Send Orders for Reprints to reprints@benthamscience.net

RESEARCH ARTICLE

Computational Identification of Lysine Glutarylation Sites Using Positive-Unlabeled Learning

Zhe Ju[1,*] and Shi-Yun Wang[1]

[1]*College of Science, Shenyang Aerospace University, Shenyang 110136, P.R. China*

ARTICLE HISTORY

Received: December 25, 2019
Revised: April 12, 2020
Accepted: April 13, 2020

DOI:
10.2174/1389202921666200511072327

Abstract: ***Background*:** As a new type of protein acylation modification, lysine glutarylation has been found to play a crucial role in metabolic processes and mitochondrial functions. To further explore the biological mechanisms and functions of glutarylation, it is significant to predict the potential glutarylation sites. In the existing glutarylation site predictors, experimentally verified glutarylation sites are treated as positive samples and non-verified lysine sites as the negative samples to train predictors. However, the non-verified lysine sites may contain some glutarylation sites which have not been experimentally identified yet.

***Methods*:** In this study, experimentally verified glutarylation sites are treated as the positive samples, whereas the remaining non-verified lysine sites are treated as unlabeled samples. A bioinformatics tool named PUL-GLU was developed to identify glutarylation sites using a positive-unlabeled learning algorithm.

***Results*:** Experimental results show that PUL-GLU significantly outperforms the current glutarylation site predictors. Therefore, PUL-GLU can be a powerful tool for accurate identification of protein glutarylation sites.

***Conclusion*:** A user-friendly web-server for PUL-GLU is available at http://bioinform.cn/pul_glu/.

Keywords: Post-translational modification, glutarylation, support vector machine, positive-unlabeled learning, protein acylation, site predictors.

1. INTRODUCTION

Protein post-translational modifications (PTMs) are crucial steps in protein synthesis and regulate various biological processes such as protein signaling, localization, and degradation. Among the various types of PTMs, acetylation, succinylation, malonylation, 2-hydroxyisobutyrylation, butyrylation, crotonylation, *etc*., can all occur at the ε-amino groups of specific lysine residues [1-6] and are known as lysine acylation modification. Recently, Tan *et al.* [7] discovered a new type of lysine acylation modification, named glutarylation, which is found in both prokaryotic and eukaryotic cells. Lysine glutarylation is a dynamic and evolutionarily conserved modification process, in which a glutaryl group attaches to specific lysine residues of a substrate protein. Similar to succinylation and acetylation, lysine glutarylation has been found to play a crucial role in metabolic processes and mitochondrial functions, such as fatty acid metabolism, amino acid metabolism and cellular respiration [6, 7]. Previous studies have shown that glutarylation of carbamoyl phosphate synthase 1 (CPS1) inhibits its activity but can be reversed by SIRT5 [7]. More importantly, molecular evidence suggested that abnormal glutarylation was closely related to several metabolic disorders, including diabetes, neurodegenerative diseases, glutaric acidemia type I and cancer [7]. Therefore, research on glutarylation would be beneficial for drug discovery. Although some research work has been done to reveal the biological functions of glutarylation, the regulatory mechanism of glutarylation in cells is still largely unknown.

In order to further investigate the molecular mechanisms of glutarylation, a fundamental and critical task is to identify glutarylation sites with high accuracy. Although several large-scale proteomics methods such as mass spectrometry [7, 8] have been applied to detect glutarylation sites, these experimental approaches are not only time-consuming but also expensive. The majority of lysine glutarylation substrates and glutarylation sites still remain largely unknown. Therefore, it is urgent and necessary to develop computational methods to identify the potential glutarylated proteins and the corresponding glutarylation sites. Up to now, a few computational tools have been proposed to identify glutarylation sites. Ju and He [9] proposed the first glutarylation site predictor named GlutPred based on maximum relevance minimum redundancy (mRMR) feature selection algorithm. Xu *et al.* [10] developed a predictor, iGlu-Lys, by using the position-specific propensity matrix (PSPM) features around lysine-centered peptides and SVM algorithm. Huang *et al.* [11] proposed a prediction model by incorporating maximal dependence decomposition (MDD)-identified substrate motifs into an integrated SVM classifier. The cross-validation showed that amino acid composition features were most effective in discriminating between glutarylation and non-glutarylation sites. Recently, Albarakati *et al.* [12] developed a novel predictor, RF-GlutarySite, by using the physiochem-

*Address correspondence to this author at the College of Science, Shenyang Aerospace University, Shenyang 110136, P.R. China;
Tel: +86 024 89723442; E-mail: juzhe1120@hotmail.com

ical and sequence-based features and random forest (RF) algorithm.

