genes

MDPI

Article

PupStruct: Prediction of Pupylated Lysine Residues Using Structural Properties of Amino Acids

Vineet Singh [1,*], Alok Sharma [2,3,4,*], Abdollah Dehzangi [5,6] and Tatushiko Tsunoda [3,7,8]

1 Faculty of Science Technology and Environment, University of the South Pacific, Suva, Fiji
2 Institute for Integrated and Intelligent Systems, Griffith University, Brisbane, QLD 4111, Australia
3 Laboratory for Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences, Yokohama 230-0045, Japan; tsunoda@bs.s.u-tokyo.ac.jp
4 School of Engineering and Physics, Faculty of Science Technology and Environment, University of the South Pacific, Suva, Fiji
5 Department of Computer Science, Rutgers University, Camden, NJ 08102, USA; i.dehzangi@rutgers.edu
6 Center for Computational and Integrative Biology, Rutgers University, Camden, NJ 08102, USA
7 Laboratory for Medical Science Mathematics, Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Tokyo 113-0033, Japan
8 Department of Medical Science Mathematics, Medical Research Institute, Tokyo Medical and Dental University, Tokyo 113-8510, Japan
* Correspondence: vineet.singh@usp.ac.fj (V.S.); alok.sharma@griffith.edu.au (A.S.); Tel.: +679-9436172 (V.S)

Received: 22 October 2020; Accepted: 23 November 2020; Published: 28 November 2020

Abstract: Post-translational modification (PTM) is a critical biological reaction which adds to the diversification of the proteome. With numerous known modifications being studied, pupylation has gained focus in the scientific community due to its significant role in regulating biological processes. The traditional experimental practice to detect pupylation sites proved to be expensive and requires a lot of time and resources. Thus, there have been many computational predictors developed to challenge this issue. However, performance is still limited. In this study, we propose another computational method, named PupStruct, which uses the structural information of amino acids with a radial basis kernel function Support Vector Machine (SVM) to predict pupylated lysine residues. We compared PupStruct with three state-of-the-art predictors from the literature where PupStruct has validated a significant improvement in performance over them with statistical metrics such as sensitivity (0.9234), specificity (0.9359), accuracy (0.9296), precision (0.9349), and Mathew's correlation coefficient (0.8616) on a benchmark dataset.

Keywords: post-translational modification (PTM); lysine pupylation; structural features; protein sequences; amino acids; prediction

1. Introduction

Post-translational modifications (PTM) are referred to as changes of protein composition through the addition of small molecules to specific sites (amino acid residues) on the body of a protein. Those modifications are responsible for protein function regulation, cell functioning, subcellular localization, biological practices, and protein turnover in health and disease [1–5]. Different PTMs have been identified so far, including ubiquitination [6], methylation [7,8], acetylation [9], glycation [10], prolyl isomerization [11], succinylation [12–16], crotonylation [17], and phosphoglycerylation [18].

While there have been a lot of studies done on these PTMs, another modification named pupylation [19,20] has attracted much attention in the research community. The process whereby prokaryotic ubiquitin-like protein (Pup) [21,22] attaches to substrates for degradation via an isopeptide

bond causing the modification of specific lysine residues is pupylation [23]. Although Pup and Ubiquitin (Ub) are similar, their amino acid sequence or structure are different. While Ub has 76 amino acids, Pup proteins are small, ranging from 60 to 70 residues in length [24]. Pupylation plays a key role in regulating various cellular procedures, such as signal transduction and protein degradation in prokaryotic cells [25,26]. The pupylation and ubiquitylation are the same in function [27], but the enzymology involved in ubiquitylation requires an activating enzyme, conjugating enzyme, and protein ligase. In contrast, pupylation only requires deamidase of Pup (DOP) [28] and proteasome accessory factor A (PafA) [19,24,29]. First, the C-terminal glutamine of Pup is deamidated to glutamate via DOP, then the deamidated Pup is attached to a specific lysine of substrate proteins by PafA. The tagging with Pup can render proteins as substrates for proteasomal degradation. The depupylation event in actinobacteria and the fact that some members harbor the pupylation gene locus without encoding proteasomal subunits proposes the assumption that pupylation might fulfill a larger role in regulation and cellular signaling [30]. Most prokaryotic pupylation remains unknown [31,32].

To understand the fundamentals of pupylation, it is critical to involve biological markers at the cellular level for detecting pupylation sites. Identifying a pupylated site through the traditional experimental process is demonstrated to be expensive, complex, inefficient, and time-consuming. To overcome these disadvantages, computational methods are more preferred and a prediction tool is needed.

There are a number of computational models developed with a different technique, but there are a lot of improvements that can be done for better performance. Some of these methods include the first proposed technique GPS-PUP, which employed a group-based prediction system (GPS) sequence encoding [33], and EnsemblePup [34] which incorporated the bi-profile Bayes feature extraction with support vector machine (SVM). Features such as position-specific scoring matrix (PSSM), secondary structure, amino acid index property (AAindex), conservation scores, and structural disorder score were employed with an SVM classifier to develop PrePup [35]. IMP-PUP [36] constructed features based on the composition of k-spaced amino acid pairs on a semi-supervised self-training SVM algorithm, while pbPUP [37] was developed with the profile-based composition of k-spaced amino acid pair (pbCKSAAP) encoding with the SVM classifier. PUL-PUP [38] made use of the SVM algorithm and positive-unlabeled learning with a composition of k-spaced amino acid pairs feature (CKSAAP), iPUP [39] also incorporated CKSAAP features. The structural, sequential, and evolutionary hallmarks features which included protein secondary structures, physicochemical properties, binary features, PSSM, and amino acid pairs and SVM classifier was employed to develop PupPred [40]. EPuL [41] incorporated only positive and unlabeled samples. The progress and challenges faced in protein pupylation sites prediction were discussed in [20]. CIPPN [42] was developed using a neural network and, most recently, PSSM-PUP [43] employed PSSM, which was converted into bigram probabilities for feature extraction with an LibSVM classifier was developed.

The benchmark datasets from the PupDB database [44] is used in most of the studies. While many of the methods used the composition of k-spaced amino acid pairs features, there are only three methods, namely PrePup [35], PUL-PUP [38], and PupPred [40], which involve secondary structural features. Despite several methods being presented so far, their performance in identifying pupylated lysine residues remains limited, and therefore better techniques are necessary to determine the pupylated and non-pupylated lysine residues correctly.

