

Sequence analysis

MDD–SOH: exploiting maximal dependence decomposition to identify S-sulfenylation sites with substrate motifs

Van-Minh Bui^{1,†}, Cheng-Tsung Lu^{1,†}, Thi-Trang Ho¹ and Tzong-Yi Lee^{1,2,*}

¹Department of Computer Science and Engineering and ² Innovation Center for Big Data and Digital Convergence, Yuan Ze University, Taoyuan 320, Taiwan

*To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: John Hancock

Received on June 22, 2015; revised on September 2, 2015; accepted on September 18, 2015

Abstract

S-sulfenylation (S-sulphenylation, or sulfenic acid), the covalent attachment of S-hydroxyl (–SOH) to cysteine thiol, plays a significant role in redox regulation of protein functions. Although sulfenic acid is transient and labile, most of its physiological activities occur under control of S-hydroxylation. Therefore, discriminating the substrate site of S-sulfenylated proteins is an essential task in computational biology for the furtherance of protein structures and functions. Research into S-sulfenylated protein is currently very limited, and no dedicated tools are available for the computational identification of SOH sites. Given a total of 1096 experimentally verified S-sulfenylated proteins from humans, this study carries out a bioinformatics investigation on SOH sites based on amino acid composition and solvent-accessible surface area. A TwoSampleLogo indicates that the positively and negatively charged amino acids flanking the SOH sites may impact the formulation of S-sulfenylation in closed three-dimensional environments. In addition, the substrate motifs of SOH sites are studied using the maximal dependence decomposition (MDD). Based on the concept of binary classification between SOH and non-SOH sites, Support vector machine (SVM) is applied to learn the predictive model from MDD-identified substrate motifs. According to the evaluation results of 5-fold cross-validation, the integrated SVM model learned from substrate motifs yields an average accuracy of 0.87, significantly improving the prediction of SOH sites. Furthermore, the integrated SVM model also effectively improves the predictive performance in an independent testing set. Finally, the integrated SVM model is applied to implement an effective web resource, named MDD-SOH, to identify SOH sites with their corresponding substrate motifs.

Availability and implementation: The MDD-SOH is now freely available to all interested users at <http://csb.cse.yzu.edu.tw/MDDSOH/>. All of the data set used in this work is also available for download in the website.

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

Contact: francis@saturn.yzu.edu.tw

1 Introduction

Protein S-sulfenylation (also known as S-sulphenylation), is a reversible post-translation modification (PTM) involving covalent

attachment of hydroxide to the thiol group of cysteine residues. Oxidation of the thiol functional group in cysteine (Cys–SH) to sulfenic (Cys–SOH), sulfinic (Cys–SO₂H) and sulfonic acids

(Cys-SO₃H) has emerged over the last decade as an important method of controlling protein functions through both physiologic and oxidative stress conditions (Leonard and Carroll, 2011; Poole and Nelson, 2008; Roos and Messens, 2011; Wani et al., 2011). More than 200 sulfenic modifications have been identified from transcription factors, signaling proteins, metabolic enzymes, proteostasis regulators and cytoskeletal components. Due to the labile nature and low abundance of *in vivo* S-sulfenylation, the detail characteristics and mechanisms of S-sulfenylation still need to be clarified. To our knowledge, the protein database of human, mouse or rat only consists of ~2% cysteine residues. Both the degree of S-sulfenylation within proteins and quantification of protein sulfenic modification are necessary for understanding oxidative PTM regulation.

Several chemo-proteomic strategies exist for identifying S-sulfenylated sites with various clickable probes (Qian et al., 2013; Szychowski et al., 2010; Wang et al., 2014; Weerapana et al., 2010; Zheng et al., 2013), but these are very expensive. In the laboratory, the lack of the information about sites and the multiplicity of redox changes both lead to false-positive identifications (Yang et al., 2014). As for the increasing number of experimentally identified S-sulfenylated peptides, building a resource database is essential for further biological investigation of S-sulfenylated proteins and the substrate specificities of S-sulfenylation sites. With the development of powerful bioinformatics technologies, these approaches are adopted for prognostication of individual cysteine residues to S-sulfenylation. Although several published algorithms and public servers are available to analyze and predict the reactive state of cysteine (Dosztányi et al., 2003; Marino and Gladyshev, 2012) and oxidative yet disulfide cysteine (Mucchielli-Giorgi et al., 2002; Sun et al., 2012), no unconfirmed bio-reports exist on the specific information of S-sulfenylated targets and substrate sites, except for chemical approaches with over 1000 S-sulfenylation sites on >700 proteins in intact cells (Yang et al., 2014). Other potential novel consensus S-sulfenylation motifs and substrate site specificities remain unclear.

This work concentrates on identifying S-sulfenylation sites with potential substrate motifs. Depending on the *in silico* characterization of protein, the sequential and structural features including amino acid composition (AAC), amino acid pair composition (AAPC), position-specific scoring matrix (PSSM), position weight matrix (PWM), amino acid substitution matrix (BLOSUM62) and accessible surface area (ASA), are applied to discriminate between the SOH sites and non-SOH sites. Additionally, this study proposes an iteratively statistical method for recognizing S-sulfenylation sites and promising consensus motifs by maximal dependence decomposition (MDD) (Burge and Karlin, 1997). Based on the most significant dependencies between positions, MDD allows a large group of aligned sequence to be separated into subgroups containing characteristic motifs. For each subgroup, support vector machine (SVM) was adopted to build the predictive model for each MDD cluster. Consequently, the ability to predict results between S-sulfenylation sites and non-S-sulfenylation sites is significantly increased through evaluation of 5-fold cross-validation. Furthermore, an independent testing set, which is blind to the training dataset, was generated from the experimental data of published database to demonstrate the effectiveness of the proposed models in the evaluation of 5-fold cross-validation. To facilitate the study of protein S-sulfenylation, the MDD-identified substrate motifs were employed to implement a web-based resource for identifying S-sulfenylation sites with their corresponding substrate motifs.

