




CysModDB: a comprehensive platform with the integration of manually curated resources and analysis tools for cysteine posttranslational modifications

Yanzheng Meng [†], Lin Zhang [†], Laizhi Zhang, Ziyu Wang, Xuanwen Wang, Chan Li, Yu Chen, Shipeng Shang and Lei Li 

Corresponding authors: Lei Li, Faculty of Biomedical and Rehabilitation Engineering, University of Health and Rehabilitation Sciences, Qingdao 266071, China. Tel./Fax: +86 532 8581 2983; E-mail: lileime@hotmail.com; Shipeng Shang, School of Basic Medicine, Qingdao University, Qingdao 266071, China.

Tel.: +86 532 8595 1111; Fax: +86 532 8581 2983; E-mail: bio_shangsp@hotmail.com

[†]Yanzheng Meng and Lin Zhang contributed equally to this work.

Abstract

The unique chemical reactivity of cysteine residues results in various posttranslational modifications (PTMs), which are implicated in regulating a range of fundamental biological processes. With the advent of chemical proteomics technology, thousands of cysteine PTM (CysPTM) sites have been identified from multiple species. A few CysPTM-based databases have been developed, but they mainly focus on data collection rather than various annotations and analytical integration. Here, we present a platform-dubbed CysModDB, integrated with the comprehensive CysPTM resources and analysis tools. CysModDB contains five parts: (1) 70 536 experimentally verified CysPTM sites with annotations of sample origin and enrichment techniques, (2) 21 654 modified proteins annotated with functional regions and structure information, (3) cross-references to external databases such as the protein–protein interactions database, (4) online computational tools for predicting CysPTM sites and (5) integrated analysis tools such as gene enrichment and investigation of sequence features. These parts are integrated using a customized graphic browser and a Basket. The browser uses graphs to represent the distribution of modified sites with different CysPTM types on protein sequences and mapping these sites to the protein structures and functional regions, which assists in exploring cross-talks between the modified sites and their potential effect on protein functions. The Basket connects proteins and CysPTM sites to the analysis tools. In summary, CysModDB is an integrated platform to facilitate the CysPTM research, freely accessible via <https://cysmoddb.bioinfo.org/>.

Keywords: posttranslational modification, chemical proteomics, PTM cross-talk, cysteine modification, database

Introduction

Cysteine contains a thiol side chain with high nucleophilicity and redox sensitivity, which makes cysteine susceptible to many reactive molecules and generates different cysteine PTM types (CysPTMs). The modification types can be classified into three categories due to their characteristics: oxidation posttranslational modification (PTM) [1], lipid PTM [2, 3] and metabolite PTM [4]. The oxidation PTM means cysteine oxidation by reactive oxygen species, reactive nitrogen species, reactive sulfur species or glutathione (GSH) [1]. It includes s-nitrosylation, s-sulfenylation, s-sulfinylation, s-sulfonylation, s-glutathionylation, s-disulfidation and s-persulfidation. The lipid PTM refers to cysteine lipidation [3, 5], including s-palmitoylation and s-prenylation. The metabolite

PTM covers a series of nonenzymatic modifications caused by reactive metabolites [6], such as s-itaconation, s-succination and s-carbonylation. Cysteine residues play various functional roles such as metal-binding, enzyme activation and structural stabilization, and cysteine modifications may regulate protein structures and functions [7]. It has been revealed that CysPTMs are associated with many diseases like cancer and neurodegenerative disorder [8, 9].

Identifying CysPTM sites on proteomes is the foundation of exploring their functional roles in biological activities. Nevertheless, it is challenging to directly detect CysPTM sites on a proteomic scale due to their low abundance and significant dynamic changes. Unlike some PTM types (e.g. tyrosine phosphorylation

Yanzheng Meng is an undergraduate student at the School of Basic Medicine, Qingdao University, Qingdao, China. He is an intern at the University of Health and Rehabilitation Sciences.

Lin Zhang is an undergraduate student at the College of Computer Science and Technology, Qingdao University, Qingdao, China.

Laizhi Zhang is an undergraduate student at the School of Basic Medicine, Qingdao University, Qingdao, China.

Ziyu Wang is an undergraduate student at the School of Basic Medicine, Qingdao University, Qingdao, China.

Xuanwen Wang is an undergraduate student at the College of Computer Science and Technology, Qingdao University, Qingdao, China.

Chan Li is a master graduate student at the School of Basic Medicine, Qingdao University, Qingdao, China.

Yu Chen is an associate professor at the College of Computer Science and Technology, Qingdao University, Qingdao, China. His research interests include software engineering, bioinformatics, data mining and machine learning.

Shipeng Shang received his PhD degree from Harbin Medical University, China. He is an assistant professor at the School of Basic Medical Science, Qingdao University, China. His research interests include bioinformatics and databases.

Lei Li received his PhD degree from Nanyang Technological University, Singapore. He is a professor at the University of Health and Rehabilitation Sciences, China. His research interests include bioinformatics, systems biology and proteomics.

Received: February 25, 2022. **Revised:** August 27, 2022. **Accepted:** September 26, 2022

© The Author(s) 2022. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

and lysine acylation) that can be easily enriched by antibodies, the enrichment of the CysPTM sites is mainly based on chemical proteomics techniques, where various chemical probes have been invented [10]. For instance, a chemical probe named cysteine-reactive phosphate tag (CPT) was developed for detecting cysteine oxidation, and around 34 000 oxidation sites were identified on ~9400 mouse proteins [11]. Nevertheless, the CPT probe cannot distinguish different oxidation subtypes. In contrast, the DYn-2 sulfenic acid probe is specific to s-sulfenylation, and 1105 s-sulfenylated sites were identified using this probe from human RKO cells [12].

As a flood of CysPTM sites were identified, several databases have been developed to host CysPTM information (e.g. RedoxDB [13], dbPTM [14], dbSNO [15], dbGSH [16], SwissPalm [17] and iCysMod [18]). Nevertheless, they have a few limitations. First, these databases cover specific or limited CysPTM types. For example, many newly discovered types (e.g. s-itaconation [19] and s-succination [20]) have not been collected. Second, they include insufficient information. Generally, the information on identification techniques or sample origin is omitted, which is essential to estimate PTM data quality, compare the PTM profiles identified using different probes and guide further functional investigation [21]. Third, few visualization browsers are provided to show the CysPTM sites on protein sequences, or they cannot be used to estimate the potential effect of the modified sites on protein functions. In summary, a platform is missing that integrates data collection, annotations of modified sites and proteins and analytical tools to characterize the CysPTM sites, such as PTM cross-talks, sequence features and their effects on protein structures and functions.