In this study, we propose a new predictor, named PupStruct, which utilizes structural features such as accessible surface area (ASA), secondary structure (helix, strand, and coil), and backbone torsion angles for predicting pupylated lysines. The peptide comprising 13 amino acids upstream and downstream of lysine residue was employed for feature extraction. The benchmark dataset PupDB database [44] consisting of 153 proteins was used with a high number of non-pupylated lysines over the pupylated lysines. To reduce data imbalance, we used a k-nearest neighbors cleaning treatment [45] and employed a support vector machine with a radial basis kernel function for pupylation prediction. Structural features that contribute to the better overall performance of PupStruct in comparison to

other methods was used. Finally, PupStruct was compared with two benchmark predictors ([36] and [38]) which showed a significantly improved performance over them. PupStruct is able to predict pupylated lysines with 0.9234 sensitivity, 0.9359 specificity, 0.9296 accuracy, and 0. 8616 Mathew's correlation coefficient.

2. Materials and Methods

We propose a computational method named PupStruct which employs nine structural features, including accessible surface area, secondary structure (helix, strand, and coil), and backbone torsion angles. The following sections discuss the benchmark dataset, different features, feature exaction for each lysine, and the support vector machine classifier used for pupylation site prediction.

2.1. Protein Dataset

As stated in the introduction, for this study, we have taken the protein sequences from PupDB databases [44]. It contained 153 protein sequences whose lysine residues are either pupylated or non-pupylated. We examined each protein sequence and retrieved whether it was composed of pupylation or non-pupylation residues. We attained 181 positive lysines (pupylated) and 2290 negative lysines (non-pupylated) which were used for this study. The next section explains various structural features computed from each of the protein sequences.

2.2. Structural Features

We retrieved each protein sequence and computed nine different features related to the accessible surface area, secondary structure, and backbone torsion angles. These features are also used in other existing predictors [12,18,42,46,47]. To achieve this, we employed the toolbox SPIDER2 [48,49] which has previously obtained good outcomes for prediction using accessible surface area [50–53], secondary structure [54–57], and backbone torsion angles [50,58–61]. SPIDER2 has also been used to extract the structural properties for other predictions [12,18,61–64]. The details for these structural properties are explained in the succeeding sections.

2.2.1. Accessible Surface Area (ASA)

ASA refers to the accessible area of each amino acid to a solvent of the protein in 3D configuration [65–67]. Since the value of an amino acid involves the protein configuration, the predicted *ASA* value of individual amino acids displays vital information regarding the protein structure. We executed SPIDER2 on each protein sequence to compute an estimated numeric *ASA* value for each amino acid in the protein with known 3D structures [48]. It is wise to note that the predicted *ASA* value entirely depends on the sequence information which is mainly used by SPIDER2 for computation.

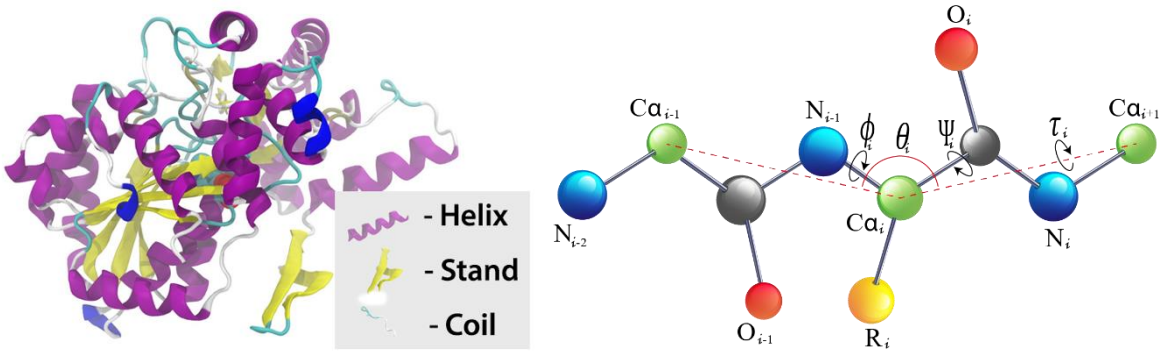
2.2.2. Secondary Structure

This property presents significant information on the local 3D structure of proteins. This can be inferred as amino acid's contribution to each of the defined local structures of proteins, namely helix (*ph*), strand (*pe*), and coil (*pc*), as is shown in Figure 1a. Again, we executed SPIDER2 to predict the prospect contribution of each amino acid to the three mentioned local structures, namely *ph, pe* and *pc*, which results in three discrete numerical vectors of these local structures [68]. Furthermore, SPIDER2 also gives the local structure with the highest probability as one $L \times 3$ matrix, where L depicts the protein length, and the three columns are the corresponding probabilities contribution to each local structure *ph, pe* and *pc*. Hence, to simplify, we denote this matrix as *SSPre* [69].

2.2.3. Local Backbone Torsion Angles

The secondary structure gives important information on local configuration of amino acids of protein [70], whereas torsion angles between neighboring amino acids supplement predicted *ASA* and

secondary structure with vital information about the local structure of proteins. Since the predicted secondary structure is a distant output, the backbone torsion angle ϕ and Ψ continuously provides information on local amino acids interaction along the protein backbone [71,72]. Recently, two new angles are identified based on the dihedral angles θ, between three $C\alpha$ atoms ($C\alpha_{i-1}$–$C\alpha_i$–$C\alpha_{i+1}$) and τ, rotated about the $C\alpha_1$–$C\alpha_{i+1}$ bond [50]. To attain these four angles, we executed SPIDER2 [49] on every protein sequence and achieved four numerical vectors, namely ϕ, Ψ, θ and τ. The illustration of ϕ, Ψ, θ and τ is shown in Figure 1b.



(**a**) (**b**)

Figure 1. Illustrations of the secondary structure of a protein and local backbone torsion angles in amino acids (**a**) the helix, strand and coil for a protein (picture source: freepik.com) while (**b**) illustrates torsion angles associated with the protein backbone. Dihedral angle for the different bonds are discussed above.

2.3. Feature Extraction for Lysine Residues

Structural features are used to identify the pupylated and non-pupylated sites by employing 6 upstream and 6 downstream amino acids from and including the lysine residue K as shown in Figure 2 which adds the window size equal to 13. The mirroring effect [18,47,73–75] was employed to fill the amino acids in the absence of 6 amino acids either upstream or downstream from the lysine residue, i.e., if the lysine residue is located near the *N* or *C* terminus as shown in Figure 3. To obtain the best window size, we constructed training dataset using 11- to 41-residue window sizes and trained the PupStruct predictor but the best result was obtained by window size 13.

