ECRONICON OPEN ACCESS

EC PROTEOMICS AND BIOINFORMATICS
Review Article

Recent Progress and Challenges for Protein Pupylation Sites Prediction

Md Mehedi Hasan[1]* and Mst Shamima Khatun[2]

[1]Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, Kawazu, Iizuka, Fukuoka, Japan
[2]Laboratory of Bioinformatics, Department of Statistics, University of Rajshahi, Rajshahi, Bangladesh

***Corresponding Author**: Md Mehedi Hasan, Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, Kawazu, Iizuka, Fukuoka, Japan.

Received: October 23, 2017; **Published:** November 03, 2017

Abstract

Lysine pupylation is a type of post-translational modification of protein that contributes to the cellular function in microbial organisms. Identification of pupylation sites is an important step for understanding the function of microbial proteins. Nowadays, *in silico* approaches for identifying the potential pupylation sites becomes gradually popular, due to various limitations of experimental methods. The purpose of this review is to discuss the recent progress in the prediction of protein pupylation sites from the published methods, datasets, and online resources. We discussed the challenges and limitations for future endeavors to develop novel tools. We also deduced why species-species classifier is necessary to predict pupylation substrates. Therefore, this review would be a useful guideline for understanding the importance of pupylation sites prediction.

Keywords: *Pupylation Sits Prediction; Machine Learning; Feature Encoding*

Introduction

The prokaryotic ubiquitin-like protein (Pup) is a small protein transformer related to the post-translational modifications (PTMs) and similar manner of ubiquitin. The Pup is covalently attached to a lysine that contributes target protein for proteosomal degradation by forming isopeptide bonds [1-4], in the tagging system referred as pupylation. To date, the Pup homologs has been presented in bacteria by the orders of Nitrospirales and Actinomycetales species [5,6]. The microbial Pup gene has been identified [3,7], but the function of Pup is not fully identified in prokaryotes until recently [2,3,8]. Although, the mark of proteasome Pup degradation has been promptly accumulating in both of the *in vivo* [9] and *in vitro* [10,11] systems.

Pupylation and ubiquitylation are functionally analogous in distinct pathway by a chemically [12,13]. In microbial species, pupylation process involves two homologous sequential action but their enzymology is different. First, the Dop (deamidase of Pup) enzyme is deamidated the C-terminal glutamine of Pup to glutamate [12,14]. Then A (PafA) catalyzes enzyme makes a formation with isopeptide bond between the side chain by attaching to the specific lysine [15,16]. The type of covalent bonds and reaction series differ between the pathways of pupylation and ubiquitylation [17]. The identification of pupylation sites will be an essential foundation for revelation of the mechanism and function of protein pupylation. A number of proteomic experimental technologies have been performed to identify lysine pupylated proteins based on the molecular signature of pupylated sites [18-22]. For systematically investigating of the protein pupylation and its relevant function, a prerequisite is needed to establish a reliable and comprehensive dataset. However, until now a vast number pupylation sites have been remain undiscovered. Due to the experimental verification of pupylated substrates is time-consuming, labor-intensive and biased toward the abundant proteins. Thus, in computational prediction of protein pupylation sites can be served as an alternative strategy for whole proteome annotation.