In this study, we present an integrative platform-dubbed CysModDB. It contains 70 536 CysPTM sites experimentally identified on 21 654 proteins across 12 organisms, from prokaryotes to eukaryotes. These sites consist of 12 different CysPTM types, which is the most extensive compared with previous databases. CysModDB is composed of five parts: (1) PTM site annotations, including flanking regions, sample origin and enrichment techniques, (2) protein annotations, covering sequences with highlighted CysPTM positions, functional regions, subcellular locations and structure information, (3) cross-references to external databases for protein pathways (Reactome [22]), protein-protein interactions (STRING [23]) and protein PTMs (dbPTM [14]), (4) online computational tools for predicting CysPTM sites and (5) an analysis toolkit, composed of gene enrichment analysis, regulatory network and investigation of sequence features. The five parts are integrated using a customized graphic browser and a Basket. The browser uses graphs to represent the distribution of modified sites with different CysPTM types on protein sequences, the CysPTM co-occurrences and the mapping of these sites to the protein structures and functional regions, which assists in exploring cross-talks between the modified sites and the potential effect of these sites on protein functions. The Basket links selected proteins to the analysis toolkit or external resources for analyzing sequence features and cross-talks of these CysPTMs. In brief, CysModDB is a comprehensive platform with manually curated resources and analysis tools for cysteine modifications.

Materials and methods

Data collection

The data collection pipeline is shown in [Supplementary Figure S1](#). Specifically, we retrieved literature in the PubMed database using 'cysteine' or 'cysteine proteomics' combined with specific

CysPTM names (or synonyms). They include s-nitrosylation (s-nitrosothiols, s-nitrosocysteine or s-nitrosation), s-sulfenylation (cysteine sulfenic acids, s-sulfenation or s-sulphenylation), s-sulfinylation (cysteine sulfinic acid or s-sulfination), s-sulfonylation (cysteine sulfonic acid or s-sulfonation), s-glutathionylation, s-disulfidation (disulfide bonds), s-persulfidation (s-sulfhydration or persulfide), s-palmitoylation (s-acylation), s-prenylation (farnesylation or geranylgeranylation), s-carbonylation (cysteine alkylation or HNE-modified cysteine), s-itaconation and s-succination (s-(2-succino) cysteine). Over 300 articles were obtained, and most were published between 2009 and 2021 ([Supplementary Table S1](#)). After manually scanning these papers, the data from high-throughput proteomics studies were collected as the CysPTM data source ([Figure 1A](#)). We further extracted essential information from the literature, including identification approaches, experimental sample names, protein names and cysteine positions with PTM types.

We explored online prediction algorithms for the CysPTM sites from literature ([Figure 1A](#)) and retained 10 accessible tools to date ([Figure 1B](#)), such as DeepCSO [24], DeepGSH [25], GPS-Palm [26], iPreny-PseAAC [27], iSulf-Cys [28], Mul-SNO [29], PreSNO [30], SIMLIN [31], SulCysSite [32] and pCysMod [33].

Data processing

The CysPTM data extracted from the literature were annotated and integrated with external resources ([Supplementary Figure S2](#)). The CysPTM information was clustered into two levels: modified sites and proteins ([Figure 1B](#)). The information for each site includes the CysPTM type and the related modification category, the identification approach, the name of the sample (or cell line) where the modification was detected, flanking sequence regions from the corresponding protein in the UniProtKB database [34] and the references (i.e. PubMed IDs). The information for each protein includes gene and protein names, UniProt AC, organism, functional description, subcellular location, protein sequence with highlighted CysPTM sites, functional regions, secondary structure [34] and tertiary structure from the AlphaFold database [35, 36]. In addition, every protein was cross-referenced with three external databases (i.e. Reactome Pathway Database, STRING and dbPTM [14, 22, 23]), providing information about involved pathways, protein-protein interactions and PTM cross-talks.

Development of online analysis tools

CysModDB includes four online bioinformatics tools: gene ontology (GO) enrichment analysis, regulatory network, sequence logo and composition heatmap ([Figure 1B](#)). The former two tools were developed for modified proteins, while the latter two were used for modified sequences. The GO enrichment analysis facilitates the enrichment analysis of the CysPTM proteins in specific categories, and the regulatory network reveals the potential regulatory relationships between the modified proteins. The sequence logo and the composition heatmap graphically represent multiple sequence alignment results.

The GO enrichment analysis was developed using the Enrichr web API [37] for statistical protein enrichment analysis. The output would show the results of GO:biological process (BP), GO:molecular function (MF) and GO:cellular component (CC). Additionally, the regulatory network was based on the STRING API [23]. Both GO analysis results and regulatory network are visualized by Echarts [38].

The sequence logo was based on WebLogo [39] and generated using the flanking sequence of modified cysteines. Each logo consists of stacks of symbols, and one stack corresponds to a

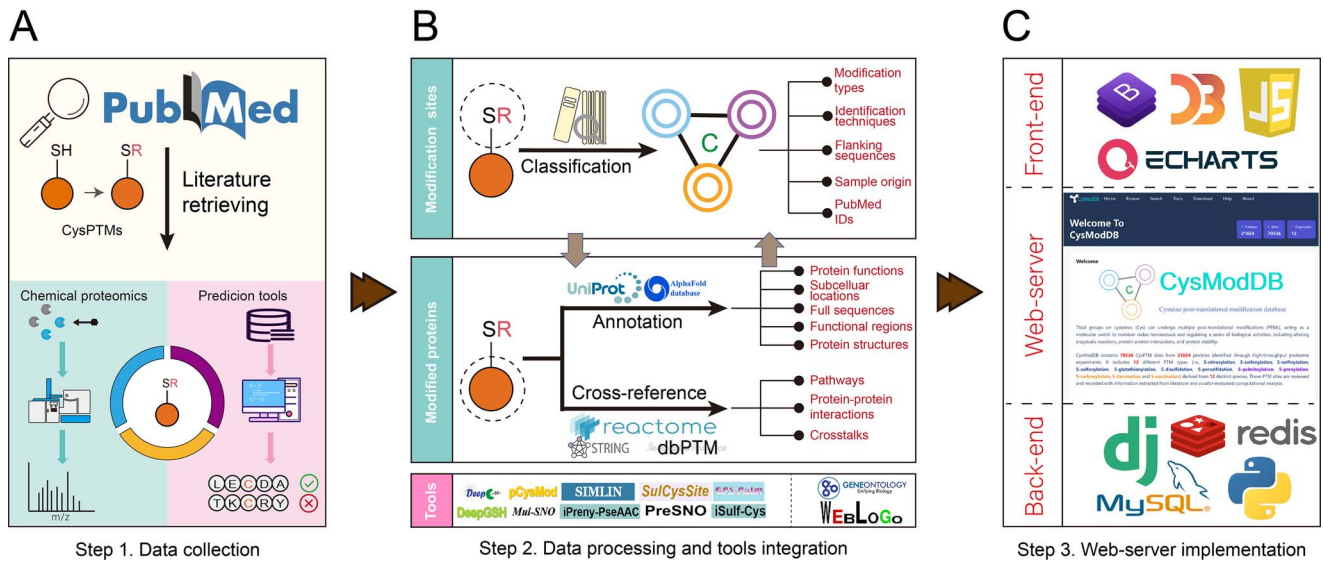


Figure 1. The construction procedure of the CysModDB. (A) Information collection. The experimentally identified CysPTMs were manually retrieved from the literature. The identification approaches and online prediction tools were collected as well. (B) The pipeline for data processing and the integration of online tools. (C) The framework of the web-server construction.

position of the sequence. The height of symbols within the stack indicates the relative frequency of each amino at that position.