Let us consider peptide S, consisting of 6 upstream and 6 downstream amino acids, including lysine reside K in the middle, that can be stated as:

$$S = \{A_{-6},\ A_{-5},\ A_{-4},\ A_{-3},\ A_{-2},\ A_{-1},\ K,\ A_1,\ A_2,\ A_3,\ A_4,\ A_5,\ A_6\} \quad (1)$$

Where A_{-i} (for $1 \le i \le 6$) are upstream and A_i (for $1 \le i \le 6$) are downstream amino acids. Therefore, the lysine residue consists of 13 amino acids in total, including *K*. Each peptide *S* will contain a pupylated or non-pupylated lysine which means the *K* can have a class label *x* as $x = \{0, 1\}$ where $x = 1$ then *S* denotes pupylated residue and if $x = 0$ then *S* denotes non-pupylated residue. Moreover, each amino acid A_i (for $-6 \le i \le 6$; $A_0 = K$) can be deliberated by the structural features as:

$$A_i = \{ASA,\ ph,\ pe,\ pc,\ \phi, \Psi, \theta,\ \tau\} \quad (2)$$

It is worth noting that the structural features *ASA*, *ph*, *pe*, *pc*, ϕ, Ψ, θ and τ are numeric vales and each represents a sole value for each amino acid A_i. Thus, A_i can be expressed in an 8-dimensional feature vector. The numeric values were normalized and then placed in a vector form. This implies that each segment *S* (of 13 amino acids) is represented by 104 structural features (of 13 amino acids *x* 8). These structural features of lysine are used to predict pupylated or non-pupylated sites in line with the peptide *S*.

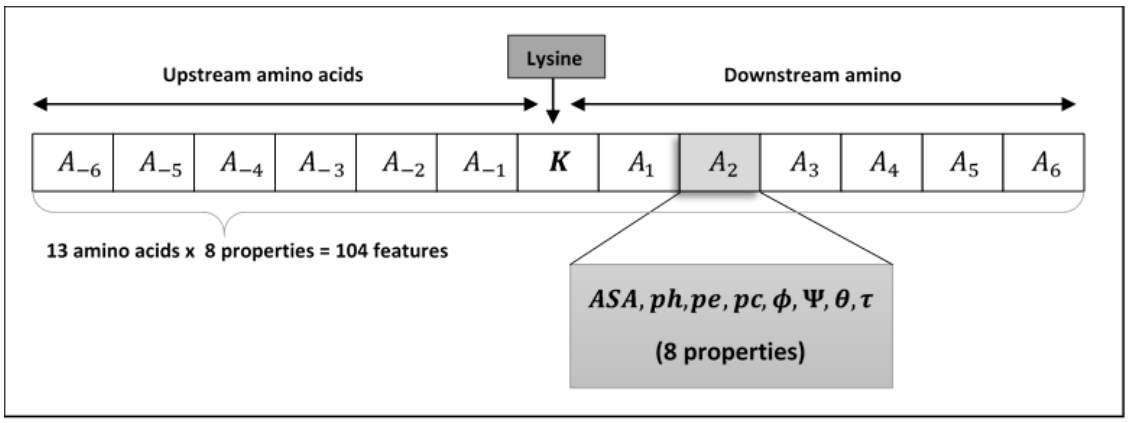


Figure 2. Shows arrangement of lysine residue's neighboring amino acids with ample upstream and downstream amino acids.

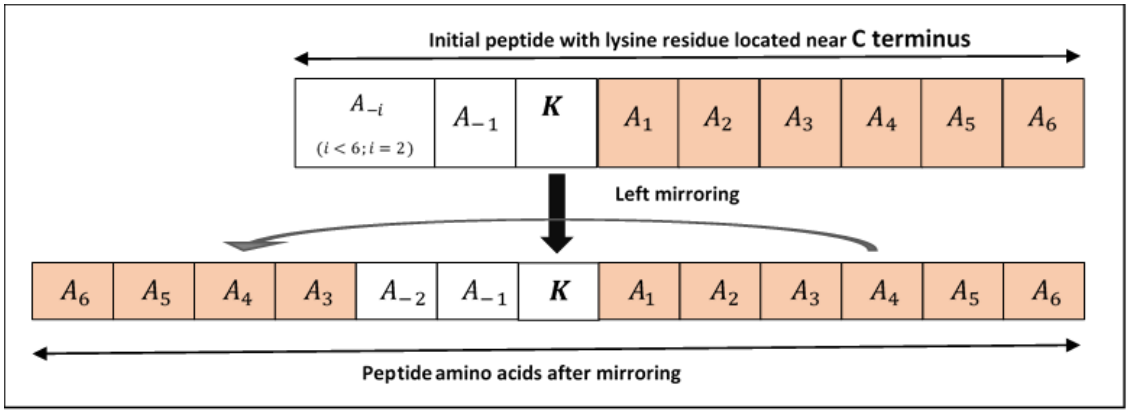


Figure 3. Shows lysine with insufficient amino acids. Left mirroring is illustrated to get sufficient upstream amino acids.

2.4. Support Vector Machine for Classification

The support vector machine (SVM) [76,77] is engaged in both regression and classification applications and is also used in many state-of-the-art predictors for pupylation sites [35–39,41,43]. The literature shows that the SVM produces a lower prediction error compared to other classifiers when large numbers of features are considered, as in this study there are 104 features. It is also proven that SVM has produced best results and mostly used in the areas of bioinformatics research including genomics, protein function prediction, protease functional site recognition, chemogenomics, transcription initiation site prediction and gene expression data classification [78–80]. SVM operates by discovering the maximum difference among the two hyperplanes demonstrating linear boundaries of two different classes. The dimension of the hyperplane is influenced by the number of features, therefore, kernel functions which can be either polynomial, radial or linear was employed to deal with non-linear boundaries between classes [81–83]. We used the LIBSVM [84] classifier developed Chih-Chung Chang and Chih-Jen Lin on the Matlab platform developed the MathWorks, Inc from Natick, Massachusetts, United States, with a radial basis kernel function to determine the margin between pupylated and non-pupylated lysine residues. The radial basis kernel function was fine-tuned with a gamma set to 0.5 and cost value set to 3 (nu-SVR).

3. Results

Having a computational method which aims to predict pupylation sites requires a severe assessment of its performance. This section discusses the statistical metrics, evaluation strategy,

balancing dataset, oversampling dataset, and the comparison of the proposed PupStruct with other recent state-of-the-art predictors from the literature.

3.1. Performance Measures

In this study, we have incorporated five metrics to compare the performance of PupStruct with other state-of-the-art predictors in terms of predicting pupylated and non-pupylated lysines. These five metrics are sensitivity (Sn), specificity (Sp), accuracy (Acc), precision (Pre) and Matthew's correlation coefficient (Mcc), which are widely used in the literature [35,37–39,85]

Sensitivity, which is one of the key measures, evaluates the correctness of identifying pupylation sites. The predictor attaining high sensitivity shows that it can accurately detect the pupylated lysines (positive instances). In other words, sensitivity value 0 shows the predictor's inability to detect any pupylated lysine residues (true positives), whereas the value 1 depicts a predictor's capability to correctly identify all pupylated lysines. Specificity gauges the predictor's capability to detect non-pupylation sites (true negative). It varies between 0 and 1 where value 0 shows predictor's inability to predict non-pupylation sites and value 1 indicates predictor's ability to predict non-pupylation sites. Accuracy calculates the total number of correctly classified pupylated and non-pupylated lysine residues which ranges between 0 and 1 where 0 means a least accurate predictor and 1 means the best accurate predictor. Precision, which is another assessment measure, is a fraction of correctly identified pupylated sites over the sum of correctly identified pupylated and non-pupylated sites. Mathew's correlation coefficient (MCC) scales the classification quality of the predictor, which ranges from −1 to +1. A predictor with (MCC) value −1 implies a totally negative correlation, whereas +1 means a completely positive correlation.

