

# PRESS: PROTEIN S-SULFENYLATION SERVER

Marianna Sakka, Grigorios Tzortzis, Michalis D. Mantzaris, Nick Bekas, Tahsin F. Kellici,  
Aristidis Likas, Dimitrios Galaris, Ioannis P. Gerothanassis and Andreas G. Tzakos

## SUPPLEMENTARY INFORMATION

### TABLE OF CONTENTS

<b>Section 1. Protein Datasets .....</b>	<b>2</b>
<i>Homology modeling. ....</i>	<i>3</i>
<i>Estimation of evolutionary conserved amino acid positions in proteins using ConSurf Server. ...</i>	<i>3</i>
<i>Identification of conserved redox sensitive cysteines in different species using multiple     sequence alignment via the Clustal Omega server. ....</i>	<i>4</i>
<b>Section 2. Tools developed for screening the cysteine environment.....</b>	<b>5</b>
<b>Section 3. Development of the SVM method .....</b>	<b>6</b>
<i>Evaluation Metrics for the SVM method .....</i>	<i>6</i>
<i>Training and Validation of the SVM method .....</i>	<i>7</i>
<b>Section 4. Statistics .....</b>	<b>10</b>
<b>Section 5. How PRESS works.....</b>	<b>11</b>
<b>Supplementary Tables .....</b>	<b>12</b>
<b>Supplementary Figures.....</b>	<b>18</b>
<b>References .....</b>	<b>30</b>

## Section 1. Protein Datasets

Six protein datasets were utilized in our study, which are described below:

- 1. Training dataset.** Consists of 204 PDB files (with 405 protein chains) obtained from the PDB repository (Berman, et al., 2000). To retrieve these files we searched (using the advanced search utility of the PDB repository) for entries containing the term “CSO” (cysteine sulfenic acid). Hence, each PDB file in our dataset contained at least one CSO residue (i.e. a cysteine that has undergone S-sulfenylation). To avoid structure redundancy, results were filtered to retrieve single representatives in cases of 90% sequence identity, using the “retrieve only representatives at sequence identity” option. The CSO and CYS (cysteine thiol) residues in the files were extracted and filtered out according to the following criteria. We removed all residues annotated as “missing” in each PDB file, since no atom information is provided for such residues. Moreover, to get a cleaner collection and avoid overestimating the potential of our predictive algorithm, we removed duplicate residues. A CYS (resp. CSO) is considered to be a duplicate if another CYS (resp. CSO) in the collection has exactly the same neighbors in the protein sequence within a window of size  $W$  (here  $W=11$ ) centered at the CYS (resp. CSO). In the end, 243 CSO residues and 777 CYS residues were retained and used for SVM training. For this dataset **Supplementary Table S1** is provided, consisting of manually curated information regarding the proteins used for the generation of the SVM algorithm. The following information has been incorporated on the relevant table: PDB ID, S-sulfenylation position (CSO position), X-ray resolution, Gene ID, Protein name and synonyms, related crystal structure reference (PMID), independent references confirming experimentally protein S-sulfenylation (PMID and RedoxDB), independent methods used to validate the modified cysteine site, and the existence of other reduced cysteine residues in the crystal structure. Collectively, S-sulfenylation sites have been verified through information related to: the usage of cryogenic conditions for crystal formation, crystallization of the reduced, oxidized and disulfide forms of the protein during  $H_2O_2$  exposure, redox control of protein's biophysical or biochemical activity using mutations, alkylating reagents (N-ethylmaleimide etc.) or reducing conditions, spectrophotometric-NMR titration and rate of oxidation by hydrogen peroxide, dimedone-based probing and immunoblotting, proteomics-LC/MS.
- 2. Endogenous mouse liver dataset.** *Qualitative data* of S-sulfenylated proteins from in vivo experiments identified by Dimedone-CSO labeling, representing the basal levels of endogenous cysteine oxidation in mouse liver (Original dataset acquired by (Gould, et al., 2015) from supp. Table S3.SOH peptides). A subset of 14 PDB files of S-sulfenylated proteins comprising 19 CSO residues and 166 CYS residues were used for SVM validation. In addition, 14 randomly selected proteins from the original set devoid of a PDB file, were used for homology modeling using Phyre2 (Kelley, et al., 2015). The generated 14 modeled PDB files comprising 14 CSO residues and 196 CYS residues were also included for SVM validation.
- 3. Endogenous RKO SulfenM dataset.** *Qualitative data* of S-sulfenylated proteins identified by Dimedone based-CSO labeling (DYn-2) in RKO cells cultured under physiological conditions, in 95% air-5%CO<sub>2</sub> (i.e. 20% oxygen). This dataset represents the basal levels of endogenous cysteine oxidation in RKO cells (Original SulfenM dataset acquired by (Yang, et al., 2014) from supp. data 2). A subset of 158 PDB files of S-sulfenylated proteins comprising 259 CSO residues and 1924 CYS residues were used for SVM validation. In addition, 16 randomly selected proteins from the original set devoid of a PDB file, were used for homology modeling using Phyre2. The generated 16 modeled PDB files comprising 29 CSO residues and 65 CYS residues were also included for SVM validation.

4. **H<sub>2</sub>O<sub>2</sub> RKO SulfenQ dataset.** *Quantitative data* of S-sulfenylated proteins in RKO cells stimulated by exogenous H<sub>2</sub>O<sub>2</sub> exposure (Original H<sub>2</sub>O<sub>2</sub> RKO SulfenQ dataset acquired by (Yang, et al., 2014) from supp. data 3). S-sulfenylation was quantified by light/heavy DYn-2 labeling ratio in matched oxidized cysteines between Control untreated cells (light DYn-2) and stimulated cells (heavy DYn-2). A subset of 58 PDB files of S-sulfenylated proteins comprising 108 CSO residues and 877 CYS residues were used for SVM validation. In addition 15 randomly selected proteins from the original set devoid of a PDB file, were used for homology modeling using Phyre2. The generated 15 modeled PDB files comprising 24 CSO residues and 85 CYS residues were also included for SVM validation.
5. **EGF A431 SulfenQ dataset.** *Quantitative data* of S-sulfenylated proteins in A431 cells stimulated by EGF incubation (Original EGF A431 SulfenQ dataset acquired by (Yang, et al., 2014) from supp. data 4). S-sulfenylation was quantified by light/heavy DYn-2 labeling ratio in matched oxidized cysteines between Ctrl untreated cells (light DYn-2) and stimulated cells (heavy DYn-2). A subset of 44 PDB files of S-sulfenylated proteins comprising 96 CSO residues and 470 CYS residues were used for SVM validation. In addition 15 randomly selected proteins from the original set devoid of a PDB file, were used for homology modeling using Phyre2. The generated 15 modeled PDB files comprising 23 CSO residues and 74 CYS residues were also included for SVM validation.
6. **X-ray dataset.** S-sulfenylated proteins acquired from the PDB repository in the form of PDB files containing CSO residues. This is the latest release of protein PDB files from the Protein Data Bank that have not been included in the training dataset (they were released at a later date than the PDB files of the training set). A total of 18 PDB files, containing 31 CSO and 246 CYS residues, were utilized for SVM validation.

### ***Homology modeling.***

Homology modeling for proteins whose three dimensional structures are not available was performed using the Phyre2 server (Kelley, et al., 2015). Phyre2 server is freely available at <http://www.sbg.bio.ic.ac.uk/phyre2/>. The amino acid sequences were downloaded as FASTA files from the UniProt database (UniProt and Consortium, 2015). The intensive mode was selected for the modeling, since in some cases the CSO residue may be present in a region for which there are no templates. Ab initio modeling is required for these regions. The models with highest alignment scores were then submitted to PRESS. The PDB file produced by Phyre2 must be modified to be used by PRESS and made compatible with the PDB format v3.10. Specifically, the PDB must contain the amino acid sequence of residues in each chain of the macromolecule (SEQRES records), missing coordinates in the three dimensional structure must be defined by REMARK-465 in the PDB, a chain name must be provided for each chain and the PDB file must end with the TER symbol.

### ***Estimation of evolutionary conserved amino acid positions in proteins using ConSurf Server.***

A subset of 9 protein X-ray structures from the training dataset (**Table S1**, pdbids: 1GSN, 1I9T, 2IBY, 2NQA, 2V5B, 2VRN, 3IQU, 3UBW, 3ZJE), where cysteine S-sulfenylation has been experimentally verified, was selected and evolutionary conservation of amino acid positions in the proteins was assessed by the ConSurf server (Ashkenazy, et al., 2010; Glaser, et al., 2003; Landau, et al., 2005). The degree to which an amino acid position is evolutionarily conserved is strongly dependent on its structural and functional importance; rapidly evolving positions are variable while slowly evolving positions are conserved. Thus,

conservation analysis of positions among members from the same family can often reveal the importance of each position in the protein structure or function. Interestingly, along this analysis for the selected cases a very high conservation was revealed for the S-sulfenylated cysteine (CSO), whereas in some cases high sequence conservation was also evident for the neighboring environment in the primary sequence. However, in some cases, although the CSO residue was conserved, low sequence conservation was recorded for the neighboring residues in the primary sequence. This result clearly highlights that identification of redox sensitive cysteines cannot be solely based on the neighboring residues in the primary sequence, but additional criteria should be considered, i.e. the 3D (spatial) neighboring structural features. Therefore, we further projected the conserved residues onto the 3D structures of the studied proteins containing the CSO residue (**Figure S1**). Interestingly, this analysis indicated that in the 3D environment surrounding the CSO residue, several amino acids are highly conserved. Thus, information related to the CSO formation is encoded on the 3D space surrounding the redox sensitive cysteine.

***Identification of conserved redox sensitive cysteines in different species using multiple sequence alignment via the Clustal Omega server.***

Table S2 is provided which consists of proteins taken from Table S1 with conserved redox sensitive cysteine residues in different species. Conservation of redox sensitive cysteines in different species has been verified by multiple sequence alignment using the Clustal Omega server (Sievers, et al., 2011). Clustal Omega server is freely available at <http://www.clustal.org/omega/>. A subset of 28 proteins from the training set was identified to possess conserved oxidized cysteine residues experimentally verified in other species (Table S2-Verified CSO in species and Figures S2 and S3). A second subset of 23 proteins was identified to possess conserved redox sensitive cysteine residues in other species, which have not been experimentally verified in them (Table S2-Conserved CYS in species and Figures S4 and S5). Proteins with conserved and verified CSO in different species were manually curated by identifying redox sensitive homologous proteins in RedoxDB (Sun, et al., 2012), using the Blast tool. The database RedoxDB is available at <http://biocomputer.bio.cuhk.edu.hk/RedoxDB/>. Proteins from other species where the CSO site has been conserved but not experimentally verified in them were manually curated by identifying homologous proteins in different species using Uniprot database at <http://www.uniprot.org/>.

## Section 2. Tools developed for screening the cysteine environment

We have developed three tools, available in the PRESS server, to assist users in analyzing the *surrounding environment of cysteines* contained in proteins that interest them. We enable the users to acquire a variety of information for the desired proteins by simply providing to our tools their PDB files. Note that cysteine residues which are annotated as “missing” in a PDB file are ignored by our tools.

### Tool 1. Primary sequence amino acid neighbors

For each cysteine contained in the PDB files its *neighboring amino acid residues in the primary protein sequence* lying within a window of size  $W=11$ , centered at the cysteine, are extracted and returned to the user. As complementary information, the frequency of each amino acid in each window position, counted over all identified cysteine neighbors, is given. Moreover, a sequence logo (Schneider and Stephens, 1990) summarizing the characteristics of the CYS neighbors is provided. In the sequence logo amino acids are organized into seven categories using the following coloring scheme: Black: Aliphatic (A, G, I, L, P, V); Orange: Aromatic (F, W, Y); Red: Acidic (D, E); Blue: Basic (R, K, H); Purple: Hydroxylic (S, T); Green: Sulfur-containing (C, M); Magenta: Amidic (N, Q). Different amino acid letters at the same position (x axis) in the sequence logo are scaled according to the frequency of the corresponding amino acid. The total height of each stack of letters (y axis), depicts the entropy of each position, expressed in bits of information. The higher the entropy (i.e. the less the information content) the less the height of the stack. Sequence logos are produced using our own MatLab implementation.