The composition heatmap was built using two modules, i.e. position probability matrix (PPM) and position weight matrix (PWM), from the seqlogo python package [39]. PPM describes the probability of each amino acid on each position of the sequences. PWM illustrates the pattern of the amino acid distribution around the modified cysteines. The PPM and PWM for each Basket can be separately calculated through the following formulas.

$$M_{PPM} = \begin{pmatrix} P_{1,1} & P_{1,2} & \cdots & P_{1,n} \\ P_{2,1} & P_{2,2} & \cdots & P_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ P_{m,1} & P_{m,2} & \cdots & P_{m,n} \end{pmatrix},$$

$$M_{PWM} = \log_2 \left(\frac{M_{PPM}}{b_m} \right),$$

where $P_{m,n}$ is the probability of the amino acid m at the position n of the sequences. PWM is the PPM converted into log-likelihood, where b_m is the probability of amino acid m in the proteome. In this module, m is up to 20, and the range of n value is from -15 to $+15$.

Web-server implementation

Figure 1C shows various web applications for developing the front and back ends. The form layout of the front end was arranged using HTML5 and Bootstrap 5. Data and statistical results were shown using two JavaScript packages: D3 and Echarts. Notably, an interactive graphic browser was developed to visualize the PTM sites on proteins. Protein and PTM annotations were outputted using ProtVista and Echarts [38, 40]. Protein tertiary structures with CysPTM annotations were visualized by the interactive viewer 3Dmol.js [41]. On the back end, input data were processed using the Python-based framework Django. All the data were stored and organized by MySQL and Redis (Figure 1C). AJAX was used for the communication between the front-end and the back-end. JQuery was applied to improve interactive development and browser compatibility. Form validation and CSRF validation were added to prevent potential security risks.

Results

Figure 1 shows the construction procedure of CysModDB, which includes three steps. The first step is information collection. The experimentally identified CysPTM sites and identification approaches were retrieved from the literature, as well as online computational programs for predicting CysPTM sites. The second step includes data processing and the integration of online tools. The third step is the web-server construction for storing, showing and visualizing the collected information. The details of each step can be found in 'Materials and Methods'. In the following, we present the features and functions of the database.

Data summary and statistics

CysModDB contains 70 536 experimentally identified CysPTM sites on 21 654 proteins across 12 organisms. These sites are annotated with 12 modification types and classified into three PTM categories according to the modification characteristics: seven in the oxidation PTM category, two in the lipid PTM category and three in the metabolite PTM category (Figure 2). Among the three categories, oxidation includes the largest number of PTM sites and proteins (60 702 sites and 25 317 proteins), perhaps because oxidation is a common biological reaction throughout the lifespan (Figure 3A) [42]. Interestingly, the number of s-sulfonylation sites in the oxidation category is minimal (89 sites from 61 proteins), probably because it is at the highest oxidation state. Oxidized cysteines *in vivo* can be identified using advanced chemical proteomics techniques based on the thiol blocking strategy, like the CPT tags [11]. It should be noted that some tags (e.g. CPT) can identify but not distinguish multiple oxidation types, and the modified sites identified using these tags were grouped and annotated as 'not elsewhere classified' (nec.) (Figure 3A). Additionally, CysModDB includes three metabolite PTM types (i.e. s-itaconation [19], s-succination [20] and s-carbonylation [43]), which are excluded in previous databases [18].

We investigated the distribution of different modified sites clustered into different PTM types and organisms (Figure 3B). The majority of CysPTM sites are from the human and mouse

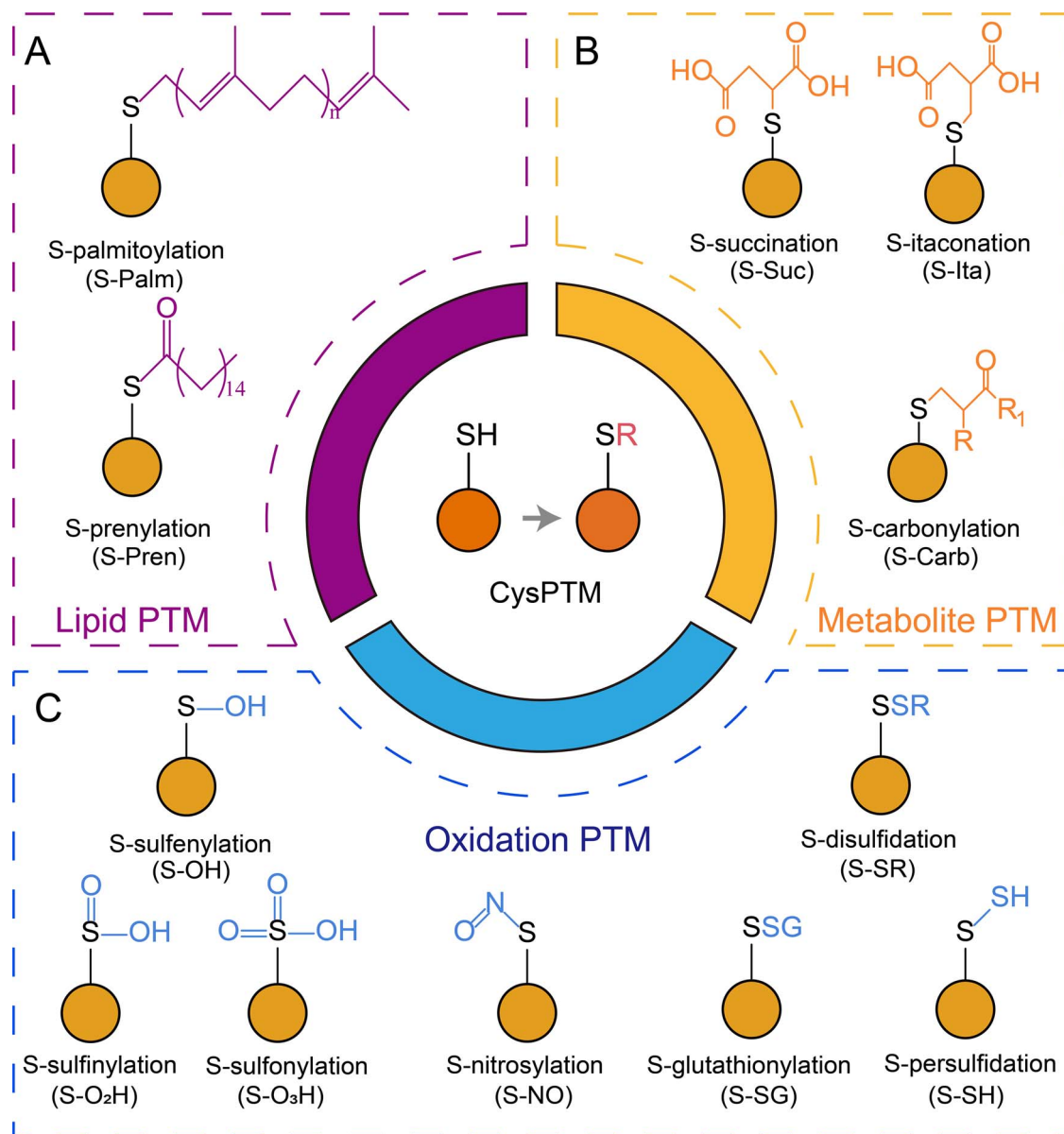


Figure 2. Classification of 12 CysPTM types into three categories. (A) Two types in the lipid PTM category, (B) three types in the metabolite PTM category and (C) seven types in the oxidation PTM category.

