BMC Bioinformatics

SOFTWARE Open Access

SIMLIN: a bioinformatics tool for prediction of S-sulphenylation in the human proteome based on multi-stage ensemble-learning models

Xiaochuan Wang[1,2†], Chen Li[3,4†], Fuyi Li[1,4], Varun S. Sharma[3], Jiangning Song[1,4,5*] and Geoffrey I. Webb[1*]

Abstract

Background: S-sulphenylation is a ubiquitous protein post-translational modification (PTM) where an S-hydroxyl (−SOH) bond is formed via the reversible oxidation on the Sulfhydryl group of cysteine (C). Recent experimental studies have revealed that S-sulphenylation plays critical roles in many biological functions, such as protein regulation and cell signaling. State-of-the-art bioinformatic advances have facilitated high-throughput in silico screening of protein S-sulphenylation sites, thereby significantly reducing the time and labour costs traditionally required for the experimental investigation of S-sulphenylation.

Results: In this study, we have proposed a novel hybrid computational framework, termed *SIMLIN*, for accurate prediction of protein S-sulphenylation sites using a multi-stage neural-network based ensemble-learning model integrating both protein sequence derived and protein structural features. Benchmarking experiments against the current state-of-the-art predictors for S-sulphenylation demonstrated that *SIMLIN* delivered competitive prediction performance. The empirical studies on the independent testing dataset demonstrated that *SIMLIN* achieved 88.0% prediction accuracy and an AUC score of 0.82, which outperforms currently existing methods.

Conclusions: In summary, *SIMLIN* predicts human S-sulphenylation sites with high accuracy thereby facilitating biological hypothesis generation and experimental validation. The web server, datasets, and online instructions are freely available at http://simlin.erc.monash.edu/ for academic purposes.

Keywords: Protein post-translational modification, S-sulphenylation, Bioinformatics software, Machine learning, Ensemble learning

Background

Post-translational modifications (PTMs) of the cellular proteome provide a dynamic regulatory landscape that include both rapid reversible modifications and long-lasting irreversible modifications to cellular perturbations [1]. In particular, reactive oxygen species (ROS), which are highly reactive and toxic molecules generated during mitochondrial metabolism, have been shown to play important signalling roles in the presence of oxidative stress and cellular pathophysiology in various complex diseases when their levels are altered in periods of cellular stress [2–5]. In the redox environment, S-sulphenylation (i.e. S-sulfenylation), a type of PTM that occurs at cysteine residues, is a fleeting and reversible covalent oxidation of cysteinyl thiols (Cys-SH) towards supheric acids (Cys-SOH) in the presence of hydrogen peroxide, which thereby acts as a rapid sensor of oxidative stress [6–12]. Thus far, a number of experiments have validated that S-sulphenylation plays important roles in regulating protein functions under both physiologic and oxidatively stressed conditions [7, 9–11, 13–19]. Despite the lack of knowledge regarding the specific functionality of this redox modification in human cell systems, it has been reported that S-sulphenylation is involved in many signal transduction processes, such as the

* Correspondence: Jiangning.Song@monash.edu; Geoff.Webb@monash.edu
†Xiaochuan Wang and Chen Li contributed equally to this work.
[1]Monash Centre for Data Science, Faculty of Information Technology, Monash University, Melbourne, VIC 3800, Australia
Full list of author information is available at the end of the article

© The Author(s). 2019 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

deubiquitinase activity in ovarian tumors and growth factor stimulation [11, 17, 20]. Furthermore, including S-sulphenylation, more than 200 sulfenic modifications that have been identified in various situations, such as transcription factors, signaling proteins, metabolic enzymes, proteostasis regulators, and cytoskeletal components [17]. Although only approximately 2% of proteins in the human, mouse, and rat proteomes contain cysteine residues [21], it is essential to understand the underlying mechanisms that contribute to the residues' critical roles in various biological processes, such as S-sulphenylation, regulation of oxidative PTMs, and the quantification of sulfenic modification processes [6, 7, 9, 10, 14–16].

Despite the significant progress in selective labelling methods for S-sulphenylation using β-dicarbonyl compounds dimedone and analogues, it remains challenging to accurately characterize protein S-sulphenylation sites experimentally, due to their intrinsic instability and low abundance of cysteine residues [6–8, 11, 17, 20, 22]. Moreover, experimental identification of S-sulphenylation is labour-intensive and particularly difficult due to its intrinsically unstable nature and the diversity of the redox reaction [7, 8, 11]. Therefore, in order to assist biologists with characterization of S-sulphenylation sites and S-sulphenylated sequences, it is imperative to construct a generalizable computational tool for highly accurate prediction of protein S-sulphenylation sites.

To date, several algorithms for S-sulphenylation prediction have been published, including MDD-SOH, SOHSite [6, 7], SOHPRED [23], Press [24], iSulfCys [25], SulCysSite [26], PredSCO [27], the predictor by Lei et al [28], and SVM-SulfoSite [29]. Among these computational tools, to the best of our knowledge, the most representative algorithm for S-sulphenylation prediction is MDD-SOH, along which the training dataset in this study was assembled. MDD-SOH is a two-stage ensemble learning model based only on SVM classifiers built upon the previous "SOHSite" project [6, 7]. Despite the progress of computational methods for S-sulphenylation prediction, the prediction performance needs to be further improved, due to the low abundance of cysteine residues and the insufficient number of experimentally verified S-sulphenylation sites.

In this study, we propose a novel bioinformatics tool for improved prediction of protein S-sulphenylation sites, named *SIMLIN*, integrating a number of protein sequence-derived and protein structural features based on the sequence motifs previously identified in [6, 7]. *SIMLIN* is a two-layer framework consisting of Support Vector Machine (SVM) and Random Forests (RF) in the first layer and neural network models in the second layer. To further improve the prediction accuracy of *SIMLIN*, an incremental feature selection method was employed, based on by the mRMR approach implemented in the R package "mRMRe" [30]. The constructed SVM and RF models, trained on different feature clusters plus the selected feature set, were used as the input for the neural network in the second layer. Empirical assessment on the independent testing dataset demonstrated that *SIMLIN* achieved a prediction accuracy of 88% and an AUC score of 0.82, outperforming the existing methods for S-sulphenylation site prediction.

Implementation

Figure 1 provides an overview of the framework of *SIMLIN*, which consists of four major steps: (i) data collection, (ii) feature calculation and selection, (iii) model training, and (iv) performance evaluation. During the data collection process, we collected experimentally verified S-sulphenylation sites from the study of Bui et al. [7]. The negative dataset (defined as proteins without experimentally validated S-sulphenylation sites) was extracted from the UniProt database [31]. Refer to the section 2.1 for more details regarding data collection and pre-processing. For feature extraction, a variety of protein sequence and structural features were extracted and selected using the MDL (minimum descriptive length) technique [32] and mRMR (minimum-redundancy maximum-relevancy) algorithm [30, 33]. A detailed description and statistical summary of the calculated features are provided in the Section 2.2. To construct accurate predictive models, at the 'Model Construction' step, a generalized ensemble framework of *SIMLIN* was developed by integrating various machine-learning algorithms including Artificial Neural Networks (ANNs) [34, 35], SVMs with various kernel functions [36, 37], and RFs [38]. To evaluate and compare the prediction performance of *SIMLIN* with the existing methods, at the last step, we assessed the prediction performance of different algorithms on both 10-fold stratified cross-validation sets and independent datasets assembled in the previous study of Bui et al [7].

