Huang et al. BMC Bioinformatics 2019, 19(Suppl 13):384
https://doi.org/10.1186/s12859-018-2394-9

BMC Bioinformatics

RESEARCH Open Access

Characterization and identification of lysine glutarylation based on intrinsic interdependence between positions in the substrate sites

Kai-Yao Huang[1,2], Hui-Ju Kao[2,3], Justin Bo-Kai Hsu[4], Shun-Long Weng[5,6,7] and Tzong-Yi Lee[1,2*]

From 17th International Conference on Bioinformatics (InCoB 2018)
New Delhi, India. 26-28 September 2018

Abstract

Background: Glutarylation, the addition of a glutaryl group (five carbons) to a lysine residue of a protein molecule, is an important post-translational modification and plays a regulatory role in a variety of physiological and biological processes. As the number of experimentally identified glutarylated peptides increases, it becomes imperative to investigate substrate motifs to enhance the study of protein glutarylation. We carried out a bioinformatics investigation of glutarylation sites based on amino acid composition using a public database containing information on 430 non-homologous glutarylation sites.

Results: The TwoSampleLogo analysis indicates that positively charged and polar amino acids surrounding glutarylated sites may be associated with the specificity in substrate site of protein glutarylation. Additionally, the chi-squared test was utilized to explore the intrinsic interdependence between two positions around glutarylation sites. Further, maximal dependence decomposition (MDD), which consists of partitioning a large-scale dataset into subgroups with statistically significant amino acid conservation, was used to capture motif signatures of glutarylation sites. We considered single features, such as amino acid composition (AAC), amino acid pair composition (AAPC), and composition of k-spaced amino acid pairs (CKSAAP), as well as the effectiveness of incorporating MDD-identified substrate motifs into an integrated prediction model. Evaluation by five-fold cross-validation showed that AAC was most effective in discriminating between glutarylation and non-glutarylation sites, according to support vector machine (SVM).

Conclusions: The SVM model integrating MDD-identified substrate motifs performed well, with a sensitivity of 0.677, a specificity of 0.619, an accuracy of 0.638, and a Matthews Correlation Coefficient (MCC) value of 0.28. Using an independent testing dataset (46 glutarylated and 92 non-glutarylated sites) obtained from the literature, we demonstrated that the integrated SVM model could improve the predictive performance effectively, yielding a balanced sensitivity and specificity of 0.652 and 0.739, respectively. This integrated SVM model has been implemented as a web-based system (MDDGlutar), which is now freely available at http://csb.cse.yzu.edu.tw/MDDGlutar/.

Keywords: Protein glutarylation, Intrinsic interdependence, Maximal dependence decomposition

* Correspondence: francislee0215@gmail.com
[1]School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen 518172, China
[2]Warshel Institute for Computational Biology, The Chinese University of Hong Kong, Shenzhen 518172, China
Full list of author information is available at the end of the article

Background

Protein post-translational modifications (PTM) consist of chemical modifications that play an important role in many cellular processes in biology, such as regulating the activity, localization, and protein interactions. Some PTMs occur at amino acid side chains of a protein molecule, generally by covalent enzymatic modification. Most PTMs are enzymatically controlled and regulated. Phosphorylation is the most well-known example of a PTM, which is the attachment of a phosphoryl group to a serine (S), threonine (T), or tyrosine (Y) residue of a protein via protein kinases [1]. Crucial protein PTM occurring at the ε-amino groups of specific lysine residues (K) have proven to be hallmarks of active chromatin, and major regulators of gene expression, protein-protein interactions, and protein processing and degradation. These include 2-hydroxyisobutyrylation [2], acetylation [3], butyrylation [4], crotonylation [5], malonylation [6], propionylation [7] and succinylation [8]. In addition, malonylation, succinylation, and glutarylation, which are collectively referred to as lysine acylation, are highly dynamic PTMs and appear conserved across evolutionarily related species [3].

Lysine glutarylation, which is the addition of a glutaryl group to a lysine residue of a protein molecule, is an important posttranslational modification. It plays a crucial role in mitochondrial functions and metabolic processes both in eukaryotic and prokaryotic cells, such as amino acid metabolism, fatty acid metabolism and cellular respiration [6]. Previous studies have indicated that the activity of carbamoyl phosphate synthase 1 was inhibited through protein glutarylation [9]. However, lysine acyltransferases (KATs) can bind specificity to its substrates in malonylation and succinylation but lacking evidence for glutarylation. Because of the similarities in biological psychology, we surmised the mechanism of glutarylation in much the same way that can be enzymatically catalyzed by KATs and removed by lysine deacylases (KDACs) as acylation. Moreover, owing to the labile nature and low abundance of in vivo glutarylation sites, further research is needed to clarify the characteristics of glutarylation and its mechanisms. Therefore, there is an urgent requirement in bioinformatics for a practical approach to investigate the potential substrate motifs of protein glutarylation sites to be designed.

In this study, we used a similar concept as that previously developed to predict protein functional sites using in silico characterization of substrate specificity [10–14]. We selected sequence-based features to discriminate between glutarylation sites and non-glutarylation sites, such as amino acid composition (AAC), amino acid pair composition (AAPC), and composition of k-spaced amino acid pairs (CKSAAP). Additionally, we applied a chi-square test, as part of maximal dependence decomposition (MDD), to measure the interdependence between positions in the substrate sites. Further, MDD can partition large-scale PTM sites data into subgroups according to the most significant dependencies between amino acid compositions surrounding the substrate sites. We explored promising consensus motifs for glutarylation sites by applying MDD [15]. Subsequently, for each subgroup containing one of the MDD-identified substrate motifs, we built a predictive model using support vector machine (SVM). Furthermore, to assess the effectiveness of the proposed models by five-fold cross-validation, we created an independent test dataset by extracting experimental data from the literature, which was completely blind to the training dataset. To facilitate the study of protein glutarylation, we are motivated to design a public system, named MDDGlutar (http://csb.cse.yzu.edu.tw/MDDGlutar/), for the identification of glutarylation sites and their corresponding motifs using experimentally verified glutarylation sites curated from research articles.

Methods

Data collection and preprocessing

Protein Lysine Modifications Database (PLMD) [16] is a manually curated database of experimentally verified glutarylation sites, which contains 715 glutarylation sites of 211 proteins. The purpose of this study was to investigate potential substrate motifs based on the amino acids surrounding glutarylated lysine residues. For this reason, we extracted sequence fragments centered around experimentally verified glutarylation sites with a window length of 2n + 1, such that the fragment included n upstream and n downstream flanking amino acids. These sequence fragments of length 2n + 1 amino acids centered at the glutarylated lysine residue were regarded as the positive dataset. Alternatively, if the lysine residue has not been annotated as a glutarylation site, the fragments were regarded as the negative dataset. The determination of an appropriate window size for model construction is difficult without the well-defined information of motif signatures. Hence, we adopted different window lengths ranging from 11 to 25 for the preparation of training datasets. The performance comparison among predictive models using different window lengths was performed on the basis of SVM classifier with AAC features. As shown in Additional file 1, the cross-validation results displayed that the model trained using 21-mer window length could provide best performance than that using other window lengths.

The CD-HIT software [17] is a useful tool for clustering protein sequences based on a specified value of sequence identity. It was used to remove homologous sequence fragments from the positive and negative datasets, preventing overestimation of the predictive performance. Because of the incomplete information available concerning the experimentally validated glutarylation sites, analysis of

sequence fragments could sometimes lead to a negative sequence appearing identical to a positive sequence, potentially causing false positive or false negative predictions. Therefore, CD-HIT was applied a second time by running cd-hit-2d across the positive and negative training datasets with 100% sequence identity. If a sequence in the negative set was the same as a sequence in the positive set, only the sequence in the positive set was reserved, and the sequence in the negative set was discarded. After filtering out homologous fragments with 40% sequence identity, as shown in Table 1, the non-homologous dataset consisted of 476 positive sequences and 1918 negative sequences. The non-homologous dataset was divided into two parts, training dataset and independent dataset. The training dataset included 430 positive sequences, and a random selection of approximately 1:2 of the 860 negative sequences (approximately the ratio between the number of positive and negative sequences). The remaining sites were used as the testing dataset. Based on the binary classification, the positive (glutarylation sites) and negative (non-glutarylation sites) datasets were used to build a predictive model.