### Tool 2. Spatial (3D) amino acid neighbors

For each cysteine its *spatial* (i.e. 3D) *amino acid neighbors* lying within a user defined radius of x Angstroms are returned, along with their distance. The distance between two residues is defined as the shortest distance between any pair of their atoms (atom coordinates are specified in the PDB files). As above, the frequency of each amino acid in the identified 3D cysteine neighborhoods is reported. Actually, the 3D neighborhood is divided into equally sized intervals (according to the distance from the cysteine residue) and the amino acid occurrences in each interval are reported. For example, in the case of three intervals, the intervals are  $[0 - x/3]$ ,  $[x/3 - 2x/3]$  and  $[2x/3 - x]$ . Similar to the first tool, a sequence logo summarizing the characteristics of the 3D cysteine neighborhood is also created. Each stack of letters in the sequence logo corresponds to one of the intervals.

### Tool 3. Spatial (3D) heterogen atoms

For each cysteine residue the *water molecules, the metal ions* (Li, Be, Na, Mg, Al, K, Ca, Sc, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, Ga, Rb, Sr, Y, Zr, Mo, Ru, Rh, Pd, Ag, Cd, In, Sn, Sb, Cs, Ba, La, Ce, Pr, Sm, Eu, Gd, Tb, Dy, Ho, Er, Yb, Lu, Hf, Ta, W, Re, Os, Ir, Pt, Au, Hg, Tl, Pb, Bi, Pa, U), *or even all types of heterogen atoms* lying within a user defined radius of x Angstroms are obtained from the PDB files, along with their distance (calculated as above). The occurrences of water molecules (resp. metal ions) in the cysteine neighborhood are counted as in the case of the spatial amino acid neighbors and visualized using a bar chart.

### Section 3. Development of the SVM method

We have devised a Support Vector Machine (SVM)-based (Bishop, 2006; Cortes and Vapnic, 1995) method to assist researchers in determining whether a cysteine in a protein (provided in PDB file format) is prone to S-sulfenylation. The method automatically predicts which CYS in a protein are sensitive to oxidation and CSO formation. To construct the SVM prediction model, we exploit a number of features that capture information related to the surrounding environment of CSO and CYS residues and can contribute in their effective discrimination. Specifically, for each CYS (resp. CSO) the following information is extracted:

- The neighboring amino acid residues in the protein sequence (*primary sequence neighbors*) lying within a window of size  $W=11$ , centered at the CYS (resp. CSO). Based on this, the frequency (number of occurrences) of each amino acid in the neighborhood is calculated and utilized during SVM training.
- Its *spatial* (i.e. 3D) amino acid *neighbors* lying within a radius of 5 Angstroms. The frequency of each amino acid in this spatial neighborhood is computed and utilized during SVM training. The distance between two residues is defined as the shortest distance between any pair of their atoms.
- The absolute solvent accessibility (ASA) value using the DSSP program (Kabsch and Sander, 1983).
- The secondary structure (SS) using the DSSP program (Kabsch and Sander, 1983). Secondary structure is categorized as helix, sheet and coil. Letters G, H returned by DSSP are considered as helix, letters B, E as sheet, while T and blank as coil. Letters S, I are ignored.

The predictive algorithm is available to users as a tool in the PRESS server. Experimental results analyzing the performance of the SVM-based algorithm are presented in the following sections.

#### *Evaluation Metrics for the SVM method*

Here we briefly present the metrics used to assess the performance of our approach. Results are evaluated in terms of sensitivity, specificity and accuracy, defined as:

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

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

$$\text{Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2} ,$$

where  $TP$ ,  $TN$ ,  $FP$  and  $FN$  stand for true positive, true negative, false positive and false negative, respectively. In our case CSO residues form the positive control, whereas CYS residues the negative control. Finally, the accuracy metric shown above is actually a *balanced accuracy* metric and its advantage is that it avoids inflated performance estimates on imbalanced datasets.

## ***Training and Validation of the SVM method***

To evaluate the performance of the predictive algorithm, multiple sets of experiments are conducted. SVM is implemented using the LIBSVM toolbox (Chang and Lin, 2011). For all experiments the radial basis kernel function (rbf) is adopted.

At first, we use the protein data retrieved from the PDB repository (the training dataset in Section 1) and employ 10-fold cross validation to train, as well as, assess the performance of our algorithm. Hence, the dataset is split in ten equally sized parts and nine parts are used for training, while the remaining part serves as the test set. The process is repeated ten times, so that each of the ten parts will serve once as the test set, and performance scores are computed as the average over the ten runs. Notice that this dataset is highly imbalanced (243 CSO residues and 777 CYS residues). This imbalance is taken into account during training by assigning different penalty parameters to the two classes in the SVM formulation (Chang and Lin, 2011). During training, the C regularizer of the SVM and the gamma parameter of the rbf kernel are also optimized.

Different combinations of the four feature types mentioned at the beginning of this section were tried and the results are illustrated in Table S3**Error! Not a valid bookmark self-reference.** (sequence refers to primary sequence neighbors, spatial to spatial neighbors, ASA to absolute solvent accessibility and SS to secondary structure). As can be seen, when only sequence or spatial neighbors are used, CSO and CYS residues are not effectively distinguished. A considerable boost in performance is observed when these features are combined with the ASA feature (sequence + ASA, spatial + ASA). It seems that spatial neighbors together with ASA (spatial + ASA) convey critical information for the discrimination of CSO and CYS residues. When these two features are further combined with those related to primary sequence neighbors and/or secondary structure (sequence + spatial + ASA, sequence + spatial + ASA + SS) a further improvement is possible. In a nutshell, *the proposed prediction method achieves* sensitivity of 79.9%, specificity of 73.6% and an overall accuracy of 76.8%, when all feature types are considered.

Let us note that to construct the prediction tool available at the PRESS server, we trained an SVM model by combining all feature types (i.e. sequence + spatial + ASA + SS), using the protein data of the training dataset.

In order to investigate how the performance of the predictive algorithm varies when the *maximum allowed* sequence identity filter in the dataset *is reduced* below the 90% barrier, that was used in the above experiments (see the first dataset description in Section1), we retrieved from the PDB repository three additional collections of PDB files, where the *maximum allowed* sequence identity filter was set to 30%, 50% and 70%, respectively (the retrieval was done analogously to that described in Section 1 for the first dataset). Actually, the PDB files in the three new collections are subsets of those in the original collection. In each case, if structures are present with sequence identity up to the specified limit, a single representative is retrieved in the collection. To conduct the experiments, we employed exactly the same procedure of 10-fold cross validation as before and utilized the combination of all feature types (i.e. sequence + spatial + ASA + SS).

Table S4 shows only a *slight* drop in performance as the maximum allowed sequence identity in the data is decreased. These results provide solid evidence that the four chosen features successfully capture the information required for efficient discrimination between CSO and CYS and that our method can achieve high prediction accuracy.

The rest of the datasets (datasets 2-6 in Section 1) are only used for validating our SVM method. Specifically, the SVM model obtained by training using the first dataset from Section 1 was directly

employed to get the predictions for these datasets. Results are reported in Table S5



, where a high success rate is observed. For each subset the total number of S-sulfenylated cysteine sites is presented under the column entitled #CSO. All other cysteines present in the proteins comprise the negative controls and their total number for each subset is presented under the column entitled #CYS. All the datasets presented in Table S5, apart from the X-ray dataset, include homology modeled subsets, as described in Section 1.

It has to be pointed out that the original sets of S-sulfenylated proteins reported by Yang and et al., 2014 and Gould et al., 2015, from which datasets 2-5 were derived and used for SVM validation, were experimentally verified using Dimedone and Dimedone-based probes (DYN-2) for CSO detection. Although this detection method is the optimal chemical method used today for proteomic studies, the concentration used for such sulfenic acid-probes has to be carefully optimized against cell number and type, as non-specific binding of the probe is possible upon enhanced concentrations of the probe ((Akter, et al., 2015), [Supplementary Figure S1](#); (Gould, et al., 2015), [Figure 2G](#); (Reddie, et al., 2008), [Figure 5a](#)). Moreover, steric features may hinder Dimedone-based labeling of some S-sulfenylated sites that could lead to false negative results (Yang, et al., 2014). Due to the possible non-specific binding of the sulfenic acid-probes, quantifiable changes acquired in data sets of S-sulfenylated proteins after H<sub>2</sub>O<sub>2</sub> or EGF stimulation (based on light/heavy DYN-2 ratio) represent “more accurate” datasets for validation of the predictive algorithm compared to the endogenous sets. Endogenous sets represent only qualitative data of S-sulfenylated proteins and thus it is more likely to include false-positive CSO sites due to non-specific labeling. Thus, using such information to train a CSO prediction algorithm can lead to potential false positive scores due to this inherent limitation. In support of this notion, as illustrated in Table S5, our SVM algorithm returned higher scores for CSO prediction (i.e. sensitivity) in H<sub>2</sub>O<sub>2</sub> stimulated dataset compared to the endogenous set. In addition, S-sulfenylated proteins tested from EGF-stimulated A431 cells returned lower prediction scores. The lower prediction scores acquired from EGF data set could be attributed to the different cell type and to the fact that EGF represents a milder oxidative environment, activating fewer and more selective redox regulated pathways and thus resembling more to an endogenous state. In fact, experimental results by (Yang, et al., 2014) reported significant quantifiable changes in over 90% of the detected S-sulfenylated sites in H<sub>2</sub>O<sub>2</sub>-stimulated cells whereas in EGF, less than 50% of S-sulfenylated sites had significant changes. The rest of the detected oxidized sites were unchanged. These results show that approximately 10% of the detected CSO sites in H<sub>2</sub>O<sub>2</sub> protein set and 50% of the EGF protein set are possible endogenous CSO sites (i.e. DYN-2 binding sites) that did not respond to oxidant stimulation indicating possible non-specific labeling. These results are in agreement with the predictions made by our server for these two subsets. Finally, the above findings and remarks could have extended implications in the development of relevant tools for S-sulfenylome prediction when the training of the predictive algorithms is based solely in the protein sequence of S-sulfenylated peptides verified by Dimedone-based probes. Since this information could be inherently biased due to possible false positive and negative entries, the use of structural data of verified S-sulfenylated proteins by x-ray analysis represents a better training set for a predictive algorithm.

