PLOS ONE

RESEARCH ARTICLE

GSHSite: Exploiting an Iteratively Statistical Method to Identify S-Glutathionylation Sites with Substrate Specificity

Yi-Ju Chen[1], Cheng-Tsung Lu[2], Kai-Yao Huang[2], Hsin-Yi Wu[1], Yu-Ju Chen[1]*, Tzong-Yi Lee[2,3]*

1 Institute of Chemistry, Academia Sinica, Taipei, Taiwan, **2** Department of Computer Science and Engineering, Yuan Ze University, Taoyuan, Taiwan, **3** Innovation Center for Big Data and Digital Convergence, Yuan Ze University, Taoyuan, Taiwan

* francis@saturn.yzu.edu.tw (TYL); yujuchen@gate.sinica.edu.tw (YJC)

Citation: Chen Y-J, Lu C-T, Huang K-Y, Wu H-Y, Chen Y-J, Lee T-Y (2015) GSHSite: Exploiting an Iteratively Statistical Method to Identify S-Glutathionylation Sites with Substrate Specificity. PLoS ONE 10(4): e0118752. doi:10.1371/journal.pone.0118752

Academic Editor: Dinesh Gupta, International Centre for Genetic Engineering and Biotechnology (ICGEB), INDIA

Received: May 12, 2014

Accepted: January 6, 2015

Published: April 7, 2015

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The authors sincerely appreciate the Ministry of Science and Technology of Taiwan for financially supporting this research under Contract Number MOST 103-2221-E-155-020-MY3 and 103-2633-E-155-002 to TYL and 100-2628-M-001-003-MY4 to YiJC. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

S-glutathionylation, the covalent attachment of a glutathione (GSH) to the sulfur atom of cysteine, is a selective and reversible protein post-translational modification (PTM) that regulates protein activity, localization, and stability. Despite its implication in the regulation of protein functions and cell signaling, the substrate specificity of cysteine *S*-glutathionylation remains unknown. Based on a total of 1783 experimentally identified *S*-glutathionylation sites from mouse macrophages, this work presents an informatics investigation on *S*-glutathionylation sites including structural factors such as the flanking amino acids composition and the accessible surface area (ASA). TwoSampleLogo presents that positively charged amino acids flanking the *S*-glutathionylated cysteine may influence the formation of *S*-glutathionylation in closed three-dimensional environment. A statistical method is further applied to iteratively detect the conserved substrate motifs with statistical significance. Support vector machine (SVM) is then applied to generate predictive model considering the substrate motifs. According to five-fold cross-validation, the SVMs trained with substrate motifs could achieve an enhanced sensitivity, specificity, and accuracy, and provides a promising performance in an independent test set. The effectiveness of the proposed method is demonstrated by the correct identification of previously reported *S*-glutathionylation sites of mouse thioredoxin (TXN) and human protein tyrosine phosphatase 1b (PTP1B). Finally, the constructed models are adopted to implement an effective web-based tool, named GSHSite (http://csb.cse.yzu.edu.tw/GSHSite/), for identifying uncharacterized GSH substrate sites on the protein sequences.

Introduction

S-glutathionylation is a redox-dependent PTM that involves the covalent attachment of glutathione (GSH) to the thiol group of cysteine residues and regulates by physiological GSH level and GSH/GSSG ratio through spontaneous or enzymatic reaction [1–5]. In addition to

Competing Interests: The authors have declared that no competing interests exist.

regulating redox signaling, *S*-glutathionylation also serves to modulate cancer migration, cell death and survival, energy metabolism and glycolysis, as well as protein folding and degradation from bacteria to human [2]. The various targets of *S*-glutathionylation also participate in pathogenesis of many diseases, such as neurodegenerative diseases, metabolic disorders and cancers [6,7]. Due to the labile nature and low abundance of *in vivo S*-glutathionylation, the detail characteristics and mechanisms of *S*-glutathionylation still await to be clarified. To our knowledge, the protein database of human, mouse, or rat possess only consist of approximate 2% cysteine residues. However, only cysteine residue containing lower pKa in a basic environment perhaps in close three-dimensional proximity to Arg, Lys, or His residues is more accessible by GSH modification [8].

To date, numerous methods have been directed toward mass spectrometry-based *S*-glutathionylomics using various biological systems to investigate and identify more than thousands of *S*-glutathionylated targets and sites [9–11]. As for the growing number of experimentally-identified *S*-glutathionylated peptides, a curated database is of urgent need to facilitate further biological investigation of *S*-glutathionylated proteins and the substrate specificities of *S*-glutathionylation sites. Bioinformatic approaches are powerful tools for prediction of the susceptibility of individual cysteine residues to *S*-glutathionylation. Although several algorithms and public servers have been developed to analyze and predict the reactive state of cysteine [12,13] and oxidative yet disulfide cysteines [14,15], few specific information of *S*-glutathionylation targets and sites is reported. Sun et al. have computationally identified *S*-glutathionylation motifs by functional annotation and *S*-glutathionylation sites prediction by collecting 43 experimentally *S*-glutathionylated proteins and 227 corresponding sites [16]. Other potential novel consensus *S*-glutathionylation motifs and substrate site specificities remains unclear.

To further investigate potential *S*-glutathionylation motifs in primary amino acid sequence, the *in silico* characterization, i.e. amino acid composition (AAC) and accessible surface area (ASA), of protein *S*-glutathionylation sites is applied to distinguish the *S*-glutathionylation sites *versus* non-*S*-glutathionylation sites. In this study, we anticipate to characterize the *S*-glutathionylation sites with the consideration of substrate specificity of GSH. This study presents a statistical method for identifying *S*-glutathionylation sites and potential consensus motifs by maximal dependence decomposition (MDD) [17]. With the application of MDD, a large group of aligned sequences can be moderated into subgroups that capture the most significant dependencies between positions. By further evaluation using five-fold cross-validation, the support vector machine (SVM) models trained with MDD-clustered subgroups could improve predictive accuracy when compared to the model without MDD clustering. Moreover, the experimental *S*-glutathionylation data from published database (independent set) are used to test the effectiveness of the models in cross-validation. To facilitate the study of protein *S*-glutathionylation, the identified substrate motifs were exploited to implement a web-based resource for identifying *S*-glutathionylation sites with potential motifs.

Materials and Methods

Data collection and preprocessing of training set and independent test set

With the MS-based high-throughput *S*-glutathionylomic data, the experimentally verified *S*-glutathionylated cysteines from mouse macrophages [18] constituted the positive data of training set, and non-*S*-glutathionylated cysteines on these *S*-glutathionylated proteins were used as the negative data. As shown in Table 1, 1783 positive and 8423 negative data on 1005 *S*-glutathionylated proteins were obtained. In order to avoid a biased prediction performance for a

Table 1. Data statistics of training set and independent testing set.

Species	S-glutathionylation sites (Positive data)	Non-S- glutathionylation sites (Negative data)
Training set		
Su et al., 2014 (PMID: 24333276)		
Mouse	1783	8423
Independent testing set		
RedoxDB		
Mouse	20	186
Other	222	887
SGDB		
Mouse	4	62
Other	71	327
Combined non-redundant database		
Mouse	19	213
Other	254	1054

doi:10.1371/journal.pone.0118752.t001

binary classification between positive and negative data, the negative training data was balanced with the positive training data. A *K*-means clustering method based on sequence identity [19,20] was employed for acquiring a subset that represented the whole negative data set. The number of corresponding positive data was set as the value of *K*, which denoted the number of samples obtained from the negative set. This resulted in an equal number of positive and negative sequence fragments from the training data (Table 1).

For independent testing set, the experimentally verified *S*-glutathionylation sites were mainly extracted from RedoxDB [14] and SGDB [16]. A total of 20 *S*-glutathionylated cysteines on 11 proteins from mouse, extracted from RedoxDB were used as the positive data set, while the remaining 186 non-*S*-glutathionylated cysteines on these proteins functioned as the negative data set. SGDB contributes 4 *S*-glutathionylated cysteines on 4 proteins from mouse (positive data set), while other 62 non-*S*-glutathionylated cysteines were used as the negative data set. This study focused on the sequence-based analysis of substrate specificity of cysteine *S*-glutathionylation. The ability to distinguish the *S*-glutathionylated cysteine from the non-*S*-glutathionylated cysteine of the identified motifs would be evaluated based on cross-validation.