2 Materials and methods

2.1 Data collection and pre-processing

With the newly discovered S-sulfenyl-mediated redox mechanisms of transcription factor H1F1A by SIRT6 (Yang et al., 2014), experimentally verified S-sulfenylated cysteines from humans were used as the positive set, and all non-S-sulfenylated cysteines on these S-sulfenylated proteins were used as the negative data. As shown in Supplementary Table S1, the main dataset was gathered from Carroll Lab (PMID: 25175731) with 1443 positive and 10 521 negative data on 987 S-sulfenylated proteins. In addition, several experimentally verified S-sulfenylation sites were manually curated from RedoxDB (Sun et al., 2012), UniProtKB and other literatures. For the RedoxDB database, a total of 102 S-sulfenylated cysteine on 92 proteins, from human and others were used as the positive dataset, while the rest of the cysteine residues on S-sulfenylated proteins were regarded as the negative data (non-S-sulfenylated cysteine). From the UniProtKB database, one S-sulfenylated protein was from human, and the remaining one from others, where the 17 S-sulfenylated sites and 97 non-S-sulfenylated sites were considered as positive and negative data, respectively. The protein sulfenic acid was detected by X-ray Crystallography (Furdui and Poole, 2014). The positive dataset comprised 33 S-sulfenylated cysteines, while the negative data set contained 143 non-S-sulfenylated cysteines. After the removal of redundant data, a total of 1434 unique S-sulfenylation sites were obtained from 1096 proteins.

This study focuses on the sequence-based characterization of substrate site motifs of S-sulfenylated cysteines. Thus, a window length of $2n + 1$ was utilized to extract sequence fragments that centered at the experimentally verified SOH sites and contained n upstream and n downstream flanking amino acids. Given 1096 S-sulfenylated proteins, the sequence fragments containing window length of $2n + 1$ amino acids and centering at cysteine residue with the annotation of S-sulfenylation were regarded as the positive data set (S-sulfenylated sites). Based on a window size of 21 ($n = 10$), the data set contained 1434 positive data and 10 476 negative data. To avoid overestimating the predictive performance, the CD-HIT program (Li and Godzik, 2006) was employed to remove homologous sequence fragments from the positive and negative datasets. CD-HIT is an effective tool for clustering protein sequences based on a specified sequence similarity value. One instance sequence was chosen to represent each cluster. Owing to the incomplete information of experimentally validated S-sulfenylation sites, based on the analysis of sequence fragments, some negative data could be identical to positive data in this work, potentially causing false positive or false negative predictions. Therefore, CD-HIT was applied again, by running CD-HIT-2D across positive and negative training data with 100% sequence identity. Table 1 shows the data statistics after eliminating the homologous fragments using CD-HIT based on various values of sequence identity (ranging from 100 to 40%). After having filtered out homologous fragments with 40% sequence identity (by running CD-HIT and psi-CD-HIT), the final dataset comprised 1247 positive sequences and 9446 negative sequences.

Based on binary classification, the positive and negative datasets were used to generate the predictive model. The 5-fold cross-validation was then applied to evaluate the power of distinguishing between the S-sulfenylated and non-S-sulfenylated cysteines. However, the predictive performance might be overestimated due to an overfitting of training dataset. To assess the real case for the predictive performance of the proposed models, an independent testing set, definitely blind to the training set, was generated. The dataset for independent testing was generated by randomly selecting from the

Table 1. Data statistics after using CD-HIT

Threshold	Number of proteins	S-sulfonylation sites (positive data)	Non-S-sulfonylation sites (negative data)
100%	1096	1434	10476
90%	1026	1406	10322
80%	997	1375	10177
70%	970	1340	9935
60%	949	1307	9739
50%	926	1279	9648
40%	901	1247	9446
Training data		1031	8028
Independent testing data		216	1418

final dataset (1247 positive sequences and 9446 negative sequences), and the remaining data were used as the training dataset. The training dataset contained 1031 positive and 8028 negative sequence fragments. The final independent testing dataset contained 216 positive and 1418 negative data (Table 1).

2.2 Feature investigation and encoding

To build the predictive models, SVM was adopted to discriminate between S-sulfonylation and non-S-sulfonylation sites based on sequence-based features such as 20D binary coding (AA), AAC, BLOSUM62, ASA, AAPC, PSSM and PWM. The fragment sequences, which contained the S-sulfonylated cysteine in the center, were extracted with window size equal to 21 from positive and negative data (Lee *et al.*, 2011a,b,c). Orthogonal binary coding is one of the most popular coding methods of converting amino acids into numeric vectors, called 20D binary code. For instance, Alanine (A) was encoded as '10000000000000000000', and Cysteine (C) was encoded as '01000000000000000000'. The number of feature vectors was $(2n+1) \times 20$ to represent the flanking amino acids surrounding the S-sulfonylation sites. The training data contained k vectors $\{x_i, i=1, 2, \dots, k\}$ corresponding to the k fragment sequences. To classify the positive and negative cysteine, a label was applied for each vector to mark the class of its corresponding protein. For composition of amino acid around the S-sulfonylation sites, the vector x_i had 21 elements for AAC and 441 elements for AAPC. Some rare amino acids and non-existing '-' residues were used to represent <21-mer fragment sequences at an N- or C-terminus. The BLOSUM62 was built on the alignments of amino acid sequences with no >62% identity between two peptide sequences with 21 amino acids. Using the SulfoSite method (Chang *et al.*, 2009), PWM was determined with non-homologous training data. The PWM described the frequent occurrence of amino acids surrounding the S-sulfonylation sites, and was utilized in encoding the fragment sequences. Each residue of a training dataset was represented by a matrix of $m \times w$ elements, where w is the window size equal to 21, and m represents 21 elements including 20 types of amino acids and one for the terminal signal.