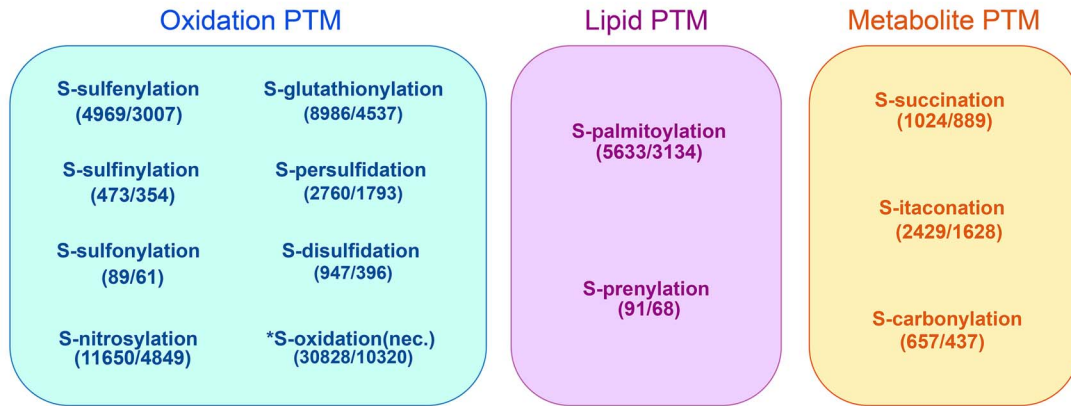
species (human/mouse: 19 072/40 403 sites), covering all three categories. Additionally, s-glutathionylation was widely investigated across six organisms, whereas s-prenylation and s-carbonylation were explored in a single organism and require further analysis. Besides, we collected 32 approaches developed to identify CysPTM sites and 10 online computational programs for predicting the CysPTM sites (Figure 3C). All the prediction tools focus on oxidation and lipid PTM categories, probably due to numerous related CysPTM sites identified. In contrast, no predictor has been developed for the recently reported metabolite PTM types. Supplementary Table S2 lists the detailed information of the 10 prediction models, including data size, feature encodings and algorithms. We compared the models and found that the early ones were based on traditional machine learning algorithms (e.g. support vector machine and random forest), while the late ones relied on advanced algorithms (e.g. XGBoost and deep neural network). For example, four models for predicting s-sulfenylation sites have been developed, i.e. iSulf-Cys [28], SulCys-Site [32], SIMLIN [31] and DeepCSO [24] (Supplementary Table S2).

The latest model DeepCSO, superior to others, was constructed using deep-learning-based long short-term memory instead of traditional machine learning algorithms used in the other models (Supplementary Table S2).

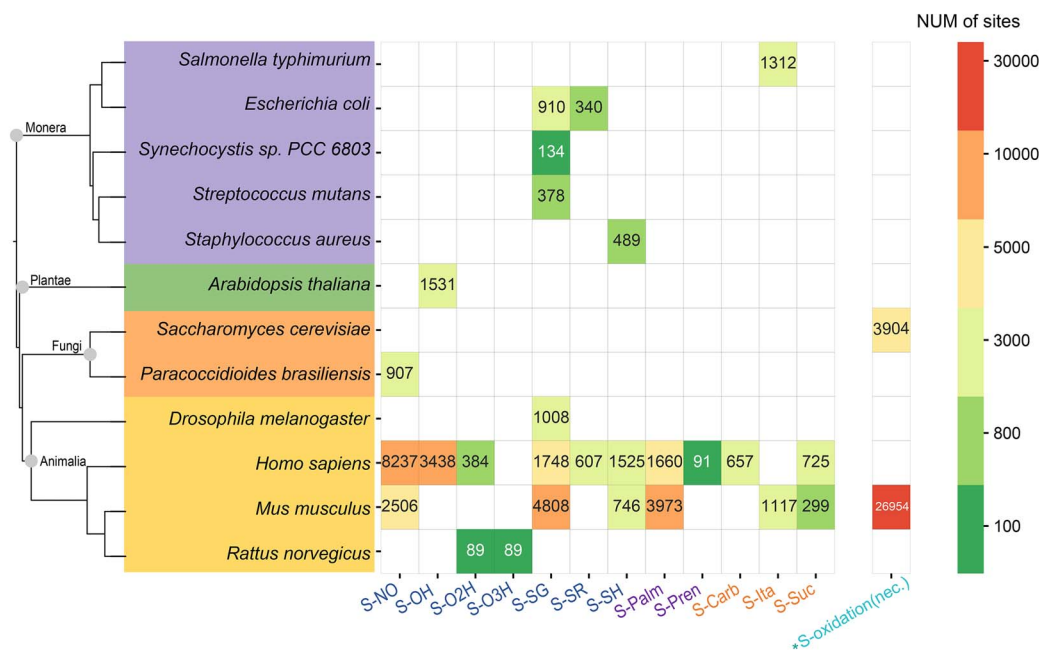
Data query and result display

The 'Home' page briefly introduces CysModDB and data statistics (Figure 4A). The 'Statistics' diagram is an interactive Sankey diagram using Echarts [38] and shows the number of CysPTM sites annotated with different PTM types from distinct species. The diagram links the CysPTM sites to the organisms and the CysPTM categories through weighted lines, where the line thickness is proportional to the number of CysPTM sites. Additionally, the data are accessible via either the 'Browse' page or the 'Search' page (Figure 4B). The 'Browse' page contains two parts: 'Browse by PTM type' and 'Browse by organisms', where each item is clickable to query the results (Figure 4B). There are three options on the 'Search' page (i.e. simple, advanced and multiple searches), where gene name, protein name, UniProt AC and organism name are

A



B



C

PTM type	S-NO	S-OH	S-O ₂ H	S-O ₃ H	S-SG	S-SR	S-SH	S-Palm	S-Pren	S-Carb	S-Ita	S-Suc
Identification approaches	3	6	2	2	4	1	3	4	2	2	2	1
Prediction tools	2	5	1	0	1	0	1	1	1	0	0	0

Figure 3. The data statistics of CysModDB. (A) The statistics of the CysPTM types in the three categories. The numbers of modified sites and proteins are shown for each type. s-oxidation (nec.) means oxidized cysteine not elsewhere classified. (B) The distribution of different modified sites clustered into different PTM types and organisms. (C) The summary of experimental identification techniques and the available online tools for predicting CysPTM sites.

supported as queries (Figure 4C). We exemplified the usage via 'glyceraldehyde-3-phosphate dehydrogenase' (GAPDH), an essential enzyme in glycolytic metabolism. The result page showed the table containing Gene Name, Protein Name, PTM categories, Organism, UniProt AC and CMID (i.e. CysModDB ID) (Figure 4D). Notably, GAPDH has a few isoforms with different UniProt AC.