After the cross-validation of training set, the model with highest accuracy was further evaluated by using an independent test set. However, the positive data of independent test set may include the sequences that were homologous to training data. As for classification, the prediction performance of the trained models may be overestimated owing to the over-fitting of a training set. To this, the homologous sequences between training set and independent test set were removed. With reference to the reduction of the homology of the training set in MASA [19], two *S*-glutathionylated protein sequences with more than 30% identity were defined as homologous sequences. Two homologous sequences were specified to re-align the fragment sequences using a window length of $2n+1$, centered on the *S*-glutathionylation sites using BL2SEQ [21]. For two fragment sequences with 100% identity, only one *S*-glutathionylation site on homologue fragment sequence in training set was kept while the other in testing set was discarded. Redundancy was removed by retaining only one record in the event of finding multiple records of the same site position and accession number. The non-redundant negative data were generated using the same approach as positive one. After the removal of redundant data, 254 positive sequence fragments and 1054 negative sequence fragments with cysteine residues were obtained for independent testing.

Features investigation

Aside from the composition of flanking amino acids (AA), the **accessible surface areas** (ASA) around the *S*-glutathionylation sites were also investigated. Amino acid sequences with a cysteine in the center were individually extracted from positive and negative training sets using a window of length $2n+1$ centered on substrate sites, where n was set to ten in this study. An orthogonal binary coding scheme was adopted to transform amino acids into numeric vectors, in the so-called 20-dimensional binary coding. For example, glycine was encoded as "10000000000000000000;" alanine was encoded as "01000000000000000000," and so on. The number of feature vectors represented the flanking amino acids surrounding the *S*-glutathionylation site was $(2n+1) \times 20$. A total of p vectors $\{x_i, i = 1,\ldots, p\}$ were used to represent all p sequence fragments in the training data. Each vector in positive or negative cysteines was labeled with the class of its corresponding protein (e.g. positive or negative). For the composition of 20 amino acids surrounding the *S*-glutathionylation sites, the vector x_i had 20 elements for the amino acid composition (AAC) and 441 elements for the amino acid pair composition (AAPC). The 20 elements were defined as the occurrence frequencies of 20 amino acids in a sequence fragment, and the 400 elements were defined as the occurrence frequencies of 400 amino acid pairs in a sequence fragment. When the fragment sequences at N- or C-terminus are less than 21-mer, non-existing residues were filled with "X" in the corresponding position. A total of 21 types of amino acids and 441 types of amino acid pairs were presented in our setting. Using the BLOcks SUbstitution Matrix (BLOSUM62) matrix [22], the given substitution scores were derived from the alignments of amino acid sequences that had no more than 62% identity between two peptide sequences with 21 amino acids. S1 Fig. displays in detail how to generate the 441 AAPC combining BLOSUM62 features for each sequence fragment.

Refer to the method of SulfoSite [23], the positional weighted matrix (PWM) of amino acids around the *S*-glutathionylated cysteines was determined using non-homologous training data. The PWM specified the relative frequency of amino acids that surrounded the *S*-glutathionylation sites, and was utilized in encoding the fragment sequences. A matrix of $m \times w$ elements was used to represent each residue of a training dataset, where w stands for the window size and m consists of 21 elements including 20 types of amino acids and one for terminal signal. In addition, WebLogo [24,25] was adopted to generate the graphical sequence logo for the relative frequency of the corresponding amino acid at each position around the *S*-glutathionylation sites.

In the viewpoint of structural environment, several amino acid residues of a protein can be mutated without changing its structure, and two proteins may have similar structures with different amino acid compositions. Position Specific Scoring Matrix (PSSM) profiles, which have been extensively utilized in protein secondary structure prediction, subcellular localization and other bioinformatics problems [20,26–28], are adopted herein with significant improvement. The PSSM profiles were obtained by PSI-BLAST [29] against non-redundant sequences of *S*-glutathionylation sites. The matrix of $(2n+1)\times20$ elements had rows centered on substrate site, extracted from the PSSM profile, where $2n+1$ represented the window size and 20 represented the position specific scores for each type of amino acid. After that, the $(2n+1)\times20$ matrix was transformed into a 20×20 matrix by summing up the rows that were associated with the same type of amino acid. Finally, every element in 20×20 matrix was divided by the window length $2n+1$ and then normalized using the formula: $\frac{1}{1+e^{-x}}$.

A side-chain of amino acid that undergoes post-translational modification prefers to be accessible on the surface of a protein [30]. Thus, the solvent-**accessible surface area** (ASA) was used to evaluate the characteristics of *S*-glutathionylation sites. Since most of the experimental *S*-glutathionylated proteins did not have corresponding protein tertiary structures in PDB [31],

an effective tool, RVP-Net [32,33], was applied to compute the ASA value from the protein sequence. RVP-net applied a neutral network to predict the real ASA of residues based on information about their neighborhood. The measurement was with a mean absolute error of 18.0–19.5%, which was defined as the absolute difference between the predicted and experimental values of relative ASA per residue [33]. The computed ASA was the percentage of the solvent-accessible area of each amino acid on the protein. The full-length protein sequences with experimentally identified *S*-glutathionylation sites were inputted to RVP-Net to compute the ASA value of all of the residues. The ASA values of amino acids around the *S*-glutathionylation sites were extracted and normalized to be between zero and one.

Data clustering by maximal dependence decomposition

The aim of this study was to investigate the motifs of *S*-glutathionylation sites based on the amino acid sequences. Due to the difficulty of detecting the conserved motifs for the sequence data with a larger size, this work applied maximal dependence decomposition (MDD) [17] to cluster all sequences of *S*-glutathionylation site into subgroups, which had statistically obvious motifs. MDDLogo has been reported that the grouping of protein sequences into smaller groups is prior to computationally identifying PTM sites [34–40]. As illustrated in Fig 1A, the sequence fragments of GSH sites are extracted using a window of length $2n+1$. Since n is set to