## Section 4. Statistics

This section offers a general overview on the features that discriminate a non-redox sensitive cysteine (CYS) from a cysteine that can undergo S-sulfenylation (CSO). Figures S6-S11 and Tables S6 and S7 illustrate this discrimination based on the following information extracted from the training dataset (Section 1):

1. **Primary sequence neighbors.** Presentation of 5 amino acid residues before (position: -1, -2, -3, -4, -5) and after (position: +1, +2, +3, +4, +5) the CSO (resp. CYS) in the primary structure. Sequence logos, described in Section 2, were produced and presented in Figure S6. To further support this statistical analysis, a frequency plot of the neighboring amino acids in the primary sequence for the respective CSO/CYS residues is presented in Table S6. Statistically significant differences between the distribution of the neighboring amino acids of CSO and CYS residues in each position were evaluated by Pearson's chi-squared test of independence (a difference is considered significant when  $p\text{-value} < 0.05$ ). From the results it can be seen that only positions +2 and +4 exhibit differences in the amino acid distribution that are statistically significant. Hence, those two positions convey significant information for discriminating CSO and CYS residues.
2. **Spatial (3D) amino acid neighbors.** Presentation of the amino acid residues lying within a 20 Angstroms radius of the CSO (resp. CYS) residues, using 2 Angstroms in size intervals. Sequence logos, described in Section 2, were produced and presented in Figure S7. To further support this statistical analysis, a frequency plot of the neighboring amino acids in the surrounding 3D environment is also presented in Table S7. Statistically significant differences between the distribution of the neighboring amino acids of CSO and CYS residues in each interval were evaluated by Pearson's chi-squared test of independence (a difference is considered significant when  $p\text{-value} < 0.05$ ). From the results it can be seen that seven out of ten intervals exhibit differences in the amino acid distribution that are statistically significant. Apparently, the 3D neighborhood conveys significant information for discriminating CSO and CYS residues.
3. **Water molecules.** Presents the frequency appearance of the *water molecules* surrounding the CSO (resp. CYS) within a 20 Angstrom radius (Figure S8).
4. **Metal ions.** Presents the frequency appearance of the *metal ions* surrounding the CSO (resp. CYS) within a 20 Angstrom radius (Figure S9).
5. **Solvent accessibility.** Presents the *solvent accessibility* of the CSO (resp. CYS) calculated by the DSSP software (Kabsch and Sander, 1983). On each box, the central mark presents the median, the edges of the box present the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually. Whisker length  $w=1.5$ . Points are drawn as outliers if they are larger than  $q_3+w(q_3-q_1)$  or smaller than  $q_1-w(q_3-q_1)$ , where  $q_1$  and  $q_3$  are the 25th and 75th percentiles, respectively.  $w=1.5$  corresponds to approximately  $\pm 2.7\sigma$  and 99.3 coverage if the data are normally distributed. Outliers are plotted individually as red crosses (Figure S10).
6. **Secondary structure.** Presents the secondary structural features, calculated by the DSSP software, for the CYS (resp. CSO) (Figure S11). B = beta bridge, E = beta bulges, G =  $3_{10}$  helix, H =  $\alpha$  helix, S = regions of high curvature, T = turn, ~ = no structure (Kabsch and Sander, 1983).

## Section 5. How PRESS works

This section provides an overview of the PRESS server. The PRESS server is freely available at: <http://press-sulfenylation.cse.uoi.gr/>.

- **Upload tab**

At this tab the user can upload one or more PDB files of his/her choice, which can later be analyzed using the methods available at the “Tools” tab and the “SVM Prediction” tab. For proteins without available three dimensional structures (i.e. without a PDB file), homology modeling can be employed to generate a PDB file and upload it to the server (see the instructions in Section 1 on homology modeling). For demonstration purposes, at the bottom of the “Upload” tab, we provide a link to a zip file (containing two PDB files) which can be readily used to test the functionality of the server.

- **Tools tab**

In this tab, PRESS offers three different methods for analyzing the uploaded PDB files. The available methods include:

- a) Sequence Neighbors** – Presents the 10 amino acid residues surrounding each cysteine residue in the uploaded files in the protein sequence (see Tool 1 in Section 2).
- b) 3D Neighbors** – Presents the amino acid residues surrounding each cysteine residue in the uploaded files in 3D space, within a user-defined radius (see Tool 2 in Section 2).
- c) 3D HetW Neighbors** – Presents the metal ions, water molecules or all types of heterogeneous atoms, surrounding every cysteine residue in the uploaded files in 3D space within a user-defined radius (see Tool 3 in Section 2).

The desired method of analysis can be selected via a dropdown menu. By running the selected method, results are automatically presented. The log file of the analysis can be downloaded by clicking the “*Download log file*” option. Results can be downloaded in fasta format and are summarized in sequence logo or bar chart images, which are also available for download in different formats.

- **SVM prediction tab**

At this tab the user can get predictions regarding the cysteines contained in the uploaded PDB files. Two examples demonstrating the effectiveness of PRESS on CSO prediction are illustrated in Figure S12 (Figure S12 is a snapshot of the PRESS SVM prediction tab). The first example presents the SVM prediction on the human endothelial growth factor (EGF) receptor protein (pdbid 5czh). EGF receptor has been verified to be S-sulfenylated in the catalytic cysteine CYS 797, an event that enhances its kinase activity underlying the role of S-sulfenylation in EGFR signaling (Paulsen, et al., 2012). Mutations or overexpression of EGFR in breast and lung cancers, has motivated the development of inhibitors that covalently modify this site and are currently under evaluation in clinical trials (Paulsen, et al., 2012). Interestingly, PRESS successfully predicted that out of 6 cysteines present in the PDB file 5czh, only one cysteine is prone to S-sulfenylation (CSO) and that is CYS 797 (Figure S12-A). The second example presents the SVM prediction on human monoacylglyceride lipase (pdbid 3hju) containing 4 cysteines. PRESS predicts that both CYS 201 and 208 are prone to S-sulfenylation (Figure S12-B), which was, very recently, experimentally verified (Dotsey, et al., 2015).

## Supplementary Tables

**Table S1. Manually curated information regarding the training dataset used for the generation of the SVM algorithm.** The proteins comprising the SVM training dataset (Section 1) have been incorporated on the relevant table with the following information: PDB ID, S-sulfenylation position (CSO position), X-ray resolution, Gene ID, Protein name and synonyms, related crystal structure reference (PMID), independent references confirming experimental protein S-sulfenylation (PMID and RedoxDB), independent methods used to validate the modified cysteine site, and the existence of other reduced cysteine residues in the crystal structure. Table S1 is provided as an attached excel file.

**Table S2. Proteins from Table S1 with conserved redox sensitive cysteine residues in different species.** A subset of 28 proteins from the training dataset (Section 1) was identified to possess conserved oxidized cysteine residues experimentally verified in other species (Verified CSO in species). A second subset of 23 more proteins was identified to possess conserved redox sensitive cysteine residues in other species but have not been experimentally verified, in them (Conserved CYS in species). Proteins with conserved and verified CSO in deferent species were manually curated by identifying redox sensitive homologous proteins in RedoxDB using the Blast tool. Proteins where the redox sensitive CYS has been conserved in other species but not experimentally verified yet in them were manually curated by identifying homologous proteins in deferent species using Uniprot database. Conservation of redox sensitive cysteines has been verified by multiple sequence alignment using the Clustal Omega server. Table S2 is provided as an attached excel file.

**Table S3.** Prediction results on the training dataset (Section 1) for different feature combinations using 10-fold cross validation. Details are presented in Section 3. Sequence refers to primary sequence neighbors, spatial to spatial neighbors, ASA to absolute solvent accessibility and SS to secondary structure.

Features	Accuracy (%)	Sensitivity (%)	Specificity (%)
Sequence	58.5	56.7	60.3
Spatial	64.3	67.8	60.8
sequence + ASA	73.5	73.7	73.3
spatial + ASA	76.7	75.9	77.5
sequence + ASA + SS	73.3	72.4	74.2
spatial + ASA + SS	76.7	75.1	78.3
sequence + spatial + ASA	<b>76.8</b>	79.8	73.7
sequence + spatial + ASA + SS	<b>76.8</b>	79.9	73.6

**Table S4.** Prediction results on subsets of the original training dataset (Section 1) with different levels of maximum allowed sequence identity, using 10-fold cross validation. Details are presented in Section 3.

<b>Dataset</b>	<b>Accuracy (%)</b>	<b>Sensitivity (%)</b>	<b>Specificity (%)</b>
90%	76.8	79.9	73.6
70%	76.7	80.8	72.5
50%	74.9	78.6	71.2
30%	73.8	75.3	72.4

**Table S5.** Prediction results on various protein datasets (Section 1), used to validate the developed SVM algorithm. Details are presented in Section 3.

DATA SETS	# of PDBs	Sensitivity	Specificity	Accuracy	#CSO	#CYS
<b>H<sub>2</sub>O<sub>2</sub></b>	<b>73</b>	<b>73.5 %</b>	<b>66.6 %</b>	<b>70.1 %</b>	<b>132</b>	<b>962</b>
<b>EGF DATA SET</b>	<b>59</b>	<b>63.7 %</b>	<b>67.1 %</b>	<b>65.5 %</b>	<b>119</b>	<b>544</b>
<b>ENDOGENOUS</b>	<b>202</b>	<b>60.4 %</b>	<b>68.9 %</b>	<b>64.7 %</b>	<b>321</b>	<b>2351</b>
ENDOGENOUS SULFENM	174	59.7 %	71.1 %	65.4 %	288	1989
ENDOGENOUS MOUSE LIVER	28	66.7 %	56.6 %	61.6 %	33	362
<b>X-ray</b>	<b>18</b>	<b>83.9 %</b>	<b>82.9 %</b>	<b>83.4 %</b>	<b>31</b>	<b>246</b>

**Table S6.** Frequency plot of amino acid neighbors in the primary protein sequence of cysteines. Results present the frequency of appearance (%) of each amino acid in a window position of five residues before (-5,-4,-3,-2,-1) and after (+1,+2,+3,+4,+5) the CSO/CYS residues. Aminoacid distributions for each position were analyzed by Pearson's chi-squared test of independence and statistically significant differences between CSO and CYS neighbors are presented for \*p<0.05.