The detailed information on mouse GAPDH (UniProt AC: P16858) can be visited from the 'Detailed page' via the hyperlink of its CMID. This page shows the annotations on this protein and its

CysPTM sites (Figure 4E). The top part contains the summary of the protein and the identified CysPTM types so that the users can take a quick overview. Below the summary, an interactive graphic browser shows the protein sequence with the annotations of the modified cysteines and PTM types, protein functional regions and secondary structures. The Zoom bar can be adjusted to focus on CysPTM site(s) in a specific region, enabling the investigation of the potential functions. For instance, four modifications of cysteine at the 150th position (i.e. C150) are shown in the second

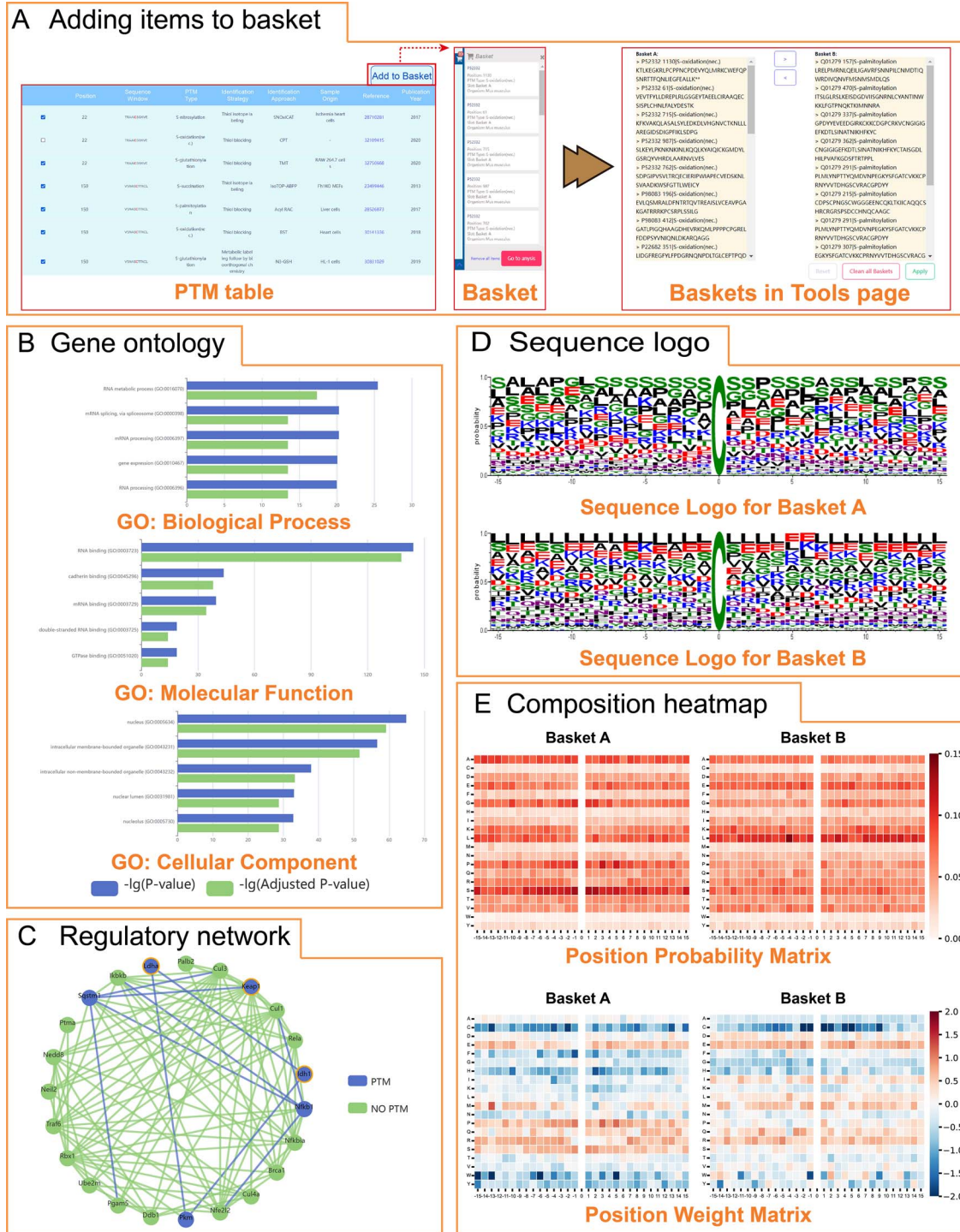


Figure 5. The detailed analysis procedure in CysModDB. (A) The 'Basket' was designed to add items of interest for later analysis. (B) The enrichment analysis results for GO (FDR-adjusted P -value < 0.01). (C) Regulatory network result for some s-itaconation proteins (the input proteins were marked as orange circle). (D) Sequence logos for the s-itaconation sites (1117 sites; Basket A) and the rest modified sites (2181 sites; Basket B) in the mouse proteins. (E) Heatmaps of position probability matrix and position weight matrix for the same data in (D).

FDR-adjusted P -value = 3.0×10^{-8} , RNA binding (GO:0003723; FDR-adjusted P -value = 1.3×10^{-60}) and nucleus (GO:0005634; FDR-adjusted P -value = 2.0×10^{-26}), indicating s-itaconation may play a role in regulating transcription. Indeed, s-itaconation can suppress inflammation by metabolism regulation and anti-inflammatory signal pathway [19]. Here we selected three s-itaconated proteins to generate a small regulatory network. The proteins included LDHA and IDH1, related to energy metabolism,

and Keap1, an anti-inflammatory transcription factor. Figure 5C shows that all three proteins interact with a series of other proteins, suggesting that s-itaconation may regulate macrophage activity via multiple pathways.