However, since the parameters of the predictive model were optimized, its predictive performance might be overestimated because of over-fitting the training dataset. In order to evaluate the actual performance of the proposed models, we generated an independent dataset, blind to the training datasets. The independent testing dataset was generated by extracting glutarylated peptides excluding those represented in training dataset. Similar to the training dataset, based on a window size of 21 (n = 10), the independent testing dataset contained 46 positive and 92 negative sequences (Table 1).

Investigation and encoding of sequence-based features

The research in this study focused on the analysis of sequence-based features including amino acid composition (AAC), amino acid pair composition (AAPC), and composition of k-spaced amino acid pairs (CKSAAP). To construct an SVM prediction model, fragment sequences must be transformed into numeric vectors based on various features. The training dataset contains k vectors $\{\boldsymbol{x_i}, \boldsymbol{i} = \boldsymbol{1}, \boldsymbol{2}..., \boldsymbol{k}\}$, which represent the k sequence fragments of length corresponding to the specified window length. To classify sites as glutarylation and non-glutarylation sites, a label was applied to each data to mark the class of its corresponding protein.

Table 1 Data statistics of training and testing datasets after the removal of homologous sequences using CD-HIT program

Sequence identity cut-off	Number of glutarylation sites	Number of non-glutarylation sites
Raw Data	715	4145
90%	667	3675
80%	631	3317
70%	597	3037
60%	556	2767
50%	534	2539
40%	476	1918
Training data	430	860
Independent testing data	46	92

Amino acid composition (AAC) is a widely-used sequence feature for calculating the occurring frequency of twenty amino acids within a given sequence fragment. There are 21 types of amino acids that need to be considered for feature encoding. The vector $\boldsymbol{x_i}$ represented the 20 native amino acids and 1 rare amino acid. Some rare amino acids and non-existing "X" residues were used to represent less than 21-mer fragment sequences at an N- or C-terminus [18]. Given a sequence fragment k, $\boldsymbol{f_k}(n)$ represents the number of occurrences of native amino acids, where n stands for one of the 20 types of native amino acids. Hence, the composition for each of the twenty amino acids $\boldsymbol{P_k}(n)$ is computed as follows [19]:

$$P_k(n) = \frac{f_k(n)}{\sum_{n=1}^{20} f_k(n)} n = 1, 2, ..., 20.$$

The AAC vector of a sequence fragment $\boldsymbol{x_k}$ is then define as

$$x_k = [P_k(1), P_k(2), ..., P_k(20)].$$

To encode the composition of each of the twenty amino acids around the glutarylation sites, the 21-dimensional vector $\boldsymbol{x_k}$ included 21 elements specifying the frequencies of 21 amino acids normalized by the total number of amino acids in a fragmented sequence. The composition of amino acid pairs (AAPC) [20], transforms a sequence fragment into a 441-dimensional vector, which includes 441 elements specifying the numbers of occurrences of 441 amino acid pairs divided by the total number of amino acid pairs in a fragmented sequence [21]. CKSAAP [22] is a widely used sequence encoding method that has been applied with great success to many PTM prediction problems, such as O-glycosylation [23], palmitoylation [24], ubiqutination [8], phosphorylation [25], pupylation [26], methylation [27], N-formylation [28] and crotonylation [29]. In this study, we also employed CKSAAP to classify lysine residues into glutarylation and non-glutarylation sites. For example, a pair between glycine (G) and alanine (A), separated by one (k = 1) amino acid of any type, is represented as GxA. To represent a sequence fragment, for each k (k = 1, 2 and 3), the 441-dimensional feature vector $\boldsymbol{x_k}$ needs to be computed, where each component is the frequency of the

corresponding k-spaced amino acid pair appearing in this sequence fragment, respectively.

Capturing the intrinsic interdependence between positions

Following previously described methods [30], a dependency graph model was developed to fully capture the intrinsic interdependence between base positions in a DNA splice site. Hence, we used a chi-square test, as employed in maximal dependence decomposition (MDD) [31], to measure the interdependence between positions in the substrate sites. To perform the dependence testing on a pair of amino acids at the *i*-th and *j*-th positions of a glutarylated site, we built a 21×21 matrix by counting, from a sample of fragment sequences, the observed number $\boldsymbol{Y_{mn}}$ of fragment sequences where the *i*-th amino acid $\boldsymbol{X_i}$ was *m* and the *j*-th amino acid $\boldsymbol{X_j}$ was *n*. The test statistic used can be described as follows:

$$\chi^2(X_i, X_j) = \sum_{m=1}^{21}\sum_{n=1}^{21}\frac{(Y_{mn}-E_{mn})^2}{E_{mn}}$$

where

$$E_{mn} = \frac{Y_{mc}Y_{rn}}{Y}$$

$\boldsymbol{Y_{mc}}$ and $\boldsymbol{Y_{rn}}$ are row sums and column sums of the matrix, respectively. $\boldsymbol{E_{mn}}$ is the expected number of fragment sequences, from a sample of fragment sequences, in which the *i*-th amino acid $\boldsymbol{X_i}$ is *m* and the *j*-th amino acid $\boldsymbol{X_j}$ is *n*. We have selected a window size of 21-mer fragment sequences in the training data to capture the intrinsic interdependence between positions. The amino acids upstream to the glutarylated lysine ($\boldsymbol{P_0}$) were marked as position $\boldsymbol{P_{-10}}$ to $\boldsymbol{P_{-1}}$, whereas those downstream were marked as positions $\boldsymbol{P_{+1}}$ to $\boldsymbol{P_{+10}}$. After constructing, from a glutarylated site, the matrix for each pair ($\boldsymbol{P_i}$,$\boldsymbol{P_j}$) of amino acids at distinct positions of the glutarylated site, we measured the independence for $\boldsymbol{P_i}$ and $\boldsymbol{P_j}$ with the chi-square test $\boldsymbol{\chi^2(P_i, P_j)}$. As stated previously, the proposed method can fully capture the intrinsic interdependence between two positions surrounding glutarylation sites.

Detection of motif signatures by maximal dependence decomposition

Three types of lysine modifications have been recently described which are collectively referred to as lysine acylation: malonylation, succinylation, and glutarylation. A previous study has reported sirtuin 5 (SIRT5) as a lysine deacylase (KDAC) that has potent demalonylase, desuccinylase and deglutarylase activities, both in vitro and in vivo [9, 32, 33]. In contrast, it is worth noting that protein malonylation and succinylation can be enzymatically catalyzed by lysine acyltransferases (KATs). Despite the lack of direct evidence that KAT enzymes bind specificity to its substrates and results in glutarylation. Because of the similarities in biological psychology, we surmised the mechanism of glutarylation in much the same way that a substrate can be glutarylated by one or more KATs.

The increasing number of experimentally identified glutarylation peptides warrants further investigation of substrate motifs to facilitate the study of protein. However, no tool was available so far to predict glutarylation sites, and analysis of their respective substrate motifs remains limited. Thus, the aim of this study was to explore motif signatures of protein glutarylation based on the amino acids surrounding substrate sites. Maximal dependence decomposition (MDD) [31] was utilized to cluster all fragment sequences into subgroups in order to detect those motifs that were statistically conserved among largescale sequence data. The clustering method was performed using MDDLogo [15], which has been demonstrated to increase the effectiveness of PTM sites identification by dividing a group of protein sequences into smaller subgroups before performing the computational identification of the PTM sites [21, 34–44]. In this investigation, a chi-square test $\boldsymbol{\chi^2(A_i, A_j)}$ was used to iteratively evaluate the interdependence between two positions, $\boldsymbol{A_i}$ and $\boldsymbol{A_j}$, which are flanking the substrate site, based on the occurrence of amino acids. Amino acids, 20 in total, were categorized into 5 groups, based on their biochemical properties: polar, acidic, basic, hydrophobic, and aromatic. A contingency table describes the frequency of the presence of each of the twenty amino acids in positions $\boldsymbol{A_i}$ and $\boldsymbol{A_j}$. The chi-square test was defined as:

$$\chi^2(A_i, A_j) = \sum_{m=1}^{5}\sum_{n=1}^{5}\frac{(X_{mn}-E_{mn})^2}{E_{mn}}$$

where X_{mn} is the number of target sequences having amino acids of group *m* in position A_i and having amino acids of group *n* in position A_j, for each pair (A_i, A_j) and $i \neq j$. E_{mn} was determined as $\frac{X_{nR}-X_{Cn}}{X}$, where $X_{mR} = X_{m1} + \ldots + X_{m5}$, $X_{Cn} = X_{1n} + \ldots + X_{5n}$ and X stands for the total number of target sequences. If there is a significant dependence (determined as a X^2 value higher than 34.3, proportional to a cutoff level of $P = 0.01$ with 16 degrees of freedom) between two positions, it followed the description of Burge and Karlin [31]. After the recursive chi-square testing, MDD algorithm can divide a group of target sequences into subsets that capture the most significant dependencies of positions on each other. When executing MDDLogo, a parameter, i.e., the maximum cluster size, should be set. If the size of a subgroup is less than the specified value of maximum cluster size, the subgroup will not be divided any further.