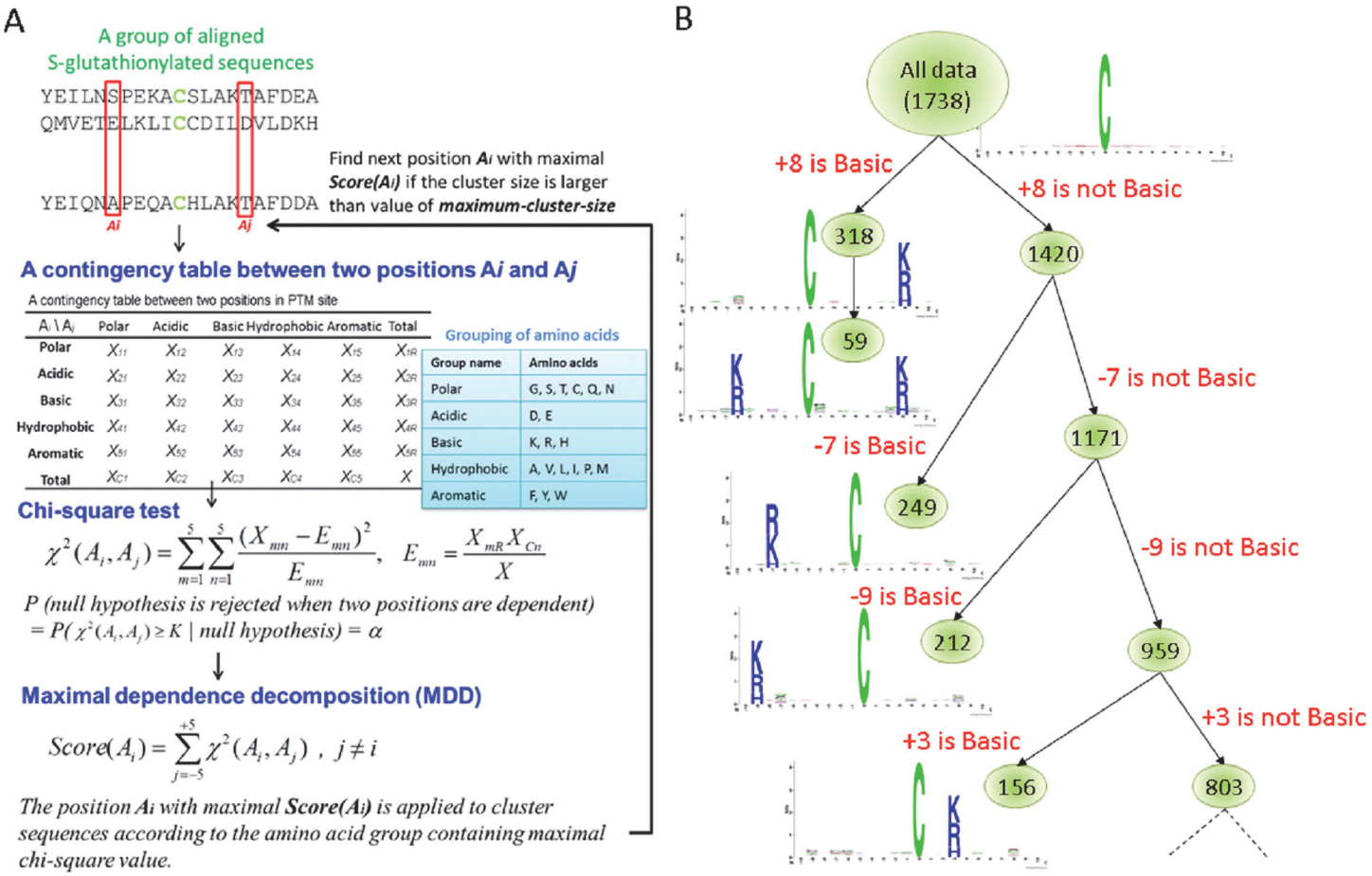

Fig 1. The analytical flowchart of MDDLogo application. (A) Using chi-square test to detect the maximal dependence of position, and (B) Tree-like visualization of MDDLogo-clustering result.

doi:10.1371/journal.pone.0118752.g001

10, the position *Ai* is defined from the range of -10 to +10. Then, a contingency table of the amino acids occurrence between two positions *Ai* and *Aj* is generated. In order to extract the motifs that had conserved biochemical property of amino acids, the 20 types of amino acids were categorized into five groups, including polar, acidic, basic, hydrophobic, and aromatic groups (S1 Table). Next, MDDLogo adopted chi-square test $\chi^2(A_i, A_j)$ to evaluate the dependence of amino acid occurrence between two positions *Ai* and *Aj* surrounding the GSH sites. The chi-square test was defined as:

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

where X_{mn} represented the number of sequences that had the amino acids of group m in position A_i and had the amino acids of group n in position A_j, for each pair (A_i, A_j) with i≠j. E_{mn} was calculated as $\frac{X_{mR} \cdot X_{Cn}}{X}$, where $X_{mR} = X_{m1} + \ldots + X_{m5}$, $X_{Cn} = X_{1n} + \ldots + X_{5n}$, and X denoted the total number of sequences. If a strong dependence were detected (defined as a X^2 value was larger than 34.3, corresponding to a cutoff level of P = 0.01 with 16 degrees of freedom) between two positions, it will be proceeded as described by Burge and Karlin [17]. A dependence value for each position *Ai*, Score(A_i), is calculated as follows:

$$Score(A_i) = \sum_{j=-10}^{+10}\chi^2(A_i, A_j), j \neq i \quad (2)$$

The position *Ai* with a maximal dependence value, Score(Ai), is applied to cluster sequences. When applying MDD to cluster sequences, a parameter, i.e., the maximum-cluster-size, should be set. If the size of a subgroup is less than the cutoff value of maximum-cluster-size, the subgroup will not be divided any further. The MDD process terminates after all of the subgroup sizes are less than the value of the specified maximum-cluster-size.

MDD clustering is a recursive process that divides all sequences into tree-like subgroups. After the detection of maximal dependence of flanking positions, as the example illustrated in Fig 1B, position +8 had the maximal dependence with the occurrence of basic amino acids. Subsequently, all data can be divided into two subgroups: one had the occurrence of basic amino acids in position +8 and the other lacked the occurrence of basic amino acids in position +8. The MDD clustering was a recursively process to divide the positive sets into tree-like subgroups. When applying MDDLogo to cluster the sequences of a positive set, a parameter, i.e., the maximum-cluster-size, should be set. If the size of a subgroup was less than the maximum-cluster-size, the subgroup will not be divided any more. In order to obtain an optimal minimum cluster size, MDDLogo was executed using various values. For this investigation, each subgroup resulting from MDDLogo was represented using WebLogo [24] for determining if they presented conserved motifs for the substrate specificity of *S*-glutathionylation.

Model construction and evaluation

The support vector machine (SVM) was adopted to learn the predictive model from the positive and negative data of the training set. Based on binary classification, the concept behind SVM was to map the input samples into a higher dimensional space using a kernel function, followed by finding a hyper-plane that can discriminating the two classes with maximal margin and minimal error. A public SVM library, LIBSVM [41], was employed to generate the predictive models trained with various features. The radial basis function (RBF) $K(S_i, S_j) = \exp(-\gamma||S_i - S_j||^2)$ was used as the kernel function of the SVMs. The LIBSVM library could output a value of probability estimated ranging from 0 to 1 for each prediction. According to that, the values of

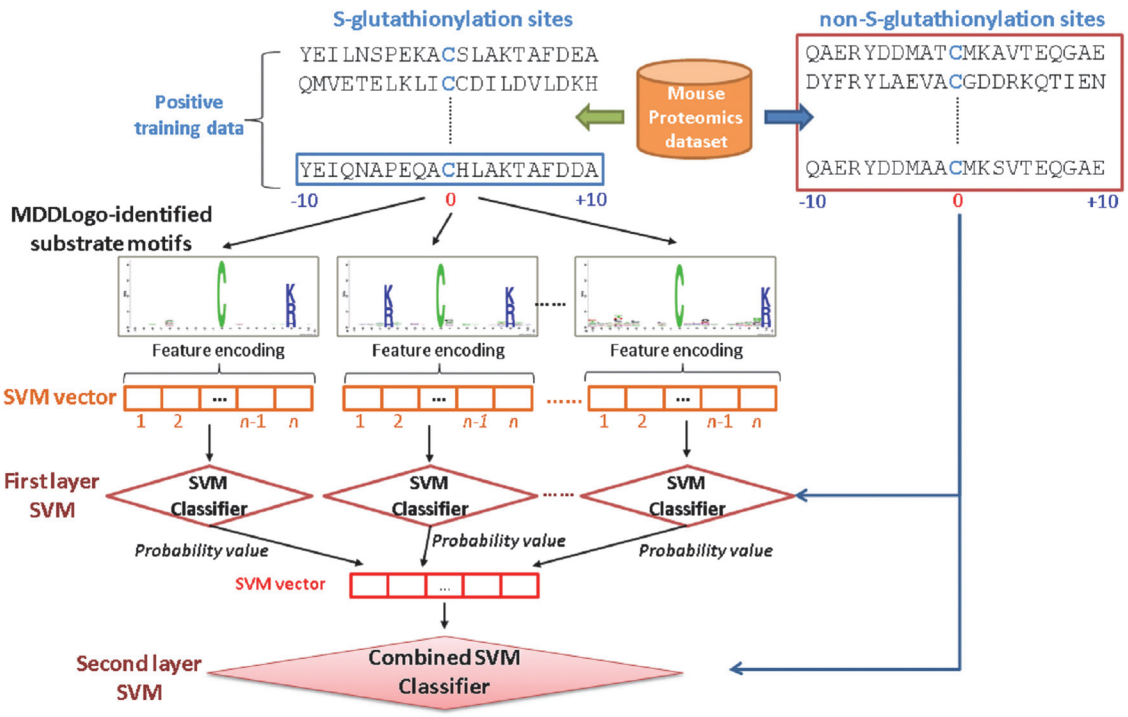

Fig 2. The conceptual diagram of two-layered SVMs trained with MDDLogo- identified substrate motifs.

doi:10.1371/journal.pone.0118752.g002

probability estimated from each SVM classifier trained with the best feature corresponding to a specific motif were adopted to form an input vector for second-layered SVM.