Note that in the aforementioned four existing prediction methods, the experimentally verified glutarylation sites were treated as the positive samples and the remaining non-verified lysine sites were treated as the negative samples to train classifiers to predict glutarylation sites from unknown proteins. However, due to the limitations of experimental technique and condition, the remaining non-verified lysine sites might contain some glutarylation sites which have not been experimentally identified yet. Thus, the existing predictors were actually built on the noisy dataset. As a result, the accuracy of the existing predictors would not be as good as they were supposed to be.

In contrast to previous methods, experimentally verified glutarylation sites were treated as positive samples and the remaining non-verified lysine sites were treated as unlabeled samples in our study. A novel glutarylation site predictor was developed by using a positive unlabeled (PU) learning technique [13]. Specifically, the algorithm had five stages: stage 1, the composition of k-spaced amino acid pairs (CKSAAP), binary encoding (BE), and amino acid factors (AAF) were combined to encode the glutarylation site; stage 2, the crucial features were refined out using the maximum relevance and minimum redundancy (mRMR) feature selection method [14]; stage 3, a reliable negative set was selected from the unlabeled set by a maximum distance rule; stage 4, the reliable negative set was expanded and a series of SVM classifiers with RBF (Radial Basis Function) kernels were trained iteratively; stage 5, a final SVM model was trained on the positive set and the selected reliable negative set by 10-fold cross-validation. This method was called PUL-GLU (PU Learning for GLUtarylation sites prediction). The experimental results showed that the accuracy of PUL-GLU was 79.77% on the training set evaluated by 10-fold cross-validation and 76.65% on the independent test set.

As demonstrated by a series of recent publications [15, 16] and summarized in three comprehensive review papers [17-19], to develop a really useful predictor for a biological system, one needs to follow "Chou's 5-steps rule" [17] to go through the following five steps: (1) select or construct a valid benchmark dataset to train and test the predictor; (2) represent the samples with an effective formulation that can truly reflect their intrinsic correlation with the target to be predicted; (3) introduce or develop a powerful algorithm to conduct the prediction; (4) properly perform cross-validation tests to objectively evaluate the anticipated prediction accuracy; (5) establish a user-friendly web-server for the predictor that is accessible to the public. The description of how to deal with these five steps is given below.

2. MATERIALS AND METHODS

2.1. Dataset

Benchmark dataset was collected from the recent literature (Ju and He, 2018) in this study. The identity of these proteins was reduced to 40% by the CD-HIT program [20]. The training set consisted of 167 proteins with 590 experimentally annotated lysine glutarylation sites and 3498 non-annotated lysine sites; the independent test set consisted of 20 proteins with 56 experimentally annotated lysine glutarylation sites and 428 non-annotated lysine sites. Sliding window method was used to encode every lysine residue K of the dataset because glutarylation only occurred in lysine residues K. Based on our previous work [9], the window size of every training peptide was selected as 35 here. That means every lysine residue in the training dataset and the testing dataset was represented as a peptide segment of length 35 with 17 residues upstream and 17 residues downstream of lysine residue K. The training set and independent testing set are provided in Supplementary Material **S1**.

2.2. Feature Construction

2.2.1. Amino Acid Factors

Physicochemical properties of amino acids play a crucial role in the identification of PTMs site. By using multivariate statistical analyses, 544 physicochemical properties of amino acids in the AAIndex have been transformed into five multidimensional patterns of attributes. The five multidimensional patterns of attributes reflect the polarity, secondary structure, molecular volume, codon diversity, and electrostatic charge, respectively [21]. These five generated attributes are named amino acid factors (AAF). For a given peptide, it can be encoded as a 35*5=175-dimensional vector by AAF.

2.2.2. Binary Encoding

The information of amino acid composition and position can be effectively characterized by binary encoding (BE) [22]. Considering 21 amino acids were ordered as 'ACDEFGHIKLMNPQRSTVWYX' ('X' means virtual amino acid), each amino acid residue in a given peptide was translated into a 21-dimensional binary vector. For example, amino acid 'A' is encoded as (100000000000000000000), …, and 'X' is encoded as (000000000000000000001). Thus, every training peptide can be expressed as a 35*21=735-dimensional vector by BE encoding.

2.2.3. Composition of k-spaced Amino Acid Pairs

The composition of k-spaced amino acid pairs (CKSAAP) can reflect the short linear motif information by calculating the occurrence frequency of the amino acid pairs in a given sequence fragment [23, 24]. An amino acid pair separated by any k amino acid residues is known as the k-spaced amino acid pair. For example, the CKSAAP of a given peptide for k=1 yields a 441-dimensional numeric vector defined as:

$$(N_{AxA} / N_{Total}, N_{AxC} / N_{Total}, ..., N_{XxX} / N_{Total})_{441} \quad (1)$$

where 'x' represents any one of the 21 amino acids, and N_{Total} represents the total number of 1-spaced amino acid pairs. Here, CKSAAP with k= 0, 1, 2, 3 and 4 was utilized to encode the training peptides as 2205-dimensional feature vectors.