PSSM profiles were generated from PSI-BLAST (Altschul *et al.*, 1997) against non-redundant sequences of S-sulfonylation sites. This matrix of score values can represent the multiple sequence alignment of proteins which may have similar structures with different AACs. Extracting from the PSSM profile, the matrix of $(2n+1) \times 20$ elements had rows centered on the substrate site, where $2n+1$ represents the window size, and 20 is the number of position-specific scores for each type of amino acid.

The structural feature of ASA was investigated based on the accessibility of a side-chain of amino acid on the surface of a protein

that experienced post-translational modification (Pang *et al.*, 2007). RVP-Net (Ahmad *et al.*, 2003a,b), an effective tool, was adopted to calculate the ASA value from the protein sequence due to the lack of most experimental S-sulfonylated protein tertiary structures in Protein Data Bank (PDB) (Berman *et al.*, 2000). RVP-Net can predict the real ASA of a residue based on information about the neighborhood by using a neural network. The possible mean absolute error, given by the absolute difference between the predicted and experimental values of relative ASA per residue, was 18.0–19.5%, for each measurement (Ahmad *et al.*, 2003a,b). The value of ASA was the percentage of the solvent-accessible area of each amino acid on the protein. The input data of the RVP-Net were the full-length protein sequences to compute the ASA value of all of the residues. The ASA values of amino acids around the S-sulfonylation sites were then extracted and normalized to the range 0–1.

Each hybrid feature was formed by combining two or more single features. Based on the performance of each feature, the single feature with the best performance was incorporated with other single features to enhance predictive power. Supplementary Figure S1 describes the conceptual flowchart for combining the PSSM and BLOSUM62 features for each sequence fragment. Before the construction of SVM classifier, all the numeric data need to be scaled into values ranging from -1 to +1 to improve prediction effectiveness.

2.3 Model construction and evaluation

The positive and negative training data sets for the predictive model were built using SVM. Based on binary classification, a kernel function transforms the input samples into a higher dimensional space, and then finds a hyper-plane to discriminate between the two classes with maximal margin and minimal error. This study employed a public SVM library (LIBSVM) (Chang and Lin, 2011) to implement the predictive model for distinguishing the S-sulfonylation sites from non-S-sulfonylation sites. The radial basis function (RBF):

$$K(S_i, S_j) = \exp(-\gamma \|S_i - S_j\|^2) \quad (1)$$

was adopted as the kernel function for learning in the SVM classifier. Two supporting factors to enhance the performance are gamma and cost. The RBF kernel is determined by the gamma parameter, while the cost parameter controls the hyper-plane softness. Each feature was used to generate a predictive model by LIBSVM library; then, the best feature was selected as an input vector for the second-layered SVM based on the values of previous estimated probability from each SVM classifier.

To choose the best final model, the 5-fold cross-validation was carried out for each different feature to evaluate the predictive performances. The training data were divided into five approximately equal-sized subgroups. The ratio of test and training sets was 1:4, and the cross-validation process was run five times. The five validation results were then combined to generate a single estimation. Cross-validation evaluation improves the reliability of evaluation, because it considers all original data, in both the training and testing data sets, in general, and tests each subset only once (Lu *et al.*, 2011). To gauge the effectively predictive performance of training model, the following measures were used: sensitivity (Sn), specificity (Sp), accuracy (Acc) and Matthews Correlation Coefficient (MCC):

$$Sn = \frac{TP}{TP + FN} \quad (2)$$

$$Sp = \frac{TN}{TN + FP} \quad (3)$$

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \quad (4)$$

$$MCC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}} \quad (5)$$

where TP, TN, FP and FN represent the numbers of true positives, true negatives, false positives and false negatives, respectively. Sensitivity is the percentage of correct predictions from positive data (*S*-sulfenylated cysteines), while specificity represents that from negative data (non-*S*-sulfenylated cysteines). Accuracy reflects the overall proportion of correctly predicted positive data (*S*-sulfenylated cysteines) and negative data (non-*S*-sulfenylated cysteines). For binary classifications, accuracy is sometimes not useful when the two classes are of very different sizes. Therefore, the MCC is typically considered as a balanced measure, even if the two classes are of very different sizes (Matthews, 1975). The MCC value is ranging from -1 to $+1$, while the values of other three measures range from 0 to 1 . A coefficient value of $+1$ represents a perfect prediction, while the values 0 and -1 represent random and opposite predictions, respectively. A higher positive MCC value indicates a better prediction for correctly classifying positive and negative data. Finally, after selecting the best predictive model with the highest MCC in this work, an independent test was carried out on the final model with best performance in cross-validation evaluation.

