover, an independent test dataset was used to further measure its performance, and the prediction results achieved the Acc of 83.75%, MCC of 0.49, which was superior to previous predictors. The better performance and stability of our proposed method showed it is an effective way to predict pupylation sites.

Keywords: fuzzy undersampling; machine learning; principal component analysis; protein pupylation; sequence information

1. Introduction

Pupylation is a prokaryotic analog of ubiquitination whose prokaryotic ubiquitin-like protein (Pup) separates intracellular proteins under the action of enzymes and specifically modifies the target protein [1,2]. Its proteasome-independent functions help in the regulation of DNA repair mechanisms, particularly recent advances on its regulatory role for careful series of functions that gives rise to a therapeutic solution in cancer chemotherapy [3–5]. Pup is an identified posttranslational small modifier in prokaryotes [1,2] that usually attaches to the substrate lysine via isopeptide bonds, and this process is called pupylation. Although the functions of pupylation and ubiquitylation are similar, the enzymologies involved in these processes are different [1,2,6]. Since experimental methods are laborious, it is essential to improve the current computational methodologies to provide direction for further research on intriguing research questions. An example of this is the works on understanding the stability of ERCC1 DNA repair protein, a biomarker of several advanced

cancer diseases. This protein functions in multiple DNA repair pathways and certain mutations in this protein and its partner had drastic consequences for the protein complex stability [3–9]. Functional and structural studies have determined the significance of the complex integrity and ubiquitination/deubiquitination events for controlling the function of the protein during DNA damage response [1,2,6–8]. These findings persuade new therapeutic solution in cancer chemotherapy. Thus, it is important to devise accurate computational methodologies for scientists to ultimately predict the probability of different regulatory pathways that control protein function and disease severities [6,8,9]. Several proteomics methods have been proposed for identifying pupylated proteins and pupylation sites [10–17]. However, pupylation sites of many experimentally identified pupylated proteins are still unknown. Accurate identification of pupylation sites is an essential first step to better understand the mechanism underlying protein pupylation. A number of large-scale proteomics techniques have been used for predicting pupylation sites, which are time-consuming and laborious. Therefore, a series of other effective and accurate prediction methods have been proposed for predicting pupylation sites.

The Group-based Prediction System was used by Lin to design the predictor of GPS-PUP for predicting the pupylation sites [11]. The predictor iPup [12] was designed by combining the composition of k-spaced amino acid pair (CKSAAP) feature and support vector machines (SVMs). Zhao et al. [13] created a predictive model with five features and also adopted feature selection methods to find the optimal feature set. Chen et al. [14] proposed a model PupPred, which was also an SVM-based predictor. It used amino acid pair features to encode lysine central peptides and combined a series of features to improve its predictive performance. Hasan et al. [15] constructed a predictor called pbPUP that used the profile-based composition of k-spaced amino acid pairs (pbCKSAAP) to represent the sequence information around the pupylation site. Jiang et al. [16] created a predictor called PUL-PUP that combined the positive-unlabeled learning technique with CKSAAP to predict pupylation sites. A structured and searchable database PupDB was used for integrating information of pupylated proteins and sites, protein structures, and functional annotations for the management and analysis of pupylation sites due to a large number of newly identified pupylated proteins and sites [16]. Recently, Nan et al. [17] proposed the enhanced positive-unlabeled learning algorithm for predicting pupylation sites.

In this study, the Pearson correlation coefficient was used to evaluate the relevance among amino acid pairs based on amino acid composition information. Then, according to the descending order, the amino acid pair features were added by a step of 20 to get the best prediction results by means of jackknife test. Then the selected 320 amino acid pairs were combined with TOP-n-gram [18], adapted normal distribution bi-profile Bayes (ANBPB) [19], and parallel correlation pseudo amino acid composition (PC-PseAAC) [20] to construct a multiple feature vector for the query protein peptide sequence. Because of the imbalance between the number of positive and negative samples (183:2258), the K-means principal component analysis (KPCA) oversampling technique, firstly proposed by Jia and Zuo [21], was applied on the positive training dataset for oversampling, and the synthetic samples were added to the original positive training dataset as a new positive training dataset. The fuzzy undersampling (FUS) method [21,22] was applied to the negative dataset to reduce noise negative samples, and the selected negative samples were used as a new negative training dataset. At last, the 377 pupylated sites (positive samples) and 365 non-pupylated lysine sites (negative samples) were used to train and test our model. Moreover, the performance of our method was verified by means of jackknife, 10-fold cross-validation test and an independent dataset. When compared with other existing predictors, all of the results showed better performance and stability, proving as an effective way to predict pupylation sites.

2. Results and Discussion

2.1. Pearson Correlation Coefficient for Feature Selection

The Pearson Correlation Coefficient [23,24] is a measurement for the linear relationship among distance variables. If a positive linear correlation exists between two variables, their metric approaches 1. The Pearson Correlation Coefficient was used to find out the most closely related amino acids. First, extracting amino acid composition (AAC) was used for feature extraction on the training dataset. Second, the Pearson Correlation Coefficient was employed to calculate the correlation coefficient among amino acids based on AAC values, and its values were sorted in a descending order. Finally, the amino acid pair composition (AAPC) was selected according to the order. The value of Pearson Correlation Coefficient was 1 for the amino acid pair AA, CC, ..., YY, and so these 20 amino acid pairs of the same amino acid were reserved to combine with others. The prediction on the jackknife test on the combination of 20, 40, ..., 380, and 400 features was performed by a step of 20, and the detailed results are shown in Table S1. The best results were obtained when the first AAPC (320) features were selected; sensitivity (Sn) was 81.42%, specificity (Sp) was 76.44%, accuracy (Acc) was 78.10%, and Matthews Correlation Coefficient (MCC) was 0.55. However, the prediction results were not monotonous; this might be due to the same parameter used for different dimension features.

2.2. Combination of KPCA and FUS for Training Set Balancing

As described in Section 2.1, the training dataset comprised 183 positive samples and 2258 negative samples which is the same as [15,17] to guarantee the contrast of the experiment.