The clustering process of MDD shall be terminated when all the subgroup sizes are less than the specified value of maximum cluster size [15].

Construction of predictive model

A support vector machine (SVM) [45] is a discriminative classifier for pattern recognition and data classification. As shown in Fig. 1, we employed a public SVM library (LIBSVM) [46] to implement the predictive model for distinguishing glutarylation sites from non-glutarylation sites. Based on this binary classification of samples as positive or negative, a kernel function transformed the input samples into a higher dimensional space. Subsequently, a hyperplane is determined for discriminating between the two classes with maximal margin and minimal error by a separating hyperplane. As described in a number of previous works [21, 47–52], the radial basis function (RBF): defined as $(\boldsymbol{S_i}, \boldsymbol{S_j}) = \mathbf{exp}\,(-\boldsymbol{\gamma}\|\boldsymbol{S_i} - \boldsymbol{S_j}\|^2)$, is a reasonably best choice for SVM classifier learning. The RBF kernel is determined by a gamma ($\boldsymbol{\gamma}$) parameter, while the cost © parameter controls the hyper-plane softness. The two supporting parameters could be optimized by a Python program (grid.py) of LIBSVM to improve the predictive accuracy. In this study, LIBSVM was applied to build a predictive model for each feature, and the best feature was selected as the training feature to construct a predictive model for each MDD-clustered subgroup.

Evaluation of predictive performance

We evaluated the predictive performance of the proposed models trained with various features using five-fold cross-validation. First, the training data were randomly divided into five subgroups of approximately equal size, one was used as validation data and the remaining four subgroups were used as training data. The five validation results were then combined to generate a single estimation. Cross-validation evaluation improves the reliability of evaluation, because it considers all the original data, both from the training and testing data sets, and tests each subset only once [47]. To gauge the effective predictive performance of the training model, the following measures were used: sensitivity (Sn), specificity (Sp), accuracy (Acc) and Matthews Correlation Coefficient (MCC):

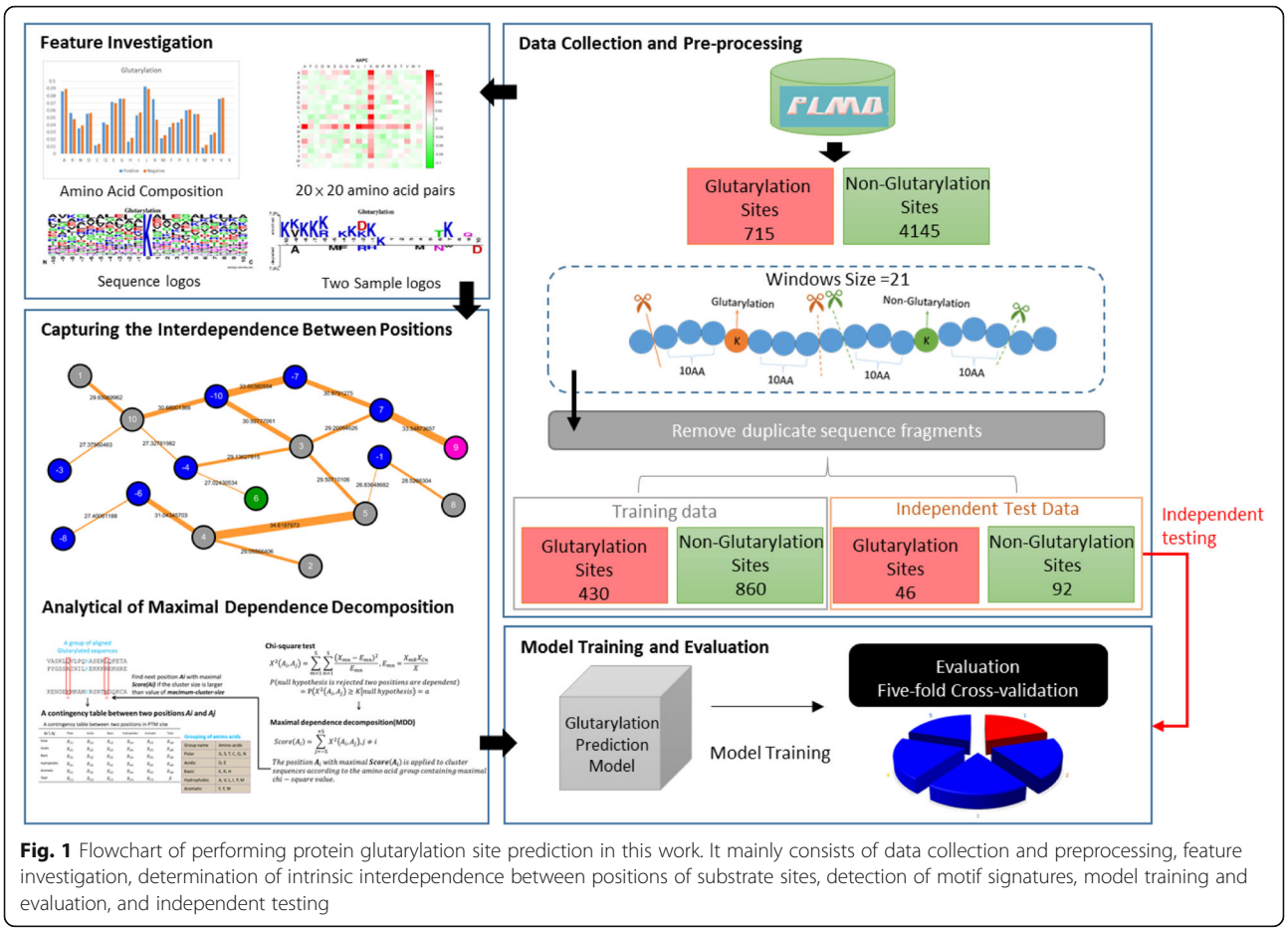

Fig. 1 Flowchart of performing protein glutarylation site prediction in this work. It mainly consists of data collection and preprocessing, feature investigation, determination of intrinsic interdependence between positions of substrate sites, detection of motif signatures, model training and evaluation, and independent testing

$$Sn = \frac{TP}{TP + FN}$$

$$Sp = \frac{TN}{TN + FP}$$

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

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

where TP, TN, FP, and FN represent the numbers of true positives, true negatives, false positives and false negatives, respectively. To evaluate how well the model distinguished between glutarylation sites and non-glutarylation sites, five-fold cross-validation was used to assess the predictive performance of the models. To compare with other approaches, we have utilized five-fold cross validation to evaluate the performance of the proposed method. Meanwhile, we also test the stability of predictive performance by using other k values (ranging from 6 to 10) for k-fold cross validation. The results presented in Additional file 2 indicated that the SVM model trained with AAC feature could provide a stable performance on different k values of k-fold cross validation. The predictive model with the best performance in the cross-validation evaluation was considered as a final model. Finally, an independent test was carried out on this final model.