Prior to the construction of a final model, the predictive performance of models using different features was evaluated by performing five-fold cross validation. As shown in Fig 2, firstly, the training data was divided into five groups by splitting each dataset into five approximately equal sized subgroups. During cross-validation, one subgroup was regarded as the test set, and the remaining four subgroups were regarded as the training set. The cross-validation process was repeated five times, in which each subgroup was used as a test set once. The five validation results were then combined to produce a single estimation. The advantage of cross-validation evaluation was that all original data were regarded as both training set and testing set, and each data was used for testing exactly once [42]. The following measures were then used to gauge the predictive performance of the trained models:

$$\text{Sensitivity (Sn)} = \text{TP}/(\text{TP} + \text{FN}) \tag{3}$$

$$\text{Specificity (Sp)} = \text{TN}/(\text{TN} + \text{FP}) \tag{4}$$

$$\text{Accuracy (Acc)} = (\text{TP} + \text{TN})/(\text{TP} + \text{FP} + \text{TN} + \text{FN}) \tag{5}$$

$$\text{Matthews Correlation Coefficient (MCC)} = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}} \tag{6}$$

where TP, TN, FP and FN represented the numbers of true positives, true negatives, false positives and false negatives, respectively. After the selection of the predictive model with best

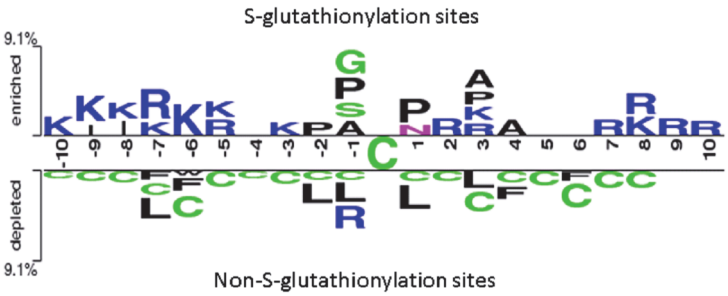

Fig 3. TwoSampleLogo presents the compositional biases of amino acids around *S*-glutathionylation sites compared to the non-*S*-glutathionylation sites in mouse macrophages. The significant amino acids around *S*-glutathionylated cysteine residue is enriched from the positive dataset and presented in upper panel ($p < 0.01$). Relatively, the high frequency of amino acids around non-*S*-glutathionylated cysteines is depleted from the negative dataset and presented in lower panel.

doi:10.1371/journal.pone.0118752.g003

performance, an independent testing was carried out to further evaluate the predictive performance of the best model (two-layered SVM).

Results and Discussion

Positively charged and higher solvent accessible amino acids neighboring with the *S*-glutathionylation site

To explore the potential consensus motifs of *S*-glutathionylation, in this study, we focused on the sequence-based analysis of substrate specificity for *S*-glutathionylation. Here, a web-based tool TwoSampleLogo [43], that detected and displayed statistically significant differences in position-specific symbol compositions between two sets of multiple sequence alignments, was applied. In the preliminarily evaluation of the amino acid frequency neighboring the *S*-glutathionylated cysteine, the non-homologous *S*-glutathionylated cysteines were centered on position 0, and the flanking amino acids (-10 ~ +10) were graphically visualized as sequence logos. Contrast between 1783 *S*-glutathionylation sites and 8423 non-*S*-glutathionylation sites, the TwoSampleLogo revealed that the most pronounced feature of *S*-glutathionylation sites was the abundance of charged amino acids, especially the positively charged Lysine (K) and Arginine (R), at positions -10 ~ -5, -3, +2, +3, and +7 ~ +10 ($p < 0.01$, Fig 3, upper panel). Another interesting feature was the absence of positively charged residues at position -2, -1, +1, and +4 that was immediately adjacent to the *S*-glutathionylation sites. Comparatively, another featured characteristic, such as neutral amino acids Leucine (L), Phenylalanine (F), and Tryptophan (W), locating around non-*S*-glutathionylated cysteines at position -7, -6, -2, -1, +1, +3, +4, and +6, was depleted in the negative dataset (Fig 3, lower panel). Cysteine (C) residues also randomly located around non-*S*-glutathionylated cysteines from -10 ~ +8. This investigation also implicated that the notable difference of amino acid characteristic in sequence located around position -7, -6, -1, and +3. The analysis also revealed that the distant amino acids in sequence had significant difference between *S*-glutathionylation sites and non-*S*-glutathionylation sites, indicating that positively charged amino acids may be close to *S*-glutathionylated cysteines in three-dimensional structure.

In addition to the composition of amino acids in linear sequences, we further used RVP-Net algorithm to analyze the correlation of *S*-glutathionylation sites and solvent **accessible surface area** (ASA). As shown in S2 Fig, the comparison of average percentage of ASA in the 21-mer window (-10 ~ +10) showed that the cysteine residues had the lowest ASA on both *S*-glutathionylated and non-*S*-glutathionylated cysteines, suggesting low preference of solvent

accessibility in *S*-glutathionylation sites. Moreover, the adjacent amino acids neighboring the centered *S*-glutathionylation sites had relatively higher preference of solvent-accessible surface area than that of non-*S*-glutathionylation sites. The result suggested that the flanking amino acids having hydrophilic characteristics may regulate the *S*-glutathionylation on cysteine residues due to the relative surface solvent accessibility.

Cross-validation performance of training features

To determine what features provide the best performance to identify the *S*-glutathionylation sites compared to the non-*S*-glutathionylation sites, the predictive models were trained with various features, such as 20D binary code, BLOSUM62, AAC, AAPC, ASA, PWM, and PSSM. Using cross validation, four predictive powers, including sensitivity (Sn), specificity (Sp), accuracy (Acc), and Matthews correlation coefficient (MCC), were also evaluated. As shown in Table 2, the SVM models that were trained with 20D binary code generated the predictive sensitivity, specificity, accuracy, and MCC at 0.63, 0.63, 0.63, and 0.20, respectively. The similar quality was presented from that analyzed by BLOSUM62, AAC, AAPC, and PSSM. However, the model trained with ASA and PWM had the lowest predictive accuracy at 0.57 and relatively lower sensitivity, specificity, and MCC at 0.55, 0.57, and 0.1, respectively, which was probably caused by the low ASA value of cysteines. The specificity of the model trained with AAPC was equal to that with BLOSUM62 at 0.66 and slightly superior to that with other features. Given that AAPC was regarded as the best feature for training a model for discrimination of 1783 *S*-glutathionylation sites, the predictive sensitivity, specificity, accuracy, and MCC of the best model were 0.65, 0.66, 0.66, and 0.24, respectively. Additionally, the predictive power of the model trained with the hybrid combination of BLOSUM62 and AAPC provided the best performance of the predictive sensitivity, specificity, accuracy, and MCC at 0.66, 0.67, 0.67, and 0.26, respectively. Thus, BLOSUM62 combined with AAPC was selected as the training feature for the construction of two-layered SVM model.

MDD-clustered substrate motifs and the cross-validation performances

To improve the detection of the conserved motifs from large-scale *S*-glutathionylation data set, we further applied the maximal dependence decomposition (MDD) to cluster all 1783 identified *S*-glutathionylated peptide sequences. Here, 12 subgroups of *S*-glutathionylation motifs can be obtained from the most significant dependencies of amino acid composition between specific positions (Table 3 and Fig 4). According to the chi-square test of the dependence of

Table 2. Five-fold cross validation results on single SVM model trained with various features.

Training features	Sn	Sp	Acc	MCC
20D Binary code	0.63	0.63	0.63	0.20
BLOSUM62	0.63	0.66	0.65	0.22
Amino Acid Composition (AAC)	0.63	0.65	0.65	0.22
Amino Acid Pair Composition (AAPC)	0.65	0.66	0.66	0.24
Accessible Surface Area (ASA)	0.55	0.57	0.57	0.10
Position Weight Matrix (PWM)	0.57	0.58	0.57	0.11
Position-specific scoring matrix (PSSM)	0.64	0.65	0.65	0.22
BLOSUM62 + AAPC	0.66	0.67	0.67	0.26

Total 1783 cysteine sequences were applied in positive and negative data. Sn, sensitivity; Sp, specificity; Acc, accuracy; MCC, Matthews Correlation Coefficient.

doi:10.1371/journal.pone.0118752.t002

Table 3. The 12 MDDLogo-clustered subgroups and their performances of five-fold cross-validations from 1783 *S*-glutathionylation sites in mouse data set.