The ratio of positive and negative samples was 1:12, leading to biased results and inaccurate data. The KPCA and fuzzy undersampling (FUS) oversampling and undersampling methods were first proposed by Jia and Zuo [21]. They were effective in identifying protein O-GlcNAcylation sites. Therefore, in this study, KPCA and FUS were also employed to solve the imbalance between positive and negative training datasets. First, FUS was used to remove the redundant negative samples, and the remaining 365 nonpupylation protein peptides were used as the new negative training dataset. Then, KPCA was applied on the positive samples to divide them into three clusters (k = 3). The experiment was repeated several times. The results of clustering were added to the original positive samples so that the ratio of the positive samples and the negative samples was approximately 1:1. The following results were obtained based on KPCA and FUS.

The performance illustrated that the results obtained from the original unbalanced training dataset were biased toward the larger number of classes. For example, on the 320-dimensional selected feature vector, Sn was 0 and Sp was 100%; it was the worst performance for the prediction model. However, after using KPCA and FUS on the training dataset, Sn was 81.42% and Sp was 76.44% shown in Table 1, indicating that the unbalanced data had a great influence on the experimental results.

Table 1. Comparison of original imbalance dataset and balanced dataset.

Method	Sn (%)	Sp (%)	Acc (%)	MCC
Without resampling	0	100	92.50	NaN
KPCA oversampling	25.54	99.87	92.67	0.33
KPCA oversampling and FUS undersampling	81.42	76.44	78.10	0.55

2.3. Predictive Performance Improvement

A variety of feature extraction methods can be found on the Web server (http://bioinformatics.hitsz.edu.cn/Pse-in-One2.0/) [25]. After several trials, TOP-n-gram, ANBPB, and PC-PseAAC feature extraction methods were used to combine with 320-dimensional features extracted by AAPC, and the results are shown in Table 2.

Table 2. Predictive performance of TOP-n-gram, adapted normal distribution bi-profile Bayes (ANBPB), and parallel correlation pseudo amino acid composition (PC-PseAAC) using the jackknife test.

Feature	Sn (%)	Sp (%)	Acc (%)	MCC
AAPC(320)	81.42	76.44	78.1	0.55
AAPC(320) + TOP-n-gram	80.33	100	93.43	0.86
AAPC(320) + ANBPB	64.48	100	88.14	0.74
AAPC(320) + PC-PseAAC	84.15	72.60	76.46	0.54
AAPC(320) + TOP-n-gram + ANBPB	70.49	98.36	89.05	0.75
AAPC(320) + TOP-n-gram + PC-PseAAC	87.43	75.34	79.38	0.59
AAPC(320) + ANBPB + PC-PseAAC	75.41	100	91.79	0.82
AAPC(320) + TOP-n-gram + ANBPB + PC-PseAAC	94.54	100	98.18	0.96

The Pearson correlation coefficient was used to filter out the AAPC (320) dimension features from the 400-dimensional features so as to remove the redundant information. Although the results were improved, the value of MCC was slightly lower. AAPC (320) was combined with other features to increase the value of MCC. When AAPC (320) was combined with TOP-n-gram and ANBPB, Sn, Acc, and MCC improved greatly. Especially when AAPC (320) was combined with TOP-n-gram, MCC increased by 0.3037, but Sn slightly reduced. When AAPC (320) was combined with PC-PseAAC, Sn increased by 2.73%. Combining AAPC (320) with TOP-n-gram and PC-PseAAC, Sn increased by 6.01%. After several experiments, AAPC (320) was combined with TOP-n-gram, ANBPB, and PC-PseAAC, with the MCC of 0.96, which was higher by 0.41 than the previous value of AAPC (320); the accuracy was 98.18%. Furthermore, Sn and Sp increased by 13.12% and 23.56% compared with AAPC (320). The results indicated that the predictor achieved better prediction performance.

2.4. Comparison between the Proposed Method and Other Prediction Methods

The proposed method was compared with other methods by tenfold cross-validation, including EPuL [17], PUL-PUP [16], PSoL [26], and SVM balance [15] on the training dataset to evaluate the effectiveness of the proposed method for pupylation site prediction. Table 3 presents the comparison among EPuL, PUL-PUP, PSoL, and SVM balance.

Table 3. Comparison of the proposed method with EPuL, PUL-PUP, PSoL, and support vector machine (SVM) balance on 10-fold cross-validation test.

Method	Sn (%)	Sp (%)	Acc (%)	MCC	AUC
Proposed method	90.53	99.8	95.09	0.91	0.96
* EPuL	84.21	95.45	90.24	0.81	0.93
* PUL-PUP	82.24	91.57	88.92	0.74	0.92
* PSoL	67.5	73.6	70.55	0.42	0.8
* SVM balance	76.71	63.65	69.88	0.4	0.77

* The results of the comparison are from EPuL [17].

The SVM balance method can solve the imbalance of datasets, but its negative samples obtained from unannotated lysine sites are inaccurate. PUL-PUP and PSoL use unreliable negative datasets, whose features are not optimal. EPuL uses a series of options to obtain the reliably negative datasets; however, its accuracy is low because it does not integrate other features and the imbalance dataset is selected randomly. Figure 1 shows that the proposed method achieved a higher accuracy compared with EPuL for running 10 times of tenfold cross-validation results, with Sn of 6.32% and Sp of 4.35%. Furthermore, the Acc and MCC of the proposed method were 95.09% and 0.9063, respectively, which were 4.85% and 0.0963 higher than those of EPuL. For the EPuL web-server is unavailable,

we listed the detailed 10-fold cross-validation results of our method, PUL-PUP, PSoL, and SVM in Tables S2–S5.

2.5. Performance on the Independent Test Dataset

In this study, the proposed method achieved a better performance, indicating that the Pearson correlation coefficient, KPAC, and FUS were effective in predicting the pupylation sites. An independent test dataset containing 20 proteins, which included 29 experimentally validated pupylation sites and 408 nonannotated pupylation sites, was selected to further verify the effectiveness of the proposed method. Although the positive and negative datasets of the independent test dataset were imbalanced, it was obtained from the real proteins and reflected the true distribution of pupylated sites and nonpupylated sites. A comparison of results with those of the existing methods, including the EPuL, PUL-PUP, PSoL, and SVM balance, is listed in Figure 1. The prediction results of PUL-PUP, PSoL, and SVM-balance are directly from EPuL [17]. we listed results of our method, PUL-PUP, PSoL, and SVM balance on the independent dataset in Table S6–S9.