Results

Composition of amino acids around glutarylation sites

To explore potential consensus motifs, the frequency of occurrence around glutarylation sites of each of the 20 amino acids was investigated based on 430 fragment sequences using a 21-mer window length. Fig 2a indicates that, at glutarylation sites, lysine (K) and arginine ® residues occur more frequently, while asparagine (N), histidine (H), methionine (M), phenylalanine (F), and proline (P) residues occur less frequently. Additionally,

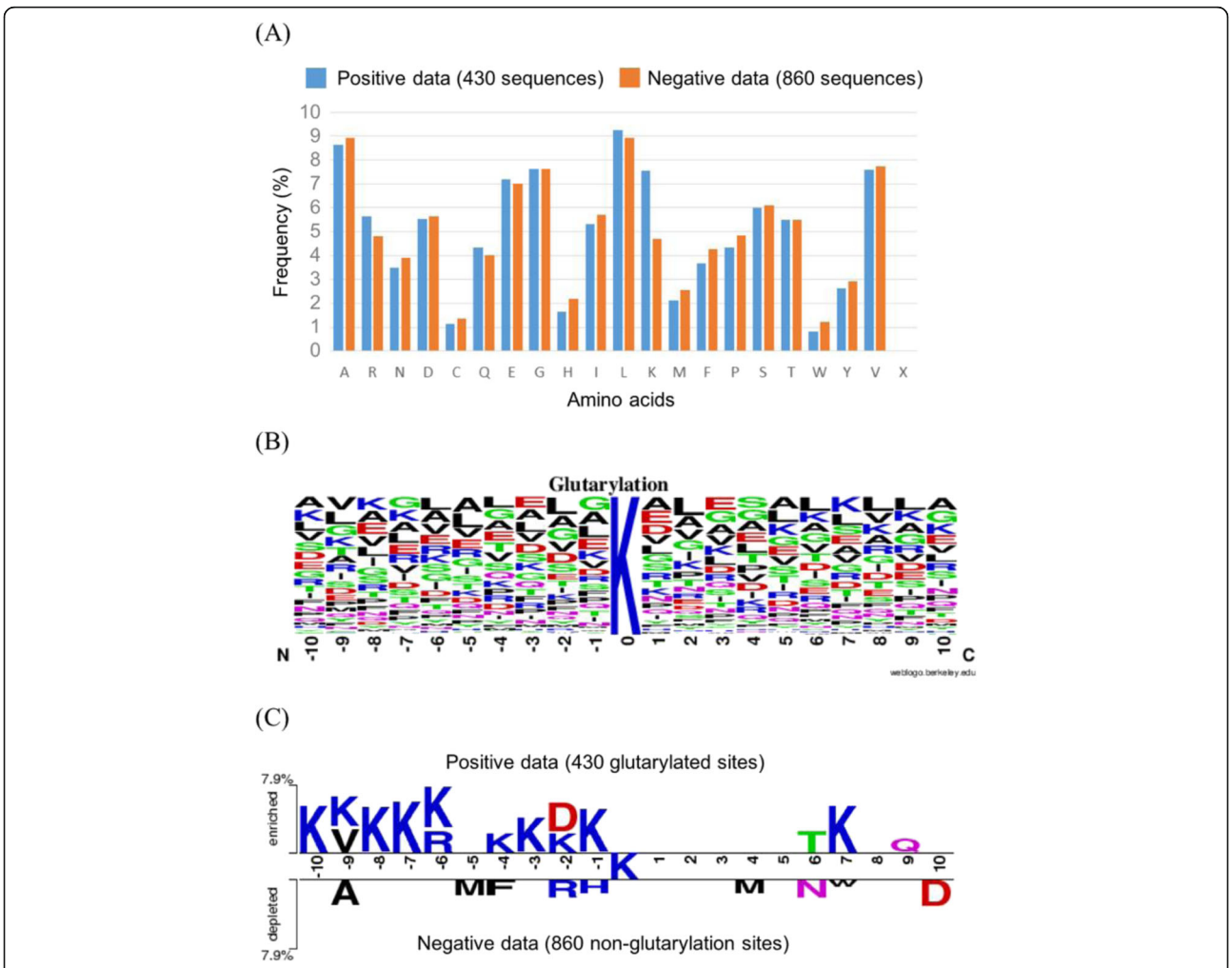

Fig. 2 Composition of amino acids around glutarylation sites. **a** Comparison of AAC between 430 positive and 860 negative sequences. **b** Position-specific AAC of 430 glutarylated sequences. **c** Comparison of position-specific AAC between glutarylated and non-glutarylated sequences based on TwoSampleLogo analysis

WebLogo [53] was utilized to compute the position-specific amino acid composition for glutarylation sites (Fig. 2b). However, it is difficult to compare the amino acid composition between glutarylation and non-glutarylation sites at a specific position. Thus, the Two-SampleLogo tool [54] was employed to detect differences in position-specific symbol compositions between the glutarylated and non-glutarylated instances. Lysine was placed in the middle of the fragment sequences, and positions of the flanking amino acids ranged from − 10 to + 10. Comparison of the 430 glutarylated sites and 860 non-glutarylated sites in Fig. 2c indicates that positively charged amino acids such as lysine (K) residues had the highest ratios at position + 7, upstream on the peptide compared to the glutarylation site (with $P < 0.01$). It also shows that polar amino acids such as threonine (T) and glutamine (Q) are slightly more abundant than expected at positions + 6 and + 9. Position − 2 was a special case, exhibiting the highest proportion of acidic residues; namely aspartate (D). This analysis shows that, in a sequence, the distance between amino acid with different properties plays a vital role in distinguishing between glutarylated and non-glutarylated sequences.

Performance evaluation of the trained models

To determine sequence-based features to discriminate between glutarylation sites and non-glutarylation sites, SVM models were built using various sequence-based features, including AAC, AAPC and CKSAAP. Each predictive model was evaluated using five-fold cross-validation based on four measures: sensitivity (Sn), specificity (Sp), accuracy (Acc), and Matthews correlation coefficient (MCC). As shown in Table 2, the SVM model trained with AAC had the highest MCC, 0.22, and relatively high sensitivity, specificity, and accuracy, with values of 0.62, 0.61, and 0.62, respectively. The SVM model trained using AAPC with a 441-dimensional vector did not perform well, with a sensitivity of 0.61, specificity of 0.48, accuracy of 0.53, and MCC value of 0.09. On the other hand, the composition of 3-spaced amino acid pairs (CKSAAP) was found to be the worst feature for predicting glutarylation sites, with a sensitivity of 0.66, specificity of 0.41, accuracy of 0.49, and MCC of 0.07.

Table 2 Five-fold cross validation results on SVM models trained with various features

Training features	Sensitivity	Specificity	Accuracy	MCC
Amino Acid Composition (AAC)	62.0%	61.3%	61.6%	0.22
Amino Acid Pair Composition (AAPC)	61.3%	48.1%	52.5%	0.09
CKSAAP[a], K = 1	62.0%	51.7%	55.1%	0.13
CKSAAP[a], K = 2	58.8%	49.8%	52.8%	0.08
CKSAAP[a], K = 3	66.0%	41.2%	49.4%	0.07

[a]*CKSAAP* Composition of k-spaced amino acid pairs

Based on the evaluation of five-fold cross-validation, Fig. 3 presents the comparison of ROC curves between the predictive models trained using various features. The results indicated that the SVM model trained using the AAC feature yielded the best prediction outcomes. To examine the robustness of the predictions from the classification methods, the five-fold cross-validation results of the models were trained by random-forest and decision tree, which also have been provided in Additional files 3 and 4, respectively. According to various evaluation criteria, the SVM model trained using AAC displayed the best overall performance among various predictive models.

The intrinsic interdependence between positions in substrate sites

In this study, we used a dependency graph method to fully display the intrinsic interdependence between positions in glutarylation sites. As shown in Fig. 4, the highest chi-square test result was obtained between position + 4 and + 5, indicating a strong interdependence between the two positions. As inferred from Fig. 2c, positively charged amino acids are frequently found in the upstream region of glutarylated sites, at positions starting from − 10 to − 1. A strong interdependence between positions − 10 and − 7 was also observed, according to the upstream consensus region of glutarylation sites. Similarly, we found that acidic amino acids were enriched at position + 9, and that the pair of amino acids in positions + 9 and + 7 showed strong interdependence, according to the downstream consensus region flanking glutarylation sites. This implies that KAT enzymes recognize substrates likely on conserved motifs of amino acids at specific positions, in agreement with previous biological knowledge.