The occurrence of a PTM site is usually affected by the residues around the modified site. In other words, certain amino acid types may be preferred as flanking residues for a specific PTM type. The preference for flanking residues can be identified using

Table 1. A comparison between CysModDB and other available related databases

	CysModDB	iCysMod	SwissPalm	dbSNO
Modification types	12 types: S-NO, S-OH, S-O ₂ H, S-O ₃ H, S-SG, S-SH, S-SR, S-Palm, S-Pren, S-Carb, S-Ita, S-Suc, S-oxidation(nec.)	8 types: S-NO, S-OH, S-O ₂ H, S-SG, S-SH, S-SR, S-Palm, S-oxidation(nec.)	1 type: S-Palm	1 type: S-NO
Data main source	Literature retrieving	Database collection	Literature retrieving	Literature retrieving
Number of PTM sites	70 536	85 747	7459	4165
Protein information				
Protein functions	✓	✓	✓	✓
Full sequences	Provided with PTM positions highlighted	✓	✓	✓
Subcellular localization	✓	–	✓	✓
Secondary structure	Experimentally determined	Prediction by NetSurfP-2.0	–	Experimentally determined
Tertiary structure	Extracted from AlphaFold DB	–	–	Extracted from PDB
Functional regions	Extracted from Uniprot	–	–	Extracted from InterPro
Pathways	Link to Reactome	–	–	Link to KEGG
Protein–protein interactions	Link to STRING	–	–	–
PTM site information				
PTM types	✓	✓	✓	✓
PubMed IDs	✓	✓	✓	✓
Flanking sequences	✓	✓	–	✓
PTM cross-talks	Link to dbPTM	–	–	Link to dbPTM
Sample origin	✓	–	✓	–
PTM detection techniques	✓	–	✓	–
Visualization	Interactive	Interactive	–	Static
Online analysis tools	GO analysis, regulatory network, sequence logo and composition heatmaps	–	–	SNO-containing protein regulatory network
External prediction tools	10	–	–	–
Data acquisition	Direct download	By request	Direct download	Direct download
File formats for download	tsv, xml, fasta and json	tsv	tsv and json	tsv

two tools: sequence logo and composition heatmap. For example, we compared the sequence characteristics of s-itaconation and other CysPTM types on the s-itaconation-containing proteins. We saved the s-itaconation sites (1117 sites) in Basket A and put the rest (2181 sites) into Basket B. Figure 5D shows the sequence logos for the Baskets A and B data, respectively. The residue serine (S) is preferable in the s-itaconation logo, whereas leucine (L) is preferred in another logo, suggesting the enrichment of S and depletion of L around the s-itaconation sites. We further compared them using the position heatmap tool based on the PPM and PWM, which shows the pattern of the amino acid distribution around the modified cysteine [49]. The S enrichment and L depletion around the s-itaconation sites compared with other PTM types were evident on both heatmaps (Figure 5E). The PWM heatmap shows that S is enriched in the s-itaconation sequence compared with its occurrence frequency of the mouse proteome. All the results are downloadable. Overall, CysModDB integrates a series of tools for data analysis.

Other modules and functionalities

The ‘Help’ page provides a detailed user guide for all sections in CysModDB via pictures and animations. Moreover, the modification data are freely downloadable on the ‘Download’ page with four available file formats (i.e. tsv, fasta, xml and json).

Additionally, the ‘About’ page provides the contact information of the platform developers.

Discussion and conclusions

With the identification of numerous CysPTM types and their significant roles in life activities, it is necessary to establish a comprehensive platform integrated with CysPTM data resources and online analysis tools for the community. CysModDB is such a platform, containing 70 536 CysPTM sites on 21 654 proteins across 12 organisms and covering 12 PTM types. Extensive information from external databases and literature is included to annotate the CysPTM sites and related proteins. CysModDB includes a customized graphic browser to visualize the distribution of modified sites with different CysPTM types on protein sequences and mapping these sites to protein structures and functional regions, which assists in exploring cross-talks between the modified sites and the potential influence of the CysPTM sites on protein functions. Online analysis tools are integrated, including gene enrichment, regulatory network, investigation of sequence features and online computational classifiers for predicting CysPTM sites.

Compared with the reported CysPTM databases (Table 1), CysModDB contains more PTM types, richer annotations and information visualization and more data formats for downloading. Specifically, it includes experimental identification

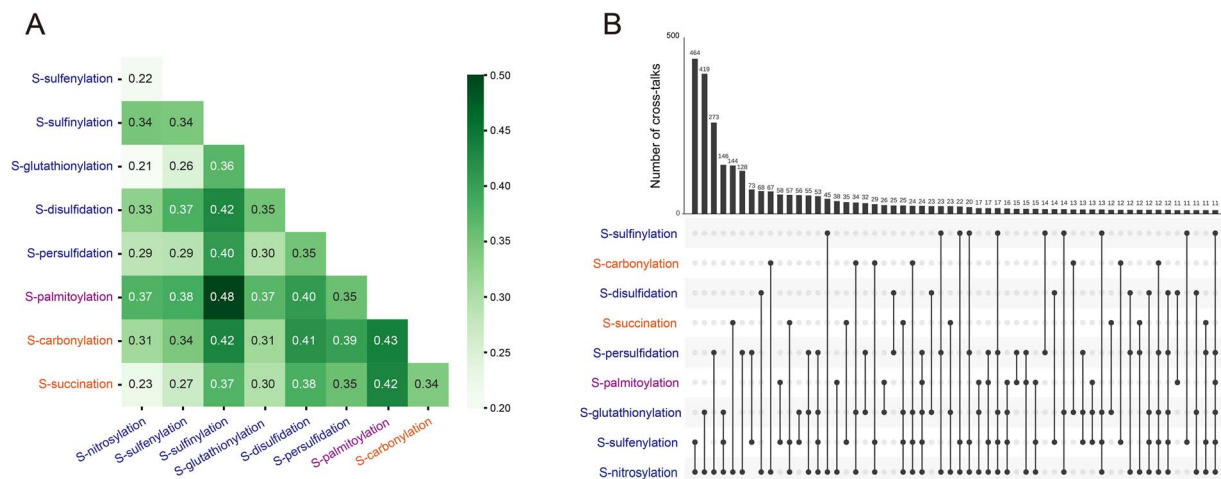


Figure 6. CysPTM features and cross-talks in the human proteome. **(A)** The sequence similarity of different CysPTM types. Each cell contains the Euclidean distance value between the PPM matrices of every two CysPTM types. **(B)** The cross-talks among the CysPTM types identified in the human proteome (cross-talk number < 10 is not shown), visualized using EVenn [56].

techniques and sample origin information, showing the PTM heterogeneity in different cell types and the target preference of different identification techniques. Moreover, a few analysis tools are included to assist in analyzing the modified sites or types, such as comparing the patterns between different PTM types. We compared the number of CysPTM sites between CysModDB and iCysMod [18] and found that the latter contained more than 10 000 sites. Most of these 10 000 PTM sites were annotated in iCysMod with human oxidation but derived from the mouse species *in vivo* [11] (private communications with Dr Zexian Liu, the senior author of the iCysMod article). Besides, some cysteine modification data collected by other databases are excluded in CysModDB due to annotation inaccuracy caused by technical limitations [21]. For instance, sodium arsenate was applied to detect s-sulfenylation, but it was later found to identify disulfides as well, and therefore, the modified sites recognized by sodium arsenate cannot be annotated with s-sulfenylation only [1, 50].