The overall accuracy of the proposed method was the best among the five models. Especially, Sn reached 100% and Acc was 24% higher than that of EPuL.

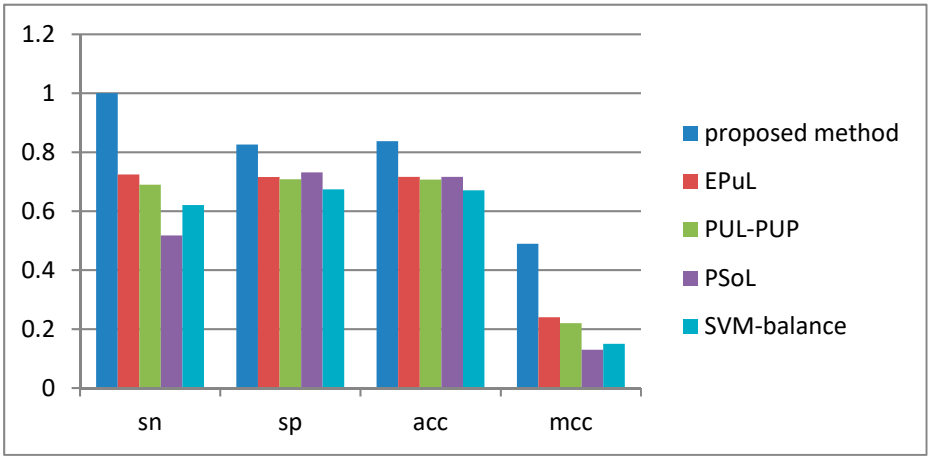

Figure 1. Comparison of performance between our model, EPuL, PUL-PUP, PSoL, and SVM balance on an independent test dataset.

3. Materials and Methods

This study involved the following steps: (1) extracting amino acid composition (AAC); (2) calculating Pearson Correlation Coefficient and sorting the values; (3) selecting the top n amino acid pair composition; (4) combining with other features; (5) balancing the training datasets; and (6) verifying the model. The conceptual diagram of the pupylation site prediction model is shown in Figure 2.

Figure 2. Conceptual diagram of the pupylation site prediction model.

3.1. Formatting of Mathematical Components

The training and test datasets constructed by Tung [12] and lately used by Nan et al. [17], were also adopted in this study. The training dataset consisted of 183 experimentally validated pupylation sites and 2258 artificially generated nonannotated lysine sites from 162 proteins of Mycobacterium smegmatis (*M. smegmatis*), Mycobacterium tuberculosis (*M. tuberculosis*) and Escherichia coli (*E. coli*). The independent test dataset contained 29 experimentally verified lysine pupylation sites and 408 nonannotated lysine sites from 20 proteins. As the independent test dataset was highly unbalanced, it reflected the real effects of different methods [10–17]. Similar to the previous findings, the length of each protein peptide was 21 in the training and test datasets. The training and test datasets are available in the supplementary material.

3.2. Information from the Protein Peptide Sequence

3.2.1. Amino Acid Composition (AAC)

AAC [27] is a widely used method for feature extraction, varied protein sequence analysis, and prediction. It was used to calculate the content of amino acids in the amino acid fragment. For a given peptide fragment containing 20 natural amino acids and pseudo amino acids "X," only 20 natural amino acids were used to construct the 20-dimensional feature vector of AAC for encoding because the frequency of pseudo amino acid "X" always approached zero or equaled zero. AAC can be defined as follows:

$$\mathrm{AAC} = [f_1, f_2, \ldots, f_{20}] \quad (1)$$

where f_i where f_i represents the frequency of the occurrence of the *ith* amino acid in the 20 natural amino acids $\{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$ and expressed as follows:

$$f_i = \frac{N_i}{N_{\text{total}}} \tag{2}$$

where N_i indicates the number of amino acid *ith* in the peptide fragment, and N_{total} indicates the total length of the peptide fragment. The length of the amino acid peptide was 20 in the study.

3.2.2. Amino Acid Pair Composition

Amino acid pair composition (AAPC) [28] is generally used to obtain the correlation between one amino acid and other amino acids in a protein sequence. In this study, it was used to calculate the frequency of occurrence of amino acids in pupylation fragments. The AAPC feature vector is defined as follows:

$$\text{AAPC} = [x_1, x_2, \ldots, x_i, \ldots, x_{400}]^{\text{T}} \ (0 < i \leq 400) \tag{3}$$

where x_i indicates the occurrence correlation of the *ith* amino acid pairs with other amino acid pairs (AA, AC, ..., YY).

$$x_i = \frac{n_{ij}}{n_{total}} \tag{4}$$

where n_{ij} indicates the number of amino acid pair *ijth* in the peptide fragment, and n_{total} indicates the total number of amino acid pair. The length of the amino acid pair was 400 in the study.

3.2.3. Adapted Normal Distribution bi-profile Bayes

ANBPB [19] is a combination of bi-profile Bayes (BPB) [29,30] and the standard normal distribution, which uses a probability vector to encode the peptide fragment.

$$P_j = [p_1, p_2, \ldots, p_n, p_{n+1}, \ldots, p_{2n}]^T \tag{5}$$

where $P_j(j = 1, 2, \ldots, n)$ is the posterior probability of each amino acid in the *jth* position of the positive training dataset; $P_j(j = n+1, n+2, \ldots, 2n)$ is the posterior probability of each amino acid in the *jth* position of the negative training dataset. The occurrence frequency of each amino acid at each position is encoded as random variables X_{ij} $(i = 1, 2, \ldots, 21;\ j = 1, 2, \ldots, 21)$, which are independent and identically distributed in the binomial distribution $b = (m,\ p)$, in this study; *m = 183/2258* is the number of samples in the positive/negative dataset, $p = 1/21$ is the probability of occurrence of each amino acid at each position. According to the Moivre–Laplace's theorem, if m is large enough, $\frac{X_{ij} - mp}{\sqrt{mp(1-p)}}$ approximately obeys the standard normal distribution $N(0, 1)$. If V_j is used to represent the standard deviation of the random variable X_{ij} $(i = 1, 2, \ldots, 21)$, the normalized variable of the random variable X_{ij} can be defined as follows:

$$X'_{ij} = \frac{X_{ij} - mp}{\sqrt{V_j}}. \tag{6}$$

Therefore, p_j $(i = 1, 2, \ldots n, n+1, \ldots, 2n)$ can be encoded by the adapted normal distribution as follows:

$$p_j = P(X \leq X_{ij}) = \varphi\left(X'_{ij}\right) \tag{7}$$

where the formula for $\varphi(x)$ is given by $\varphi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$.