2.2.4. The Feature Space

In accordance with our previous work [9], the AAF, BE and CKSAAP were integrated to encode the training samples. Overall, each sample in the benchmark dataset was encoded as a 35*5+35*21+2205=3115-dimensional feature.

Since the integrated encoding generated a high-dimensional feature vector, the maximum relevance and minimum redundancy (mRMR) feature selection method [14] and incremental feature selection (IFS) algorithm were used to remove the redundant features. Firstly, each component of 3115 features was ranked by the mRMR method. Then, the IFS algorithm was used to select 50 features with the highest score in each iteration. Here, the top 300 features were selected as optimal input features based on our previous work [9].

2.3. Prediction Method

2.3.1. Support Vector Machine

To facilitate its description, the training set is denoted as $\{(x_i, t_i), i = 1,2,...,l\}$. The SVM can be formulated as follows:

$$\min_{\omega,\xi} \frac{1}{2} \| \omega \|^2 + C \sum_{i=1}^{l} \xi_i$$

$$\text{s.t. } t_i(\omega \cdot \Phi(x_i) + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0, \quad i = 1, 2, ..., l \tag{2}$$

where $\Phi(x)$ is the non-linear mapping, and x_i (i=1, 2,...,l) are slack variables. C is the parameter determining the trade-off between model complexity (margin size) and classification errors. The Gaussian kernel function $K(x_i, x_j) = \Phi(x_i)^{\mathrm{T}} \Phi(x_j) = \exp(-\gamma \| x_i - x_j \|^2)$ is used in the SVM. The Libsvm toolkit [25] was utilized to carry out the SVM models. Here, C and γ were set to the default values in Libsvm (*i.e.*, $C = 1$ and $\gamma = 1/300$).

2.3.2. Positive-unlabeled Learning for Bioinformatics

In many fields, the obtainment of negative examples is usually costly and even not possible. Hence, many PU learning algorithms have been developed to deal with the problems which were lacking in negative examples. PU learning originated in text classification [25-28], and has been successfully applied to many biological problems in recent years. For example, Wang *et al.* [13] developed a PU learning algorithm named PSoL to finding non-coding RNA genes; Zhao *et al.* [26] designed a PU learning algorithm, AGPS, for gene function prediction; Cerulo *et al.* [27] used a PU learning algorithm named PosOnly for the derivation of gene regulatory networks; Yang *et al.* [28] designed a PU learning algorithm named PUDI for disease gene identification; Yang *et al.* [29] proposed an ensemble-based PU learning method for identifying disease gene by integrating multiple PU learning classifiers; Li *et al*. [30] proposed a positive unlabelled (PU) learning-based method, PA2DE (V2.0), based on the AlphaMax algorithm for protein glycosylation site prediction [31-33].

2.3.3. Development of PUL-GLU

As mentioned above, in this study, experimentally verified glutarylation sites were treated as positive samples and the remaining non-verified lysine sites were treated as unlabeled samples to build a classifier. In this way, the training dataset is divided into two parts: (1) the positive training dataset P and (2) the unlabeled training dataset U. Thus the prediction of glutarylation sites became learning from positive and unlabeled samples. An effective positive-unlabeled learning algorithm, PSoL [13], was used to construct PUL-GLU. The flowchart of PUL-GLU is shown in Table **1**. There are three stages in it:

Stage 1. Selection of initial reliable negatives:

PUL-GLU selected the initial reliable negative set RN^0 from the unlabeled set U based on the formula (3). The formula (3) ensures that the selected initial negative set has the highest reliability because it is farthest from the positive example set.

$$RN^0 = \arg\max_{\substack{N \subset U \\ |N| = |P|}} d(N, P) \tag{3}$$

where $d(N, P)$ is defined as follows:

$$d(N, P) = \min_{p \in P} \sum_{n \in N} \| n - p \| \tag{4}$$

Stage 2. Expansion of the reliable negative example set:

The initial negative set is gradually extended by iteratively trained SVM classifiers. Let RN^i be the current reliable negative training set; and U^i be the current unlabeled set at the ith iteration. An SVM classifier f^i was firstly trained on P and RN^i; then, f^i was used to classify U^i and calculate its decision value. To ensure the purity of the selected negative set, the selected negative samples with the decision value less than a threshold T (here, T was set to -0.2) were selected as the newly predicted negative set N^i_{pred}. To avoid the imbalance problem, the size of N^i_{pred} was controlled less than $|P|$, and RN^i is replaced with the negative support vectors N^i_{SV}. At the (i+1)th iteration, $U^{i+1} = U^i \setminus N^i_{pred}$; $RN^{i+1} = N^i_{pred} \cup N^i_{SV}$. An SVM classifier f^{i+1} was trained on P and RN^{i+1}. As the number of iterations increases, RN^i may contain more and more false-positive examples, therefore, iteration should be terminated if the size of U^i goes below a threshold $r*|P|$ (here r was set to 2).