MDDLogo Cluster	Number of positive data	C (cost)	γ (gamma)	Sn	Sp	Acc	MCC
GSH1	259	0.03125	0.0078125	0.67	0.69	0.68	0.28
GSH2	59	32768	0.5	0.71	0.76	0.75	0.39
GSH3	249	32768	0.5	0.69	0.70	0.70	0.30
GSH4	212	32768	0.5	0.68	0.69	0.69	0.29
GSH5	156	0.03125	0.0078125	0.65	0.66	0.66	0.21
GSH6	147	0.03125	0.0078125	0.70	0.71	0.71	0.33
GSH7	125	2048	8	0.72	0.75	0.74	0.37
GSH8	96	0.03125	0.0078125	0.69	0.70	0.70	0.31
GSH9	76	32768	0.5	0.74	0.80	0.79	0.45
GSH10	73	0.03125	0.0078125	0.66	0.68	0.68	0.27
GSH11	51	32768	0.5	0.67	0.71	0.70	0.29
GSH12	280	0.03125	0.0078125	0.65	0.69	0.68	0.26
All data	1783	0.03125	0.0078125	0.66	0.67	0.67	0.26
Combined MDDLogo-clustered motifs	**1783**	**0.03125**	**0.0078125**	**0.69**	**0.71**	**0.71**	**0.32**

C, cost value; γ, gamma value; Sn, sensitivity; Sp, specificity; Acc, accuracy; MCC, Matthews Correlation Coefficient.

doi:10.1371/journal.pone.0118752.t003

five amino acid groups in flanking positions, 11 out of all MDDLogo-clustered subgroups had the conserved motifs of positively charged amino acids (K, R and H) at a specific position. In particular, 59 *S*-glutathionylated peptides in the second subgroup extracted from the 259 *S*-glutathionylated peptides in first subgroup had two positively charged amino acids on conserved motifs at two specific positions -6 and +8. Based on these significant frequency of sequence logos, the result was consistent with a previous data describing that the *S*-glutathionylated cysteine located in a basic environment perhaps in close three-dimensional proximity to K/R/H is more accessible [8]. However, the 12th subgroup containing the remaining 280 *S*-glutathionylation sites which did not have any conserved motif.

Furthermore, we evaluated all of the *S*-glutathionylation sites and these 12 MDDLogo-clustered subgroups for their predictive performance by five-fold cross-validation. The predictive and average value of cross-validation performance in each subgroup was displayed in Table 3. Among them, subgroup GSH9, which had a conserved K/R/H at position -10, contained the highest predictive power at 0.74, 0.80, 0.79, and 0.45 for sensitivity, specificity, accuracy, and MCC, respectively. Moreover, the subgroup GSH2 presenting a conserved K/R/H at position -6 and +8 yielded the next best specificity, accuracy, and MCC at 0.76, 0.75, and 0.39. The predictive performance in all of these 12 subgroups of MDDLogo-clustered SVMs was presented higher sensitivity, specificity, accuracy, and MCC than that of all 1783 *S*-glutathionylation sites without any clustering. On the other hand, the SVM model trained with the combined MDDLogo-clustered motifs generated an enhanced performance of sensitivity, specificity, accuracy, and MCC at 0.69, 0.71, 0.71, and 0.32, compared with all 1783 *S*-glutathionylation sites without any clustering which contained lower performance at 0.65, 0.66, 0.66, and 0.24, respectively. This analysis indicated that the *S*-glutathionylated sequences in a large-scale data set can be alternatively clustered by MDD method, which significantly enhanced the signal of amino acids motif and improved the performance of the predictive model. Thus, the two-layered SVM model combining all MDDLogo-identified substrate motifs was utilized to implement a web-based prediction tool in website.

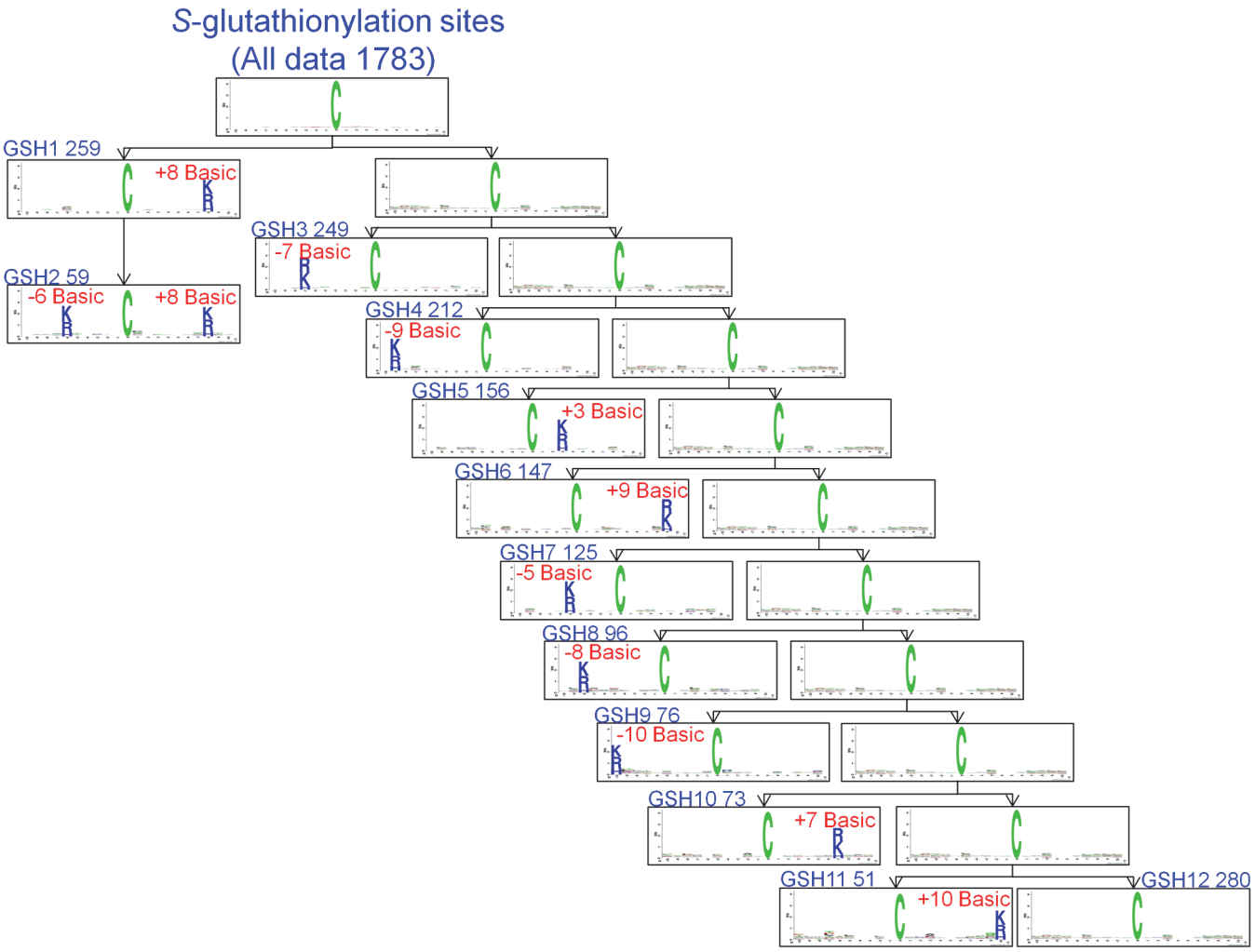

Fig 4. The MDDLogo-clustered subgroups from 1783 *S*-glutathionylation sites in mouse data set.

doi:10.1371/journal.pone.0118752.g004

Evaluation of *S*-glutathionylation predictive models using independent test set

To evaluate effectiveness of the investigated features that achieved the best accuracy in cross-validation, an independent test set of *S*-glutathionylation was used to test the MDDLogo-clustered models training. The independent test set was composed of the experimentally verified *S*-glutathionylation data from multiple species, which contains a total of 254 positive data and 1054 negative data in 170 *S*-glutathionylated proteins. As shown in Fig 5, the MDD-clustered models could perform with a sensitivity of 0.57, a specificity of 0.58, an accuracy of 0.58, and the MCC of 0.12 in independent test set. Additionally, the two-layered SVM models using all the MDDLogo-clustered substrate motifs accomplished a sensitivity of 0.81, a specificity of 0.83, an accuracy of 0.83, and the MCC of 0.56. Overall, the independent testing demonstrated that the MDD-clustered models had higher estimated specificity comparing to sensitivity. Therefore, greater prediction power can be obtained by using MDDLogo-clustered SVM models than that by single SVM model. The detailed independent testing results were presented in S2 Table.