2.3 Data grouping by MDD

This work investigated the substrate motifs of *S*-sulfenylation sites based on the amino acid sequences. MDD (Burge and Karlin, 1997) was utilized to cluster all fragment sequences into subgroups to detect the statistically conserved motifs from large-scale sequence data. The clustering was performed using a public MDD clustering resource, MDDLogo (Lee et al., 2011a,b,c), which splits a larger group into smaller subgroups with significantly conserved motifs before computationally identifying PTM sites (Bretana et al., 2012; Chen et al., 2014; Huang et al., 2005a,b; Lee et al., 2011a,b,c, 2012; Wong et al., 2007). The dependence of amino acid occurrence between two positions, A_i and A_j , that surround the *S*-sulfenylated cysteine, was evaluated using chi-square test $\chi^2(A_i, A_j)$. Based on the biochemical property of amino acids, the 20 amino acids were categorized into five groups, namely polar, acidic, basic, hydrophobic and aromatic groups (Supplementary Table S2). A contingency table in Supplementary Figure S2A describes the frequency of existence of amino acids between two positions A_i and A_j . The chi-square test was defined as:

$$\chi^2(A_i, A_j) = \sum_{m=1}^5 \sum_{n=1}^5 \frac{(X_{mn} - E_{mn})^2}{E_{mn}} \quad (6)$$

where X_{mn} represented the number of sequences at both the position A_i of the group m and position A_j of group n , for each pair (A_i, A_j) with $i \neq j$. E_{mn} was calculated as $\frac{X_{mR} \times X_{Cn}}{X}$, where $X_{mR} = X_{m1} + \dots + X_{m5}$, $X_{Cn} = X_{1n} + \dots + X_{5n}$, and X denoted the total number of sequences. If a strong dependence (defined as $\chi^2 > 34.3$, corresponding to a cutoff level of P -value = 0.01 with 16 degrees of freedom) was detected between two positions, then the decomposition was proceeded (Burge and Karlin, 1997). As shown in Supplementary Figure S2B, the group of basic amino acids kept the maximal dependent value at position -2 . Consequently, all data can be separated into two subgroups: one with basic amino acids in

position -2 , and the other with no basic amino acids in position -2 . The maximum cluster size is an important parameter when applying MDDLogo to cluster the sequences of the positive set. The recursive process of MDD clustering was performed to divide the positive sets into tree-like subgroups repeatedly until all subgroups are smaller than the maximum cluster size. To obtain an optimal minimum cluster size, MDDLogo was run using various values for that parameter. For this investigation, each subgroup generated by MDDLogo was represented using WebLogo (Crooks et al., 2004) to determine whether they presented conserved motifs for the substrate specificity of *S*-sulfenylation.

These MDDLogo-identified substrate motifs could be adopted to create a two-layered SVM prediction model for the identification of SOH sites with their corresponding substrate motifs. As illustrated in Figure 1, the LIBSVM library was employed at the first layer of SVM to generate a predictive model based on the investigation results of the best features. The LIBSVM library outputs a probability value estimate ranging from 0 to 1 for each prediction. Therefore, the probability values estimated from each SVM classifier trained according to a specific motif were adopted to form an input vector for the second layer of SVM.

3 Results and discussion

3.1 Impact of flanking AAC to *S*-sulfenylation sites

This investigation analyzed the frequency of occurrence of 20 amino acids surrounding the *S*-sulfenylation site on fragment sequences of proteins to find the potential consensus motifs. TwoSampleLogo (Vacic et al., 2006) is an effective web-based tool to detect statistically noteworthy differences in position-specific symbol compositions between the *S*-sulfenylated and non-*S*-sulfenylated data sets. The Cysteine amino acid was placed in the middle of the fragment sequences, and positions of the flanking amino acids were described in range from -10 to $+10$. The comparison between 1031 *S*-sulfenylation sites and 8028 non-*S*-sulfenylation sites in Figure 2 demonstrates that the positively charged amino acids, such as Lysine (K) and Arginine (R), had the highest ratios at positions -10 , -7 , -6 , -4 , -2 , and $[+4, +8]$ (with $P < 0.01$). Additionally, no positively charged residues were found at positions -9 , -3 , -1 , $+1$, $+2$, $+3$ close to the *S*-sulfenylation sites. Furthermore, negatively charged residues (Glutamic acid – E) were found at positions -4 , -3 , $+1$, $+3$, $+4$. Position -1 was a special case with the highest proportion of the polar group of residues, namely Serine (S) and Asparagine (N). In contrast, the non-*S*-sulfenylation database contained many instances of neutral amino acids Leucine (L), Cysteine (C), Histidine (H), Methionine (M), Phenylalanine (F) and Tyrosine (Y) at positions in range $[-9, +7]$, while only Arginine (R) residue appeared at three positions (-1 , $+1$ and $+2$) near the non-*S*-sulfenylation sites. This investigation found a significant difference in the amino acid sequences located around position -6 , -2 , -1 , $+3$ and $+4$ in the two sets. The analysis shows that the distance among amino acid characteristic in a sequence plays a vital role in distinguishing between *S*-sulfenylation sites and non-*S*-sulfenylation sites. Finally, positively charged amino acids may be contiguous to *S*-sulfenylated cysteine in a 3D structure.

The significant amino acids around *S*-sulfenylated 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*-sulfenylated cysteines is depleted from the negative dataset and presented in lower panel.