Stage 3. Acquisition of the final classifier:

Let *RN* be the representative reliable negative training set. A final SVM classifier *f* was trained on *P* and *RN*.

2.4. Cross-validation and Performance Assessment

Jackknife test, K-fold cross-validation, and independent dataset test are three of the most common strategies for the evaluation of the performance of a predictor [17]. Although the jackknife test is the most objective among three evaluation methods, it is the most time-intensive. Therefore, to reduce computational time, we adopted a 10-fold cross-validation test to evaluate the proposed model. The 10-fold cross-validation is repeated 10 times. In addition, an independent dataset test was also adopted to further evaluate our method.

Table 1. The flowchart of the PUL-GLU algorithm.

Input: - positive training set P - unlabeled set U
Output: - final SVM classifier f
Stage 1: Selection of initial reliable negative set: - $RN^0 = \arg\max_{\substack{N \subset U \\ \|N\|=\|P\|}} d(N,P)$
Stage 2: Expansion of reliable negative set: - $i = 0$; - Repeat - $U = U \setminus RN^i$; - Train SVM f^i on P and RN^i; - Classify U by f^i; - N^i_{pred} is the predicted negative set, where $\| N^i_{pred} \| \leq \| P \|$ and $f(N^i_{pred}) < -0.2$; - $RN^{i+1} = N^i_{pred} \cup N^i_{SV}$ where N^i_{SV} is the negative SVs of f^i; - $i = i + 1$; - until $\| U \| \leq 2^* \| P \|$;
Stage 3: Acquisition of the final classifier: - A final SVM classifier f was trained on P and RN.

Five widely-accepted measurements, including sensitivity (Sn), specificity (Sp), precision (Pre), accuracy (ACC), and Matthew's correlation coefficient (MCC), were used to evaluate the prediction performances of PUL-GLU, which are defined as:

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{5}$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{6}$$

$$\mathrm{Pre} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{7}$$

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \tag{8}$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FN} \times \mathrm{FP}}{\sqrt{(\mathrm{TP} + \mathrm{FN}) \times (\mathrm{TN} + \mathrm{FP}) \times (\mathrm{TP} + \mathrm{FP}) \times (\mathrm{TN} + \mathrm{FN})}} \tag{9}$$

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

3. RESULTS AND DISCUSSION

3.1. Performance of PUL-GLU

To select the representative reliable negative samples, a negative set expansion was implemented on the training set. As a result, the representative reliable negative set RN contains 1199 reliable non-glutarylated lysine sites (Supplementary Material **S1**). Finally, PUL-GLU was trained on the positive set P and the representative reliable negative set RN. The 10-fold cross-validation of PUL-GLU on P and RN is shown in Table **2**. As shown in Table **2**, the prediction values for Sn, Sp, Pre, ACC, and MCC values reached 66.56%,

86.43%, 70.71%, 79.88% and 0.5384, respectively. The prediction performance of PUL-GLU was much higher than the SVM model trained on positive samples and randomly selected negative samples. This result indicated that the selected representative reliable negative samples could be more effective than those selected randomly. To avoid the overestimation of PUL-GLU, it was performed on the entire training set. The performance of PUL-GLU also achieved a satisfactory performance with an MCC of 0.35.

We compared PUL-GLU with existing glutarylation site predictors. As shown in Table **2**, in the training dataset, PUL-GLU reaches the highest MCC values of 0.5384 by 10-fold cross-validation. Although iGlu-Lys achieved the highest value of Sp (95.2%), the value of Sn (50.4%) was much lower than that of PUL-GLU (66.6%). It suggests that iGlu-Lys tends to identify a query lysine site as a non-glutaryllysine, and can predict less glutaryllysine sites than PUL-GLU. Moreover, the Sn value of MDDGlutar (66.7%) is slightly higher than that of PUL-GLU (66.6%), but the Sp value of MDDGlutar (61.9%) is much lower than that of PUL-GLU (86.4%). It indicates that PUL-GLU can predict more non-glutaryllysine sites than MDDGlutar at a similar level of Sn. As PUL-GLU and GlutPred were trained on the same training dataset with the same features, the better performance of PUL-GLU suggested that by using the extracted reliable non-glutarylated lysine sites to train model, the prediction performance has been improved effectively. In short, PUL-GLU outperforms the current glutarylation site predictors remarkably on the training dataset.