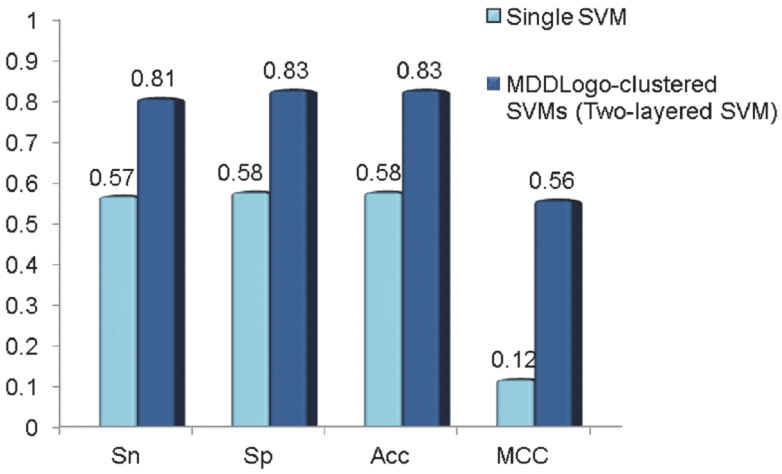

Fig 5. Comparison of independent testing performance between single SVM and MDDLogo-clustered SVM models. Sn, sensitivity; Sp, specificity; Acc, accuracy; MCC, Matthews Correlation Coefficient.

doi:10.1371/journal.pone.0118752.g005

Implementation of web-based tool for identification of *S*-glutathionylation sites

With the time-consuming and laboratory-intensive experimental workflow, even though a protein can be *S*-glutathionylated, precise identification of the *S*-glutathionylation sites on the substrate is still experimentally challenging. Therefore, developing an effective prediction tool can efficiently help identify potential *S*-glutathionylation sites. Following evaluation by cross-validation and an independent test, the MDDLogo-clustered models trained with combination of BLOSUM62 and AAPC are utilized in the construction of web-based prediction system, GSHsite. After the users submit their uncharacterized protein sequences, GSHSite efficiently returns the predictions including *S*-glutathionylated position, the flanking amino acids, and the matched MDDLogo-clustered motif. In addition, the protein sequences in FASTA format or protein name, gene name, and accession number can be used to search and predict. After submitting the information of peptide sequence or protein, the detail information and annotation of target protein, *S*-glutathionylation sites in published literature, and predicted *S*-glutathionylation motifs will be presented.

A mouse thioredoxin (TXN, THIO_MOUSE), which contains one *S*-glutathionylation site at Cys-73 [44], is used to demonstrate the effectiveness of GSHSite. As presented in Fig 6, GSHSite is able to correctly predict the experimentally verified *S*-glutathionylation site at positions 73. The matched MDD-clustered motif is also provided for the future investigation of substrate site specificity. The second case study was performed on human protein tyrosine phosphatase 1B (PTP1B, PTN1_HUMAN), which contains one *S*-glutathionylation site at cys-215 [45] and is not included in the training data set. The experimentally verified *S*-glutathionylation site at position 215 was also correctly predicted by GSHSite.

Characteristic relationship between *S*-nitrosylation and *S*-glutathionylation

Because the *S*-nitrosylation and *S*-glutathionylation can target on the same cysteine residues of proteins, here, we further investigated the relationship between these two modifications. Based on the experimentally verified data set from dbSNO [46] and the study of Su. et al. [18], total 495 consistent sites from mouse were presented in total 1783 *S*-glutathionylation sites and total

Fig 6. A case study of S-glutathionylation site prediction for mouse thioredoxin (THIO_MOUSE). The website presents (A) protein information and annotation, (B) S-glutathionylation sites from published experiment, and (C) prediction result of potential consensus motifs.

doi:10.1371/journal.pone.0118752.g006

2159 *S*-nitrosylation sites (S3 Table). We further analyzed the potential consensus motifs to explore the substrate specificity of these two modifications presenting by TwoSampleLogo. Contrast between 495 identically common sites and 1711 un-modified sites, the TwoSampleLogo revealed that the most pronounced features neighboring the modified sites were the abundance of the positively charged K/R at positions -7, +8, and +9, hydrophobic amino acids Isoleucine (I)/Proline (P) at position -6 ~ -4, +1, and +3, and polar amino acids Serine (S)/Threonine (T)

at position -10, -1, +6 and +7 ($p < 0.01$, S3A Fig, upper panel). Comparatively, three amino acids, including hydrophobic amino acid Methionine (M), locating around un-modified cysteines at position -2, positively charged amino acid K at position -1, and polar amino acid Cysteine (C) randomly located from position -6 to +5, were depleted in the negative dataset (S3A Fig lower panel). This result implicated that the positively charged amino acids in distant sequence and hydrophobic amino acids surrounding *S*-glutathionylation and *S*-nitrosylation sites have notable difference of amino acid characteristics comparing with non-modification sites. For instance, the Cys73 on thioredoxin (Trx, THIO_MOUSE) in case 1 study was categorized into C $(X)_7$K/R/H motif (GSH1 group in Table 3). Trx is one of the well-studied regulators to catalyze the transnitrosylation and denitrosylation of specific targets, including Cys215 on PTP1B (case 2 study, GSH3 group in Table 3), depending on the redox status of different cysteine residues [47–49]. In addition, Trx can also be *S*-glutathionylated at Cys73 and functioned as the deglutathionylase to regulate the enzymatic activity [50,51]. Due to the detail mechanism of *S*-nitrosylation and *S*-glutathionylation is still unclear *in vivo*, in this study, we proposed that the identified motifs for substrate specificity may help shed light to the study of the site-specific interplay between these two modifications.

After depleting the identically common cysteines, 1664 and 1288 cysteines were respectively presented in *S*-nitrosylation only and *S*-glutathionylation only (S4 Table). We further investigated the characteristic difference using TwoSampleLogo between these two modifications. This investigation also implicated that the positively charged amino acids R and polar amino acids Glutamine (Q), S, and C in sequence (around position -10 ~ -7, -4 ~ -1, +1 ~ +3, +8, and +10), had notable difference of amino acid characteristics between *S*-glutathionylation sites and *S*-nitrosylation sites (S3B Fig, upper panel). Comparatively, more hydrophobic amino acids in sequence had significant difference surrounding *S*-nitrosylation sites. The result indicated that polar and positively charged amino acids might be close to *S*-glutathionylated cysteines in three-dimensional structure.

To further understand the characteristics and categories of potential biological processes of these proteins, we also analyzed the annotation of Gene Ontology (GO) for cross talk of 328 *S*-glutathionylated and *S*-nitrosylated proteins by DAVID software ($p < 0.01$). S5 Table showed that most identically common proteins modifying by *S*-glutathionylation and *S*-nitrosylation contributed for translation and generation of precursor metabolites and energy. Moreover, more proteins involving in structural constituent of ribosome, structural molecule activity, and nucleotide binding were presented. In addition to the mitochondrial proteins, most proteins were located in cytosol and ribonucleoprotein complex. Similar biological processes and molecular functions were presented in only *S*-glutathionylated proteins (S6 Table). For the *S*-nitrosylation only, 133 of 974 proteins (14%) involved in oxidation reduction, 246 proteins (25%) played roles in nucleotide binding, and 333 (34%) proteins located in mitochondrion (S7 Table).