To date, a few numbers of computational methods for prediction pupylation site has been established [23-27]. Xue., *et al*. proposed a computational predictor GPS-PUP [25], which was used a Group-based Prediction System (GPS) sequence encoding, including motif length selection, weight training, and matrix mutation to improve the performances'. Xiaowei., *et al*. proposed a computational prediction EnsemblePup [28]. It was utilized the Bi-profile Bayes feature extraction as an encoding scheme with Support Vector Machine (SVM) classifier. Xiaowei., *et al*. proposed another computational predictor PrePup [23]. It was based on multiple feature encoding such as amino acid index property (AAindex), position-specific scoring matrix (PSSM) conservation scores, structural disorder score, secondary structure, solvent accessibility, and feature space with a SVM classifier. Tung Chun-Wei developed a SVM based predictor iPUP, by exploiting a single sequence encoding, i.e. composition of k-spaced amino acid pair (CKSAAP) [24]. To train the classifier SVM together with a backward feature selection method was used. The CKSAAP is a broadly used algorithm in protein bioinformatics [29-31]. Chen., *et al*. [26] developed another SVM-based predictor PupPred. This predictor showcase variety of features including binary features, physicochemical properties, amino acid pairs, protein secondary structures, PSSM and with a k-nearest neighbor algorithm. The authors demonstrated that the encoding of amino acid pairs and the implementation of F-measures for feature selection with the SVM-based classifier contributed to the improved performance of PupPred. We developed pbPUP predictor for predicting pupylation site based on the profile-based composition of k-spaced amino acid pair (pbCKSAAP) encoding with SVM classifier [27]. The pbCKSAAP is also widely used method in protein bioinformatics research [32,33]. Jiang and Cao developed a predictor PUL-PUP using positive-unlabeled learning with a SVM algorithm [30]. Ju., *et al*. established another predictor IMP-PUP based on semi-supervised self-training with SVM algorithm [31]. Recently, Nan., *et al.* developed a predictor EPul based on an enhanced positive-unlabeled learning algorithm [34]. All the available predictors datasets were collected from the PupDB database [35]. Our developed pbPUP predictor achieved an overall performance improvement in comparison to several other predictors on a comprehensive independent test set. Although, in predicting of pupylation sites the significant progress has been achieved, there has still room for performance improvement.

Notwithstanding the accessibility of various prediction tools of pupylation site, an important issue is comprehensively evaluating the comparing performances and the weaknesses and strengths of the tools. The above tools have also some limitations when applied to whole proteomes species as a training model. The most important issue is that the regulation mechanism of pupylation can differs between species of prokaryotic proteins. Therefore, the structural or sequences patterns surrounding the pupylation sites may significantly differ in different prokaryotic species. However, all of the existing tools of prediction pupylation sites disregarded the differences between species by considering combined all species as a generic predictor to build a simplified model. Therefore, to generate more accurate models for the efficient identification of species-specific pupylation sites predictor is necessary and perquisite.

The aim of this review is to provide informative and practical observations about more accurate prediction of protein pupylation. We discussed which tool serves the best performance, in which aspect the existing predictors can be enhanced, as well as the most significant features contribute to the prediction. We also discussed why species-species predictor is necessary to predict pupylation sites. In general, our aimed to examine: whether a universal best predictor exists that can be used to prediction of pupylated proteins.

Materials and Methods

A brief flowchart of computational framework in prediction of protein pupylation sites is shown figure 1.

Data collection and preprocessing

Experimentally verified five-species (i.e. *M. smegmatis, M. tuberculosis, E. coli, C. glutamicum*, and *R. erythropolis*) pupylation datasets were collected from a popular pupylation site database [35]. At first, with a 30% identity cutoff the sequence redundancy was removed in the datasets using CD-HIT [36]. Experimentally examined pupylated lysine residues were regarded as pupylated sites (i.e. positive samples); whether other lysine residues were regarded as non-pupylated sites (i.e. negative samples). Then 1:2 (pupylated vs. non-pupylated) non-pupylated samples were randomly pooled from the remaining lysine residues. In the remaining lysine residues that have not yet been verified as pupylation sites, which could contain pupylated sites. The numbers of pupylated proteins and pupylated sites for each dataset are presented in table 1.

Figure 1: *The flowchart of protein pupylation site prediction.*

Dataset	Number of pupylated Proteins	Number of pupylation Sites
M. smegmatis	75	84
M. tuberculosis	55	60
E. coli	51	69
C. glutamicum	55	65
R. erythropolis	31	31
Total	267	309

Table 1: *Statistics of the pupylated protein and pupylation sites used in this study.*

Prediction techniques under assessment

In this review our main principle to include an algorithm in the comparison analysis is that such method has been executed as either an online implementation or performance of corresponding features. Until now, nine predictors have been established for analyzing pupylation proteins: GPS-PUP [25], EnsemblePup [28], PrePup [23], iPUP [24], PupPred [23], pbPUP [27], PUL-PUP [30], IMP-PUP [31], and EPuL [34]. The all of existing prediction models analyzed pupylation proteins by using a general predictor model, i.e. combined the existing experimentally verified pupylation proteins. Among the nine existing perdition model, GPS-PUP analyzed the pupylation proteins by group based prediction algorithm, EnsemblePup used Bi-Profile Bayes and other 7 models were used SVM. More information about these exiting methods is summarized in table 2.