Data collection and pre-processing

Both benchmark and independent test datasets in this study were extracted from the 'SOHSite' web server, constructed by Bui et al. [6, 7]. Sequence redundancy of the dataset was removed in this study (using 30% as the sequence identity threshold), which was reported to be the most complete dataset for S-sulphenylation to date through the integration of experimentally validated S-sulphenylation sites from four different resources: (i) the human S-sulphenylation dataset assembled using a chemoproteomic workflow involving the S-sulfenyl-mediated redox regulation [11], by which the S-sulphenylation cysteines were identified; (ii) the RedoxDB database [39], which curates the protein oxidative modifications including S-sulphenylation sites; (iii) the UniProt database [31], and (iv) related literature. Considering the frequent updates of UniProt, based on the gene names provided in

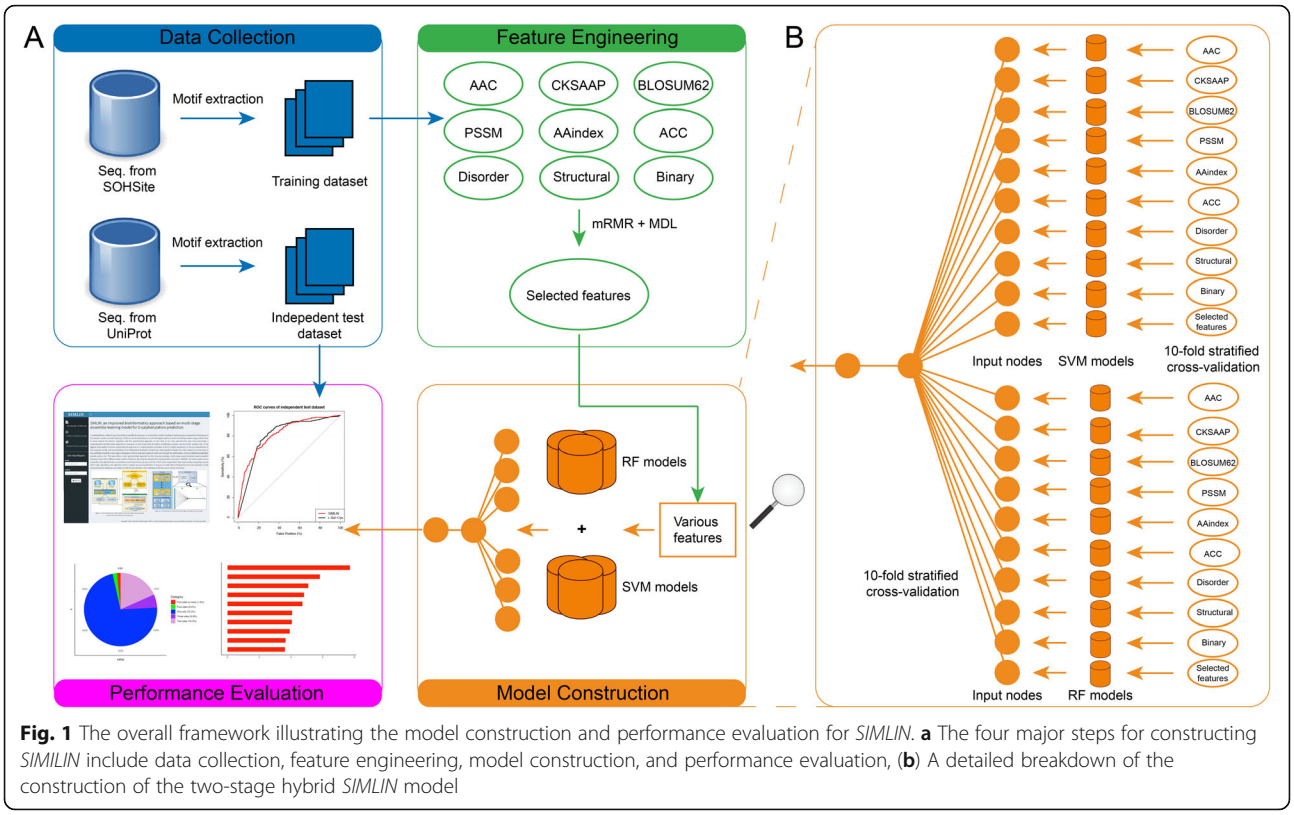

Fig. 1 The overall framework illustrating the model construction and performance evaluation for *SIMLIN*. **a** The four major steps for constructing *SIMILIN* include data collection, feature engineering, model construction, and performance evaluation, (**b**) A detailed breakdown of the construction of the two-stage hybrid *SIMLIN* model

the datasets, we further mapped these proteins to the UniProt database (downloaded November 2016). The canonical protein sequences harboring experimentally verified S-sulphenylation sites were retrieved and downloaded from the UniProt database. Motifs of 21 amino acids with the S-sulphenylation site in the center and flanked by 10 amino acids each side were then extracted from the protein sequences. The highly homologous motifs have been further removed to maximize the sequence diversity according to [7, 13]. The resulting dataset contains a total of 1235 positive samples (i.e. with S-sulphenylation sites) and 9349 negative samples (i.e. without S-sulphenylation sites). Table 1 provides a statistical summary of the benchmark and independent test datasets, respectively.

Feature extraction and calculation

To numerically represent the sequence motifs in the datasets, we calculated and extracted both sequence-based and structural features [40]. In total nine types of sequence-derived and structural features were extracted and used, including the composition of *k*-spaced amino acid pairs (CKSAAP) [41], motif binary representations [42], amino acid substitution matrix (BLOSUM62) [43], protein specific scoring matrix (PSSM) by PSI-BLAST [44], amino acid index (AAindex) [45], amino acid composition (AAC), surface accessibility (ACC) based on protein secondary structure prediction, protein predicted disordered region, and protein predicted secondary structure. The detailed information about each type of features and its feature dimensionality is shown in Table 2.

Composition of k-spaced amino acid pairs (CKSAAP)

The CKSAAP encoding theme has been widely applied [46–49], which represents a protein sequence using the compositions of amino acid pairs spaced by the *k* residues [41, 50, 51]. The composition of each possible *k*-spaced amino acid pair *i* can be therefore calculated based on the following formula:

Table 1 The statistics of datasets employed in this study

	Number of positive motifs	Number of negative motifs	Total
Training dataset	1019	7937	8956
Independent test dataset	216	1412	1628
Total	1235	9349	10,584

Table 2 The sequence and structural features extracted and the feature dimensionalities

Feature type	Feature Cluster	Dimension
Sequence	AAC	20
	CKSAAP	2400
	BLOSUM62	441
	PSSM	400
	AAindex	1344
	Binary	441
Structural	Predicted protein disordered region	20
	Predicted protein secondary structure	84
	Predicted surface accessibility	147
Total		5297

$$CKSAAP[i = 1, 2, 3, ...(k_{max} + 1) \times 400] = N_i/(W - k - 1), \quad (1)$$

where N_i is the number of the k-spaced amino acid pair i, W denotes the window size, and k_{max} represents the maximum space considered — which has been optimized as $k_{max} = 5$ in this study [42]. In total, the CKSAAP scheme generated a feature vector of 2400 dimensions for each motif.

Motif one-hot encoding (binary)

Each motif was also presented using a binary encoding scheme [42], where each amino acid in the motif was denoted using a 21-dimensional vector organized via the alphabetic order of 20 natural amino acids and a gap-filling residue "X". The value 1 was used to denote that the amino acid was in fact in the motif and was placed in its corresponding position in the vector, while other positions in the vector were filled with "0". For instance, the residue C (cysteine) is denoted as {0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0}. Therefore, for a motif with 21 amino acids, a total of 441 (21 × 21) features were generated using the motif binary representation scheme.

Amino acid substitution matrix (BLOSUM62)

The BLOSUM62 is a widely used amino acid substitution matrix based on sequence alignment [43, 52] and has been employed in a variety of bioinformatic studies [6, 22, 53–55]. For each amino acid, a 21-dimensional vector consisting of substitution scores of all 20 amino acids and an additional terminal signal constitute the matrix. For each motif, a 21 × 21 matrix was used and a total number of 441 features were added.

Position-specific scoring matrix (PSSM)

Using the UniRef90 dataset from the UniProt database, we performed PSI-BLAST (version 2.2.26) search to generate the PSSM for each motif in our dataset to represent the sequence conservation and similarity scores. PSSM has been widely applied in a variety of bioinformatics studies as a crucial sequence feature type. Similar to the feature representation of BLOSUM62, 441 features were finally generated for each motif.