Conclusion

In this study, we reported a systematic informatics investigation on the *S*-glutathionylation substrate specificity from experimentally verified *S*-glutathionylomic data. The analysis of position-specific amino acids composition reveals that the most pronounced feature of *S*-glutathionylation sites is the abundance of positively charged amino acids at surrounding positions, especially on the positions from -5 to +3. This investigation also implicates that the distant amino acids in sequence (around position -7 and -6), which may be close to *S*-glutathionylation cysteines in three-dimensional structure, have notable difference between *S*-glutathionylation sites and non-*S*-glutathionylation sites. Moreover, the flanking amino acids around *S*-

glutathionylation sites have higher preference of solvent-accessible surface area than that around non-*S*-glutathionylation sites. According to the five-fold cross-validation, the model trained with the combined features of BLOSUM62 and amino acid pair composition gets the highest sensitivity, specificity, accuracy, and MCC.

Due to the abundance of experimental data, this study focuses on investigating the motifs of *S*-glutathionylation sites based on the amino acid sequences. However, it is difficult to explore the conserved motifs from large-scale S-glutathionylome data set. Thus, this work applies MDDLogo algorithm to cluster all sequences of *S*-glutathionylation site into 12 subgroups. According to the chi-square test of the dependence in flanking positions, surprisingly, all of the MDD-clustered subgroups have the conserved motifs of positively charged amino acids (K, R and H) at a specific position. Particularly, subgroups GSH2 have the conserved motifs of positively charged amino acids at two specific positions (-6 and +8). Although the newly identified motifs could not be experimentally verified, it still worthy to be noticed that MDD clustering can help the biologist investigating the potential substrate motifs of *S*-glutathionylation sites. More noteworthy is that the MDD-clustered motifs can be applied to improve the predictive power of computationally identifying *S*-glutathionylation sites with various substrate specificities. According to the evaluation of five-fold cross-validation, the models trained with combined MDD-clustered motifs are increased for the predictive accuracy of 0.71, comparing to the model trained without MDD clustering. This analysis indicates that the *S*-glutathionylated sequences with a larger size can be alternatively clustered by MDD method in order to enhance the signal of amino acids motif and improve the performance of the predictive model.

Finally, the independent testing indicates that the predictive model by MDDLogo-clustered SVMs can generate the best performance compared with the single SVM model. The acquisition of additional experimentally verified *S*-glutathionylation data is needed to re-calibrate more accurate MDD-clustered motifs. The proposed method can be improved by considering the motifs that are intrinsically included in the test data. Consequently, the models with MDD clustering method are applied to implement a novel web-based tool, named GSHSite, for identifying cysteine *S*-glutathionylation. Correct prediction on two experimentally verified *S*-glutathionylated proteins demonstrated the effectiveness of GSHSite. In this web-based tool, the detail information, annotation, and 3D structure provided from PDB of proteins are also included. This approach not only provides the prediction yet experimental *S*-glutathionylation site information, but also can be used to explore the potential substrate specificity of *S*-glutathionylation.

Availability

The proposed method is implemented as a web-based resource, which is now freely available to all interested users at http://csb.cse.yzu.edu.tw/GSHSite/. All of the data set used in this work is also available for download in the website.

Supporting Information

S1 Fig. The encoding scheme of the amino acid pair composition (AAPC) combined with BLOSUM62 feature.
(TIF)

S2 Fig. Comparison of solvent-accessible surface area between *S*-glutathionylation and non-*S*-glutathionylation sites.
(TIF)

S3 Fig. TwoSampleLogo presents the compositional biases of amino acids around *S*-glutathionylation sites compared to the *S*-nitrosylation sites in mouse dataset. (A) The identically common cysteines for *S*-glutathionylation and *S*-nitrosylation in upper panel were compared with un-modified cysteines in lower panel ($p < 0.01$). (B) The significant amino acids around *S*-glutathionylated cysteine residue were enriched from the positive dataset and presented in upper panel ($p < 0.01$). Relatively, the high frequency of amino acids around *S*-nitrosylated cysteines were depleted from the negative dataset and presented in lower panel.
(TIF)

S1 Table. The amino acids group of MDDLogo used in this study.
(DOCX)

S2 Table. The detailed results of training and independent testing comparison between our method.
(DOCX)

S3 Table. The number of proteins and sites in each *S*-glutathionylation and *S*-nitrosylation data.
(DOCX)

S4 Table. The detail information in *S*-glutathionylation and *S*-nitrosylation data by two-layered SVMs analysis.
(DOCX)

S5 Table. The top 10 distributions of Gene Ontology (GO) annotations for cross talk of *S*-glutathionylated and *S*-nitrosylated proteins by DAVID analysis ($p < 0.01$).
(DOCX)

S6 Table. The top 10 distributions of GO annotations for only *S*-glutathionylated proteins by DAVID analysis ($p < 0.01$).
(DOCX)

S7 Table. The top 10 distributions of GO annotations for only *S*-nitrosylated proteins by DAVID analysis ($p < 0.01$).
(DOCX)

Author Contributions

Conceived and designed the experiments: TYL Yu-Ju Chen. Performed the experiments: Yi-Ju Chen CTL. Analyzed the data: CTL KYH. Wrote the paper: Yi-Ju Chen HYW TYL.

References

1. Ghezzi P (2013) Protein glutathionylation in health and disease. Biochim Biophys Acta 1830: 3165–3172. doi: 10.1016/j.bbagen.2013.02.009 PMID: 23416063
2. Pastore A, Piemonte F (2012) S-Glutathionylation signaling in cell biology: Progress and prospects. Eur J Pharm Sci 46: 279–292. doi: 10.1016/j.ejps.2012.03.010 PMID: 22484331
3. Dalle-Donne I, Rossi R, Colombo G, Giustarini D, Milzani A (2009) Protein S-glutathionylation: a regulatory device from bacteria to humans. Trends Biochem Sci 34: 85–96. doi: 10.1016/j.tibs.2008.11.002 PMID: 19135374
4. Dalle-Donne I, Rossi R, Giustarini D, Colombo R, Milzani A (2007) S-glutathionylation in protein redox regulation. Free Radic Biol Med 43: 883–898. PMID: 17697933
5. Gallogly MM, Mieyal JJ (2007) Mechanisms of reversible protein glutathionylation in redox signaling and oxidative stress. Curr Opin Pharmacol 7: 381–391. PMID: 17662654