Considering Equations (3)–(7) for each metric, let's look at a dataset with $+P$ as a number of pupylated sites and $-P$ as a number of non-pupylated sites. Therefore, each metric can be expressed as:

$$Sensitivity = \frac{+P_+}{+P_+ + +P_-} \tag{3}$$

$$Specificity = \frac{-P}{-P_+ + -P_-} \tag{4}$$

$$Precision = \frac{+P_+ + -P_+}{+P + -P} \tag{5}$$

$$Accuracy = \frac{+P_+ + -P_+}{+P + -P} \tag{6}$$

$$MCC = \frac{(-P_+ \times +P_+) - (-P_- \times +P_-)}{\sqrt{(+P_+ + +P_-)(+P_+ + -P_-)(-P_- + +P_-)(-P_+ + -P_-)}} \tag{7}$$

where $+P_+$ is number of pupylated sites classified correctly, $+P_-$ represents the number of pupylated sites incorrectly classified, $-P_+$ is the number of non-pupylated sites predicted correctly and $-P_-$ represents the number of incorrectly predicted non-pupylated sites by the predictor.

The perfect predictor should achieve the highest in all the five metrics. However, at least sensitivity should be greater when comparing it with other predictors. A lower value of sensitivity shows that it cannot correctly predict pupylated lysine residues and, therefore, it is not fit for lysine pupylation detection.

3.2. Evaluation Strategy

To accurately evaluate the effectiveness of the PupStruct predictor in terms of the statistical metrics, we used a cross-validation method. Two most common cross-validation approaches are n-fold cross-validation and jackknife. An independent test set is used for evaluation purposes. The jackknife method is considered to be the least arbitrary and yields unique outcomes for a dataset [86] but we

deployed the n-fold cross-validation scheme for this study which involves less processing time and also commonly used in the literature [15,35,36,38,39,41]. The n-fold cross-validation technique is employed in the following steps:

1. Partition data samples randomly into n parts of roughly equal size with roughly similar negative and positive samples on each fold.
2. Take out one-fold as test set or validation data and the remaining n-1 folds as training data.
3. Use the training data set to fine-tune the parameters of the predictor.
4. Use the test set to compute the five statistical metrics.
5. Repeat Step 1 to Step 4 for the remaining n folds and calculate the average of each performance metric.

We carried out 6-, 8- and 10-fold cross-validations to evaluate PupStruct predictor and recorded the result.

3.3. Filtering Out the Imbalance Data

Our benchmark dataset comprised 163 protein sequences which has 181 pupylation sites (positive sample set) and 2290 non-pupylation sites (negative sample set). The difference between the number of positive set and negative set of around 12 times creates a huge imbalance between the classes. Although this may be biologically realistic to have a number of non-pupylated lysines greater than pupylated lysines, this inconsistency can cause severe bias in machine learning. We applied the k-nearest neighbors cleaning technique [45] to tackle this problem, which is mostly used in the literature.

For this, we first set the cut-off K value equal to 12 since the negative and positive sample ratio was about 12:1, thus we eliminated any negative sample which had at least one positive sample within its 12-nearest neighbors. We consequently increased the k value until we achieved similar numbers of positive and negative samples. Eventually, the number of negative samples was significantly reduced to 188 samples by k value of 48. After the filtering process, the filtered negative samples and all positive samples were used to perform n-fold cross-validation to evaluate the PupStruct predictor.

3.4. PupStruct vs. Other Existing Predictors

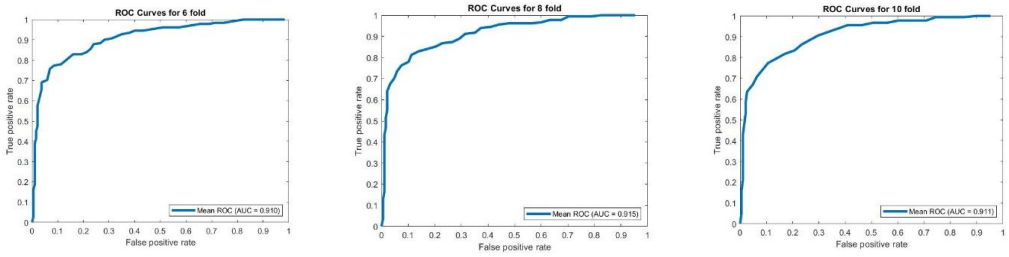
The proposed PupStruct was compared with recent two proposed predictors IMP-PUP [36] and PUL-PUP [38]. It can be noted that PUL-PUP also used structural features. The software packages were given for the two methods. It is worth noting that both predictors used the same dataset thus, many of the proteins may be used in their training set. Therefore, the software was re-run and tested using the test set respectively to the set used in PupStruct evaluation process. For PUL-PUP [38], since the code didn't execute, we retrieve the features from the method and used the same train and test set from PupStruct to calculate the performance. The performance reported is based on test data which correspond to the test set kept aside during the n-fold cross-validation procedure means that we keep aside the test set during the n-fold cross-validation procedure. Test data was not used to adjust the training parameters of the model.

Table 1 reports the performance of the predictors. It is clearly witnessed that the proposed PupStruct is performing better than all the benchmark predictors in metrics in likes of sensitivity, accuracy and MCC. The sensitivity was improved by 14%, accuracy by 11%, specificity by 7%, and precision by 9%. Moreover, MCC was significantly improved by 21% compared to IMP-PUP [36].

To give more insight of the performance of PupStruct, we generated ROC curve to measure AUC (area under the curve) and calculated the average AUC values for 6-, 8-, and 10-fold cross validations which was recorded at 0.910, 0.915 and 0.911 respectively which indicates stable performance of PupStruct. The results of the ROC-AUC analysis are shown in Figure 4.

Table 1. Shows comparison of performance assessment of PupStruct and two benchmark predictors for 6-, 8-, 10-fold cross-validation. The highest values in each metric are in bold.