Amino acid index (AAindex)

AAindex is a collective database that provides a variety of physical and chemical properties of amino acids [45]. A number of bioinformatics studies have benefited from use of these amino acid properties provided in the AAindex database [46, 48, 56]. Due to the high diversity of the properties offered in the AAindex database, Saha et al. [57] further categorized these indices into eight clusters, which were used for the AAindex feature set for each motif in our study. Therefore, we utilized a selected set of AAindex (i.e., a vector of 1344 dimensions (21 × 8 × 8) [52] attributes to represent each motif.

Amino acid composition (AAC)

For the ACC encoding, each motif is represented as a 20-dimensional vector, where each dimension denotes the number of occurrence of each amino acid within the given motif and is further normalized (i.e. divided by the length of the motif [22]).

Predicted protein disordered region

Given the strong relationships between protein disordered regions and PTMs [58–63], we also integrated the predicted disordered region of a protein as a feature set. To do so, we conducted protein disordered region prediction using DISOPRED (Version 3.1) [64] based on protein sequence. Each amino acid is given a predictive score by DISOPRED, which indicates the likelihood of being located in the protein's disordered region. For a sequence motif of 21 residues, a 20-dimensional vector of predicted scores (i.e. 10 scores for the upstream and 10 scores for the downstream amino acids, respectively) was constructed.

Predicted protein secondary structure

PSIPRED (Version 3.5) [65, 66] was employed to predict protein secondary structure based on the protein's amino acid sequence. The predictive outputs of PSIPRED contain four scores for each residue including the predicted structural class (i.e. C, coil; E, beta strand; and H, alpha helix) and the probabilities of each structural class. As a result, for a motif with 21 amino acids, an 84-dimensional (including three probabilities and the recommendation for each residue) vector was generated for the predicted protein secondary structure feature.

Predicted surface accessibility (ACC)

The surface accessibility feature was calculated using the NetSurfP-1.1 algorithm [67] based on the protein sequences. Each residue in the protein is represented using seven predictive scores, indicating the accessibility (i.e. if this residue is buried), relative surface accessibility, absolute surface accessibility, Z-fit score, probability of this residue being in alpha-helices, beta-strands, and coils. Note that the predictive scores of each category generated by NetSurfP range widely. Therefore, we employed the Min-Max method to normalize the prediction scores of each type [35]. The formula we used for the data normalization was as follows:

$$V_{ij} = \frac{V_{ij} - min_{j\in\{1\ldots m\}}\{V_{ij}\}}{max_{j\in\{1\ldots m\}}\{V_{ij}\} - min_{j\in\{1\ldots m\}}\{V_{ij}\}}, \quad (2)$$

where V_{ij} represents the value i of the feature category vector j, and m denotes the number of observations represented in the vector j. As a result, all values were rescaled to the range between 0 and 1.

Feature selection

As shown in Table 2, a total of 5297 sequence and structural features were calculated and extracted. Such high-dimensional feature vectors might contain misleading and noisy information, which would lead to biased model training. Furthermore, it would require considerable time and effort to build computational models based on such high-dimensional feature set. Therefore, we employed the mRMR (minimum Redundancy Maximum Relevance) [30, 33] package and forward incremental feature selection to eliminate noisy and less informative features from the original feature vector. To perform feature selection, we first applied mRMR to calculate and rank the importance score of each feature. Then, based on the feature importance ranking provided by mRMR, we initiated an empty set and added one feature from the original feature set at a time. The AUC values based on the current feature set were evaluated for both RF and SVM independently, and the resulting feature subset was formed using the features that resulted in higher AUC values for both SVM and RF models. Each feature was incrementally added into the optimized feature set based on the scores of feature importance provided by the mRMR until the curve of AUC values achieved its peak. As described, by applying this forward stepwise sequential variable elimination, the feature with the highest importance was selected. According to the RF algorithm, the global permuted importance is based on the out-of-bag sample B of the tree t in the forest F for each feature X_j and is defined as follows [22, 35, 38]:

$$f_{imp}(X_j) = \frac{\sum_{i\in B} I(y_i = y_i') - I(y_i = y_{ij}')}{|B|}. \quad (3)$$

Model construction

As shown in Fig. 1, the development of *SIMLIN* consists of two major stages after feature selection: (i) employing SVM and RF models based on different feature types (Table 2) to generate the input for the neural network models, and (ii) training of the neural network model based on the optimized RF and SVM models to deliver the final predictive outputs. During the first stage, ten RF and SVM models were constructed based on the nine types of features and the selected feature set. 10-fold stratified cross-validation was performed on the training dataset to select the best model (i.e. with highest AUC values) for each feature type. During the second stage, we built a neural network model which consists of three layers including an input layer, a hidden layer, and an output layer. The first layer harbours 20 nodes to take the output of the best RF and SVM models as the input based on the 10-fold stratified cross-validation performed during the first stage, while the hidden and output layers only have one node (denoted as H_1 and O_1, respectively). Furthermore, in the hidden layer, in addition to H_1, two extra nodes, B_1 and B_2, were auto-generated nodes by the neural network algorithm for the purpose of model balancing. Lastly, the O_1 node in the output layer represents the prediction outcome from the entire algorithm.

We applied a number of software packages to implement *SIMLIN* in our study, including the Python-based machine learning package "scikit-learn" [68], and various R packages of SVM (combining "kernelab" and "e1071") and neural network model ("nnet") [35, 69]. The feature selection techniques employed in our study, including mRMR and MDL, were implemented based on the R packages "mRMRe" and "discretization" [70–72], respectively. Additionally, R packages "caret" [73] and "fscaret" [74] have been used in combination for the control of overall workflow for model training and parameter optimization.

Prediction performance evaluation

We applied widely used measures to evaluate and compare the prediction performance of *SIMLIN*, including the Area Under the Curve (AUC), Accuracy, Sensitivity, Specificity and Matthew's Correlation Coefficient (MCC) [75–77]. During the model training process, AUC was used as the main measure for parameter optimization. The performance measures used are defined as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN},$$

$$Sensitivity = \frac{TP}{TP + FN},$$

$$Specificity = \frac{TN}{TN + FP},$$

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

where *TP*, *TN*, *FP*, and *FN* denote the numbers of true positives, true negatives, false positives and false negatives, respectively. In this study, the S-sulphenylation sites were regarded as the positives, while the non-S-sulphenylation sites were considered as the negatives for the statistics of AUC, specificity and sensitivity.

Results and discussion

Motif conservation analysis and feature selection

We first performed the motif conservation analysis using both benchmarking and independent test datasets. Two sequence logos with the human proteome as the background set generated by pLogo are shown in Fig. 2. In general, the over- and under-represented amino acids surrounding the central cysteine are similar across the benchmarking and independent test datasets. In accordance with the conclusion by Biu et al., amino acids such as leucine (L), lysine (K), glutamate (E), and aspartate (D) are over-represented, while cysteine (C), serine (S), and phenylalanine (F) are under-represented.

Prior to the construction of *SIMLIN*, based on the calculated and extracted features (Table 2), we generated another feature set which contains selected features from the original combined features (i.e. AAC, CKSAAP, BLOSUM62, PSSM, AAindex, ACC, Protein predicted disordered region, Protein secondary structure prediction, and Binary) using stepwise forward sequential variable elimination. As a result, the AUC achieved its highest value of 0.72 (sensitivity: 0.95; specificity: 0.19; accuracy: 86.6%; MCC: 0.182) when 166 features were selected. Among the selected 166 features, 110 (66.3%) and 56 (33.7%) were sequence and structural features, respectively. A detailed breakdown list of these features in terms of feature types and names is available in supplementary material (Additional file 1: Table S1).