3.2. Comparison of PUL- GLU with Other Predictors on the Independent Test Set

To further evaluate the effectiveness of PUL-GLU, it was compared with the other current methods on the independent test set. It should be pointed out that RF-GlutarySite [12] did not provide a shared web-server. Hence, RF-GlutarySite was not compared with PUL-GLU. The compared results of existing predictors are shown in Table **3**. Although iGlu-Lys achieved the best performance on the independent test dataset, the prediction results were overestimated. In fact, the training set of iGlu-Lys contains all of the samples of our independent test set; whereas PUL-GLU, GlutPred and MDDGlutar were trained and tested on the same dataset. As PUL-GLU was trained by the PU learning algorithm, the performance of PUL-GLU outperforms GlutPred and MDDGlutar. The results of the independent test and cross-validation both demonstrated that PUL-GLU could be an effective predictor for the prediction of glutarylation sites.

3.3. Prediction Server of PUL-GLU

Building a user-friendly online server can provide convenience for the related experimental researchers to further

Table 2. 10-fold cross-validation performance of PUL-GLU and other methods.

Methods	Sn(%)	Sp(%)	Pre(%)	ACC(%)	MCC
SVM[1]	61.73±0.83	76.45±1.45	56.36±1.53	71.59±0.98	0.3738±0.0173
GlutPred	64.80±0.99	76.60±0.28	31.84±0.49	74.90±0.32	0.3194±0.0087
iGlu-Lys[2]	50.4±0.88	95.2±0.14	—	88.38±0.15	0.5098±0.0072
MDDGlutar[2]	67.7	61.9	—	63.8	0.28
RF-GlutarySite	74.9	69.7	71.2	72.3	0.45
PUL-GLU	66.56±0.73	86.43±0.28	70.71±0.45	79.88±0.29	0.5384±0.69
PUL-GLU[3]	71.69	75.07	32.66	74.58	0.3533

[1] SVM trained on 590 positive samples and 1199 randomly extracted negative samples.

[2] The values of Pre were not reported for iGlu–Ly and MDDGlutar, therefore, no comparison could be made with respect to this parameter.

[3] PUL-GLU was performed on the entire training dataset.

Table 3. Comparison with other predictors on the independent test dataset.

Methods	Sn(%)	Sp(%)	Pre(%)	ACC(%)	MCC
SVM[1]	51.79	76.87	22.66	73.97	0.2078
GlutPred	51.79	78.50	23.97	75.41	0.2238
iGlu-Lys	89.09	97.67	83.05	96.69	0.8416
MDDGlutar	49.09	84.62	29.03	80.58	0.2715
PUL-GLU	58.93	78.97	26.83	76.65	0.2785

[1] SVM trained on 590 positive samples and 1199 randomly extracted negative samples.

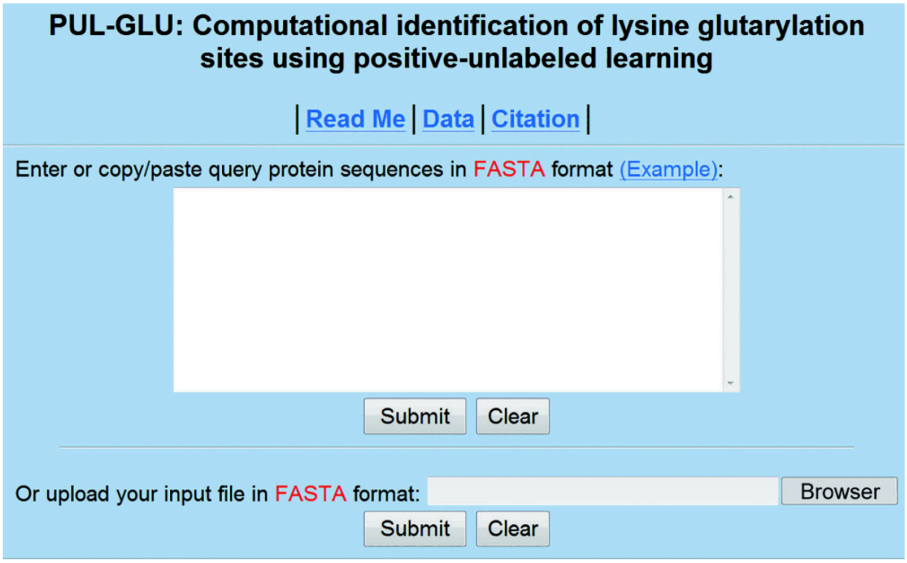

Fig. (1). The prediction interface of the web-server PUL-GLU. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

Table 4. The top 20 most likely glutarylation sites in non-validated lysine sites.