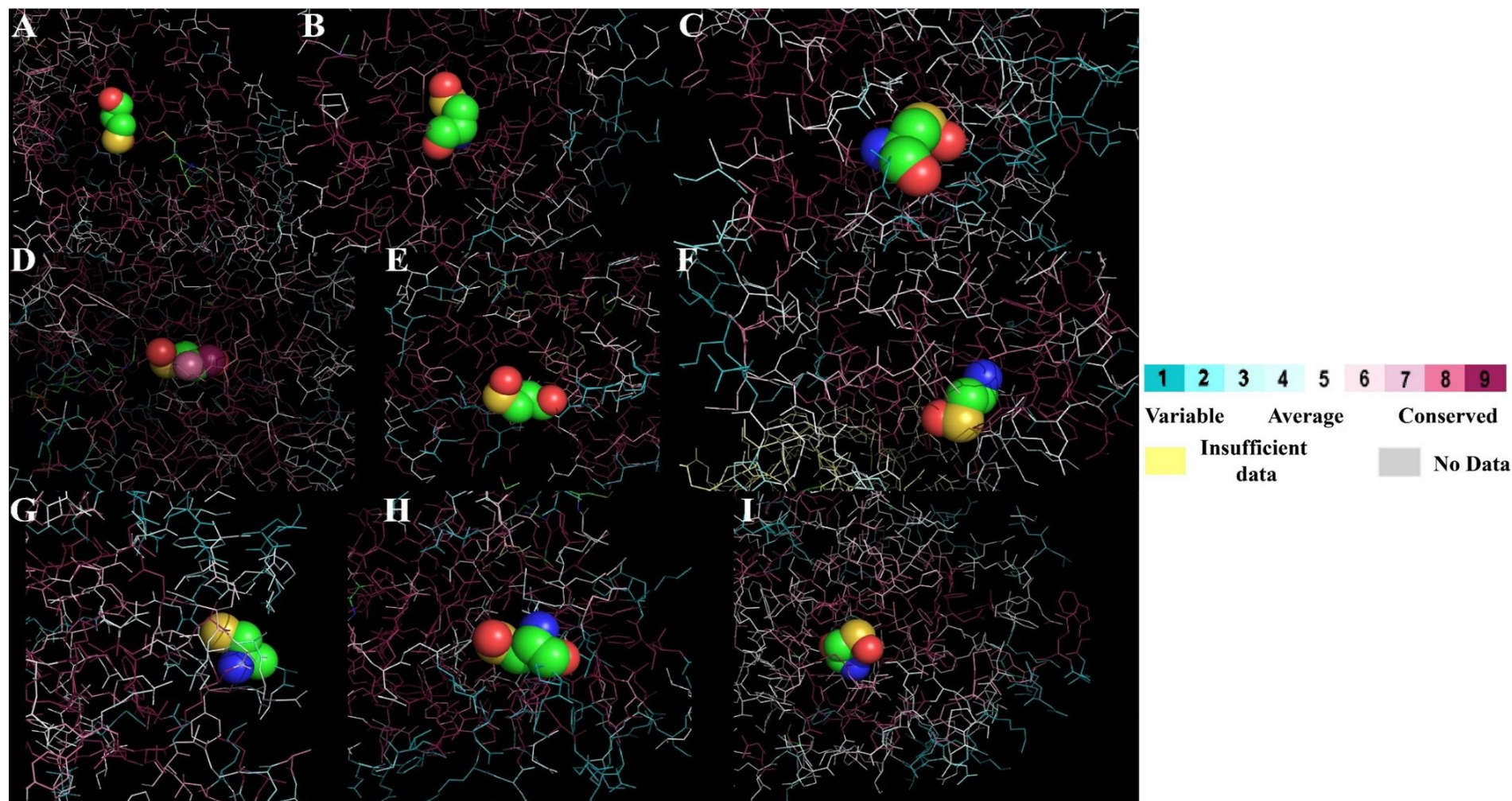
Amino acid	Window Position (CSO / CYS)										
	-5	-4	-3	-2	-1	0	1	2	3	4	5
A	9.8 / 6.7	7.6 / 9.4	5.7 / 7.5	9.1 / 8.7	8.7 / 7.2	0 / 0	6.3 / 5.6	9.1 / 7.5	5.3 / 8.7	11.5 / 7.2	5.3 / 7.6
C	2.9 / 1.8	1.0 / 2.0	2.9 / 4.4	2.4 / 2.1	1.9 / 1.2	100/ 100	0.9 / 1.5	0.5 / 2.4	1.9 / 5.3	0.0 / 2.3	0.9 / 2.3
D	6.8 / 5.3	3.4 / 7.5	4.8 / 5.8	7.2 / 4.2	4.3 / 4.3	0 / 0	5.2 / 5.2	2.4 / 6.4	3.3 / 4.7	4.3 / 5.5	6.2 / 6.3
E	5.3 / 6.0	6.3 / 3.6	7.2 / 5.2	5.7 / 7.8	1.4 / 4.9	0 / 0	7.5 / 5.6	5.2 / 6.4	7.1 / 3.6	7.1 / 3.7	4.7 / 5.5
F	4.3 / 3.6	5.3 / 4.1	4.3 / 3.9	5.7 / 3.0	3.3 / 4.0	0 / 0	2.8 / 3.1	2.8 / 4.0	4.8 / 3.3	2.4 / 3.8	4.7 / 4.8
G	6.3 / 9.8	5.8 / 7.5	9.1 / 8.2	7.2 / 8.4	11.0 / 10.0	0 / 0	9.9 / 10.4	12.3 / 9.1	13.8 / 10.0	6.2 / 11.1	9.5 / 8.1
H	2.4 / 2.6	3.9 / 3.3	3.4 / 1.9	1.4 / 1.9	4.8 / 2.2	0 / 0	3.8 / 3.3	4.3 / 1.8	3.3 / 2.7	4.7 / 3.5	2.4 / 3.2
I	6.8 / 5.3	5.8 / 6.3	6.3 / 5.2	5.3 / 4.9	7.7 / 7.3	0 / 0	9.0 / 5.5	6.6 / 4.9	4.8 / 4.9	6.6 / 4.3	5.2 / 5.5
K	3.4 / 5.9	5.3 / 3.5	2.4 / 5.1	4.3 / 4.3	4.3 / 3.7	0 / 0	2.8 / 3.4	2.8 / 5.2	5.2 / 4.4	2.8 / 4.0	4.3 / 4.8
L	13.0 / 10.6	10.6 / 9.6	11.5 / 8.2	6.7 / 10.3	8.6 / 8.6	0 / 0	7.5 / 9.1	6.6 / 9.4	12.4 / 11.1	13.3 / 8.5	12.3 / 9.8
M	0.5 / 2.3	1.9 / 2.7	1.4 / 1.9	4.3 / 2.5	2.4 / 1.8	0 / 0	1.4 / 1.5	1.9 / 1.8	0.0 / 1.4	2.8 / 3.0	3.8 / 3.4
N	3.4 / 3.6	2.9 / 5.4	4.8 / 4.6	5.7 / 5.1	5.7 / 5.7	0 / 0	1.9 / 4.2	2.8 / 4.8	2.4 / 3.6	4.7 / 3.7	3.8 / 4.0
P	4.8 / 3.3	2.9 / 5.1	2.4 / 3.6	5.3 / 4.3	4.3 / 3.3	0 / 0	8.0 / 4.0	7.6 / 5.1	5.2 / 5.8	4.3 / 4.9	5.7 / 5.4
Q	1.9 / 3.5	4.8 / 4.1	5.3 / 4.6	5.3 / 2.2	2.9 / 3.7	0 / 0	3.3 / 5.0	3.3 / 4.0	2.4 / 3.8	2.8 / 3.4	3.8 / 5.2
R	8.2 / 6.9	6.3 / 4.4	3.8 / 6.1	3.3 / 5.1	1.9 / 6.6	0 / 0	6.1 / 4.9	4.7 / 4.5	5.2 / 5.8	3.3 / 7.2	5.2 / 4.1
S	4.3 / 5.9	6.8 / 5.6	6.3 / 7.6	6.7 / 6.1	6.2 / 6.6	0 / 0	5.7 / 7.0	6.6 / 4.9	6.7 / 4.7	9.5 / 5.5	4.7 / 3.2
T	5.3 / 6.3	5.8 / 4.5	8.7 / 4.9	4.8 / 6.6	7.7 / 5.1	0 / 0	6.6 / 7.6	9.5 / 4.9	6.7 / 4.4	3.8 / 5.8	4.7 / 5.5
V	5.3 / 5.3	8.7 / 6.6	6.3 / 6.4	7.2 / 7.9	8.1 / 7.2	0 / 0	6.6 / 7.6	6.2 / 7.6	5.7 / 5.9	5.2 / 8.2	9.5 / 7.1
W	0.0 / 1.1	1.0 / 2.1	1.0 / 1.6	0.5 / 1.3	1.0 / 1.8	0 / 0	1.4 / 1.0	2.4 / 1.0	2.4 / 1.8	1.4 / 2.1	0.5 / 1.1
Y	5.3 / 4.2	3.9 / 2.7	2.4 / 3.3	1.9 / 3.3	3.8 / 4.8	0 / 0	3.3 / 4.5	2.4 / 4.3	1.4 / 4.1	3.3 / 2.3	2.8 / 3.1
Pearson's chi-squared test											
p-value	0.5086	0.3657	0.6108	0.3305	0.3719		0.6698	<b>*0.0247</b>	0.1228	<b>*0.0079</b>	0.9717



**Table S7.** Frequency plot of amino acid neighbors in the surrounding 3D environment of cysteines. Results present the frequency of appearance (%) of each amino acid at the indicated distance intervals (in Angstroms) from the CSO/CYS residues. Aminoacid distributions for each interval were analyzed by Pearson's chi-squared test of independence and statistically significant differences between CSO and CYS neighbors are presented for \*p<0.05.

Amino acid	Distance Intervals in Angstroms (CSO / CYS)									
	0.0-2.0	2.0-4.0	4.0-6.0	6.0-8.0	8.0-10.0	10.0-12.0	12.0-14.0	14.0-16.0	16.0-18.0	18.0-20.0
A	7.4 / 6.5	5.8 / 7.0	6.7 / 6.7	8.0 / 7.8	9.0 / 7.6	8.5 / 7.9	9.2 / 7.6	8.2 / 8.0	8.0 / 8.0	7.8 / 8.3
C	1.7 / 1.5	2.2 / 4.9	1.6 / 3.2	1.4 / 2.8	1.4 / 2.0	1.8 / 2.2	1.3 / 2.0	1.3 / 1.9	1.6 / 1.9	1.7 / 1.8
D	4.7 / 4.7	4.2 / 3.8	2.9 / 3.9	4.5 / 4.3	5.0 / 4.6	5.0 / 5.0	5.1 / 5.1	4.8 / 5.1	5.0 / 5.6	5.5 / 5.5
E	4.5 / 5.3	6.3 / 4.1	5.4 / 4.5	5.2 / 4.7	5.2 / 5.2	5.4 / 5.6	5.4 / 5.4	5.6 / 6.1	6.3 / 5.6	6.5 / 6.3
F	3.1 / 3.6	5.2 / 6.5	5.0 / 5.3	5.6 / 5.1	4.9 / 5.2	5.3 / 4.7	4.5 / 4.8	4.6 / 4.1	4.2 / 4.3	4.0 / 4.0
G	10.4 / 10.1	9.5 / 7.2	8.8 / 7.3	6.6 / 7.6	8.5 / 7.2	6.9 / 7.5	7.6 / 7.7	7.0 / 8.0	7.1 / 7.5	7.6 / 7.9
H	4.3 / 2.7	4.7 / 3.0	2.8 / 2.6	2.7 / 2.7	2.6 / 2.9	2.3 / 2.6	2.3 / 2.7	2.5 / 2.4	2.1 / 2.7	2.6 / 2.6
I	8.3 / 6.4	5.7 / 6.2	6.6 / 6.7	7.0 / 7.1	6.8 / 6.9	7.4 / 6.2	7.0 / 6.4	6.1 / 6.2	6.2 / 5.9	5.7 / 5.3
K	3.5 / 3.6	3.7 / 3.6	4.1 / 3.4	4.1 / 3.9	4.2 / 3.8	3.8 / 4.8	4.4 / 4.9	4.3 / 5.1	4.9 / 4.8	5.3 / 5.1
L	8.0 / 9.0	8.5 / 11.3	9.3 / 10.3	11.4 / 10.6	11.4 / 10.8	10.9 / 9.6	11.1 / 9.5	10.5 / 9.1	10.1 / 9.3	9.7 / 8.6
M	1.9 / 1.6	2.0 / 2.6	4.5 / 3.0	2.7 / 2.6	3.0 / 2.7	2.4 / 2.7	2.0 / 2.7	2.3 / 2.4	2.5 / 2.3	2.4 / 2.2
N	3.8 / 4.9	4.6 / 3.0	3.2 / 4.3	3.1 / 3.7	3.3 / 3.5	3.5 / 3.7	3.6 / 3.9	3.6 / 3.7	3.8 / 4.1	3.9 / 4.1
P	6.1 / 3.6	3.9 / 3.8	4.0 / 4.4	3.9 / 4.6	4.0 / 4.6	4.6 / 4.7	4.6 / 4.9	4.7 / 4.9	5.2 / 5.2	4.5 / 5.4
Q	3.1 / 4.4	3.6 / 3.1	4.8 / 3.5	2.8 / 3.3	3.1 / 3.5	2.8 / 3.6	3.1 / 3.5	3.6 / 3.6	3.3 / 3.4	3.6 / 3.8
R	4.0 / 5.7	5.4 / 5.2	4.8 / 5.2	5.2 / 5.2	4.1 / 5.2	4.7 / 5.3	4.5 / 4.9	5.8 / 5.3	4.8 / 5.3	5.9 / 5.4
S	5.9 / 6.7	6.5 / 4.8	6.1 / 5.6	6.1 / 5.1	5.0 / 4.9	5.6 / 5.3	5.2 / 5.4	5.7 / 5.8	5.5 / 5.8	5.4 / 5.7
T	7.3 / 6.3	5.9 / 4.8	5.9 / 4.7	5.6 / 5.0	5.2 / 5.3	4.7 / 4.9	5.3 / 5.5	5.3 / 6.0	5.1 / 5.6	4.7 / 5.5
V	7.3 / 7.3	5.9 / 7.8	8.6 / 8.6	7.3 / 7.8	7.8 / 8.2	8.6 / 7.9	8.6 / 7.4	8.2 / 6.8	9.2 / 7.2	7.8 / 7.2
W	1.2 / 1.4	2.1 / 2.5	1.5 / 2.2	1.9 / 2.1	2.0 / 2.1	1.5 / 2.2	1.9 / 1.8	1.8 / 1.7	1.5 / 1.7	1.6 / 1.7
Y	3.5 / 4.7	4.3 / 4.8	3.4 / 4.6	4.9 / 4.0	3.5 / 3.8	4.3 / 3.6	3.3 / 3.9	4.1 / 3.8	3.6 / 3.8	3.8 / 3.6
Pearson's chi-squared test										
p-value	0.6515	*0.0001	*0.0024	*0.0067	0.1682	*0.0004	*0.0001	*0.0017	*0.0015	0.1860

## Supplementary Figures



**Figure S1. ConSurf server analysis of evolutionary conserved amino acid positions in proteins.** Proteins selected from Table S1 with PDB ID's A) 1GSN; B) 1I9T; C) 2V5B; D) 2IBY; E) 3UBW; F) 2VRN; G) 3IQU; H) 2NQA and I) 3ZJE, were analyzed by Consurf Server as described in Section 1. Surface residues are shaded according to degree of conservation. The nine-color conservation scores are projected onto the 3D structure of the query protein and the colored residues are shown as sticks. The CSO residue is shown as space filling spheres. ConSurf results indicate that several amino acids are highly conserved in the 3D space surrounding the CSO residue. ConSurf server is freely available at <http://consurf.tau.ac.il/>.