To estimate carefully the composition of amino acids in sequences, the correlation between *S*-sulfenylation sites and the

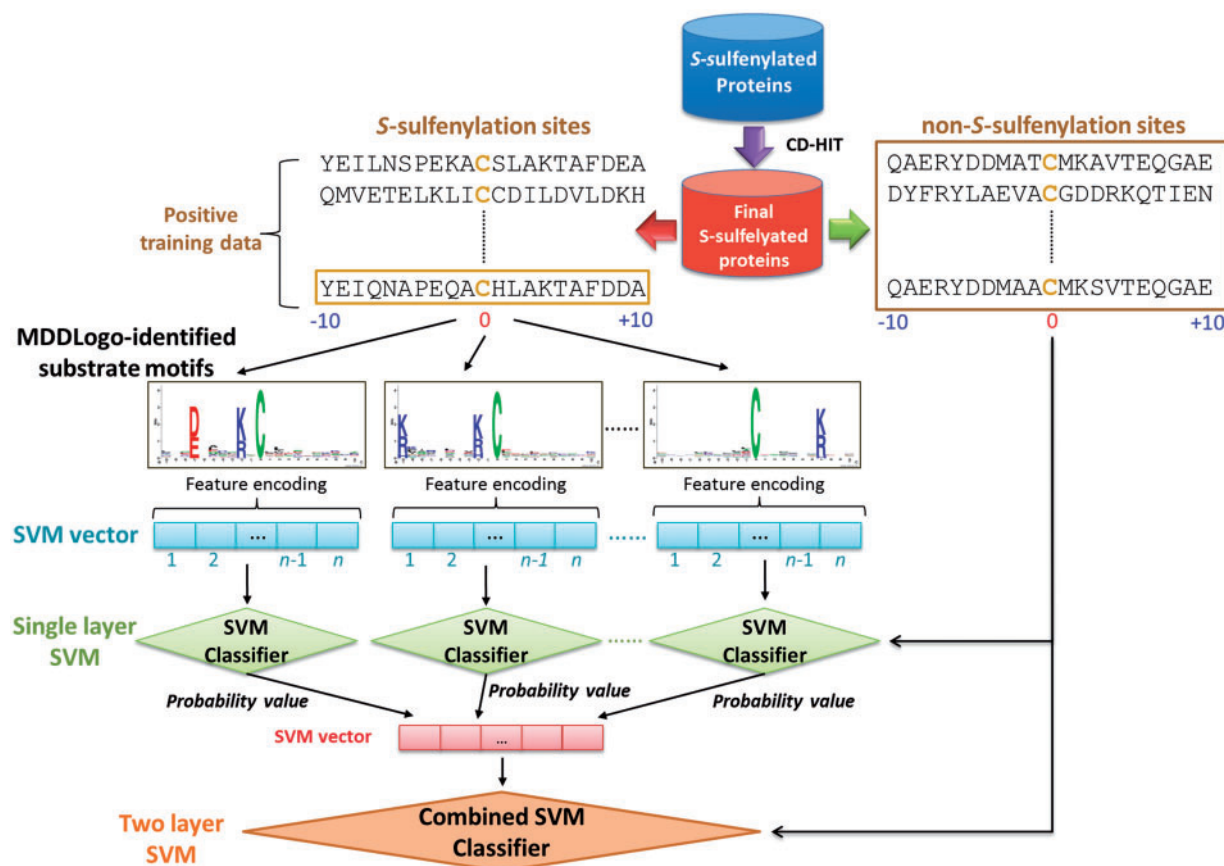


Fig. 1. The conceptual diagram of two-layered SVMs trained with MDDLogo-identified substrate motifs

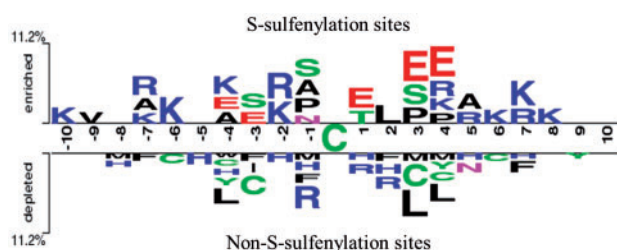


Fig. 2. TwoSampleLogo presents the compositional biases of amino acids around S-sulfenylation sites compared with the non-S-sulfenylation sites

solvent ASA was evaluated using the RVP-Net algorithm. As shown in Supplementary Figure S3, the average percentage of ASA of cysteine residue was the lowest on both the S-sulfenylated and non-S-sulfenylated cysteines based on the comparison of average proportion of ASA in the 21-mer window ($[-10, +10]$). This finding indicates that solvent accessibility at S-sulfenylated sites is usually lower than that of flanking regions. Additionally, the neighboring amino acids surrounding the S-sulfenylation sites reacted to the solvent-ASA higher than non-S-sulfenylation sites. Following investigations of AAC (Fig. 2) and solvent accessibility (Supplementary Fig. S3), the important influence of hydrophilic amino acid group may regulate the S-sulfenylated cysteine residues.

3.2 Performance of 5-fold cross-validation on training data

To identify the best feature for discriminating between S-sulfenylation sites and non-S-sulfenylation sites, the SVM models

were trained using various features, namely AA, AAC, AAPC, BLOSUM62, ASA, PSSM and PWM. As shown in Table 2, the SVM model trained with ASA had the lowest MCC value at 0.14, and relatively low sensitivity, specificity and accuracy at 0.61, 0.61 and 0.61, respectively. In contrast, PSSM was found to be the best feature for generating an SVM model, with sensitivity at 0.68, specificity at 0.68, accuracy at 0.68, and MCC at 0.24. The features BLOSUM62, AA and AAPC were found to be combined well with PSSM to improve predictive power. For others cases in this table, the results of AAC and PWM provided quite similar performance. The predictive model based on the hybrid combination of BLOSUM62 with PSSM provided the best predictive performance in Sn, Sp, Acc and MCC at 0.68, 0.70, 0.70 and 0.27, respectively. This hybrid feature had a slightly higher sensitivity than PSSM. Therefore, the hybrid feature of PSSM with BLOSUM62 was selected as the training feature for the construction of two-layered SVM model.