Fold	Predictor	Sensitivity	Specificity	Precision	Accuracy	MCC
6	PUL-PUP	0.5586	0.7547	0.6897	0.6586	0.3219
	IMP-PUP	0.7785	0.8611	0.8407	0.8205	0.6437
	PupStruct	0.9228	0.9309	0.9317	0.9270	0.8563
8	PUL-PUP	0.5753	0.7919	0.7308	0.6856	0.3826
	IMP-PUP	0.7767	0.8610	0.8422	0.8197	0.6423
	PupStruct	**0.9234**	**0.9359**	**0.9349**	**0.9296**	**0.8616**
10	PUL-PUP	0.6082	0.7190	0.6946	0.6646	0.3380
	IMP-PUP	0.7784	0.8611	0.8429	0.8203	0.6441
	PupStruct	0.9173	0.9409	0.9398	0.9296	0.8611



(**a**) (**b**) (**c**)

Figure 4. Shows mean Receiver operating characteristic (ROC) curves of PupStruct predictions on 6-, 8- and 10-folds (**a**) ROC curves for 6 folds, (**b**) ROC curves for 8 folds, (**c**) ROC curves for 10 folds. This encouraging result demonstrates the proficiency of proposed PupStruct predictor to distinguish the pupylated and non-pupylated lysine residues accurately. It appears that structural information of amino acids provides essential information about nearby modified lysines. Finally, the SVM classifier with a radial basis kernel function appears to discover the maximal separation of both hyperplanes when structural characteristics are employed. All these structural features together with the classifier plays a key role in predicting pupylated and non-pupylated lysine residues.

Our PupStruct predictor's software package can be accessed from https://github.com/vinzsingh09/PupStruct.

4. Discussion

We further analyzed each feature (accessible surface area, secondary structure (helix, strand, and coil), and backbone torsion angles) to gauge their contribution towards the predictor. We used different features to train and test the model and recorded the result for comparison. Initially, we used each group of features for training, that is *ASA* then secondary structure (helix, strand, and coil) and backbone torsion angles. Next, we used individual features (*ASA*, *ph*, *pe*, *pc*, ϕ, Ψ, θ, τ) separately and recorded their result contributing to the predictor. Finally, we used a combination of some of these features which are contributing the most towards the predictor. The result shown in Table 2 is for six-fold cross validation.

It is clearly observed from Table 2 that *ASA* and secondary structure (*ph*, *pe*, *pc*, also known as *SSpre*) contributes the most towards the performance. It observed that *SSPre* has contributed the most towards specificity and precision, while *ASA* contributes the most towards sensitivity and MCC, which are the most important metrics. However, combining the two reduces the performance. Protein's shape is determined by amino acid sequence in the polypeptide chain. When exposed to the cytosol (water-based solution in which proteins floats) or lumen (inside space of a tubular structure), polypeptide chain assumed to localized organization to secondary structure that optimizes interactions between side chains of amino acids with each other and water. The polypeptide backbone folds

into spirals (helix) and ribbons (stand). These properties provide very important information about the amino acid and extracting the helix, stand, and coil (*SSPre*) values contributed the most to the performance of PupStruct [87,88]. In the literature, the accessible surface area (*ASA*) of a protein is always considered as a determining factor in protein folding and stability work. *ASA* is surface characterized around a protein by a hypothetical centre of a solvent sphere with the surface of the molecule. Based on the *ASA* value, amino acid residues can be determined as buried or exposed. This makes *ASA* a crucial feature contributing towards the performance [88]. When considering the individual feature only, then *ASA*, coil (*pc*), followed by Tau from the local torsion angle contributes the most towards the predictor. Eventually, using all the features gave the best result, which is shown in Table 1, which means that each feature demonstrated some contribution towards the predictor.

Table 2. Shows features and what percentage it contributed towards the predictor.

Feature	Sn (%)	Sp (%)	Pre (%)	Acc (%)	MCC (%)
ASA	86.70251	87.83602	87.60489	87.26766	0.74812
Ph, Pe, Pc (SSPre)	81.75627	92.89773	91.29129	87.56934	0.754546
Helix (*Ph*)	65.08961	93.59879	90.65543	79.61739	0.615526
Strand (*Pe*)	43.15412	95.39141	91.2274	70.2561	0.461857
Coil (*Pc*)	81.72043	89.78495	89.19853	85.81879	0.723357
Local Torsion angle	75.71685	69.80843	71.79877	72.77045	0.457679
Phi	76.73835	78.88889	78.17483	77.80965	0.560354
Psi	75.66308	76.88172	76.33012	76.26917	0.526891
Theta	61.21864	81.49425	77.14761	71.3354	0.438473
Tau	80.10753	81.1828	80.70276	80.64075	0.615214
ASA + Sspre	77.921147	88.25605	86.53159	83.18623	0.66673

5. Conclusions

This study presented a new computational method named PupStruct for identifying pupylation sites in protein sequences. PupStruct utilizes structural information of amino acids around the lysine residue and uses the k-nearest neighbour approach to solve the imbalance data issue. The analysis of which features contribute how much to the predictor was crucial information for training. Finally, the support vector machine (LIBSVM) with a radial basis kernel function to identify maximal separation between pupylated and non-pupylated lysine residue showed that PupStruct performed better than the existing predictors.

Author Contributions: V.S. and A.S.; methodology, V.S.; software, V.S; validation, V.S., A.S. and A.D.; formal analysis, V.S., A.D. and A.S.; investigation, V.S.; resources, A.S. and T.T.; data curation, V.S. and T.T.; writing—original draft preparation, V.S. and A.S.; writing—review and editing, V.S., A.S., and A.D.; visualization, V.S.; supervision, A.S.; project administration, V.S.; funding acquisition, A.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by JST CREST Grant Number JPMJCR 1412, Japan; JSPS KAKENHI Grant Numbers JP17H06307, JP17H06299, and JP20H03240; Grant-in-Aid for Scientific Research (JP16H06299) from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

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

References

1. Haj-Yahya, M.; Lashuel, H.A. Protein semisynthesis provides access to tau disease-associated post-translational modifications (PTMs) and paves the way to deciphering the tau PTM code in health and diseased states. *J. Am. Chem. Soc.* **2018**, *140*, 6611–6621. [CrossRef] [PubMed]
2. Mann, M.; Jensen, O.N. Proteomic analysis of post-translational modifications. *Nat. Biotechnol.* **2003**, *21*, 255–261. [CrossRef] [PubMed]
3. Deribe, Y.L.; Pawson, T.; Dikic, I. Post-translational modifications in signal integration. *Nat. Struct. Mol. Biol.* **2010**, *17*, 666–672. [CrossRef] [PubMed]
4. Hart, G.W.; Ball, L.E. Post-translational Modifications: A Major Focus for the Future of Proteomics. *Mol. Cell. Proteom.* **2013**, *12*, 3443. [CrossRef]