3.2.4. TOP-n-gram

Top-n-gram [18], which includes the evolutionary information extracted from the frequency profiles, can be viewed as a novel profile-based building block of proteins. The multiple sequence alignments yielded by PSI-BLAST [31] are used to calculate the protein sequence frequency profiles, which are combined with the *n* most frequent amino acids in each amino acid frequency profile, and the frequency profiles are converted into Top-n-grams. The occurrence times of each TOP-n-gram are used to convert protein fragments into fixed feature vectors, and then the corresponding vectors are substituted into SVM. Several basic building blocks have been investigated as the words of "protein sequence language," including Ngrams [29,32], patterns [33], motifs [34], and binary profiles [18].

3.2.5. Parallel Correlation Pseudo Amino Acid Composition

Parallel correlation pseudo amino acid composition (PC-PseAAC) was also considered to obtain more order information for a fragment [20]. PC-PseAAC is an approach incorporating the contiguous local sequence-order information and the global sequence-order information into the feature vector of the protein sequence. Given a protein sequence P, the PC-PseAAC feature vector of P is defined as follows [25]:

$$P = [x_1\ x_2 x_3 \ldots x_{20} x_{21} \ldots x_{20+\lambda}\ ]^T \tag{8}$$

where

$$x_u = \begin{cases} \frac{f_u}{\sum_{i=1}^{20} f_i + w\sum_{j=1}^{\lambda} \Theta_j} (1 \le u \le 20) \\ \frac{w\Theta_{u-20}}{\sum_{i=1}^{20} f_i + w\sum_{j=1}^{\lambda} \Theta_j} (20+1 \le u \le 20+\lambda) \end{cases} \tag{9}$$

where *fi (i = 1,2, . . . ,20)* is the normalized occurrence frequency of the 20 amino acids in the protein sequence P; the parameter λ is an integer, representing the highest counted rank (or tier) of the correlation along a protein sequence; w is the weight factor ranging from 0 to 1; and θj *(j = 1, 2, · · · , λ)* is called the *j-tier* correlation factor reflecting the sequence-order correlation between all the *jth* most contiguous residues along a protein chain and defined as follows:

$$\Theta_\lambda = \frac{1}{L-\lambda} \sum_{i=1}^{L-\lambda} \Theta(R_i, R_{i+\lambda}) (0 < \lambda < 1) \tag{10}$$

where the correlation function is given by:

$$\Theta(R_i, R_j) = \frac{1}{3}\left\{ \left[H_1(R_j) - H_1(R_i)\right]^2 + \left[H_2(R_j) - H_2(R_i)\right]^2 + \left[M(R_j) - M(R_i)\right]^2 \right\} \tag{11}$$

where $H_1(R_i)$, $H_2(R_i)$, *and* $M(R_i)$ are the hydrophobicity value, hydrophilicity value, and side-chain mass of the amino acid R_i, respectively. Before substituting the values of hydrophobicity, hydrophilicity, and side-chain mass into [18], they should all be subjected to a standard conversion as described by the following equation [25]:

$$H_1(\mathrm{i}) = \frac{H_1^0(\mathrm{i}) - \sum_{\mathrm{i}=1}^{20} \frac{H_1^0(\mathrm{i})}{20}}{\sqrt{\frac{\sum_{\mathrm{i}=1}^{20} [H_1^0(\mathrm{i}) - \sum_{\mathrm{i}=1}^{20} \frac{H_1^0(\mathrm{i})}{20}]^2}{20}}} \tag{12}$$

$$H_2(\mathrm{i}) = \frac{H_2^0(\mathrm{i}) - \sum_{\mathrm{i}=1}^{20} \frac{H_2^0(\mathrm{i})}{20}}{\sqrt{\frac{\sum_{i=1}^{20} [H_2^0(\mathrm{i}) - \sum_{i=1}^{20} \frac{H_2^0(\mathrm{i})}{20}]^2}{20}}} \tag{13}$$

$$M(\mathrm{i}) = \frac{M^0(\mathrm{i}) - \sum_{i=1}^{20} \frac{M^0(\mathrm{i})}{20}}{\sqrt{\frac{\sum_{\mathrm{i}=1}^{20} [M^0(\mathrm{i}) - \sum_{\mathrm{i}=1}^{20} \frac{M^0(\mathrm{i})}{20}]^2}{20}}} \tag{14}$$

where $H_1^0(i)$ is the original hydrophobicity value of the *ith* amino acid; $H_2^0(i)$ is the corresponding original hydrophilicity value; and $M^0(i)$ *is* the mass of the *ith* amino acid side chains.