MDD-identified motif signatures for glutarylated substrate sites

In this study, MDDLogo was used to explore motif signatures by dividing the positive training dataset (430 sites) into six subgroups. Each subgroup possessed the potential substrate specificity, containing statistically significant dependencies of amino acid composition in specific positions. Fig. 5 provides a tree-like visualization of MDD-clustered subgroups with statistically significant motifs for the 430 non-homologous glutarylation sites. On the left subtree, one motif (subgroup Group1) was detected based on the occurrence of basic amino acids (K, R, and H) at position − 8, with the highest dependence value among all the MDD-clustered subgroups. In parallel, the remaining dataset (328 sites) was further examined for maximal dependency in the occurrence of

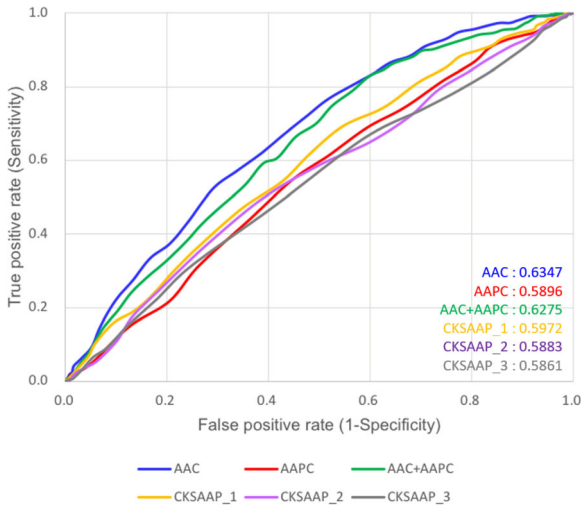

Fig. 3 Comparison of ROC curves among the SVM models trained using various features based on five-fold cross-validation

amino acids at other positions. Subgroup Glutar2 (59 sites) had a similar motif of basic amino acids at position − 6. This result was consistent with the observation in two-sample logo, that basic and hydrophobic amino acids are common in the upstream region of glutarylated lysine. Additionally, subgroups Glutar3 (60 sites) had basic amino acids at position − 10. Subgroups Glutar4 (55 sites) and Glutar5 (62 sites) had acidic amino acids at positions + 3 and + 4, respectively. Finally, the remaining 92 positive sequences formed the sixth subgroup (Glutar6), which contained a slight conservation of amino acids at positions + 3 and + 4.

Effectiveness of incorporating MDD-identified motifs into the identification of glutarylation sites

In order to evaluate the predictive power of MDD-identified substrate motifs in discriminating between glutarylation sites and non-glutarylation sites, LIBSVM was utilized to generate a predictive model for each subgroup based on AAC, the most informative feature. Based on the five-fold cross-validation, Fig. 6 provides the comparison of ROC curves between SVM model trained using all dataset and that trained from MDD-clustered subgroups. In addition, Table 3 provides the predictive sensitivity, specificity, accuracy, and MCC for each subgroup, based on their five-fold cross-validation performance. The values of ROC are also given in Table 3. It shows that subgroup Glutar1 with K/R motif at position − 8 showed a best performance at sensitivity, specificity, accuracy and MCC values of 0.81, 0.64, 0.69, and 0.42, respectively. Subgroup Glutar6 had predictive sensitivity, specificity, accuracy, and MCC values of 0.72, 0.63, 0.66, and 0.33, respectively, which is comparable to those of subgroup Glutar1. Subgroup Glutar3, whose motif was only slightly conserved, performed badly in general with relatively low sensitivity, 0.60; specificity, 0.59; accuracy, 0.60; and MCC, 0.18. Overall, the six subgroups, which contained conserved motifs of amino acids at specific positions, yielded promising accuracy as well as a balanced sensitivity and specificity. To use the

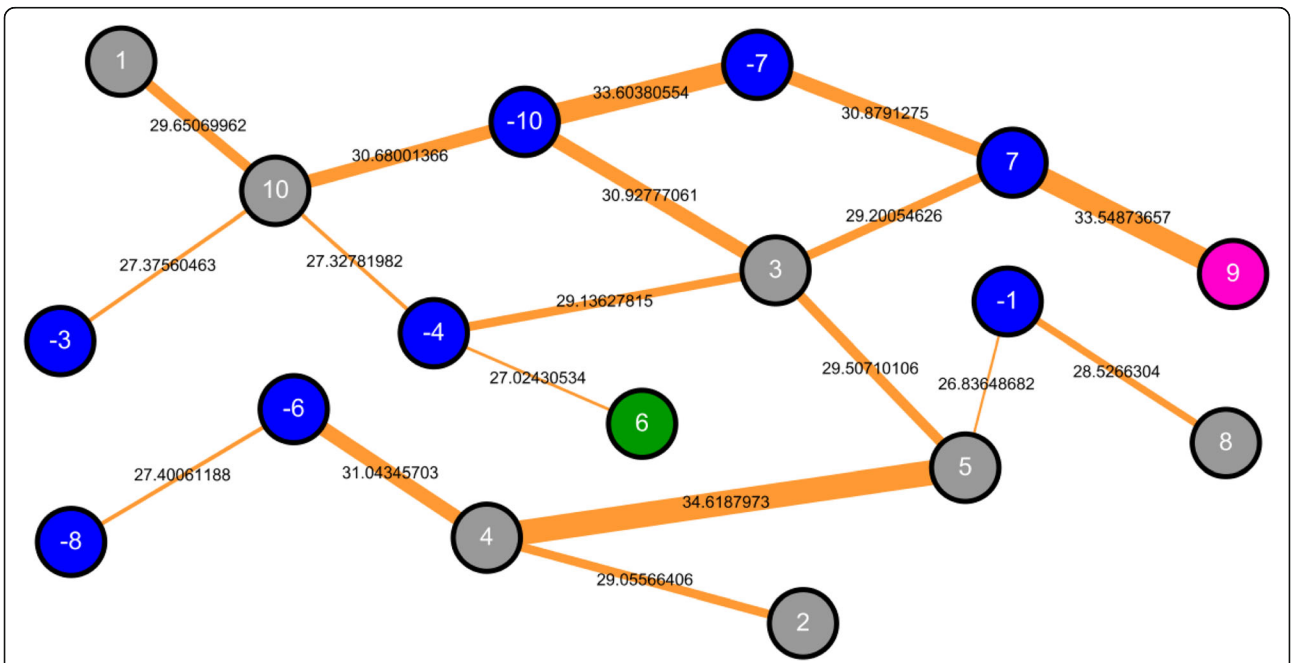

Fig. 4 The intrinsic interdependence between positions around glutarylation sites. The number shown in a circle stands for the position around glutarylation site. A higher value displayed on an edge means a more significant dependence of AAC between two circles (positions)

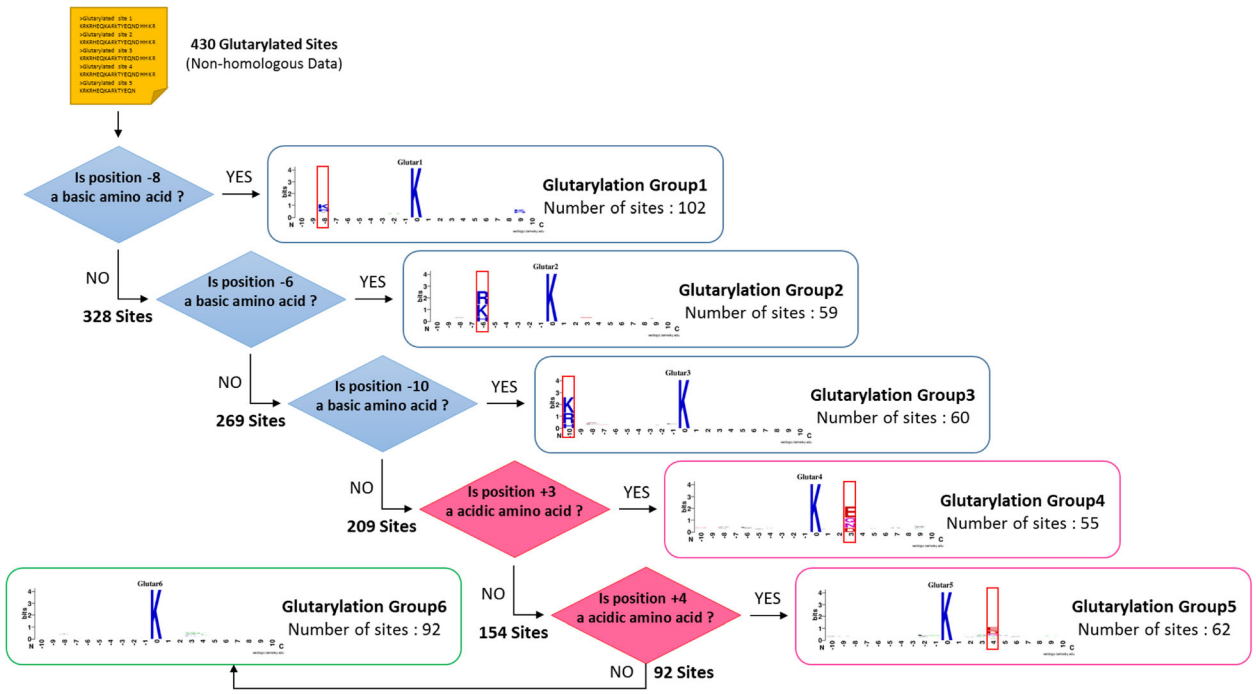

Fig. 5 A hierarchical MDD-clustering process on the detection of motif signatures from 430 glutarylated sequences