5. Walsh, C.T.; Garneau-Tsodikova, S.; Gatto Jr, G.J. Protein posttranslational modifications: The chemistry of proteome diversifications. *Angew. Chem. Int. Ed.* **2005**, *44*, 7342–7372. [CrossRef]
6. Qiu, W.-R.; Xiao, X.; Lin, W.-Z.; Chou, K.-C. iUbiq-Lys: Prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a gray system model. *J. Biomol. Struct. Dyn.* **2015**, *33*, 1731–1742. [CrossRef]
7. Liu, Z.; Xiao, X.; Qiu, W.-R.; Chou, K.-C. iDNA-Methyl: Identifying DNA methylation sites via pseudo trinucleotide composition. *Anal. Biochem.* **2015**, *474*, 69–77. [CrossRef]
8. Lan, F.; Shi, Y. Epigenetic regulation: Methylation of histone and non-histone proteins. *Sci. China Ser. C Life Sci.* **2009**, *52*, 311–322. [CrossRef]
9. Hou, T.; Zheng, G.; Zhang, P.; Jia, J.; Li, J.; Xie, L.; Wei, C.; Li, Y. LAceP: Lysine acetylation site prediction using logistic regression classifiers. *PLoS ONE* **2014**, *9*, e89575. [CrossRef]
10. Singh, R.; Barden, A.; Mori, T.; Beilin, L. Advanced glycation end-products: A review. *Diabetologia* **2001**, *44*, 129–146. [CrossRef]
11. Wulf, G.; Finn, G.; Suizu, F.; Lu, K.P. Phosphorylation-specific prolyl isomerization: Is there an underlying theme? *Nat. Cell Biol.* **2005**, *7*, 435–441. [CrossRef] [PubMed]
12. López, Y.; Sharma, A.; Dehzangi, A.; Lal, S.P.; Taherzadeh, G.; Sattar, A.; Tsunoda, T. Success: Evolutionary and structural properties of amino acids prove effective for succinylation site prediction. *BMC Genom.* **2018**, *19*, 105–114. [CrossRef] [PubMed]
13. Jia, J.; Liu, Z.; Xiao, X.; Liu, B.; Chou, K.-C. pSuc-Lys: Predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach. *J. Theor. Biol.* **2016**, *394*, 223–230. [CrossRef] [PubMed]
14. Jia, J.; Liu, Z.; Xiao, X.; Liu, B.; Chou, K.-C. iSuc-PseOpt: Identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset. *Anal. Biochem.* **2016**, *497*, 48–56. [CrossRef]
15. Zhao, X.; Ning, Q.; Chai, H.; Ma, Z. Accurate in silico identification of protein succinylation sites using an iterative semi-supervised learning technique. *J. Theor. Biol.* **2015**, *374*, 60–65. [CrossRef]
16. Park, J.; Chen, Y.; Tishkoff, D.X.; Peng, C.; Tan, M.; Dai, L.; Xie, Z.; Zhang, Y.; Zwaans, B.M.; Skinner, M.E. SIRT5-mediated lysine desuccinylation impacts diverse metabolic pathways. *Mol. Cell* **2013**, *50*, 919–930. [CrossRef]
17. Tan, M.; Luo, H.; Lee, S.; Jin, F.; Yang, J.S.; Montellier, E.; Buchou, T.; Cheng, Z.; Rousseaux, S.; Rajagopal, N. Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. *Cell* **2011**, *146*, 1016–1028. [CrossRef]
18. Chandra, A.; Sharma, A.; Dehzangi, A.; Ranganathan, S.; Jokhan, A.; Chou, K.-C.; Tsunoda, T. PhoglyStruct: Prediction of phosphoglycerylated lysine residues using structural properties of amino acids. *Sci. Rep.* **2018**, *8*, 17923. [CrossRef]
19. Striebel, F.; Imkamp, F.; Özcelik, D.; Weber-Ban, E. Pupylation as a signal for proteasomal degradation in bacteria. *Biochim. et Biophys. Acta (BBA) Bioenerg.* **2014**, *1843*, 103–113. [CrossRef]
20. Hasan, M.M.; Khatun, M.S. Recent progress and challenges for protein pupylation sites prediction. *EC Proteom. Bioinform.* **2017**, *2*, 36–45.
21. Tamura, N.; Yun, H.; Tamura, T. Ubiquitin-like protein involved in proteasomal protein degradation in bacteria. *Seikagaku. J. Jpn. Biochem. Soc.* **2009**, *81*, 896.
22. Pearce, M.J.; Mintseris, J.; Ferreyra, J.; Gygi, S.P.; Darwin, K.H. Ubiquitin-like protein involved in the proteasome pathway of Mycobacterium tuberculosis. *Science* **2008**, *322*, 1104–1107. [CrossRef]
23. Sutter, M.; Striebel, F.; Damberger, F.F.; Allain, F.H.-T.; 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 Lett.* **2009**, *583*, 3151–3157. [CrossRef] [PubMed]
24. Burns, K.E.; Liu, W.-T.; Boshoff, H.I.; Dorrestein, P.C.; Barry, C.E. Proteasomal protein degradation in Mycobacteria is dependent upon a prokaryotic ubiquitin-like protein. *J. Biol. Chem.* **2009**, *284*, 3069–3075. [CrossRef] [PubMed]
25. Chen, X.; Solomon, W.C.; Kang, Y.; Cerda-Maira, F.; Darwin, K.H.; Walters, K.J. Prokaryotic ubiquitin-like protein pup is intrinsically disordered. *J. Mol. Biol.* **2009**, *392*, 208–217. [CrossRef] [PubMed]
26. Janssen, G.V.; Zhang, S.; Merkx, R.; Schiesswohl, C.; Chatterjee, C.; Darwin, K.H.; Ovaa, H. Discovery and Optimization of Inhibitors for the Pup Proteasome System in Mycobacterium tuberculosis. *bioRxiv* **2019**. [CrossRef]