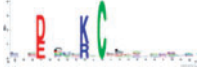
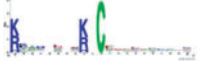
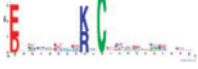

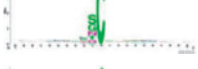

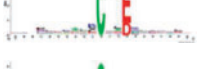



3.3 MDDLogo-identified substrate motifs

In this investigation, the MDD was adopted to detect the conserved motifs by clustering S-sulfenylation data set (1031 sequences) into 10 subgroups. Each subgroup represents an S-sulfenylated motif that contains statistically significant dependencies of AAC between specific positions (Table 3). Overall, eight motifs out of all MDDLogo-clustered subgroups were detected based on the positively charged amino acids (K, R and H) and negatively charged amino acids (D, E). Subgroups SOH1, SOH2, SOH3 and SOH4 represented the occurrence of positively charged amino acids

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

Training features	C (cost)	γ (gamma)	Sn	Sp	Acc	MCC
20D Binary coding (AA)	32	0.03125	0.65	0.65	0.65	0.20
Amino acid substitution matrix (BLOSUM62)	32	0.000030517578125	0.64	0.65	0.65	0.19
AAC	8192	0.0078125	0.62	0.63	0.63	0.17
AAPC	32768	0.5	0.65	0.66	0.66	0.21
ASA	2	0.000030517578125	0.61	0.61	0.61	0.14
PWM	32	2	0.60	0.62	0.62	0.16
PSSM	2	0.000488281	0.68	0.68	0.68	0.24
PSSM + AA	2	0.0078125	0.68	0.68	0.68	0.24
PSSM + AAPC	2	0.0078125	0.68	0.68	0.68	0.24
PSSM + BLOSUM62	32	0.00012207	0.68	0.70	0.70	0.27

Table 3. The 10 MDDLogo-identified substrate motifs and their performances of 5-fold cross-validation

MDDLogo-clustered subgroup	Sequence logo of substrate motif	Number of positive data	C (cost)	G (gamma)	Sn	Sp	Acc	MCC
All data		1,031	32	0.00012207	0.68	0.69	0.69	0.25
SOH1		29	32	0.0000305175781250	0.93	0.94	0.94	0.76
SOH2		30	8	0.0001220703125	0.87	0.88	0.88	0.6
SOH3		26	8	0.000030517578125	0.92	0.92	0.92	0.72
SOH4		115	32	0.000030517578125	0.9	0.92	0.92	0.7
SOH5		276	2	0.00048828125	0.81	0.82	0.82	0.46
SOH6		85	32	0.000030517578125	0.87	0.88	0.88	0.61
SOH7		59	8	0.000030517578125	0.83	0.87	0.86	0.54
SOH8		83	8	0.0001220703125	0.86	0.87	0.87	0.58
SOH9		49	32	0.000030517578125	0.8	0.86	0.85	0.5
SOH10		279	8	0.000030517578125	0.72	0.73	0.73	0.3
Combined all MDDLogo-identified motifs		1031	32	0.00012207	0.85	0.87	0.87	0.58

(lysine and arginine) in position -2 with maximal dependence, whereas the other subgroups had no occurrence of positively charged amino acids in position -2 . In particular, one special motif was discovered with a significant percentage of amino acid belonging to the polar group (over 27%). Finally, the tenth subgroup

(SOH10), containing the remaining 279 *S*-sulfenylation sites, had a little conservation of amino acids at position -1 .
To improve the reliability of our study, the 5-fold cross-validation evaluation was adopted on all of the *S*-sulfenylation sites and on the 10 MDDLogo-clustered subgroups. Table 3 displays the

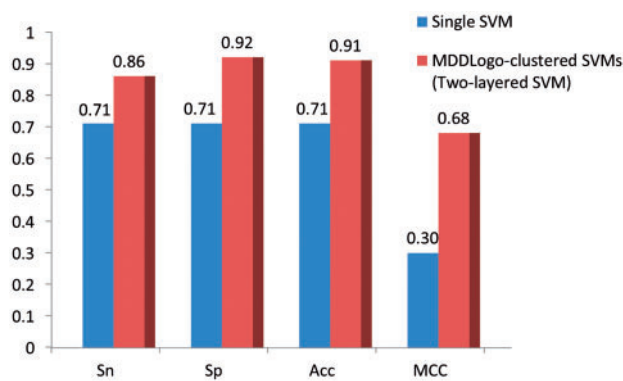


Fig. 3. Comparison of independent testing performance between single SVM and MDDLogo-clustered SVM models. Sn, sensitivity; Sp, specificity; Acc, accuracy; MCC, Matthews correlation coefficient

predictive results of 5-fold cross-validation in each group. As noted earlier, position -2 of the first four groups hid behind the significant influence leading to their performance >0.87 . Among these four groups, SOH1 and SOH3 were found to have the best sensitivity, specificity, accuracy and MCC, with respective values of 0.93, 0.94, 0.94 and 0.76 for SOH1, and 0.92, 0.92, 0.92 and 0.72 for SOH3. The SVM-based predictive performance of these 10 MDDLogo-clustered subgroups was higher than that of all 1031 unclustered S-sulfenylation sites. The SVM model constructed at the second SVM layer, combining MDDLogo-identified motifs, had sensitivity, specificity, accuracy and MCC values of 0.85, 0.87, 0.87 and 0.58, respectively, which were all much higher than those of the unclustered 1,031 S-sulfenylation sites at 0.68, 0.69, 0.69 and 0.25, respectively. In summary, the combined MDDLogo-clustered models might be expected to enhance the performance, and could be implemented a web-based prediction tool in a website.