Tools	Web-server	Working server	Algorithm	Dataset size (Pupylation sites/proteins)	Ratio of Training set Neg. vs Pos.	Ratio of Independent test Neg. vs Pos.	Window size	Time for processing a sequence
GPS-PUP	http://pup.biocuckoo.org	Yes	GPS	127/109	1:total	-	15	Within 10 second
Ensemble Pup	http://210.47.24.217:8080/EnsemblePup/	No	Bi-profile Bayes	127/109	1:3	-	17	-
PrePup	http://210.47.24.217:8080/PrePup/	No	SVM	127/109	1:2	1:total	21	-
iPUP	http://cwtung.kmu.edu.tw/ipup	Yes	SVM	215/182	1:total	1:total	25	Within 20 seconds
PupPred	http://bioinfo.ncu.edu.cn/PupPred.aspx	No	SVM	215/182	1:1	-	27	-
pbPUP	http://protein.cau.edu.cn/pbPUP/	Yes	SVM	275 / 237	1:2	1:total	57	Within 5 minutes
PUL-PUP	http://59.73.198.144:8080/EPuL	No	SVM	162/183	1:1	1:total	21	-
IMP-PUP	https://juzhe1120.github.io/	No	SVM	162/183	1:total	1:total	21	-
EPuL	http://59.73.198.144:8080/EPuL	No	SVM	162/183	1:1	1:total	21	-

Table 2: *Summary of pupylation site prediction tools compared in this study.*

Model evaluation

To evaluate the performances of existing methods, four widely used measurements were considered including specificity (Sp), accuracy (Ac), sensitivity (Sn) and Matthews correlation coefficient (MCC). The following formulas are used for calculating the Sn, Sp, Ac, and MCC.

$$Sn = nTP/(nTP+nFN) \quad (1)$$

$$Ac = (nTP+nTN)/(nTP+nTN+nFP+nFN) \quad (2)$$

$$Sp = nTN/(nTN+nFP) \quad (3)$$

$$MCC = (nTP\times nTN-nFP\times nFN)/\sqrt{((nTN+nFN)\times(nTP+nFP)\times(nTP+nFN)\times(nTN+nFP))} \quad (4)$$

where, nTP, nFP, nFN and nTN represent the numbers of true positives, false positives, false negatives and, true negatives respectively.

Results and Discussion

Performance comparison of different prediction methods

The majority of the existing algorithms cited in this review used protein sequence information, evolutionary information and some residue properties and/or structural properties. We compared the predictive performances of different pupylation site predictors, including GPS-PUP [25], iPUP [24], PupPred [23], and our previous predictor pbPUP [27]. The exhaustive comparison of the predictive results obtained from different schemes is almost impossible because they use different training and testing samples, different positive and negative samples and different assessment procedures. Performance comparison is further complicated because many methods are not publicly available such as, EnsemblePup [28], PrePup [23], PUL-PUP [30], IMP-PUP [31], and EPuL [34]. Performance comparison with existing predictors, 71 pupylated proteins containing 86 pupylation and 1136 putative non-pupylation sites that constitute an independent dataset was used. Among these proteins, 20 proteins were extracted from iPUP [24] and 51 proteins were retrieved from a recently published article [20]. Currently, there exist four computational predictors to predict pupylation lysine sites, which are iPUP, GPS-PUP, PupPred and pbPUP. These predictors employed different training datasets for predicting pupylation sites. The independent dataset was used for making a fair comparison of the performance of these different predictors. As shown in table 3, the pbPUP predictor achieved improved performance than other exiting predictors. We found that all the existing predictors performances were very lower (Table 3). A possible reason is that existing computational tools are developed as a generic model by combining the data of all species. From the comparison results, we conclude that across all species presently no universal generic best predictor exists for pupylation site prediction.