Uniprot_AC	Site	SVM Score	Uniprot_AC	Site	SVM Score
P32020	432	3.04	Q8BMS1	411	1.63
Q8BMS1	413	2.82	Q8BWT1	211	1.63
P42125	242	2.51	P26443	386	1.62
Q8BMS1	414	2.48	Q8BMS1	262	1.60
Q8C196	906	2.36	P54869	342	1.59
P32020	442	2.25	Q8C196	856	1.56
Q8C196	908	2.21	Q9D819	238	1.53
Q61425	206	1.93	Q8BMS1	249	1.53
Q61176	39	1.65	Q9D172	155	1.53
Q8BMS1	284	1.64	Q61425	202	1.52

investigate the molecular mechanisms of glutarylation. Therefore, PUL-GLU has been implemented as a web-server. The prediction server for PUL-GLU is available at http://bioinform.cn/pul_glu. The style of PUL-GLU is similar to the published webserver iGlu-Lys [10]. As shown in Fig. (**1**), PUL-GLU accepts single query protein or multiple query proteins in FASTA format. Or users can upload query proteins in FASTA format as a text document for the prediction of glutarylation sites. The predicted results will be written to a CSV-formatted file.

3.4. Prediction of the Most Likely Glutaryllysine in Non-annotated Lysine Residues

As mentioned earlier, there are 646 experimentally validated glutarylation sites and 3926 non-validated lysine sites in the training dataset. However, the non-validated lysine residues may contain some glutarylation sites which have not been experimentally identified yet. To find the most likely glutarylation sites from those non-validated lysine residues, all 3926 non-validated lysine sites in the training dataset have been re-predicted by PUL-GLU algorithm. The top 20 most likely glutaryllysine in non-validated lysine residues are listed in Table **4**. Here, we just give a possible hypothesis, it remains to be experimentally identified whether those lysine residues can be glutarylated or not. The completed prediction results are given in Supplementary Material **S2** and may provide clues for studying glutarylation sites.

CONCLUSION

In this study, we developed a bioinformatics tool named PUL-GLU for the prediction of glutarylation sites using the PU learning algorithm and multiple sequence features. To

the best of our knowledge, this is the first time PU learning has been applied to predict the glutarylation sites. Experimental results have shown that PUL-GLU outperformed the current glutarylation site predictors. A web-server for PUL-GLU was built, which could provide a great convenience for experimental researchers to investigate glutarylation.

AUTHORS' CONTRIBUTIONS

Zhe Ju wrote the manuscript and was involved in all the experimental steps. Shi-Yun Wang constructed the online web-server of PUL-GLU. Both the authors approved the final version of this manuscript.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

HUMAN AND ANIMAL RIGHTS

No Animals/Humans were used for studies that are the basis of this research.

CONSENT FOR PUBLICATION

Not applicable.

AVAILABILITY OF DATA AND MATERIALS

The data supporting the findings of the article is available in the Data repository at 123.206.31.171/GlutPred/, reference number [9].

FUNDING

This work was supported by the National Natural Science Foundation of China (No. 11701390); the Natural Science Foundation of Liaoning Province (No. 2019-BS-187); and the Scientific Research Fund Project in Liaoning Province Department of Education (No. JYT19027).

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

Declared none.

SUPPLEMENTARY MATERIAL

Supplementary material is available on the publisher's website along with the published article.

REFERENCES

[1] Chen, Y.; Sprung, R.; Tang, Y.; Ball, H.; Sangras, B.; Kim, S.C.; Falck, J.R.; Peng, J.; Gu, W.; Zhao, Y. Lysine propionylation and butyrylation are novel post-translational modifications in histones. *Mol. Cell. Proteomics,* **2007**, *6*(5), 812-819. http://dx.doi.org/10.1074/mcp.M700021-MCP200 PMID: 17267393

[2] Tan, M.; Luo, H.; Lee, S.; Jin, F.; Yang, J.S.; Montellier, E.; Buchou, T.; Cheng, Z.; Rousseaux, S.; Rajagopal, N.; Lu, Z.; Ye, Z.; Zhu, Q.; Wysocka, J.; Ye, Y.; Khochbin, S.; Ren, B.; Zhao, Y. Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. *Cell,* **2011**, *146*(6), 1016-1028. http://dx.doi.org/10.1016/j.cell.2011.08.008 PMID: 21925322

[3] Zhang, Z.; Tan, M.; Xie, Z.; Dai, L.; Chen, Y.; Zhao, Y. Identification of lysine succinylation as a new post-translational modification. *Nat. Chem. Biol.,* **2011**, *7*(1), 58-63. http://dx.doi.org/10.1038/nchembio.495 PMID: 21151122