As most of the CysPTM types in CysModDB are identified from the human species, we examined sequence preferences around the modification sites of different CysPTM types in humans. The results are similar to the previous study in the iCysMod database [18] (data not shown). Furthermore, we investigated the differences between the CysPTM types by calculating the Euclidean distance between the PPMs of every two CysPTM types. A short distance indicates similar sequence features. Figure 6A shows that the distance between s-nitrosylation and s-glutathionylation is the shortest, followed by that between s-nitrosylation and s-sulfenylation, suggesting that s-nitrosylation has similar features to s-glutathionylation and s-sulfenylation. Indeed, the numbers of cross-talks between s-nitrosylation and these two oxidation subtypes are noteworthy compared with the cross-talks between any other two types (Figure 6B), which is consistent with the previous study [18, 51]. Interestingly, the distance between s-nitrosylation and s-succination is relatively short, suggesting both share certain sequence features (Figure 6A). Figure 6A also shows that s-palmitoylation is far from any other CysPTM type, probably because s-palmitoylation requires catalysis by enzymes with unique features, whereas the rest modification types do not [2].

Different CysPTM types can competitively co-occupy at the identical position. Such co-occurrences include pairwise cross-talks between two different CysPTM types at the same position and multiple cross-talks with more than two CysPTM

types. We investigated the cross-talks from the human CysPTM sites in CysModDB (Figure 6B; Supplementary Table S3). There are 1987 sites involved in pairwise cross-talks and 1262 sites involved in multiple cross-talks (Supplementary Table S3). s-nitrosylation contributed to the most PTM sites and formed the most cross-talks to other oxidation types (e.g. 464 to s-sulfenylation, 419 to s-glutathionylation, 273 to s-persulfidation), which might match the fact that s-nitrosylation is the initial state of many oxidation types [52]. Additionally, as the three CysPTM types (i.e. s-glutathionylation: 1748, s-palmitoylation: 1525 and s-persulfidation: 1660) have a similar number of PTM sites, we investigated their pairwise cross-talks. The number of cross-talks between s-glutathionylation and s-persulfidation was 353, double than those for the other two pairs (196 and 165). This observation suggests the high similarity of both oxidation types and the difference between them and s-palmitoylation.

A few shortcomings of CysModDB still need to be addressed in the future. First, the identification method lacks detailed description and may be hard to understand. It is better to visualize them with graphics and provide the chemical structure of probes. Second, the online analysis tools require manual operations, which is inconvenient for large data analysis. We will develop an API to analyze the data in batches or enable users to upload data. Third, the online analysis tools are limited, and we will integrate other powerful tools such as network analysis. In addition, we will collect more CysPTM data with richer annotations. As some CysPTM types (e.g. s-itaconation [19] and s-succination [20]) still lack computational prediction tools, we will develop related predictors based on advanced algorithms such as ensemble learning, multi-feature fusion and deep learning [53–55]. Overall, CysModDB is a comprehensive online platform integrating various data and tools to provide an almost one-stop solution for investigating CysPTM, and we anticipate that it is helpful for both experimental and computational biologists.

Key Points

- CysModDB is a comprehensive platform including 70 536 cysteine PTM sites with 12 different types, covering the

largest number of PTM types compared with previous databases.

- CysModDB comprises several parts: PTM site and protein annotations, cross-reference to external resources, online computational tools for predicting CysPTM sites and integrated analysis tools. They are integrated by a customized graphic browser and a 'Basket'.
- CysModDB is user-friendly, in which the information is easily accessible, and data are downloadable in various file formats.

Supplementary Data

Supplementary data are available online at <https://academic.oup.com/bib>.

Funding

National Natural Science Foundation of China (grant 31770821, 32071430); Shandong Training Program of Innovation and Entrepreneurship for Undergraduates (grant S202011065118).