Predictor	Threshold	Ac (%)	Sn (%)	Sp (%)	MCC (%)
GPS-PUP	High	83.89	19.76	88.74	6.73
iPUP	High	81.13	29.06	84.90	9.56
PupPred	High	88.93	9.19	94.77	4.33
pbPUP	High	82.87	30.13	88.56	13.97

Table 3: *The prediction performance of existing tools on the independent test dataset.*

Species-specific pupylation site analysis

The patterns of sequence surrounding the pupylation sites in the 5 species datasets could be partly explained while missing performances in existing generic pupylation site predictors. Initially, we investigated the two sample logo to analyze sequence patterns information for determining statistically significant amino acid surrounding pupylation sites, based on five pupylation families: M. smegmatis, *M. tuberculosis, E. coli, C. glutamicum*, and *R. erythropolis*. The graphical sequence logo representations showing the distinct patterns or conserved sequence motifs between pupylation and candidate non-pupylation sites (Figure 2). From the sequence logo, we observed that, a number of amino acid residues that are significantly enriched around pupylation sequences. We also find that pupylation sequences are very dissimilar in five different species. For example, in sequence logo at position +1 and -1 residues were different in all of five species

(Figure 2). Another example is that, 'P' (proline) at position +3 only found in *E. coli* sequence, whereas residue is not favored in other species sequence. 'R' (arginine) residue tends enriched residues sequence, whereas *C. glutamicum*, and *R. erythropolis* had not any depleted residues. The sequence logo suggested that pupylation and candidate non-pupylation fragments have a considerable difference among the species sequence. Altogether, the result highlights the necessity of precise candidate pupylation site recognition by developing species-specific predictors.

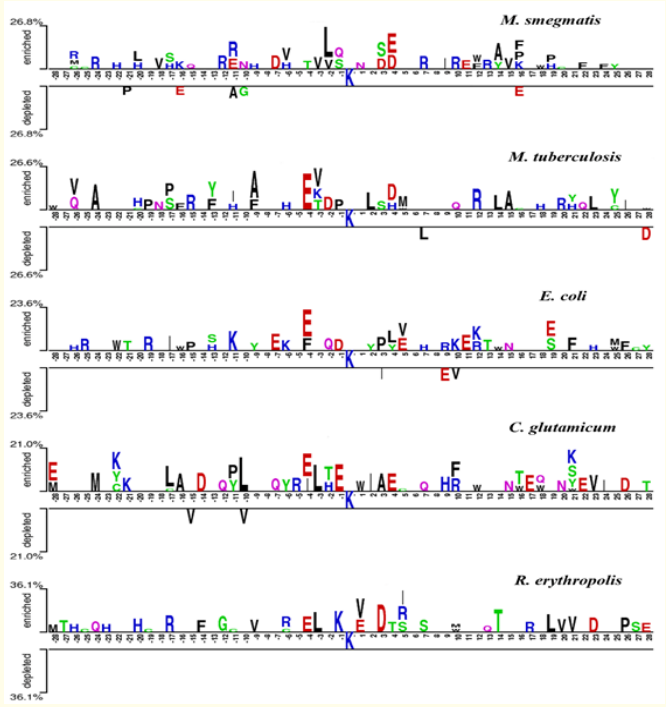

Figure 2: *The Two-Sample-Logo representation of position-specific residue composition surrounding the pupylation sites and non-pupylation sites, based on five pupylation families: M. smegmatis, M. tuberculosis, E. coli, C. glutamicum, and R. erythropolis. It has also showed that for the position between pupylated and non-pupylated residues in above five pupylation families the compositional amino acids had no significance differences, especially those located in the positions of ~-28 to -1 and +1 to +28.*

Furthermore, we conducted a chi-square goodness of fit test to classify the amino acid residues in the different pupylation sequences of the five examined datasets. For the 5 species including, *M. smegmatis, M. tuberculosis, E. coli, C. glutamicum*, and *R. erythropolis*, the total number of collected pupylation sites were 84, 60, 69, 65 and 31, respectively. The occurrences of amino acid residues at different fragment positions (~-5 to +5) with p-values were calculated and corrected by Bonferroni test (Table S1). We found that for calculating amino acid frequencies at each fragment window positions, a p-value of lower than 0.01 indicates that the amino acids of the 5 species-specific pupylation sequences are significantly different. Both Figure 2 and statistical test yielded significant differences in the sequence patterns of the 5 species around pupylation samples. Therefore, we recommend the scientists to make an accurate species-specific classifier to identify pupylation substrates.