[4] Choudhary, C.; Weinert, B.T.; Nishida, Y.; Verdin, E.; Mann, M. The growing landscape of lysine acetylation links metabolism and cell signalling. *Nat. Rev. Mol. Cell Biol.,* **2014**, *15*(8), 536-550. http://dx.doi.org/10.1038/nrm3841 PMID: 25053359

[5] Dai, L.; Peng, C.; Montellier, E.; Lu, Z.; Chen, Y.; Ishii, H.; Debernardi, A.; Buchou, T.; Rousseaux, S.; Jin, F.; Sabari, B.R.; Deng, Z.; Allis, C.D.; Ren, B.; Khochbin, S.; Zhao, Y. Lysine 2-hydroxyisobutyrylation is a widely distributed active histone mark. *Nat. Chem. Biol.,* **2014**, *10*(5), 365-370. http://dx.doi.org/10.1038/nchembio.1497 PMID: 24681537

[6] Hirschey, M.D.; Zhao, Y. Metabolic regulation by lysine malonylation, succinylation, and glutarylation. *Mol. Cell. Proteomics,* **2015**, *14*(9), 2308-2315. http://dx.doi.org/10.1074/mcp.R114.046664 PMID: 25717114

[7] Tan, M.; Peng, C.; Anderson, K.A.; Chhoy, P.; Xie, Z.; Dai, L.; Park, J.; Chen, Y.; Huang, H.; Zhang, Y.; Ro, J.; Wagner, G.R.; Green, M.F.; Madsen, A.S.; Schmiesing, J.; Peterson, B.S.; Xu, G.; Ilkayeva, O.R.; Muehlbauer, M.J.; Braulke, T.; Mühlhausen, C.; Backos, D.S.; Olsen, C.A.; McGuire, P.J.; Pletcher, S.D.; Lombard, D.B.; Hirschey, M.D.; Zhao, Y. Lysine glutarylation is a protein posttranslational modification regulated by SIRT5. *Cell Metab.,* **2014**, *19*(4), 605-617. http://dx.doi.org/10.1016/j.cmet.2014.03.014 PMID: 24703693

[8] Xie, L.; Wang, G.; Yu, Z.; Zhou, M.; Li, Q.; Huang, H.; Xie, J. Proteome-wide lysine glutarylation profiling of the *Mycobacterium tuberculosis* H37Rv. *J. Proteome Res.,* **2016**, *15*(4), 1379-1385. http://dx.doi.org/10.1021/acs.jproteome.5b00917 PMID: 26903315

[9] Ju, Z.; He, J.J. Prediction of lysine glutarylation sites by maximum relevance minimum redundancy feature selection. *Anal. Biochem.,* **2018**, *550*, 1-7. http://dx.doi.org/10.1016/j.ab.2018.04.005 PMID: 29641975

[10] Xu, Y.; Yang, Y.; Ding, J.; Li, C. iGlu-Lys: A predictor for lysine glutarylation through amino acid pair order features. *IEEE Trans. Nanobioscience,* **2018**, *17*(4), 394-401. http://dx.doi.org/10.1109/TNB.2018.2848673 PMID: 29994125

[11] Huang, K.Y.; Kao, H.J.; Hsu, J.B.; Weng, S.L.; Lee, T.Y. Characterization and identification of lysine glutarylation based on intrinsic interdependence between positions in the substrate sites. *BMC Bioinformatics,* **2019**, *19*(Suppl. 13), 384. http://dx.doi.org/10.1186/s12859-018-2394-9 PMID: 30717647

[12] Al-Barakati, H.J.; Saigo, H.; Newman, R.H.; Kc, D.B. RF-GlutarySite: a random forest based predictor for glutarylation sites. *Mol Omics,* **2019**, *15*(3), 189-204. http://dx.doi.org/10.1039/C9MO00028C PMID: 31025681

[13] 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*(21), 2590-2596. http://dx.doi.org/10.1093/bioinformatics/btl441 PMID: 16945945

[14] Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.,* **2005**, *27*(8), 1226-1238. http://dx.doi.org/10.1109/TPAMI.2005.159 PMID: 16119262

[15] Du, X.; Diao, Y.; Liu, H.; Li, S. MsDBP: Exploring DNA-binding proteins by integrating multiscale sequence information *via* chou's five-step rule. *J. Proteome Res.,* **2019**, *18*(8), 3119-3132. http://dx.doi.org/10.1021/acs.jproteome.9b00226 PMID: 31267738

[16] Kabir, M.; Ahmad, S.; Iqbal, M.; Hayat, M. iNR-2L: a two-level sequence-based predictor developed *via* Chou's 5-steps rule and general PseAAC for identifying nuclear receptors and their families. *Genomics,* **2019**, *112*(1), 276-285. http://dx.doi.org/10.1016/j.ygeno.2019.02.006