Model constructions in the two stages of *SIMILN*

At the first stage of *SIMILN* construction, we built nine SVM and RF models based on the nine clusters of calculated features (Table 2), respectively. Additionally one SVM and RF models were also constructed using the set of selected features (Additional file 1: Table S1). The RF and SVM models were constructed and assessed via 10-fold stratified cross-validation and the average AUC values are shown in Table 3. For the RF models, to reach the optimal performance, the number of trees was set to the nearest integer of the subspace dimensionality of the classification task, which is the square root of the predictors' number. For the SVM models, different kernels were used including the polynomial, radial sigma, and linear kernels for each feature set. The AUC-based performance optimization and kernel selection was performed automatically by the R packages "caret" and "kernelab". The best-performing kernels and their corresponding AUC values were listed in Table 3. It can be seen from Table 3 that SVM and RF models provided competitive performance when using different types of features; however, the RF model outperformed the SVM model on the selected feature set. As shown in Fig. 3, the outputs of the 20 constructed models (i.e. ten RF and ten SVM models; the first layer) were used as inputs for the second layer, i.e. the neural network model, where the nodes, from I_1 to I_{20} took the output of the 20 models based on the outputs of RF and SVM models.

At the second stage a Feed-Forward Neural Network with three layers - including an input layer (20 nodes), a hidden layer (3 nodes) and an output layer (1 node) — was constructed using the R package 'nnet' and subsequently evaluated. Similar to the RF and SVM construction, 10-fold

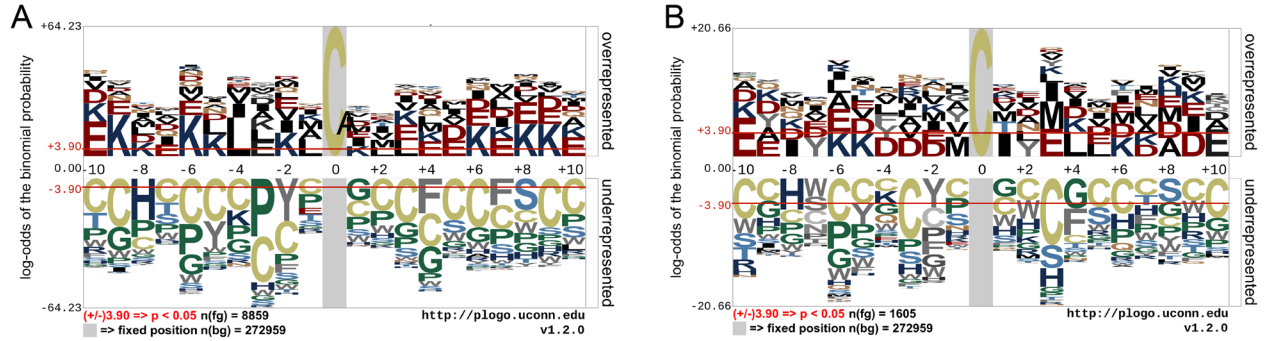

Fig. 2 Motif conservation analysis of S-sulphenylation using the human proteome as the background set for (**a**) benchmarking and (**b**) independent datasets

Table 3 The AUC values of RF and SVM models constructed using different feature sets at the first stage

Feature sets	AUC	
	RF (class weight balanced)	SVM (kernel function)
AAC	**0.68**	0.63 (Polynomial kernel)
AAindex	0.69	0.69 (Radial basis function kernel with grid search hyperparameter tuning)
ACC	**0.71**	0.64 (Radial basis function kernel)
BINARY	0.59	**0.71** (Polynomial kernel)
BLOSUM62	0.68	**0.74** (Radial basis function kernel)
CKSAAP	**0.66**	0.63 (Polynomial kernel)
DISOPRED	0.54	**0.55** (Linear kernel)
PSIPRED	**0.62**	0.60 (Polynomial kernel)
PSSM	**0.73**	0.71 (Polynomial kernel)
Selected features (mRMR+forward consequential elimination)	**0.75**	0.72 (Linear kernel)

The bold font shows the highest performance of each feature among the RF and SVM

stratified cross-validation was employed using the training dataset for building the neutral network model. During the training process, two parameters (i.e. the number of units in the hidden layer and the weight decay for optimising the performance and minimizing the overfitting) were automatically adjusted and evaluated by the network model. The values of the two parameters were adjusted automatically and the resulting performance including AUC, sensitivity, and specificity are given in Table 4. Generally, the performance achieved using different numbers of units in the hidden layer and weight decay values was satisfactory. Based on the performance, the number of units and the weight decay were set to 1 and 0.1 in the final neural network model, respectively (Additional file 1: Table S2). This was for the purpose of minimizing the number of nodes in the hidden layer while maximining the AUC value and convergence rate.

Independent test and performance comparison with existing methods

We assessed and compared the prediction performance of *SIMLIN* with state-of-the-art methods for S-sulphenylation prediction on the independent test dataset. The compared approaches included MDD-SOH, SOHSite [6, 7], SOHPRED, PRESS, iSulf-Cys, SulCysSite. We also noticed that several new computational frameworks have been published recently, including PredSCO [27], the predictor by Lei et al [28], and SVM-SulfoSite [29]. However, due to the inaccessibility of source codes or implemented webservers, we

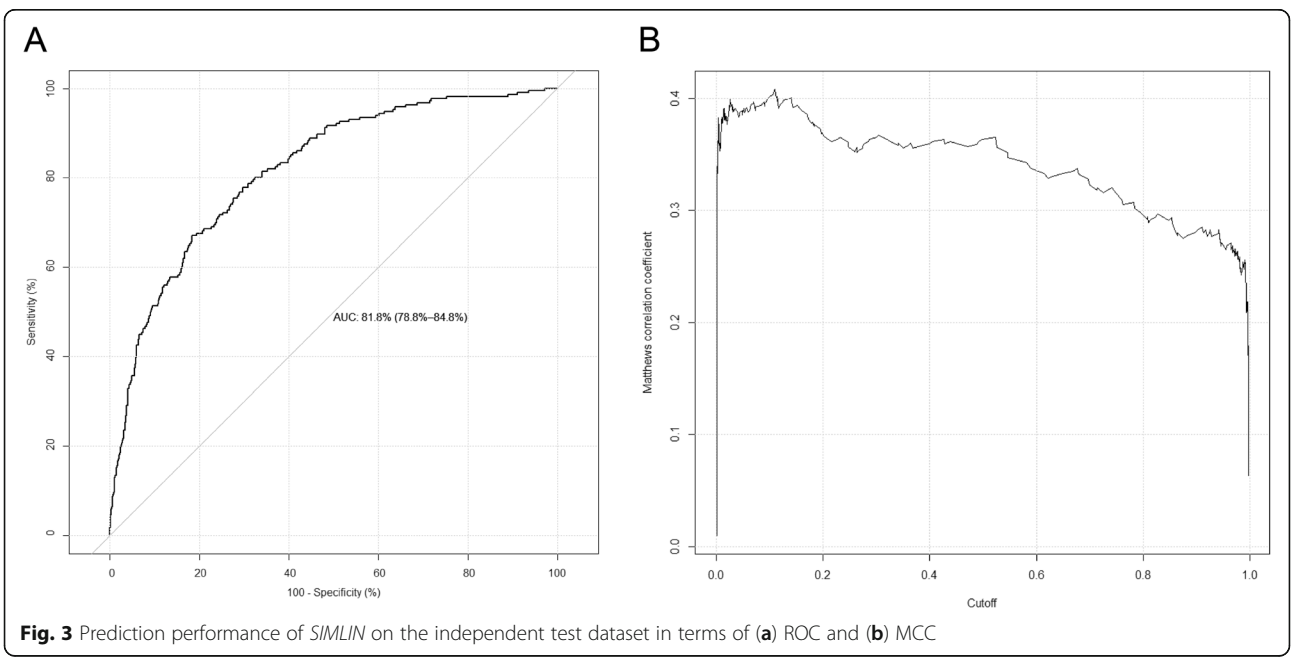

Fig. 3 Prediction performance of *SIMLIN* on the independent test dataset in terms of (**a**) ROC and (**b**) MCC

Table 4 Prediction performance of the neural network model with different units in the hidden layer via 10-fold stratified cross-validation test