Position	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
A	1.63E-29	2.23E-52	1.29E-39	3.12E-21	2.24E-37	7.73E-33	7.23E-87	6.03E-03	1.39E-23	6.19E-16
C	9.19E-128	2.42E-176	3.19E-29	N/A	6.28E-28	4.11E-203	1.82E-99	1.04E-252	3.15E-71	1.69E-11
D	5.14E-13	5.31E-53	N/A	1.81E-05	2.28E-176	6.91E-101	3.18E-139	5.01E-123	4.29E-03	3.94E-38
E	1.29E-16	3.28E-17	2.21E-33	4.65E-68	3.46E-58	1.19E-13	5.03E-71	9.81E-12	1.25E-22	3.53E-17
F	3.71E-71	3.39E-25	4.19E-37	2.15E-77	2.78E-16	7.49E-3156	4.01E-25	3.21E-75	3.21E-18	8.21E-42
G	7.10E-15	1.19E-39	5.71E-53	5.05E-37	7.06E-36	3.51E-18	8.61E-78	1.04E-49	6.43E-39	N/A
H	1.48E-204	N/A	3.14E-183	2.12E-23	5.27E-205	N/A	N/A	N/A	3.57E-27	7.11E-103
I	2.18E-37	2.76E-24	6.05E-44	6.65E-54	4.35E-102	5.49E-27	8.22E-69	4.58E-82	2.84E-48	3.19E-09
K	1.6E-105	8.51E-14	3.84E-87	5.12E-104	7.42E-66	7.11E-231	7.11E-48	1.96E-131	6.30E-135	4.11E-23
L	3.31E-21	4.91E-25	8.29E-18	1.13E-39	8.14E-39	2.69E-114	4.08E-59	1.62E-09	8.81E-09	9.13E-03
M	N/A	3.71E-101	6.77E-19	2.17E-107	9.15E-28	N/A	N/A	N/A	1.81E-54	1.71E-06
N	8.71E-22	6.03E-31	1.07E-69	3.46E-217	2.26E-208	8.04E-27	3.02E-39	2.66E-78	5.40E-41	8.13E-13
P	4.09E-08	4.34E-61	5.30E-27	3.69E-62	6.58E-177	9.31E-109	8.11E-14	4.08E-19	6.14E-22	7.53E-11
Q	1.32E-17	2.10E-24	3.79E-91	7.85E-15	1.86E-14	1.18E-31	7.81E-16	1.49E-46	3.17E-21	1.09E-104
R	2.62E-71	3.94E-26	3.39E-211	4.48E-26	3.03E-176	4.23E-88	2.37E-17	1.79E-34	2.48E-41	7.12E-17
S	2.71E-61	5.21E-28	2.82E-13	5.31E-18	7.57E-07	1.84E-59	8.87E-42	3.11E-42	2.19E-18	8.81E-05
T	5.01E-17	5.01E-53	2.87E-53	2.41E-69	4.02E-93	5.03E-68	3.58E-29	2.26E-19	7.19E-15	3.08E-48
V	2.01E-73	1.06E-11	4.01E-24	2.41E-53	1.50E-55	3.56E-19	1.46E-18	4.43E-08	4.43E-09	9.04E-03
W	2.08E-115	2.74E-113	1.16E-23	5.04E-15	6.13E-256	N/A	1.72E-16	615E-107	5.01E-29	5.12E-31
Y	9.43E-36	7.41E-39	4.54E-58	2.41E-16	9.87E-26	8.08E-28	5.83E-158	4.47E-41	1.65E-156	3.29E-22