information provided by all six motifs to identify glutarylation sites with substrate specificity, the six SVM models trained from MDD-clustered subgroups were incorporated into an integrated SVM model. The values of probability estimated from five SVM models according to a specific motif signature were combined to form an input vector for the integrated SVM classifier. As shown in Table 3, based on five-fold cross-validation, the predictive performance of the integrated SVM model was significantly improved as compared to that of the single SVM model trained from all datasets without MDD clustering. The integrated SVM model presented a sensitivity, specificity, accuracy, and MCC of 0.68, 0.62, 0.64, and 0.28, respectively. In summary, the integrated SVM model combining all MDD-identified motif signatures can enhance the performance of glutarylation site identification and could be implemented as a web-based prediction resource.

Implementation of MDDGlutar web interface

Because experimentation is a time-consuming and labor-intensive process, development of an effective prediction system can aid the study of glutarylation sites. However, no method dedicated to the characterization of potential substrate motifs of glutarylated sites currently exists. Thus, we were inspired to develop a user-friendly web tool, named MDDGlutar, for the identification of glutarylation sites with their substrate motifs. The generated SVM model, combining all MDD-identified substrate motifs and the AAC feature set, was adopted to implement the prediction function on the website. After submitting protein sequences in the FASTA format, MDDGlutar returns the prediction results, including glutarylation sites, their flanking amino acids, and the corresponding substrate motif signatures. A case study on mouse aspartate aminotransferase (UniProt ID:AATM_MOUSE) was utilized to demonstrate the effectiveness of MDDGlutar. The mouse aspartate aminotransferase contains six verified glutarylation sites at

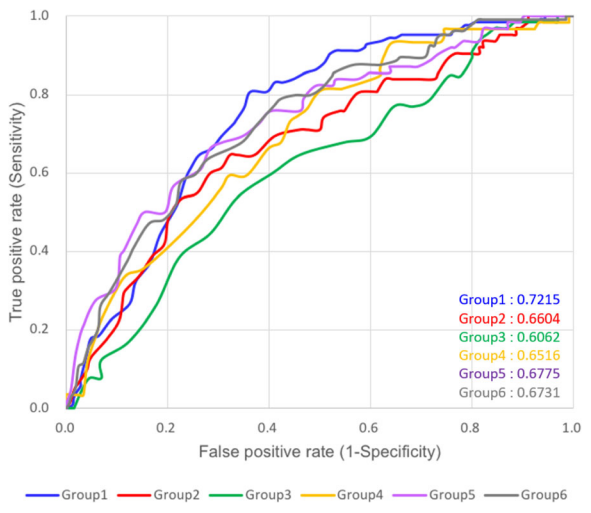

Fig. 6 ROC curves of six SVM models trained from MDD-clustered subgroups based on five-fold cross-validation

Table 3 Five-fold cross-validation results for six SVM models trained from MDD-identified motifs

Dataset	Number of positive data	Number of negative data	Sn	Sp	Acc	MCC
All Data	430	860	62.0%	61.3%	61.6%	0.22
Glutar 1	102	204	80.6%	63.7%	69.4%	0.42
Glutar 2	59	120	64.5%	67.7%	66.7%	0.31
Glutar 3	60	120	60.0%	59.2%	59.5%	0.18
Glutar 4	55	110	66.1%	60.2%	62.1%	0.25
Glutar 5	62	121	75.8%	59.7%	65.1%	0.33
Glutar 6	92	185	72.1%	62.5%	65.7%	0.33
Combined result	430	860	67.7%	61.9%	63.8%	0.28

Lys-59, Lys-90, Lys-296, Lys-302, Lys-309, and Lys-396 [9]. As presented in Fig. 7, MDDGlutar could achieve an accurate prediction at the six validated glutarylation sites, according to the corresponding motif signatures.

Discussions

An independent test set of glutarylation sites was also taken from PLMD [16], which consisted of 46 positive sites and 92 negative sites. It was used to further evaluate the predictive power of the single SVM model trained using all the training data and the integrated SVM model trained using the six MDD-identified motifs. As shown in Table 4, the single SVM model yielded a sensitivity of 0.609, a specificity of 0.685, an accuracy of 0.659, and an MCC of 0.28. Meanwhile, the performance of the integrated SVM model achieved a sensitivity of 0.652, a specificity of 0.739, an accuracy of 0.710, and an MCC of 0.38. However, the prediction ability of the proposed method, upon independent testing, showed that it can outperform other prediction methods at a specified level of false positive rate (1-specificity).

To further demonstrate the effectiveness of the proposed model, the independent test dataset was used to compare the model with existing prediction tool.

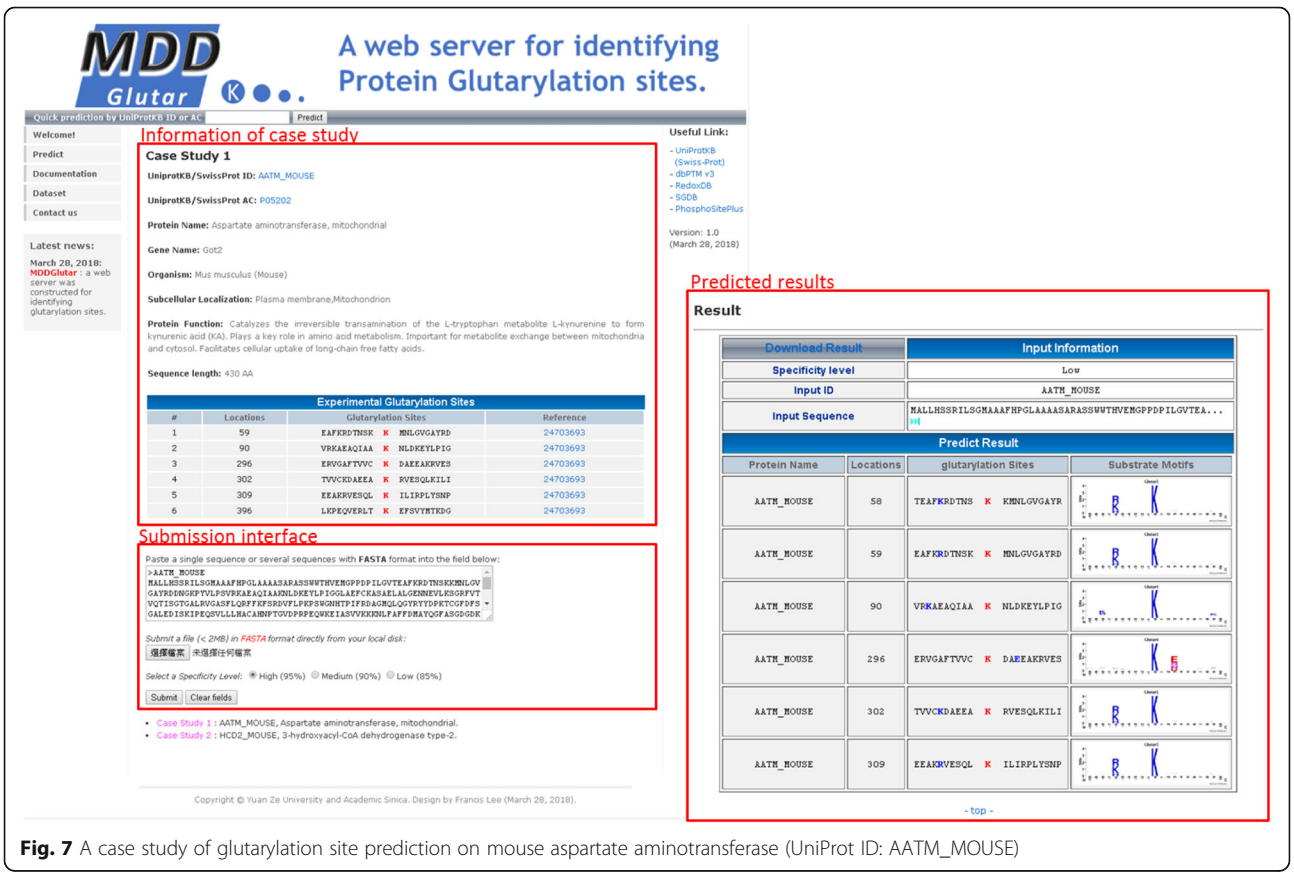

Fig. 7 A case study of glutarylation site prediction on mouse aspartate aminotransferase (UniProt ID: AATM_MOUSE)