#Units in the hidden layer	Decay	AUC	Sensitivity	Specificity
1	0	0.999842 ± 3.15E-4	0.999685 ± 6.30E-4	1
	0.0004	0.999994 ± 6.30E-5	0.999887 ± 3.62E-4	1
	0.1	1	0.999874 ± 3.68E-4	1
3	0	0.999874 ± 3.35E-4	0.999723 ± 6.84E-4	1
	0.0004	0.999987 ± 8.85E-5	0.999937 ± 2.76E-4	1
	0.1	1	0.999874 ± 3.80E-4	1
5	0	0.999793 ± 5.90E-4	0.999685 ± 7.02E-4	0.999902 ± 9.80E-4
	0.0004	0.999869 ± 7.28E-4	0.999912 ± 4.48E-4	0.999704 ± 2.20E-3
	0.1	1	0.999899 ± 3.44E-4	1

were not able to compare their prediction results on our independent test dataset with the performance of *SIMLIN*. From Table 5 and Fig. 3, it is clear that generally *SIMLIN* outperformed the compared approaches. Compared to MDD-SOH, an important advantage of *SIMLIN* is that it does not require any pre-classified motifs. iSulf-Cys is another computational framework that employs a similar approach to create a unified predictive model, but it only used SVM models with three major encoding features (AAindex, binary and PSAAP) for model construction. The overall performance of iSulf-Cys is lower than *SIMLIN*. On the 95% CI the accuracy of iSulf-Cys is 0.7155 ± 0.0085; while *SIMLIN* achieved a prediction accuracy of 0.88 (0.857–0.892) on the 95% CI. The MCC value of *SIMLIN* was also higher than iSulf-Cys (0.39 vs. 0.3122). The SulCysSite model is mainly developed based on the multistage RFs with four major features (AAindex, binary amino acid codes, PSSM, and compositions of profile-based amino acids). Although SulCysSite achieved an AUC of 0.819, it used a biased approach whose final decision was dependent on a complex series of rules, each of which can only cover a small subset. In general, *SIMLIN* outperformed all the compared methods in terms of sensitivity, MCC, and AUC, demonstrating its ability to accurately predict human S- sulphenylation sites.

Table 5 Performance comparison with existing approaches for S-sulphenylation prediction on the independent test

Method	Sensitivity	Specificity	MCC	Accuracy	AUC
SOHPRED	0.73	0.74	0.34	N.A.[b]	0.80
PRESS	0.68	0.69	0.27	73.8%	N.A.
iSulf-Cys	0.73	0.64	0.31	66.8%	0.72
SulCysSite	0.77	0.71	N.A.	72.0%	0.76
SIMLIN	**0.88**	0.56	0.39	**88.0%**	0.82
MDD–SOH[a]	0.85	**0.87**	**0.58**	87.0%	N.A.

[a]The performance values of MDD-SOH were extracted from the study of Bui et al [6]
[b]N.A.: not available
The bold font shows the highest performance of each feature among the RF and SVM

Proteome-wide prediction and functional enrichment analysis

In order to more effectively portray the distribution of predicted S-sulphenylation sites and their potential molecular functions, we performed human proteome-wide S-sulphenylation site prediction using the protein sequences collected from the UniProt database (Version Sep 2017) and our proposed *SIMLIN* framework. We first conducted statistical analysis on the distribution of predicted S-sulphenylation sites in proteins followed by a Gene Ontology (GO) enrichment analysis to reveal the potential cellular localization, biological function, and signalling/metabolic pathways involved in the predicted S-sulphenylation sites using the DAVID biological functional annotation tool (Version 6.8) [78, 79].

Figure 4a-d display the top ten enriched candidates of our gene ontology and pathway enrichment analysis, in terms of molecular function, biological process and cellular component. Figure 4e shows the distribution of numbers of predicted S-sulphenylation sites in the human proteome. In terms of molecular function, the ATPase related activities (i.e., ATPase activity, coupled to movement of substances with a significant p-value of 8.5×10^{-21}; ATPase activity, coupled to transmembrane movement of substances - 8.5×10^{-21}; ATPase activity - 3.42×10^{-14}) have been found to be significantly enriched in proteins with predicted S-sulphenylation sites (Fig. 4a). An example of such relationship has been demonstrated in the study by Wojdyla et al. [80] where Acetaminophen (APAP) treatment has been shown to influence the ATP production, and the APAP-induced S-sulphenylation may act as one contributing fact to such effect. All enriched biological processes shown in Fig. 4b are metabolic processes, which indicate the important roles of S-sulphenylation in metabolism [11]. For instance, one S-sulphenylation occurring at C212 of a fatty acid synthase (FASN) protein may play a role in blocking an active site (C161), which is responsible for fatty acid synthase (Fig. 3B; fatty acid metabolic process - 5.82×10^{-17}) [11, 81]. While for cellular

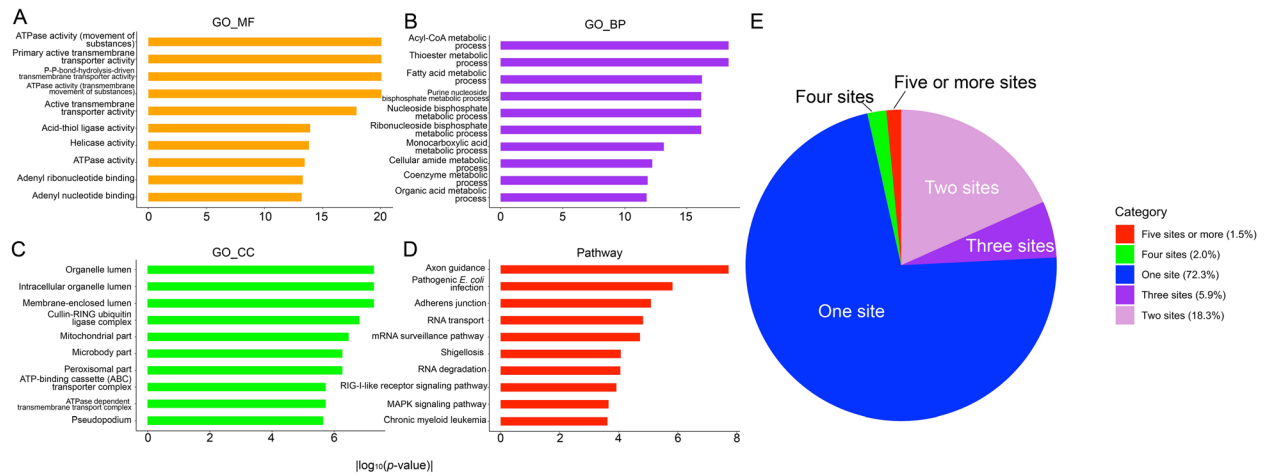

Fig. 4 Gene ontology enrichment analysis of the predicted protein S-sulphenylation sites in the human proteome using *SIMLIN*: top 10 significant (**a**) molecular function terms (GO_MF), (**b**) biological process terms (GO_BP), (**c**) cellular component terms (GO_CC), (**d**) pathways; and (**e**) distribution of the numbers of predicted S-sulphenylation sites

component category (Fig. 4c), the top three localisations are organelle (5.30×10^{-08}), intracellular organelle (5.30×10^{-08}) and membrane-enclosed lumens (5.30×10^{-08}), which is consistent with the analysis of Bui et al [6, 7] RNA transport is an important process associated with protein synthesis, which consists of 14 proteins enriched in S-sulphenylation and S-nitrosylation sites [80], highlighting the necessity of protein S-sulphenylation sites in RNA transport (Fig. 4d; 1.50×10^{-05}). Figure 3e shows the distribution of the numbers of predicted S-sulphenylation site contained in each protein. Expectedly, most of the proteins (72.3%) only contain one predicted site; while only 1.5% of the human proteome harbour five or more predicted sites. A full list of the predicted S-sulphenylation sites on human proteome is freely available on the *SIMLIN* webserver.

Case study of predicted S-sulphenylation using *SIMLIN*

As aforementioned, compared with the dataset used for training *SIMLIN*, three more S-sulphenylation sites have been recently identified and added to the UniProt database, including BRF2_HUMAN (position 361 of Q9HAW0) [82], PTN7_HUMAN (position 361 of P35236; by similarity according to UniProt) and UCP1_HUMAN (position 254 of P25874; by similarity according to UniProt). *SIMLIN* precisely predicted all of these three S-sulphenylation sites, with the possibility scores of 0.997, 0.999 and 0.998, respectively, illustrating the predictive power and capacity of *SIMLIN* for predicting human S-sulphenylation sites.