Supplementary Table S1: *The p-values were calculated using Chi-square test and corrected by Bonferroni for the amino acid occurrence frequencies at each window positions (~-5 to +5) for pupylated sequences. Five model species includes M. smegmatis, M. tuberculosis, E. coli, C. glutamicum, and R. erythropolis.*

[a]*The N/A indicates among the 5 species at least one corresponding amino acid missing on the sequence fragments.*

The online implementation services

A user friendly public interface web implementation or a downloadable software package is essential for users. As listed in table 1, there were only 9 predictors provided online implementations along with their research publication. However, some of prediction tools with website are not accessible to users for unknown reasons, especially for the output formats. In particular, we compared the exiting tools using the following criteria: (i) whether the web implementation supports batch sequences prediction; (ii) whether the method prediction has probability scores; and (iii) restrictions of the exiting implements. The comparison performances are summarized in table 1. Among the existing tools EnsemblePup [28], PrePup [23], PUL-PUP [30], IMP-PUP [31], and EPuL [34] did not provide web-implementation to implement their algorithm. The GPS-PUP [25] did not provide the information of flanking window positions, predicted probability scores, and cutoff thresholds. The iPUP [24] server did not include the prediction pupylation scores in the output page. Users of pbPUP [27] can submit protein sequences in RAW or FASTA format. The processing time of pbPUP was < 5 minutes for one sequence, which was slightly longer time than the existing tools. The prediction output of pbPUP contains 4 items: protein name, residue position, prediction score and annotation of pupylation site with a text format. To the user perspective of the output results of the pupylation tool should include at least the position of the predicted pupylation site, flanking window positions and assessment scores or probability of the predicted pupylation site. Additionally, it is required that the predictor supports stringency modification from output of the developed software's. Especially for large-scale predictions, user control of the prediction inflexibility is important, because users are interested in predictions with above a certain confidence threshold.

Biological and functional aspects: pupylation site prediction

The biological function of pupylation should be context specific. The position of pupylation site on the protein sequence (and thus protein structure) can decide the function of a pupylation site. For example, the function of a pupylation site on the enzyme's activity site should be different from a site on the protein-protein interaction interface. That's why we need to know the position of pupylation sites. Many pupylation sites are functionally unimportant because they are not associated with other important functional sites. On the other hand, genetic screening result in some mutations with altered phenotypes. The explanation of the mechanism underlying these mutations sometimes could be hard if we only consider the typical functional sites (e.g. enzyme activity site). The change in pupylation sites could be one interesting clue, especially for some proteins from signaling pathways. In general, the pupylation site information can enrich the functional annotation of protein sequences.

In pupylation analysis, the consensus motif sequence is calculated based on the most frequent residues, either each nucleotide or amino acid, originate at each position in a sequence fragments. Pupylation substrates are susceptible to reversible and dynamic modifications of proteins during protein biosynthesis. Some substrates do not exhibit any significant consensus motif. Even for the one with consensus motif, the prediction accuracy by using the consensus motif alone is not satisfactory. The reason is the consensus motif is a simplified presentation of binary-encoding-based classifier, which ignores many details. For example, if one sequence position has 70% L, 25% of A and 5% of F, the significant consensus motif usually gives "L or A". Two facts were ignored here: 1) L is much more likely than A. 2) It could be F, though not such often. Therefore, for *in silico* analysis of pupylation sites, it is a problem if we test specific-species dataset using a different species model. In this case false discovery rate could increase, because their sequence patterns are different each other.