3.3. Pearson Correlation Coefficient

As a measure of the variability among variables, the correlation coefficient can indicate a certain correlation between two variables at the same time. The Pearson correlation coefficient [23,24] is used to judge whether two datasets are on a line and measure the linear relationship according to distance. For the Pearson correlation coefficient, if a positive linear correlation exists between two variables, their metric value approaches 1. Similarly, if the two variables have a negative linear correlation, their metric value approaches –1. Let *(X, Y)* be a two-dimensional random variable vector, and the Pearson correlation coefficient is defined as follows:

$$\rho = \frac{\mathrm{cov}\ (\ \mathrm{X},\ \mathrm{Y}\ )}{\sqrt{\mathrm{var(X)}}\ \sqrt{\mathrm{var(Y)}}} \tag{15}$$

Let $(\mathrm{X_i}, \mathrm{Y_i})$ (i = 1, 2 . . . n) be a random sample of $(\mathrm{X},\ \mathrm{Y})$, and another form of the Pearson correlation coefficient can be expressed as follows:

$$\rho = \frac{\sum_{\mathrm{i}=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\left[\sum_{\mathrm{i}=1}^{n}\left(X_i - \overline{X}\right)^2\ \sum_{\mathrm{i}=1}^{\mathrm{n}}\left(Y_i - \overline{Y}\right)^2\right]^{1/2}} \tag{16}$$

where $\overline{X}$ is the mean of X_i, and $\overline{Y}$ is the mean of Y_i, $(i = 1, 2 \ldots\ n)$.

In this study, AAC was used for feature extraction on the training dataset, and the values of Pearson Correlation Coefficients were sorted in a descending order. The positive and negative training datasets were extracted in AAPC. The feature selection was performed on the 400-dimensional feature vector, and the best results were obtained when the first AAPC (320) features were selected.

3.4. Processing of Positive and Negative Datasets

3.4.1. KPCA Oversampling Technique

The KPCA oversampling technique [21] is based on the K-means clustering algorithm and principal component analysis (PCA) [35,36]. First, the initial clustering center K was formed by randomly selecting class samples from the positive training dataset.

$$K(\mathrm{InitialCenter}(j)), j = 1, 2, \ldots n \tag{17}$$

Second, the Euclidean distance was used to divide the positive training samples into k clusters, defined as follows:

$$I_j = \left|\left|K(j)\ - K\left(\mathrm{InitialCenter}(n)\right)\ \right|\right|^2 \tag{18}$$

where $K(j)$ represents j original positive training samples, and I_j is the distance between j original positive training samples and n clusters.

Finally, the average values of all peptide fragments of the original positive training sample were calculated in each cluster and then clustering centers were adjusted. The initial clustering center K was not picked twice, and the process was performed iteratively. Next, each cluster of positive training samples $K = [k_1,\ k_2, \ldots, k_{m1}]^T$ was normalized by $K' = \frac{k_i - \mu}{\sigma}$ to be $K' = \left[k'_1, k'_2, \ldots,\ k'_{m1}\right]^T$. Then, the covariance matrix $\mathrm{R} = \left(r_{ij}\right)_{m*n}$ *of* K' was obtained and the following equations were used to add the synthesized positive samples to the positive training dataset.

$$K' = \left[k'_1, k'_2, \ldots,\ k'_{m1}\right]^T \tag{19}$$

$$r_{ij} = \frac{\left(\sum_{i=1}^{m} X_{kj} * X_{kj}\right)}{\left(\sum_{k=1}^{m} (X_{ki})^2 \sum_{k=1}^{m} \left(X_{kj}\right)^2\right)^{\frac{1}{2}}} \tag{20}$$

$$Y = [y_1, y_2, \ldots, y_{m1}]^T \tag{21}$$

$$y_i = (y_{i1}, y_{i2}, \ldots y_{i\mathrm{Dim}}), i = 1, 2, \ldots, \mathrm{Dim} \tag{22}$$

$$y_{ij} = k_i' * A(:, j), i = 1, 2, \ldots, m1, j = 1, 2, \ldots, \mathrm{Dim} \tag{23}$$

where *m1* is the number of positive samples in the *nth* cluster, *Dim* is the number of features in the positive samples, and *A* is the eigenvalue matrix of the covariance matrix arranged in a descending order. In this study, the value of *K* was set to 3 and KPCA was used on the positive samples. Finally, the experiment of clustering was repeated 4 times and their results were added to 183 positive training datasets as the new positive training dataset, including 377 samples. The number of the new positive training samples and that of the new negative training samples obtained from 365 samples using the FUS approach 1:1.

3.4.2. Fuzzy Undersampling Method

The FUS method [22] uses the fuzzy membership function to extract the information hidden in the training samples. In this study, the FUS approach was used to reduce the number of negative training samples to keep the positive and negative datasets in balance.

First, the mean and standard deviation of each feature of the positive and negative training samples were calculated as follows:

$$C_{\mathrm{pos}}^{j} = \frac{\sum_{i=1}^{\mathrm{PosNum}} \mathrm{Pos}(i, j)}{\mathrm{PosNum}}, j = 1, 2, \ldots, \mathrm{Dim} \tag{24}$$

$$C_{\mathrm{Neg}}^{j} = \frac{\sum_{i=1}^{\mathrm{NegNum}} \mathrm{Neg}(i, j)}{\mathrm{NegNum}}, j = 1, 2, \ldots, \mathrm{Dim} \tag{25}$$

$$\sigma_{\mathrm{Pos}}^{j} = \sqrt{\frac{1}{\mathrm{PosNum}} \sum_{i=1}^{\mathrm{PosNum}} \left(\mathrm{Pos}(i, j) - C_{\mathrm{Pos}}^{\mathrm{j}}\right)^2}, j = 1, 2, \ldots, \mathrm{Dim} \tag{26}$$

$$\sigma_{\mathrm{Neg}}^{j} = \sqrt{\frac{1}{\mathrm{NegNum}} \sum_{i=1}^{\mathrm{NegNum}} \left(\mathrm{Neg}(i, j) - C_{\mathrm{Neg}}^{\mathrm{j}}\right)^2}, j = 1, 2, \ldots, \mathrm{Dim} \tag{27}$$

where PosNum is the number of positive training samples, NegNum is the number of negative training samples, $Pos(i, j)$ is the value of the *jth* feature of the *ith* positive training sample, and $Neg(i, j)$ is the value of the *jth* feature of the *ith* negative training sample. *Dim* is the number of training samples features.