Implementation and usage of the *SIMLIN* webserver

The open-access web application for *SIMLIN* was implemented using the Shiny framework (Version 1.3.0.403) in R language combining with Node.js (Version 0.10.21) and is freely available for academic use at http://simlin.erc.monash.edu/. The *SIMLIN* server resides on a Linux server, equipped with dual AMD Opteron CPUs, 8 GB memory, and 10 GB disk space. *SIMLIN* accepts both individual protein and a sequence file with the size limit of 1 MB as the input in FASTA format. An 'Example' link has been provided to demonstrate the predictive functionality of the service and guide users to conveniently use it. As the training dataset of *SIMLIN* was collected from the human proteome, the prediction results delivered by *SIMLIN* should be interpreted at the users' discretion if the input protein is from other species rather than *Homo sapiens*. A graphical illustration of the *SIMLIN* webserver in terms of input and output is provided in Fig. 5.

Conclusion

In light of the biological importance of S-sulphenylation, it is imperative to develop easy-to-use computational approaches for the accurate identification of S-sulphenylation sites. In this article, we present *SIMLIN*, a hybrid computation al framework integrating RF, SVM, and neural network models and sequence and structural features of S-sulphenylated motifs and proteins. Performance assessment on both cross-validation and independent test sets demonstrated that *SIMLIN* achieved outstanding prediction performance compared to state-of-the-art computational approaches (MDD-SOH, SOHSite, SOHPRED, PRESS, iSulf-Cys, and SulCysSite) for S-sulphenylation prediction. A user-friendly webserver has also been implemented to provide high-quality predictions of human S-sulphenylation sites using the optimised hybrid *SIMLIN* framework. Proteome-wide prediction of S-sulphenylation sites for the entire

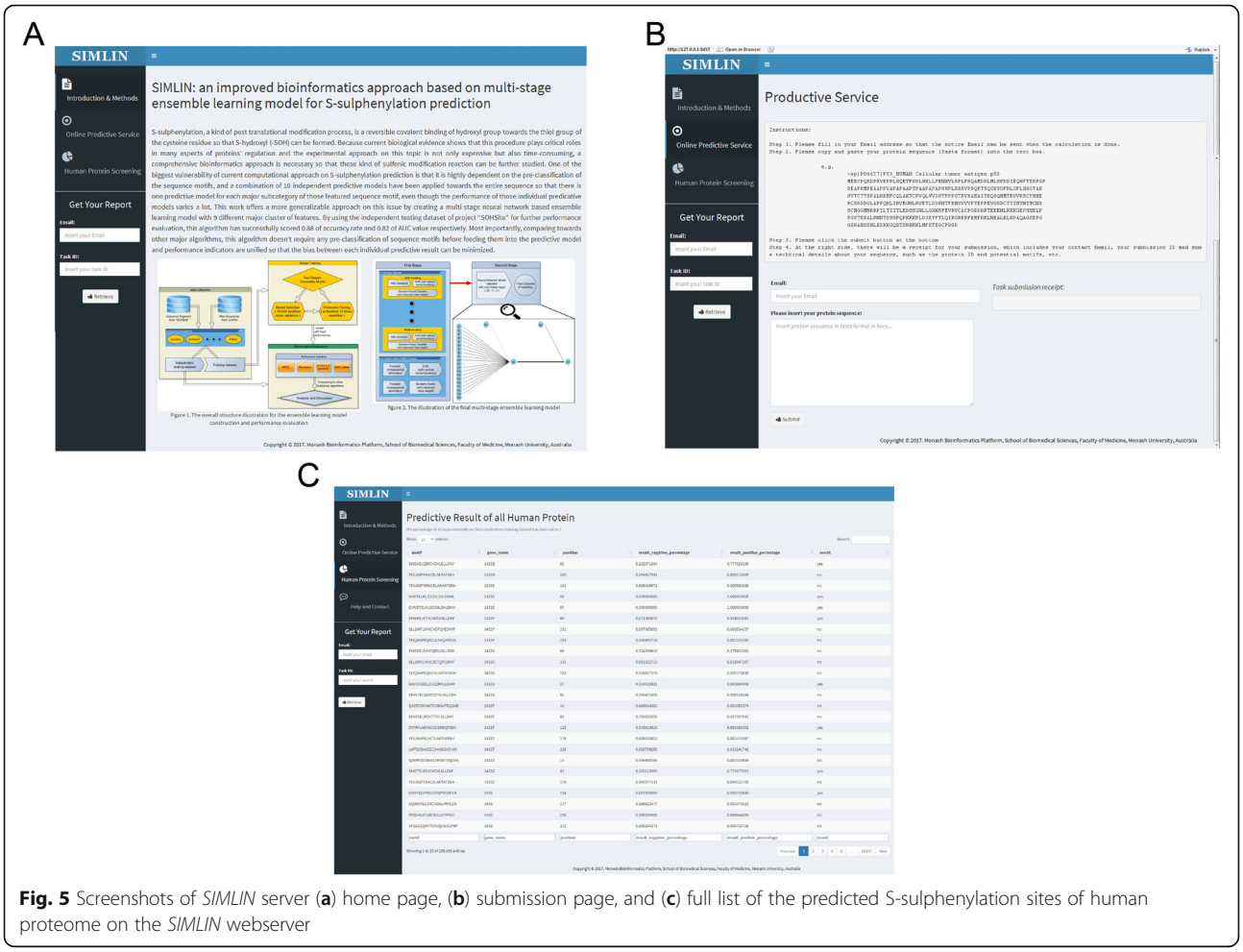

Fig. 5 Screenshots of *SIMLIN* server (**a**) home page, (**b**) submission page, and (**c**) full list of the predicted S-sulphenylation sites of human proteome on the *SIMLIN* webserver

human proteome extracted from the UniProt database, has been made available at the *SIMLIN* webserver, aiming to provide highly accurate S-sulphenylation sites and facilitate biologists' efforts for experimental validation, hypothesis generation, and data analysis. We anticipate that *SIMLIN* will be explored as a useful tool for human S-sulphenylation prediction. This effective framework can also be generally applied to address the prediction problem of other protein PTMs.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12859-019-3178-6.

Additional file 1: Table S1. A detailed summary of the selected sequence and structural features using the MDL and mRMR feature selection methods. **Table S2.** The assigned weights of each node in the final neural network model.

Abbreviations

AAC: amino acid composition; ACC: accuracy; ACC: surface accessibility; ANN: artificial neural network; AUC: area under the ROC curve; CKSAAP: composition of k-spaced amino acid pairs; FN: false negative; FP: false positive; GO: gene ontology; MCC: Matthews' Correlation Coefficient; MDL: minimum descriptive length; mRMR: minimum Redundancy Maximum Relevance; PSSM: protein-specific scoring matrix; PTM: post-translational modification; RF: Random Forest; SVM: Support Vector Machine; TN: true negative; TP: true positive

Acknowledgments

We acknowledge the anonymous reviewers' constructive comments, which have greatly helped to improve the scientific quality of this study.

Authors' contributions

G.W. and J.S. conceived the project and designed the experiments. X.W. and C.L. performed the model construction, data analysis and drafted the manuscript. F.L. and V.S. provided useful comments and assisted with the data analysis and webserver construction. All authors read, revised and approved the final manuscript.

Funding

This work was supported by grants from the Australian Research Council (ARC) (LP110200333 and DP120104460), National Health and Medical Research Council of Australia (NHMRC) (1144652, 490989), the National Institute of Allergy and Infectious Diseases of the National Institutes of Health (R01 AI111965), and a Major Inter-Disciplinary Research (IDR) Grant Awarded by Monash University (201402). C.L. is currently supported by an NHMRC CJ Martin Early Career Research Fellowship (1143366). The funding bodies ARC, NHMRC, NIH, and Monash University had no role in the design of the study; collection, analysis, and interpretation of data; or in writing the manuscript.

Availability of data and materials
The datasets of this study are available at http://simlin.erc.monash.edu/.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
J.S. is an Associate Editor of BMC Bioinformatics.