3.4 Evaluation of S-sulfenylation predictive models using independent testing set

An independent testing set of S-sulfenylated proteins was used to evaluate the effectiveness of the MDDLogo-clustered SVM models that achieved the best accuracy in 5-fold cross-validation. The independent testing set comprised 216 positive data and 1418 negative data. Figure 3 presents the test results, indicating significant differences in the ratio of sensitivity, specificity, accuracy and MCC between the two models at 0.71, 0.71, 0.71, 0.30 respectively for Single SVM, and 0.86, 0.92, 0.91, 0.68 respectively for Two-layered SVM. Supplementary Table S3 lists the detailed independent testing results, in which the MDDLogo-clustered SVM models had better predictive performance than the Single SVM model.

3.5 Implementation of web-based prediction tool

Researchers have applied various methods to work with S-sulfenylated proteins in the laboratory (Furdui and Poole, 2014; Yang *et al.*, 2014), but often encounter difficulties such as expensive equipment, time-consuming processing and complex laboratory-intensive experimental workflows. Therefore, this work developed an effective prediction tool to overcome the challenges in identifying the S-sulfenylation sites. Based on the evaluated results of cross-validation and independent test, the MDDLogo-clustered SVM model trained using PSSM and BLOSUM62 was implemented to develop the web-based prediction system, named MDD-SOH. The system enables users to submit the protein sequences of interest and return predictive results including S-sulfenylation position, the

flanking amino acids and the matched MDDLogo-clustered motif. The input data for MDD-SOH can be the protein sequence in FASTA format or the protein name, gene name and accession number.

The effectiveness of MDD-SOH was tested using following three cases: 'Fatty acid-binding protein, epidermal' (FABP5, FABP5_HUMAN), Peroxiredoxin (PRDX6, PRDX6_HUMAN) and peroxiredoxin HYR1 (HYR1, GPX3_YEAST). The first case is Fatty acid-binding protein, epidermal (FABP5, FABP5_HUMAN), which contains one S-sulfenylation site at Cys-127 (Yang *et al.*, 2014). As presented in Supplementary Figure S4, MDD-SOH was predicted exactly at position 127 of the S-sulfenylation protein. Moreover, the system also displayed the matched MDDLogo-clustered motif. The second test was carried out on Peroxiredoxin protein on human (PRDX6, PRDX6_HUMAN), which contains one S-sulfenylation site at Cys-91 (Yang *et al.*, 2014), and which was not in the training data set. Consequently, the position 91 of the experimentally-verified S-sulfenylation site was also correctly predicted by MDD-SOH. Finally, peroxiredoxin HYR1 (HYR1, GPX3_YEAST) has Cysteine in position 36 (Seo and Carroll, 2011), which was also correctly predicted.

3.6 Functional investigation of S-sulfenylated proteins

To further understand the characteristics and function of proteins, Database for Annotation, Visualization and Integrated Discovery (DAVID) software provides a comprehensive set of functional annotation tools for investigators to exploit biological meaning behind large list of genes (<http://david.abcc.ncifcrf.gov/home.jsp>). According to the GO annotations, the distributions of the biological processes, molecular functions and cellular components of S-sulfenylated proteins were presented in Supplementary Table S4. Following the analytic results, 113 of 1096 S-sulfenylation proteins involve in RNA processing with 10.8% corresponding *P*-value ($7.65E-32$). Additionally, there are many proteins took part in RNA binding, nucleotide binding, ribonucleotide binding, purine ribonucleotide binding and purine nucleotide binding with 148 (14.15%), 273 (26.10%), 217 (20.75%) and 223 (21.32%), respectively. Furthermore, S-sulfenylation proteins also played an important role in construction of cellular component such as nuclear lumen (21.03%), non-membrane-bounded organelle (29.73%), intracellular non-membrane-bounded organelle (29.73%), etc.

Supplementary Table S5 describes the distribution of KEGG pathways for S-sulfenylated proteins. Comparing with the GO enrichment analysis, few of these proteins were involved in the metabolic pathways. The statistics show that only 27 of 1096 proteins (2.58%) formed a pathway for the pathogenic *Escherichia coli* infection. Finally, the investigation of protein domains shows that the most abundant domain in S-sulfenylated proteins is the Thioredoxin-like fold with 1.74% (7.47×10^{-6}), as shown in Supplementary Table S6.

4 Conclusion

This study identified the S-sulfenylation sites of proteins, and investigated the statistically significant conserved motifs. Through TwoSampleLogo, the analysis of position-specific amino acids composition between S-sulfenylation and non-S-sulfenylation site identified the relationship between flanking amino acids with the S-sulfenylation cysteine site. Moreover, this investigation also found that the solvent-accessible surface of amino acid surrounding S-sulfenylation sites have a tendency to higher than that around

non-S-sulfonylation sites. Based on the results of 5-fold cross-validation, the hybrid feature of BLOSUM62 and PSSM was estimated as the best feature with the highest proportion of sensitivity, specificity, accuracy and MCC. As stated previously, the main purpose is to find the conserved motifs of S-sulfonylation sites based on the amino acid sequences. By using MDDLogo, all S-sulfonylation proteins were clustered into 10 subgroups corresponding with 10 conserved motifs. The MDD-identified motif can thus be adopted to train the model to enhance significantly the predictive performance of the S-sulfonylation site. The sets were evaluated through independent testing using two models, namely MDDLogo-clustered SVM and Single SVM without MDD. As expected, MDDLogo-clustered SVM model, which contained the significant conserved motifs, achieved the better performance. Consequently, this model was employed to set up a novel web-based resource, named MDD-SOH, to identify S-sulfonylation sites and their corresponding substrate motifs.

Funding

This work was supported by the Ministry of Science and Technology of Taiwan under Contract Number MOST 103-2221-E-155-020-MY3 and 104-2221-E-155-036-MY2 to T.Y.-L.

Conflict of Interest: none declared.