**Figure S2. Conserved verified CSO sites in different species using the Clustal Omega server.** Multiple sequence alignment for the proteins: A) Inosine Monophosphate Dehydrogenase; B) Coronin 1A; C) 1-pyrroline-5-carboxylate dehydrogenase; D) protease 1; E) probable histone acetyltransferase MYST1; F) Nitrile hydratase subunit alpha; G) Rhodanese-like protein; H) peptide methionine sulfoxide reductase msra/msrb; I) 3-hydroxyanthranilate 3,4-dioxygenase; J) Aldehyde dehydrogenase, mitochondrial; K) Acetyl-CoA acetyltransferase; L) Peroxiredoxin-4, from Table S2. Verified CSO in species, as described in Section 1. Clustal Omega server is freely available at <http://www.clustal.org/omega/>.



**Figure S3. Conserved verified CSO sites in different species using the Clustal Omega server.** Multiple sequence alignment for the proteins: M) Protein Epsilon; N) receptor-type tyrosine-protein phosphatase S; O) arylamine N-acetyltransferase; P) BCL10-interacting card protein; Q) hydroxycinnamoyl-CoA shikimate/quinic, hydroxycinnamoyltransferase; R) mitogen-activated protein kinase 1; S) aldo-keto reductase family 1 member B10; T) cathepsin B-like peptidase (C01 family); U) probable manganese-dependent inorganic pyrophosphatase; V) proto-oncogene tyrosine-protein kinase SRC; W) protein (complement C3DG); X) Hemoglobin Beta chain; Y) Severin; Z) protein chain elongation factor EF-TU; AA) S-formylglutathione hydrolase; AB) transferase malate synthase A, from Table S2. Verified CSO in species, as described in Section 1.



sp|P05164|PERM\_HUMAN  
sp|P11247|PERM\_MOUSE

PRIKNQADCI PFFRSCPACPGSNITIRNQINALTSFVDASMYVGSEEP LARNLRNMSNQL  
 PRIKNQKDCI PFFRSCPACRNNITIRNQINALTSFVDASGVYGS EDP LARKLRNLTNQL  
 \*\*\*\*\* :\*\*\*\*\*:\*\*\*\*\*:\*\*\*\*\*:\*\*\*\*\*

sp|P00390|GSHR\_HUMAN  
sp|A2TIL1|GSHR\_CALJA  
sp|P47791|GSHR\_MOUSE  
sp|P70619|GSHR\_RAT

AVASYDYLVIGGGSGGLASARRAAELGARAADVESHKLGTCVNVGCVPKKVMWNTAVHS  
 AVVSYDYLVIGGGSGGLASARRAAELGARAADVESHKLGTCVNVGCVPKKVMWNTAVHS  
 DTSSFYDLVIGGGSGGLASARRAAELGARAADVESHKLGTCVNVGCVPKKVMWNTAVHS  
 -----VNVGCVPKKVMWNTAVHS  
 \*\*\*\*\*

```
sp|O13016|PTN1_CHICK
sp|P18031|PTN1_HUMAN
sp|P20417|PTN1_RAT
sp|P35821|PTN1_MOUSE
```

DFGVPEPASFLNFLFKVRESGSLNPEYGPVVVHCSAGIGRSGTFCFLVDTCLLMLDKRKD  
DFGVPEPASFLNFLFKVRESGSLSPHEGPPVVHCSAGIGRSGTFCFLADTCLLMLDKRKD  
DFGVPEPASFLNFLFKVRESGSLSPHEGPIVVHCSAGIGRSGTFCFLADTCLLMLDKRKD  
DFGVPEPASFLNFLFKVRESGSLSPHEGPVVHCSAGIGRSGTFCFLADTCLLMLDKRKD  
\*\*\*\*\* \* \*

sp|A1A547|PGRP3\_MOUSE  
tr|D3ZYV3|D3ZYV3\_RAT  
sp|Q96LB9|PGRP3\_HUMAN  
tr|E1BJ76|E1BJ76\_BOVIN

AQSLIQCAVAKGYLTSNYLLMGHSDVSNILSPGQALYNIKTWPHFKH  
AQS~~L~~IQCAVAVEGYLASNLYLLMGHSDVSNILSPGQALYNIKTWPHFKH  
AQD~~L~~IQCAVVEGYLTSPNYLLMGHSDVVNLSPGQALYNIISTWPHFKH  
AQLTHCSVKGYLPVNYYLVGHSDVTDLSPGRALYNIKTWPHFRQ  
\*\* . \* : \* : \* . \*\*\* . \*\*\*\* : \*\*\*\*\* : \*\*\*\*\* : \*\*\*\*\* : \*\*\*\*\* ::

```
tr|A0A0A9ZAM7|A0A0A9ZAM7_LYGHE
tr|E1C234|E1C234_CHTCK
sp|Q8R0F3|SUMF1_MOUSE
tr|D4A7I8|D4A7I8_RAT
sp|Q8NBK3|SUMF1_HUMAN
tr|F1P978|F1P978_CANLF
sp|Q0P5L5|SUMF1_BOVIN
```

```

GGSFMCNHSFCNRYRCSRGRYHSSSTHHGQTNTGFRVCVKSLPVY--
GGSYMCHKSYCYRYYRCAARSQNTPDSSASNLGFRCAADALPDPQ
GGSYMCHKSYCYRYYRCAARSQNTPDSSASNLGFRCAADHLPDPL
GGSYMCHKSYCYRYYRCAARSQNTPDSSASNLGFRCAADHLPATAN
GGSYMCHRSYCYRYYRCAARSQNTPDSSASNLGFRCAADRLPTMD
GGSYMCHKSYCYRYYRCAARSQNTPDSSASNLGFRCAADRQPTTG
GGSYMCHKSYCYRYYRCAARSQNTPDSSASNLGFRCAADHLPPTG
***:***:.* *****.* :.: :.:*****.

```

sp|P30306|MPIP2\_MOUSE  
sp|P30305|MPIP2\_HUMAN  
tr|E1BL76|E1BL76\_BOVIN

TMVALLTGKFSNIVEKFVIVDCRYPY EYEGGHIKAVNLP LERDAETFL LQR PIMPCSLD  
TMVALLTGKFSNIVDKFVIVDCRYPY EYEGGHIKTAVNLP LERDAESFL LKS PIAPCSLD  
TVVALLTGKFSHIVEKFVIVDCRYPY EYEGGHIKTAVNLP LERDAETFL LQS PITPCSLD  
\* \* \* \* \*

```
sp|P01112|RASH_HUMAN
sp|Q61411|RASH_MOUSE
sp|P20171|RASH_RAT
sp|P08642|RASH_CHICK
```

MTEYKLVVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILD  
TAG  
MTEYKLVVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILD  
TAG  
MTEYKLVVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILD  
TAG  
MTEYKLVVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILD  
TAG  
\*\*\*\*\*

sp|Q6AZA0|THIL\_DANRE  
sp|Q6NU46|THILA\_XENLA  
sp|Q8QZT1|THIL\_MOUSE  
sp|P17764|THIL\_RAT  
sp|P24752|THIL\_HUMAN  
sp|Q29RZ0|THIL\_BOVIN

TINKVCASGMKSIIMLASQSLMCGHQDVMVAGGMESMSQVVPY-----IMAREAPPYGGVKM  
TINKVCASGMKSVMLAAQSLMCGHQDVMVAGGMESMSNPY-----CMSRGATPYGGVKL  
TVNKVCASGMKIMMASQSLMCGHQDVMVAGGMESMSNPY-----VMSRGATPYGGVKL  
TVNKVCASGMKAIMMASQSLMCGHQDVMVAGGMESMSNPY-----VMSRGATPYGGVKL  
TINKVCASGMKAIMMASQSLMCGHQDVMVAGGMESMSNPY-----VMNRGSTPYGGVKL  
TINKVCASGMKAIMMASQSLMCGHQDVMVAGGMESMSNPY-----VMNRGATPYGGVKL  
\* : \* \* \* \* \* : \* : \* : \* \* \* : \* : \* : \* \* \* : \* : \* : \* \* \* : \* : \*

```
tr|O93417|O93417_CHICK
tr|F1MB04|F1MB04_BOVIN
sp|P42574|CASP3_HUMAN
sp|P70677|CASP3_MOUSE
sp|P55213|CASP3_RAT
```

[illegible]

sp|Q80XL1|CBLC\_MOUSE  
sp|G3V8H4|CBLC\_RAT  
sp|Q9ULV8|CBLC\_HUMAN  
tr|A6QPZ5|A6QPZ5\_BOVI

ACHPVEPGPTMQALRSTLDLTCSGHVSVEFFDVFTRLFPWPPTLLRNWQLLAVNHPGYMA  
 SCHPVEPGPTIMQALRSTLDLTCSGHVSVEFFDIFTRLFPWPPTLLKNWQLLAVNHPGYMA  
 TCHPVEPGCTALALRTTIDLTCSGHVSIFEFDVFTRLFPWPPTLLKNWQLLAVNHPGYMA  
 ICHPVEPGSTALARSTIDLTCSGHVSIFEFDIFTRLFPWPPTLLKNWQLLAVNHPGYMA  
 \*\*\*\*\*  
 \*\*\*:\*.\*:\*.\*:\*.\*:\*.\*:\*.\*:\*.\*:\*.\*:\*.\*:\*.\*:\*.\*:\*.\*:\*.\*:\*.\*:\*.\*:\*.\*:

sp|Q9TV66|HCN4\_RABBIT  
sp|Q9Y3Q4|HCN4\_HUMAN  
sp|O70507|HCN4\_MOUSE  
sp|Q9JKA7|HCN4\_RAT

QVEQYMSFHKLPPDTRQRIHDYYEHRYQGKMFDDEESILGELSEPLREEIINFNCRKLVAS  
QVEQYMSFHKLPPDTRQRIHDYYEHRYQGKMFDDEESILGELSEPLREEIINFNCRKLVAS  
QVEQYMSFHKLPPDTRQRIHDYYEHRYQGKMFDDEESILGELSEPLREEIINFNCRKLVAS  
QVEQYMSFHKLPPDTRQRIHDYYEHRYQGKMFDDEESILGELSEPLREEIINFNCRKLVAS  
\*\*\*\*\*

**Figure S4. Conserved non-verified redox sensitive cysteines in other species using the Clustal Omega server.** Multiple sequence alignment for the proteins: A) Myeloperoxidase; B) Glutathione reductase, mitochondrial; C) Tyrosine-protein phosphatase non-receptor type 1; D) Peptidoglycan recognition protein 3; E) Sulfatase-modifying factor 1; F) M-phase inducer phosphatase 2; G) GTPase Hras; H) Acetyl-CoA acetyltransferase, mitochondrial; I) Caspase-3; J) E3 ubiquitin-protein ligase CBL-C; K) Potassium/sodium hyperpolarization-activated cyclic nucleotide-gated channel 4, from Table S2. Conserved CYS in species, as described in Section 1.