6. Dalle-Donne I, Milzani A, Gagliano N, Colombo R, Giustarini D, Rossi R (2008) Molecular mechanisms and potential clinical significance of S-glutathionylation. Antioxid Redox Signal 10: 445–473. PMID: 18092936
7. Mieyal JJ, Chock PB (2012) Posttranslational modification of cysteine in redox signaling and oxidative stress: Focus on S-glutathionylation. Antioxid Redox Signal 16: 471–475. doi: 10.1089/ars.2011.4454 PMID: 22136616
8. Grek CL, Zhang J, Manevich Y, Townsend DM, Tew KD (2013) Causes and consequences of cysteine S-glutathionylation. J Biol Chem 288: 26497–26504. doi: 10.1074/jbc.R113.461368 PMID: 23861399
9. Lind C, Gerdes R, Hamnell Y, Schuppe-Koistinen I, von Lowenhielm HB, Holmgren A, et al. (2002) Identification of S-glutathionylated cellular proteins during oxidative stress and constitutive metabolism by affinity purification and proteomic analysis. Arch Biochem Biophys 406: 229–240. PMID: 12361711
10. Newman SF, Sultana R, Perluigi M, Coccia R, Cai J, Pierce WM, et al. (2007) An increase in S-glutathionylated proteins in the Alzheimer's disease inferior parietal lobule, a proteomics approach. J Neurosci Res 85: 1506–1514. PMID: 17387692
11. Chiang BY, Chou CC, Hsieh FT, Gao S, Lin JC, Lin SH, et al. (2012) In vivo tagging and characterization of S-glutathionylated proteins by a chemoenzymatic method. Angew Chem Int Ed Engl 51: 5871–5875. doi: 10.1002/anie.201200321 PMID: 22555962
12. Dosztányi Z, Magyar C, Tusnády GE, Cserző M, Fiser A, Simon I (2003) Servers for sequence–structure relationship analysis and prediction. Nucl Acids Res 31: 3359–3363. PMID: 12824327
13. Marino SM, Gladyshev VN (2012) Analysis and Functional Prediction of Reactive Cysteine Residues. J Biol Chem 287: 4419–4425. doi: 10.1074/jbc.R111.275578 PMID: 22157013
14. Sun M-a, Wang Y, Cheng H, Zhang Q, Ge W, Guo D (2012) RedoxDB—a curated database for experimentally verified protein oxidative modification. Bioinformatics 28: 2551–2552. PMID: 22833525
15. Mucchielli-Giorgi MH, Hazout S, Tufféry P (2002) Predicting the disulfide bonding state of cysteines using protein descriptors. Proteins 46: 243–249. PMID: 11835499
16. Sun C, Shi Z-Z, Zhou X, Chen L, Zhao X-M (2013) Prediction of S-Glutathionylation Sites Based on Protein Sequences. PLoS ONE 8: e55512. doi: 10.1371/journal.pone.0055512 PMID: 23418443
17. Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. J Mol Biol 268: 78–94. PMID: 9149143
18. Su D, Gaffrey MJ, Guo J, Hatchell KE, Chu RK, Clauss TRW, et al. (2014) Proteomic identification and quantification of S-glutathionylation in mouse macrophages using resin-assisted enrichment and isobaric labeling. Free Radic Biol Med 67: 460–470. doi: 10.1016/j.freeradbiomed.2013.12.004 PMID: 24333276
19. Shien DM, Lee TY, Chang WC, Hsu JB, Horng JT, Hsu PC, et al. (2009) Incorporating structural characteristics for identification of protein methylation sites. J Comput Chem 30: 1532–1543. doi: 10.1002/jcc.21232 PMID: 19263424
20. Lee TY, Chen SA, Hung HY, Ou YY (2011) Incorporating distant sequence features and radial basis function networks to identify ubiquitin conjugation sites. PLoS One 6: e17331. doi: 10.1371/journal.pone.0017331 PMID: 21408064
21. Tatusova TA, Madden TL (1999) BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. FEMS Microbiol Lett 174: 247–250. PMID: 10339815
22. Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 89: 10915–10919. PMID: 1438297
23. Chang WC, Lee TY, Shien DM, Hsu JB, Horng JT, Hsu PC, et al. (2009) Incorporating support vector machine for identifying protein tyrosine sulfation sites. J Comput Chem.
24. Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: a sequence logo generator. Genome Res 14: 1188–1190. PMID: 15173120
25. Schneider TD, Stephens RM (1990) Sequence logos: a new way to display consensus sequences. Nucleic Acids Res 18: 6097–6100. PMID: 2172928
26. Hsu JB, Bretana NA, Lee TY, Huang HD (2011) Incorporating evolutionary information and functional domains for identifying RNA splicing factors in humans. PLoS One 6: e27567. doi: 10.1371/journal.pone.0027567 PMID: 22110674
27. Xie D, Li A, Wang M, Fan Z, Feng H (2005) LOCSVMPSI: a web server for subcellular localization of eukaryotic proteins using SVM and profile of PSI-BLAST. Nucleic Acids Res 33: W105–110. PMID: 15980436
28. Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292: 195–202. PMID: 10493868

29. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402. PMID: 9254694

30. Pang CN, Hayen A, Wilkins MR (2007) Surface accessibility of protein post-translational modifications. J Proteome Res 6: 1833–1845. PMID: 17428077

31. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. (2000) The Protein Data Bank. Nucleic Acids Res 28: 235–242. PMID: 10592235

32. Ahmad S, Gromiha MM, Sarai A (2003) RVP-net: online prediction of real valued accessible surface area of proteins from single sequences. Bioinformatics 19: 1849–1851. PMID: 14512359

33. Ahmad S, Gromiha MM, Sarai A (2003) Real value prediction of solvent accessibility from amino acid sequence. Proteins 50: 629–635. PMID: 12577269

34. Lee TY, Chen YJ, Lu CT, Ching WC, Teng YC, Huang HD (2012) dbSNO: a database of cysteine S-nitrosylation. Bioinformatics 28: 2293–2295. doi: 10.1093/bioinformatics/bts436 PMID: 22782549

35. Bretana NA, Lu CT, Chiang CY, Su MG, Huang KY, Lee TY, et al. (2012) Identifying protein phosphorylation sites with kinase substrate specificity on human viruses. PLoS One 7: e40694. doi: 10.1371/journal.pone.0040694 PMID: 22844408

36. Lee TY, Lin ZQ, Hsieh SJ, Bretana NA, Lu CT (2011) Exploiting maximal dependence decomposition to identify conserved motifs from a group of aligned signal sequences. Bioinformatics 27: 1780–1787. doi: 10.1093/bioinformatics/btr291 PMID: 21551145

37. Lee TY, Bretana NA, Lu CT (2011) PlantPhos: using maximal dependence decomposition to identify plant phosphorylation sites with substrate site specificity. BMC Bioinformatics 12: 261. doi: 10.1186/1471-2105-12-261 PMID: 21703007

38. Wong YH, Lee TY, Liang HK, Huang CM, Wang TY, Yang YH, et al. (2007) KinasePhos 2.0: a web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns. Nucleic Acids Res 35: W588–594. PMID: 17517770

39. Huang HD, Lee TY, Tzeng SW, Wu LC, Horng JT, Tsou AP, et al. (2005) Incorporating hidden Markov models for identifying protein kinase-specific phosphorylation sites. J Comput Chem 26: 1032–1041. PMID: 15889432

40. Huang HD, Lee TY, Tzeng SW, Horng JT (2005) KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites. Nucleic Acids Res 33: W226–229. PMID: 15980458

41. Chang C-C, Lin C-J (2011) LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2: 1–27.

42. Lu CT, Chen SA, Bretana NA, Cheng TH, Lee TY (2011) Carboxylator: incorporating solvent-accessible surface area for identifying protein carboxylation sites. J Comput Aided Mol Des 25: 987–995. doi: 10.1007/s10822-011-9477-2 PMID: 22038416

43. Vacic V, Iakoucheva LM, Radivojac P (2006) Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. Bioinformatics 22: 1536–1537. PMID: 16632492

44. Wu C, Liu T, Chen W, Oka S-i, Fu C, Jain MR, et al. (2010) Redox Regulatory Mechanism of Transnitrosylation by Thioredoxin. Mol Cell Proteomics: (In press).

45. Chen Y-Y, Chu H-M, Pan K-T, Teng C-H, Wang D-L, Wang AHJ, et al. (2008) Cysteine S-Nitrosylation Protects Protein-tyrosine Phosphatase 1B against Oxidation-induced Permanent Inactivation. J Biol Chem 283: 35265–35272. doi: 10.1074/jbc.M805287200 PMID: 18840608

46. Lee T-Y, Chen Y-J, Lu C-T, Ching W-C, Teng Y-C, Huang H-D, et al. (2012) dbSNO: a database of cysteine S-nitrosylation. Bioinformatics 28: 2293–2295. doi: 10.1093/bioinformatics/bts436 PMID: 22782549

47. Wu C, Parrott AM, Liu T, Jain MR, Yang Y, Sadoshima J, et al. (2011) Distinction of thioredoxin transnitrosylation and denitrosylation target proteins by the ICAT quantitative approach. J Proteomics 74: 2498–2509. doi: 10.1016/j.jprot.2011.06.001 PMID: 21704743

48. Wu C, Liu T, Chen W, Oka S-i, Fu C, Jain MR, et al. (2010) Redox Regulatory Mechanism of Transnitrosylation by Thioredoxin. Mol Cell Proteomics 9: 2262–2275. doi: 10.1074/mcp.M110.000034 PMID: 20660346

49. Benhar M, Forrester MT, Hess DT, Stamler JS (2008) Regulated Protein Denitrosylation by Cytosolic and Mitochondrial Thioredoxins. Science 320: 1050–1054. doi: 10.1126/science.1158265 PMID: 18497292

50. Casagrande S, Bonetto V, Fratelli M, Gianazza E, Eberini I, Massignan T, et al. (2002) Glutathionylation of human thioredoxin: A possible crosstalk between the glutathione and thioredoxin systems. Proc Natl Acad Sci U S A 99: 9745–9749. PMID: 12119401

51. Silva GM, Netto LES, Discola KF, Piassa-Filho GM, Pimenta DC, Bárcena JA, et al. (2008) Role of glutaredoxin 2 and cytosolic thioredoxins in cysteinyl-based redox modification of the 20S proteasome. FEBS J 275: 2942–2955. doi: 10.1111/j.1742-4658.2008.06441.x PMID: 18435761