Considering previously published prediction tools, only one predictor of glutarylation sites, GlutPred [55], was freely available. We compared our predictive performance to that of GlutPred based on independent testing datasets. As it can be seen in Table 4, GlutPred yielded higher specificity than our model, i.e., 0.91. However, the high true-negative prediction of GlutPred resulted in lower sensitivity in the identification of glutarylation sites. It is worth noting that the comparison cloud was controversial because the same data were used as source for the two studies, which means that a majority of the testing data was also present in the training data of GlutPred. However, although the proposed method could not provide better specificity than GlutPred, the results of the independent testing demonstrated that the integrated SVM model (MDDGlutar) could provide a promising performance with balanced sensitivity and specificity in the prediction of glutarylation sites.

Conclusion

In this study, we proposed a bioinformatics method for characterization and identification of glutarylation sites using substrate site specificity. The investigation using two-sample logo revealed that the most conspicuous feature of glutarylation sites is an enrichment of positively charged amino acids (K and R) upstream of the glutarylated lysine, as well as of basic amino acids at positions − 10. Based on five-fold cross-validation, the SVM model trained with the feature AAC achieved the highest sensitivity, specificity, accuracy, and MCC. As stated previously, the main purpose of this study was to explore the substrate motifs of glutarylation sites based on amino acid sequences. First, we measured the interdependence between two positions to fully capture the intrinsic interdependence in the neighboring region of glutarylation sites. After application of MDDLogo on positive training dataset, the glutarylated sequences were clustered into six subgroups corresponding to statistically significant motif signatures. The MDD-identified motifs could thus be employed to develop an integrated SVM model, significantly enhancing the predictive performance of glutarylation sites identification. An independent testing dataset was further prepared for evaluating two models: the integrated SVM model and the single SVM model without MDD implementation. The independent testing results demonstrated that the integrated model, which combined six MDD-identified motifs, provided a better predictive performance with balanced sensitivity and specificity. Consequently, the proposed model was employed to build a web-based resource, named MDDGlutar, to identify glutarylation sites and their corresponding substrate motifs.

Table 4 Performance comparison between proposed methods and an existing tool (GlutPred) based on independent testing dataset

Methods	TP	FN	TN	FP	Sn	Sp	Acc	MCC
Single SVM	28	18	63	29	60.9%	68.5%	65.9%	0.28
Integrated SVM	30	16	68	24	65.2%	73.9%	71.0%	0.38
GlutPred	25	21	84	8	54.3%	91.3%	79.0%	0.50

Additional files

Additional file 1: Table S1. Performance comparison among the SVM models trained using different window lengths. (DOCX 13 kb)

Additional file 2: Table S2. Results of k-fold cross-validation with *k* ranging from 5 to 10. (DOCX 15 kb)

Additional file 3: Table S3. Five-fold cross validation results of Random Forest models trained using various features. (DOCX 14 kb)

Additional file 4: Table S4. Five-fold cross validation results of Decision Tree classifiers trained using various features. (DOCX 14 kb)

Abbreviations

AAC: Amino acid composition; AAPC: Amino acid pair composition; Acc: Accuracy; CKSAAP: Composition of k-spaced amino acid pair; KAT: Lysine acyltransferase; KDACs: Lysine deacylases; MCC: Matthews correlation coefficient; MDD: Maximal dependence decomposition; PLMD: Protein lysine modifications database; PTM: Post-translational modification; RBF: Radial basis function; Sn: Sensitivity; Sp: Specificity; SVM: Support vector machine

Funding

Publication costs were funded by the Warshel Institute for Computational Biology, The Chinese University of Hong Kong, Shenzhen, China.

Availability of data and materials

The MDDGlutar tool and datasets used in this work can be accessed at the following URL: http://csb.cse.yzu.edu.tw/MDDGlutar/.

About this supplement

This article has been published as part of *BMC Bioinformatics Volume 19 Supplement 13, 2018: 17th International Conference on Bioinformatics (InCoB 2018): bioinformatics*. The full contents of the supplement are available online at https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-19-supplement-13.

Authors' contributions

TYL and KYH conceived and designed the experiments. KYH, HJK, JBKH and SLW performed the experiments and analyzed the data. KYH wrote the manuscript with revision by TYL and SLW. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen 518172, China. [2]Warshel Institute for Computational Biology, The Chinese University of Hong Kong, Shenzhen 518172, China. [3]Department of

Computer Science and Engineering, Yuan Ze University, Taoyuan city 320, Taiwan. [4]Department of Medical Research, Taipei Medical University Hospital, Taipei city 110, Taiwan. [5]Department of Medicine, Mackay Medical College, New Taipei City 252, Taiwan. [6]Mackay Medicine, Nursing and Management College, Taipei 112, Taiwan. [7]Department of Obstetrics and Gynecology, Hsinchu Mackay Memorial Hospital, Hsin-Chu 300, Taiwan.