sp Q6DE87 CHK1_XENLA	LH
sp Q8AYC9 CHK1_CHICK	LH
sp Q14757 CHK1_HUMAN	LH
sp Q35280 CHK1_MOUSE	LH
sp Q91ZN7 CHK1_RAT	LH

sp|P51004|PAPO1\_XENLA  
sp|Q61183|PAPOA\_MOUSE  
sp|P25500|PAPOA\_BOVIN  
sp|P51003|PAPOA\_HUMAN

```
sp|P25324|THTR_CHICK
sp|Q16762|THTR_HUMAN
sp|P00586|THTR_BOVIN
sp|P24329|THTR_RAT
sp|P52196|THTR_MOUSE
```

tr|A0A0A1WRH1|A0A0A1WRH1\_BACCU  
sp|Q6NY98|MCE1\_DANRE  
tr|Q9IA93|Q9IA93\_XENLA  
sp|O55236|MCE1\_MOUSE  
sp|O60942|MCE1\_HUMAN

sp|Q13363|CTBP1\_HUMAN  
sp|Q9Z2F5|CTBP1\_RAT  
sp|O88712|CTBP1\_MOUSE

tr|E1BZ59|E1BZ59\_CHICK  
tr|E1BKC3|E1BKC3\_BOVIN  
sp|P21580|TNAP3\_HUMAN  
sp|Q60769|TNAP3\_MOUSE

sp|Q9WVM8|AADAT\_MOUSE  
sp|Q64602|AADAT\_RAT  
sp|Q8N5Z0|AADAT\_HUMAN  
sp|Q5E9N4|AADAT\_BOVIN  
tr|E1C9H5|E1C9H5\_CHICK  
tr|E7F5G8|E7F5G8\_DANRE

sp|O88622|PARG\_MOUSE  
sp|Q9QYM2|PARG\_RAT  
sp|Q86W56|PARG\_HUMAN  
sp|O02776|PARG\_BOVIN

sp|P31947|1433S\_HUMAN  
sp|O70456|1433S\_MOUSE  
sp|Q0VC36|1433S\_BOVIN  
sp|O77642|1433S\_SHEEP

```
sp|P30291|WEE1_HUMAN
tr|A7MBC3|A7MBC3_BOVIN
sp|P47810|WEE1_MOUSE
sp|Q63802|WEE1_RAT
```

sp|Q9D289|TPC6B\_MOUSE  
sp|Q86SZ2|TPC6B\_HUMAN  
sp|Q32L78|TPC6B\_BOVIN  
tr|D3ZES2|D3ZES2\_RAT  
tr|Q5XJB3|Q5XJB3\_DANRE

LHSGITHRDIKPENLLD~~ERD~~QLKISDFGLATVFRHNGKERLLNKMCGTLPYVAP~~ELIK~~  
LHSMGITHRDLKPENLLD~~ERD~~NLKISDFGLATVFKHNGRERLLNKMCGTLPYVAP~~ELLR~~  
LHGIGITHRDIKPENLLD~~ERD~~NLKISDFGLATVFRYNNRERLLNKMCGTLPYVAP~~ELLK~~  
LHGIGITHRDIKPENLLD~~ERD~~NLKISDFGLATVFRHNNRERLLNKMCGTLPYVAP~~ELLK~~  
LHGIGITHRDIKPENLLD~~ERD~~NLKISDFGLATVFRHNNRERLLNKMCGTLPYVAP~~ELLK~~  
\* \* \* \* \*

[illegible]

SKPLTATCRKGVTACHIALAAYLCGKPDVAVDGWSWSEFHRAPPQYKVTTELK---  
 SQPLIATCRKGVTACHVALAAYLCGKPDVAVDGWSWSEFRRAPPESRVSQGKSEKA  
 TKPLIATCRKGVTACHIALAAYLCGKPDVAIDGWSWFEFHRAPPETWVSQGKGGKA  
 SQPLIATCRKGVTACHIALAAYLCGKPDVAVDGWSWSEFRRAPPETRVVSQGKSGKA  
 SQPLIATCRKGVTACHVALAAYLCGKPDVAVDGWSWSEFRRAPPETRVVSQGKSGKA  
 \* \* \* \* \*

U ERFFEIVGVHCTHGFNRTGFLIASYLVRLDYSIEAALAIFAEARPPGIYKQDYIDELFR  
KTPTELIGVHCTHGFNRTGFLICAYLVEKMDWSIEAAVAFAQARPPGIYKGDYIKELFR  
RNPTELIGVHCTHGFNRTGFLICAFVLEKMDWSIEAAVATFAQARPPGIYKADYIKELFR  
RSPELIGVHCTHGFNRTGFLICAFVLEKMDWSIEAAVATFAQARPPGIYKGDYIKELFR  
RNPPTELIGVHCTHGFNRTGFLICAFVLEKMDWSIEAAVATFAQARPPGIYKGDYIKELFR  
\* \* \* \* \*

MGSSHLNKGKPLGVRPPIMNGPLHPRPLVALLDGRDCTVEMPILKDVATVAFCDASTQ  
-----MSGVRPPIMNGPMHPRPLVALLDGRDCTVEMPILKDVATVAFCDASTQ  
MGSSHLNKGKPLGVRPPIMNGPMHPRPLVALLDGRDCTVEMPILKDVATVAFCDASTQ  
\*\*\*\*\*

[illegible][illegible]

SKSEDRRKEQCEVRHQRTERKIPKYIPPNLPPEKKWLGTPIEEMRKM**PRCGIHL**PSLRPS  
PKSEDRRKEQCEVRHQRAERKIPKYVPPNLPDKKWLGTPIEEMRKM**PRCGVRL**PLLRPS  
PKAEDRRKEQWETKHQRTERKIPKYVPPHLPDKKWLGTPIEEMRRM**PRCGIRL**PLLRPS  
PKAEDKRKEQCEMKHQRTERKIPKYIPPHLSPDKKWLGTPIEEMRRM**PRCGIRL**PPLRPS

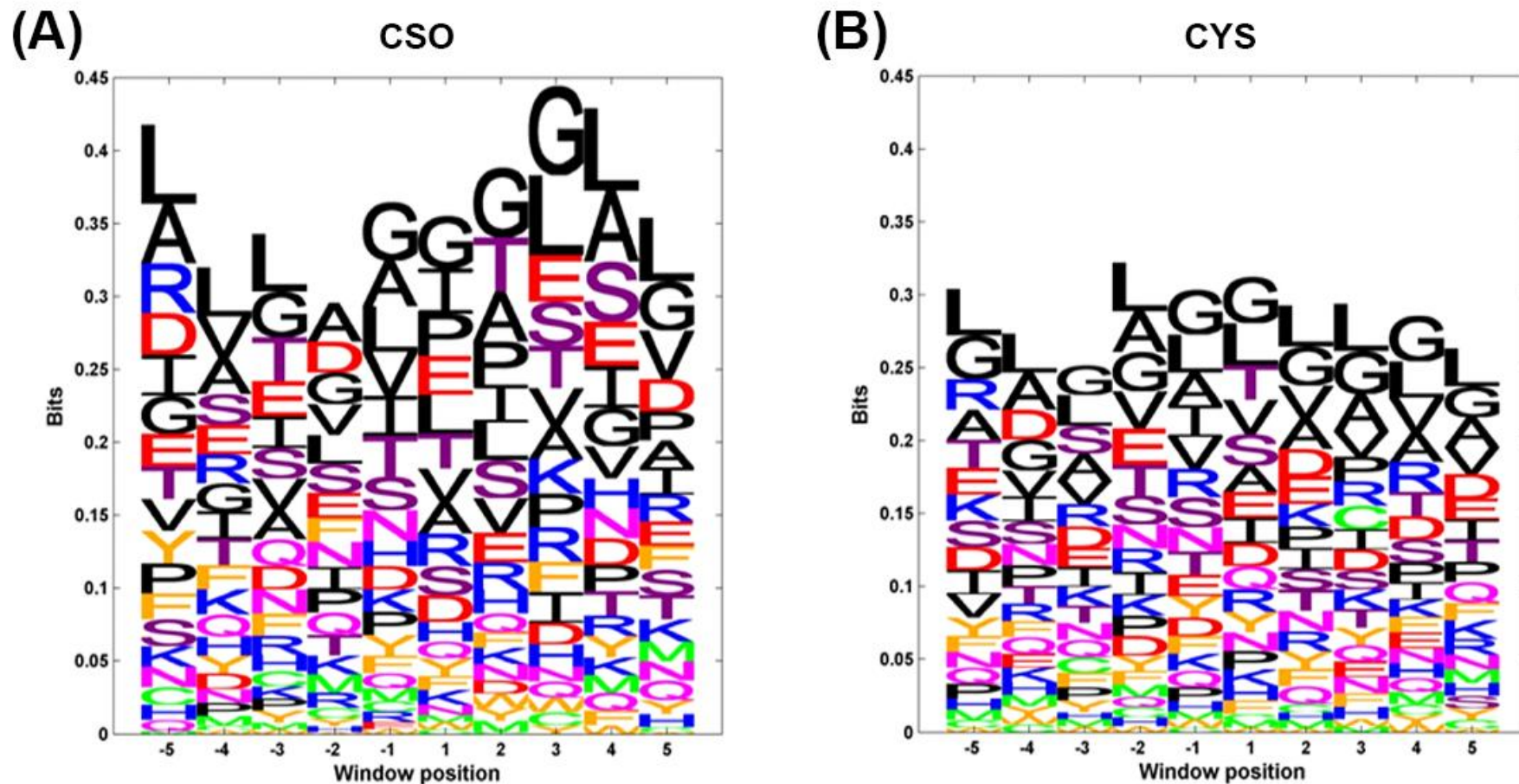
MERASLIQAKLAEQAERYEDMAAFMKGAVEKGEELSCEERNLLSVAYKNVVGGRAAWR  
MERASLIQAKLAEQAERYEDMAAFMKSAVEKGEELSCEERNLLSVAYKNVVGGRAAWR  
MERASLIQAKLAEQAERYEDMAAFMKSAVEKGEELSCEERNLLSVAYKNVVGGRAAWR  
MERASLIQAKLAEQAERYEDMAAFMKSAVEKGEELSCEERNLLSVAYKNVVGGRAAWR  
\*\*\*\*\*

SRFLANEVLQENYTHLPKADIFALALTVVCAAGAEPLPRNGDQWHEIRQGRLPRIQVLS  
SRFLANEVLQENYTHLPKADIFALALTVVCAAGAEPLPRNGDQWHEIRQGRLPRIQVLS  
SRFLANEVLQENYSHLPKADIFALALTVVCAAGAEPLPRNGEQWHEIRQGRLPRIQVLS  
SRFLANEVLQENYSHLPKADIFALALTVVCAAGAEPLPRNGDQWHEIRQGRLPRIQVLS  
\*\*\*\*\*