Future perspectives

An accurate prediction of pupylation substrate requires detailed information of the structures and functions of pupylated proteins. There are still remaining many issues need to be resolved. To assist knowledge discoveries through intensive analysis of vast amounts of pupylation data, further improvements are required. The followings are important perspectives for pupylation sites prediction. At first, the sequences or structural patterns around the pupylation sites may significantly differ in different species. However, all of the existing predictors pupylation disregarded the differences between species by considering combined all species as a generic model to build a simplified prediction model. Therefore, the next generation of computational methods needs to generate more accurate models for the efficient identification of species-specific prediction sites. Secondly, the performance of prediction is acceptable, but there is still more

room to improve. With the growth of lysine pupylation data, more robust prediction tools need to be developed. To further improve the prediction accuracy of lysine pupylation sites, scientists need to develop new tools, including the introduction of new classification algorithms and new features. Finally, all the current lysine pupylation site predictors are developed based merely on sequence information. With the increase of pupylation site data whose structures are known, we might take structural-based protein pupylation site analyses and forecasts into account for more comprehensive understanding of protein pupylation site patterns. We therefore anticipate that better prediction methods of pupylation site with improved performance will continue to emerge as increasing amounts of pupylation data.

Conclusions

In this review, the major observations from our analysis are first, across all species no generic best predictor exists for predicting pupylation sites. Secondly, to predict potential pupylation sites in different species scientist should make species-specific classifiers. Finally, the performance of the prediction tools developed is acceptable, but the prediction performance can be further improved by integrating different sequence encoding schemes. Altogether, in living cells, combining computational and tentative methods will certainly accelerate the accumulation of our knowledge on protein lysine pupylation.

Acknowledgments

This work was supported by a Grant-in-Aid for Scientific Research (B) (16H02898) and Grant-in-Aid for Young Scientists (B) (26870432) from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

Bibliography

1. Burns KE and Darwin KH. "Pupylation versus ubiquitylation: tagging for proteasome-dependent degradation". *Cellular Microbiology* 12.4 (2010): 424-431.

2. Ikeda F and Dikic I. "Atypical ubiquitin chains: new molecular signals. 'Protein Modifications: Beyond the Usual Suspects' review series". *EMBO Reports* 9.6 (2008): 536-542.

3. Pearce MJ., *et al*. "Ubiquitin-like protein involved in the proteasome pathway of Mycobacterium tuberculosis". *Science* 322.5904 (2008): 1104-1107.

4. Agarwal KL., *et al*. "Feline gastrin. An example of peptide sequence analysis by mass spectrometry". *Journal of the American Chemical Society* 91.11 (1969): 3096-3097.

5. Iyer LM., *et al*. "Unraveling the biochemistry and provenance of pupylation: a prokaryotic analog of ubiquitination". *Biology Direct* 3 (2008): 45.

6. Jastrab JB., *et al*. "An adenosine triphosphate-independent proteasome activator contributes to the virulence of Mycobacterium tuberculosis". *Proceedings of the National Academy of Sciences of the United States of America* 112.14 (2015): E1763-E1772.

7. Tamura N., *et al*. "[Ubiquitin-like protein involved in proteasomal protein degradation in bacteria]". *Seikagaku The Journal of Japanese Biochemical Society* 81.10 (2009): 896-899.

8. Burns KE and Darwin KH. "Pupylation: A Signal for Proteasomal Degradation in Mycobacterium tuberculosis". *Sub-Cellular Biochemistry* 54 (2010): 149-157.

9. Sutter M., *et al*. "A distinct structural region of the prokaryotic ubiquitin-like protein (Pup) is recognized by the N-terminal domain of the proteasomal ATPase Mpa". *FEBS Letters* 583.19 (2009): 3151-3157.

10. Kraut DA and Matouschek A. "Pup grows up: in vitro characterization of the degradation of pupylated proteins". *The EMBO Journal* 29.7 (2010): 1163-1164.

11. Imkamp F., *et al*. "Dop functions as a depupylase in the prokaryotic ubiquitin-like modification pathway". *EMBO Reports* 11.10 (2010): 791-797.

12. Striebel F., *et al*. "Bacterial ubiquitin-like modifier Pup is deamidated and conjugated to substrates by distinct but homologous enzymes". *Nature Structural and Molecular Biology* 16.6 (2009): 647-651.

13. Hecht N and Gur E. "Development of a fluorescence anisotropy-based assay for Dop, the first enzyme in the pupylation pathway". *Analytical Biochemistry* 485 (2015): 97-101.

14. Yun HY., *et al*. "Rhodococcus prokaryotic ubiquitin-like protein (Pup) is degraded by deaminase of pup (Dop)". *Bioscience, Biotechnology, and Biochemistry* 76.10 (2012): 1959-1966.