[17] Chou, K.C. Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.,* **2011**, *273*(1), 236-247. http://dx.doi.org/10.1016/j.jtbi.2010.12.024 PMID: 21168420

[18] Chou, K.C. Advance in predicting subcellular localization of multi-label proteins and its implication for developing multi-target drugs. *Curr. Med. Chem.,* **2019**, *26*, 4918-4943.

http://dx.doi.org/10.2174/0929867326666190507082559 PMID: 31060481

[19] Chou, K.C. Impacts of pseudo amino acid components and 5-steps rule to proteomics and proteome analysis. *Curr. Topics Med. Chem.,* **2019**. *19*(25), 2283-2300. http://dx.doi.org/10.2174/1568026619666191018100141

[20] Li, W.; Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics,* **2006**, *22*(13), 1658-1659. http://dx.doi.org/10.1093/bioinformatics/btl158 PMID: 16731699

[21] Atchley, W.R.; Zhao, J.; Fernandes, A.D.; Drüke, T. Solving the protein sequence metric problem. *Proc. Natl. Acad. Sci. USA,* **2005**, *102*(18), 6395-6400. http://dx.doi.org/10.1073/pnas.0408677102 PMID: 15851683

[22] Sagara, J.I.; Shimizu, S.; Kawabata, T.; Nakamura, S.; Ikeguchi, M.; Shimizu, K. The use of sequence comparison to detect 'identities' in tRNA genes. *Nucleic Acids Res.,* **1998**, *26*(8), 1974-1979. http://dx.doi.org/10.1093/nar/26.8.1974 PMID: 9518491

[23] Ju, Z.; Cao, J.Z. Prediction of protein N-formylation using the composition of k-spaced amino acid pairs. *Anal. Biochem.,* **2017**, *534*, 40-45. http://dx.doi.org/10.1016/j.ab.2017.07.011 PMID: 28709899

[24] Ju, Z.; Wang, S.Y. Prediction of citrullination sites by incorporating k-spaced amino acid pairs into Chou's general pseudo amino acid composition. *Gene,* **2018**, *664*, 78-83. http://dx.doi.org/10.1016/j.gene.2018.04.055 PMID: 29694908

[25] Chang, C.C.; Lin, C.J. Libsvm: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.,* **2011**, *2*, 27. http://dx.doi.org/10.1145/1961189.1961199

[26] Yu, H.; Han, J.; Chang, K.C. **2002**, PEBL: positive example based learning for web page classification using svm. *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 239-248. http://dx.doi.org/10.1145/775047.775083

[27] Liu, B.; Dai, Y.; Li, X.; Lee, W.S.; Yu, P.S. Building text classifiers using positive and unlabeled examples. *In: Data Mining,* Third IEEE International Conference on, IEEE **2003**, pp. 179-186.

[28] Liu, B.; Lee, W.S.; Yu, P.S.; Li, X. Partially supervised classification of text documents. *ICML*; *Citeseer*, **2002**, *2*, 387-394.

[29] Zhao, X.M.; Wang, Y.; Chen, L.; Aihara, K. Gene function prediction using labeled and unlabeled data. *BMC Bioinformatics,* **2008**, *9*, 57. http://dx.doi.org/10.1186/1471-2105-9-57 PMID: 18221567

[30] Cerulo, L.; Elkan, C.; Ceccarelli, M. Learning gene regulatory networks from only positive and unlabeled data. *BMC Bioinformatics,* **2010**, *11*, 228. http://dx.doi.org/10.1186/1471-2105-11-228 PMID: 20444264

[31] Yang, P.; Li, X.L.; Mei, J.P.; Kwoh, C.K.; Ng, S.K. Positive-unlabeled learning for disease gene identification. *Bioinformatics,* **2012**, *28*(20), 2640-2647. http://dx.doi.org/10.1093/bioinformatics/bts504 PMID: 22923290

[32] Yang, P.; Li, X.; Chua, H.N.; Kwoh, C.K.; Ng, S.K. Ensemble positive unlabeled learning for disease gene identification. *PLoS One,* **2014**, *9*(5), e97079. http://dx.doi.org/10.1371/journal.pone.0097079 PMID: 24816822

[33] Li, F.; Zhang, Y.; Purcell, A.W.; Webb, G.I.; Chou, K.C.; Lithgow, T.; Li, C.; Song, J. Positive-unlabelled learning of glycosylation sites in the human proteome. *BMC Bioinformatics,* **2019**, *20*(1), 112. http://dx.doi.org/10.1186/s12859-019-2700-1 PMID: 30841845