Second, the mean and standard deviation of the positive and negative samples were used to establish a membership function, defined as follows:

$$u_{\mathrm{Pos}}^{j}(i) = \mathrm{GaussMF}\left(\mathrm{Data}(i, j); C_{\mathrm{Pos}}^{j}, \sigma_{\mathrm{Pos}}^{j}\right) \tag{28}$$

$$u_{\mathrm{Neg}}^{j}(i) = \mathrm{GaussMF}\left(\mathrm{Data}(i, j); C_{\mathrm{Neg}}^{j}, \sigma_{\mathrm{Neg}}^{j}\right) \tag{29}$$

$$\mathrm{GaussMF}(x; C, \sigma) = \exp\left(-\frac{1}{2}\left(\frac{x - C}{\sigma}\right)^2\right) \tag{30}$$

where $Data(i,j)$ is the jth eigenvalue of the ith sample among the training samples $(i = 1,2,\ldots,PosNum, PosNum+1,\ldots,PosNum+NegNum,\ j = 1,2,\ldots,Dim)$.

$$\mathrm{Fval}(i,j) \ = u^{j}_{\mathrm{Pos}}(i) \ + \left(1 - u^{j}_{\mathrm{Neg}}(i)\right). \tag{31}$$

Finally, all the positive training samples were saved and the negative training samples with high scores were removed. The score function is defined as follows:

$$\mathrm{Score}(i) \ = \sum_{\mathrm{j=1}}^{\mathrm{Dim}} \mathrm{Fval}(i,j), (i = \ \mathrm{PosNum}+1,\ldots,\mathrm{PosNum}+\mathrm{NegNum}). \tag{32}$$

Then, FUS was applied on the negative training samples, and the number of negative training samples was reduced from 2258 to 365, which could decrease the inaccuracy caused by imbalanced datasets.

3.5. SVMs Implementation and Parameter Selectiont

SVM is a supervised learning method, which applies statistical theory to complete classification and regression in the area of bioinformatics [15,37–43]. In this study, the LibSVM package [43,44] was used to establish and evaluate the performance of the prediction model. The kernel function used the radial basis kernel function (RBF) $K\left(S_i, S_j\right) \ = e^{\left(-\gamma \| S_i - S_j \|^2\right)}$, and a grid search strategy based on 10 times of tenfold cross-validation and jackknife was employed to find the optimal parameters. The optimal parameters of tenfold cross-validation and jackknife are C = 0.70711 and γ= 1.4142, respectively.

Four measurements, including sensitivity (Sn), specificity (Sp), accuracy (Acc), and Matthews correlation coefficient (MCC), were used to evaluate the performance of the proposed predictor [44], which are shown as follows:

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP} \ + \ \mathrm{FN}} \tag{33}$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN} \ + \ \mathrm{FP}} \tag{34}$$

$$\mathrm{Acc} = \frac{\mathrm{TP} \ + \ \mathrm{TN}}{\mathrm{TP} \ + \ \mathrm{TN} \ + \ \mathrm{FP} \ + \ \mathrm{FN}} \tag{35}$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \ \times \ \mathrm{TN} \ - \ \mathrm{FN} \ \times \ \mathrm{FP}}{\sqrt{(\mathrm{TP} \ + \ \mathrm{FN}) \ \times \ (\mathrm{TP} \ + \ \mathrm{FP}) \ \times \ (\mathrm{TN} \ + \ \mathrm{FP}) \ \times \ (\mathrm{TN} \ + \ \mathrm{FN})}} \tag{36}$$

where TP, FP, TN, and FN are the number of true positives, false positives, true negatives, and false negatives, respectively.

4. Conclusions

In this study, the Pearson correlation coefficient was used to evaluate the relevance between any two amino acids, which were employed to obtain an optimal combination of amino acid pairs. Then, the AAPC (320) information was combined with TOP-n-gram, ANBPB, and PC-PseAAC to construct multiple feature vectors. Moreover, KPAC and FUS were used to solve the imbalance between positive and negative datasets. Finally, tenfold cross-validation, jackknife test, and an independent test were used to verify the proposed model. The prediction results showed that the proposed method was more accurate than the previous methods in predicting the pupylation sites. However, this study only considered the content of amino acids and neglected the location of amino acids in pupylated proteins. Amino acid environment such as surface areas and buried residues should be considered in future studies to explore the effects of amino acids at different positions. It should be pointed out user-friendly and publicly accessible web-servers will significantly enhance their impacts, we shall make efforts in our future work to provide a web-server for the prediction method presented

in this paper. The MATLAB (matrix laboratory) package of our prediction method is available as Supplementary material.

Supplementary Materials: The following are available online at http://www.mdpi.com/1420-3049/23/12/3097/s1, Excel S1, the training and test datasets used in this study; Table S1: The results of 20, 40, ..., 380, 400 of features by the 20 step length by jackknife test; Table S2: The prediction results of our method for the training dataset; Table S3: The prediction results of PUL-PUP for the training dataset; Table S4: The 10-fold cross-validation prediction results of PSoL on the training dataset; Table S5: The 10-fold cross-validation prediction results of SVM balance on the training dataset; Table S6: The prediction results of our method for the test dataset; Table S7: The prediction results of PUL-PUP for the test dataset; Table S8: The prediction results of PSoL for the test dataset; Table S9. The prediction results of SVM_balance for the test dataset.

Author Contributions: Conceptualization, T.L. (Tao Li) and Y.C.; methodology, T.L. (Tao Li); validation, T.L. (Taoying Li) and C.J.; formal analysis, T.L. (Tao Li); data curation, T.L. (Tao Li); writing—original draft preparation, T.L. (Tao Li); writing—review and editing, T.L. (Taoying Li) and C.J.; project administration, T.L. (Tao Li); funding acquisition, T.L. (Tao Li).