Received: 23 May 2018 Accepted: 25 September 2018
Published: 4 February 2019

References

1. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S. The protein kinase complement of the human genome. Science. 2002;298(5600):1912–34.
2. Dai L, Peng C, Montellier E, Lu Z, Chen Y, Ishii H, Debernardi A, Buchou T, Rousseaux S, Jin F, et al. Lysine 2-hydroxyisobutyrylation is a widely distributed active histone mark. Nat Chem Biol. 2014;10(5):365–70.
3. Choudhary C, Weinert BT, 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):536–50.
4. Chen Y, Sprung R, Tang Y, Ball H, Sangras B, Kim SC, Falck JR, Peng J, Gu W, Zhao Y. Lysine propionylation and butyrylation are novel post-translational modifications in histones. Mol Cell Proteomics. 2007;6(5):812–9.
5. Tan M, Luo H, Lee S, Jin F, Yang JS, Montellier E, Buchou T, Cheng Z, Rousseaux S, Rajagopal N, et al. Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. Cell. 2011;146(6):1016–28.
6. Hirschey MD, Zhao Y. Metabolic regulation by lysine Malonylation, Succinylation, and Glutarylation. Mol Cell Proteomics. 2015;14(9):2308–15.
7. Kebede AF, Nieborak A, Shahidian LZ, Le Gras S, Richter F, Gomez DA, Baltissen MP, Meszaros G, Magliarelli HF, Taudt A, et al. Histone propionylation is a mark of active chromatin. Nat Struct Mol Biol. 2017; 24(12):1048–56.
8. 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.
9. Tan M, Peng C, Anderson KA, Chhoy P, Xie Z, Dai L, Park J, Chen Y, Huang H, Zhang Y, et al. Lysine glutarylation is a protein posttranslational modification regulated by SIRT5. Cell Metab. 2014;19(4):605–17.
10. Sharma R, Raicar G, Tsunoda T, Patil A, Sharma A. OPAL: prediction of MoRF regions in intrinsically disordered protein sequences. Bioinformatics. 2018; 34(11):1850–8.
11. Lopez Y, Sharma A, Dehzangi A, Lal SP, Taherzadeh G, Sattar A, Tsunoda T. Success: evolutionary and structural properties of amino acids prove effective for succinylation site prediction. BMC Genomics. 2018;19(Suppl 1):923.
12. Lopez Y, Dehzangi A, Lal SP, 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.
13. Dehzangi A, Lopez Y, Lal SP, 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(2):e0191900.
14. Dehzangi A, Lopez Y, Lal SP, Taherzadeh G, Michaelson J, Sattar A, Tsunoda T, Sharma A. PSSM-Suc: accurately predicting succinylation using position specific scoring matrix into bigram for feature extraction. J Theor Biol. 2017; 425:97–102.
15. Lee TY, Lin ZQ, Hsieh SJ, Bretana NA, Lu CT. Exploiting maximal dependence decomposition to identify conserved motifs from a group of aligned signal sequences. Bioinformatics. 2011;27(13):1780–7.
16. Lin H, Su X, He B. Protein lysine acylation and cysteine succination by intermediates of energy metabolism. ACS Chem Biol. 2012;7(6):947–60.
17. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22(13):1658–9.
18. Bui VM, Weng SL, Lu CT, Chang TH, Weng JT, Lee TY. SOHSite: incorporating evolutionary information and physicochemical properties to identify protein S-sulfenylation sites. *BMC Genomics*. 2016;17 Suppl 1:9.
19. Sahu SS, Panda G. A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction. Comput Biol Chem. 2010;34(5–6):320–7.
20. Park KJ, Kanehisa M. Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. Bioinformatics. 2003;19(13):1656–63.
21. Wong YH, Lee TY, Liang HK, Huang CM, Wang TY, Yang YH, Chu CH, Huang HD, Ko MT, Hwang JK. KinasePhos 2.0: a web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns. Nucleic Acids Res. 2007;35(Web Server issue):W588–94.
22. Chen K, Kurgan LA, Ruan J. Prediction of flexible/rigid regions from protein sequences using k-spaced amino acid pairs. BMC Struct Biol. 2007;7:25.
23. Zencheck WD, Xiao H, Weiss LM. Lysine post-translational modifications and the cytoskeleton. Essays Biochem. 2012;52:135–45.
24. Wang XB, Wu LY, Wang YC, Deng NY. Prediction of palmitoylation sites using the composition of k-spaced amino acid pairs. Protein Eng Des Sel. 2009;22(11):707–12.
25. Zhao X, Zhang W, Xu X, Ma Z, Yin M. Prediction of protein phosphorylation sites by using the composition of k-spaced amino acid pairs. PLoS One. 2012;7(10):e46302.
26. Tung CW. Prediction of pupylation sites using the composition of k-spaced amino acid pairs. J Theor Biol. 2013;336:11–7.
27. Hasan MM, 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(6):e0129635.
28. Ju Z, Cao JZ. Prediction of protein N-formylation using the composition of k-spaced amino acid pairs. Anal Biochem. 2017;534:40–5.
29. Ju Z, He JJ. Prediction of lysine crotonylation sites by incorporating the composition of k-spaced amino acid pairs into Chou's general PseAAC. J Mol Graph Model. 2017;77:200–4.
30. Chen TM, Lu CC, Li WH. Prediction of splice sites with dependency graphs and their expanded bayesian networks. Bioinformatics. 2005;21(4):471–82.
31. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. J Mol Biol. 1997;268(1):78–94.
32. Peng C, Lu Z, Xie Z, Cheng Z, Chen Y, Tan M, Luo H, Zhang Y, He W, Yang K, et al. The first identification of lysine malonylation substrates and its regulatory enzyme. Mol Cell Proteomics. 2011;10(12):M111 012658.
33. Park J, Chen Y, Tishkoff DX, Peng C, Tan M, Dai L, Xie Z, Zhang Y, Zwaans BM, Skinner ME, et al. SIRT5-mediated lysine desuccinylation impacts diverse metabolic pathways. Mol Cell. 2013;50(6):919–30.
34. Lee TY, Chen YJ, Lu CT, Ching WC, Teng YC, Huang HD, Chen YJ. dbSNO: a database of cysteine S-nitrosylation. Bioinformatics. 2012;28(17):2293–5.
35. Chen YJ, Lu CT, Su MG, Huang KY, Ching WC, Yang HH, Liao YC, Chen YJ, Lee TY. dbSNO 2.0: a resource for exploring structural environment, functional and disease association and regulatory network of protein S-nitrosylation. Nucleic Acids Res. 2015;43(Database issue):D503–11.
36. Bretana NA, Lu CT, Chiang CY, Su MG, Huang KY, Lee TY, Weng SL. Identifying protein phosphorylation sites with kinase substrate specificity on human viruses. PLoS One. 2012;7(7):e40694.
37. Lee TY, Chen YJ, Lu TC, Huang HD, Chen YJ. SNOSite: exploiting maximal dependence decomposition to identify cysteine S-nitrosylation with substrate site specificity. PLoS One. 2011;6(7):e21849.
38. Lee TY, Bretana NA, Lu CT. PlantPhos: using maximal dependence decomposition to identify plant phosphorylation sites with substrate site specificity. BMC Bioinformatics. 2011;12:261.
39. Huang HD, Lee TY, Tzeng SW, Wu LC, Horng JT, Tsou AP, Huang KT. Incorporating hidden Markov models for identifying protein kinase-specific phosphorylation sites. J Comput Chem. 2005;26(10):1032–41.
40. Huang HD, Lee TY, Tzeng SW, Horng JT. KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites. Nucleic Acids Res. 2005;33(Web Server issue):W226–9.
41. Chen YJ, Lu CT, Huang KY, Wu HY, Chen YJ, Lee TY. GSHSite: exploiting an iteratively statistical method to identify s-glutathionylation sites with substrate specificity. PLoS One. 2015;10(4):e0118752.
42. Wu HY, Lu CT, Kao HJ, Chen YJ, Chen YJ, Lee TY. Characterization and identification of protein O-GlcNAcylation sites with substrate specificity. BMC Bioinformatics. 2014;15 Suppl 16:S1.
43. Lu CT, Lee TY, Chen YJ. An intelligent system for identifying acetylated lysine on histones and nonhistone proteins. Biomed Res Int. 2014;2014: 528650.
44. Huang KY, Lu CT, Bretana N, Lee TY, Chang TH. ViralPhos: incorporating a recursively statistical method to predict phosphorylation sites on virus proteins. BMC Bioinformatics. 2013;14 Suppl 16:S10.
45. Vapnik VN. An overview of statistical learning theory. IEEE Trans Neural Netw. 1999;10(5):988–99.
46. Chang CC, Lin CJ. LIBSVM: A Library for Support Vector Machines. Acm T Intel Syst Tec. 2011;2(3):27.

47. Lu CT, Chen SA, Bretana NA, Cheng TH, Lee TY. Carboxylator: incorporating solvent-accessible surface area for identifying protein carboxylation sites. J Comput Aided Mol Des. 2011;25(10):987–95.
48. Lee TY, Chen SA, Hung HY, Ou YY. Incorporating distant sequence features and radial basis function networks to identify ubiquitin conjugation sites. PLoS One. 2011;6(3):e17331.
49. Kumari B, Kumar R, Kumar M. PalmPred: an SVM based palmitoylation prediction method using sequence profile information. PLoS One. 2014;9(2):e89246.
50. Chang WC, Lee TY, Shien DM, Hsu JB, Horng JT, Hsu PC, Wang TY, Huang HD, Pan RL. Incorporating support vector machine for identifying protein tyrosine sulfation sites. J Comput Chem. 2009;30(15):2526–37.
51. Kao HJ, Weng SL, Huang KY, Kaunang FJ, Hsu JB, Huang CH, Lee TY. MDD-carb: a combinatorial model for the identification of protein carbonylation sites with substrate motifs. BMC Syst Biol. 2017;11(Suppl 7):137.
52. Weng SL, Kao HJ, Huang CH, Lee TY. MDD-palm: identification of protein S-palmitoylation sites with substrate motifs based on maximal dependence decomposition. PLoS One. 2017;12(6):e0179529.
53. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 2004;14(6):1188–90.
54. Vacic V, Iakoucheva LM, Radivojac P. Two sample logo: a graphical representation of the differences between two sets of sequence alignments. Bioinformatics. 2006;22(12):1536–7.
55. Ju Z, He JJ. Prediction of lysine glutarylation sites by maximum relevance minimum redundancy feature selection. Anal Biochem. 2018;550:1–7.