27. Burns, K.E.; Darwin, K.H. Pupylation versus ubiquitylation: Tagging for proteasome-dependent degradation. *Cell. Microbiol.* **2010**, *12*, 424–431. [CrossRef]
28. Imkamp, F.; Striebel, F.; Sutter, M.; Özcelik, D.; Zimmermann, N.; Sander, P.; Weber-Ban, E. Dop functions as a depupylase in the prokaryotic ubiquitin-like modification pathway. *EMBO Rep.* **2010**, *11*, 791–797. [CrossRef]
29. Burns, K.E.; Cerda-Maira, F.A.; Wang, T.; Li, H.; Bishai, W.R.; Darwin, K.H. "Depupylation" of prokaryotic ubiquitin-like protein from mycobacterial proteasome substrates. *Mol. Cell* **2010**, *39*, 821–827. [CrossRef]
30. Barandun, J.; Delley, C.L.; Weber-Ban, E. The pupylation pathway and its role in mycobacteria. *BMC Biol.* **2012**, *10*, 1–9. [CrossRef]
31. Poulsen, C.; Akhter, Y.; Jeon, A.H.W.; Schmitt-Ulms, G.; Meyer, H.E.; Stefanski, A.; Stühler, K.; Wilmanns, M.; Song, Y.H. Proteome-wide identification of mycobacterial pupylation targets. *Mol. Syst. Biol.* **2010**, *6*, 386. [CrossRef] [PubMed]
32. 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. *Nat. Struct. Mol. Biol.* **2009**, *16*, 647–651. [CrossRef] [PubMed]
33. Liu, Z.; Ma, Q.; Cao, J.; Gao, X.; Ren, J.; Xue, Y. GPS-PUP: Computational prediction of pupylation sites in prokaryotic proteins. *Mol. BioSyst.* **2011**, *7*, 2737–2740. [CrossRef] [PubMed]
34. Zhao, X.; Zhang, J.; Ning, Q.; Sun, P.; Ma, Z.; Yin, M. Identification of protein pupylation sites using bi-profile Bayes feature extraction and ensemble learning. *Math. Probl. Eng.* **2013**, *2013*, 1–7. [CrossRef]
35. 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 Res. Int.* **2013**, *2013*, 1–9. [CrossRef] [PubMed]
36. Ju, Z.; Gu, H. Predicting pupylation sites in prokaryotic proteins using semi-supervised self-training support vector machine algorithm. *Anal. Biochem.* **2016**, *507*, 1–6. [CrossRef]
37. Hasan, M.M.; Zhou, Y.; Lu, X.; Li, J.; Song, J.; Zhang, Z. Computational identification of protein pupylation sites by using profile-based composition of k-spaced amino acid pairs. *PLoS ONE* **2015**, *10*, e0129635. [CrossRef]
38. Jiang, M.; Cao, J.-Z. Positive-Unlabeled learning for pupylation sites prediction. *BioMed Res. Int.* **2016**, *2016*, 1–5. [CrossRef]
39. Tung, C.-W. Prediction of pupylation sites using the composition of k-spaced amino acid pairs. *J. Theor. Biol.* **2013**, *336*, 11–17. [CrossRef]
40. 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]
41. Nan, X.; Bao, L.; Zhao, X.; Zhao, X.; Sangaiah, A.; Wang, G.-G.; Ma, Z. EPuL: An enhanced positive-unlabeled learning algorithm for the prediction of pupylation sites. *Molecules* **2017**, *22*, 1463. [CrossRef]
42. Bao, W.; You, Z.-H.; Huang, D.-S. CIPPN: Computational identification of protein pupylation sites by using neural network. *Oncotarget* **2017**, *8*, 108867–108879. [CrossRef] [PubMed]
43. Singh, V.; Sharma, A.; Chandra, A.; Dehzangi, A.; Shigemizu, D.; Tsunoda, T. Computational Prediction of Lysine Pupylation Sites in Prokaryotic Proteins Using Position Specific Scoring Matrix into Bigram for Feature Extraction. In Proceedings of the Public-Key Cryptography—PKC 2018, Rio De Janeiro, Brazil, 25–29 March 2018; Springer Science and Business Media LLC: Berlin/Heidelberg, Germany, 2019; pp. 488–500.
44. Tung, C.-W. PupDB: A database of pupylated proteins. *BMC Bioinform.* **2012**, *13*, 1–5. [CrossRef] [PubMed]
45. Chou, K.-C. Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.* **2011**, *273*, 236–247. [CrossRef]
46. Wang, Y.; Ding, Y.; Wen, H.; Lin, Y.; Hu, Y.; Zhang, Y.; Xia, Q.; Lin, Z. QSAR modeling and design of cationic antimicrobial peptides based on structural properties of amino acids. *Comb. Chem. High. Throughput Screen.* **2012**, *15*, 347–353. [CrossRef]
47. López, Y.; Dehzangi, A.; Lal, S.P.; Taherzadeh, G.; Michaelson, J.; Sattar, A.; Tsunoda, T.; Sharma, A. SucStruct: Prediction of succinylated lysine residues by using structural properties of amino acids. *Anal. Biochem.* **2017**, *527*, 24–32. [CrossRef]
48. Heffernan, R.; Paliwal, K.; Lyons, J.; Dehzangi, A.; Sharma, A.; Wang, J.; Sattar, A.; Yang, Y.; Zhou, Y. Improving prediction of secondary structure, local backbone angles and solvent accessible surface area of proteins by iterative deep learning. *Sci. Rep.* **2015**, *5*, 11476. [CrossRef] [PubMed]