Author details
[1]Monash Centre for Data Science, Faculty of Information Technology, Monash University, Melbourne, VIC 3800, Australia. [2]Division of Cancer Epidemiology, Cancer Council Victoria, Melbourne, VIC 3004, Australia. [3]Institute of Molecular Systems Biology, Department of Biology, ETH Zürich, 8093 Zürich, Switzerland. [4]Infection and Immunity Program, Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Melbourne, VIC 3800, Australia. [5]ARC Centre of Excellence for Advanced Molecular Imaging, Monash University, Melbourne, VIC 3800, Australia.

Received: 4 October 2019 Accepted: 28 October 2019
Published online: 21 November 2019

References
1. Venne AS, Kollipara L, Zahedi RP. The next level of complexity: crosstalk of posttranslational modifications. Proteomics. 2014;14(4–5):513–24.
2. Liguori I, Russo G, Curcio F, Bulli G, Aran L, Della-Morte D, Gargiulo G, Testa G, Cacciatore F, Bonaduce D, Abete P. Oxidative stress, aging, and diseases. Clin Interv Aging. 2018;13:757–72.
3. Sharma K. Mitochondrial hormesis and diabetic complications. Diabetes. 2015;64(3):663–672
4. Zhao X, Drlicab K. Reactive oxygen species and the bacterial response to lethal stress. Curr Opin Microbiol. 2014.
5. Ristow M. Unraveling the truth about antioxidants: mitohormesis explains ROS-induced health benefits. Nat Med. 2014;20(7):709–11.
6. Bui VM, Lu CT, Ho TT, Lee TY. MDD-SOH: exploiting maximal dependence decomposition to identify S-sulfenylation sites with substrate motifs. Bioinformatics. 2016;32(2):165–72.
7. Bui VM, Weng SL, Lu CT, Chang TH, Weng JT, Lee TY: SOHSite: incorporating evolutionary information and physicochemical properties to identify protein S-sulfenylation sites. BMC Genomics 2016, 17 Suppl 1:9.
8. Leonard SE, Carroll KS. Chemical 'omics' approaches for understanding protein cysteine oxidation in biology. Curr Opin Chem Biol. 2011;15(1):88–102.
9. Leonard SE, Reddie KG, Carroll KS. Mining the thiol proteome for sulfenic acid modifications reveals new targets for oxidation in cells. ACS Chem Biol. 2009;4(9):783–99.
10. Paulsen CE, Carroll KS. Cysteine-mediated redox signaling: chemistry, biology, and tools for discovery. Chem Rev. 2013;113(7):4633–79.
11. Yang J, Gupta V, Carroll KS, Liebler DC. Site-specific mapping and quantification of protein S-sulphenylation in cells. Nat Commun. 2014; 5:4776.
12. Beedle AE, Lynham S, Garcia-Manyes S. Protein S-sulfenylation is a fleeting molecular switch that regulates non-enzymatic oxidative folding. Nat Commun. 2016;7:12490.
13. Chen X, Qiu JD, Shi SP, Suo SB, Huang SY, Liang RP. Incorporating key position and amino acid residue features to identify general and species-specific ubiquitin conjugation sites. Bioinformatics. 2013;29(13):1614–22.
14. Furdui CM, Poole LB. Chemical approaches to detect and analyze protein sulfenic acids. Mass Spectrom Rev. 2014;33(2):126–46.
15. Mucchielli-Giorgi MH, Hazout S, Tuffery P. Predicting the disulfide bonding state of cysteines using protein descriptors. Proteins. 2002;46(3):243–9.
16. Paulsen CE, Truong TH, Garcia FJ, Homann A, Gupta V, Leonard SE, Carroll KS. Peroxide-dependent sulfenylation of the EGFR catalytic site enhances kinase activity. Nat Chem Biol. 2011;8(1):57–64.
17. Radivojac P, Vacic V, Haynes C, Cocklin RR, Mohan A, Heyen JW, Goebl MG, Iakoucheva LM. Identification, analysis, and prediction of protein ubiquitination sites. Proteins. 2010;78(2):365–80.
18. Sun C, Shi ZZ, Zhou X, Chen L, Zhao XM. Prediction of S-glutathionylation sites based on protein sequences. PLoS One. 2013;8(2):e55512.
19. Yang J, Gupta V, Tallman KA, Porter NA, Carroll KS, Liebler DC. Global, in situ, site-specific analysis of protein S-sulfenylation. Nat Protoc. 2015; 10(7):1022–37.
20. Kulathu Y, Garcia FJ, Mevissen TE, Busch M, Arnaudo N, Carroll KS, Barford D, Komander D. Regulation of A20 and other OTU deubiquitinases by reversible oxidation. Nat Commun. 2013;4:1569.
21. Hess DT, Matsumoto A, Kim SO, Marshall HE, Stamler JS. Protein S-nitrosylation: purview and parameters. Nat Rev Mol Cell Biol. 2005;6(2): 150–66.
22. Lee TY, Chen SA, Hung HY, Ou YY. Incorporating distant sequence features and radial basis function networks to identify ubiquitin conjugation sites. PLoS One. 2011;6(3):e17331.
23. Xiaofeng Wang, Renxiang Yan, Jinyan Li, Jiangning Song. SOHPRED: a new bioinformatics tool for the characterization and prediction of human S-sulfenylation sites. Molecular BioSystems. 2016;12(9):2849–58.
24. Marianna Sakka, Grigorios Tzortzis, Michalis D. Mantzaris, Nick Bekas, Tahsin F. Kellici, Aristidis Likas, Dimitrios Galaris, Ioannis P. Gerothanassis, Andreas G. Tzakos. PRESS: PRotEin S-Sulfenylation server. Bioinformatics. 2016;32(17):2710–12.
25. Yan Xu, Jun Ding, Ling-Yun Wu, Bin Liu. iSulf-Cys: Prediction of S-sulfenylation Sites in Proteins with Physicochemical Properties of Amino Acids. PLOS ONE. 2016;11(4):e0154237.
26. Md. Mehedi Hasan, Dianjing Guo, Hiroyuki Kurata. Computational identification of protein S-sulfenylation sites by incorporating the multiple sequence features information. Molecular BioSystems. 2017;13(12):2545–50.
27. Deng L, Xu X, Liu H. PredCSO: an ensemble method for the prediction of S-sulfenylation sites in proteins. Mol Omics. 2018;14(4):257–65.
28. Lei G-C, Tang J, Du P-F. Predicting S-sulfenylation sites using physicochemical properties differences. Lett Org Chem. 2017;144:8.
29. Al-Barakati HJ, McConnell EW, Hicks LM, Poole LB, Newman RH, Kc DB. SVM-SulfoSite: a support vector machine based predictor for sulfenylation sites. Sci Rep. 2018;8(1):11288.
30. Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell. 2005;27(8):1226–38.
31. The UniProt Consortium. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2017;45(D1):D158–69.
32. Fayyad UM, Irani KB: Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. Ijcai-93, Vols 1 and 2 1993:1022–1027.
33. De Jay N, Papillon-Cavanagh S, Olsen C, El-Hachem N, Bontempi G, Haibe-Kains B. mRMRe: an R package for parallelized mRMR ensemble feature selection. Bioinformatics. 2013;29(18):2365–8.
34. Zhang GQP. Neural networks for classification: a survey. Ieee T Syst Man Cy C. 2000;30(4):451–62.
35. Venables WN, Ripley BD. Modern applied statistics with S, 4th edn: springer; 2002.
36. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273–97.
37. Karatzoglou A, Smola A, Hornik K, Zeileis A: kernlab - An S4 Package for Kernel Methods in R. J Stat Softw 2004, **11**(9).
38. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
39. Sun MA, Wang Y, Cheng H, Zhang Q, Ge W, Guo D. RedoxDB--a curated database for experimentally verified protein oxidative modification. Bioinformatics. 2012;28(19):2551–2.
40. Spanig S, Heider D. Encodings and models for antimicrobial peptide classification for multi-resistant pathogens. BioData Min. 2019;12:7.
41. Chen K, Kurgan L, Rahbari M. Prediction of protein crystallization using collocation of amino acid pairs. Biochem Biophys Res Commun. 2007; 355(3):764–9.
42. Chen Z, Zhou Y, Song J, Zhang Z. hCKSAAP_UbSite: improved prediction of human ubiquitination sites by exploiting amino acid pattern and properties. Biochim Biophys Acta. 2013;1834(8):1461–7.
43. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A. 1992;89(22):10915–9.
44. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389–402.
45. Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M. AAindex: amino acid index database, progress report 2008. Nucleic Acids Res. 2008;36(Database issue):D202–5.