References

- Alcock LJ, Perkins MV, Chalker JM. Chemical methods for mapping cysteine oxidation. *Chem Soc Rev* 2018;**47**:231–68.
- Linder ME, Deschenes RJ. Palmitoylation: policing protein stability and traffic. *Nat Rev Mol Cell Biol* 2007;**8**:74–84.
- Xu N, Shen N, Wang X, et al. Protein prenylation and human diseases: a balance of protein farnesylation and geranylgeranylation. *Sci China Life Sci* 2015;**58**:328–35.
- Diskin C, Ryan TAJ, O'Neill LAJ. Modification of proteins by metabolites in immunity. *Immunity* 2021;**54**:19–31.
- Peng T, Thion E, Hang HC. Proteomic analysis of fatty-acylated proteins. *Curr Opin Chem Biol* 2016;**30**:77–86.
- Qin W, Yang F, Wang C. Chemoproteomic profiling of protein-metabolite interactions. *Curr Opin Chem Biol* 2020;**54**:28–36.
- Bak DW, Bechtel TJ, Falco JA, et al. Cysteine reactivity across the subcellular universe. *Curr Opin Chem Biol* 2019;**48**:96–105.
- Jeong A, Suazo KF, Wood WG, et al. Isoprenoids and protein prenylation: implications in the pathogenesis and therapeutic intervention of Alzheimer's disease. *Crit Rev Biochem Mol Biol* 2018;**53**:279–310.
- Weiss JM, Davies LC, Karwan M, et al. Itaconic acid mediates crosstalk between macrophage metabolism and peritoneal tumors. *J Clin Invest* 2018;**128**:3794–805.
- Couvertier SM, Zhou Y, Weerapana E. Chemical-proteomic strategies to investigate cysteine posttranslational modifications. *Biochim Biophys Acta* 2014;**1844**:2315–30.
- Xiao H, Jedrychowski MP, Schweppe DK, et al. A quantitative tissue-specific landscape of protein redox regulation during aging. *Cell* 2020;**180**:968, e924–983.e24.
- Yang J, Gupta V, Carroll KS, et al. Site-specific mapping and quantification of protein S-sulphenylation in cells. *Nat Commun* 2014;**5**:4776.
- Sun MA, Wang Y, Cheng H, et al. RedoxDB—a curated database for experimentally verified protein oxidative modification. *Bioinformatics* 2012;**28**:2551–2.
- Li Z, Li S, Luo M, et al. dbPTM in 2022: an updated database for exploring regulatory networks and functional associations of protein post-translational modifications. *Nucleic Acids Res* 2022;**50**:D471–9.
- Chen YJ, Lu CT, Su MG, et al. dbSNO 2.0: a resource for exploring structural environment, functional and disease association and regulatory network of protein S-nitrosylation. *Nucleic Acids Res* 2015;**43**:D503–11.
- Chen YJ, Lu CT, Lee TY, et al. dbGSH: a database of S-glutathionylation. *Bioinformatics* 2014;**30**:2386–8.
- Blanc M, David F, Abrami L, et al. SwissPalm: protein palmitoylation database. *F1000Res* 2015;**4**:261.
- Wang P, Zhang Q, Li S, et al. iCysMod: an integrative database for protein cysteine modifications in eukaryotes. *Brief Bioinform* 2021;**22**. <https://doi.org/10.1093/bib/bbaa1400>.
- O'Neill LAJ, Artyomov MN. Itaconate: the poster child of metabolic reprogramming in macrophage function. *Nat Rev Immunol* 2019;**19**:273–81.
- Merkley ED, Metz TO, Smith RD, et al. The succinated proteome. *Mass Spectrom Rev* 2014;**33**:98–109.
- Qu Z, Greenlief CM, Gu Z. Quantitative proteomic approaches for analysis of protein S-nitrosylation. *J Proteome Res* 2016;**15**:1–14.
- Jassal B, Matthews L, Viteri G, et al. The reactome pathway knowledgebase. *Nucleic Acids Res* 2020;**48**:D498–503.
- Szklarczyk D, Gable AL, Nastou KC, et al. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res* 2021;**49**:D605–12.
- Lyu X, Li S, Jiang C, et al. DeepCSO: a deep-learning network approach to predicting cysteine S-sulphenylation sites. *Front Cell Dev Biol* 2020;**8**:594587.
- Li S, Yu K, Wang D, et al. Deep learning based prediction of species-specific protein S-glutathionylation sites. *Biochim Biophys Acta Proteins Proteomics* 2020;**1868**:140422.
- Ning W, Jiang P, Guo Y, et al. GPS-Palm: a deep learning-based graphic presentation system for the prediction of S-palmitoylation sites in proteins. *Brief Bioinform* 2021;**22**:1836–47.
- Xu Y, Wang Z, Li C, et al. iPreny-PseAAC: identify C-terminal cysteine prenylation sites in proteins by incorporating two tiers of sequence couplings into PseAAC. *Med Chem* 2017;**13**:544–51.
- Xu Y, Ding J, Wu LY. iSulf-Cys: prediction of S-sulphenylation sites in proteins with physicochemical properties of amino acids. *PLoS One* 2016;**11**:e0154237.
- Zhao Q, Ma J, Wang Y, et al. Mul-SNO: a novel prediction tool for S-nitrosylation sites based on deep learning methods. *IEEE J Biomed Health Inform* 2021;**26**:2379–87.
- Hasan MM, Manavalan B, Khatun MS, et al. Prediction of S-nitrosylation sites by integrating support vector machines and random forest. *Mol Omics* 2019;**15**:451–8.
- Wang X, Li C, Li F, et al. SIMLIN: a bioinformatics tool for prediction of S-sulphenylation in the human proteome based on multi-stage ensemble-learning models. *BMC Bioinformatics* 2019;**20**:602.
- Hasan MM, Guo D, Kurata H. Computational identification of protein S-sulphenylation sites by incorporating the multiple sequence features information. *Mol Biosyst* 2017;**13**:2545–50.
- Li S, Yu K, Wu G, et al. pCysMod: prediction of multiple cysteine modifications based on deep learning framework. *Front Cell Dev Biol* 2021;**9**:617366.
- UniProt C. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* 2021;**49**:D480–9.
- Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583–9.
- Varadi M, Anyango S, Deshpande M, et al. AlphaFold protein structure database: massively expanding the structural

- coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res* 2022;**50**:D439–44.
37. Kuleshov MV, Jones MR, Rouillard AD, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* 2016;**44**:W90–7.
 38. Li D, Mei H, Shen Y, et al. ECharts: a declarative framework for rapid construction of web-based visualization. *Vis Informatics* 2018;**2**:136–46.
 39. Crooks GE, Hon G, Chandonia JM, et al. WebLogo: a sequence logo generator. *Genome Res* 2004;**14**:1188–90.
 40. Watkins X, Garcia LJ, Pundir S, et al. ProtVista: visualization of protein sequence annotations. *Bioinformatics* 2017;**33**:2040–1.
 41. Rego N, Koes D. 3Dmol.js: molecular visualization with WebGL. *Bioinformatics* 2015;**31**:1322–4.
 42. Held JM. Redox systems biology: harnessing the sentinels of the cysteine redoxome. *Antioxid Redox Signal* 2020;**32**:659–76.
 43. Curtis JM, Hahn WS, Long EK, et al. Protein carbonylation and metabolic control systems. *Trends Endocrinol Metab* 2012;**23**:399–406.
 44. Zhang Y, Qin W, Liu D, et al. Chemoproteomic profiling of itaconations in *Salmonella*. *Chem Sci* 2021;**12**:6059–63.
 45. Chouchani ET, James AM, Methner C, et al. Identification and quantification of protein S-nitrosation by nitrite in the mouse heart during ischemia. *J Biol Chem* 2017;**292**:14486–95.
 46. Duan J, Zhang T, Gaffrey MJ, et al. Stoichiometric quantification of the thiol redox proteome of macrophages reveals subcellular compartmentalization and susceptibility to oxidative perturbations. *Redox Biol* 2020;**36**:101649.
 47. Klopfenstein DV, Zhang L, Pedersen BS, et al. GOATOOLS: a Python library for gene ontology analyses. *Sci Rep* 2018;**8**:10872.
 48. Wang H, Wang Z, Li Z, et al. Incorporating deep learning with word embedding to identify plant ubiquitylation sites. *Front Cell Dev Biol* 2020;**8**:572195.
 49. He Y, Shen Z, Zhang Q, et al. A survey on deep learning in DNA/RNA motif mining. *Brief Bioinform* 2021;**22**. <https://doi.org/10.1093/bib/bbaa1229>.
 50. Tyther R, Ahmeda A, Johns E, et al. Proteomic profiling of perturbed protein sulfenation in renal medulla of the spontaneously hypertensive rat. *J Proteome Res* 2010;**9**:2678–87.
 51. Martinez-Ruiz A, Lamas S. Signalling by NO-induced protein S-nitrosylation and S-glutathionylation: convergences and divergences. *Cardiovasc Res* 2007;**75**:220–8.
 52. Gorelenkova Miller O, Mielej JJ. Sulfhydryl-mediated redox signaling in inflammation: role in neurodegenerative diseases. *Arch Toxicol* 2015;**89**:1439–67.
 53. Bao W, Yuan CA, Zhang Y, et al. Mutli-features prediction of protein translational modification sites. *IEEE/ACM Trans Comput Biol Bioinform* 2018;**15**:1453–60.
 54. Bao W, Yang B, Chen B. 2-hydr_Ensemble: lysine 2-hydroxyisobutyrylation identification with ensemble method. *Chemom Intel Lab Syst* 2021;**215**:104351.
 55. Chen Z, Liu X, Li F, et al. Large-scale comparative assessment of computational predictors for lysine post-translational modification sites. *Brief Bioinform* 2019;**20**:2267–90.
 56. Chen T, Zhang H, Liu Y, et al. EVenN: easy to create repeatable and editable Venn diagrams and Venn networks online. *J Genet Genomics* 2021;**48**:863–6.