49. Yang, Y.; Heffernan, R.; Paliwal, K.; Lyons, J.; Dehzangi, A.; Sharma, A.; Wang, J.; Sattar, A.; Zhou, Y. SPIDER2: A Package to Predict Secondary Structure, Accessible Surface Area, and Main-Chain Torsional Angles by Deep Neural Networks. In *Methods in Molecular Biology*; Springer Science and Business Media LLC: Berlin/Heidelberg, Germany, 2017; Volume 1484, pp. 55–63.
50. Lyons, J.; Dehzangi, A.; Heffernan, R.; Sharma, A.; Paliwal, K.; Sattar, A.; Zhou, Y.; Yang, Y. Predicting backbone Cα angles and dihedrals from protein sequences by stacked sparse auto-encoder deep neural network. *J. Comput. Chem.* **2014**, *35*, 2040–2046. [CrossRef]
51. Heffernan, R.; Dehzangi, A.; Lyons, J.; Paliwal, K.; Sharma, A.; Wang, J.; Sattar, A.; Zhou, Y.; Yang, Y. Highly accurate sequence-based prediction of half-sphere exposures of amino acid residues in proteins. *Bioinformatics* **2015**, *32*, 843–849. [CrossRef]
52. Ooi, T.; Oobatake, M.; Nemethy, G.; Scheraga, H.A. Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. *Proc. Natl. Acad. Sci. USA* **1987**, *84*, 3086–3090.
53. Wodak, S.J.; Janin, J. Analytical approximation to the accessible surface area of proteins. *Proc. Natl. Acad. Sci. USA* **1980**, *77*, 1736–1740.
54. Faraggi, E.; Zhang, T.; Yang, Y.; Kurgan, L.; Zhou, Y. SPINE X: Improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. *J. Comput. Chem.* **2012**, *33*, 259–267. [CrossRef]
55. McGuffin, L.J.; Bryson, K.; Jones, D.T. The PSIPRED protein structure prediction server. *Bioinformatics* **2000**, *16*, 404–405. [CrossRef] [PubMed]
56. Senior, A.W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Žídek, A.; Nelson, A.W.R.; Bridgland, A.; et al. Improved protein structure prediction using potentials from deep learning. *Nature* **2020**, *577*, 706–710. [CrossRef] [PubMed]
57. Kuhlman, B.; Bradley, P. Advances in protein structure prediction and design. *Nat. Rev. Mol. Cell Biol.* **2019**, *20*, 681–697. [CrossRef] [PubMed]
58. Fang, C.; Shang, Y.; Xu, D. Prediction of Protein Backbone Torsion Angles Using Deep Residual Inception Neural Networks. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2019**, *16*, 1020–1028. [CrossRef]
59. Xu, G.; Wang, Q.; Ma, J. OPUS-TASS: A protein backbone torsion angles and secondary structure predictor based on ensemble neural networks. *Bioinformatics* **2020**. [CrossRef]
60. Gao, J.; Yang, Y.; Zhou, Y. Grid-based prediction of torsion angle probabilities of protein backbone and its application to discrimination of protein intrinsic disorder regions and selection of model structures. *BMC Bioinform.* **2018**, *19*, 29. [CrossRef]
61. Li, H.; Hou, J.; Adhikari, B.; Lyu, Q.; Cheng, J. Deep learning methods for protein torsion angle prediction. *BMC Bioinform.* **2017**, *18*, 1–13. [CrossRef]
62. Sharma, R.; Raicar, G.; Tsunoda, T.; Patil, A.; Sharma, A. OPAL: Prediction of MoRF regions in intrinsically disordered protein sequences. *Bioinformatics* **2018**, *34*, 1850–1858. [CrossRef]
63. Reddy, H.M.; Sharma, A.; Dehzangi, A.; Shigemizu, D.; Chandra, A.A.; Tsunoda, T. GlyStruct: Glycation prediction using structural properties of amino acid residues. *BMC Bioinform.* **2019**, *19*, 55–64. [CrossRef]
64. Shamim, M.T.A.; Anwaruddin, M.; Nagarajaram, H. Support Vector Machine-based classification of protein folds using the structural properties of amino acid residues and amino acid residue pairs. *Bioinformatics* **2007**, *23*, 3320–3327. [CrossRef]
65. Pan, B.-B.; Yang, F.; Ye, Y.; Wu, Q.; Li, C.; Huber, T.; Su, X.-C. 3D structure determination of a protein in living cells using paramagnetic NMR spectroscopy. *Chem. Commun.* **2016**, *52*, 10237–10240. [CrossRef] [PubMed]
66. Lins, L.; Thomas, A.; Brasseur, R. Analysis of accessible surface of residues in proteins. *Protein Sci.* **2003**, *12*, 1406–1417. [CrossRef] [PubMed]
67. Tarafder, S.; Ahmed, T.; Iqbal, S.; Hoque, T.; Rahman, M. RBSURFpred: Modeling protein accessible surface area in real and binary space using regularized and optimized regression. *J. Theor. Biol.* **2018**, *441*, 44–57. [CrossRef] [PubMed]
68. Dehzangi, A.; López, Y.; Lal, S.P.; Taherzadeh, G.; Sattar, A.; Tsunoda, T.; Sharma, A. Improving succinylation prediction accuracy by incorporating the secondary structure via helix, strand and coil, and evolutionary information from profile bigrams. *PLoS ONE* **2018**, *13*, e0191900. [CrossRef] [PubMed]
69. Dehzangi, A.; López, Y.; Taherzadeh, G.; Sharma, A.; Tsunoda, T. SumSec: Accurate prediction of Sumoylation sites using predicted secondary structure. *Molecules* **2018**, *23*, 3260. [CrossRef]

70. Faraggi, E.; Yang, Y.; Zhang, S.; Zhou, Y. Predicting continuous local structure and the effect of its substitution for secondary structure in fragment-free protein structure prediction. *Structure* **2009**, *17*, 1515–1527. [CrossRef]
71. Xue, B.; Dor, O.; Faraggi, E.; Zhou, Y. Real-value prediction of backbone torsion angles. *Proteins Struct. Funct. Bioinform.* **2008**, *72*, 427–433. [CrossRef]
72. Dor, O.; Zhou, Y. Real-SPINE: An integrated system of neural networks for real-value prediction of protein structural properties. *Proteins: Struct. Funct. Bioinform.* **2007**, *68*, 76–81. [CrossRef]
73. Schumacher, T.N.; Mayr, L.M.; Minor, D.L.; Milhollen, M.A.; Burgess, M.W.; Kim, P.S. Identification of D-Peptide Ligands Through Mirror-Image Phage Display. *Science* **1996**, *271*, 1854–1857. [CrossRef]
74. Meinnel, T.; Dian, C.; Giglione, C. Myristoylation, an Ancient Protein Modification Mirroring Eukaryogenesis and Evolution. *Trends Biochem. Sci.* **2020**, *45*, 619–632. [CrossRef]
75. Guptasarma, P. Reversal of peptide backbone direction may result in the mirroring of protein structure. *FEBS Lett.* **1992**, *310*, 205–210. [CrossRef]
76. Meyer, D.; Leisch, F.; Hornik, K. *Benchmarking Support Vector Machines*; WU Vienna University of Economics and Business: Vienna, Austria, 2002.
77. Mangasarian, O.L.; Musicant, D.R. Active support vector machine classification. In *Proceedings of Advances in Neural Information Processing Systems*; Neural Information Processing Systems (NIPS): Denver, CO, USA, 2001; pp. 577–583.
78. Byvatov, E.; Schneider, G. Support vector machine applications in bioinformatics. *Appl. Bioinform.* **2003**, *2*, 67–77.
79. Wang, J.-Y. *Application of Support. Vector Machines in Bioinformatics*; National Taiwan University: Taipei, Taiwan, 2002.
80. Huang, S.; Cai, N.; Pacheco, P.P.; Narrandes, S.; Wang, Y.; Xu, W. Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. *Cancer Genom. Proteom.* **2018**, *15*, 41–51. [CrossRef]
81. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2006.
82. Amari, S.-i.; Wu, S. Improving support vector machine classifiers by modifying kernel functions. *Neural Netw.* **1999**, *12*, 783–789. [CrossRef]
83. Tharwat, A. Parameter investigation of support vector machine classifier with kernel functions. *Knowl. Inf. Syst.* **2019**, *61*, 1269–1302. [CrossRef]
84. Control, C.F.D. Prevention, Antibiotic Resistance. US Department of Health & Human Services. 2013. Available online: https://www.cdc.gov/drugresistance/about.html (accessed on 3 July 2019).
85. Bao, W.; Jiang, Z. Prediction of Lysine Pupylation Sites with Machine Learning Methods. In *Proceedings of the International Conference on Intelligent Computing*; Springer: Cham, Switzerland, 2017; pp. 408–417.
86. Hajisharifi, Z.; Piryaiee, M.; Beigi, M.M.; Behbahani, M.; Mohabatkar, H. Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test. *J. Theor. Biol.* **2014**, *341*, 34–40. [CrossRef] [PubMed]
87. Schneider, J.P.; Lombardi, A.; DeGrado, W.F. Analysis and design of three-stranded coiled coils and three-helix bundles. *Fold. Des.* **1998**, *3*, R29–R40. [CrossRef]
88. Ausaf Ali, S.; Hassan, I.; Islam, A.; Ahmad, F. A Review of Methods Available to Estimate Solvent-Accessible Surface Areas of Soluble Proteins in the Folded and Unfolded States. *Curr. Protein Pept. Sci.* **2014**, *15*, 456–476. [CrossRef]