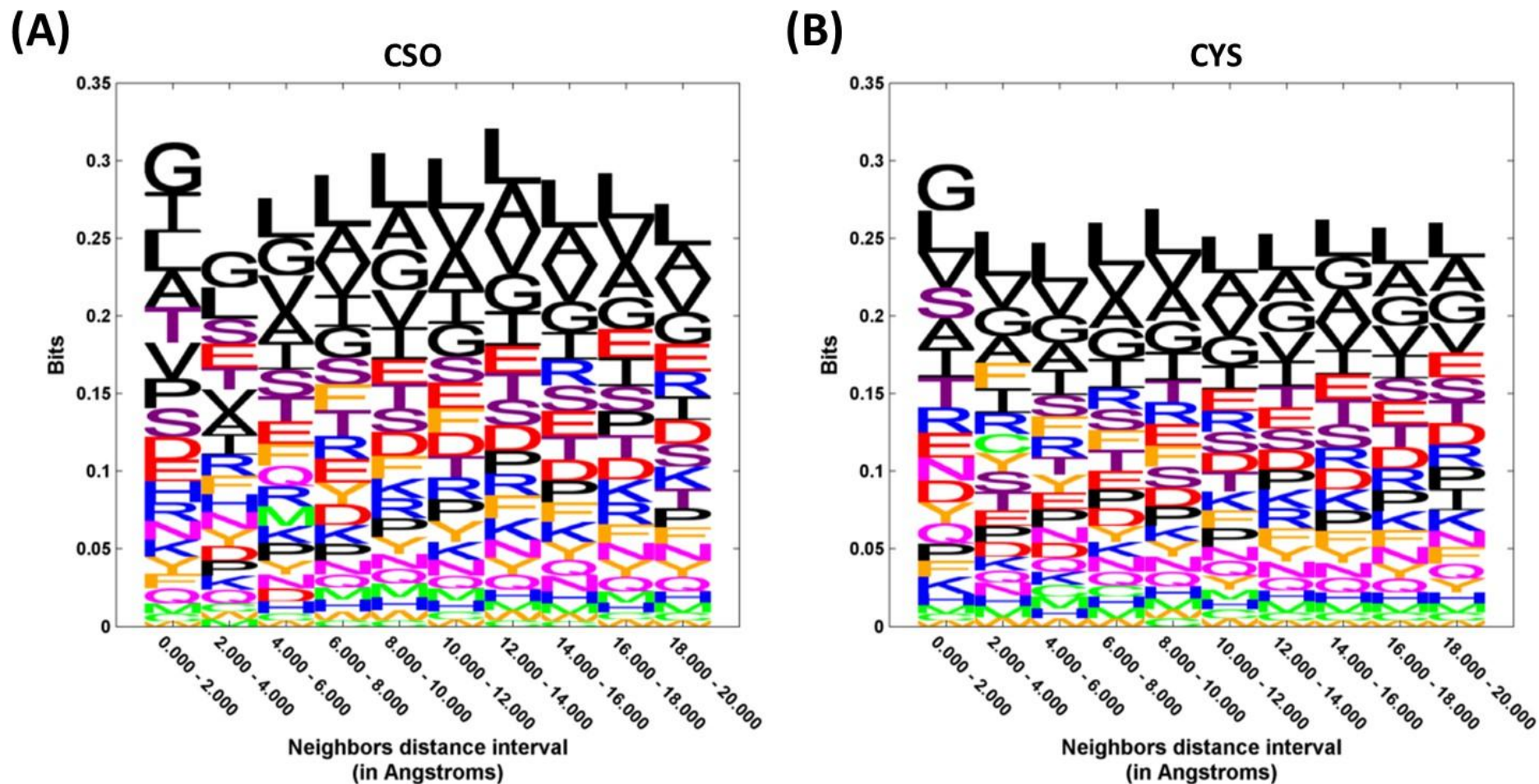
[illegible]

**Figure S5. Conserved non-verified redox sensitive cysteines in other species using the Clustal Omega server.** Multiple sequence alignment for the proteins: L) Serine/threonine-protein kinase Chk1; M) Poly(A) polymerase alpha; N) Thiosulfate sulfurtransferase; O) mRNA-capping enzyme; P) C-terminal-binding protein 1; Q) Tumor necrosis factor alpha-induced protein 3; R) Kynurenine/alpha-aminoadipate aminotransferase, mitochondria; S) Poly(ADP-ribose) glycohydrolase; T) 14-3-3 protein sigma; U) Wee1-like protein kinase; V) Trafficking protein particle complex subunit 6B, from Table S2. Conserved CYS in species, as described in Section 1.



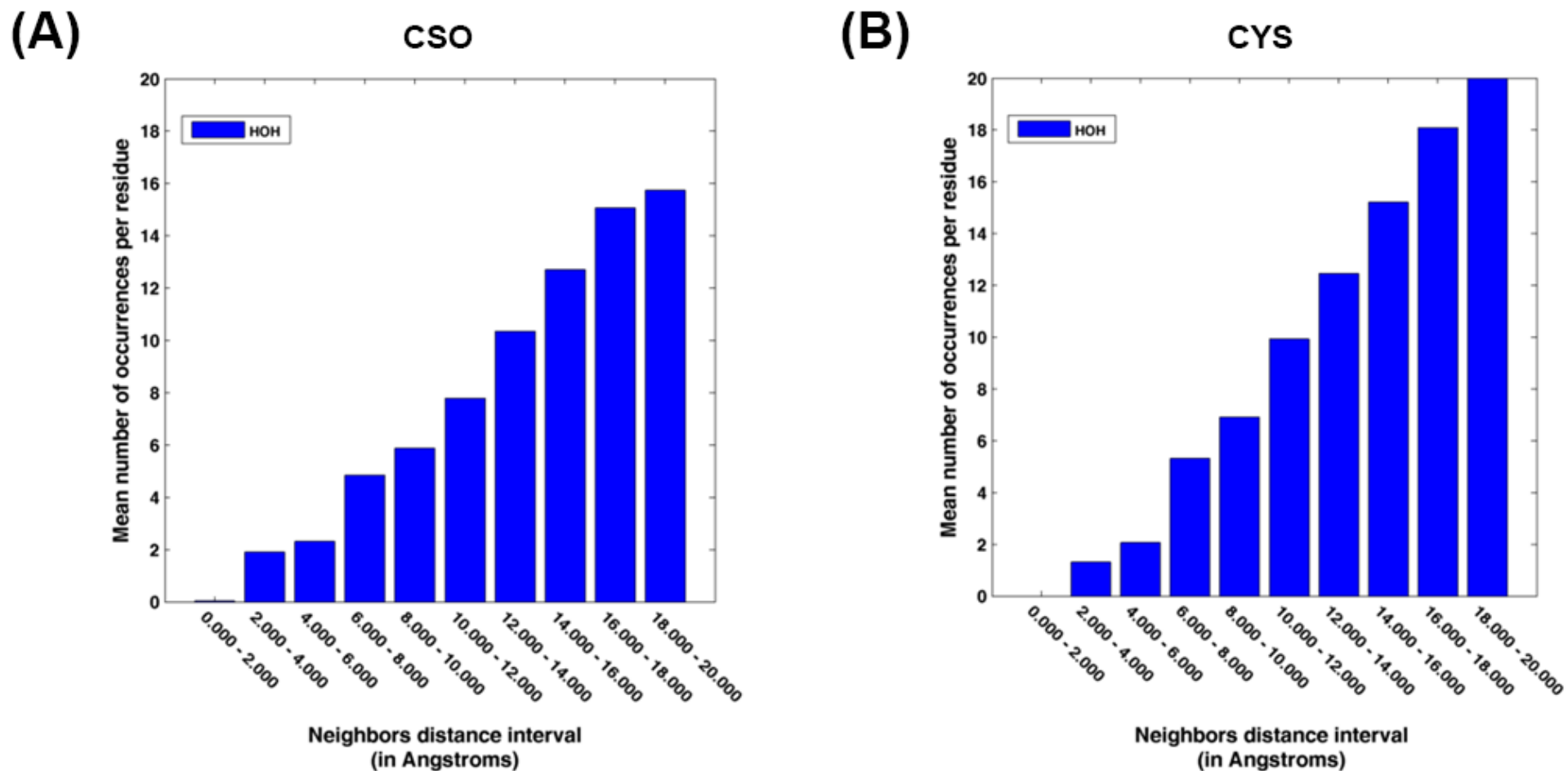


**Figure S6.** Sequence logos of the neighboring amino acid residues of cysteines, in the primary protein sequence. (A) Sequence logo of cysteines that have undergone S-sulfenylation (CSO) and (B) that are *not* prone to S-sulfenylation (CYS). The relative size of the letters indicates their frequency in each position. The total height of each stack of letters (y axis), depicts the entropy of each position, expressed in bits of information. The higher the entropy (i.e. the less the information content) the less the height of the stack. Amino acids are colored according to their chemical properties: aliphatic residues (A, G, I, L, P, V) are colored in black; aromatic residues (F, W, Y) are colored in orange; acidic (D, E) residues in red; basic residues (R, K, H) in blue; hydroxylic residues (S, T) in purple; sulfur-containing residues (C, M) in green and amidic residues (N, Q) are colored in magenta. Sequence logos (Schneider and Stephens, 1990) were produced using our own MatLab implementation.

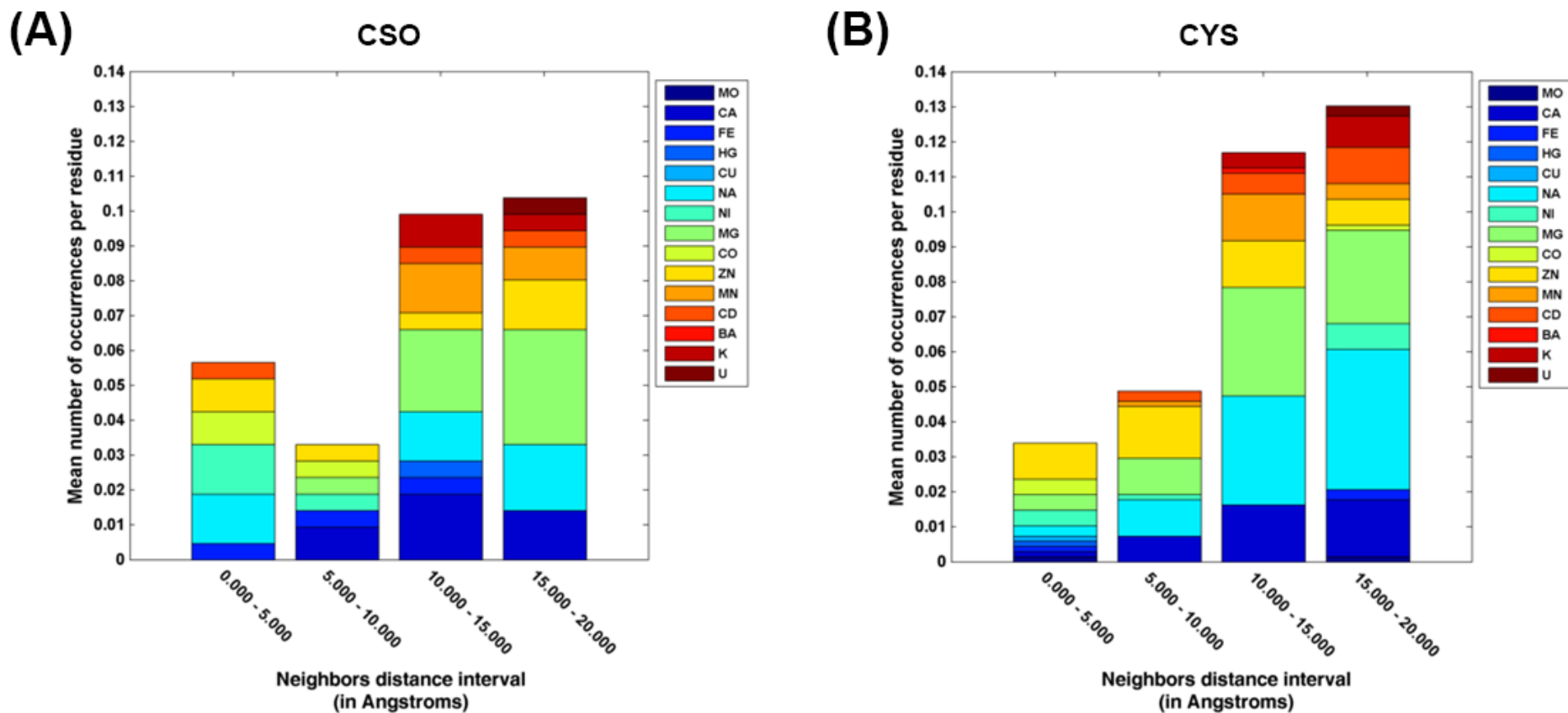


**Figure S7.** Sequence logos of the neighboring amino acid residues of cysteines, in the surrounding 3D environment. (A) Sequence logo of cysteines that have undergone S-sulfenylation (CSO) and (B) that are *not* prone to S-sulfenylation (CYS). The relative size of the letters indicates the frequency of each amino acid at the indicated distance intervals (in Angstroms) from the cysteine residues. The total height of each stack of letters (y axis), depicts the entropy of each position, expressed in bits of information. The higher the entropy (i.e. the less the information content) the less the height of the stack. Amino acids are colored according to their chemical properties: aliphatic residues (A, G, I, L, P, V) are colored in black; aromatic residues (F, W, Y) are colored in orange; acidic (D, E) residues in red; basic residues (R, K, H) in blue; hydroxylic residues (S, T) in purple; sulfur-containing residues (C, M) in green and amidic residues (N, Q) are colored in magenta. Sequence logos (Schneider and Stephens, 1990) were produced using our own MatLab implementation.