Funding: This research was funded by the National Science and Technology Major Project of China, grant number 2017YFC1404602.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

Pup	prokaryotic ubiquitin-like protein
KPCA	k-means principal component analysis oversampling technique
FUS	fuzzy undersampling method
Sn	sensitivity
Sp	specificity
Acc	accuracy
MCC	Matthews correlation coefficient
CKSAAP	composition of k-spaced amino acid pairs feature
pbCKSAAP	profile-based composition of k-spaced amino acid pairs
AAC	extracting amino acid composition
M. smegmatis	Mycobacterium smegmatis
M. tuberculosis	Mycobacterium tuberculosis
E. coli	Escherichia coli
AAPC	amino acid pair composition
ANBPB	adapted normal distribution bi-profile Bayes
BPB	bi-profile Bayes
PC-PseAAC	parallel correlation pseudo amino acid composition
PCA	principal component analysis
SVM	support vector machine
RBF	radial basis kernel function

References

1. Herrmann, J.; Lerman, L.O.; Lerman, A. Ubiquitin and ubiquitin-like proteins in protein regulation. *Circ. Res.* **2007**, *100*, 1276–1291. [CrossRef] [PubMed]
2. Welchman, R.L.; Gordon, C.; Mayer, R.J. Ubiquitin and ubiquitin-like proteins as multifunctional signals. *Nat. Rev. Mol. Cell. Bio.* **2005**, *6*, 599–609. [CrossRef] [PubMed]
3. Bauman, J.E.; Austin, M.C.; Schmidt, R.; Kurland, B.F.; Vaezi, A.; Hayes, D.N.; Mendez, E.; Parvathaneni, U.; Chai, X.; Sampath, S.; et al. ERCC1 is a prognostic biomarker in locally advanced head and neck cancer: Results from a randomised, phase II trial. *Br. J. Cancer* **2013**, *109*, 2096–2105. [CrossRef] [PubMed]
4. Lee, S.M.; Falzon, M.; Blackhall, F.; Spicer, J.; Nicolson, M.; Chaudhuri, A.; Middleton, G.; Ahmed, S.; Hicks, J.; Crosse, B. Randomized Prospective Biomarker Trial of ERCC1 for Comparing Platinum and Nonplatinum Therapy in Advanced Non–Small-Cell Lung Cancer: ERCC1 Trial (ET). *J. Clin. Oncol.* **2017**, *35*, 402–411. [CrossRef] [PubMed]

5. Faridounnia, M.; Wienk, H.; Kovacic, L.; Folkers, G.E.; Nicolass, G.J.; Kaptein, R.; Jan, H.J.; Boelens, R. The Cerebro-oculo-facio-skeletal Syndrome Point Mutation F231L in the ERCC1 DNA Repair Protein Causes Dissociation of the ERCC1-XPF Complex. *J. Biol. Chem.* **2015**, *33*, 20541–20555. [CrossRef] [PubMed]
6. Pearce, M.J.; Mintseris, J.; Ferreyra, J.; Steven, P.G.; Heran, D.K. Ubiquitin-Like Protein Involved in the Proteasome Pathway of Mycobacterium tuberculosis. *Science.* **2008**, *5904*, 1104–1107. [CrossRef] [PubMed]
7. Perez-Oliva, A.B.; Lachaud, C.; Szyniarowski, P.; Muñoz, I.; Macartney, T.; Hickson, I.; Rouse, J.; Alessi, D.R. USP45 deubiquitylase controls ERCC1-XPF endonuclease-mediated DNA damage responses. *EMBO J.* **2015**, *34*, 326–343. [CrossRef] [PubMed]
8. Zhang, L.; Gong, F. The emerging role of deubiquitination in nucleotide excision repair. *DNA Repair* **2016**, *43*, 34–37. [CrossRef] [PubMed]
9. Cuijk, L.V.; Van Belle, G.J.; Turkyilmaz, Y.; Poulsen, S.L.; Janssens, R.C.; Theil, F.A.; Sabatella, M. SUMO and ubiquitin-dependent XPC exchange drives nucleotide excision repair. *Nat. Commun.* **2015**, *6*, 7499. [CrossRef] [PubMed]
10. Tung, C.W. PupDB: A database of pupylated proteins. *BMC Bioinf.* **2012**, *1186*, 1471–2015. [CrossRef] [PubMed]
11. Liu, Z.; Ma, Q.; Cao, J.; Gao, X.; Ren, J.; Xue, Y. GPS-PUP: Computational prediction of pupylation sites in prokaryotic proteins. *Mol. Biosystems.* **2011**, *7*, 2737–2740. [CrossRef] [PubMed]
12. Tung, C.W. Prediction of pupylation sites using the composition of k-spaced amino acid pairs. *J. Theor. Biol.* **2013**, *336*, 11–17. [CrossRef] [PubMed]
13. Zhao, X.W.; Dai, J.Y.; Ning, Q.; Ma, Z.Q.; Yin, M.H.; Sun, P.P. Position-Specific Analysis and Prediction of Protein Pupylation Sites Based on Multiple Features. *Biomed Res. Int.* **2013**, *2013*, 1–9. [CrossRef] [PubMed]
14. Chen, X.; Qiu, J.D.; Shi, S.P.; Suo, S.B.; Liang, R.P. Systematic Analysis and Prediction of Pupylation Sites in Prokaryotic Proteins. *PLoS ONE* **2013**, *8*, e74002. [CrossRef] [PubMed]
15. Hasan, M.M.; Zhou, Y.; Lu, X.T.; Li, J.Y.; Song, J.N.; Zhang, Z.D. Computational Identification of Protein Pupylation Sites by Using Profile-Based Composition of k-Spaced Amino Acid Pairs. *PLoS ONE* **2015**, *10*, e0129635. [CrossRef] [PubMed]
16. Jiang, M.; Cao, J.Z. Positive-Unlabeled learning for pupylation sites prediction. *Biomed. Res. Int.* **2016**, *16*, 1–5. [CrossRef] [PubMed]
17. Nan, X.G.; Bao, L.L.; Zhao, X.S.; Zhao, X.W.; Sangaiah, A.K.; Wang, G.G.; Ma, Z.Q. EPuL: An Enhanced Positive-Unlabeled Learning Algorithm for the Prediction of Pupylation Sites. *Molecules* **2017**, *22*, 1463. [CrossRef] [PubMed]
18. Liu, B.; Wang, X.; Lin, L.; Dong, Q.; Wang, X. A discriminative method for protein remote homology detection and fold recognition combining Top-n-grams and latent semantic analysis. *BMC Bioinform.* **2008**, *9*, 510. [CrossRef] [PubMed]
19. Jia, C.Z.; Liu, T.; Wang, Z.P. O-GlcNAcPRED: A sensitive predictor to capture protein O-GlcNAcylation sites. *Mol. Biosyst.* **2013**, *9*, 2909–2913. [CrossRef] [PubMed]
20. Chou, K.C. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins: Structure Function Bioinform* **2001**, *43*, 246–255. [CrossRef] [PubMed]
21. Jia, C.Z.; Zuo, Y. Computional prediction of protein O-GlcNAc modification. *Methods Mol. Biol.* **2018**, *1754*, 235–246. [PubMed]
22. Hosseinzadeh, M.; Eftekhari, M. Using Fuzzy Undersampling and Fuzzy PCA to Improve Imbalanced Classification through Rotation Forest Algorithm. *CSSE Int. Symp. Cmpt. Sci. Software Eng.* **2015**, 1–7.
23. Kruskal, W.H. Ordinal measures of association. *Journal of the American Statistical Association* **1958**, *53*, 814–861. [CrossRef]
24. Lee Rodgers, J.; Nicewander, W.A. Thirteen ways to look at the correlation coefficient. *Am. Stat.* **1988**, *42*, 59. [CrossRef]
25. Liu, B.; Wu, H. Pse-in-One 2.0: A web server for generating comprehensive modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res.* **2017**, *10*, 4236.
26. Wang, C.; Ding, C.; Meraz, R.F.; Holbrook, S.R. PSoL: A positive sample only learning algorithm for finding non-coding RNA genes. *Bioinformatics* **2006**, *22*, 2590. [CrossRef] [PubMed]
27. Bui, V.M.; Weng, S.L.; Lu, C.T.; Cheng, T.L.; Chang, T.H.; Weng, T.Y.; Lee, T.Y. SOHSite: Incorporating evolutionary information and physicochemical properties to identify protein S-sulfenylation sites. *BMC Genomics* **2016**, *17*, 9. [CrossRef] [PubMed]