References

- Ahmad, S. *et al.* (2003a) Real value prediction of solvent accessibility from amino acid sequence. *Proteins*, **50**, 629–635.
- Ahmad, S. *et al.* (2003b) RVP-net: online prediction of real valued accessible surface area of proteins from single sequences. *Bioinformatics*, **19**, 1849–1851.
- Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
- Berman, H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
- Bretana, N.A. *et al.* (2012) Identifying protein phosphorylation sites with kinase substrate specificity on human viruses. *PLoS One*, **7**, e40694.
- Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78–94.
- Chang, C.-C. and Lin, C.-J. (2011) LIBSVM: a library for support vector machines. *ACM Trans. Intel. Syst. Technol.*, **2**, 1–27.
- Chang, W.C. *et al.* (2009) Incorporating support vector machine for identifying protein tyrosine sulfation sites. *J. Comput. Chem.*, **30**, 2526–37.
- Chen, Y.J. *et al.* (2014) dbGSH: a database of S-glutathionylation. *Bioinformatics*, **30**, 2386–2388.
- Crooks, G.E. *et al.* (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
- Dosztányi, Z. *et al.* (2003) Servers for sequence–structure relationship analysis and prediction. *Nucleic Acids Res.*, **31**, 3359–3363.
- Furdui, C.M. and Poole, L.B. (2014) Chemical approaches to detect and analyze protein sulfenic acids. *Mass Spectrom. Rev.*, **33**, 126–146.
- Huang, H.D. *et al.* (2005a) KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites. *Nucleic Acids Res.*, **33**, W226–W229.
- Huang, H.D. *et al.* (2005b) Incorporating hidden Markov models for identifying protein kinase-specific phosphorylation sites. *J. Comput. Chem.*, **26**, 1032–1041.
- Lee, T.Y. *et al.* (2011a) PlantPhos: using maximal dependence decomposition to identify plant phosphorylation sites with substrate site specificity. *BMC Bioinformatics*, **12**, 261.
- Lee, T.Y. *et al.* (2012) dbSNO: a database of cysteine S-nitrosylation. *Bioinformatics*, **28**, 2293–2295.
- Lee, T.Y. *et al.* (2011b) SNOsite: exploiting maximal dependence decomposition to identify cysteine S-nitrosylation with substrate site specificity. *PLoS One*, **6**, e21849.
- Lee, T.Y. *et al.* (2011c) Exploiting maximal dependence decomposition to identify conserved motifs from a group of aligned signal sequences. *Bioinformatics*, **27**, 1780–1787.
- Leonard, S.E. and Carroll, K.S. (2011) Chemical ‘omics’ approaches for understanding protein cysteine oxidation in biology. *Curr. Opin. Chem. Biol.*, **15**, 88–102.
- Li, W. and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.
- Lu, C.T. *et al.* (2011) Carboxylator: incorporating solvent-accessible surface area for identifying protein carboxylation sites. *J. Comput. Aided Mol. Des.*, **25**, 987–995.
- Marino, S.M. and Gladyshev, V.N. (2012) Analysis and functional prediction of reactive cysteine residues. *J. Biol. Chem.*, **287**, 4419–4425.
- Matthews, B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.
- Mucchielli-Giorgi, M.H. *et al.* (2002) Predicting the disulfide bonding state of cysteines using protein descriptors. *Proteins*, **46**, 243–249.
- Pang, C.N. *et al.* (2007) Surface accessibility of protein post-translational modifications. *J. Proteome Res.*, **6**, 1833–1845.
- Poole, L.B. and Nelson, K.J. (2008) Discovering mechanisms of signaling-mediated cysteine oxidation. *Curr. Opin. Chem. Biol.*, **12**, 18–24.
- Qian, Y. *et al.* (2013) An isotopically tagged azobenzene-based cleavable linker for quantitative proteomics. *ChemBioChem*, **14**, 1410–1414.
- Roos, G. and Messens, J. (2011) Protein sulfenic acid formation: from cellular damage to redox regulation. *Free Radic. Biol. Med.*, **51**, 314–326.
- Seo, Y.H. and Carroll, K.S. (2011) Quantification of protein sulfenic acid modifications using isotope-coded dimedone and iododimedone. *Angew. Chem.-Int. Edit.*, **50**, 1342–1345.
- Sun, M.-A. *et al.* (2012) RedoxDB—a curated database for experimentally verified protein oxidative modification. *Bioinformatics*, **28**, 2551–2552.
- Szychowski, J. *et al.* (2010) Cleavable biotin probes for labeling of biomolecules via the azide – alkyne cycloaddition. *J. Am. Chem. Soc.*, **132**, 18351–18360.
- Vacic, V. *et al.* (2006) Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. *Bioinformatics*, **22**, 1536–1537.
- Wang, C. *et al.* (2014) A chemoproteomic platform to quantitatively map targets of lipid-derived electrophiles. *Nat. Methods*, **11**, 79–85.
- Wani, R. *et al.* (2011) Isoform-specific regulation of Akt by PDGF-induced reactive oxygen species. *Proc. Natl. Acad. Sci.*, **108**, 10550–10555.
- Weerapana, E. *et al.* (2010) Quantitative reactivity profiling predicts functional cysteines in proteomes. *Nature*, **468**, 790–795.
- Wong, Y.H. *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–W594.
- Yang, J. *et al.* (2014) Site-specific mapping and quantification of protein S-sulphenylation in cells. *Nat. Commun.*, **5**, 4776.
- Zheng, T. *et al.* (2013) Single-stranded DNA as a cleavable linker for bioorthogonal click chemistry-based proteomics. *Bioconjug. Chem.*, **24**, 859–864.