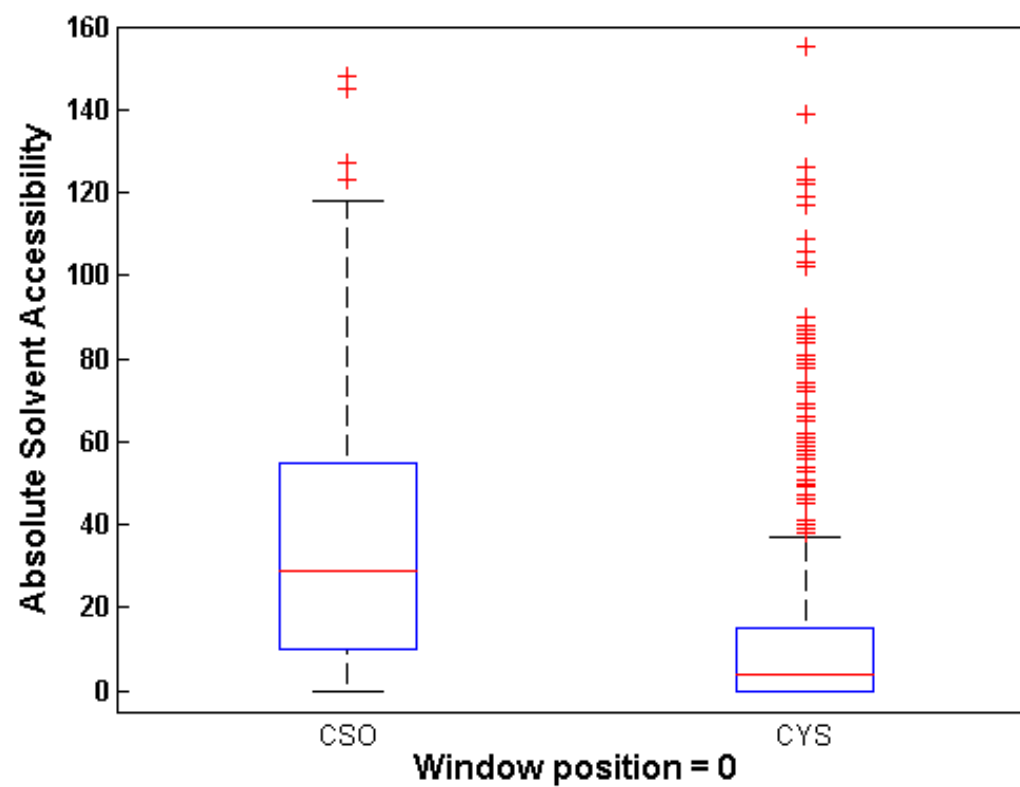




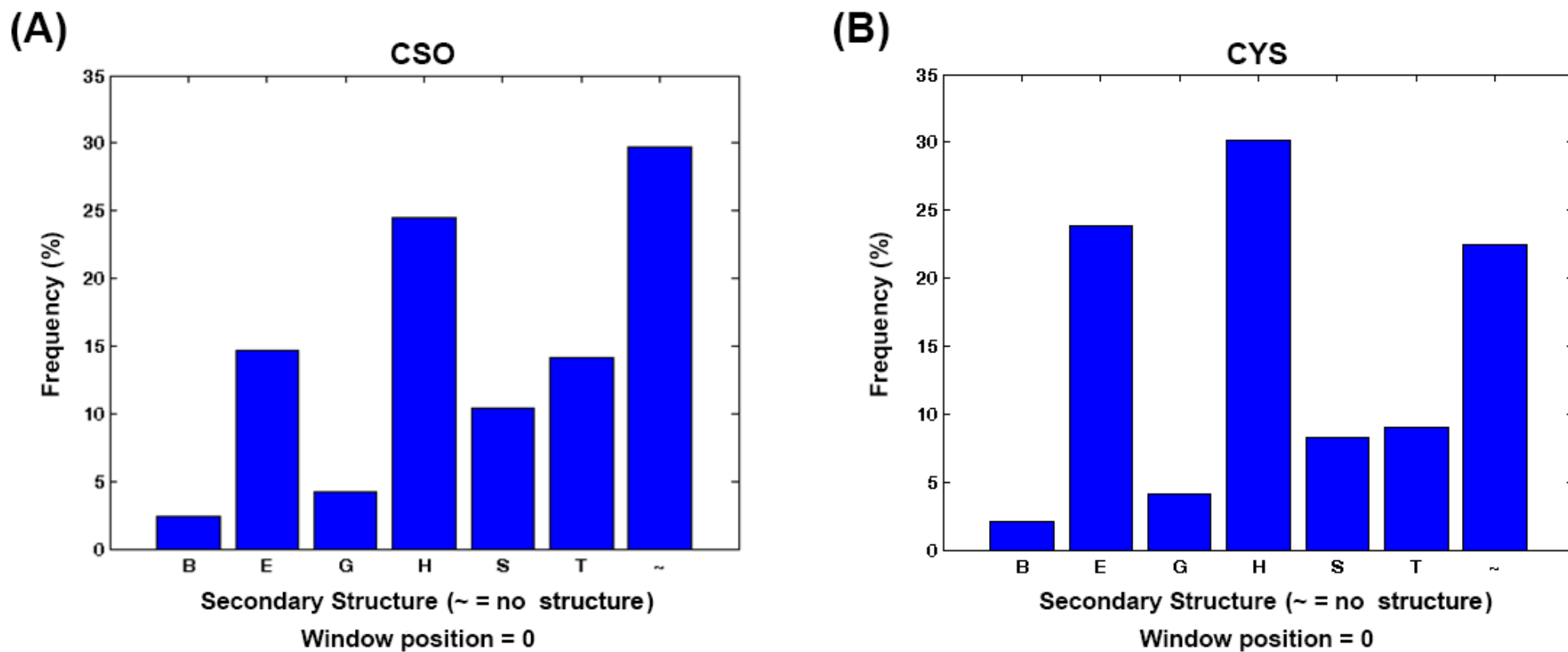
**Figure S8.** The frequency of water (HOH) molecules in the surrounding 3D environment of cysteines. (A) Water molecules in cysteines that have undergone S-sulfenylation (CSO) and (B) that are *not* prone to S-sulfenylation (CYS). Bars present the frequency of water molecules at the indicated distance intervals (in Angstroms) from the cysteine residues.



**Figure S9.** The frequency of metal ions in the surrounding 3D environment of cysteines. **(A)** Metal ions in cysteines that have undergone S-sulfenylation (CSO) and **(B)** that are *not* prone to S-sulfenylation (CYS). Coloured bars present the frequency of metal ions at the indicated distance intervals (in Angstroms) from the cysteine residues. The corresponding metals are depicted in the left inset.



**Figure S10.** Solvent accessibility of protein cysteines that have undergone S-sulfenylation (CSO) and cysteines that are *not* prone to S-sulfenylation (CYS), calculated using the DSSP software.



**Figure S11:** Secondary structural features of protein cysteines. The frequency (%) of cysteines, **(A)** that have undergone S-sulfenylation (CSO) and **(B)** that are *not* prone to S-sulfenylation (CYS), in different secondary protein structures was calculated using the DSSP software (B = beta bridge, E = beta bulges, G =  $3_{10}$  helix, H =  $\alpha$  helix, S = regions of high curvature, T = turn, ~ = no structure).

**(A)**

**PRESS** | *PROtein S- Sulfenylation server* Home Upload

## Result.

Operation completed successfully!!!  
o files contain warnings  
o files contain errors

For more details see the log file.  
[Download log file](#)

Prediction:

```
>5czh_C775 ( A )  
DNPHVCRLGI  
Predicted class -> CYS  
  
>5czh_C781 ( A )  
RLLGICLTSTV  
Predicted class -> CYS  
  
>5czh_C797 ( A )  
LMPFGCLLDYV  
Predicted class -> CSO
```

[Download prediction file](#)

**(B)**

**PRESS** | *PROtein S- Sulfenylation server* Home Upload

## Result.

Operation completed successfully!!!  
o files contain warnings  
o files contain errors

For more details see the log file.  
[Download log file](#)

Prediction:

```
>3hju_C32 ( A )  
GQYLFCRYWKP  
Predicted class -> CYS  
  
>3hju_C201 ( A )  
SDPLICRAGLK  
Predicted class -> CSO  
  
>3hju_C208 ( A )  
AGLKVCFGIQL  
Predicted class -> CSO
```

[Download prediction file](#)

**Figure S12.** (A) PRESS prediction test on the human EGF receptor which was experimentally verified to be S-sulfenylated on CYS 797 (Paulsen, et al., 2012) using dimedone-based CSO labeling. Successful prediction by PRESS of 5 cysteines and one S-sulfenylation site in CYS 797 of the human EGF receptor with pdbid: 5czh. (B) PRESS prediction test on the human monoacylglyceride lipase with pdbid 3hju containing 4 cysteines. PRESS predicts that CYS 201 and 208 are prone to S-sulfenylation, which was, very recently, experimentally verified by Dotsey, et al., 2015.

## References

- Akter, S., *et al.* (2015) DYn-2 Based Identification of Arabidopsis Sulfenomes, *Molecular & cellular proteomics : MCP*, **14**, 1183-1200.
- Ashkenazy, H., *et al.* (2010) ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids, *Nucleic acids research*, **38**, W529-533.
- Berman, H.M., *et al.* (2000) The Protein Data Bank, *Nucleic acids research*, **28**, 235-242.
- Bishop, C.M. (2006) *Pattern recognition and machine learning*. Information science and statistics. Springer, New York.
- Chang, C.C. and Lin, C.J. (2011) LIBSVM: A library for support vector machines, *ACM Transactions on Intelligent Systems and Technology*, **3**, 1-27.
- Cortes, C. and Vapnic, V. (1995) Support-vector networks, *Machine Learning*, **20**, 273-297.
- Glaser, F., *et al.* (2003) ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information, *Bioinformatics*, **19**, 163-164.
- Gould, N.S., *et al.* (2015) Site-Specific Proteomic Mapping Identifies Selectively Modified Regulatory Cysteine Residues in Functionally Distinct Protein Networks, *Chemistry & biology*, **22**, 965-975.
- Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers*, **22**, 2577-2637.
- Kelley, L.A., *et al.* (2015) The Phyre2 web portal for protein modeling, prediction and analysis, *Nature protocols*, **10**, 845-858.
- Landau, M., *et al.* (2005) ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures, *Nucleic acids research*, **33**, W299-302.
- Paulsen, C.E., *et al.* (2012) Peroxide-dependent sulfenylation of the EGFR catalytic site enhances kinase activity, *Nature chemical biology*, **8**, 57-64.
- Reddie, K.G., *et al.* (2008) A chemical approach for detecting sulfenic acid-modified proteins in living cells, *Molecular bioSystems*, **4**, 521-531.
- Schneider, T.D. and Stephens, R.M. (1990) Sequence logos: a new way to display consensus sequences, *Nucleic acids research*, **18**, 6097-6100.
- Sievers, F., *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega, *Molecular systems biology*, **7**, 539.
- Sun, M.A., *et al.* (2012) RedoxDB--a curated database for experimentally verified protein oxidative modification, *Bioinformatics*, **28**, 2551-2552.
- UniProt and Consortium (2015) UniProt: a hub for protein information, *Nucleic acids research*, D204-212.
- Yang, J., *et al.* (2014) Site-specific mapping and quantification of protein S-sulphenylation in cells, *Nature communications*, **5**, 4776.