28. Xu, Y.; Ding, J.; Wu, L.Y. iSulf-Cys: Prediction of S-sulfenylation Sites in Proteins with Physicochemical Properties of Amino Acids. *PLoS ONE* **2016**, *11*, 4. [CrossRef] [PubMed]
29. Song, J.N.; Thomson, B.A. Cascleave: Towards more accurate prediction of caspase substrate cleavage sites. *Bioinformatics* **2010**, *26*, 752–760. [CrossRef] [PubMed]
30. 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* **2013**, *8*, 2. [CrossRef] [PubMed]
31. Liu, B.; Fang, L.Y.; Liu, F.L.; Wang, X.L.; Chen, J.J.; Chou, K.C. Identification of Real MicroRNA Precursors with a Pseudo Structure Status Composition Approach. *PLoS ONE* **2015**, *10*, 3. [CrossRef] [PubMed]
32. Sokal, R.R.; Thomson, B.A. Population structure inferred by local spatial autocorrelation: An example from an Amerindian tribal population. *Am. J. Phys. Anthropol.* **2006**, *129*, 121–131. [CrossRef] [PubMed]
33. Kawashima, S.; Kanehisa, M. AAindex: Amino acid index database. *Nucleic Acids Res.* **2000**, *28*, 374. [CrossRef] [PubMed]
34. Lin, H.; Deng, E.Z.; Ding, H.; Chen, W.; Chou, K.C. iPro54-PseKNC: A sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. *Nucleic Acids Res.* **2014**, *42*, 12961–12972. [CrossRef] [PubMed]
35. Leslie, C.S.; Eskin, E.; Cohen, A.; Weston, J.; Noble, W.S.S. Mismatch string kernels for discriminative protein classification. *Bioinformatics* **2004**, *20*, 467–476. [CrossRef] [PubMed]
36. Jia, C.Z.; Zuo, Y.; Zou, Q. O-GlcNAcPRED-II: An integrated classification algorithm for identifying O-GlcNAcylation sites based on fuzzy undersampling and a K-means PCA oversampling technique. *Bioinformatics* **2018**, *34*, 2029–2036. [CrossRef] [PubMed]
37. Si, J.N.; Yan, R.X.; Wang, C.; Zhang, Z.D.; Su, X.D. TIM-Finder: A new method for identifying TIM-barrel proteins. *BMC Struct. Biol.* **2009**, *9*, 1–10. [CrossRef] [PubMed]
38. Xu, L.L.; Li, J.T.; Shu, Y.M.; Peng, J.H. SAR Image Denoising via Clustering-Based Principal Component Analysis. *IEEE T. Geosci. Remote* **2014**, *52*, 6858–6869.
39. Yan, R.X.; Si, J.N.; Wang, C.; Zhang, Z.D. DescFold: A web server for protein fold recognition. *BMC Bioinform.* **2009**, *10*, 416. [CrossRef] [PubMed]
40. Liao, Z.J.; Wan, S.X.; He, Y.; Zou, Q. Classification of Small GTPases with Hybrid Protein Features and Advanced Machine Learning Techniques. *Curr. Bioinform.* **2018**, *13*, 492–500. [CrossRef]
41. Liao, Z.J.; Li, D.P.; Wang, X.R.; Li, L.S.; Zou, Q. Cancer Diagnosis Through IsomiR Expression with Machine Learning Method. *Curr. Bioinform.* **2018**, *13*, 57–63. [CrossRef]
42. Li, D.P.; Ju, Y.; Zou, Q. Protein Folds Prediction with Hierarchical Structured SVM. *Curr. Proteomics* **2016**, *13*, 79–85. [CrossRef]
43. Chang, C.C.; Lin, C.J. LIBSVM: A Library for Support Vector Machines. *ACM T. Intel. Syst. Tec.* **2011**, *2*, 3. [CrossRef]
44. Ju, Z.; Cao, J.Z.; Gu, H. Predicting lysine phosphoglycerylation with fuzzy SVM by incorporating k-spaced amino acid pairs into Chou's general PseAAC. *Theor. Biol.* **2016**, *397*, 145–150. [CrossRef] [PubMed]

Sample Availability: Not available.