15. Sutter M., *et al*. "Prokaryotic ubiquitin-like protein (Pup) is coupled to substrates via the side chain of its C-terminal glutamate". *Journal of the American Chemical Society* 132.16 (2010): 5610-5612.

16. Guth E., *et al*. "Mycobacterial ubiquitin-like protein ligase PafA follows a two-step reaction pathway with a phosphorylated pup intermediate". *The Journal of Biological Chemistry* 286.6 (2011): 4412-4419.

17. Maupin-Furlow JA. "Prokaryotic ubiquitin-like protein modification". *Annual Review of Microbiology* 68 (2014): 155-175.

18. Cerda-Maira FA., *et al*. "Reconstitution of the Mycobacterium tuberculosis pupylation pathway in Escherichia coli". *EMBO Reports* 12.8 (2011): 863-870.

19. Festa RA., *et al*. "Prokaryotic ubiquitin-like protein (Pup) proteome of Mycobacterium tuberculosis". *PloS one* 5.1 (2010): e8589.

20. Kuberl A., *et al*. "Pupylated proteins in Corynebacterium glutamicum revealed by MudPIT analysis". *Proteomics* 14.12 (2014): 1531-1542.

21. Watrous J., *et al*. "Expansion of the mycobacterial "PUPylome"". *Molecular BioSystems* 6.2 (2010): 376-385.

22. Poulsen C., *et al*. "Proteome-wide identification of mycobacterial pupylation targets". *Molecular Systems Biology* 6 (2010): 386.

23. Zhao X., *et al*. "Position-specific analysis and prediction of protein pupylation sites based on multiple features". *BioMed Research International* (2013): 109549.

24. Tung CW. "Prediction of pupylation sites using the composition of k-spaced amino acid pairs". *Journal of Theoretical Biology* 336 (2013): 11-17.

25. Liu Z., *et al*. "GPS-PUP: computational prediction of pupylation sites in prokaryotic proteins". *Molecular Biosystems* 7.10 (2011): 2737-2740.

26. Chen X., *et al*. "Systematic analysis and prediction of pupylation sites in prokaryotic proteins". *PloS one* 8.9 (2013): e74002.

27. Hasan MM., *et al*. "Computational Identification of Protein Pupylation Sites by Using Profile-Based Composition of k-Spaced Amino Acid Pairs". *PloS one* 10.6 (2015): e0129635.

28. Zhao XW., *et al*. "Identification of Protein Pupylation Sites Using Bi-Profile Bayes Feature Extraction and Ensemble Learning". *Mathematical Problems in Engineering* (2013).

29. Hasan MM., *et al*. "SuccinSite: a computational tool for the prediction of protein succinylation sites by exploiting the amino acid patterns and properties". *Molecular Biosystems* 12.3 (2016): 786-795.

30. Jiang M and Cao JZ. "Positive-Unlabeled Learning for Pupylation Sites Prediction". *BioMed Research International* (2016): 4525786.

31. Ju Z and Gu H. "Predicting pupylation sites in prokaryotic proteins using semi-supervised self-training support vector machine algorithm". *Analytical Biochemistry* 507 (2016): 1-6.

32. Hasan MM., *et al*. "A systematic identification of species-specific protein succinylation sites using joint element features information". *International Journal of Nanomedicine* 12 (2017): 6303-6315.

33. Hasan MM., *et al*. "Computational identification of protein S-sulfenylation sites by incorporating the multiple sequence features information". *Molecular BioSystems* (2017).

34. Nan X., *et al*. "EPuL: An Enhanced Positive-Unlabeled Learning Algorithm for the Prediction of Pupylation Sites". *Molecules* 22.9 (2017): E1463.

35. Tung CW. "PupDB: a database of pupylated proteins". *BMC Bioinformatics* 13 (2012): 40.

36. Fu L., *et al*. "CD-HIT: accelerated for clustering the next-generation sequencing data". *Bioinformatics* 28.23 (2012): 3150-3152.

Volume 2 Issue 1 November 2017