46. Li F, Zhang Y, Purcell AW, Webb GI, Chou K-C, Lithgow T, Li C, Song J. Positive-unlabelled learning of glycosylation sites in the human proteome. BMC Bioinformatics. 2019;20(1):112.
47. Song J, Wang Y, Li F, Akutsu T, Rawlings ND, Webb GI, Chou K-C. iProt-sub: a comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites. Brief Bioinform. 2018;20(2):638–58.
48. Li F, Li C, Wang M, Webb GI, Zhang Y, Whisstock JC, Song J. GlycoMine: a machine learning-based approach for predicting N-, C- and O-linked glycosylation in the human proteome. Bioinformatics. 2015;31(9):1411–9.
49. Li F, Chen J, Leier A, Marquez-Lago T, Liu Q, Wang Y, Revote J, Smith AI, Akutsu T, Webb GI, et al. DeepCleave: a deep learning predictor for caspase and matrix metalloprotease substrates and cleavage sites. Bioinformatics. 2019. https://doi.org/10.1093/bioinformatics/btz721.
50. Chen Z, Zhao P, Li F, Leier A, Marquez-Lago TT, Wang Y, Webb GI, Smith AI, Daly RJ, Chou KC, et al. iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences. Bioinformatics. 2018;34(14):2499–502.
51. Chen Z, Zhao P, Li F, Marquez-Lago TT, Leier A, Revote J, Zhu Y, Powell DR, Akutsu T, Webb GI, et al. iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. Brief Bioinform. 2019. https://doi.org/10.1093/bib/bbz041.
52. Eddy SR. Where did the BLOSUM62 alignment score matrix come from? Nat Biotechnol. 2004;22(8):1035–6.
53. Gao J, Thelen JJ, Dunker AK, Xu D. Musite, a tool for global prediction of general and kinase-specific phosphorylation sites. Mol Cell Proteomics. 2010; 9(12):2586–600.
54. Wang Y, Song J, Marquez-Lago TT, Leier A, Li C, Lithgow T, Webb GI, Shen HB. Knowledge-transfer learning for prediction of matrix metalloprotease substrate-cleavage sites. Sci Rep. 2017;7(1):5755.
55. Li F, Li C, Marquez-Lago TT, Leier A, Akutsu T, Purcell AW, Ian Smith A, Lithgow T, Daly RJ, Song J, et al. Quokka: a comprehensive tool for rapid and accurate prediction of kinase family-specific phosphorylation sites in the human proteome. Bioinformatics. 2018;34(24):4223–31.
56. Li F, Li C, Revote J, Zhang Y, Webb GI, Li J, Song J, Lithgow T. GlycoMine(struct): a new bioinformatics tool for highly accurate mapping of the human N-linked and O-linked glycoproteomes by incorporating structural features. Sci Rep. 2016;6:34595.
57. Saha I, Maulik U, Bandyopadhyay S, Plewczynski D. Fuzzy clustering of physicochemical and biochemical properties of amino acids. Amino Acids. 2012;43(2):583–94.
58. Bah A, Forman-Kay JD. Modulation of intrinsically disordered protein function by post-translational modifications. J Biol Chem. 2016;291(13): 6696–705.
59. Collins MO, Yu L, Campuzano I, Grant SG, Choudhary JS. Phosphoproteomic analysis of the mouse brain cytosol reveals a predominance of protein phosphorylation in regions of intrinsic sequence disorder. Mol Cell Proteomics. 2008;7(7):1331–48.
60. Darling AL, Uversky VN. Intrinsic disorder and posttranslational modifications: the darker side of the biological dark matter. Front Genet. 2018;9:158.
61. Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, Obradovic Z, Dunker AK. The importance of intrinsic disorder for protein phosphorylation. Nucleic Acids Res. 2004;32(3):1037–49.
62. Lin Y, Currie SL, Rosen MK. Intrinsically disordered sequences enable modulation of protein phase separation through distributed tyrosine motifs. J Biol Chem. 2017;292(46):19110–20.
63. van der Lee R, Buljan M, Lang B, Weatheritt RJ, Daughdrill GW, Dunker AK, Fuxreiter M, Gough J, Gsponer J, Jones DT, et al. Classification of intrinsically disordered regions and proteins. Chem Rev. 2014;114(13):6589–631.
64. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Mol Biol. 2004;337(3):635–45.
65. Buchan DW, Minneci F, Nugent TC, Bryson K, Jones DT: Scalable web services for the PSIPRED Protein Analysis Workbench. Nucleic Acids Res 2013, 41(Web Server issue):W349–W357.
66. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol. 1999;292(2):195–202.
67. Petersen B, Petersen TN, Andersen P, Nielsen M, Lundegaard C. A generic method for assignment of reliability scores applied to solvent accessibility predictions. BMC Struct Biol. 2009;9:51.
68. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:6.
69. Ripley BD, Hjort NL. Pattern recognition and neural networks. NY, USA: Cambridge University Press New York; 1995.
70. Tay FEH, Shen L. A modified Chi2 algorithm for discretization. IEEE Trans Knowl Data Eng. 2002;14(3):5.
71. Pawlak Z. Rough sets. Int J Computer Info Sci. 1982;11(5):16.
72. Chmielewski MR, Grzymala-Busse JW. Global discretization of continuous attributes as preprocessing for machine learning. Int J Approx Reason. 1996; 15(4):13.
73. Kuhn M. Building predictive models in R using the caret package. J Stat Softw. 2008;28(5):26
74. Szlek Jakub, Paclawski Adam, Lau Raymond,Jachowicz Renata and Mendyk Aleksander. Heuristic modeling of macromolecule release from PLGA microspheres. International Journal of Nanomedicine. 2013:8(1):4601-4611.
75. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta. 1975;405(2):442–51.
76. Li F, Wang Y, Li C, Marquez-Lago TT, Leier A, Rawlings ND, Haffari G, Revote J, Akutsu T, Chou K-C, et al. Twenty years of bioinformatics research for protease-specific substrate and cleavage site prediction: a comprehensive revisit and benchmarking of existing methods. Brief Bioinform. 2018. https://doi.org/10.1093/bib/bby077.
77. Mei S, Li F, Leier A, Marquez-Lago TT, Giam K, Croft NP, Akutsu T, Smith AI, Li J, Rossjohn J, et al. A comprehensive review and performance evaluation of bioinformatics tools for HLA class I peptide-binding prediction. Brief Bioinform. 2019. https://doi.org/10.1093/bib/bbz051.
78. Huang da W, Sherman BT, Lempicki RA: Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 2009, 37(1):1–13.
79. Huang da W, Sherman BT, Lempicki RA: Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 2009, 4(1):44–57.
80. Wojdyla K, Wrzesinski K, Williamson J, Fey SJ, Rogowska-Wrzesinska A. Acetaminophen-induced S-nitrosylation and S-sulfenylation signalling in 3D cultured hepatocarcinoma cell spheroids. Toxicol Res (Camb). 2016;5(3):905–20.
81. Pappenberger G, Benz J, Gsell B, Hennig M, Ruf A, Stihle M, Thoma R, Rudolph MG. Structure of the human fatty acid synthase KS-MAT didomain as a framework for inhibitor design. J Mol Biol. 2010;397(2):508–19.
82. Gouge J, Satia K, Guthertz N, Widya M, Thompson AJ, Cousin P, Dergai O, Hernandez N, Vannini A. Redox signaling by the RNA polymerase III TFIIB-related factor Brf2. Cell. 2015;163(6):1375–87.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions